

Building a Precise Video Language with Human–AI Oversight

Siyuan Cen^{1*} Hewei Wang^{1*} Chancharik Mitra^{1*} Isaac Li¹ Yuhan Huang¹ Yu Tong Tiffany Ling¹
Irene Pi¹ Shihang Zhu¹ Yili Han¹ Yilun Du² Deva Ramanan¹ Zhiqiu Lin¹

¹Carnegie Mellon University ²Harvard University

Abstract

Video–language models (VLMs) learn to reason about the dynamic visual world through natural language. We introduce a suite of open datasets, benchmarks, and recipes for scalable oversight that enable precise video captioning. First, we define a structured specification for describing subjects, scenes, motion, spatial, and camera dynamics, grounded by hundreds of carefully defined visual primitives developed with professional video creators such as filmmakers. Next, to curate high-quality captions, we introduce a critique-based human–AI (CHAI) oversight framework, where trained human experts provide correctional critiques to revise model-generated pre-captions into improved post-captions. This division of labor improves annotation accuracy and efficiency by offloading text generation to models, allowing humans to better focus on verification. Additionally, these critiques and preferences between pre- and post-captions provide rich supervision for improving open-source models (Qwen3-VL) on caption generation, reward modeling, and critique generation through SFT, DPO, and inference-time scaling. Our ablations show that critique quality in precision, recall, and constructiveness, ensured by our oversight framework, directly governs downstream performance. With modest expert supervision, the resulting model outperforms even closed-source models such as Gemini-3.1-Pro. Finally, we apply our approach to re-caption large-scale professional videos (e.g., films, commercials, games) and fine-tune video generation models such as Wan to better follow detailed prompts of up to 400 words, achieving finer control over cinematography including camera motion, angle, lens, focus, point of view, and framing. Overall, our results show that precise specification and human–AI oversight are key to achieving professional-level video understanding and generation.

1. Introduction

The limits of my language mean the limits of my world.

— Ludwig Wittgenstein, *Tractatus Logico-Philosophicus* [66]

Video–language models (VLMs) learn a world model of

visual dynamics through language supervision that teaches them *what* is present, *where* it is, and *how* it unfolds over time. Prior work [2, 4, 19, 46] shows that high-quality language supervision improves alignment and control in image-text models, yet research on curating precise video annotations remains scarce. Existing video–text datasets [11, 12, 25, 38, 43, 59, 69] often lack a clear specification of what to describe, producing inconsistent captions that cause models to hallucinate or miss key visual details. To differentiate our goal from that of prior captioning work, we use the term video **language-building** as the umbrella task that includes the following three key components (Figure 1): (1) a clear *specification* of what to describe and (2) a scalable *oversight framework* for high-quality annotation, and (3) *post-training strategies* for scaling model capability with modest expert supervision.

(1) Precise specification (Figure 2). Describing a video is inherently ambiguous without explicit guidelines, as one may focus on what the subject does, how the camera moves, or how the framing changes over time. Prior cognitive studies show that language affects visual perception [41, 45, 65], making a precise specification essential for consistent video description. To build this foundation, we collaborate with professional content creators such as filmmakers who rely on a shared vocabulary [35, 54] to coordinate complex creative workflows that often require teamwork. Together, we develop a *comprehensive specification* that formalizes this vocabulary from first principles, defining a structured framework that spans *subjects, scenes, motion, spatial framing, and camera dynamics*. Unlike popular datasets [11, 25, 59] that lack structured annotation policies, our specification defines hundreds of visual and motion primitives (developed in the concurrent CameraBench-Pro technical report to be released) and training guidelines that together improve caption consistency and coverage in human evaluations.

(2) Oversight framework (Figure 3). Writing detailed video descriptions is cognitively demanding even with clear guidelines. Even a short 5-second clip can contain multiple subjects entering and exiting the frame, each performing distinct actions. Describing such video dynamics can take hundreds of words and more than ten minutes to complete. As a

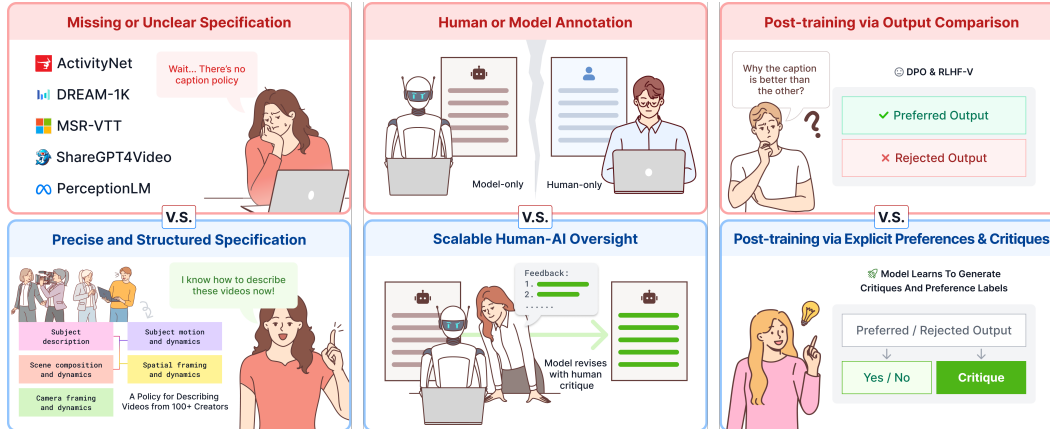


Figure 1. **Recipe for precise video language.** We present a recipe for producing high-quality video captions (blue, bottom row) and compare it with prior work (red, top row). (1) While prior video–text datasets lack a clear policy to teach annotators and models *what* and *how* to describe, we develop a **structured specification** with professional video creators to collect descriptions that support comprehensive video understanding. (2) Instead of relying solely on human or model annotators, we introduce a scalable **human–AI oversight** framework where humans critique model-generated captions with LLM-assisted writing to ensure annotation quality. (3) Our framework naturally produces explicit preferences and feedback for **post-training** models, leading to better performance via SFT, DPO, and inference-time scaling. Figure 5 shows how our model curates detailed video descriptions that fine-tune video generators to better follow detailed prompts with professional-level control over camera dynamics, shot composition, and cinematography. We release our recipe, data, and models to support future research in precise video language.

result, many recent video–text datasets [11, 12, 23, 60, 70] instead rely on video–language models [3, 48] to write captions following human-provided instructions. While these models are highly fluent writers, they often hallucinate visual details, such as confusing left and right, inventing non-existent objects or motion, or misidentifying effects like fisheye lens or dolly zoom. On the other hand, in language tasks such as summarization and code review, models have begun to outperform humans in fluency and coverage, motivating research into *scalable oversight* methods [5, 8, 51] that let humans supervise models that may exceed their own capabilities in certain task-relevant skills. We extend this idea to video captioning by implementing an oversight framework where models first generate high-recall **pre-captions**, and trained human experts focus on **critiquing** rather than writing from scratch, guiding models to produce improved **post-captions**. Our user study in Appendix A shows that this approach improves annotation accuracy, writing quality, and efficiency over prior work [14, 78] that rely on manual caption editing, presumably because shifting limited human attention from text *generation* to *verification* allows for more effective use of cognitive resources.

(3) **Post-training strategies** (Figure 4). Popular post-training methods such as DPO [50] and GRPO [21] rely on preference-based supervision, comparing candidate outputs ranked by humans or reward models. Our data engine not only provides such preference signals but also adds explicit language feedback that explains how to improve weaker outputs into stronger ones through triplets of (**pre-caption**, **critique**, **post-caption**). We find that explicitly training on

these preferences and critiques greatly improves standard supervised fine-tuning and offline RL methods such as DPO and RLHF-V [78]. Crucially, our ablations show that successful post-training depends on the high-quality critiques produced by our framework, which are more accurate, complete, and constructive than those collected in prior work such as in OpenAI’s GDC [51, 85].

Benchmarks and findings. Using our collected triplets on $\sim 4k$ videos spanning films, games, commercials, and user-generated clips, we release the first unified benchmark for precise video–language understanding that jointly evaluates (1) *caption generation*, (2) *reward modeling*, and (3) *critique generation*. Key findings are:

- 💡 (1) Current video-language models capture subject appearance and scene context well but struggle with fine-grained aspects such as subject motion and camera dynamics.
- (2) Explicit preference and critique supervision improves standard SFT and RL methods, enabling open-source models to outperform closed-source Gemini-3.1 with modest expert supervision.
- (3) Critique quality (precision, recall, and constructiveness) is crucial for post-training success. We provide practical guidelines on quality control and workflow design, and release the oversight pipeline to facilitate future research on scalable human-AI data curation.

Application: improving text-to-video generation. We apply our models to re-caption [4] large-scale professional

videos, sourcing diverse, high-quality footage from films, commercials, music videos, and games. This re-captioning helps fine-tuned video generation models such as Wan [58] to better follow detailed prompts of up to 400 words, achieving finer control over motion, camera, visual composition, and cinematography. We hope this serves as a promising first step toward professional-grade text-to-video generation with precise language control. See Figure 5 for examples.

Summary. Our contributions are threefold: (1) a *comprehensive specification* that defines what and how to describe in videos developed with professional video creators; (2) a *scalable oversight framework* that shifts human effort from writing to verification via critique-based correction; and (3) *post-training recipes* that explicitly leverage these preference and feedback signals to further improve the model. We hope our released data, framework, and benchmarks pave the way for scalable human–AI data curation for professional video understanding and generation.

2. Related Work

Video–text datasets. Video captioning has evolved from short, loosely aligned descriptions [25, 69] to longer, detailed ones [14, 16, 24, 31, 35, 43, 62, 67, 82, 83]. However, current datasets curated by humans or models [7, 11, 12, 23, 39, 56, 70, 74, 84] often suffer from incomplete coverage, visual hallucinations, and writing problems. Our human evaluations show that these issues largely stem from the lack of clear specification and effective oversight. In addition, many foundation models are trained on closed-source datasets [3, 10, 73] that cannot be verified for quality.

Scalable oversight. As AI systems begin to match or surpass humans in complex tasks such as coding, mathematics, and writing, a key question arises: how can people continue to supervise models that are stronger than themselves in many of the skills required for these tasks? Scalable oversight [5, 51] in natural language processing studies this challenge through the lens of human–AI collaboration; rather than human-only supervision, models assist humans in providing high-quality oversight through error detection in tasks such as question answering, summarization, and code review [42]. However, most video–text datasets still rely on either human-written [25, 59, 69] or model-generated captions [11, 70] without cooperation. We adapt the principle of scalable oversight to curate video captions, allowing models to generate fluent captions while humans focus on visual details, each specializing in what they do best.

Post-training strategies. Post-training aligns model behavior with human preferences or task-specific goals through supervised fine-tuning, reinforcement learning, or inference-time scaling [6, 9, 21, 32, 33, 36, 44, 49, 50, 52, 77]. In multimodal learning, recent work extends these methods to vision–language models using self-critique or reflection [1, 37, 80], yet most rely on synthetic or image-only

feedback [13, 20, 28, 30, 36, 40, 63, 64, 68, 75, 79, 86]. Ours is the first to show that high-quality feedback provides an effective supervision signal for video understanding.

3. Specification For Video Captioning

How the captioning task is framed directly affects what annotators perceive and choose to describe [22, 55]. We first identify common issues in prior datasets that lack consistent annotation policies and then present our structured specification built with professional video creators. See Figure 2 for an overview.

Issue: lack of specification. Without clear instruction, annotators may not know what to describe or how much detail to include. To examine this, we manually evaluate eight widely used video–text datasets: MSR-VTT [69], ActivityNet [25], ShareGPT4Video [11], UltraVideo [70], VDC [7], Dream1K [59], PerceptionLM (PE-Video) [14], and TUNA-Bench [24]. We find that most datasets do not provide a detailed policy for annotators (the only exception being [24], whose guideline is not public). We observe three major issues caused by the lack of specification and provide detailed error examples in Appendix A:

- (1) **Imprecise terminology.** Without clear guidelines, annotators may lack the proper vocabulary to describe visual or motion effects. For example, they often confuse *camera translation* (camera physically moves forward or sideways) with *zoom* (focal length change without moving the camera) or *rotation* (camera pivots in place). They also misuse cinematography terms, calling a *full shot* (entire subject in view) a *close-up* (small part of the subject), an *aerial shot* (high altitude) a *bird’s-eye view* (top-down angle), or describing *fisheye distortion* (straight lines bending outward from a wide lens) as a “circular” scene. See Figure 6 for such mistakes made by untrained annotators.
- (2) **Missing information.** Without clear rules on what to include, annotators often miss key elements essential for understanding a shot. Some captions focus only on the subject and its actions, while others describe only the scene or camera movement, leading to incomplete descriptions. For instance, although many datasets attempt to capture camera work [14, 24], their captions frequently omit crucial details such as camera shake, focus changes, and tracking movement.
- (3) **Subjective descriptions.** Without an objective annotation policy, captions often include personal opinions or emotional language, such as calling a scene “inspiring” or “informative,” which other annotators may not agree with. Such subjective phrasing varies across annotators and distracts from describing the actual visual content.

Building specification with content creators. The lack of clear specification among prior datasets motivates us to learn from professionals who already use one. We view video captioning as a form of *visual storytelling*: the best

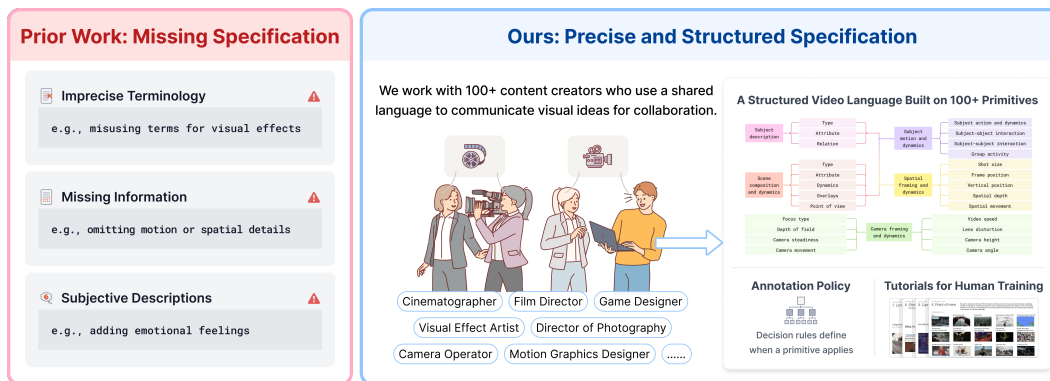


Figure 2. **Specification is crucial in video language.** Prior datasets (red, left column) that lack a clear specification often misuse visual and motion terms, omit key information, and include subjective or emotional descriptions that distract from the observable video content. Over the course of a year, we worked (blue, right column) with more than 100 professional video creators, including filmmakers, cinematographers, visual effects artists, and game designers, who rely on a shared language to communicate visual intent precisely in their work. Together, we define a structured specification of video language that spans subjects, scenes, motion, spatial, and camera dynamics, grounded by hundreds of primitives such as POVs, overlays, playback speed, subject actions and transitions, shot size and angle, and camera lens and movement. Each primitive is paired with clear decision rules and training materials to help both professionals and non-professionals apply them consistently. Detailed descriptions of these primitives and guidelines are provided in our concurrent work CameraBench-Pro (which will be released separately). We find this specification effective in professional workflows and will release it publicly to support future research in precise video understanding and generation.

caption should let someone who has never seen the video picture it vividly [22]. This idea naturally aligns with how content creators work. In filmmaking, for example, directors must tell camera operators exactly how the camera should move, what angle to use, and how the frame should look to achieve the intended shot [35, 54]. Such teamwork depends on a precise shared language to communicate visual intent before the shot is made. We adopt this practice to build a structured, teachable video language with over 100 creators from the US and China working in film, game design, and motion graphics, with 2–5+ years of professional experience. Through a year-long collaboration, we take a bottom-up approach to formalize this language. We start by asking professional creators to describe diverse videos—from films and games to user-generated clips—and collect the aspects they naturally mention. These are then grouped into a structured framework of five major aspects with clear subcategories: *subject* (types, attributes, relations), *scene* (point of view, overlays, setting, time of day), *motion* (actions, interactions, group activities), *spatial* (shot size, frame position, spatial depth, movement), and *camera* (playback speed, camera height, angle, lens, focal plane, steadiness, movement). Each aspect captures both what appears and how it changes over time. We note that our framework focuses only on *objectively observable* visual aspects, rather than subjective impressions or emotional interpretations.

Primitives for precise specification. Common cinematic terms are often used inconsistently across practitioners and datasets. To address this, the concurrent work CameraBench-Pro (to be released) defines hundreds of fundamental vi-

sual and motion **primitives** through bottom-up collaboration with professional video creators, establishing rigorous definitions, video examples, corner cases, and decision rules for each primitive. For example, CameraBench-Pro clarifies that *bird’s-eye view* refers to a strict top-down angle rather than any high vantage point, and extends playback speed beyond *slow motion* and *time-lapse* to include *speed ramp* and *stop motion*. The full taxonomy includes over 200 primitives spanning camera motion (~50, overlapping with CameraBench [35]), camera setup (~100), and video cinematography (~70). Each static property (e.g., focal plane, shot size) is labeled at both the start and end of a shot to capture temporal change, with decision rules resolving common ambiguities such as subject identification and dynamic framing. We refer readers to the CameraBench-Pro technical report for the complete taxonomy, training guidelines, and labeling platform.

Structured captioning policy. While primitives define the building blocks of our specification, they alone cannot describe how these elements unfold within a video. A complete caption must link multiple primitives into a coherent description; for example, a transition from a *bird’s-eye view* to a *level angle* through camera *tilting upward*, or a *shallow depth of field* shot where the focus shifts from *foreground* to *midground* through a *rack focus*. To capture such structure, we develop detailed guidelines that extend primitive-based labeling into full captioning, which specifies how to describe each aspect under different conditions. For instance, when a video lacks a clear or consistent *subject*, annotators note the absence (e.g., an establishing shot) or describe how the sub-

ject changes over time (e.g., a revealing shot). [Appendix I](#) provides the full guidelines in detail.

4. Human-AI Oversight Framework

With or without a clear specification, writing detailed video descriptions is difficult and prone to error. Even a short five-second clip can include several subjects and camera or spatial movements that must be described precisely and in order. Oversight is therefore key to keeping annotations accurate and consistent. To study its effect on data quality, we first analyze datasets with limited or no oversight and the issues that result, then present our **CHAI** framework (Critique-based Human-AI oversight). [Figure 3](#) provides a summary.

Issue: lack of oversight. Most existing video-text datasets rely on either human-written or model-generated captions without systematic quality control. We provide a detailed report of them in [Appendix A](#) and summarize the major issues caused by insufficient oversight below:

- (4) **Poor writing.** Human captions often contain typos, grammatical mistakes, or awkward phrasing. Beyond these surface-level issues, untrained annotators may describe events out of temporal order or use vague references when multiple subjects appear, such as saying “the first person hugs the second person” instead of “the person on the left hugs the one on the right.”
- (5) **Visual hallucinations.** Video captioning models frequently make confident but incorrect statements about the video, e.g., describing actions or objects that do not exist, or claiming the camera is static when it is clearly moving.
- (6) **Inaccurate details.** Both human and AI annotators struggle with subtle visual and motion details, such as mistaking a hand on the right side of the frame for the subject’s right hand (which is actually the left), or missing a small handheld motion that makes the camera non-static.

Quality data requires effective oversight. Our human evaluations ([Appendix A](#)) show that even recent benchmarks such as TUNA-Bench from Kling[24], which claims to have a specification (though unreleased), still suffer from issues caused by limited oversight. In our case, oversight is even more critical because following our caption policy to write an extremely detailed caption (200–400 words) from scratch is cognitively demanding even for trained experts. To efficiently collect high-quality captions, we design an oversight framework that combines (a) *human-AI collaboration* to reduce writing and hallucination errors, and (b) *screening, training, and incentives* to ensure annotators are skilled and motivated to keep details accurate.

(a) **Human-AI collaboration for data curation.** Inspired by scalable oversight in NLP [5, 51], we design a simple but effective workflow that divides tasks between humans and models so each focuses on what it does best. Since

large language models are already more efficient writers than most humans [8], our workflow lets models write fluent text that closely follows human instruction, while humans focus on visual fact-checking of model outputs. [Figure 3](#) shows the steps: (1) Humans first label all visual and motion **primitives**¹ that are important but easy to miss when writing captions from scratch (e.g., camera shake, focal-plane shifts, point of view, shot size, and overlays such as framing or subtitles); (2) a video-language model drafts a **pre-caption** from these labels, covering as many relevant details as possible according to our captioning guideline; (3) humans review the pre-caption and write a **correctional critique** explaining what is wrong or missing and how to fix it; (4) the model incorporates this critique to produce a refined **post-caption**; and (5) humans repeat the critique if needed until the caption is fully accurate. The appendix provides more details on this process, including how annotators first complete *subject* and *scene* post-captions, which are then used to prompt the model to generate more accurate *motion* and *spatial* pre-captions describing how subjects and scene elements move and are arranged within the frame. The appendix also includes concrete examples and interface screenshots. Our preliminary study shows that annotators find this division of labor faster and less mentally demanding, leading to captions that are (1) more accurate, as humans focus on *verification* rather than *generation*; (2) more complete, as models better follow the comprehensive guidelines and make fuller use of all labeled primitives; and (3) more fluent, since all text is polished by the model.

(b) **Screening, training, and incentives.** While this human-AI workflow is effective, both humans and models can still miss fine spatial or motion details; for example, for a camera-facing person, “moving to his left” appears on the frame’s right, which is easy to confuse with “moving to the frame’s left.” To ensure precise descriptions, we focus on selecting, training, and motivating the most capable experts. We recruit annotators only with prior experience in content creation, such as filmmaking, motion design, or game capture. All applicants must complete six rounds of multiple-choice exams based on our primitives, covering over 150 videos that test understanding of camera motion, setup, and video cinematography. Only the top 3% of applicants (from over 600) who rank in the top 20% after six exams are selected. They then complete a month of (paid) training on our labeling platform, practicing primitive-level tasks to build the discrimination skills needed for professional-quality captioning. We also encourage annotators to raise questions during training if they disagree with any ground-truth answer or are unsure how to handle edge cases. High performers are promoted to captioning roles and begin by shadowing

¹To reduce cognitive load, primitive labeling is done on a separate platform with its own quality-controlled process, allowing annotators to focus purely on captioning during the main workflow.

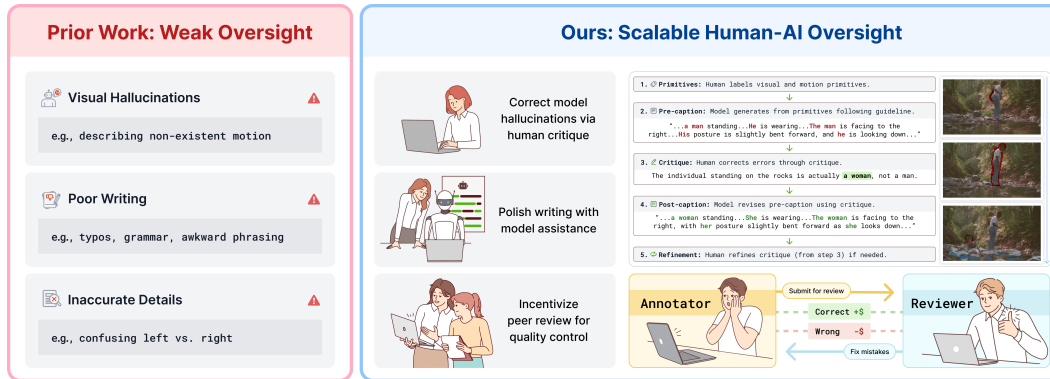


Figure 3. **Oversight is crucial for quality annotations.** Even with clear task guidelines, writing detailed video descriptions remains error-prone for both humans and models. Our human evaluations show that prior datasets (red, left column) suffer from recurring issues: human-written captions often contain typos, grammatical errors, awkward phrasing, and events described out of order; while model-written captions frequently exhibit visual hallucinations, describing non-existent subjects or motion. To address these issues, we introduce a critique-based **human–AI oversight** framework (blue, right column) that combines the strengths of both. The model first generates a comprehensive **pre-caption** following task guidelines, and the human writes a **critique** that corrects hallucinations and guides the model toward a refined **post-caption**. This division of labor (1) eliminates *writing problems* since the language model handles text generation and editing, and (2) allows humans to focus on *verification* rather than writing a full caption from scratch or worrying about phrasing. Lastly, since both humans and models still struggle with fine details such as spatial relations (e.g., left vs. right), we implement (3) a **quality control system** with a peer-review bonus: annotators earn bonuses when reviewers find no errors, and reviewers earn bonuses for valid corrections. This incentive keeps both sides motivated for accuracy. Implementation details—including annotator screening, training, and interface design—are provided in the appendix. The appendix includes the full critiques (which are more detailed) and additional examples.

expert critiques on 100 gold-standard captions authored and reviewed by the paper’s authors (some of whom are also content creators). To further improve accuracy, we introduce a reviewer role (promoted from top annotators with a strong track record) that checks every critique and post-caption and corrects errors using the same workflow. Annotators can see how their work is corrected and may appeal to the managers (the paper’s authors) if they disagree. This peer-review process is reinforced by an accuracy-based bonus: annotators earn rewards for error-free annotations, and reviewers for correcting mistakes, keeping both roles motivated to maintain precise captions. See [Appendix C](#) for implementation details.

5. Evaluation and Post-Training

Our oversight framework naturally yields triplets of (*pre-caption*, *critique*, *post-caption*) that support evaluation and post-training for video–language models.

Benchmarks and tasks. We curate $\sim 20k$ triplets in total, allocating 5k as a held-out benchmark and using the rest for training. This scale far exceeds recent public benchmarks (typically $\sim 1k$), such as DREAM-1K [59] and TUNA-Bench [24], and uniquely covers five aspects per video: *subject*, *scene*, *motion*, *spatial*, and *camera*. The appendix provides additional statistics such as video duration, FPS, and word count. To our knowledge, ours is the first unified benchmark that evaluates not only *video captioning* but also

reward modeling and *critique generation*, which have become increasingly important for post-training [20, 42]. We expect all three tasks to play a crucial role in advancing video understanding; for example, the appendix shows that strong reward and critique models enable inference-time scaling. Concretely, the tasks are:

- (1) **Caption generation** (Video \rightarrow Caption): generate a long caption that follows our policy for the five aspects. We report the reference-based BLEU-4 metric in the main paper and defer ROUGE and LLM-as-judge results to the appendix.
- (2) **Reward modeling** ([Video, Caption] \rightarrow Score): predict which caption better matches the video by checking whether the post-caption receives a higher score than the pre-caption. We report binary accuracy (chance = 0.5).
- (3) **Critique generation** ([Video, Caption] \rightarrow Critique): produce a correctional critique that identifies errors or omissions and explains how to fix them. We report BLEU-4 against the reference critique in the main paper, and include ROUGE and a “critique-guided revision” proxy metric [20] in the appendix.

Preliminaries: offline post-training. Given our collected triplets, we first consider offline **SFT**, which directly trains the model to generate the preferred outputs (post-captions). We also explore offline RL methods like **DPO** [50], which adds a contrastive objective that rewards preferred outputs and penalizes rejected ones with a KL reg-

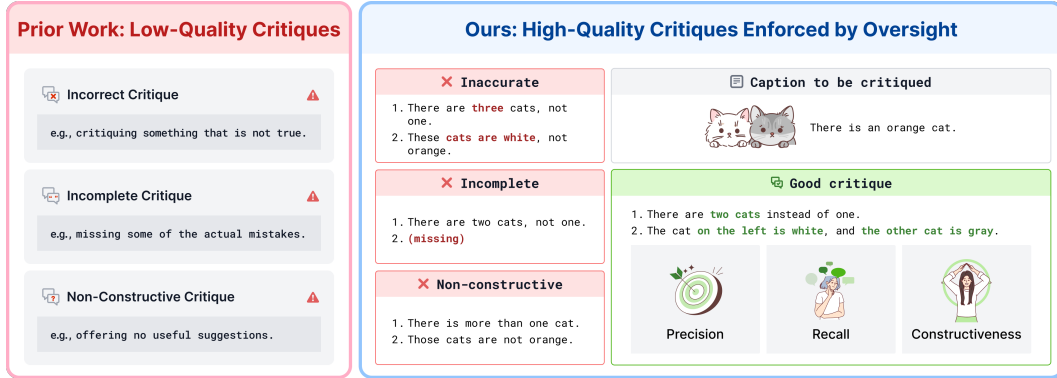


Figure 4. **Critique quality matters for post-training.** Prior work [51, 85] often collects critiques that are inaccurate (e.g., hallucinating information), incomplete (e.g., skipping mistakes), or non-constructive (e.g., noting errors without explaining how to fix them). Our curation framework instead requires each critique to directly guide the model in producing the final post-caption, forcing annotators to write critiques that are accurate, complete, and constructive. Appendix E shows that removing any of these qualities leads to substantially worse results.

Table 1. **Model performance across captioning, reward modeling, and critique generation.** We note that: (1) Current models perform well on *subject* and *scene* but struggle with *motion* and *camera* aspects. (2) Our fine-tuned model achieves state-of-the-art performance across all tasks, outperforming the close-sourced Gemini-2.5/3.1. (3) Adding explicit *preference* and *critique* signals (full-data post-training) improves both SFT and RL. Note that Gemini-3.1 does not support logprobs for reward modeling. Appendix F provides experimental details.

