

# InstanceCap: Improving Text-to-Video Generation via Instance-aware Structured Caption

Tiehan Fan<sup>1\*</sup> Kepan Nan<sup>1\*</sup> Rui Xie<sup>1</sup> Penghao Zhou<sup>2</sup>

Zhenheng Yang<sup>2</sup> Chaoyou Fu<sup>1</sup> Xiang Li<sup>3</sup> Jian Yang<sup>1</sup> Ying Tai<sup>1✉</sup>

<sup>1</sup> Nanjing University <sup>2</sup> ByteDance <sup>3</sup> Nankai University

<https://github.com/NJU-PCALab/InstanceCap>

## Abstract

*Text-to-video generation has evolved rapidly in recent years, delivering remarkable results. Training typically relies on video-caption paired data, which plays a crucial role in enhancing generation performance. However, current video captions often suffer from insufficient details, hallucinations and imprecise motion depiction, affecting the fidelity and consistency of generated videos. In this work, we propose a novel instance-aware structured caption framework, termed InstanceCap, to achieve instance-level and fine-grained video caption for the first time. Based on this scheme, we design an auxiliary models cluster to convert original video into instances to enhance instance fidelity. Video instances are further used to refine dense prompts into structured phrases, achieving concise yet precise descriptions. Furthermore, a 22K InstanceVid dataset is curated for training, and an enhancement pipeline that tailored to InstanceCap structure is proposed for inference. Experimental results demonstrate that our proposed InstanceCap significantly outperform previous models, ensuring high fidelity between captions and videos while reducing hallucinations.*

## 1. Introduction

Recently, text-to-video (T2V) generation with advanced diffusion transformers (DiT) [2, 8, 10–12, 15, 18, 21, 23, 25, 32, 36] have attracted significant attention for the ability to generate realistic, long-duration videos based on text prompts. Video-caption paired data is typically used in training and plays a crucial role in enhancing generation performance. Current video recaption methods often incorporate multimodal large language models to produce detailed captions, which however usually suffer from hallucinations, leading to inconsistencies between captions and

video content. Consequently, creating *consistent video-caption pairs with accurate details and precise motion depiction* for T2V generation remains a significant challenge.

As shown in Figure 1, current video recaption methods can be broadly categorized into three types: 1) Short captions, such as Panda-70M [4], lack sufficient coverage of video content, leading to low fidelity. 2) Dense captions, like ShareGPT4Video [3], enrich textual content but suffer from hallucination issues, often generating meaningless or inaccurate video content. 3) Coarse-level structured captions, exemplified by MiraData [9], improve video quality but provide coarse-level details. Moreover, the redundancy introduced by MLLM across structures diminishes its overall effectiveness. To this end, achieving accurate captions remains two crucial challenges: 1) *High fidelity between caption and video*: Retain as much of the original video’s objects, textures, and motion information as possible. 2) *Accurate content in caption*: Enable MLLM model to generate precise content, minimizing hallucinations and repetition.

To address the challenges, we propose a novel instance-aware structured caption framework, termed InstanceCap, to achieve instance-level and fine-grained video caption for the first time. Our structure is specifically designed to incorporate *instances, background, and camera movement*. For each instance, we specify *class, appearance, actions, motion, and position*. To enhance the fidelity and accuracy of video captions, we focus on two key aspects: 1) *From Global Video to Local Instances*: For each instance, we propose an auxiliary models cluster (AMC) to isolate it from the original video and obtain the corresponding position and category information. This operation minimizes interference from unrelated regions while retaining as much of the original video’s information as possible. 2) *From Dense prompt to Structured Phrases*: We leverage multimodal large language models (MLLMs) in an improved Chain-of-Thought (CoT) process to obtain concise yet accurate descriptions of textures, camera movement, actions and motion for each instance. This reduces the probability of hallucinations and irrelevant content produced by the lan-

\* Equal contributions. Ying Tai is the corresponding author.