Method	Caption Generation						Reward Modeling						Critique Generation					
	Subject	Scene	Motion	Spatial	Camera	Avg	Subject	Scene	Motion	Spatial	Camera	Avg	Subject	Scene	Motion	Spatial	Camera	Avg
<i>Open-source models</i>																		
PerceptionLM [14]	8.2	4.8	5.0	7.0	7.5	6.5	38.2	32.4	29.9	34.9	39.9	35.1	2.5	1.5	2.0	1.8	2.2	2.0
OmniVinci [76]	2.8	5.2	3.5	5.5	3.0	4.0	35.9	42.9	37.5	32.8	34.7	36.8	1.2	2.2	1.8	2.5	1.3	1.8
VideoChat-R1.5 [71]	6.5	5.8	4.2	2.8	5.0	4.9	42.5	44.3	41.0	43.3	49.3	44.1	1.0	2.5	1.8	2.0	2.2	1.9
VideoCap-R1 [43]	7.5	6.2	5.5	6.8	4.0	6.0	52.7	46.7	45.5	45.9	44.6	47.1	2.0	1.5	2.8	1.2	2.0	1.9
SkyReels-V2 [10]	1.8	4.0	2.5	4.5	3.2	3.2	52.7	58.0	55.2	51.0	59.9	55.4	2.2	1.0	1.5	2.8	1.5	1.8
OwlCap [87]	4.8	5.5	3.8	5.2	2.5	4.4	48.4	51.3	49.7	47.4	55.2	50.4	1.5	2.5	2.0	1.2	1.8	1.8
video-SALMONN-2 [56]	2.5	1.5	2.0	3.5	3.8	2.7	53.1	61.9	57.8	56.1	61.2	58.0	1.8	1.0	2.5	1.5	2.2	1.8
ShareGPT4Video [11]	5.5	6.8	3.5	6.0	4.5	5.3	49.5	42.5	46.7	44.2	51.4	46.9	2.0	1.5	1.8	2.5	1.0	1.8
MotionSight [18]	4.0	3.5	5.8	4.2	2.2	3.9	45.3	48.2	53.1	46.7	49.3	48.5	2.8	1.5	1.0	2.0	2.2	1.9
MMR1 [27]	2.0	4.2	2.8	5.0	3.5	3.5	55.2	60.2	59.5	57.8	58.2	58.2	1.5	2.0	2.5	1.0	1.5	1.7
SynPO [16]	6.8	7.2	5.0	3.8	6.5	5.9	46.0	43.9	44.6	45.0	51.8	46.3	1.2	2.5	1.8	2.2	1.8	1.9
VideoPASTA [26]	3.0	2.8	1.8	3.8	1.5	2.6	56.9	56.7	54.0	51.0	49.5	53.6	2.0	1.0	1.5	2.5	2.2	1.8
Omni-Captioner [39]	8.0	7.8	4.5	7.5	6.5	6.9	41.0	43.9	42.5	42.5	50.1	44.0	1.5	2.2	2.5	1.0	1.8	1.8
InternVL-3.5 [61]	3.8	2.4	1.8	3.2	1.8	2.6	46.0	63.7	61.6	59.5	63.1	58.8	2.8	2.0	2.5	2.2	1.5	2.2
Granite 3.3 [57]	2.0	1.8	0.8	2.2	0.8	1.5	41.2	44.3	38.9	42.5	46.7	42.7	1.0	1.5	0.8	1.2	2.0	1.3
Qwen2.5-VL-7B [3]	5.1	1.4	4.3	4.6	4.6	4.0	31.0	27.2	35.0	26.1	64.9	36.8	2.2	2.8	1.8	1.5	2.0	2.1
Qwen2.5-VL-72B [3]	7.0	6.5	3.0	6.5	6.0	5.8	33.3	28.8	36.1	27.8	63.9	38.0	3.0	2.5	2.8	2.2	1.8	2.5
Qwen3-VL-8B-Instruct [72]	1.2	3.2	1.4	4.8	2.6	2.6	47.0	63.9	60.5	51.7	56.1	55.8	2.5	3.0	2.0	2.8	1.8	2.4
Qwen3-VL-72B-Instruct [72]	5.4	6.0	3.0	7.8	4.2	5.3	55.7	67.3	66.3	57.8	52.9	60.0	3.2	2.5	2.8	3.5	2.0	2.8
<i>Closed-source models</i>																		
GPT-4o [47]	4.7	3.0	3.8	4.5	4.9	4.2	49.2	53.7	51.1	51.9	56.6	52.5	1.5	2.0	2.5	1.2	1.8	1.8
GPT-5	5.9	6.3	4.3	5.8	5.1	5.5	55.7	61.6	58.2	59.0	62.9	59.5	3.0	2.5	3.5	2.8	2.2	2.8
Gemini-2.5-Pro [15]	6.3	6.8	4.2	7.2	5.5	6.0	58.2	64.9	57.3	62.9	66.5	62.0	3.5	2.8	2.5	3.2	3.0	3.0
Gemini-3.1-Pro	6.1	6.0	1.4	7.2	4.6	5.1	-	-	-	-	-	-	3.8	2.8	3.5	3.4	2.8	3.3
<i>Caption-only post-training (Qwen3-VL-8B-Instruct)</i>																		
RLHF-V (Caption)	10.2	9.5	6.8	13.5	10.8	10.2	57.8	67.0	45.8	44.2	42.5	51.5	3.2	2.8	3.0	4.5	3.2	3.3
DPO (Caption)	10.8	10.0	7.2	13.0	11.2	10.4	58.2	66.5	46.2	43.8	43.0	51.5	3.5	2.5	2.8	4.2	3.5	3.3
SFT (Caption)	14.5	13.2	9.0	18.2	14.8	13.9	59.5	68.8	47.4	45.7	44.4	53.2	3.8	2.5	3.5	5.2	3.8	3.8
SFT + RLHF-V (Caption)	15.2	14.0	9.5	19.0	15.5	14.6	60.2	69.5	48.0	46.2	45.0	53.8	4.2	3.2	3.8	5.5	4.0	4.1
SFT + DPO (Caption)	15.8	13.5	9.3	18.6	15.8	14.6	59.8	69.8	47.8	45.5	45.2	53.6	3.8	3.5	3.2	5.0	4.2	3.9
<i>Full data post-training (Qwen3-VL-8B-Instruct)</i>																		
RLHF-V (All)	14.2	14.0	8.6	20.8	13.5	14.2	83.5	86.2	73.8	80.4	77.1	80.2	21.8	21.5	22.4	25.0	30.2	24.2
DPO (All)	13.8	14.5	9.0	20.2	14.1	14.3	82.8	87.4	72.5	79.6	78.5	80.2	20.5	22.8	21.6	25.8	29.4	24.0
SFT (All)	16.8	17.5	10.2	24.1	16.5	17.0	89.8	94.2	80.1	86.0	84.1	86.8	23.3	24.7	23.5	29.1	34.1	26.9
SFT + RLHF-V (All)	17.5	18.1	11.0	25.2	17.2	17.8	90.2	94.0	80.5	86.3	84.5	87.1	24.1	25.2	24.0	29.8	34.5	27.5
SFT + DPO (All)	17.9	17.8	10.8	24.8	17.8	17.8	89.5	94.5	79.8	85.7	84.8	86.9	23.8	25.8	23.2	30.2	33.8	27.4

ularization term. **RLHF-V** [78] further extends DPO by increasing gradients on the edited text segments between rejected and preferred outputs. Our data naturally supports these methods since each triplet provides both preferred (post-caption) and rejected (pre-caption) outputs.

Ours: training with explicit preferences and critiques.

While DPO and RLHF-V only compare preferred and rejected outputs implicitly, we further consider explicitly training models to generate (1) critiques and (2) preference labels.

- (1) **Critiques:** We train the model to generate a critique for each (video, caption) pair using reference critiques when available. If a caption is already correct, the target critique is “*The caption is accurate and requires no edits.*” This training enables the model to perform more effective self-critique during inference.
- (2) **Preference labels:** We train the model to classify each (video, caption) pair into {Yes, No}, where “Yes” denotes the preferred (post-)caption and “No” the rejected (pre-)caption. At inference time, following VQAScore [34], the model’s probability of “Yes” is used as a reward score, which we find more reliable than prompting for Likert-scale ratings.

Findings. Table 1 presents our main results. (1) Current models perform well on *subject* and *scene* but struggle with *motion* and *camera* aspects, which are likely underrepresented in training data. (2) Our fine-tuned model achieves state-of-the-art performance, surpassing both open- and closed-source baselines. (3) Adding explicit *preference* and *critique* signals consistently improves both SFT and RL methods across all tasks. Additional results in the appendix show that (4) inference-time scaling brings further gains without extra human supervision on reward modeling and caption generation tasks.

Critique quality is the key. Effective critiques must be *accurate*, *complete*, and *constructive*, explaining not only what is wrong but also how to fix it (see Figure 4 for an example). Our ablations in Appendix E confirm that all three properties matter for post-training success. In contrast, prior work such as MM-RLHF [85] and OpenAI’s GDC [51] often collects feedback that identifies errors but offers no corrections.

6. Improving Professional Video Generation

High-quality captions are crucial for improving visual generation [4]. We use our post-trained model to re-caption large-scale professional videos and fine-tune Wan2.2 [58] for better prompt following. We manually collect about ~150K videos from YouTube channels covering films, commercials, music videos, and game footage under standard (non-commercial) licenses.

Re-captioning improves text-to-video. Figure 5 shows that our fine-tuned Wan model better follows detailed

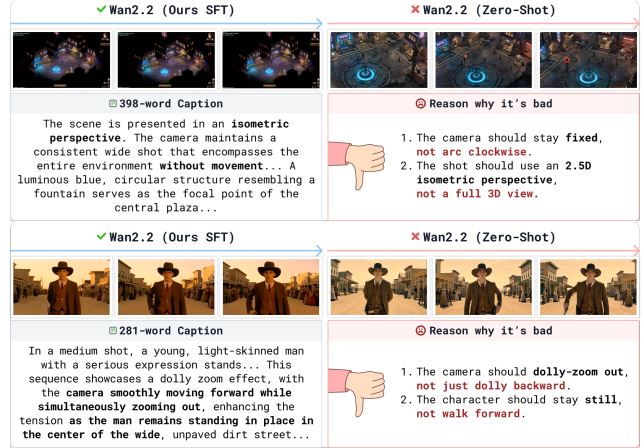


Figure 5. **Re-captioning improves video generation.** After fine-tuning on large-scale videos re-captioned by our post-trained Qwen3-VL, Wan2.2 can better follow detailed prompts (up to 400 words) more faithfully and achieve finer control over camera motion and video cinematography, such as dolly zoom movements and isometric (2.5D) views. Human and automated evaluations in the appendix confirm that this pipeline outperforms both zero-shot Wan and Wan fine-tuned using zero-shot Qwen3-VL captions.

prompts and supports finer control such as *dolly zoom* and *isometric* (2.5D) perspectives. Additional examples in the appendix demonstrate control over camera angles (*worm’s-eye*, *Dutch angle*), movements (*rolling*), lenses (*fish-eye*), playback speed (*speed ramp*), height transitions (*underwater to above-water*), focus shifts (*rack focus*), and framing (*overlays*, *shot size*, and *revealing shots*). Evaluations on 200 samples show that fine-tuning on our re-captioned data significantly improves prompt following compared to Wan2.2 fine-tuned with zero-shot Qwen3-VL captions (Appendix H).

7. Conclusion

Limitations and future work. Our CHAI framework shows that high-quality data can be curated efficiently through human-AI collaboration, even with modest academic resources. Future work can further scale this process by using stronger critique models to assist human annotators [5, 51]. While this work focuses on video understanding, building benchmarks for video generation is a natural next step.

Summary. We present an open recipe, dataset, and benchmarks for precise video language, combining clear specification, scalable oversight, and effective post-training. We hope this work fosters broader progress in scalable human-AI collaboration for high-quality data curation in video understanding and generation.

References

- [1] Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models. *arXiv preprint arXiv:2408.11791*, 2024. 3
- [2] Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences. *arXiv preprint arXiv:2506.02095*, 2025. 1
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 7
- [4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2023. 1, 2, 8
- [5] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilè Lukošiušė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022. 2, 3, 5, 8
- [6] Tianning Chai, Chancharik Mitra, Brandon Huang, Gautam Rajendrakumar Gare, Zhiqiu Lin, Assaf Arbelle, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, Deva Ramanan, et al. Activation reward models for few-shot model alignment. *arXiv preprint arXiv:2507.01368*, 2025. 3
- [7] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashishth Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 3, 13, 15
- [8] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025. 2, 5
- [9] Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025. 3
- [10] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 3, 7
- [11] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2025. 1, 2, 3, 7, 13, 14, 15
- [12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 1, 2, 3
- [13] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14239–14250, 2024. 3
- [14] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Hanoona Rasheed, Peize Sun, Po-Yao Huang, Daniel Bolya, Suyog Jain, Miguel Martin, Huiyu Wang, Vivian Lee, Andrew Westbury, Salman Khan, Philipp Krähenbühl, Piotr Dollár, Lorenzo Torresani, Kristen Grauman, and Christoph Feichtenhofer. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. Preprint. 2, 3, 7, 13, 15, 16
- [15] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Bliestein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7, 36
- [16] Jisheng Dang, Yizhou Zhang, Hao Ye, Teng Wang, Siming Chen, Huicheng Zheng, Yulan Guo, Jianhuang Lai, and Bin Hu. Synpo: Synergizing descriptiveness and preference optimization for video detailed captioning. *arXiv preprint arXiv:2506.00835*, 2025. 3, 7
- [17] Daniel Deutsch, George Foster, and Markus Freitag. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, 2023. 33, 36
- [18] Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. Motionsight: Boosting fine-grained motion understanding in multimodal llms. *arXiv preprint arXiv:2506.01674*, 2025. 7
- [19] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36: 35544–35575, 2023. 1
- [20] Brian Gordon, Yonatan Bitton, Andreea Marzoca, Yasumasa Onoe, Xiao Wang, Daniel Cohen-Or, and Idan Szepes. Unblocking fine-grained evaluation of detailed captions: An explaining autorater and critic-and-revise pipeline. *arXiv preprint arXiv:2506.07631*, 2025. 3, 6
- [21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 3
- [22] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhat-tacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer, 2020. 3, 4
- [23] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan.

- Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024. 2, 3
- [24] Fanheng Kong, Jingyuan Zhang, Hongzhi Zhang, Shi Feng, Daling Wang, Linhao Yu, Xingguang Ji, Yu Tian, Victoria W., and Fuzheng Zhang. Tuna: Comprehensive fine-grained temporal understanding evaluation on dense dynamic videos. *arXiv preprint arXiv:2505.20124*, 2025. 3, 5, 6, 13, 15, 16
- [25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1, 3, 13, 14, 15
- [26] Yogesh Kulkarni and Pooyan Fazli. Videopasta: 7k preference pairs that matter for video-llm alignment. *arXiv preprint arXiv:2504.14096*, 2025. 7
- [27] Sicong Leng, Jing Wang, Jiayi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, et al. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *arXiv preprint arXiv:2509.21268*, 2025. 7
- [28] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5290–5301, 2024. 3, 33
- [29] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 34
- [30] Pengxiang Li, Zhi Gao, Bofei Zhang, Tao Yuan, Yuwei Wu, Mehrtash Harandi, Yunde Jia, Song-Chun Zhu, and Qing Li. Fire: A dataset for feedback integration and refinement evaluation of multimodal models. *Advances in Neural Information Processing Systems*, 37:101618–101640, 2024. 3
- [31] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*, 2025. 3
- [32] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models, 2023. 3
- [33] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2024. 3
- [34] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 8, 33
- [35] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhang Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. 2025. 1, 3, 4, 16, 18
- [36] Shihong Liu, Zhiqiu Lin, Samuel Yu, Ryan Lee, Tiffany Ling, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. *arXiv preprint arXiv:2309.05950*, 2023. 3
- [37] Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025. 3
- [38] Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwei Yu, Jixuan Chen, Boqiang Zhang, Nianzu Yang, Pandeng Li, Yinglu Li, Zuan Gao, et al. What is a good caption? a comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *arXiv preprint arXiv:2502.14914*, 2025. 1
- [39] Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*, 2025. 3, 7
- [40] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023. 3
- [41] Martin Maier and Rasha Abdel Rahman. Native language promotes access to visual consciousness. *Psychological Science*, 29(11):1757–1772, 2018. 1
- [42] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024. 3, 6
- [43] Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025. 1, 3, 7
- [44] Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. Enhancing few-shot vision-language classification with large multimodal model features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2760–2772, 2025. 3
- [45] Eduardo Navarrete, Michele Miozzo, and Francesca Peressotti. Language can shape the perception of oriented objects. *Scientific reports*, 10(1):8409, 2020. 1
- [46] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pages 291–309. Springer, 2024. 1
- [47] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7, 36
- [48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances*

- in *Neural Information Processing Systems*, 35:27730–27744, 2022. 2
- [49] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails of vision-language models. *arXiv preprint arXiv:2401.12425*, 2024. 3
- [50] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 2, 3, 6
- [51] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022. 2, 3, 5, 7, 8, 28, 29, 30
- [52] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv e-prints*, abs/1707.06347, 2024. 3
- [53] Tomas Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024. 40
- [54] Raymond Spottiswoode. *A grammar of the film: An analysis of film technique*. Univ of California Press, 1969. 1, 4
- [55] Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd international ACM SIGACCESS conference on computers and accessibility*, pages 1–15, 2021. 3
- [56] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. videosalmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. 3, 7
- [57] Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, et al. Granite vision: a lightweight, open-source multimodal model for enterprise intelligence. *arXiv preprint arXiv:2502.09927*, 2025. 7
- [58] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wenten Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 8
- [59] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 1, 3, 6, 13, 15, 16
- [60] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025. 2
- [61] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internv13. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 7
- [62] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025. 3
- [63] Yubo Wang, Xiang Yue, and Wenhui Chen. Critique fine-tuning: Learning to critique is more effective than learning to imitate. *arXiv preprint arXiv:2501.17703*, 2025. 3
- [64] Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xianguyu Zhang, et al. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025. 3
- [65] Jonathan Winawer, Nathan Witthoft, Michael C Frank, Lisa Wu, Alex R Wade, and Lera Boroditsky. Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19):7780–7785, 2007. 1
- [66] Ludwig Wittgenstein. *Tractatus logico-philosophicus*. 1922. 1
- [67] Peiran Wu, Yunze Liu, Zhengdong Zhu, Enmin Zhou, and Shawn Shen. Ugc-videocaptioner: An omni ugc video detail caption model and new benchmarks. *arXiv preprint arXiv:2507.11336*, 2025. 3
- [68] Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. Visco: Benchmarking fine-grained critique and correction towards self-improvement in visual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9527–9537, 2025. 3
- [69] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 3, 13, 14, 15
- [70] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025. 2, 3, 13, 14, 15
- [71] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025. 7
- [72] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang,

- Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 7, 36
- [73] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025. 3
- [74] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261, 2024. 3
- [75] Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025. 3
- [76] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025. 7
- [77] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 3
- [78] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rllhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 2, 8
- [79] Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. Diverse ai feedback for large language model alignment. *Transactions of the Association for Computational Linguistics*, 13:392–407, 2025. 3
- [80] Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chengguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, et al. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*, 2024. 3
- [81] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aavek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 17
- [82] Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9050–9061, 2025. 3
- [83] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 3
- [84] Shi-Xue Zhang, Hongfa Wang, DuoJun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. Vcapsbench: A large-scale fine-grained benchmark for video caption quality evaluation. *arXiv preprint arXiv:2505.23484*, 2025. 3
- [85] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025. 2, 7, 8, 28, 29, 32
- [86] Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haiyan Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, et al. Omnialign-v: Towards enhanced alignment of mllms with human preference. *arXiv preprint arXiv:2502.18411*, 2025. 3
- [87] Chunlin Zhong, Qiuxia Hou, Zhangjun Zhou, Shuang Hao, Haonan Lu, Yanhao Zhang, He Tang, and Xiang Bai. Owlcap: Harmonizing motion-detail for video captioning via hmd-270k and caption set equivalence reward. *arXiv preprint arXiv:2508.18634*, 2025. 7

Paper Title

Supplementary Material

Outline

Below is the outline of the supplement:

- **Section A** provides a detailed error analysis of prior datasets and crowdsourced annotations.
- **Section B** presents our full specification, including creator demographics and captioning policy.
- **Section C** details our complete oversight pipeline, including annotator screening, training, platform design, and promotion criteria.
- **Section D** shows additional examples of (pre-caption, critique, post-caption) triplets from our dataset.
- **Section E** analyzes how critique quality (precision, recall, constructiveness) affects post-training success and examines the issue of non-constructive critiques in prior work.
- **Section F** provides additional ablations of evaluation metrics and reward scoring formats, details SFT dataset preparation, explains adversarial negative critique generation, and reports hyperparameters.
- **Section G** shows inference-time scaling results.
- **Section H** presents details of the web-scraped video set and the fine-tuning experiments for text-to-video generative models.
- **Section I** provides the complete captioning guidelines for human annotators.
- **Section J** presents Python-style pseudocode for converting labeled primitives into model instructions for generating pre-captions.

A. Error Analysis of Prior Datasets

We worked with our top-performing annotators to manually evaluate eight widely used video–text datasets and benchmarks released between 2016 and 2025: MSR-VTT [69], ActivityNet [25], Dream1K [59], VDC [7], UltraVideo [70], ShareGPT4Video [11], TUNA-Bench [24], and PerceptionLM [14]. Through this human study, we identify major issues supported by quantitative results and visual examples. We begin with a summary of the main issues, followed by detailed analyses for each dataset and recommendations for improving future dataset curation.

Issues in prior datasets. Without clear specification and proper oversight, video–text datasets often end up with captions that are inconsistent, incomplete, or even incorrect. When there is no clear specification, annotators are unsure what to describe or how much detail to include. When there is no oversight, these mistakes remain uncorrected and

spread through the dataset. We analyze both problems below, with examples from prior datasets.

Issue: lack of specification. Without clear rules on what and how to describe, captions often become inconsistent and inaccurate:

- **(1) Imprecise terminology.** For example, the caption may misuse terms for visual effects.
- **(2) Missing information.** For example, the caption may omit motion or spatial details.
- **(3) Subjective descriptions.** For example, the caption may introduce emotional interpretations.

Issue: lack of oversight. Curating long, detailed video descriptions is inherently demanding and error-prone. Without proper oversight, both humans and models can produce captions with errors that go unchecked:

- **(4) Poor writing.** For example, there may be typos, grammatical mistakes, or awkward phrasing. More subtle issues include describing multiple temporal events in the wrong order or mentioning multiple subjects without giving clear references based on visual attributes.
- **(5) Visual hallucinations.** For example, the caption may hallucinate non-existent objects or motions.
- **(6) Inaccurate details.** For example, the caption may confuse subtle visual details.

Detailed reports. Below we summarize each dataset, discussing their specification and oversight frameworks, and highlighting representative issues. For each dataset, we provide an overview, report its specification and oversight method, and list key issues we identify. In addition to prior datasets, we also report statistics for our own dataset, which includes three versions: (1) *Crowdsourced*: captions collected from untrained workers who were not given our specification but simply told to describe the subject, background, motion, and camera in detail; (2) *ours (without human)*: the pre-caption generated by our label-then-caption approach, where expert-labeled primitives are provided to Gemini-2.5-Pro or Qwen-2.5-VL-72B for converting into more accurate captions; and (3) *ours*: the final human-revised version used for training and evaluation after a second-stage quality check.

Human rating rubrics. We apply the same Likert-scale rating policy across all datasets for the five aspects we covered. In addition, we report an Overall rating that summarizes the coverage and accuracy across all aspects, except for MSR-VTT and ActivityNet Captions, which focus solely on human activities. However, scores across datasets are not directly comparable because the video distributions, caption detail levels, and covered aspects differ:

5. Completely correct; no changes needed.
4. Excellent, with only minor adjustments required (e.g., a few inaccurate, missing, or hallucinated words). Roughly one sentence may need addition, modification, or deletion.

3. Mostly correct but with notable omissions, hallucinations, or errors (e.g., more than two sentences require rewriting, deletion, or addition).
2. Mostly incorrect, requiring substantial revisions (e.g., more than half of the caption is incorrect and must be redone).
1. Almost entirely wrong, requiring a complete rewrite (e.g., the entire caption is irrelevant or fails to mention the specified aspect at all).

Summary of human evaluations. Table 2 summarizes our human evaluation results. The collected scores are not directly comparable across datasets since their video domains (e.g., length, genre, and complexity) and annotation specifications differ. Among prior work, TUNA-Bench appears the most reliable, likely due to its more comprehensive internal specification and quality control, though its guideline is not public and many avoidable issues such as typos and grammatical errors remain. On our videos, we find that captions from crowdworkers are generally unreliable; they tend to be much shorter than those produced under our structured specification and human-AI oversight framework (~100 words vs. ~400 words). Their errors are often severe, as crowdworkers typically lack the visual vocabulary needed to describe common cinematic or motion effects. We include illustrative examples in Figure 6.

- **MSR-VTT** [69] (2016)

- **Overview:** As one of the earliest video-text datasets, MSR-VTT includes short clips (10–30 seconds each) annotated with a single short sentence.
- **Specification:** No public annotation policy was released. Captions mainly describe subjects and their immediate actions, without mentioning scene context, camera motion, or spatial layout.
- **Oversight:** Captions were collected from thousands of crowdsourced AMT workers, with each clip annotated by several independent annotators. The only quality control mentioned is removing duplicate or very short sentences during post-processing.
- **Issues identified:** Crowdsourced captions suffer from severe quality issues. Many describe only part of the clip or focus on one action while ignoring others. They also often contain grammatical mistakes. Because the captions are short (typically under 10 words), we do not observe major issues with subjective descriptions or hallucinated content.

- **ActivityNet Captions** [25] (2017)

- **Overview:** ActivityNet Captions include temporally grounded human-annotated descriptions for ~20k videos (around 100k sentences in total).
- **Specification:** Annotators first wrote a paragraph enumerating all major human actions; then each sentence was tied to a start and end time. Each sentence describes a single event.

- **Oversight:** Annotators were trained through examples of good and bad annotations shown in the interface. The interface enforced minimum length (≥ 3 sentences, ≥ 5 words each) and required all sentences to be timestamped before submission.

- **Issues identified:** Like MSR-VTT, captions mainly focus on human actions. Despite better quality control, quality still varies. Many describe only part of the clip; overlapping events are merged into one sentence. Camera motion and visual style are never mentioned. Due to crowdsourced annotation, Grammar errors (e.g., missing “to”) remain frequent. Because annotators attempt to include more details than in MSR-VTT, hallucinations—especially with subjects and spatial relations—become more common.

- **ShareGPT4Video** [11] (2024)

- **Overview:** ShareGPT4Video annotates 40k videos automatically using GPT-4V without any human-in-the-loop, by sending static frames sampled every 2 seconds to the model.
- **Specification:** The authors provide only a brief instruction to GPT-4V, e.g., “conveying changes in actions, behaviors, environments, states and attributes of objects, and camera movements between adjacent frames.”
- **Oversight:** No quality control was applied to the generated captions.
- **Issues identified:** We manually evaluate only the summary captions from all individual frames. GPT-4V is good at describing general scene context but still hallucinates often, especially for spatial and camera details (e.g., confusing left vs. right). It also makes basic visual mistakes, such as treating overlay graphics as real objects, or saying an object or motion exists when it does not. There are no writing errors since all captions are model-generated, but the model frequently adds subjective phrases like “reflective ambiance,” “warm atmosphere,” or “themes of speed and connectivity” that humans may not agree with. Their videos are also relatively static, which partly explains the higher score compared to newer datasets. Because descriptions are based on static frames, camera motion is often wrong, e.g., calling a moving shot static, or the opposite. Although a short specification is provided, it is inconsistently followed (e.g., some captions mention shot size, others do not). We recommend caution when using this dataset for training video-language models.

- **UltraVideo** [70] (2025)

- **Overview:** UltraVideo includes 59k high-resolution (4K to 8K) videos and provide Qwen2.5-VL generated structured captions.
- **Specification:** The authors instruct Qwen2.5-VL to generate nine structured captions—Brief Description, Detailed Description, Background, Theme Description,

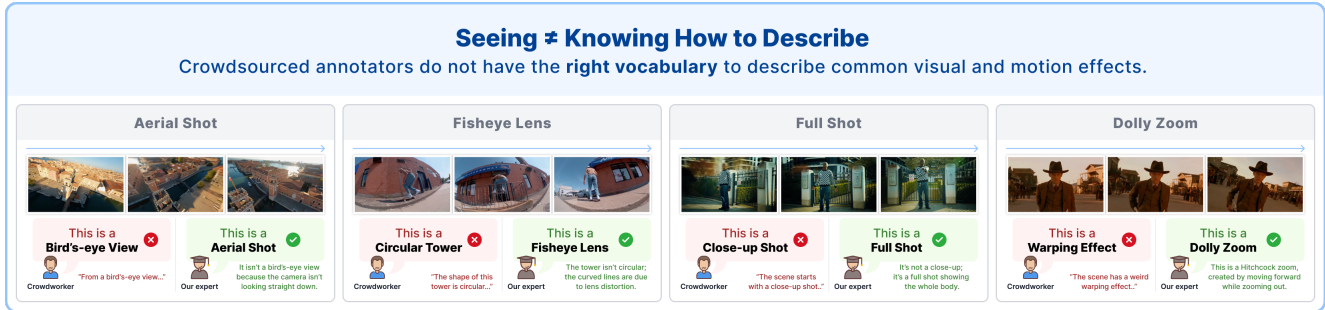


Figure 6. **Errors in crowdsourced video captions.** Crowdsourced video captions often omit key details, resulting in overly brief descriptions, and more importantly, lack the visual vocabulary needed to describe common visual or motion effects in internet videos. For example, they may describe a regular building as a “circular tower” because the camera uses a fisheye lens that bends straight lines outward; confuse a high vantage viewpoint with a “bird’s-eye view,” which actually means a top-down angle; mistake a full shot of a human body for a “close-up” showing only part of the body; or describe a “dolly zoom” shot, where the camera moves while zooming in the opposite direction, as a “weird warping effect.” These errors motivate us to work with professional video creators to define our specification and annotate the captions.

Table 2. **Comparison of video–text datasets on video statistics, caption source, task rating, and known issues.** We do not report the Overall score for MSR-VTT and ActivityNet Captions because they primarily describe human actions without other aspects. We mark only issues that occur frequently (in more than 15% of the samples we manually evaluated). To ensure robustness, we evaluate over 20 random samples per dataset and provide a website to visualize these samples, along with their original dataset IDs. We report caption collection methods from original papers, except VDC, which claims manual evaluation but appears AI-generated. We note that this analysis is not intended as a comprehensive study for several reasons. The scores are not directly comparable across datasets since their video content, annotation specifications, and caption detail levels vary significantly. For example, ShareGPT4Video achieves a similar score to UltraVideo despite using an older image-based GPT-4V model without temporal reasoning, because its videos are generally less dynamic and therefore easier to describe. Among existing datasets, we find the recent TUNA-Bench (with 1k videos) to be the highest quality across most aspects, especially subject, scene, and motion. Detailed qualitative examples and findings are presented in [Appendix A](#). We emphasize that these scores serve only as illustrative references, and we encourage readers to review the actual error cases on our website for deeper insights.

Dataset	Video Statistics			Caption		Task Rating					Specification			Oversight			
	Size	Duration	FPS	Annotation	# Word	Overall	Subject	Scene	Motion	Spatial	Camera	Precise Terminology	Complete Information	Objective Description	Clear Writing	No Hallucination	Accurate Details
MSR-VTT [69]	10k	~15s	3	Human	~10	–	1.8	1.2	3.3	1.0	1.0	X	X	✓	X	✓	X
ActivityNet Captions [25]	20k	~120s	3	Human	~50	–	1.5	1.5	3.7	1.3	1.0	X	X	✓	X	X	X
ShareGPT4Video [11]	40k	~30s	~30	AI	~280	3.7	3.8	4.7	3.8	1.5	2.1	X	X	X	✓	X	X
UltraVideo [70]	59k	~13s	~37	AI	~824	3.7	4.7	4.8	4.5	2.6	3.2	X	X	X	✓	X	X
VDC [7]	1k	~30s	~30	AI?	~500	3.6	4.4	4.6	4.4	2.4	2.7	X	X	X	✓	X	X
Dream1K [59]	1k	~9s	~30	Human	~60	3.3	2.3	1.7	4.7	1.6	2.0	X	X	X	X	✓	X
PerceptionLM [14]	120k	~15s	~28	AI → Human	~50	4.0	4.0	3.2	4.7	2.1	1.4	X	X	✓	✓	✓	X
TUNA-Bench [24]	1k	~15s	~20	Expert	~200	4.1	4.9	4.7	4.9	3.5	2.2	X	X	✓	✓	✓	X
Crowdsourced	1k	~6s	~30	Human	~110	3.2	3.4	3.5	3.5	2.4	2.5	X	X	X	X	X	X
Ours (pre-caption w/o human)	4k	~6s	~30	Expert ↔ AI	~395	3.9	3.9	3.9	3.9	3.4	4.5	✓	✓	✓	✓	X	X
Ours	4k	~6s	~30	Expert ↔ AI	~405	5.0	5.0	5.0	5.0	5.0	5.0	✓	✓	✓	✓	✓	✓

Style, Shot Type, Camera Movement, Lighting, and Video Atmosphere—plus one summarized caption by Qwen3 (average 824 words total). From manual inspection, we find heavy redundancy among the structured captions, so we focus only on the summarized caption for human evaluation.

- **Oversight:** No quality control was applied to the generated captions.
- **Issues identified:** We manually evaluate only the summarized captions. While they roughly follow our specification and include many relevant aspects, each caption focuses on different elements, making them inconsistent. Similar to GPT-4V in ShareGPT4Video, Qwen2.5-VL performs well at describing general scene context but still hallucinates frequently, especially for spatial and camera details (e.g., confusing left vs. right or misnam-

ing camera motion). It rarely mentions camera height, angle, lens distortion, or spatial depth. Hallucinations are also common when describing subjects or actions that do not exist in the video, and key visual subjects are often omitted. There are no grammatical errors since captions are model-generated, but they frequently include subjective phrases like “meditative viewing experience,” “engaging and informative tone,” or “inspire awe and wonder” which humans may not agree with.

- **VDC [7] (2024)**
 - **Overview:** VDC (Video Detailed Captioning) includes 1k videos with long, structured captions developed as part of the AuroraCap benchmark.
 - **Specification:** We cannot find any human annotation policy released by the authors. The structured detailed captions includes camera, short, background, main ob-

ject, and detailed captions. Each part only have a brief task description, e.g., “Camera caption. Describe the camera work in detail, including shot types, angles, movements, transitions, and any special effects used to enhance the video.”

- **Oversight:** Although the paper mentions “manual quality inspection is employed to ensure the quality of the video captions,” the captions appear to be automatically generated. They are too long, uniformly fluent, and free of writing errors, suggesting no human revision was applied.
 - **Issues identified:** The captions are very long, so we only focus on the detailed caption for evaluation. We refer readers to [35] for issues in their camera captions. We find frequent hallucinations and subjective language. Examples include describing a white object as “colorful,” calling a gliding shot “panning,” claiming the lighting changes when it does not, or saying the camera angle shifts when it stays fixed. Captions sometimes mention shot size but not consistently, and descriptions of camera motion are often wrong. While there are no grammar mistakes, many captions use subjective or emotional phrases such as “harmonious escape from the hustle and bustle of everyday life,” “celebrates the dynamic and ever-changing beauty of nature,” and “beauty of concentration and the art of multitasking in a modern office setting.” These patterns strongly suggest the captions are model-generated without human quality control. We do not recommend using this benchmark to evaluate video detailed captioning due to these major issues.
- **DREAM-1K [59] (2024)**
 - **Overview:** Dream1K is a human-annotated test set released with the Tarsier paper, containing 1k challenging movie clips with multiple shots, subjects, and events.
 - **Specification:** The paper provides limited details about the human captioning policy. The authors claim that the human annotation focuses on describing complex videos at different levels of granularity and camera motions such as zooming, translating, panning, and rotating. No detailed public guideline or policy is released beyond these descriptions.
 - **Oversight:** The authors claim the dataset was “carefully described in detail by human annotators,” but no explicit quality control procedure is described. Captions show clear signs of human writing, including grammar issues and uneven phrasing, but also reveal inconsistent precision across samples.
 - **Issues identified:** Although claimed to be carefully annotated, we find occasional hallucinations, especially in fine-grained details such as spatial relations (e.g., confusing reference frames like saying “to her right” instead of “to the right of the frame”) and ambiguous
 - object references (referring “the object” instead of “the person”). Camera motion is frequently omitted even when clearly visible, and subject appearance and scene context are rarely mentioned or described comprehensively, e.g., saying multicolored hair as “green”. We also observe frequent grammar mistakes that result in incomplete sentences, e.g., “The cat grabbing the bowl.”
- **PerceptionLM (PE-Video) [14] (2025)**
 - **Overview:** PerceptionLM builds on the PE-Video dataset, containing about 1M videos, of which around 120k have human-refined captions. According to the paper, each caption is first generated by a model and then refined by human annotators.
 - **Specification:** The authors instruct human annotators to improve the synthetic captions by “removing any hallucinations, correcting words that describe the video inaccurately, eliminating repetitive or redundant words to make the caption more concise, and adding any missing actions being performed in the video.” In practice, the captions mostly focus on activities and actions but sometimes include subject appearance, scene context, or camera motion, leading to inconsistent coverage across samples.
 - **Oversight:** Similar to our dataset, they first use a video captioner to produce pre-captions and then hire 200 annotators to refine them. However, there is no explicit mention of any quality-check or review procedure. We find frequent hallucinations and writing issues, suggesting the absence of systematic quality control.
 - **Issues identified:** Similar to our pipeline, they first use models to draft a pre-caption for each video and then hire annotators to correct the mistakes. However, we suspect that no model was used to polish the final human-edited captions, as many contain writing errors such as “The text say” instead of “The text says” or “siting” instead of “sitting.” Descriptions of camera motion are often inaccurate, for example confusing camera rotation (“pan across”) with camera translation. Spatial details are rarely mentioned, and when they appear, they are often wrong; for instance, confusing a person’s left hand as the right. Fine-grained visual details are also frequently misinterpreted, such as describing a green paper on a screen as a device with a green screen. When multiple subjects appear, captions use vague pronouns like “it” when there are multiple potential subjects, making references unclear. Overall, although human refinement was employed to improve accuracy, captions remain inconsistent and prone to fine-detail visual and writing errors.
 - **TUNA-Bench [24] (2025)**
 - **Overview:** TUNA-Bench consists of 1k videos with detailed captions annotated by trained human experts.
 - **Specification:** The authors claim that “we prepare a

detailed note document for instructing human annotators on annotation.” However, the actual instruction materials are not released publicly and only briefly mentioned in the paper as instructing annotators to “*Strictly follow chronological ordering of events*” and to provide “*Objective descriptions without summarization and subjective feelings.*” They also emphasize distinguishing similar objects “*by unique attributes (e.g., age, dress, etc.)*.” The element rules require descriptions of “*Camera states, including panning, rotating, zooming, following, shaking, transition, etc.,*” along with scene context, actions, and visual attributes such as “*characters’ gender, age, and dress, objects’ color, shape, and number.*”

- **Oversight:** The paper describes the annotator screening and training procedure only at a high level without releasing implementation details. The authors state that “*all annotators have TEM-4 or TEM-8 English proficiency, and have experience in video captioning annotation (e.g., several annotators have previously annotated video-caption pairs for Kling project). Prior to formal annotation, they undergo our specialized training to guarantee the quality of their annotation.*” They further claim to “*ensure annotation quality and consistency, we implemented a rigorous annotator selection and training process. Initially, all potential annotators underwent a trial annotation phase using a shared subset of videos... Through careful evaluation of their trial annotations, we selected only those annotators who demonstrated high consistency, accuracy, and thorough understanding of the annotation guidelines.*” Their captions occasionally shows writing problems such as typos and grammar mistakes, likely because they do not use LLMs to polish the human caption. Overall, the procedure appears rigorous, although no concrete examples or released materials are provided.
- **Issues identified:** Among all benchmarks we manually evaluated, TUNA-Bench provides the highest overall human caption quality and is recommended for evaluating subject motion description. Specification-wise, the authors claim to have a detailed annotation guideline, but it is not publicly released. The captions are strong in describing subject appearance, scene context, and general motion or activity, yet show common human issues in spatial and camera details. Many captions misinterpret camera movement (e.g., saying the camera is moving when it is actually rotating, or mentioning rotation without specifying whether it arcs around a subject or rotates along its own axis). Annotators often confuse “shot transitions (hard cuts)” with “zooming” and rarely describe spatial depth, sometimes labeling a “midground car” as “background.” Overlay elements such as on-screen subtitles are also frequently omitted. Minor hallucinations occur (e.g., “a man on the left”

when centered, “patterned” when the object has plain texture, or “one person” when there are two), but these are relatively infrequent given the level of detail. Overall, this dataset highlights the value of clear specification and quality control.

We emphasize that our evaluation is not a comprehensive study but a motivation for developing precise specification and oversight frameworks to enhance caption quality. We hope our study and released materials lay a foundation for future work on reliable video–language data curation.

B. Specification Details

In this section, we describe how we develop our specification in collaboration with video content creators.

Demographics of content creators. Our specification draws on direct input from professional video creators experienced in camera movement and shot composition. A core team of 30 creators (filmmakers, motion designers, cinematographers, and game artists) meets weekly with paper authors to refine the taxonomy, resolve ambiguous cases, and verify primitive definitions across hundreds of video samples. We additionally collaborate with over 100 creators selected from a pool of more than 600 applicants, who provide feedback, raise clarification questions during training, and help refine definitions and decision rules. Annotators are also drawn from this trained pool, ensuring strong visual literacy and familiarity with the specification. [Figure 7](#) summarizes the distribution of creator experience, age, geographic region, and professional domain.

Specification overview. [Figure 8](#) provides a high-level overview of our specification, organized into five major aspects and their corresponding subaspects.

From primitives to captioning. We convert labeled **primitives** (from CameraBench-Pro) into **pre-captions** by prompting the model with dynamically generated instructions that depend on each video’s labels. Unlike Socratic models [81], which use a fixed template, our prompts change according to the specific primitives. For example, if a video contains multiple people with one clear main subject, the prompt asks the model to describe that subject’s appearance, clothing, and relationships to others. If the video is a scenery shot, the prompt instead tells the model to describe the environment and note the absence of a main subject. [Appendix J](#) includes Python-style pseudocode showing how raw primitives are converted into *subject*, *scene*, *motion*, *spatial*, and *camera* pre-captions.

C. Oversight Details

In this section, we detail our oversight framework covering annotator screening, training, captioning, and human–AI team data curation and incentive structures.

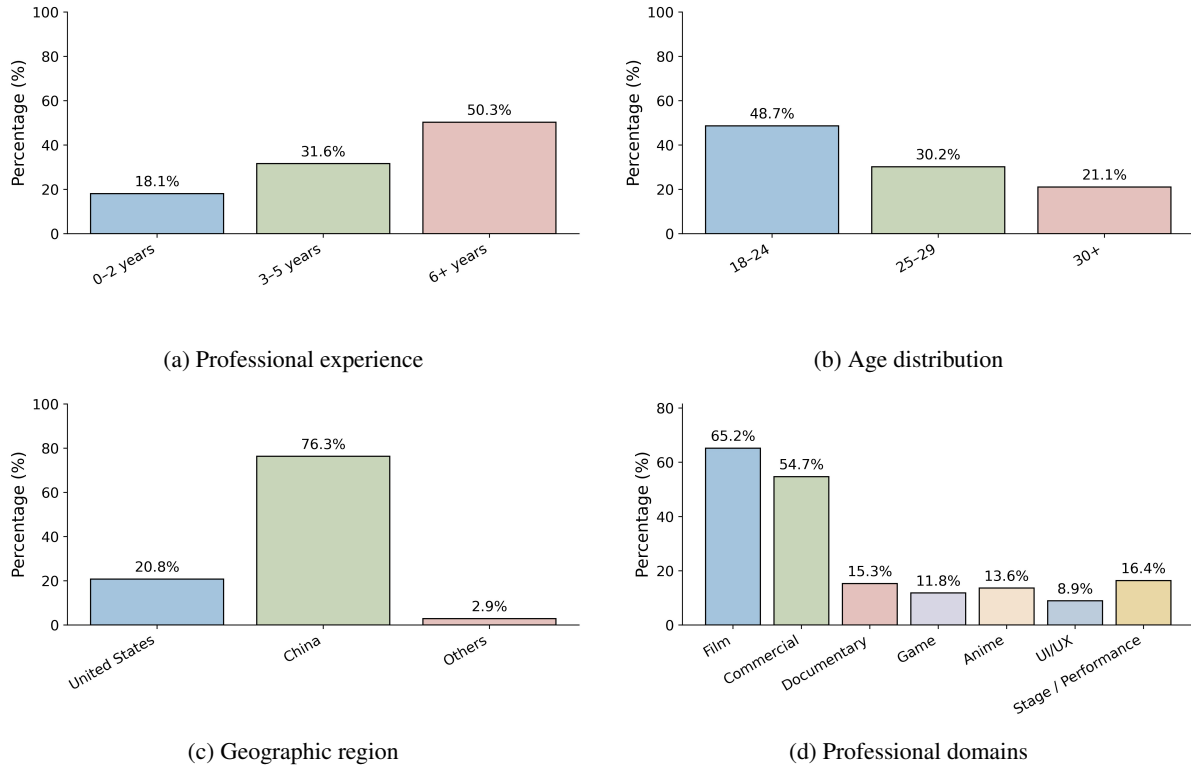


Figure 7. Demographics of video content creators.

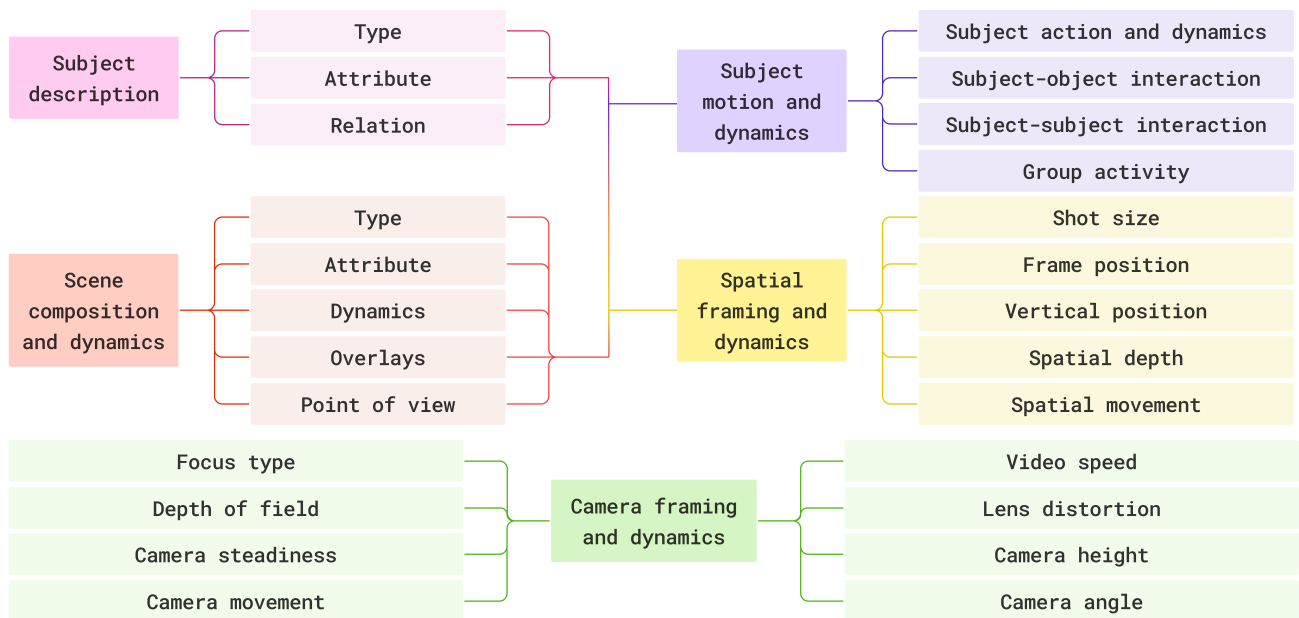


Figure 8. Overview of our specification. Note that motion and spatial captions depend on first completing the subject and scene captions. The diagram shows arrows to reflect this dependency.

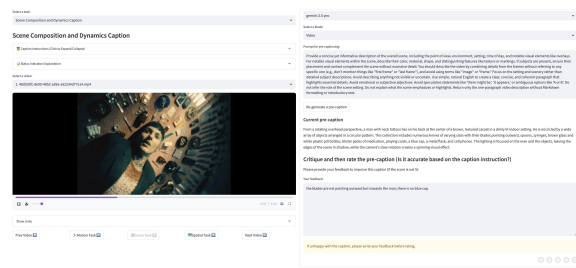
Annotator screening and training. We recruit annotators primarily from our pool of professional video content creators. Similar to CameraBench [35], we find that profes-

sionals learn our training guideline quickly and make fewer mistakes once they begin real tasks. We also include a small number of highly motivated non-professionals who spend

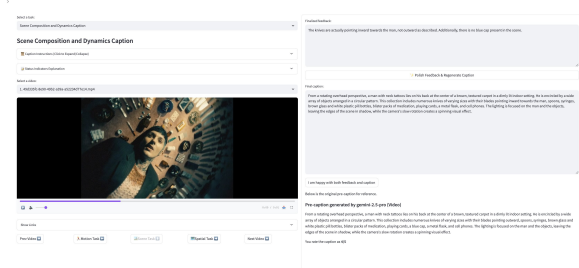
over three months training with our guideline and platform (including paper authors who originally came from computer vision backgrounds but became fluent in creative visual vocabulary through this process), confirming that the guideline is sufficient to train non-professionals to a professional level with enough effort. Screening begins with studying the training materials and completing primitive labeling for two projects: camera motion and video cinematography (including camera setup). Each project requires passing two to six multiple-choice exams. We only accept annotators who score above 90% across all primitives. In practice, most professionals reach 80–90% accuracy within two or three exams; annotators who cannot reach 90% after three attempts in any section are removed. Out of 600 applicants, roughly 50 pass all exams, and about 30 exceed 90%, forming our main annotator pool. These 30 begin with primitive labeling; only after completing more than 300 videos with a strong track record are they promoted to captioning, yielding about 20 captioners. Among them, only around 8 are eventually promoted to reviewer roles for final-stage quality checks.

Platform for human-AI caption curation. [Figure 9](#) shows screenshots of our in-house captioning platform, which supports efficient human–AI collaboration. The interface includes tools for annotators to refine critiques and captions, allows new annotators to shadow expert work, allows reviewers to approve or reject submitted captions, and provides an appeal workflow where annotators can request a regrade by emailing the reviewer and manager. We show the human-AI team structure in [Figure 10](#).

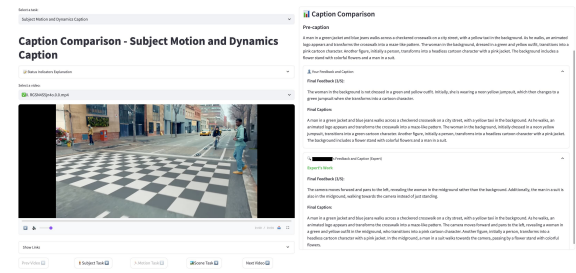
Promotion and salary. Our annotation pipeline uses a quality-based compensation system that rewards accuracy and encourages consistent performance. For captioning, annotators earn a base of \$30 per set (10 videos, 50 captions). Accuracy-based deductions of \$5, \$10, or \$15 apply if any task falls to 70%, 50%, or 30% accuracy; a \$5 bonus is awarded if all tasks reach 90% or above. Annotators with consistently low accuracy may be stopped early and asked to self-correct or complete additional onboarding tasks. Those who achieve 80–90% accuracy for three consecutive sets and complete at least one set per day may be promoted to caption reviewer. Caption reviewers earn a base of \$15 per set, with dynamic increases to \$20, \$25, \$30, or \$35 depending on annotator accuracy (all tasks $\geq 90\%$, $\geq 70\%$, $\leq 70\%$, $\leq 50\%$, or $\leq 30\%$ respectively). Manager spot-checks are conducted at random; if major issues are found, the reviewer must redo the batch, and three such incidents result in suspension.



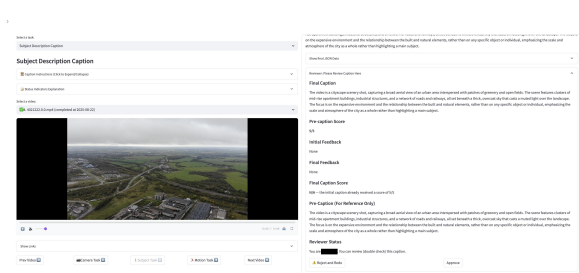
(a) Critiquing model pre-caption



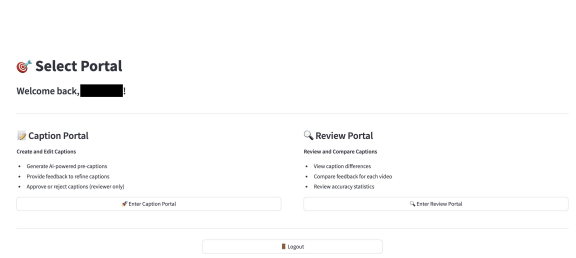
(b) Finalizing critique and post-caption



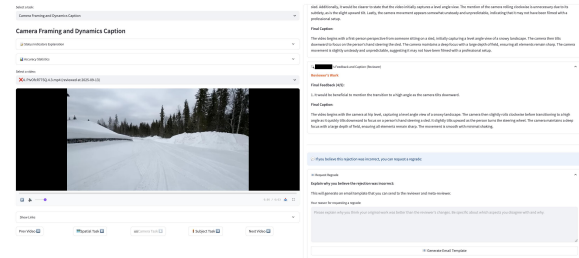
(c) Incoming annotator shadowing expert critique



(d) Reviewer reviewing annotator's work

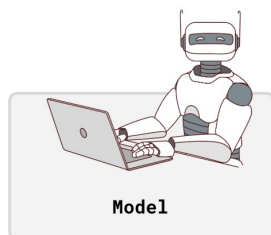


(e) Choosing caption or review portal



(f) Review portal allows annotator appealing rejection

Figure 9. **Example screenshots from our captioning platform.** (a) shows how annotators (or reviewers) use our platform to generate and critique model pre-captions, and (b) shows how they continue polishing the critique and post-caption until satisfied. (c) shows how new annotators can perform training by first completing the above steps, then comparing their critique with expert versions verified by multiple paper authors. (d) reviewers can review annotator work by examining their critique and post-caption, approving good work or rejecting and redoing incorrect submissions. (e) annotators can choose the caption or review portal, and (f) shows how they can appeal via regrade request by sending email to the reviewer.



- Saves human time by drafting initial work.
- Polishes writing under human guidance.



- Submits work for review for base pay
- Gains bonus if correct
- Loses bonus if Reviewer fixes annotator errors
- Promoted to **Reviewer** after 3 high-performing batches (~100 videos).



- Checks annotator submissions.
- Fixes annotator mistakes to earn bonus.
- Receives base pay if no errors found.
- Penalized if reviewer was wrong.



- Handles **regrade requests**.
- Performs **random audits** on reviewer work.
- Promotes high-performing annotators.

Figure 10. **Human-AI team structure.**

Pre-Caption

The video features four light-skinned men outdoors. On the far left, a man with dark curly hair and a full dark beard, wearing a light grey t-shirt and black shorts, initially holds a red and a blue water balloon. To his right, a man with short brown hair and glasses, dressed in a white t-shirt and dark shorts with a light blue stripe, receives the balloons. In the middle, a man with dark hair and a full dark beard, wearing a white t-shirt with a dark graphic print and grey shorts with red accents, initially holds an orange water balloon and then throws it and the blue water balloon (passed to him) into the air. On the far right, a man wearing a red baseball cap backward, a blue t-shirt with a graphic, and dark shorts, holds a knife in his right hand and a red water balloon in his left, which he then pops with the knife, causing water to spray.

Critique

The man with dark curly hair and a full dark beard is holding only one red water balloon, not a blue one. The man with short brown hair and glasses is actually to the left of the first man, not the right. Additionally, there is no light blue stripe on his shorts; it appears to be a light-colored logo, and he is holding a purple water balloon. There is no blue water balloon being passed to the man in the middle, so that detail should be removed. The man wearing a red baseball cap backward is holding the knife with both hands, not just his right hand, and he does not have a red water balloon in his left hand. The red water balloon is thrown by the first man, and the man with the cap pops it.

Post-Caption

The video features four light-skinned men outdoors. On the far left, a man with dark curly hair and a full dark beard, wearing a light grey t-shirt and black shorts, initially holds a red water balloon. To his left, a man with short brown hair and glasses, dressed in a white t-shirt and dark shorts with a light-colored logo, holds a purple water balloon. In the middle, a man with dark hair and a full dark beard, wearing a white t-shirt with a dark graphic print and grey shorts with red accents, initially holds an orange water balloon and then throws it into the air. On the far right, a man wearing a red baseball cap backward, a blue t-shirt with a graphic, and dark shorts, holds a knife with both hands. The red water balloon is thrown by the first man, and the man with the cap pops it, causing water to spray.

ID: bpZc7avrCT4.5.5.mp4

Pre-Caption

The video features a young male diver, wearing blue swim trunks, performing a dive. He is seen mid-air, with his back facing the camera, arms extended, and legs bent. His hair is short and light-colored. The dive concludes with him entering the water, creating a splash. The focus is on his athletic form and the precision of his dive.

Critique

The diver maintains straight legs throughout the dive, without bending them. He starts with his back to the camera, then flips mid-air, rotating his body to face the camera with his chest and face as he prepares to enter the water. The video continues beyond the splash, capturing a brief moment of him submerged beneath the surface.

Post-Caption

The video features a young male diver, wearing blue swim trunks, performing a dive. He is seen mid-air, starting with his back facing the camera, then flipping to rotate his body and face the camera with his chest and face. His legs remain straight throughout the dive, and his hair is short and light-colored. The dive concludes with him entering the water, creating a splash, and the video continues to capture a brief moment of him submerged beneath the surface, highlighting his athletic form and the precision of his dive.

ID: nb69eg95m08.2.2.mp4

Pre-Caption

The video shows the lower bodies of three individuals on an ice rink, all wearing white figure skates. The first individual, initially positioned in a doorway on the right, wears black leggings with two thin white vertical stripes on each outer leg and a dark blue and black top. A second individual, wearing black leggings and a light grey long-sleeved top, skates from left to right across the foreground. Subsequently, a third individual, dressed in black leggings featuring three prominent white vertical stripes on each outer leg and a dark blue and black top similar to the first person's, skates out from the same doorway on the right.

Critique

The video actually features four individuals, not just three. The second individual is wearing a black long-sleeved top, not a light grey one, and enters the ice rink from the center gate, skating toward the left across the scene. A third individual, dressed in black leggings without stripes and a light grey top, also skates out from the same doorway. The fourth skater, wearing all black, follows. Finally, the first skater moves after the other three, skating leftward further into the rink.

Post-Caption

The video shows the lower bodies of four individuals on an ice rink, all wearing white figure skates. The first individual, initially positioned in a doorway, wears black leggings with three thin white vertical stripes on each outer leg and a dark blue and black top. A second individual, wearing black leggings and a black long-sleeved top, enters the ice rink from the center gate and skates toward the left across the scene. A third individual, dressed in black leggings without stripes and a light grey top, skates out from the same doorway. The fourth skater, wearing all black, follows. Finally, the first skater moves after the other three, skating leftward further into the rink.