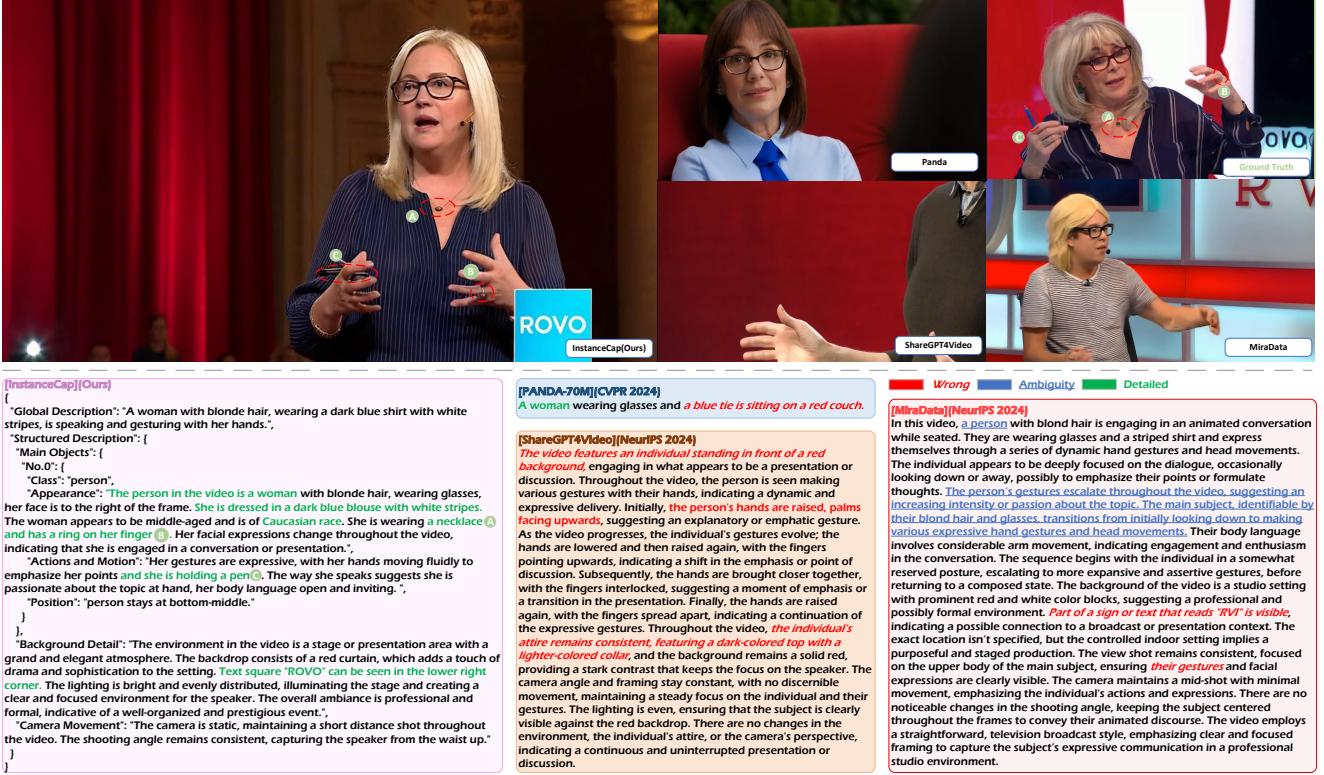


Figure 1. **Top:** Comparison of the reconstruction-via-recaption results between InstanceCap and state-of-the-art captioning methods for annotating the ground truth video. InstanceCap produces results that more closely resemble the original video, showing greater detail fidelity (highlighted by the red circle). **Bottom:** The corresponding captions generated by InstanceCap and others. **Red** denotes incorrect captions, **blue** represents ambiguous captions, and **green** indicates detailed and accurate descriptions of video. Specific visual hints are marked as A, B, and C for clarity. All videos are generated using the *same* video generation product, **Hailuo AI**<sup>1</sup>, which has robust prompt-following capabilities, clearly highlighting the effectiveness of InstanceCap.

guage model compared to traditional caption methods that directly describe video content in a complex, dense caption.

To validate the effectiveness of InstanceCap in T2V generation, we constructed a high-definition video dataset comprising 22K samples to create a training dataset with our instance-aware structured captions, named InstanceVid. At the inference stage, we also implemented a prompt enhancement pipeline tailored to our structured captioning method, enabling the generation of concise captions that better align with user needs. Our InstanceCap integrates seamlessly with existing diffusion models. Experimental results demonstrate that after finetuned with our InstanceVid, the T2V model exhibits better ability with prompt following on details and motion actions. In summary, our main contributions are as follows:

- We propose InstanceCap, the first instance-aware structured caption method for text-to-video generation. We also constructed a 22K InstanceVid dataset during training, and developed an enhancement pipeline that tailored to InstanceCap structure during inference.
- We design the AMC paradigm to convert global video

into instances, enhancing instance fidelity. Additionally, we propose an improved CoT pipeline with MLLMs to refine dense prompts into structured phrases, achieving concise yet precise descriptions.

- Extensive experiments on video reconstruction demonstrate that our InstanceCap significantly enhances the fidelity between captions and videos. T2V models finetuned on our InstanceVid further achieve more precise generation on instance details and motion actions.