ID: cd248d6e8326138876d9f828943ac894463a938ba557ccff8171d483dbe91d14.8.mp4

Pre-Caption

From a low-angle, point-of-view perspective, a man with a dark complexion surfs inside a wave. He wears a black, short-sleeved wetsuit and black shorts with a red patch on the left leg. Initially, he is in a kneeling position on his orange and red surfboard, which has a yellow nose. He then rises to a standing position, leaning back with his arms outstretched to his sides for balance.

Critique

The surfer begins in a crouching position, transitions to kneeling, and then stands upright with arms extended for balance, contrary to the initial description of starting in a kneeling position. Additionally, there is no red patch on the left leg of the shorts, and the surfer has a light skin tone, not a dark complexion.

Post-Caption

From a low-angle, point-of-view perspective, a man with a light complexion surfs inside a wave. He wears a black, short-sleeved wetsuit and black shorts. Initially, he is in a crouching position on his orange and red surfboard, which has a yellow nose. He then transitions to a kneeling position before rising to stand upright, leaning back with his arms outstretched to his sides for balance.

ID: 1yTv_SR2uUo.2.7.mp4

Figure 12. Subject caption examples.

Pre-Caption

The scene is set on the stage of a TEDx Vancouver event. A woman stands in the center of a square red rug, which is placed on a dark, raised platform. Behind her, a large screen displays the "TEDx Vancouver" logo and a live feed of her presentation. The backdrop is a large, geometrically faceted wall illuminated with a deep blue light. To the left of the stage, another "TEDx Vancouver" logo is constructed from large, white, illuminated letters, positioned next to a cluster of white spheres. The overall lighting is theatrical, focusing on the stage and leaving the surrounding area in darkness.

Critique

1. The description inaccurately states that a single large screen displays both the "TEDx Vancouver" logo and a live feed of the presentation. In reality, there are two separate screens: the "TEDx Vancouver" logo is displayed on a large screen at the top right, while a live feed of the presentation is shown on another screen at the top left.
2. The "TEDx Vancouver" logo is described incorrectly. The "TEDx" portion is in red, while "Vancouver" is constructed from large, white letters that are not illuminated.
3. The description of the stage setup is incorrect regarding the placement of the white decorative elements. The cluster of white decorations is positioned behind the "TEDx Vancouver" logo, not next to it.
4. The backdrop is a large, geometrically patterned wall illuminated with a deep blue light, not faceted.

Post-Caption

The scene is set on the stage of a TEDx Vancouver event. A woman stands in the center of a square red rug, which is placed on a dark, raised platform. Behind her, two separate screens are visible: the "TEDx Vancouver" logo is displayed on a large screen at the top right, while a live feed of her presentation is shown on another screen at the top left. The "TEDx" portion of the logo is in red, while "Vancouver" is constructed from large, white letters. The backdrop is a large, geometrically patterned wall illuminated with a deep blue light. To the left of the stage, the "TEDx Vancouver" logo is constructed from large, white letters, and a cluster of white spheres is positioned behind it. The overall lighting is theatrical, focusing on the stage and leaving the surrounding area in darkness.

ID: LNH0MFcznE.0.4.mp4

Pre-Caption

In a daytime scene set against a vast, clear blue sky, a uniquely designed aircraft flies over a barren landscape of rolling brown hills. The aircraft, featuring a long fuselage with a reddish-brown and white pattern and multiple wings, has a large, silver-colored, bulbous structure at its nose. As it travels from left to right, it appears to be generating a large, fluffy white cloud that expands behind it, contrasting with the otherwise clear sky.

Critique

1. The description of the aircraft having multiple wings is incorrect; it actually has two wings. Additionally, the paint scheme is white on the underside and reddish-brown on the top.
2. The aircraft is moving from right to left relative to the frame, not left to right.
3. The aircraft is not generating a large cloud; the cloud was already present, and the aircraft is simply flying away from it.
4. Since only the tops of the hills are visible, the detail of the "barren landscape" should be removed.
5. The bulbous structure is not the nose of the aircraft; instead, this silver object, which resembles a conical shape, is being dragged behind the tail.

Post-Caption

In a daytime scene set against a vast, clear blue sky, a uniquely designed aircraft flies over the tops of rolling brown hills. The aircraft, featuring a long fuselage with a reddish-brown top and white underside, has two wings and is towing a large, silver-colored, conical structure behind its tail. As it travels from right to left, it appears to be moving away from a large, fluffy white cloud, which contrasts with the otherwise clear sky.

ID: uMCK4nneeU.2.3.mp4

Pre-Caption

The scene takes place on a soccer field with vibrant, bright green turf marked by a white line. Initially viewed from a low angle, the setting is brightly lit. The perspective then changes to a close-up against a stark, overexposed white background, where a person's long black hair flows dynamically. The individual is wearing a red jersey with black stripes over a black long-sleeved undershirt, along with black and white cleats.

Critique

1. This is a 2D video, so please avoid mentioning details about angles or depth.
2. The final shot is not a close-up but rather a medium shot.

Post-Caption

The scene unfolds on a soccer field with vibrant, bright green turf marked by a white line. The setting is brightly lit. The perspective shifts to a medium shot to show an anime character's upper body against a stark, overexposed white background, where a person's long black hair flows dynamically. The individual is wearing a red jersey with black stripes over a black long-sleeved undershirt, along with black and white cleats.

ID: ff60lwrjnI.0.0.mp4

Pre-Caption

In bright daylight, this outdoor scene shows two red, round tomatoes being thrown towards a transparent glass surface in the foreground. A gold-colored door handle is affixed to the right side of this glass. Behind the glass, a white, paneled wall, featuring a small, circular metallic fixture, serves as the background. One of the tomatoes has a small green sticker. Initially, two hands, appearing bluish-green, release the tomatoes. Upon impact with the glass, the tomatoes splat, their red liquid and pulp spreading outwards. A dark shadow is cast onto the white paneled background from an object to the right.

Critique

The paneled wall in the background appears to be a garage door; specifying this detail would enhance clarity. The hands are not bluish-green but rather light-skinned. The shadow cast on the background seems to be that of a person, likely belonging to the individual whose arm was on the left side of the frame.

Post-Caption

In bright daylight, this outdoor scene shows two red, round tomatoes being thrown towards a transparent glass surface in the foreground. A gold-colored door handle is affixed to the right side of this glass. Behind the glass, a white, paneled garage door, featuring a small, circular metallic fixture, serves as the background. One of the tomatoes has a small green sticker. Initially, two light-skinned hands release the tomatoes. Upon impact with the glass, the tomatoes splat, their red liquid and pulp spreading outwards. A dark shadow, likely of a person, is cast onto the white paneled garage door.

ID: kxcw8lSn8xw.3.2.mp4

Figure 13. Scene caption examples.

<p>Pre-Caption</p> <p>A person swings a tennis racket forward in a forehand motion, striking a red bell pepper. Upon impact, the pepper explodes, sending a cloud of small red fragments and seeds flying forward and outward. The person completes the swing as the pieces of the pepper scatter through the air.</p>	<p>Pre-Caption</p> <p>A pair of puppet-like hands lifts a small, gold-colored clock resting on a piece of red cloth. The hands then carefully fold the cloth's corners over the clock, creating a small, square bundle, which they hold in their palms.</p>
<p>Critique</p> <p>The caption is missing a crucial detail: before swinging the tennis racket with their right hand, the person first throws the bell pepper into the air with their left hand.</p>	<p>Critique</p> <ol style="list-style-type: none"> 1. The term "lifts" in the initial description might be misleading, as the hands keep the cloth and clock at a consistent height rather than raising them. 2. Furthermore, it is important to mention that at the end of the video, the hands place the bundled clock into a drawer that was visible in the background below the hands and clock.
<p>Post-Caption</p> <p>A person throws a red bell pepper into the air with their left hand before swinging a tennis racket forward in a forehand motion with their right hand, striking the pepper. Upon impact, the pepper explodes, sending a cloud of small red fragments and seeds flying forward and outward. The person completes the swing as the pieces of the pepper scatter through the air.</p> <p>ID: kxcw@iSn8kw.1.2.mp4</p>	<p>Post-Caption</p> <p>A pair of puppet-like hands gently wraps a small, gold-colored clock resting on a piece of red cloth. The hands carefully fold the cloth's corners over the clock, creating a small, square bundle, which they hold in their palms. At the end of the video, the hands place the bundled clock into a drawer that was visible in the background below.</p> <p>ID: SqKYraJRNwo.3.3.mp4</p>
<p>Pre-Caption</p> <p>The silhouetted man initially stands still in the dimly lit room. He then makes subtle movements with his hands in front of his body, as if adjusting an object he is holding. Subsequently, he slowly turns his head to his left.</p>	<p>Pre-Caption</p> <p>The skier swiftly navigates through deep snow in a forested area, holding a selfie stick in their right hand. They extend their left hand for balance while maneuvering between snow-covered trees. As they move at high speed, they create a spray of snow that partially obscures the view, demonstrating dynamic and skillful skiing.</p>
<p>Critique</p> <p>Initially, it appears that the man is taking a photograph using a flash. He then turns to face the camera while holding his camera and walks towards it. Finally, he turns his head to his left, which corresponds to the right side of the frame.</p>	<p>Critique</p> <p>The individual in the video is not skiing but rather snowboarding.</p>
<p>Post-Caption</p> <p>The silhouetted man initially stands still in the dimly lit room, appearing to take a photograph using a flash. He then turns to face the camera while holding his camera and walks towards it. Finally, he turns his head to his left, which corresponds to the right side of the frame.</p> <p>ID: fSNFUFdVSTU.5.0.mp4</p>	<p>Post-Caption</p> <p>The snowboarder swiftly navigates through deep snow in a forested area, holding a selfie stick in their right hand. They extend their left hand for balance while maneuvering between snow-covered trees. As they move at high speed, they create a spray of snow that partially obscures the view, demonstrating dynamic and skillful snowboarding.</p> <p>ID: xhrAGJv1QJA.0.7.mp4</p>

Figure 14. Motion caption examples.

Pre-Caption

The video opens with a wide shot, the camera positioned at the subject's level. A female diver is in the center-right of the frame, standing in the middle ground on a light blue diving board. To her left, also in the middle ground, several judges are seated on high chairs. The background consists of grey tiered spectator stands, partially filled with an audience, occupying the upper portion of the frame. The bright blue swimming pool is in the foreground at the bottom. As the diver performs her dive, moving upwards from the board and then arcing downwards, the camera tilts down, transitioning to a full shot from an above-subject perspective. The diver enters the water in the bottom-center of the frame, now in the foreground, with the splash occurring in the center foreground as the pool surface comes to dominate the view.

Critique

The female diver is actually in the foreground, not the middle ground. The judges are seated on high chairs in the background, not the middle ground, and they are positioned on the left side of the frame, not specifically to the left of the diver. The pool should be described as being in the middle ground, not the foreground. Additionally, the splash occurs in the middle ground, not the foreground.

Post-Caption

The video opens with a wide shot, the camera positioned at the subject's level. A female diver is in the foreground, standing on a light blue diving board. In the background, several judges are seated on high chairs on the left side of the frame. The background consists of grey tiered spectator stands, partially filled with an audience, occupying the upper portion of the frame. The bright blue swimming pool is in the middle ground at the bottom. As the diver performs her dive, moving upwards from the board and then arcing downwards, the camera tilts down, transitioning to a full shot from an above-subject perspective. The diver enters the water in the middle ground of the frame, with the splash occurring in the middle ground as the pool surface comes to dominate the view.

ID: 1XzposKQzvs.2.8.mp4

Pre-Caption

The video begins with a view of a shattered, blue, crystalline surface occupying the foreground and filling the frame. A circular opening is positioned in the center of this surface, acting as a middle ground element that reveals a lighter blue, slightly blurred background. A yellow-orange frog then emerges from the background, moving through this central opening towards the foreground. As the frog appears, the shot transitions to a full shot of the frog, which remains centered within the frame as it moves through the opening. The camera is positioned below the frog.

Critique

The description of the blue crystalline structure should indicate that it extends from the foreground to the middle ground, rather than being confined to the foreground. Furthermore, the statement "A yellow-orange frog then emerges from the background" is incorrect. The frog actually appears in the middle ground at the opening, not the foreground. Initially, the frog emerges from the bottom of the frame and then moves to the center of the frame at the opening.

Post-Caption

The video begins with a view of a shattered, blue, crystalline surface extending from the foreground to the middle ground and filling the frame. A circular opening is positioned in the center of this surface, acting as a middle ground element that reveals a lighter blue, slightly blurred background. A yellow-orange frog then emerges from the bottom of the frame, moving through the middle ground towards the central opening. As the frog appears, the shot transitions to a full shot of the frog, which remains centered within the frame as it moves through the opening. The camera is positioned below the frog.

ID: bX86aZ304-0.3.2.mp4

Pre-Caption

The video begins with a medium shot of a man positioned in the center of the frame and in the middle ground. In the background, shelves line the top and left of the frame. A vintage television is visible in the middle ground on the bottom left, while a boombox sits in the middle ground on the right. As the man speaks, he leans forward, moving from the middle ground toward the foreground. The shot simultaneously tightens to a medium close-up, emphasizing his movement and expression. The camera remains positioned at the subject's level throughout the scene.

Critique

1. The shelves are positioned along the top, left, and right sides of the frame.
2. A vintage television is initially visible in the background on the left, but it disappears from the frame as the camera zooms in. Meanwhile, a boombox is situated in the background on the right.
3. The man is consistently in the foreground rather than the middle ground, maintaining a leaning-forward pose without changing his stance. As he speaks, he uses hand gestures to emphasize his points.

Post-Caption

The video begins with a medium shot of a man positioned in the center of the frame and in the foreground. Shelves line the top, left, and right sides of the frame. A vintage television is initially visible in the background on the left but disappears from the frame as the camera zooms in. Meanwhile, a boombox is situated in the background on the right. The man maintains a leaning-forward pose and uses hand gestures to emphasize his points as he speaks. The shot tightens to a medium close-up, highlighting his movement and expression, while the camera remains positioned at the subject's eye-level throughout the scene.

ID: kjI2gyb2hR4.0.1.mp4

Pre-Caption

In a full shot, a group of people walks away from the camera down the center of a long, cylindrical tunnel, occupying the middle ground. The tunnel's corrugated walls frame the scene on the left and right, converging toward a partially open door in the center background, from which a bright light emanates. As the group moves from the middle ground deeper into the background, more people appear in the foreground on the left and right, expanding the group's formation. Simultaneously, the white light from the center background intensifies, eventually expanding to fill the entire frame. This growing light first silhouettes the figures and then bleaches the entire scene to white, completely obscuring them. The camera is positioned at the subjects' level.

Critique

The cylindrical tunnel extends from the foreground into the background, rather than just occupying the middle ground. The group primarily remains in the foreground and middle ground, as the camera follows them at a slightly slower pace. It is a wide shot, not a full shot. The description stating that the light "bleaches the entire scene to white, completely obscuring them" is inaccurate, as the outline of the group remains visible.

Post-Caption

In a wide shot, a group of people walks away from the camera down the center of a long, cylindrical tunnel, which extends from the foreground into the background. The tunnel's corrugated walls frame the scene on the left and right, converging toward a partially open door in the center background, from which a bright light emanates. The group primarily remains in the foreground and middle ground as the camera follows them at a slightly slower pace. As the group moves deeper into the tunnel, more people appear in the foreground on the left and right, expanding the group's formation. Simultaneously, the white light from the center background intensifies, eventually expanding to fill the entire frame. This growing light first silhouettes the figures, and while it brightens the scene significantly, the outline of the group remains visible. The camera is positioned at the subjects' level.

ID: uWCGk4nne0.2.3.mp4

Figure 15. Spatial caption examples.




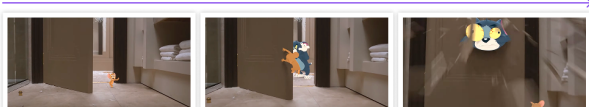
 <p>Pre-Caption</p> <p>This is a fixed, slightly unsteady shot from a ground-level, level angle. The camera uses a shallow depth of field, with the focus set on the foreground. Initially, the man lying on the ground in the background is out of focus. A woman's feet then walk into the frame and come into sharp focus.</p> <p>Critique</p> <p>Initially, the scene is out of focus, and the man is positioned in the midground, not the background. Subsequently, the woman's feet enter the frame in the foreground and stay in sharp focus.</p> <p>Post-Caption</p> <p>This is a fixed, slightly unsteady shot from a ground-level, level angle. The camera uses a shallow depth of field, with the focus set on the foreground. Initially, the scene is out of focus, with the man lying on the ground positioned in the midground. A woman's feet then walk into the frame in the foreground and come into sharp focus.</p> <p>ID: p1QedrDKzZM.2.3.mp4</p>	 <p>Pre-Caption</p> <p>From a level, ground-level perspective, the camera uses a shallow depth of field to focus on the man in the midground. The shot is slightly unsteady, with minor, unintentional shaking.</p> <p>Critique</p> <p>Please revise the description of the camera focus to: "The camera initially focuses on the grass in the foreground, utilizing a shallow depth of field. It then smoothly transitions its focus to the man in the midground using rack focus. As the man moves into the foreground, the camera employs focus tracking to consistently keep him in focus." Additionally, please mention that the camera slightly tilts upward.</p> <p>Post-Caption</p> <p>From a level, ground-level perspective, the camera initially focuses on the grass in the foreground, utilizing a shallow depth of field. It then smoothly transitions its focus to the man in the midground using rack focus. As the man moves into the foreground, the camera employs focus tracking to consistently keep him in focus. The shot is slightly unsteady, with minor, unintentional shaking, and the camera slightly tilts upward.</p> <p>ID: e_ofen9SDeM.0.2.mp4</p>
 <p>Pre-Caption</p> <p>From a smooth, high-angle aerial perspective, the camera initially moves forward, approaching a white airplane positioned on the runway. It then transitions into a clockwise arcing motion around the aircraft. The deep focus keeps the entire scene, including the airplane and the surrounding airport environment, in sharp detail.</p> <p>Critique</p> <p>The camera also moves downward continuously, which should be noted. Additionally, the term "deep focus" is inaccurate, as the background appears blurred. As the camera approaches the plane, both the ground closer to the camera and the background are out of focus.</p> <p>Post-Caption</p> <p>From a smooth, high-angle aerial perspective, the camera initially moves forward, approaching a white airplane positioned on the runway. It then transitions into a clockwise arcing motion around the aircraft while continuously moving downward. The focus highlights the airplane in sharp detail, while both the ground closer to the camera and the background appear blurred.</p> <p>ID: m1Pt1YVKP8.9.7.mp4</p>	 <p>Pre-Caption</p> <p>From a smooth, ground-level perspective, the camera moves forward and then tilts upward, shifting to a low angle as Jerry shuts the door. The shot freezes on Tom's head smashing through the door. The scene is captured with a deep focus, keeping all elements sharp.</p> <p>Critique</p> <p>The scene should actually be captured with a shallow focus, concentrating on the mid-ground door, rather than a deep focus that keeps all elements sharp.</p> <p>Post-Caption</p> <p>From a smooth, ground-level perspective, the camera moves forward and then tilts upward, shifting to a low angle as Jerry shuts the door. The shot freezes on Tom's head smashing through the door. The scene is captured with a shallow focus, concentrating on the mid-ground door.</p> <p>ID: H4AZhSSRqKk.2.19.mp4</p>

Figure 16. Camera caption examples.

Table 3. **Ablation of critiques.** We analyze three critique failure types (*Inaccurate*, *Incomplete*, and *Unhelpful*) which correspond to violations of precision, recall, and constructiveness. We also compare critiques generated by Gemini-2.5-Pro using both the video and the pre-caption with those from a blind Gemini-2.5-Pro that hallucinates critiques based only on the pre-caption.

Critique Type	Quality Metrics			Task Performance		
	Precision	Recall	Constructiveness	Caption	Reward	Critique
Blind Gemini-2.5	–	–	–	8.3	32.1	12.6
Gemini-2.5	–	–	–	11.9	59.9	16.9
Inaccurate critique	✗	✓	✓	11.3	45.5	14.1
Incomplete critique	✓	✗	✓	11.7	54.7	18.5
Unhelpful critique	✓	✓	✗	12.5	65.0	21.2
Our critique (w/o quality check)	–	–	–	13.8	70.7	23.0
Our critique (w/ quality check)	✓	✓	✓	17.0	86.8	26.9

E. Critique Quality Matters

In this section, we first identify the core properties of good critiques, show that critique quality is essential for the success of post-training, and analyze why critiques collected in prior work are often low quality due to not being constructive. See Figure 4 for a summary.

Our oversight framework enforces quality critiques. High-quality critiques are essential for successful post-training. In our workflow, annotators or reviewers must write critiques that directly guide the model to produce the final post-caption. This requirement naturally enforces three properties. First, the critique must be **accurate** and cannot include hallucinated information. Second, it must be **complete** and point out every error that needs correction. Third, it must be **constructive** by explaining not only what is wrong but also how to fix it. Because the critique is the instruction used to turn the pre-caption into the correct post-caption, these qualities are automatically maintained. To show why these properties matter, we use Gemini-2.5-Pro to take our ground-truth critiques and inject controlled errors: (1) making critiques *inaccurate* by replacing a correct point with an incorrect one (e.g., maliciously changing “*the man is wearing black, not white*” to “*the man is wearing blue, not white*”) or inserting a new point when the pre-caption is already perfect; (2) making critiques *incomplete* by removing a necessary correction (except when the critique is already “*The caption is accurate and requires no edit*”); and (3) making critiques *non-constructive* by removing the fix (e.g., changing “*the man is wearing black, not white*” to “*the man is not wearing white*”). We also compare against critiques generated by Gemini-2.5-Pro when given the video and pre-caption, by a blind Gemini-2.5-Pro model that hallucinates a critique using only the pre-caption, and by critiques written by annotators before the second-stage quality control.

Critique quality matters. Table 3 shows that weakening any of the three properties above leads to a substantial drop in performance. In contrast, adding a second stage of quality control noticeably improves results, presumably because it strengthens the critique along all three dimensions. Interestingly, critiques generated by Gemini-2.5-Pro perform poorly even when the model is given access to the video, suggesting that current state-of-the-art models still struggle to produce critiques that are accurate, complete, and constructive.

Critiques in prior work are often non-constructive. We find that critiques collected in prior work are frequently non-constructive, which may partly explain why they are not consistently helpful for downstream training or evaluation. We study OpenAI’s GDC [51], which collects critiques for topic-based summarization and text question answering tasks, and MM-RLHF [85], which collects critiques for video question answering. Using ChatGPT to classify critiques as constructive or non-constructive, combined with manual verification over ~ 200 samples per dataset (ensuring $\geq \sim 95\%$ accuracy), we estimate a lower bound on the fraction of critiques that are non-constructive. As shown in Table 4, more than half of the critiques in these datasets (and over 90% in MM-RLHF for responses requiring critique) are non-constructive. Random qualitative examples are shown in Table 5 (GDC summarization), Table 6 (GDC text QA), and Table 7 (MM-RLHF video QA). This contrasts sharply with our approach: in GDC and MM-RLHF, annotators are asked only to critique model outputs, but not to use that critique to correct them. In our oversight framework, critiques are written expressly to *fix* the model output, which naturally enforces precision, recall, and constructiveness.

Table 4. **Prior work collects non-constructive critiques.** We report the distribution of constructive versus non-constructive critiques for each benchmark task. Surprisingly, these datasets contain a high proportion of unhelpful critiques that only point out what is wrong without explaining what is correct or how to fix it.

Dataset / Benchmark	Task	Size	Critique Distribution	
			Constructive	Non-Constructive
OpenAI GDC [51]	Topic-based Summarization	1728	53%	47%
	Text Question Answering	1851	51%	49%
MM-RLHF [85]	Video Question Answering	8751	10%	90%

Question	Answer	Critique	Is the Critique Constructive?	Label
What is the purpose of the U.S red flag list?	To require that U.S companies treat customers named on the list with caution. US suppliers are to no longer use existing licenses to sell these customers products or service installed Equptment.	Replace capital "I" in "service". Use the correct spelling for "Equptment".	The critique provides specific instructions on how to improve the answer by correcting the capitalization and spelling errors. It does not require additional information from the source material to make these changes. The editor can directly implement these corrections.	Yes
What are some related fields to circle packing?	Sphere packing is a related field.	The passage also mentions ring lemma and Doyle spirals as related topics.	The critique provides specific additional related topics, "ring lemma" and "Doyle spirals," that can be directly added to the answer. This allows for a concrete improvement to the answer without needing to refer back to the source material.	Yes
Summarize the global reaction to Trump's remarks.	Botswana summoned the US ambassador to clarify if they thought of Botswana as a shithole country. Rupert Colville, UN Human Rights spokesperson, said the comments were "shocking and shameful." Haiti's US Ambassador Paul Altidor believes that the remarks Trump made about immigrants who come to the US are wrong.	The answer should include the response of the African Union.	The critique specifies that the answer should include the response of the African Union. However, it does not provide any details about what the African Union's response was. Without this information, I cannot improve the answer without referring back to the source material.	No
Why is the protagonist behaving in such a self-destructive way?	This information can't be found.	The information is not in the story but it can be implied.	The critique suggests inferring behavior, but without access to the source material, it is impossible to determine what those implications might be. The critique does not provide specific guidance or details for a concrete improvement.	No
Summarize all the information relating to the mountainous Austrian Tyrol.	Averil Demuth's first story was set in the Austrian Tyrol. The girl Trudi, the boy Hansel, the cow Lotti, the dog Berni, the hen Griselda and the raven Kraak all go up the mountain to see Riese the giant, and then all come down again. Eleanor Farjeon gave the book a positive review. The House in the Mountains was also set in the Austrian Tyrol.	The sentence is unnecessary.	The critique states that "The sentence is unnecessary," but it does not specify which sentence is unnecessary. Since there are multiple sentences, it is unclear which one should be removed without referring back to the source material.	No
Who is the protagonist of this story?	The protagonist of this story is Stephen.	Describe him more.	The critique asks for a description of the protagonist, Stephen, but does not provide any specific details or characteristics about him. Without additional information from the source material, it is not possible to expand on the description of Stephen. Therefore, the critique does not provide enough information to improve the answer without referring back to the source material.	No

Table 5. Critique analysis on OpenAI GDC [51] (Topic-based Summarization). We show random examples from the test set (1728 samples) and find that more than 47% of critiques in this task are non-constructive.

Question	Answer	Critique	Is the Critique Constructive?	Label
Why does the narrator suggest that they must terminate the patient's pregnancy?	The narrator suggests that they must terminate the patient's pregnancy because the current sector has too many male children. This is due to the fact that the State has mandated that all pregnancies must be terminated if the fetus is male.	The second sentence "mandated that all pregnancies must be terminated if the fetus is male" should be changed to "has regulates the gender distribution".	The critique provides specific guidance on how to improve the answer by suggesting a change to the wording. It implies the issue is about regulating gender distribution, allowing for concrete improvement without source material.	Yes
How fast is Jiranová?	Jiranová is a Czech runner and sprinter who has personal bests in the 400 metres, 100 metres hurdles, and 60 metres hurdles. Her 400 metres best is 53.77 seconds. Her 100 metres hurdles best is 13.37 seconds, and her 60 metres hurdles best is 8.31 seconds.	The first sentence is unnecessary for answering the question.	The critique specifies that the first sentence is unnecessary. This implies the answer should focus solely on her personal best times. The editor can improve the answer by removing the extraneous sentence.	Yes
What is the voting process in Switzerland?	Information can't be found.	This information can be found in the article.	The critique states the information exists but does not provide any specific details or corrections . It does not imply what the correct information is, so the editor cannot make a concrete improvement without the source.	No
How old was Sanuth when his cricket career took off?	Sanuth was 25 years old when his cricket career took off.	The answer is incorrect.	The critique states the answer is incorrect but does not provide any information about the correct age . Without additional details, the editor cannot determine the correct age to improve the answer.	No
How is a consensus reached?	No answer.	No answer to the question.	The critique points out the lack of an answer but does not provide any details or implications about what should be included . The editor needs the source material to construct a response.	No
Why did the Leicester mayor find the move by the health secretary disappointing?	The mayor found the move by the health secretary disappointing because it means the premier's planned speech will shift away from coronavirus and onto the lockdown, making COVID an even greater focus.	"...because it means the premier's planned speech will shift away from coronavirus and onto the lockdown..." This sentence lacks clarity and accuracy.	The critique claims the sentence lacks clarity/accuracy but does not provide specific information on how to clarify or correct it . It is impossible to improve without understanding the correct context from the source.	No

Table 6. **Critique analysis on OpenAI GDC (Text Question Answering)**. We show random examples from the test set. Out of 1851 test samples, we find that 49% of critiques in this task are non-constructive.

Question	Answer	Critique	Is the Critique Constructive?	Label
What is depicted in the frames throughout the video?	The frames depict a hand holding a small, toy car-shaped chocolate.	The description is incorrect; it is a car-shaped chocolate candy.	The critique specifies the object is a “car-shaped chocolate candy” rather than a “toy car.” This provides enough information to correct the description without referring back to the source video.	Yes
What colors are showcased in the video?	The video showcases a variety of colors, including green grass, blue sky, white clouds, and snow-capped mountains.	Correct answer, but the description and analysis are minimal.	The critique points out minimal analysis but does not specify what additional description is needed . Without specific guidance, it is impossible to improve the answer without viewing the video.	No
What kind of tool is the person holding in their hands?	The person is holding a tool, possibly a screwdriver or a similar implement.	Failure to identify the tool’s purpose.	The critique mentions a failure to identify the purpose but does not provide information about what the purpose is . It is impossible to specify the tool’s use without this info.	No
What are the two primary colors represented on the Christmas tree?	The two primary colors represented on the Christmas tree are green and purple.	Green description error.	The critique mentions a “green description error” but does not provide specific information about what the error is or how to correct it . Without details, the answer cannot be improved.	No
Who is the main focus of the video and what action do they take?	I’m unable to view the video and identify the main focus or actions taken. Could you describe the scene for me?	No answer to the question.	The critique notes the lack of answer but does not provide any information or guidance on what the correct answer should be . The editor cannot improve the answer without the source content.	No
What is the individual wearing in the video?	The individual is wearing a long-sleeved pajama top with a pattern, possibly featuring small illustrations, and matching pajama pants.	The description of the top is accurate, but the color of the pants is incorrect.	The critique indicates the pants color is incorrect but does not provide the correct color . The editor would need to view the video to determine the correct color.	No

Table 7. **Critique analysis on MM-RLHF [85] (Video Question Answering)**. We show random examples from MM-RLHF that have a non-empty critique (58% out of the 8751 test samples). Of those, we find that 90% are non-constructive.

F. Post-training ablations

In this section we provide additional details on post-training experiments. We ablate reference-based evaluation metrics to measure how well they align with human Likert scale ratings, report additional reference-based metrics for caption and critique generation, and include the task instructions used for the five captioning tasks. We also describe the formats used in the SFT dataset, including how we enforce a balanced dataset for training reward models for both captions and critiques, and finally report all training hyperparameters for offline experiments (SFT, DPO, RLHF-V).

LLM-Judge-Direct. This method directly sends the reference caption and the generated candidate caption to GPT-4o and taking the P(Yes) [34] for the question below:

Reference caption: {reference}

Candidate caption: {candidate}

Does the candidate caption accurately match the reference caption in terms of content and meaning? Answer only Yes or No.

LLM-Judge-Instruct. We also try sending the task-specific instruction to GPT-4o for computing P(Yes):

Task Instruction: {instruction}

Reference caption: {reference}

Candidate caption: {candidate}

Does the candidate caption accurately follow the task instruction and match the reference?

Answer only Yes or No.

Task instructions. The instruction for each captioning task is shown below:

- **Subject:** *Provide a concise yet informative description of the subjects in this video, including their types, appearances (e.g., clothing, facial expressions, gender, ethnicity, color, shape), and poses. When multiple subjects are present, clearly distinguish them using unique traits, position, actions, or relationships, and describe them in temporal or prominence-based order to ensure clarity.*
- **Scene:** *Provide a concise yet informative description of the overall scene, including the point of view, environment, setting, time of day, overlays, and notable visual elements. If subjects are present, the scene description should complement their descriptions by establishing their location and possible context. Aim to give enough detail to convey the setting while avoiding unnecessary information.*
- **Motion:** *Provide a concise yet informative description of the subject’s motion in this video, including individual actions, subject–object or subject–subject interactions, and group activities when a crowd is present. Event order matters—if multiple actions occur, present them in chronological order.*
- **Spatial:** *Provide a concise yet informative description of how subjects and elements are spatially framed within the scene, including the shot size of the subject (or the shot size of the scenery if there is no salient subject), their 2D position within the frame, spatial depth within the scene (foreground, middle ground, background), height relative to the camera, and any notable spatial movement.*
- **Camera:** *Provide a concise yet informative description of the video and camera configuration, including playback speed, lens distortion (if present), camera angle, camera height relative to the ground plane, camera movements (translation, rotation, zooming, steadiness, speed, intensity, and complexity), and focus (depth, focus plane, and any changes in focus).*

Ablating reference-based evaluation metrics. We compare a range of reference-based metrics to assess how well they align with human judgement. Since we collect 1–to-5 human Likert scores for all model pre-captions and also generate adversarial negative captions that automatically receive a score of 1 by hallucinating all details (explained later in this section), we can meta-evaluate a metric by computing its pairwise accuracy across all items: for every pair of captions in the evaluation set (consisting of both pre-captions and adversarial negative captions), the metric is correct if it assigns a higher score to the caption with the higher human Likert score. We follow Pairwise Accuracy with tie optimization [17, 28], which is more robust for this setting because its tie-optimization procedure handles the many tied pairs that naturally arise from discrete 1–to-5 Likert scores. The metrics we evaluate include SPICE, ROUGE-L, METEOR, CIDER, BLEU-1/2/4, as well as the two versions of LLM-Judge described above. Figure 17 shows that LLM-Judge-Instruct using GPT-5 performs best, since providing the captioning instruction slightly improves performance. For the main experiments, we continue to use BLEU-4 as the most reproducible classic metric.

Evaluating critique generation via caption-revision. For critique generation, we introduce a more reliable metric, **Critique-Revision**, which avoids comparing against the reference critique directly (since critiques vary widely in style, such as whether annotators use bullet points to point out mistakes). Instead, we use the generated critique to revise the caption with ChatGPT using the prompt shown below, and then compare the revised caption to the reference caption: Given a video caption and user feedback, please provide an improved version of the caption that addresses the

feedback. Note that the user feedback could be poorly written, so please try your best to guess what it means.

Original caption: {pre_caption}

User feedback: {critique}

Respond with the improved caption only, without quotation marks or JSON formatting.

Results with more reference-based metrics. Table 8 reports additional reference-based metrics for caption and critique generation.

Details for the offline SFT dataset. Below, we describe the eight formats used during SFT training. We also apply text cleaning by stripping newlines from all collected captions and critiques.

(1) Caption Generation

- **Input:** Video, Task Instruction
- **Output:** Caption

(2) Critique Generation

- **Input:**
 - Video, Task Instruction, Caption
 - Provide a critique of the video’s caption based on how accurately it follows the task instruction. Point out what is wrong or missing and how to fix it. If the caption is already accurate, output: “The caption is accurate and requires no edits, so it should remain exactly the same.”
- **Output:** Critique

(3) Caption Reward (VQAScore)

- **Input:**
 - Video, Task Instruction, Caption
 - Does the video’s caption accurately follow the task instruction? Please answer Yes or No only.
- **Output:** Yes (if human Likert scale is 5) or No (if human Likert scale is 1-4)
- **Note:** To prevent the reward model from collapsing to a constant prediction, we balance Yes and No examples following [29]: for each post-caption labeled Yes, we treat the corresponding pre-caption as a No example if its human score is below 5, and exclude pairs where the pre-caption already scores a perfect 5. A variant that directly predicts the Likert score via text generation performs worse. See Figure 18 for results.

(4) Critique Reward (VQAScore)

- **Input:**
 - Video, Task Instruction, Caption, Critique
 - Does this critique of the video’s caption provide accurate and constructive feedback to help the caption better follow the task instruction? Please answer Yes or No only.
- **Output:** Yes or No
- **Note:** We balance Yes and No answers by generating adversarial negative critiques that are inaccurate, incomplete, or non-constructive, following the procedure described in our critique-quality ablations.

(5) Caption Revision

- **Input:**
 - Video, Task Instruction, Caption
 - Provide an improved caption that better follows the task instruction. If the original caption is already accurate, keep it exactly the same with no edits.
- **Output:** Revised Caption

(6) Caption Revision (with Given Critique)

- **Input:**
 - Video, Task Instruction, Caption, Critique

- Provide an improved caption that better follows the task instruction. You are provided a critique of the caption with respect to the task instruction as context. If the original caption is already accurate, keep it exactly the same with no edits.

- **Output:** Revised Caption
- **Note:** During training, critique is human-generated.

(7) Critique-based Caption Revision

- **Input:**
 - Video, Task Instruction, Caption
 - Provide an improved caption that better follows the task instruction, after first writing a critique. If the caption is already accurate enough, the critique should be: ‘The caption is accurate and requires no edits, so it should remain exactly the same.’ And keep the original caption exactly the same with no edits. Write your critique after ‘Critique:’, then on a new line write the improved caption after ‘Improved Caption:’
- **Output:** (Critique, Revised Caption)

(8) Caption Scoring

- **Input:**
 - Video, Task Instruction, Caption
 - Score the video’s caption based on how well it follows the task instruction. Rate 1-5 (1=poor, 5=excellent).
- **Output:** Likert scale score (1-to-5)
- **Note:** This format is used only for the reward-scoring ablation in [Figure 18](#) and is not included in any other experiments. To train this format more effectively, because model pre-captions almost never receive a score of 1, we add adversarially generated negative captions labeled with score 1 to balance the score distribution in the training set.

Constructing negative responses for DPO and RLHF-V. Our offline dataset can easily pair each positive response with a negative one. For caption generation, we use the pre-caption when it has a score below 5, or skip it otherwise. For critique generation, we sample an adversarial critique that is inaccurate, incomplete, or non-constructive. For reward modeling, we flip the Yes/No answer.

Training details for SFT, RLHF-V, DPO. We explore three fine-tuning approaches on Qwen3-VL-8B-Instruct. For supervised fine-tuning (SFT), we perform full-parameter training with the vision tower and multi-modal projector frozen, using a learning rate of 3×10^{-5} , batch size of 10 per device with 2 gradient accumulation steps, and train for 3 epochs with cosine learning rate scheduling and 5% warmup. We use ZeRO Stage 3 optimization with BF16 precision and process videos at 8 FPS with a maximum of 128 frames. For preference optimization, we apply LoRA fine-tuning (rank 16, targeting all modules) using DPO with $\beta = 0.1$ and the sigmoid loss, training on the RLHF-V dataset for 3 epochs with a learning rate of 5×10^{-6} , batch size of 1 per device with 8 gradient accumulation steps, and cosine scheduling with 10% warmup. For reward modeling, we use identical hyperparameters to our DPO setup. All approaches use a maximum sequence length of 2048 tokens and maintain the same video processing settings across experiments.

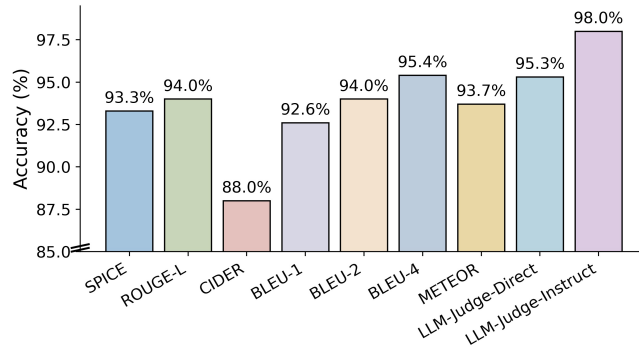


Figure 17. **Comparing reference-based metrics.** We evaluate each reference-based metric by measuring its pairwise accuracy (with tie optimization [17]) when comparing scores between any two captions in the evaluation set, which includes model pre-captions rated with human 1-to-5 Likert scores and adversarial negative captions that automatically receive a score of 1. LLM-Judge-Instruct achieves the best performance overall by sending caption instruction to GPT-5, which helps the model more reliably compare the reference caption and the candidate caption, while BLEU-4 is the most reliable classic metric.

Table 8. **More reference-based metrics for captioning and critique generation.** We additionally report ROUGE-L for both captioning and critique generation, LLM-Judge-Instruct scores for captions, and Critique-based-Revision scores for critiques.

Method	Caption Generation			Critique Generation		
	BLEU-4	ROUGE-L	LLM-Judge-Instruct	BLEU-4	ROUGE-L	Caption-Revision
<i>Open-source models</i>						
Qwen3-VL-8B-Instruct [72]	2.6	4.2	4.3	29.6	47.4	4.0
Qwen3-VL-72B-Instruct [72]	5.3	8.4	8.8	32.5	52.0	7.1
<i>Closed-source models</i>						
GPT-4o [47]	4.2	6.5	6.7	37.2	56.8	5.0
GPT-5	5.5	8.5	8.8	42.5	64.2	6.5
Gemini-2.5-Pro [15]	6.0	9.3	9.6	45.0	67.5	7.0
Gemini-3.1-Pro	5.1	7.9	8.2	46.8	70.2	7.5
<i>Caption-only post-training (Qwen3-VL-8B-Instruct)</i>						
RLHF-V (Caption)	10.2	15.4	15.8	3.3	5.1	4.2
DPO (Caption)	10.4	15.7	16.2	3.3	5.2	4.3
SFT (Caption)	13.9	20.8	21.5	3.8	5.9	4.8
SFT + RLHF-V (Caption)	14.6	21.9	22.6	4.1	6.4	5.2
SFT + DPO (Caption)	14.6	21.8	22.5	3.9	6.1	5.0
<i>Full data post-training (Qwen3-VL-8B-Instruct)</i>						
RLHF-V (All)	14.2	21.3	22.0	24.2	37.4	11.0
DPO (All)	14.3	21.5	22.2	24.0	37.1	10.8
SFT (All)	17.0	25.7	26.3	26.9	41.5	12.5
SFT + RLHF-V (All)	17.8	26.8	27.5	27.5	42.4	13.0
SFT + DPO (All)	17.8	26.7	27.4	27.4	42.2	12.8

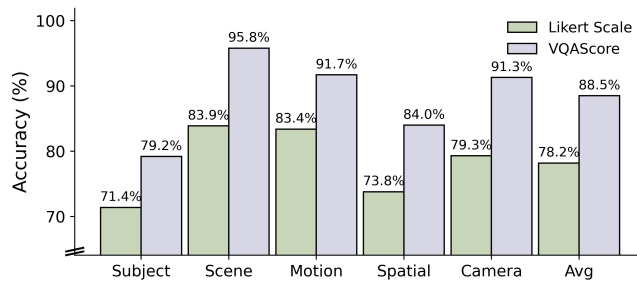


Figure 18. **Comparing reward models using VQAScore versus text generation.** We find that VQAScore (a logit-based probability score) significantly outperforms asking the model to directly output a Likert-scale score from 1 to 5 after supervised fine-tuning (SFT). Therefore, we stick to VQAScore for reward modeling in this work.

Table 9. **Test-time scaling strategies for reward modeling.** We compare four reward scoring (VQAScore) modes: **Direct** computes $P(\text{Yes})$ in a single pass; **Critique-First** generates a critique then computes $P(\text{Yes})$; **Critique-Last** computes $P(\text{Yes})$ then generates a critique; **Self-Critique-Conditioned** generates a self-critique in pass 1 and computes $P(\text{Yes} \mid \text{critique})$ in pass 2. In addition, Self-Consistency [?] replaces $P(\text{Yes})$ with the empirical ratio of Yes answers over N rollouts. Inference passes count total forward passes per sample. **Critique-Last** gives the best overall performance, showing that adding critique at inference time can yield further gains. However, Critique-First and Self-Critique-Conditioned underperform, likely because the Qwen3-VL model was not sufficiently trained on these formats to follow them reliably; increasing the scale of critique-augmented training data may improve their performance.

Mode	Passes	Reward Modeling Accuracy				
		Subject	Scene	Motion	Spatial	Camera
<i>VQAScore-based methods</i>						
Direct	1	90.8	88.3	90.4	94.9	93.6
Critique-First	1	79.1	72.9	80.7	90.7	74.6
Critique-Last	1	89.5	89.6	93.2	94.9	96.0
Self-Critique-Conditioned	2	80.4	78.1	83.5	94.9	81.7
<i>Self-Consistency using 5 rollouts</i>						
Direct	5	92.1	89.7	91.8	95.7	94.8
Critique-First	5	81.2	75.6	83.1	92.0	77.4
Critique-Last	5	92.3	91.2	94.4	96.3	97.1
Self-Critique-Conditioned	10	82.5	80.8	85.4	96.3	83.9

G. Inference-Time Scaling

G.1. Test-time Scaling for Reward Modeling

While directly generating Yes/No is already effective for reward modeling, we study whether generating a critique alongside the answer can further improve performance. We evaluate both VQAScore (probability of Yes) and Self-Consistency (ratio of Yes to No over 5 rollouts).

Results. Table 9 shows that outputting the critique after Yes/No (Critique-Last) marginally outperforms direct VQAScore, suggesting that inference-time critique can yield further gains.

G.2. Test-time Scaling for Caption Generation

We ablate different approaches to scale caption generation performance using inference-time compute. For cost estimation, we separate *generation cost* (caption, critique, or revision) from *reward cost* (scoring a completed sequence), which is significantly cheaper due to KV caching. We treat caption and critique generation as equal cost, and count a critique-based revision as $2\times$ generation cost because it requires producing a self-critique followed by a revised caption.

Parallel (best-of- N) approaches. These methods generate N candidates independently and select the best one using the reward model. All candidates can be produced in parallel, so total cost scales directly with N .

- (1) **Best-of- N Caption.** Sample N captions in parallel and select the highest-scoring one using the reward model.
 - **Generation cost:** N
 - **Reward cost:** N
- (2) **Best-of- N Revision.** Generate one caption and produce N independent revisions. The reward model scores all N revised captions, and we select the best revision.
 - **Generation cost:** $1 + N$
 - **Reward cost:** N
- (3) **Best-of- N Critique-then-Revision.** Generate one caption, produce N critiques, and revise once per critique. Each revised caption is scored with the reward model to select the best one.

Table 10. **Test-time scaling strategies for caption generation.** We compare parallel Best-of- N strategies and sequential iterative strategies. Generation cost counts caption, critique, and revision calls (each normalized to cost 1, with critique-based revision counted as cost 2), while reward cost counts calls to the reward model. The final five columns report caption performance at different test-time scales ($N \in \{1, 4, 8, 16\}$). Vanilla Best-of- N Caption yields the best performance given the same inference budget. While critique-based approaches show marginal improvement, we believe increasing the scale of critique-augmented training data to improve critique generation quality can make these approaches perform even better in future work.

Mode	Strategy	Inference Cost		Result (BLEU-4)			
		Generation	Reward	N=1	N=4	N=8	N=16
Parallel (Best-of- N)	Best-of- N Caption	N	N	16.5	20.4	22.8	25.1
	Best-of- N Revision	$1 + N$	N	16.5	18.9	20.3	21.2
	Best-of- N Critique-then-Rev.	$1 + 2N$	N	15.5	18.5	20.1	21.0
	Best-of- N Critique-based Rev.	$1 + 2N$	N	17.0	19.2	20.6	21.4
	Best-of- N Critique	$2 + N$	N	15.1	17.8	19.2	20.1
Sequential (Iterative)	Iterative Revision	$1 + N$	0	15.6	16.9	17.3	17.5
	Iterative Critique-then-Rev.	$1 + 2N$	0	16.1	17.6	18.1	18.4
	Iterative Critique-based Rev.	$1 + 2N$	0	16.8	17.5	17.8	18.0

- **Generation cost:** $1 + 2N$
 - **Reward cost:** N
- (4) **Best-of- N Critique-based Revision.** Generate one caption and run N critique-based revision steps (each step internally generates a self-critique and a revised caption). All revised candidates are scored to select the best one.
- **Generation cost:** $1 + 2N$
 - **Reward cost:** N
- (5) **Best-of- N Critique.** Generate one caption and produce N critiques. We use the *critique reward model* (trained specifically to evaluate critique quality) to score the N critiques, select the best critique, and perform a single critique-based revision based on that critique.
- **Generation cost:** $2 + N$
 - **Reward cost:** N

Sequential approaches. These methods refine a single caption over multiple dependent steps. Each iteration uses the previous output, so computation accumulates linearly with N .

- (1) **Iterative Revision.** Generate one caption and apply up to N sequential revisions.
- **Generation cost:** $1 + N$
 - **Reward cost:** 0
- (2) **Iterative Critique-then-Revision.** At each step, generate a critique of the current caption and revise accordingly, repeated for N steps.
- **Generation cost:** $1 + 2N$
 - **Reward cost:** 0
- (3) **Iterative Critique-based Revision.** Perform N critique-based revision steps, each internally generating a critique and a revision.
- **Generation cost:** $1 + 2N$
 - **Reward cost:** 0

Results. Table 10 compares these strategies across computational budgets ($N \in \{1, 4, 8, 16\}$). Vanilla best-of- N caption achieves the strongest performance at $N \geq 4$, reaching 25.1 BLEU-4 at $N = 16$, despite being the simplest parallel strategy. At $N = 1$, critique-based revision scores highest (17.0), suggesting that critique can improve a single caption, but this advantage disappears as N grows and the reward model can select from a larger candidate pool. Parallel methods consistently outperform sequential iterative approaches, which show diminishing returns as N increases (e.g., iterative critique-based revision improves only from 16.8 to 18.0 across $N = 1$ to 16). This highlights that reward-based selection over parallel candidates is more effective than sequential refinement without discriminative filtering.

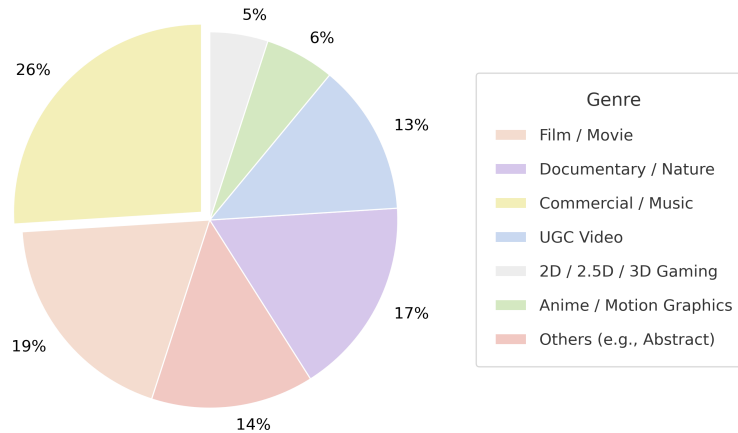


Figure 19. **Genre distribution for ~150k curated videos.**

Limitations. We focus on pure parallel and sequential strategies and do not explore hybrid search procedures. We also do not incorporate reward-guided early stopping or pruning mechanisms (e.g., discarding low-scoring candidates in a beam-search-like process). These extensions are left for future work.

H. Video Generation

In this section, we report statistics of the ~150K professional videos we curated from YouTube, summarize their genre distribution, show additional qualitative comparisons with zero-shot Wan, and include a human study on prompt following using 200 prompts generated from our test set.

Video statistics. We hire our annotators to manually source high-quality and diverse YouTube channels spanning films, commercials, music videos, user-generated content (including vlogs, first-person GoPro footage, and etc.), documentary and nature videos (including drone shots, animals, and scenery), gaming footage (3D, 2.5D isometric, 2D side view, top-down, and etc.), and others (including dashcam, screen recordings, abstract visuals, and etc.). All videos follow the standard YouTube license for non-commercial use. We then apply TransNetV2 [53] to segment the videos and use a high threshold (0.15) to discard any clips with potential hard cuts or soft transitions, resulting in ~150K videos with an average duration of 4.6 seconds (we discard clips shorter than one second). The distribution of genres (measured by channel type) is shown in Figure 19.

Captioning. We compare two approaches to caption these videos: (1) applying the zero-shot Qwen3-VL-8B-Instruct model (without any training on our dataset) to directly generate the five captions given our task instruction, and (2) using our own Qwen3-VL-8B-Instruct model fine-tuned via SFT. Because the five captions are too long to append naively (Wan2.2 accepts only 512 tokens or roughly up to 400 words), and because they contain substantial overlap (e.g., motion captions often also describe the subjects in detail), we merge them into a single summary caption that combines all subject, scene, motion, spatial, and camera information. This merging is done using ChatGPT with the prompt attached below.

Please merge the following five captions into a single, comprehensive caption that describes the video completely without any redundancy.

Caption Types:

1. Subject: Describes the subjects/people in the video
2. Scene: Describes the scene composition and environment
3. Motion: Describes the movement and dynamics of subjects
4. Spatial: Describes the spatial relationships and framing
5. Camera: Describes camera movements and framing choices

Input Captions:

{captions}

Instructions:

Table 11. **Human preference study.** Models fine-tuned on higher-quality captions receive higher average human Likert scores.

Model	Avg. Human Likert Score
Base	3.23 ± 0.78
ZS-Caption	3.64 ± 0.54
SFT-Caption	3.95 ± 0.45

Table 12. **Training hyperparameters** for Wan2.2 fine-tuning.

Hyperparameter	High-Noise Stage	Low-Noise Stage
timestep range	0 → 0.417	0.417 → 1
resolution	832×480	
num frames	49	
learning rate	1e-5	
lr schedule	cosine decay	
epochs	2	
optimizer	AdamW	
DeepSpeed stage	ZeRO-2	

1. Use the SPATIAL caption as your BASE structure - it provides the core visual description and framing
2. Merge MOTION and CAMERA captions into the spatial description to create a temporally coherent narrative that describes how things change over time
3. Add information from SUBJECT and SCENE captions ONLY if they contain unique details not already covered in the Spatial caption
4. Eliminate ALL redundant information - if the same detail appears in multiple captions, mention it only ONCE
5. Preserve the EXACT wording from the original captions - do NOT paraphrase
6. When describing temporal changes, integrate motion and camera movements in chronological order to show how the scene evolves
7. CRITICAL: Every unique detail from all five captions must appear in the final merged caption - nothing should be omitted
8. Do NOT add any information not present in the original captions
9. Return only the merged caption without any additional text or formatting

Goal: A single, temporally coherent caption based on the Spatial description, with Motion and Camera information merged chronologically, and Subject/Scene details added only when they provide new information.


Finetuning text-to-video generation models. We follow the default hyperparameters to finetune Wan2.2, as shown in Table 12.

Human study. We compare three Wan2.2-T2V-14B variants: **Base**, **ZS-Caption** (Wan2.2 finetuned using captions generated by zero-shot Qwen3-VL-8B-Instruct), and **SFT-Caption** (Wan2.2 finetuned using captions from our SFT-trained Qwen3-VL-8B-Instruct). We evaluate all three models on 200 prompts summarized from the five captions using GPT-4o. As shown in Table 11, the SFT-Caption variant achieves the best prompt following, measured using 1-to-5 human Likert scores averaged across three expert annotators. We attach qualitative samples in Figure 20 (dolly-zoom out), Figure 21 (isometric 2.5D gaming perspective), Figure 22 (camera rolling), Figure 23 (Dutch angle), Figure 24 (rack focus), Figure 25 (speed ramping effect), Figure 26 (side-view gaming perspective), Figure 27 (watermark overlays), Figure 28 (camera rising from underwater to above water), Figure 29 (changing shot size), and Figure 30 (revealing shot).

220-word Caption


In a medium shot, a young, light-skinned man with a serious expression stands in the foreground of a dusty Old West town street during what appears to be sunset, with a warm, golden-orange light illuminating the environment. He wears a dark, wide-brimmed cowboy hat, a brown suit jacket over a matching vest, a white collared shirt, and a red tie. A pocket watch chain is visible across his vest, and a pistol in a holster is strapped to his right hip. Groups of men and women dressed in late 19th-century Western attire line the left and right sides of the frame, occupying the middle ground and background, standing still and observing him. The men wear hats and suits, while the women are in long dresses, all in muted earth tones. The scene is flanked by two-story, weathered wooden buildings with porches. At first, the young man has his hands behind his back, then his right hand moves to his right hip, as if he is about to draw a gun. The camera is positioned at the subject's level throughout the scene. The shallow depth of field keeps the man in the foreground in focus, while the background is only slightly blurred. This sequence showcases a dolly zoom effect, with the camera smoothly moving forward while simultaneously zooming out, enhancing the tension as the man remains standing in place in the center of the wide, unpaved dirt street.

Wan2.2 (Ours SFT)




User rating: 4.33

Wan2.2 (Zero-Shot)



User rating: 3.67

Expert feedback:



- The camera should **dolly-zoom out**, not just **dolly backward**.
- The character should **stay still**, not **walk forward**.

Figure 20. Video generation example: dolly-zoom out.

400-word Caption

The video clip presents a disorienting and immersive top-down perspective of a person lying flat on the floor, surrounded by an extensive assortment of scattered objects. The subject, a man with tattoos on his neck and arms, is positioned squarely at the center of the frame, immediately drawing the viewer's attention as the focal point of the scene. He gazes upward toward the camera with a vacant or distressed expression, his stillness contrasting sharply with the chaotic environment around him. Visible tattoos on his neck and upper chest further contribute to a rugged and possibly troubled appearance, hinting at a deeper backstory or emotional state. He is dressed in a dark suit jacket layered over a patterned sweater, creating a visual contrast between his somewhat formal attire and the disorder of the surrounding space. His face appears pale, and his eyes are wide open, conveying an unsettling combination of confusion, distress, and perhaps the impression of a drug-induced or hallucinatory state. His mouth hangs slightly open, reinforcing the sense that he is either in shock or experiencing a disorienting episode. The camera rolls clockwise directly above him, maintaining a strict bird's-eye view that emphasizes the radial arrangement of the scattered items. There is a slight fisheye distortion from the lens, stretching the edges of the frame and adding to the uneasy, warped feeling of the shot. The objects surrounding the man are diverse and include syringes, pill bottles, spoons, ashtrays, playing cards, and various small containers and tools, as well as small glass vials, pill packets, a lighter, and a small bowl, forming a chaotic circle around his body. The floor is carpeted, and the lighting is dim and moody, casting deep shadows and giving the scene a somber, almost cinematic tone. The video begins with a top-down, fisheye lens shot, and the shot begins with a slow, deliberate rotation around the person, causing the surrounding objects to appear to spin and blur around the central figure, giving the impression of a slow, hypnotic spin. As the video progresses, the camera slowly rotates around the subject, maintaining the top-down perspective. As the video progresses, the camera slowly rotates around the subject, maintaining the top-down perspective. The camera remains at a consistent height above the floor, ensuring that the subject's face and upper body remain centered in the frame throughout the sequence. The man remains motionless throughout.

Wan2.2 (Ours SFT)




User rating: 4.33

Wan2.2 (Zero-Shot)



User rating: 3.33

Expert feedback:



- The camera is expected to **maintain a constant height above the floor**, rather than **rising upward**.
- The camera should **roll clockwise throughout**, not **move upward**.

Figure 21. Video generation example: isometric (2.5D) game perspective.

I. Human Captioning Policy

I.1. Captioning Overview

(1) Subject Description

Includes: Subject Type, Subject Attributes, Relationships Between Multiple Subjects, Subject Transitions

Example: "The camera first shows a man in a black shirt with a brown beard standing to the right of a tree, facing the camera. The camera then moves left, revealing a young boy in red looking at the tree."

(2) Subject Motion & Dynamics

Includes: Subject Action, Subject Dynamics, Subject-Object Interactions, Subject-Subject Interactions, Group Action

Example: "The video shows two teams playing soccer. A player from the red team sprints across the field and collides with an opponent in blue, causing both to fall."

400-word Caption

The video clip presents a disorienting and immersive top-down perspective of a person lying flat on the floor, surrounded by an extensive assortment of scattered objects. The subject, a man with tattoos on his neck and arms, is positioned squarely at the center of the frame, immediately drawing the viewer's attention as the focal point of the scene. He gazes upward toward the camera with a vacant or distressed expression, his stillness contrasting sharply with the chaotic environment around him. Visible tattoos on his neck and upper chest further contribute to a rugged and possibly troubled appearance, hinting at a deeper backstory or emotional state. He is dressed in a dark suit jacket layered over a patterned sweater, creating a visual contrast between his somewhat formal attire and the disorder of the surrounding space. His face appears pale, and his eyes are wide open, conveying an unsettling combination of confusion, distress, and perhaps the impression of a drug-induced or hallucinatory state. His mouth hangs slightly open, reinforcing the sense that he is either in shock or experiencing a disorienting episode. **The camera rolls clockwise directly above him, maintaining a strict bird's-eye view that emphasizes the radial arrangement of the scattered items.** There is a slight fisheye distortion from the lens, stretching the edges of the frame and adding to the uneasy, warped feeling of the shot. The objects surrounding the man are diverse and include syringes, pill bottles, spoons, ashtrays, playing cards, and various small containers and tools, as well as small glass vials, pill packets, a lighter, and a small bowl, forming a chaotic circle around his body. The floor is carpeted, and the lighting is dim and moody, casting deep shadows and giving the scene a somber, almost cinematic tone. The video begins with a top-down, fisheye lens shot, and the shot begins with a slow, deliberate rotation around the person, causing the surrounding objects to appear to spin and blur around the central figure, giving the impression of a slow, hypnotic spin. As the video progresses, the camera slowly rotates around the subject, maintaining the top-down perspective. **As the video progresses, the camera slowly rotates around the subject, maintaining the top-down perspective. The camera remains at a consistent height above the floor,** ensuring that the subject's face and upper body remain centered in the frame throughout the sequence. The man remains motionless throughout.

Wan2.2 (Ours SFT)




User rating: 4.33

Wan2.2 (Zero-Shot)



User rating: 3.33

Expert feedback:



1. The camera is expected to **maintain a constant height above the floor**, rather than **rising upward**.
2. The camera should **roll clockwise throughout**, not **move upward**.

Figure 22. Video generation example: clockwise roll.

180-word Caption

A close-up shot frames a young man with curly dark hair, positioned in the right foreground of the frame against a dark, indistinct background. The camera is held at his eye level and is slightly tilted, **creating a fixed dutch angle with a tilted horizon.** The camera employs a shallow depth of field, focusing on the foreground, and smoothly zooms in, maintaining minimal shaking throughout, with no changes in shot size and subject placement. The young man maintains a focused expression with wide-open eyes, suggesting attentiveness or surprise. The lighting changes from a warm tone to a pronounced red hue, highlighting his facial features and casting shadows that emphasize his expression. The subject exhibits slight head movement and rapid breathing, adding subtle dynamism to the scene. The dimly lit environment and subdued lighting create a moody atmosphere, with the transition in lighting suggesting a change in mood or situation. The background remains indistinct, emphasizing the subject's expression and the shift in lighting. The overall scene is intimate and focused, with minimal distractions from the surroundings.

Wan2.2 (Ours SFT)



User rating: 5.00

Wan2.2 (Zero-Shot)



User rating: 4.00

Expert feedback:



There should be a **dutch angle**, not a **level angle**.

Figure 23. Video generation example: Dutch angle.

(3) Scene Composition & Dynamics

Includes: Point-of-View, Overlays, Scene Type, Scene Attributes, Scene Dynamics

Example: "A first-person perspective of walking on a sunny beach, with ocean waves crashing onto the shore."

(4) Spatial Framing & Dynamics

Includes: Shot Size, Spatial Position, Spatial Motion

Example: "The video opens with a medium shot of a man positioned on the left side of the frame. As he walks forward toward the camera, moving from the midground to the foreground, the shot gradually transitions into a close-up of his face at the center of the frame."

(5) Camera Framing & Dynamics

Includes: Video Playback Speed, Lens Distortion, Camera Height, Angle, Focus, Depth of Field, Camera Steadiness, Camera Movement

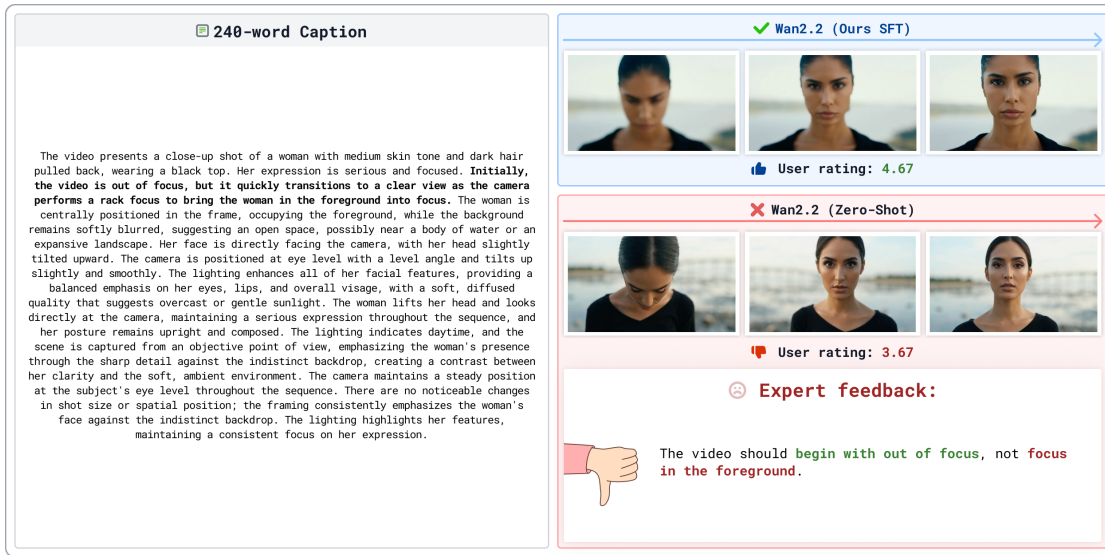


Figure 24. Video generation example: rack focus.

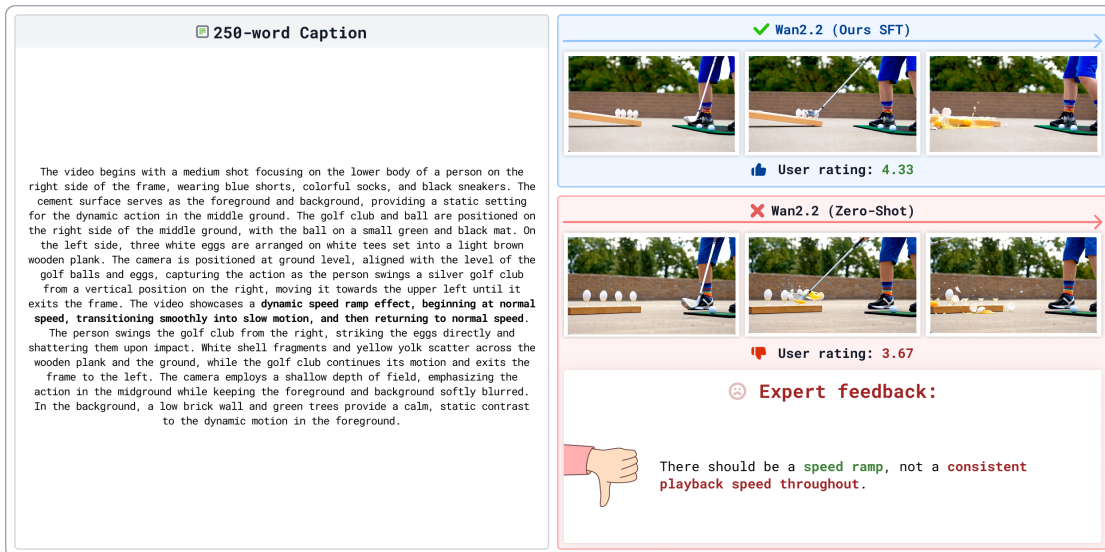


Figure 25. Video generation example: speed ramp.

Example: “The video is in slow motion, capturing a low-angle, ground-level shot of a skateboarder performing a trick. A fisheye lens introduces barrel distortion, while deep focus ensures sharp clarity throughout the frame. The camera tracks the skateboarder from the side, trucking right, with an unsteady, handheld-like motion.”

I.2. High-Level Captioning Policy

Subject Motion and Spatial Framing are dependent on the Subject/Scene descriptions.

- **Focus on the Visual:** Ensure your captions accurately reflect the visual content by describing only what is visible in the frame. If speculation is necessary (e.g., about events outside the frame or the subject’s thoughts), use phrases like “it could be that” or “this might suggest that” to indicate uncertainty. Please ignore the audio information from the video.
- **Conciseness:** Keep descriptions clear and to the point, avoiding unnecessary details.
- **Objectivity:** Avoid personal opinions or interpretations in your descriptions.

📄 230-word Caption

In a dynamic, wide, **side-view shot** set in a lush, sun-dappled forest during the day, a large, rotund, dark purple creature with a bear-like form and small, dark eyes carries a small, white, spirit-like creature on its back. The smaller creature is luminous, with a slender body, long, pointed ears, and thin limbs, giving it a delicate, ethereal appearance. Positioned in the center-right of the frame, the creatures are situated in the middle ground, moving forward on a wooden fence. The large creature sways from side to side as it takes slow, deliberate steps, while the small creature on its back remains seated, bobbing gently in sync with the larger creature's gait. The camera smoothly trucks right, tracking their movement to the right and upward as they progress through the forest. The scene is framed by large, thick-trunked trees with dense green foliage in the middle ground and background, with hazy golden light filtering through the canopy. The foreground and middle ground are covered in vibrant green grass, moss, and small purple flowers, with a nest-like structure made of woven twigs visible in the bottom-left. Particles of light drift through the air, enhancing the ethereal atmosphere. As the creatures continue, the camera rises slightly as they jump onto a rickety wooden swing, capturing the scene's dynamic motion.

✔ Wan2.2 (Ours SFT)



👍 User rating: 5.00

✘ Wan2.2 (Zero-Shot)



👎 User rating: 3.67

🗨 Expert feedback:

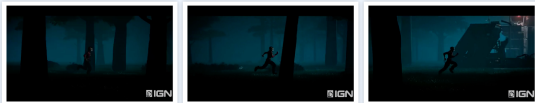
1. The shot should be **side-view perspective**, not a **regular 3D perspective**.
2. The camera should **truck right** to track the characters, not **arc counterClockwise** around them.

Figure 26. Video generation example: side-view game perspective.

📄 230-word Caption

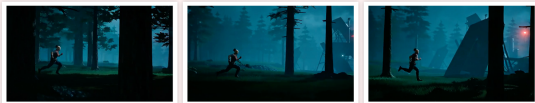
In this wide shot, a small human figure is consistently visible in the center of the middle ground, partially obscured by multiple silhouetted trees in the foreground. The individual is wearing dark pants and a dark reddish top, with light skin noticeable on their hands and part of their face, which is partially covered by a mask over the lower half. The camera, from a level angle, smoothly tracks the avatar in a side-view game, initially trucking left slightly, pausing, and then trucking right to follow the figure as it runs from left to right through a dark, wooded landscape. The figure moves at a steady pace across a grassy clearing, passing silhouetted trees and heading towards distant, dimly lit structures. As the camera progresses, various silhouetted trees pass through the foreground and middle ground, eventually revealing large, dark industrial structures in the right background. The scene is set within a dark, misty forest at night, with the camera moving to the right across a grassy clearing populated by the dark silhouettes of trees. As the character approaches the industrial area, large, angular, industrial-style structures and a single red light become visible through the dense, blue-green mist. The camera maintains a side-on perspective at the subject's height, and a white logo with the letters "IGN" is overlaid in the bottom right corner throughout the shot.

✔ Wan2.2 (Ours SFT)



👍 User rating: 4.67

✘ Wan2.2 (Zero-Shot)



👎 User rating: 4.00

🗨 Expert feedback:

There should be a **watermark "IGN"** on the bottom right, not appear **without a watermark**.

Figure 27. Video generation example: watermark addition.

- **Order Matters:** Describe subjects and events in temporal order when applicable, and prioritize the most important elements first for clarity.
- **Clarity:** Your caption should provide enough context so someone who hasn't seen the video understands what's happening.

I.3. Subject Description Policy

Instructions for Subject Description

Provide a concise yet informative description of the subjects in this video.

(1) Subject Types:

- Specify the subject's type precisely (e.g., "man," "woman," "dog," "car," "tree"). Avoid vague terms like "thing" or "item."
- If the subject type is ambiguous, provide your best judgment and explain your reasoning.

230-word Caption

The video captures a serene, tropical beach scene from a low-angle perspective, **initially with the camera positioned just below the water's surface**. In the foreground, the clear, shallow water gently ripples, revealing submerged rocks and pebbles, reflecting the surrounding greenery and sky, adding depth and texture. The scene opens with a young boy in a swim cap and swim trunks sitting cross-legged on the rocks, partially submerged in the water, focused on something in his hands, possibly a small object or toy. Behind him, two girls in swimwear stand on the shore near a tree, one in a light blue swimsuit with her back to the camera, the other in patterned shorts facing away, seemingly engaged in conversation or observation. Nearby, a man in a light-colored shirt and sunglasses sits on a large rock, leaning back and relaxing, possibly watching the children or enjoying the view. Further in the background, the shoreline extends into a beach area with thatched-roof huts, indicating a resort or recreational zone, and a blue sign reading "FISHING PROHIBITED AREA" is visible. The camera slowly emerges from underwater and settles just above the surface, capturing the gentle sway of the water and the subtle shifts in perspective as it occasionally pans slightly to the right, revealing more of the beach and its surroundings. People are scattered along the shore, strolling or lounging, and the bright, natural sunlight bathes the scene, enhancing the tranquil, tropical ambiance.

✔ Wan2.2 (Ours SFT)



👍 User rating: 4.33

✘ Wan2.2 (Zero-Shot)



👍 User rating: 3.67

⊖ Expert feedback:



The camera should rise from **underwater** to **above the water**, rather than **remain above the surface** for the entire video.

Figure 28. Video generation example: height change from underwater to above water.

280-word Caption

The video begins with a medium shot of a woman with long, straight, platinum blonde hair and light skin, wearing a dark-colored jacket, positioned centrally in the frame, moving through a dense forest environment with tall, slender trees and a ground covered in grass and scattered fallen branches. The camera is at her height, capturing her upper body and head as she advances through the wooded area. The lighting suggests daytime, though the scene is shaded by the tree canopy, creating a subdued, cool-toned atmosphere. Trees and the forest floor fill the background, visible on all sides of the frame. **The camera follows her from behind with unsteady, noticeably shaky movement**, maintaining a level angle. As she continues moving, the camera pans from behind to the side, and the woman shifts from the center right to the left, then back to the center. **The shot transitions to a close-up from the side**, with her head and shoulders filling most of the frame, slightly shifted to the right. Her face is in sharp focus, with her head slightly lowered, revealing her neutral expression. The background trees remain visible but are more blurred, occupying the upper and left portions of the frame, while the forest floor is less prominent. The camera employs a shallow depth of field, starting with focus on the background and then performing a rack focus to bring the woman in the foreground into sharp focus. As the woman continues, the camera arcs counterclockwise around her, ending in a side profile as she stops. Throughout, the camera maintains a steady, eye-level perspective, with the woman consistently in the foreground as the central subject, her dark clothing contrasting with the natural surroundings.

✔ Wan2.2 (Ours SFT)



👍 User rating: 4.67

✘ Wan2.2 (Zero-Shot)



👍 User rating: 3.33

⊖ Expert feedback:



1. The video should end with a close-up shot, not a **medium close-up**.
2. The camera should begin by following the woman from **behind**, not **from the front**.

Figure 29. Video generation example: shot size change from medium to close-up.

(2) **Visual Attributes:**

Describe the subject's key visual characteristics using specific and descriptive language. Consider the following aspects where relevant:

• **Appearance:**

People: Include details like clothing (including colors and style), hairstyle, facial hair, age (if discernible), gender, ethnicity (if relevant and clear), facial expression, and so on.

Objects: Describe their color, material, shape, and any distinguishing marks (e.g., "smooth," "rough," "furry," "metallic," "black", "red", etc.)

• **Pose/Orientation:**

Describe the subject's posture and orientation within the frame (e.g., "standing," "sitting," "lying down," "walking," "facing left," "arms raised," "facing the camera"). Pay particular attention to objects not in their usual state (e.g., a tilted lamp, a book lying open face down).

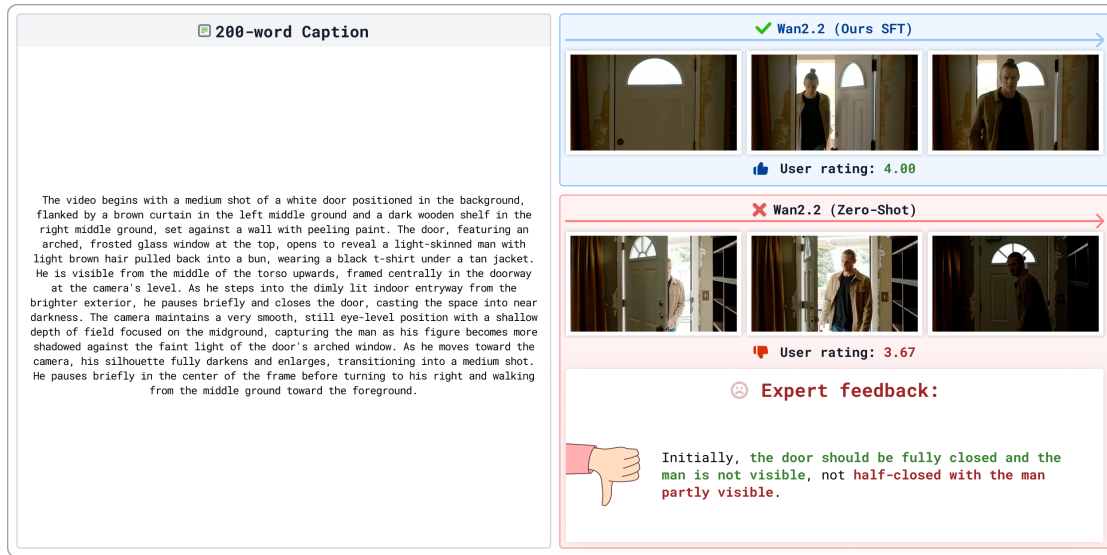


Figure 30. Video generation example: revealing shot.

(3) How to Refer to Multiple Subjects:

Sometimes there's more than one important person or thing in a picture. When that happens, make sure it's clear which one you're talking about. Here are some ways to do that:

- **Type:** The simplest way to refer to a subject is by its category, e.g., “the man,” “the dog,” or “the tree.”
- **Attributes:** If multiple subjects belong to the same category, use distinguishing features:
 - **Unique Appearance:** Highlight distinct traits, such as “the woman in the red dress,” “the man with the beard,” “the blue car,” or “the largest tree.”
 - **Location:** Specify position within the scene, e.g., “the man on the left,” “the dog in the background,” “the car in the midground,” or “the building in the middle.”
 - **Action:** Describe their activity, e.g., “the person walking,” “the child playing with a ball,” “the bird flying,” or “the cat sitting on the windowsill.”
 - **Relationship to Each Other:** For example, “the man next to the woman,” (spatial relationship), “the first man that enters the frame” (temporal relationship), or “the two cars parked side by side.”
- **Combining Descriptions:** For maximum clarity, combine multiple attributes. Example:
 - “The woman in the red dress on the left, talking on her phone.”
 - “The dog in the background, running toward the ball.”

The key is to give enough information so that anyone reading your description can easily tell which person or object you are referring to. Don't assume we know! The more detail you give, the better.

(4) Order Matters When Describing Multiple Subjects:

When describing multiple subjects, the order in which you mention them matters. Prioritize elements based on their importance in the video, ensuring a natural and logical flow.

- **Temporal Order:** If the scene unfolds over time, describe subjects in the order they appear. For example, “First, the car speeds past, then the cyclist enters the frame.”
- **Prominence-Based Order:** If temporal order isn't relevant, start with the most visually striking or important subject before moving on to less prominent ones. For example, “The video shows a bright red sports car in the foreground. In the midground, a blue sedan is right behind it.”

Subject Description Policy for Annotators

Use the following logic:

(1) Are there Shot Transitions (cuts or soft transitions)?

If Yes → No need for description (ends here).

If No → Proceed to the next step.

- (2) Are there prominent subjects that strongly draw the viewer’s attention?
 If No → Specify the type of shot and explain why there is no main subject (e.g., a scenery shot of landscape or cityscape; an establishing shot that sets the scene; a video featuring abstract visuals) (ends here).
 If Yes → Proceed to the next step.
- (3) **Determining How to Describe Subjects in a Video:**
 Before describing the subjects in a video, determine whether it features a single subject or a small group of subjects, or if it involves a complex subject scenario with multiple subjects or transitions. Identifying the scenario ensures a more fluent and logical description.
- **Single or Small Group of Subjects:**
 If the video consistently features one subject or a small group of subjects, follow the “Instructions for Subject Description.”
 When describing a small group, clearly distinguish between them by following “How to Refer to Multiple Subject.”
 - **Complex Subject Scenarios**
 If the video involves multiple subjects or complex transitions between them, please point out if the below scenarios occur:
 - **Determine the Primary Focus:**
 If there is a single clear main subject, describe the main subject in detail, explaining why this subject is the focus while others are not. Include relevant details such as appearance, actions, and positioning that make the subject stand out. Provide a less detailed overview of secondary subjects, mentioning only their general presence, or relationship to the main subject.
 If there are multiple subjects in focus, describe subjects in prominence-based order (e.g., humans before objects).
 If there is no clear main subject, give a brief overview of all subjects without excessive detail.
 - **Identify Subject Transitions:**
 If the focus shifts between subjects, specify the type of transition:
 - * Subject Revealing (a new subject enters the frame)
 - * Subject Disappearing (a subject exits or is no longer visible)
 - * Subject Switching (the focus shifts from one subject to another through rack focus or other camera movements)
 - * Other Complex Changes (subjects alternate focus multiple times)
 And explain how the transition occurs (through subject movement or camera movement).
 Identifying these scenarios ensures a more fluent and coherent description. For instance, in cases of subject switching, describe subjects in temporal order by following “Order Matters When Describing Multiple Subjects.”

I.4. Subject Motion and Dynamics Policy

Instructions for Subject Motion & Dynamics Description

Provide a concise yet informative description of the subject motion in this video. Please note that order matters. If multiple actions occur, present them in chronological order (e.g., “The bird first takes flight, then soars in a circle, and finally lands on a branch”).

Subject Motion Description Policy for Annotators

Use the following logic:

- (1) Are there Shot Transitions (cuts or soft transitions)?
 If Yes → No need for description (ends here).
 If No → Proceed to the next step.
- (2) Are there prominent subjects that strongly draw the viewer’s attention?
 If No → No need for description (ends here).
 If Yes → Proceed to the next step.
- (3) Are the subjects performing any activities or undergoing dynamic changes?
 If Yes →
- **Subject Actions and Dynamics:** Describe the actions and dynamic changes of individual subjects. Be specific about the manner of movement. Examples:
 - “A runner sprints across the finish line.” (Instead of: “A person is running.”)
 - “A hummingbird hovers delicately, wings beating rapidly as it sips nectar from a flower.” (Instead of: “A bird is flying.”)
 - “A caterpillar slowly inches its way along a leaf.” (Instead of: “An insect is moving.”)
 - “A time-lapse shows a sunflower turning its head to follow the sun across the sky.” (Instead of: “A plant is rotating.”)

- “A seed sprouts, sending a root down and a sprout up.” (Instead of “A seed is growing”)
- (4) Does the motion involve subjects interacting with objects?
If Yes →
- **Subject-Object Interactions:** Describe the interaction between a subject and an object. Be specific about the type of interaction and the object involved. Detail the effect of the interaction if relevant. Examples:
 - “A chef flips an omelet in a pan.” (Instead of: “A person is using a pan.”)
 - “A dog fetches a tennis ball thrown by its owner.” (Instead of: “A dog is playing.”)
 - “A construction worker operates a jackhammer, breaking up the pavement.” (Instead of: “A person is working.”)
 - “A car collides with a traffic sign, bending it at a sharp angle.” (Instead of: “A car crashed.”)
- (5) Does the motion involve subjects interacting with other subjects?
If Yes →
- **Subject-Subject Interactions:** Highlight interactions between different subjects. Describe the nature of the interaction and the relative movements of the subjects. Examples:
 - “Two boxers exchange blows in the ring, circling each other cautiously.” (Instead of: “People are fighting.”)
 - “A mother bird feeds worms to her chicks in the nest.” (Instead of: “Birds are together.”)
 - “Dancers perform a complex tango, their movements synchronized and graceful.” (Instead of: “People are dancing.”)
 - “A pride of lions hunts a zebra, surrounding it and closing in for the kill.” (Instead of: “Animals are interacting.”)
- (6) Is there a group engaging in collective behavior or action?
If Yes →
- **Group Activity:** Summarize collective behaviors or actions of a group. Describe the overall movement and any coordinated actions within the group. Specify the type of group if relevant. Examples:
 - “A flock of geese flies in a V-formation across the horizon.” (Instead of: “Birds are flying.”)
 - “A crowd of protesters marches down the street, carrying signs and banners.” (Instead of: “People are walking.”)
 - “A swarm of bees buzzes around a hive.” (Instead of: “Insects are moving.”)
 - “A school of fish swims in unison, changing direction as one unit.” (Instead of: “Fish are swimming.”)

I.5. Scene Composition and Dynamics Policy

Instructions for Scene Composition & Dynamics Description

Provide a concise yet informative description of the overall scene, including environment, setting, time of day, and notable visual elements. If there are subjects in this video, this scene description will also complement the subject descriptions, establishing where the subjects are and why they might be there. Your goal is to give enough context for understanding the setting while avoiding excessive detail.

(1) Describe the Overlay Elements (If any exists)

Overlays (if applicable): Identify and describe any overlay elements present in the shot that are not part of the scene. These may include text or visuals such as titles, subtitles, captions, icons, watermarks, heads-up displays (HUD), or framing elements. Clearly specify that these are overlays (not part of the scene) and describe their content and placement.

(2) Describe the Point of View (What Is the Context of the Shot?)

Point of View (POV) (if discernible): Describe how the scene is framed or captured when relevant. Examples:

- First-person: “The camera follows a person’s viewpoint as they walk through a dimly lit hallway.”
- Drone: “Aerial footage shows a city skyline stretching toward the horizon.”
- Over-the-shoulder: “A character is seen from behind, looking at a laptop screen in a dimly lit room.”
- Top-down: “A top-down oblique view of the game, with pieces arranged mid-game.”
- Dashcam: “A vehicle’s dashboard and windshield frame the road ahead, with headlights illuminating the wet pavement.”
- Objective/Neutral: “The camera provides a neutral, detached view of the scene.” (Use this when no specific POV is evident.)

(3) Describe the Setting (Where Does It Happen?)

Scene Type: Specify the general type of scene using precise and descriptive terms. Indicate whether it is indoors (e.g., “a cozy living room,” “a modern office”) or outdoors (e.g., “a bustling city street,” “a dense forest”). Avoid vague terms like “place” or “area”—be as specific as possible while ensuring clarity.

Good: “A sunlit café with large windows and wooden tables.”

Avoid: “An indoor place.”

- **Visual Attributes:** Considering the following aspects where relevant:

- **Location:** If the setting is a well-known place, state it explicitly (e.g., “Times Square,” “Grand Canyon,” “Tokyo subway station”). If the exact location is unclear, describe using its defining visual elements (e.g., “a narrow alley with graffiti-covered walls,” “a vast desert with rolling dunes,” “a dimly lit space with metal walls,” “an open area with sand and sparse vegetation”).
- **Time of Day (if discernible):** Specify whether the scene occurs during the day, night, or a transitional period like sunset or dawn, if relevant.
- **Architectural and Natural Features:** Mention buildings, roads, vegetation, water bodies, or other landscape elements that structure the scene (e.g., “a winding mountain path surrounded by tall pines,” “a bustling marketplace with food stalls and colorful banners”).
- **Weather Conditions:** If outdoors, describe weather effects (e.g., “a rainy street with wet pavement reflecting city lights,” “a snowy mountain pass covered in thick fog”).
- **Furniture and Props (for indoor scenes):** Identify relevant furnishings that establish the setting (e.g., “a wooden desk cluttered with books and a vintage lamp,” “a hospital room with a bed, medical monitors, and IV stands”).
- **Style:** If relevant, describe notable color schemes or stylistic choices (e.g., “a monochromatic, grayscale environment,” “a vibrant and colorful carnival scene with neon lights”).

(4) Describe Any Movement or Changes in the Environment (Any Changes or Motion in The Scene?)

Describe any natural or human-made movement happening at the scene level:

- **Natural Motion:** e.g., “Leaves sway in the wind,” “Waves crash against the shore,” “As the sun sets, it casts long shadows on the trees.”
- **Man-Made Motion:** e.g., “Traffic moves steadily on the highway,” “A train passes in the distance,” “Factory workers operate machinery in the background.”
- **Crowd & Background Activity:** e.g., “Pedestrians walk along a busy street,” “A crowd cheers and waves hands,” “The office starts empty, but employees gradually arrive and take their seats.”

If the scene changes, describe how it happens in the order it appears:

- **Time-Based Transitions:** e.g., “The shot begins during the day but transitions to nighttime.”
- **Movement-Based Transitions:** e.g., “The shot begins with a view of a quiet street. Then, the camera pans to reveal a hidden alley behind the main street.”

How to Refer to Multiple Scene Elements

Use precise and concise language to refer to different elements within the scene. For example:

- “In the background, a mountain range is visible.”
- “On the left side of the frame, there is a large tree.”
- “A wide river runs through the center, with a bridge arching over it.”

Prioritize the most prominent and important aspects of the scene. Start with the overall setting, then move on to more specific details.

Scene Composition & Dynamics Description Policy for Annotators

Use the following logic:

- (1) Are there Shot Transitions (cuts or soft transitions)?
If Yes → No need for description (ends here).
If No → Proceed to the next step.
- (2) Are there any Overlay Elements in this shot?
If No → No need for description.
If Yes → Describe the “Overlay Elements” following the instructions.
- (3) Is the Point of View obvious in this shot?
If No → An objective/neutral POV.
If Yes → Describe the “Point of View (Why and How Is the Shot Taken?)” following the instructions.
- (4) Are there any notable elements in the scene that should be described?
If No → For example, an extreme-close-up shot with no scene details. (ends here)
If Yes →
 - Describe the “Setting (Where Does It Happen?)” following the instructions.
 - For complex scenes (e.g., those with many elements), clearly distinguish between multiple scene elements by following “How to Refer to Multiple Scene Elements.”
- (5) Are there any changes or motion happening at the scene level?

If No → No need for further description (ends here).

If Yes →

- Describe the “Movement or Changes in the Environment” following the instructions.

I.6. Spatial Framing and Dynamics Policy

Instructions for Spatial Framing & Dynamics Description

Provide a concise yet informative description of how subjects and elements are framed within the scene, including their shot size, position, and movement within the scene or frame. Your goal is to describe the spatial composition and motion within the shot.

Framing of Subjects (How Are They Positioned in the Shot?)

(1) Describe the Subject’s Shot Size

Specify the shot size based on how much of the subject is visible:

- Close-Up: “A close-up shot of a human face.”
- Medium Shot: “A medium shot of a man’s lower body.”
- Wide Shot: “A wide shot of a man standing near the ocean.”

If the shot size does not follow a typical framing pattern or changes erratically throughout the shot, describe the overall framing instead of forcing a specific shot size. For example: “The camera tracks a skateboarder in an unsteady manner, and it mostly captures the skateboarder’s lower body.”

(2) Describe the Subject’s Position in the Scene

Position within the Frame: Describe the subject’s approximate location within the frame (e.g., “top-left corner,” “center,” “right of the frame”).

Example: “The person is in the bottom-left corner of the frame.”

Example: “The person is on the right of the frame.”

- **Depth within the Scene:** Describe the subject’s placement in relation to the foreground, midground, or background.

Example: “In the foreground, a person is sitting in front of a computer.”

- **Position within the Scene:** Describe the subject’s physical placement in the scene.

Example: “The woman in the midground stands near a window, looking outside.”

- **Height Relative to the Camera:** Describe the subject’s vertical positioning relative to the camera:

Example: “The man is framed at eye level.”

Example: “A low-angle shot captures the person from below.”

Framing of Scenery (How Is the Environment Captured?)

For scenery-focused shots without a primary subject, describe how the scene is framed:

(1) Describe the Scene’s Shot Size

Specify the type of shot capturing the environment, for example:

- Wide Shot: “A wide shot of a mountain range stretching across the horizon.”
- Close-Up: “A close-up of raindrops hitting a window.”

(2) Describe the Scene’s Spatial Composition

Spatial Positioning: specify where key elements appear within the frame (e.g., left, right, center, corners):

- “A symmetrical shot of a hallway positioned at the center of the frame, leading toward a vanishing point.”
- “A large tree stands in the left-bottom corner of the frame.”
- “A streetlamp is visible on the right side of the frame.”

Depth (Foreground, Midground, and Background Elements):

- “In the foreground, a bicycle is parked to the right against a fence, while in the background, skyscrapers rise against the sky.”

- “The midground features a river cutting through the landscape.”

Spatial Motion Within the Frame (How Do Subjects or Scene Elements Move?)

If shot size or spatial position changes within the frame, describe how these transitions happen clearly, specifying both the initial and final state.

(1) Changes in shot size and spatial position for Subjects

- “A medium shot of a man’s upper body near a doorway transitions into a close-up of his face as he walks toward the camera.”

- “A woman walking from the background to the foreground transitions from a wide shot capturing both her and the street scenery to a medium shot focusing on her lower body.”
 - “A cyclist moves from the left to the right side of the frame, maintaining a full shot throughout.”
 - “A full-body shot of a child at eye level shifts as the camera tilts upward, reframing them from a low angle looking up.”
 - “A wide shot captures a person near a park bench, who then walks diagonally from the bottom-left to the top-right corner of the frame.”
- (2) **Changes in shot size and spatial composition for Scenery shots**
- “The shot begins with an aerial view of a city skyline, then tilts downward to focus on a busy intersection.”
 - “The camera moves forward, transitioning from a wide view of a dense forest to a close-up of a single tree trunk covered in moss.”

Spatial Framing & Dynamics Description Policy for Annotators

Use the following logic:

- (1) Are there Shot Transitions (cuts or soft transitions)?
If Yes → No need for description (ends here).
If No → Proceed to the next step.
- (2) Are there prominent subjects that strongly draw the viewer’s attention?
If No → Follow “Framing of Scenery”, and note any changes in the spatial composition of scene elements according to “Spatial Motion Within the Frame”. (ends here)
If Yes → Proceed to the next step.
- (3) **Determining How to Describe Spatial Framing & Motion for Subjects**
Before describing spatial framing and motion for subjects, determine whether the video features a single subject or small group of subjects or involves a complex subject scenario with multiple subjects or transitions.
 - **Single or Small Group of Subjects:**
If the video consistently features one subject or a small group of subjects, follow the “Framing of Subjects” to describe:
 - Framing size (or shot size)
 - Frame location (position within the frame)
 - Scene Depth (foreground, midground, background)
 - Height relative to the camera (above, below, or at the height of the camera)
 Then follow “Spatial Motion Within the Frame” to describe any movement affecting any of the above aspects. (Ends here)
 - **Complex Subject Scenarios (Multiple Subjects or Subject Transitions)**
If the video contains multiple subjects or complex subject transitions, follow these steps:
 - **Determine the Primary Focus:**
If there is a single clear main subject: follow the “Framing of Subjects” to describe this subject in detail, including shot size and spatial position. Then follow “Spatial Motion Within the Frame” to describe any spatial motion and changes. Provide a less detailed overview of secondary subjects.
If the main subject is unclear: describe subjects’ spatial position and movement in prominence-based order (e.g., humans before objects). Instead of determining the shot size based on a random subject, specify it based on the most prominent subject (e.g., a human) if one is clearly dominant. Otherwise, if the subjects are relatively similar in size, use the average shot size. If the shot is even more complex, just directly say which (part of) subjects are visible and which are not.
 - If there is no clear main subject: provide a general overview of subjects’ spatial positions without excessive detail. Do not specify shot size, as it is not meaningful in this case. You may optionally describe the shot size following “Framing of Scenery” instead.
 - **Identify Subject Transitions:**
If subjects reveal, disappear, switch focus, or undergo other complex changes, describe their shot size (if relevant), spatial position, and movement accordingly. Ensure that the description follows the temporal order in which subjects appear.

I.7. Camera Framing and Dynamics Policy

Instructions for Camera Framing & Dynamics

Provide a concise yet informative description of the video’s and camera’s configuration (video speed, lens distortion, camera

angle, camera height), movements (translation, rotation, steadiness, intensity, complexity), and focus (depth, focus plane, focus changes) in this video.

Camera Framing & Dynamics Policy for Annotators

Use the following logic:

- (1) Are there Shot Transitions (cuts or soft transitions)?
If Yes → No need for description (ends here).
If No → Proceed to the next step.
- (2) Is the video played at a different speed than real-time?
If Yes → If the video speed is altered, specify:
 - Time-lapse: Events unfold significantly faster (e.g., “Clouds move rapidly across the sky.”)
 - Fast-Motion: Slightly faster than real-time (1x-3x speed).
 - Slow-Motion: Slower playback than real-time.
 - Stop-Motion: Frame-by-frame animation with discrete movements.
 - Speed-Ramp: A mix of fast and slow speeds within the same video.
 - Time-Reversed: The video plays in reverse.
- (3) Is there any noticeable lens distortion?
If Yes → Describe the distortion type and degree:
 - Fisheye: Extreme distortion with strong curvature, making the edges appear bent outward.
 - Barrel: Mild distortion causing straight lines near the edges to bow outward.
- (4) Can we tell the camera height (relative to the ground)?
If Yes →
 - Specify the height (Aerial-level, Overhead-level, Eye-level, Hip-level, Ground-level, Water-level, Underwater)
 - Mention any camera movement that causes height changes.
- (5) Can we tell the camera angle (relative to the ground)?
If Yes →
 - Specify the angle (Bird’s Eye, High Angle, Level Angle, Low Angle, Worm’s Eye)
 - Mention any camera movement that changes the camera angle.
- (6) Is there a Dutch/Canted angle (relative to the ground)?
If Yes → Describe how the Dutch Angle behaves:
 - The Dutch angle remains fixed throughout the shot.
 - The Dutch angle varies, changing due to camera rolling.
- (7) Can we tell the type of camera focus (or depth of field)?
If No → Explain if the video doesn’t seem to be filmed with a real camera, lacks a realistic depth of field effect, or is too blurry or dark to determine.
If Yes → Proceed to next step.
- (8) Is the camera using a shallow depth of field?
If No → Deep Focus.
If Yes →
 - Describe the depth of field (shallow or extremely shallow).
 - Specify which part of the frame is in focus (Foreground/Midground/Background/Out-of-Focus).
 - If the focus changes, describe the transition (Rack/Pull Focus, Focus tracking).
- (9) Is the camera moving (including shaking)?
If No → A static, fixed camera. (ends here).
If Yes → Proceed to the next step.
- (10) Is the camera shaking or wobbling?
If Yes → Describe the degree of shaking or wobbling (minimal, moderate, severe shaking).
- (11) Does the camera follow or move with any objects?
If Yes → Describe how the camera moves with the subject (e.g., Tracking Shot, Arcing, Craning).
- (12) How is the camera moving?
Describe why the camera is moving (e.g., tracking a subject, revealing a scene, creating emphasis). Describe the motion using precise movement terms:
 - Dolly In/Out: Moving forward or backward toward or away from the subject.

- Zoom In/Out: Changing focal length to create the illusion of moving closer or farther.
- Pan Left/Right: Rotating the camera horizontally.
- Truck Left/Right: Moving the camera laterally left or right.
- Tilt Up/Down: Angling the camera up or down.
- Pedestal Up/Down: Lifting or lowering the camera while keeping it level.
- Rolling Clockwise/Counterclockwise: Rotating the camera around its lens axis.

Mention the speed of movement if noticeably slow or fast. In addition, if different movements occur at different speeds, clearly distinguish them. For example: “The camera slowly dollies forward while trucking quickly to the right.”

Describe motion in temporal order if multiple movements occur. For example: “The camera first pans right, then tilts upward to follow the subject.”

However, if the movement appears too fragmented or random, avoid excessive detail.

(13) **Example (First-Person Video Game Scenario):**

Excessive Detail (Too much description):

“As the player explores, the camera moves left, then quickly tilts up, followed by a rapid pan right. The player hesitates, looking down, then abruptly swings the camera left again before slightly tilting upward and making another quick turn to the right.”

Better Description (Concise & clear):

“The first-person camera moves randomly as the player looks around, frequently changing direction without a clear pattern.”

I.8. Policy for Grading the Caption and Feedback

- 5 points: Completely correct; no changes needed.
- 4 points: Excellent, with only minor adjustments required (e.g., (1) a few inaccurate words need correction, (2) a few missing words need addition, or (3) a few hallucinated words need deletion). Overall, about one sentence’s worth of content may need to be added, changed, or deleted.
- 3 points: Mostly correct, but with some important omissions, hallucinations, or errors (e.g., more than two sentences need rewriting, deleting, or adding).
- 2 points: Mostly incorrect, requiring significant revisions (e.g., more than half of the caption is incorrect and needs to be redone).
- 1 point: Almost entirely wrong, requiring a complete rewrite (e.g., the entire caption is irrelevant).

I.9. Special Cases

Special Case 1: Missing Descriptive Details

If a caption is too brief, we can improve its readability by adding more descriptive details. As a general rule, a strong caption should be self-contained—if you read it to a friend who hasn’t seen the video, they should still be able to picture the subjects, scenery, motion, and camera work. Please aim to write captions that achieve this level of clarity. In particular, some camera captions might be too brief, such as saying the camera is focusing on the midground. Instead, we can improve readability by saying, “the camera is focusing on the woman wearing a white cap in the midground.” We should enhance the readability of these captions as much as possible. For readability issues, follow the same policy to deduct points.

Special Case 2: Formatting Errors

If the caption contains formatting errors (e.g., mentions of “first frame” or “second frame”), do not rate or correct it. Instead, regenerate a new pre-caption by retrying and modify the prompt to explicitly instruct the model not to mention “frame.” In other words, only correct factual mistakes in the video caption.

Special Case 3: Labelling Errors

If you discover that the original labels provided by humans are incorrect, please report these cases here and correct those labels in the Labelling projects.

Special Case 4: When the Polished Caption Isn’t Perfect

If the post-caption isn’t perfect, the mistakes usually fall into two categories:

- Type-A (Ignoring Feedback): ChatGPT doesn’t fully address all the points from our feedback.
- Type-B (Missed Feedback): We accidentally overlook some issues in our original feedback.

If you identify a Type-B issue, restart the feedback process. Only correct Type-A issues. In other words, ensure your original feedback is thorough and addresses all mistakes.

I.10. Common Mistakes to Avoid

- Do not say things like “this is a scene caption, so no need to mention the subject.” For pre-captions, simply correct what’s wrong and keep what’s right—no need to add unnecessary explanations.
- Focus on describing the visual content. Avoid adding extra commentary like “adds a sense of vibrancy to the scene.” However, if the pre-caption includes such descriptions and they are also accurate, it’s fine to keep them.
- Ensure consistency across the five captioning tasks. Camera motion information may be mentioned in subject, scene, or spatial captions—make sure it aligns with what’s described in the camera caption.
- If a 2D animation clearly conveys depth through layering, you may describe elements like the foreground, middle ground, and background—but be sure to specify that it is a 2D animation attempting to depict depth. If the scene appears flat and lacks depth cues, it’s best not to mention depth at all.
- A static shot is one where the camera remains entirely still—without any movement, focus change, or zoom. If any of these occur, the shot should not be described as static.
- Do not describe overlays as being in the foreground.

J. Pseudocode for Generating Pre-Captions from Labeled Primitives

This section presents the Python implementations of caption generation policies for video annotation. Each policy dynamically generates prompts based on video cinematography and camera motion primitives.

J.1. Subject Description Policy

```
class SubjectPolicy(PromptGenerator):
    """Generate prompts for subject description captions based on shot composition primitives."""

    def __init__(self):
        name = "Subject Description"
        info = "A policy that uses existing labels to prompt a human or model to provide structured captions for Subject."
        caption_fields = ["subject_description"]
        super().__init__(name, info, caption_fields)

    def get_prompt(self, data: VideoData) -> str:
        """Generate annotation prompt based on video's shot composition primitives."""

        # Base policy and format instructions
        POLICY_BASE = """Provide a concise yet informative description of the subjects in this video. Keep the description concise and clear, focusing on subject types and visual attributes. You should describe the video by combining details from the frames without referring to any specific one (e.g., don't mention things like "first frame" or "last frame"), and avoid using terms like "image" or "frame." Don't mention the background or motion unless it's necessary to distinguish subjects by location, action, or relationships. You must avoid describing what is not visible or what you are unsure about. You must use simple, natural English and ensure the description is a clear, concise, and coherent paragraph that highlights the most essential details. You must avoid subjective adjectives that convey emotions. Whenever you mention a subject, please describe its key visual attributes in detail. Return only the one-paragraph video description without Markdown formatting or introductory text.
```

Clearly identify each subject’s type, using precise terms such as "man," "woman," "dog," "car," or "tree," rather than vague words like "thing" or "item." If the subject type is ambiguous, use your best judgment and briefly explain your reasoning.

Describe key visual attributes with specific and descriptive language. For people, include details such as clothing color and style, skin tone, hairstyle, facial hair, age (if discernible), gender, ethnicity (if relevant and clear), and facial expression. For objects, describe their color, material, shape, and distinguishing features like texture or markings. Additionally, note the subject’s pose and orientation within the frame, such as standing, sitting, walking, or facing a certain direction. Pay attention to any objects that are not in their usual state, like a tilted lamp or an open book lying face down.

If there are multiple subjects to describe, ensure clarity in referring to each. The simplest way is by type, such as "the man," "the dog," or "the tree." If multiple subjects belong to the same category, distinguish them using unique appearance traits (e.g., "the woman **in** the red dress," "the man **with** the beard"), location within the scene (e.g., "the man on the left," "the car **in** the midground"), actions (e.g., "the child playing **with** a ball," "the bird flying"), or relationships to each other (e.g., "the man **next** to the woman," "the first man that enters the frame"). Also, when describing multiple subjects, the order in which they are mentioned matters. Prioritize based on relevance, starting with the largest or most centered subject. If the scene unfolds over time, describe subjects in the order they appear. If temporal order isn't relevant, begin with the most visually striking or important subject before moving to less prominent ones. The goal is to provide enough detail so that anyone reading the description can easily identify each subject.""

```
FORMAT_INST = """Please avoid using phrases like "the first frame" or "the last frame" in your
description. Instead, refer to the entire sequence simply as "the video." Your description should
integrate observations from all frames into a cohesive, temporally and logically consistent
narrative, rather than describing frames in isolation. Whenever you mention a subject, be sure to
include detailed descriptions of its key visual attributes. The final output should be a single,
fluent paragraph describing the video, with no Markdown formatting or introductory text. Don't
mention the surroundings or the subject's motion unless necessary to distinguish subjects by
location, action, or relationship."""
```

```
# Special case: Has Shot Transition
```

```
if data.Has_Shot_Transition:
    policy = POLICY_BASE
    policy += "This video contains one or more shot transitions. Please describe the subject of each
segment in a single fluent paragraph."
    return policy
```

```
# Special case: Scenery Shot (not Framing Subject)
```

```
if data.Framing_Subject == False:
    policy = "The video is a scenery shot. You do not need to describe the subject. Please concisely
specify the type of scenery shot (e.g., a landscape or cityscape scenery shot) in a single
fluent paragraph. Also explain why there is no main subject, such as the focus being on the
environment, atmosphere, or scale rather than a specific object. Just note that briefly in one
to three sentences."
    policy += "" + FORMAT_INST
    return policy
```

```
# Initialize with base policy for subject-containing videos
```

```
policy = POLICY_BASE
```

```
# Handle ambiguous framing (Framing_Subject == None)
```

```
if data.Framing_Subject == None:
    if data.Many_Subject_with_No_Clear_Focus:
        policy += "Please note that this video contains multiple subjects with no clear main focus**.
Because it does not emphasize any specific subject, please briefly describe the types of
subjects without going into too much detail. You may also describe the subjects
collectively as a group."
        policy += "" + FORMAT_INST
        return policy
```

```
# Add subject-type specific instructions
```

```
if data.Human_Shot:
    policy += "Please note that the video features salient human subjects, so the description
should focus on them."
```

```
elif data.Non_Human_Shot:
```

```

policy += "Please note that the video features salient non-human subjects, so the description
should focus on them."

# Handle subject transition cases
elif data.Change_of_Subject_Shot:
    if data.Subject_Revealing:
        policy += "Please note that the video is a revealing shot of the subject, so the
description should reflect this by explaining how the subject is revealed through either
subject movement or camera movement."
    elif data.Subject_Disappearing:
        policy += "Please note that the video features the main subjects disappearing from the frame
**, so the description should reflect this by explaining how they exit, whether through
subject movement or camera movement."
    elif data.Subject_Switching:
        policy += "Please note that the video features the main subjects switching from one to
another, so the description should reflect this by explaining how the transition occurs,
whether through subject movement or camera movement."

# Handle dynamic size variations
elif data.Clear_Subject_with_Dynamic_Shot_Size:
    policy += "Please note that the video has a main subject with dynamic size, so the description
should focus on them. Don't mention the background scene and other motion."

# Handle atypical appearance
elif data.Clear_Subject_with_Atypical_Shot_Size:
    policy += "Focus on describing the atypical appearance of the main subjects in the video.
Avoid mentioning the background or subject movements."

# Handle multiple subjects with one focus
elif data.Many_Subject_with_One_Clear_Focus:
    policy += "Please note that the video features multiple subjects with one clear main focus, so
you need to clarify who the main subject is. The description should focus on the details of
the main subject while concisely summarizing secondary subjects and describing their
relationship to the main subject if clear."

# Handle multiple different subjects in focus
elif data.Different_Subjects_in_Focus:
    policy += "Please note that the video features multiple different subjects in focus, so the
description should clearly distinguish their types and relationships."

# Handle complex/abstract scenarios
elif data.complex_shot_type == "unknown":
    policy += "Please note that the video features a complex scenario with ambiguous subjects or
it is an abstract shot. Please try your best to describe the main subjects or objects in the
video."

# Use existing shot size description if available
else:
    HAS_SHOT_SIZE = ""
    "In addition, the human-written caption below already describes the subjects (if
any) in this video, including framing information like shot size. Use this caption as a
reference to draft the subject description, but do not rely on it completely. Expand or refine
it to fully capture the subject's type, attributes, appearance, unique features, pose,
orientation, relationships between subjects, or any changes in the main focused subject, such
as revealing, disappearing, or shifting focus. However, if the description below does not
mention any subjects, please do not describe subjects and only specify the type of shot (e.g.,
a landscape or cityscape scenery shot or a FPV shot) and explain why there is no main subject.

```

```

Human-Written Caption: {shot_size_description}
    policy += "" + HAS_SHOT_SIZE.format(shot_size_description=data.shot_size_description)

    # Append format instructions and return
    policy += "" + FORMAT_INST
    return policy

```

Listing 1. Subject Description Policy

Primitives used: Has Shot Transition, Framing Subject, Many Subject with No Clear Focus, Human Shot, Non-Human Shot, Change of Subject Shot, Subject Revealing, Subject Disappearing, Subject Switching, Clear Subject with Dynamic Shot Size, Clear Subject with Atypical Shot Size, Many Subject with One Clear Focus, Different Subjects in Focus.

J.2. Scene Description Policy

```

class ScenePolicy(PromptGenerator):
    """Generate prompts for scene description captions based on point of view and overlay primitives."""

    def __init__(self):
        name = "Scene Description"
        info = "A policy that uses existing labels to prompt a human or model to provide structured captions
            for Scene."
        caption_fields = ["scene_composition_dynamics"]
        super().__init__(name, info, caption_fields)

    def get_prompt(self, data: VideoData) -> str:
        """Generate annotation prompt based on point of view and overlay information."""

        POLICY_BASE = """Provide a concise yet informative description of the overall scene, including the
            point of view, environment, setting, time of day, and notable visual elements like overlays. For
            notable visual elements within the scene, describe their color, material, shape, and
            distinguishing features like texture or markings. If subjects are present, ensure their placement
            and context complement the scene without excessive detail. You should describe the video by
            combining details from the frames without referring to any specific one (e.g., don't mention
            things like "first frame" or "last frame"), and avoid using terms like "image" or "frame." Focus
            on the setting and scenery rather than detailed subject descriptions. Avoid describing anything
            not visible or uncertain. Use simple, natural English to create a clear, concise, and coherent
            paragraph that highlights essential details. Avoid emotional or subjective adjectives. Avoid
            speculative statements like 'there might be,' 'it appears,' or ambiguous options like 'A or B.'
            Do not infer the role of the scene setting. Do not explain what the scene emphasizes or
            highlights. Return only the one-paragraph video description without Markdown formatting or
            introductory text.

```

If relevant, indicate the **point of view**, such as first-person, drone shot, or dashcam, and describe how it influences the viewer's perception. Specify the **setting** by clearly identifying whether it is indoors or outdoors, using precise language. If the location is known, state it explicitly (e.g., "Times Square" or "Tokyo subway station"). Otherwise, describe defining features such as "a narrow alley **with** graffiti-covered walls" or "a vast desert **with** rolling dunes." Mention the **time of day** and any notable **architectural** or natural features, such as buildings, roads, forests, or bodies of water. Include relevant **weather conditions** if applicable, like "a rainy street **with** wet pavement reflecting city lights" or "a snowy mountain **pass** covered **in** thick fog." For indoor settings, describe key **furniture** or **props** that establish the environment, such as "a wooden desk cluttered **with** books **and** a vintage lamp." If notable, mention the **style** of the scene, such as a monochromatic color scheme or a vibrant carnival with neon lights. If the video contains **overlay elements** such as text, titles, subtitles, captions, icons, watermarks, heads-up displays (HUD), or framing elements, specify that they are overlays (not part of the scene) and describe their content and placement.

If the scene involves **motion** or **changes**, describe natural elements like wind blowing through trees or waves crashing against the shore, as well as human-made movements such as traffic flowing on a highway

```

or pedestrians walking along a busy street."""

FORMAT_INST = """Please avoid using phrases like "the first frame" or "the last frame" in your
description. Instead, refer to the entire sequence simply as "the video." Your description should
integrate observations from all frames into a cohesive, temporally and logically consistent
narrative, rather than describing frames in isolation. Do not infer the role of the scene setting.
Do not explain what the scene emphasizes or highlights. The final output should be a single,
fluent paragraph. Focus on the setting and scenery, not on detailed descriptions of the subject."
"""

policy = POLICY_BASE
policy += "" + FORMAT_INST

# Add point of view information

pov_description = f"[Point of view information for {data.true_pov_attribute}]"
if data.true_pov_attribute == "objective_pov":
    pov_description += " (no need to mention)."
```

```

# Add overlay instruction if present
if data.Overlays:
    policy += "Please note that the video includes overlay elements, such as text or visuals like
titles, subtitles, captions, icons, watermarks, heads-up displays (HUD), or framing elements.
In your description, specify that these are overlays (not part of the scene) and describe
their content and placement."
```

```

# Add point of view context
policy += f"In addition, you do not need to infer the camera's point of view, as this information is
already provided. Please integrate the following point of view information into your caption:
Point of View Information: **{pov_description}**"
```

```

return policy
```

Listing 2. Scene Description Policy

Primitives used: Overlays, Point of View (Objective Pov, First Person Pov, Selfie Pov, Overhead Pov, Locked On Pov, Dashcam Pov, Drone Pov, Broadcast Pov, Screen Recording Pov, Third Person Over Shoulder Pov, Third Person Over Hip Pov, Third Person Full Body Game Pov, Third Person Top Down Game Pov, Third Person Side View Game Pov, Third Person Isometric Game Pov).

J.3. Subject Motion & Dynamics Policy

```

class SubjectMotionPolicy(PromptGenerator):
    """Generate prompts for subject motion description based on shot type primitives."""

    def __init__(self):
        name = "Subject Motion & Dynamics Description"
        info = "A policy that uses existing labels to prompt a human or model to provide structured captions
for Subject Motion and Dynamics."
        caption_fields = ["subject_motion_dynamics"]
        super().__init__(name, info, caption_fields)

    def get_prompt(self, data: VideoData) -> str:
        """Generate annotation prompt based on subject motion characteristics."""

        POLICY_BASE = """Provide a concise yet informative description of the subject's motion in this video,
ensuring actions are presented in chronological order if multiple movements occur (e.g., "
The bird first takes flight, then soars in a circle, and finally lands on a branch"). Focus on
the subject's motion rather than repeating details already included in the human-written subject
```

descriptions. Avoid describing anything not visible or uncertain. Use simple, natural English to create a clear, concise, and coherent paragraph that highlights essential details. Avoid emotional or subjective adjectives. Avoid speculative statements like 'there might be,' 'it appears,' or ambiguous options like 'A or B.' Return only the one-paragraph video description without Markdown formatting or introductory text.

If the subject in the video has no movement, please briefly mention that without going into too much detail.

Please only describe the content of the video. Don't mention the details of the subject's appearance unless you need to differentiate between multiple subjects by their appearance. Clearly describe the subject's motion.

Avoid abstract descriptions, such as "The car maintains a low, sleek profile **as** it maneuvers the bend, emphasizing its speed **and** agility" and "emphasizing its speed **and** agility **as** it maneuvers through the turn."

Below are detailed instructions:

Describe **individual subject actions** with clarity, specifying how they move rather than using generic descriptions. For example, instead of "a person **is** running," say "a runner sprints across the finish line."

If the subject interacts with an **object**, specify the type of interaction and its effect. Instead of "a person **is** working," say "a construction worker operates a jackhammer, breaking up the pavement."

If there are **interactions between subjects**, describe the nature of their relationship and movements relative to each other. Instead of "people are fighting," say "two boxers exchange blows **in** the ring, circling each other cautiously."

If there are collective behaviors for a group of subjects, describe **group activities** with specificity. Instead of "birds are flying," say "a flock of geese flies **in** a V-formation across the horizon." Instead of "people are walking," say "a crowd of protesters marches down the street, carrying signs **and** banners." Clearly convey the type of group, their coordinated actions, and any notable patterns in their movement.""

```
# Special case: Scenery Shot (no subjects)
```

```
if data.Framing_Subject == False:
```

```
    return "The video is a scenery shot. You do not need to describe the subject motion. Just note that briefly in one to three sentences."
```

```
policy = POLICY_BASE
```

```
# Handle ambiguous framing
```

```
if data.Framing_Subject == None:
```

```
    if data.Many_Subject_with_No_Clear_Focus:
```

```
        policy += "Please note that this video contains multiple subjects without a clear main focus **. Briefly describe the salient motions and dynamics of the primary subjects while providing a concise overview of secondary movements, or describe all subjects' collective motion if that is more appropriate."
```

```
        return policy
```

```
# Add subject-type specific instructions
```

```
if data.Human_Shot:
```

```
    policy += "Please note that the video features salient human subjects, so the description should focus on their motion and dynamics."
```

```
elif data.Non_Human_Shot:
```

```

    policy += "Please note that the video features salient non-human subjects, so the description
        should focus on their motion and dynamics."

# Handle subject transition cases
elif data.Change_of_Subject_Shot:
    if data.Subject_Revealing:
        policy += "Please note that the video is a revealing shot of the subject, so the
            description should reflect this by explaining how the subject is revealed through either
            subject movement or camera movement."
    elif data.Subject_Disappearing:
        policy += "Please note that the video features the main subjects disappearing from the frame
            **, so the description should reflect this by explaining how they exit, whether through
            subject movement or camera movement."
    elif data.Subject_Switching:
        policy += "Please note that the video features the main subjects switching from one to
            another, so the description should first describe the first subject's motion and dynamics,
            followed by the second's."

# Handle dynamic size and atypical motion
elif data.Clear_Subject_with_Dynamic_Shot_Size:
    policy += "Please note that the main subject's framing is not stable throughout the video, so
        the description should reflect how their motion and dynamics contribute to this instability."

elif data.Clear_Subject_with_Atypical_Shot_Size:
    policy += "Please note that the main subjects in this video exhibit atypical motion, posture, or
        anatomy, so the description should reflect this."

# Handle multiple subject configurations
elif data.Many_Subject_with_One_Clear_Focus:
    policy += "Please note that the video features multiple subjects with a clear main focus, so
        the description should focus on the motion and dynamics of the main subject while providing a
        concise overview of secondary subjects' movements."

elif data.Different_Subjects_in_Focus:
    policy += "Please note that the video features multiple different subjects in focus, so the
        description should clearly distinguish their types, movement patterns, and interactions."

# Add reference to existing subject description
HAS_SUBJECT_DESC = ""
In addition, the human-written caption below already describes the subjects (if
any) in this video but does not comprehensively capture their motion. Use this caption as a
reference or starting point to draft the description, but do not rely on it completely. Expand or
refine it to fully capture the subjects' motion and dynamics. However, if the caption does not
mention any subjects, do not add any description--simply note this briefly in the description.
Please note that this caption is only to help you clarify which subject's motion needs to be
analyzed, rather than adding more content based on this description.

Human-Written Caption: {subject_description}""
    policy += "" + HAS_SUBJECT_DESC. format(subject_description=data.subject_description)

return policy

```

Listing 3. Subject Motion & Dynamics Policy

Primitives used: Framing Subject, Many Subject with No Clear Focus, Human Shot, Non-Human Shot, Change of Subject Shot, Subject Revealing, Subject Disappearing, Subject Switching, Clear Subject with Dynamic Shot Size, Clear Subject with Atypical Shot Size, Many Subject with One Clear Focus, Different Subjects in Focus.

Non-primitive attribute: subject_description (human-written subject caption).

J.4. Spatial Framing & Dynamics Policy

```
class SpatialPolicy(PromptGenerator):
    """Generate prompts for spatial framing description with shot size and camera height information."""

    def __init__(self):
        name = "Spatial Framing and Dynamics Description"
        info = "A policy that uses existing labels to prompt a human or model to provide structured captions
            for Spatial Framing and Dynamics."
        caption_fields = ["spatial_framing_dynamics"]
        super().__init__(name, info, caption_fields)

    def format_shot_size(self, shot_size: str) -> str:
        """Convert shot size enum to natural language description."""
        shot_size_map = {
            "unknown": "unknown",
            "extreme_wide": "an extreme wide shot",
            "wide": "a wide shot",
            "full": "a full shot",
            "medium_full": "a medium full shot",
            "medium": "a medium shot",
            "medium_close_up": "a medium close-up shot",
            "close_up": "a close-up shot",
            "extreme_close_up": "an extreme close-up shot"
        }
        return shot_size_map[shot_size]

    def format_height_wrt_subject(self, height: str) -> str:
        """Convert height relative to subject enum to natural language."""
        height_map = {
            "unknown": "unknown",
            "above_subject": "above the subject",
            "at_subject": "at the subject's level",
            "below_subject": "below the subject"
        }
        return height_map[height]

    def get_prompt(self, data: VideoData) -> str:
        """Generate annotation prompt with shot size and height information."""

        POLICY_BASE = """Analyze the subjects and elements in this video and provide a concise yet
            informative description of how they are spatially framed within the scene, including **shot size,
            position, depth, height relative to the camera, and any changes**. Your goal is to describe the
            **spatial framing and dynamics** of the subjects and elements within the shot, considering both
            their placement within the frame and their relative positions in the scene. Ensure the
            description covers any notable spatial movements. Avoid describing anything not visible or
            uncertain. Use simple, natural English to create a clear, concise, and coherent paragraph that
            highlights essential details. Avoid emotional or subjective adjectives. Avoid speculative
            statements like 'there might be,' 'it appears,' or ambiguous options like 'A or B'. Return only
            the one-paragraph video description without Markdown formatting or introductory text.
```

First, specify the **shot size** based on the subject's size in the frame if major subjects are present. If the shot size is unclear, describe how much of the subject is visible. If no major subject exists (e.g ., a scenery shot), describe the shot size in relation to the scenery.

Next, describe the **spatial position** of subjects and elements in the video, if relevant. Indicate their approximate **2D position** within the frame using terms like **left, right, bottom left, bottom right, top right, top left, bottom, top, or center**. Additionally, describe their **3D position** within the

scene as **foreground, middle ground, or background**. Analyze as many elements as possible, and for each element mentioned, provide both its **2D and 3D position**.

Finally, describe the **camera's height relative to the subject**, if relevant. Indicate whether the camera is positioned at the subject's height, above them, or below them. We already have this information provided at the end. If it's not provided, try to describe it by yourself.

If **shot size or spatial position** changes, describe how these transitions occur clearly, specifying both the **initial and final states**."

```
HAS_SUBJECT_SCENE = ""In addition, the human-written captions below already describe the subjects (
if any) and the scenery in this video but do not capture their spatial composition and movements.
Use these captions as a reference, but do not rely on them completely. Your goal is to fully
capture the spatial framing and movements in this video. Don't write too much about the subject's
or scenery's appearance.
```

Human-Written Description for Subjects: **{subject_description}**

Human-Written Description for Scenery: **{scene_description}**""

```
policy = POLICY_BASE
policy += "" + HAS_SUBJECT_SCENE.format(
    subject_description=data.subject_description,
    scene_description=data.scene_description
)

# Determine subject status for shot size information
subject_status = None # 'has_subject', 'no_subject', 'change_of_subject', 'has_description'

# Add subject-type specific instructions
if data.Human_Shot:
    policy += "Please note that the video features salient human subjects, so you should focus on
    describing the spatial framing and movements of them."
    subject_status = "has_subject"

elif data.Non_Human_Shot:
    policy += "Please note that the video features salient non-human subjects, so you should focus
    on describing the spatial framing and movements of them."
    subject_status = "has_subject"

# Handle subject transition cases with detailed shot size info
elif data.Change_of_Subject_Shot:
    subject_status = "change_of_subject"

if data.Subject_Revealing:
    policy += "Please note that the video is a revealing shot of the subject."
    policy += f"Shot Size Information: The video begins with no subject. It then becomes {self.
    format_shot_size(data.shot_size_info['end'])} of the subject."
    if data.Subject_Height_Applicable:
        policy += f"When the subject is revealed, the camera is positioned {self.
        format_height_wrt_subject(data.height_wrt_subject_info['end'])}."

elif data.Subject_Disappearing:
    policy += "Please note that the video features main subjects disappearing from the frame."
    policy += f"Shot Size Information: The video begins with {self.format_shot_size(data.
    shot_size_info['start'])} of the subject. Then the subject disappears."
    if data.Subject_Height_Applicable:
        policy += f"Before the subject disappears, the camera is positioned {self.
        format_height_wrt_subject(data.height_wrt_subject_info['start'])}."
```

```

elif data.Subject_Switching:
    policy += "Please note that the video features main subjects switching from one to another."
    "
    policy += f"Shot Size Information: The video begins with {self.format_shot_size(data.
        shot_size_info['start'])} of the first subject. Then it becomes {self.format_shot_size(data.
        shot_size_info['end'])} of the second subject."
    if data.Subject_Height_Applicable:
        policy += f"The camera is positioned {self.format_height_wrt_subject(data.
            height_wrt_subject_info['start'])} when the first subject is in focus, and {self.
            format_height_wrt_subject(data.height_wrt_subject_info['end'])} when the second subject
            is in focus."

# Note: This policy also uses non-primitive attributes: Shot_Size_Description,
        Subject_Height_Description

# Handle other shot configurations
elif data.Clear_Subject_with_Dynamic_Shot_Size:
    policy += "Please note that the main subject's framing (shot size) is not stable throughout
        the video, so the description should emphasize this."
    subject_status = "has_subject"

elif data.Clear_Subject_with_Atypical_Shot_Size:
    policy += "Please note that the main subjects exhibit atypical posture or anatomy, so the
        description should reflect this."
    subject_status = "has_subject"

elif data.Many_Subject_with_One_Clear_Focus:
    policy += "Please note that the video features multiple subjects with a clear main focus, so
        the description should focus on the main subject."
    subject_status = "has_subject"

elif data.Different_Subjects_in_Focus:
    policy += "Please note that the video features multiple different subjects in focus, so the
        description should clearly distinguish their types and relationships."
    subject_status = "has_subject"

elif data.Many_Subject_with_No_Clear_Focus:
    policy += "Please note that this video contains multiple subjects without a clear main focus.
        Briefly describe the spatial positions and movements of salient subjects while providing a
        concise overview of secondary subjects, or describe all the spatial composition of all
        subjects collectively as a group if that is more appropriate."
    subject_status = "has_subject"

elif data.Scenery_Shot:
    policy += "Please note that the video is a scenery shot. You do not need to describe the
        subjects. Just note that briefly in one to three sentences."
    subject_status = "no_subject"

elif data.complex_shot_type == "unknown":
    policy += "Please note that the video features a complex scenario with ambiguous subjects or
        it is an abstract shot. Please try your best to describe the spatial positions and movements
        of the main subjects or objects in the video."
    subject_status = None

else:
    # Has existing shot size description
    subject_status = "has_description"

```

```

policy += "The description below already mentions the spatial framing information about the
subjects or scenery in this video. Use this caption as a reference to draft the spatial
framing and dynamics description. Simply expand on it to fully capture other spatial positions
and movements. Do not infer the any spatial framing information already mentioned below."
policy += f"Shot Size Information: {data.Shot_Size_Description}"

if data.Subject_Height_Applicable:
    if data.Height_Wrt_Subject_Changes:
        policy += f"Camera Height Relative to Subjects: The camera is initially positioned {self.
            format_height_wrt_subject(data.height_wrt_subject_info['start'])} and then changes to {
            self.format_height_wrt_subject(data.height_wrt_subject_info['end'])}."
    else:
        policy += f"Camera Height Relative to Subjects: The camera is positioned {self.
            format_height_wrt_subject(data.height_wrt_subject_info['start'])}."
    elif data.Subject_Height_Description != "":
        policy += f"Camera Height Relative to Subjects: {data.Subject_Height_Description}"

# Add shot size information based on subject status
if subject_status == "has_subject":
    if data.Shot_Size_Changes:
        policy += f"Shot Size Information: The video begins with {self.format_shot_size(data.
            shot_size_info['start'])} of the subjects. It then changes to {self.format_shot_size(data.
            shot_size_info['end'])}."
    else:
        policy += f"Shot Size Information: The video shows {self.format_shot_size(data.shot_size_info['
            start'])} of the subjects."

    if data.Subject_Height_Applicable:
        if data.Height_Wrt_Subject_Changes:
            policy += f"Camera Height Relative to Subjects: The camera is initially positioned {self.
                format_height_wrt_subject(data.height_wrt_subject_info['start'])}. It then changes to {
                self.format_height_wrt_subject(data.height_wrt_subject_info['end'])}."
        else:
            policy += f"Camera Height Relative to Subjects: The camera is positioned {self.
                format_height_wrt_subject(data.height_wrt_subject_info['start'])}."
        elif data.Subject_Height_Description != "":
            policy += f"Camera Height Relative to Subjects: {data.Subject_Height_Description}"

elif subject_status == "no_subject":
    if data.Shot_Size_Changes:
        policy += f"Shot Size Information: The video begins with {self.format_shot_size(data.
            shot_size_info['start'])} of the scenery. It then changes to {self.format_shot_size(data.
            shot_size_info['end'])}."
    else:
        policy += f"Shot Size Information: The video shows {self.format_shot_size(data.shot_size_info['
            start'])} of the scenery."

elif subject_status == None:
    policy += "Shot Size Information: The video features a complex scenario with ambiguous subjects or
it is an abstract shot. Please try your best to describe the spatial positions and movements
of the main subjects or objects in the video. Do not use shot size to describe the spatial
framing."

return policy

```

Listing 4. Spatial Framing & Dynamics Policy

Primitives used: Human Shot, Non-Human Shot, Change of Subject Shot, Subject Revealing, Subject Disappearing, Subject Switching, Clear Subject with Dynamic Shot Size, Clear Subject with Atypical Shot Size, Many Subject with One Clear

Focus, Different Subjects in Focus, Many Subject with No Clear Focus, Scenery Shot, Shot Size Is [Extreme Close Up, Close Up, Medium Close Up, Medium, Medium Full, Full, Wide, Extreme Wide], Shot Size Changes, Shot Size Changes From Small to Large, Shot Size Changes From Large to Small, Height Is Always Above Subject, Height Is Always At Subject, Height Is Always Below Subject, Height Wrt Subject Changes.

Non-primitive attributes: subject_description, scene_description, Shot_Size_Description, Subject_Height_Description.

J.5. Camera Framing & Dynamics Policy

```
class CameraPolicy(PromptGenerator):
    """Generate prompts for camera description with comprehensive camera setup and motion information."""

    def __init__(self):
        name = "Camera Framing and Dynamics Description"
        info = "A policy that uses existing labels to prompt a human or model to provide structured captions
            for Camera Framing and Dynamics."
        caption_fields = ["camera_framing_dynamics"]
        super().__init__(name, info, caption_fields)

    def format_playback_speed(self, speed: str) -> str:
        """Convert playback speed enum to natural language (reads from JSON in actual implementation)."""
        # In actual implementation, reads from labels/cam_setup/video_speed/*.json
        speed_descriptions = {
            "time_lapse": "The video is a time-lapse",
            "fast_motion": "The video is in fast motion",
            "regular": "The video is at regular playback speed (no need to mention).",
            "slow_motion": "The video is in slow motion",
            "stop_motion": "The video uses stop motion animation",
            "speed_ramp": "The video uses speed ramping (changing between fast and slow motion)",
            "time_reversed": "The video plays in reverse"
        }
        return speed_descriptions.get(speed, "Unknown playback speed")

    def format_lens_distortion(self, distortion: str) -> str:
        """Convert lens distortion type to natural language."""
        if distortion == "regular":
            return "No lens distortion (no need to mention)."
        elif distortion == "barrel":
            return "The video features mild barrel distortion where lines near the frame edges bow slightly
                outward."
        elif distortion == "fisheye":
            return "The video shows extreme fisheye distortion where most lines curve strongly outward."
        return "Unknown lens distortion"

    def format_camera_height_start(self, height: str) -> str:
        """Format camera height at start of video."""
        height_map = {
            "aerial_level": "at an aerial level",
            "overhead_level": "at an overhead level (around second floor height)",
            "eye_level": "at an eye level (above the waist)",
            "hip_level": "at a hip level (below the waist and above the knees)",
            "ground_level": "at a ground level",
            "water_level": "above water",
            "underwater_level": "underwater"
        }
        return height_map.get(height, "at an unknown height")

    def format_camera_height_end(self, height: str) -> str:
```

```

    """Format camera height at end of video (for transitions)."""
    height_map = {
        "aerial_level": "to an aerial level",
        "overhead_level": "to an overhead level (around second floor height)",
        "eye_level": "to an eye level (above the waist)",
        "hip_level": "to a hip level (below the waist and above the knees)",
        "ground_level": "to a ground level",
        "water_level": "above water",
        "underwater_level": "underwater"
    }
    return height_map.get(height, "to an unknown height")

def format_camera_angle(self, angle: str) -> str:
    """Convert camera angle enum to natural language."""
    angle_map = {
        "bird_eye_angle": "a bird's-eye view angle (looking down directly at the ground)",
        "high_angle": "a high angle (looking down from above)",
        "level_angle": "a level angle (looking straight ahead)",
        "low_angle": "a low angle (looking up from below)",
        "worm_eye_angle": "a worm's-eye view angle (looking directly up)"
    }
    return angle_map.get(angle, "an unknown angle")

def format_focus_plane(self, plane: str) -> str:
    """Convert focus plane enum to natural language."""
    plane_map = {
        "foreground": "focused on the foreground",
        "middle_ground": "focused on the midground",
        "background": "focused on the background",
        "out_of_focus": "out of focus"
    }
    return plane_map.get(plane, "focus unknown")

def format_camera_steadiness(self, steadiness: str) -> str:
    """Convert camera steadiness enum to natural language."""
    steadiness_map = {
        "static": "The camera is stationary",
        "very_smooth": "The camera movement is very smooth with no shaking",
        "smooth": "The camera movement is smooth with minimal shaking",
        "unsteady": "The camera movement is slightly unsteady with some shaking",
        "very_unsteady": "The camera movement is unsteady with noticeable shaking"
    }
    return steadiness_map.get(steadiness, "Unknown steadiness")

def format_camera_motion_speed(self, speed: str) -> str:
    """Convert camera motion speed enum to natural language."""
    speed_map = {
        "slow": "moving slowly.",
        "regular": "moving at a regular speed (no need to mention).",
        "fast": "moving quickly."
    }
    return speed_map.get(speed, "moving at unknown speed.")

def get_movement_description_simple(self, data: VideoData) -> str:
    """Generate description for simple camera movements."""
    movement_map = {
        "Roll_Clockwise": "rolling clockwise",
        "Roll_Counterclockwise": "rolling counterclockwise",

```

```

    "Forward": "moving forward",
    "Backward": "moving backward",
    "Zoom_In": "zooming in",
    "Zoom_Out": "zooming out",
    "Upward": "moving up",
    "Downward": "moving down",
    "Tilt_Up": "tilting up",
    "Tilt_Down": "tilting down",
    "Pan_Right": "panning right",
    "Pan_Left": "panning left",
    "Leftward": "moving left",
    "Rightward": "moving right",
    "Crane_Up": "craning up in an arc",
    "Crane_Down": "craning down in an arc",
    "Arc_Clockwise": "arcing clockwise",
    "Arc_Counterclockwise": "arcing counterclockwise"
}

# Get list of active movements
true_movements = [name for name, desc in movement_map.items() if getattr(data, name, False)]

if len(true_movements) == 0:
    return "The camera shows no clear or intentional movement."
elif len(true_movements) == 1:
    return f"The camera is {movement_map[true_movements[0]]}."
elif len(true_movements) == 2:
    return f"The camera is {movement_map[true_movements[0]]} and {movement_map[true_movements[1]]}."
else:
    movements_str = ", ".join([movement_map[m] for m in true_movements[:-1]])
    return f"The camera is {movements_str}, and {movement_map[true_movements[-1]]}."

def get_tracking_description(self, data: VideoData) -> str:
    """Generate description for tracking shots."""
    # This is a complex method that combines tracking types
    # Simplified version shown here
    tracking_types = data.tracking_shot_types # e.g., ["side", "pan"]

    if "side" in tracking_types:
        base = "The camera is tracking the subject from the side"
    elif "tail" in tracking_types:
        base = "The camera is following the subject from behind"
    elif "lead" in tracking_types:
        base = "The camera is leading the subject from the front"
    elif "aerial" in tracking_types:
        base = "The camera is tracking the subject from an aerial view"
    else:
        base = "The camera is tracking the subject"

    # Add size change information if applicable
    if data.Tracking_Subject_Larger_Size:
        base += ". During the tracking shot, the subject becomes larger in the frame."
    elif data.Tracking_Subject_Smaller_Size:
        base += ". During the tracking shot, the subject becomes smaller in the frame."

    return base

def get_prompt(self, data: VideoData) -> str:
    """Generate comprehensive camera annotation prompt."""

```

```
POLICY_BASE = """Provide a concise yet informative description of the video's and camera's configuration, covering video speed, lens distortion, camera angle, camera height, movements (translation, rotation, zooming, steadiness, arcing, craning, tracking, speed, complexity, and purpose), and focus (depth, focus plane, focus changes)."""
```

If **video speed** is altered, specify the type, such as *time-lapse* ("Clouds move rapidly across the sky"), *slow-motion*, *fast-motion*, or *speed ramp* (changing between fast and slow motion). If the video is *time-reversed* or *stop-motion*, note this as well.

If **lens distortion** is present, describe the type and degree. For example, *fish-eye distortion* creates extreme curvature, while *barrel distortion* causes mild outward bowing of straight lines near the edges.

Describe the **camera height** in relation to the ground, such as *eye-level*, *hip-level*, *ground-level*, *overhead-level*, *aerial-level*, *above water*, or *underwater*. If height changes due to movement, mention how it transitions. Similarly, specify the **camera angle**, such as *bird's eye*, *high angle*, *level angle*, *low angle*, or *worm's eye*, noting any shifts within the video. If a **Dutch angle** (tilted horizon) is present, indicate whether it remains fixed or varies due to camera rolling.

If discernible, describe the **camera focus and depth of field**. For example, *deep focus* keeps all elements sharp, while *shallow* or *ultra-shallow depth of field* blurs the background or foreground. If focus changes dynamically, note whether it's a *rack focus* (shifting focus between subjects) or *focus tracking* (following a subject's depth movement), and state the focus plane at each stage (foreground, midground, background, or out-of-focus). If the video lacks realistic depth of field, describe whether it appears artificial (without a physical camera) or overly blurry.

If the **camera is static**, simply state that the shot is static. If it moves, describe the **type**, **direction**, **steadiness**, and **speed** of movement. Specify movements such as *tracking* (following a subject), *arcing clockwise/counterclockwise* (circling around a subject or frame center horizontally), *craning up/down* (circling around a subject or frame center vertically), *zooming* (changing focal length), *dolly* (moving forward/backward), *truck* (moving left/right), *pedestal* (moving up/down), *panning* (rotating the camera horizontally), *tilting* (rotating the camera up/down), or *rolling* (rotating around the lens axis). If the camera is shaking or wobbling, indicate the degree (e.g., minimal, moderate, or severe). If different movements occur at different speeds, clearly distinguish them. If the camera performs multiple movements, describe them in temporal order (e.g., "The camera first pans right, then tilts upward to follow the subject")."""

```
policy = POLICY_BASE
```

```
policy += "Crucially, instead of inferring these attributes from the video, we have already provided human-labeled ground truth for some of the elements specified above. You should directly use this information in your description and should not infer any details that are not already provided. Your description should be brief, and if anything is normal or unremarkable, you do not need to include it (e.g., if the video is at regular playback speed, there is no need to mention it)."
```

```
# Add playback speed information
```

```
policy += f"Playback Speed: {self.format_playback_speed(data.playback_speed)}"
```

```
# Add lens distortion information
```

```
policy += f"Lens Distortion: {self.format_lens_distortion(data.lens_distortion)}"
```

```
# Add camera height information
```

```
if data.Height_Wrt_Ground_Applicable:
```

```
    if data.Height_Changes_From_Low_To_High or data.Height_Changes_From_High_To_Low:
```

```
        policy += f"Camera Height: The camera starts {self.format_camera_height_start(data.height_wrt_ground_info['start'])} and then moves {self.format_camera_height_end(data.height_wrt_ground_info['end'])}."
```

```

    else:
        policy += f"Camera Height: The camera is {self.format_camera_height_start(data.
            height_wrt_ground_info['start'])}."
    elif data.Overall_Height_Description != "":
        policy += f"Camera Height: {data.Overall_Height_Description}"
    else:
        policy += "Camera Height: The camera height is unclear or not significant enough to mention (
            no need to mention).\"

# Add camera angle information
if data.Camera_Angle_Applicable:
    if data.Camera_Angle_Changes:
        policy += f"Camera Angle: The camera angle is initially at {self.format_camera_angle(data.
            camera_angle_info['start'])} and then changes to {self.format_camera_angle(data.
            camera_angle_info['end'])} due to camera motion.\"
    else:
        policy += f"Camera Angle: The camera angle is at {self.format_camera_angle(data.
            camera_angle_info['start'])}.\"

# Check for Dutch angle
if data.Dutch_Angle_Varying:
    policy += \" The camera is also at a dutch angle that varies due to camera rolling.\"
elif data.Dutch_Angle_Fixed:
    policy += \" The camera is also at a fixed dutch angle during the video.\"
elif data.Camera_Angle_Description != "":
    policy += f"Camera Angle: {data.Camera_Angle_Description}\"
else:
    policy += "Camera Angle: The camera angle is unclear or not significant enough to mention (no
        need to mention).\"

# Add focus information
if data.Focus_Applicable:
    if data.Deep_Focus:
        policy += "Camera Focus: The camera uses a deep focus with a large depth of field.\"
    else:
        if data.Ultra_Shallow_Focus:
            policy += "Camera Focus: The camera uses an extremely shallow depth of field, focusing
                on a very narrow plane.\"
        else:
            policy += "Camera Focus: The camera uses a shallow depth of field, keeping a limited
                range in focus.\"

        if data.Focus_Changes:
            policy += f\" The camera starts {self.format_focus_plane(data.focus_info['start'])}, and
                later becomes {self.format_focus_plane(data.focus_info['end'])}.\"
        else:
            policy += f\" The camera is {self.format_focus_plane(data.focus_info['start'])}.\"

        if data.Rack_Pull_Focus:
            policy += \" The focus plane changes through a rack focus.\"
        elif data.Focus_Tracking:
            policy += \" The camera uses focus tracking to keep the subject in focus.\"
    elif data.Camera_Focus_Description != "":
        policy += f"Camera Focus: {data.Camera_Focus_Description}\"
    else:
        policy += "Camera Focus: The camera focus is unclear or not significant enough to mention (no
            need to mention).\"

```

```

# Add camera motion information
if data.Fixed_Camera:
    if data.Fixed_Camera_With_Shake:
        policy += "**Camera Motion:** The camera is fixed but slightly unsteady, with no intentional
            movement."
    else:
        policy += "**Camera Motion:** The camera is completely static, with no movement or shaking."
else:
    if data.Complex_Motion:
        policy += f"**Camera Motion:** {data.complex_motion_description}"
    else:
        if data.Minor_Motion:
            policy += "**Camera Motion:** The camera shows some minor movement."
        elif data.Simple_Motion:
            policy += "**Camera Motion:** The camera shows a clear movement pattern."

# Add specific movement description
policy += " " + self.get_movement_description_simple(data)

# Add tracking information
if data.Tracking_Shot:
    policy += f"**Subject Tracking:** {self.get_tracking_description(data)}"

# Add steadiness information
policy += f"**Camera Steadiness:** {self.format_camera_steadiness(data.steadiness)}."

# Add motion speed if not regular
if data.camera_motion_speed != "regular":
    policy += f"**Camera Motion Speed:** The camera is {self.format_camera_motion_speed(data.
        camera_motion_speed)}"

policy += "If possible, specify the subject that the camera focuses on when describing camera work.
    For instance, use 'focus on the man in the foreground' rather than 'focus on the foreground.'
    Likewise, if the camera follows a subject, avoid the generic phrase 'tracking the subject(s)'.
    Instead, identify the subject and describe the specific type of tracking."

return policy

```

Listing 5. Camera Framing & Dynamics Policy

Primitives used (Camera Setup): Regular Speed, Slow Motion, Fast Motion, Time Lapse, Stop Motion, Time Reversed, Speed Ramp, No Lens Distortion, Barrel Distortion, Fisheye Distortion, Height Wrt Ground Is [Ground Level, Hip Level, Eye Level, Overhead Level, Aerial Level, Water Level, Underwater Level], Height Wrt Ground Applicable, Height Changes From Low To High, Height Changes From High To Low, Camera Angle Is [Bird Eye Angle, High Angle, Level Angle, Low Angle, Worm Eye Angle], Camera Angle Applicable, Camera Angle Changes, Dutch Angle Fixed, Dutch Angle Varying, Deep Focus, Shallow Focus, Ultra Shallow Focus, Focus Applicable, Focus Changes, Focus Is [Foreground, Middle Ground, Background, Out Of Focus], Rack Pull Focus, Focus Tracking.

Primitives used (Camera Motion): Fixed Camera, Fixed Camera With Shake, Moving Camera, Clear Moving Camera, Simple Motion, Complex Motion, Minor Motion, Stable Camera Motion, Very Stable Camera Motion, Shaky Camera, Very Shaky Camera, Slow Moving Camera, Fast Moving Camera, Forward, Backward, Upward, Downward, Leftward, Rightward, Pan Left, Pan Right, Tilt Up, Tilt Down, Roll Clockwise, Roll Counterclockwise, Zoom In, Zoom Out, Arc Clockwise, Arc Counterclockwise, Crane Up, Crane Down, Tracking Shot, Aerial Tracking Shot, Arc Tracking Shot, Front Side Tracking Shot, Rear Side Tracking Shot, Lead Tracking Shot, Follow Tracking Shot, Tilt Tracking Shot, Pan Tracking Shot, Side Tracking Shot, Side Tracking Shot Leftward, Side Tracking Shot Rightward, Tracking Subject Larger Size, Tracking Subject Smaller Size.

Non-primitive attributes: complex_motion_description, Overall_Height_Description, Camera_Angle_Description, Camera_Focus_Description.