## 2. Related works

**Video recaptioning.** Advancements in text-to-video generation demand high-quality video-text datasets to build robust foundational video models for visual-language alignment. Current video recaption methods fall into two main categories: manual annotation [1, 24, 37] and end-to-end recaption using multimodal large language models (MLLM). Although manual annotation provides higher accuracy, scaling datasets to meet the needs of high-quality video generation models remains a substantial challenge. Recent

<sup>1</sup><https://hailuoai.com/video>

advances in MLLM have demonstrated impressive capabilities in video understanding and description generation. Panda [4] and InternVid [26] are with short captions, offering computational efficiency but frequently omitting crucial content and exhibit low video fidelity. OpenVid-1M [16], Vript [28] and ShareGPT4Video [3] are with dense captions, which provide richer content but face challenges: Hallucinations due to complexity, inclusion of redundant information, and text encoder truncation caused by excessively long text. MiraData [9] achieves coarse-level structured captions that attempts to mediate these issues but struggle with fine detail and redundancy across structures. Different from the existing video recaption methods, InstanceCap is the first *instance-aware structured caption* approach for text-to-video generation, ensuring high fidelity between caption and video while reducing hallucinations and repetition.

**Text-to-video generation.** Despite numerous high-quality video generation models [2, 8, 10, 11, 18, 30] perform well with simple directives, they often falter with complex prompts requiring precise instance-level details or intricate camera movements. Analyzing current video-text datasets suggests these limitations may stem from suboptimal training data quality. Traditional recaption methods have not sufficiently captured instance-specific detail granularity or provided comprehensive descriptions of camera movements. To enhance instance-level details and motion consistency, we construct a 22K InstanceVid dataset for training/fine-tuning, and develop an enhancement pipeline InstanceEnhancer that tailored to the proposed instance-aware structure during inference.

### 3. Method

In Section 3.1, we first present the InstanceCap pipeline, as shown in Figure 2. Based on this pipeline, we recaption the carefully curated dataset InstanceVid in Section 3.2, enhancing T2V models’ instance generation. Additionally, in Section 3.3, we introduce InstanceEnhancer to convert short prompts into our proposed instance-aware structured caption format during inference.

#### 3.1. InstanceCap

**Video preprocessing with auxiliary model cluster.** For continuous video processing, we implemented uniform sampling using decord<sup>2</sup>, following the methodology established in LLaVA-Video [34]. This approach enables us to extract essential temporal metadata, including duration, frame count, and timestamps, allowing MLLMs to better interpret temporal sequences in recaptioning tasks. Additionally, to enhance MLLMs’ capabilities through structured guidance, we incorporate several auxiliary models to

achieve accurate object detection, video instance segmentation and camera motion prediction, providing precise prior information to the MLLMs.

**Global description, background detail and camera movement.** When describing video content, a high-quality global description should capture primary elements, environmental context, camera movements, angles, and tonal qualities. MLLMs excel in generating high-level video summaries using Chain-of-Thought methodology. By employing carefully designed prompts with CoT, we can guide MLLMs to produce detailed background descriptions while minimizing references to foreground elements.

However, MLLMs’ limitations in processing discrete frames rather than continuous video segments make it challenging to *distinguish camera motion from instance action*. To address this, we achieve camera annotations from OpenSora [35] for basic movements (*e.g.*, zoom, rotation) and rely on MLLMs to capture subtle motion attributes (*e.g.*, intensity, speed). The integration of camera movement indicators with MLLM capabilities provides comprehensive annotations, as illustrated in Figure 10 (a).

**Structured Description on Instances.** In this subsection, we introduce the details to achieve our instance-aware structured description. To address MLLMs’ limitations in instance annotation and the suboptimal results of directly adapting *weak visual prompts* from images to videos [22, 31], we make full use of the auxiliary model cluster, including initial object detection [38], video instance segmentation with SAM2 [20], and blur non-instance regions to achieve blur background, resulting in better recaptioning outcomes compared to alternative visual prompt methods in video, as shown in Figure 11. To this end, we *decompose the global videos into local instances*.

Next, we describe how to achieve detailed and accurate description of each instance (Figure 3). To maintain instance-level precision, we deliberately constrain the information accessible to MLLMs during instance annotation. Crucially, the global video remains **invisible** due to our designed blurred backgrounds, preventing MLLMs from *confusing information across multiple instances*. This approach allows InstanceCap to focus on local instances identified through auxiliary model cluster. Furthermore, to avoid the potential limitation of MLLMs seeing only isolated instances, which could lead to overlooking inter-instance interactions and subsequent misinterpretations, we incorporate the global description mentioned in previous subsection. Specifically, we *inject the global description into the instance-annotation MLLMs*. This strategic mitigates potential biases in instance action descriptions while maintaining instance-specific accuracy.

To enhance the capability of InstanceCap in capturing instance-level details, we introduce novel insights into the improved CoT process. Our analysis of current MLLM-

<sup>2</sup><https://github.com/dmrc/decord>

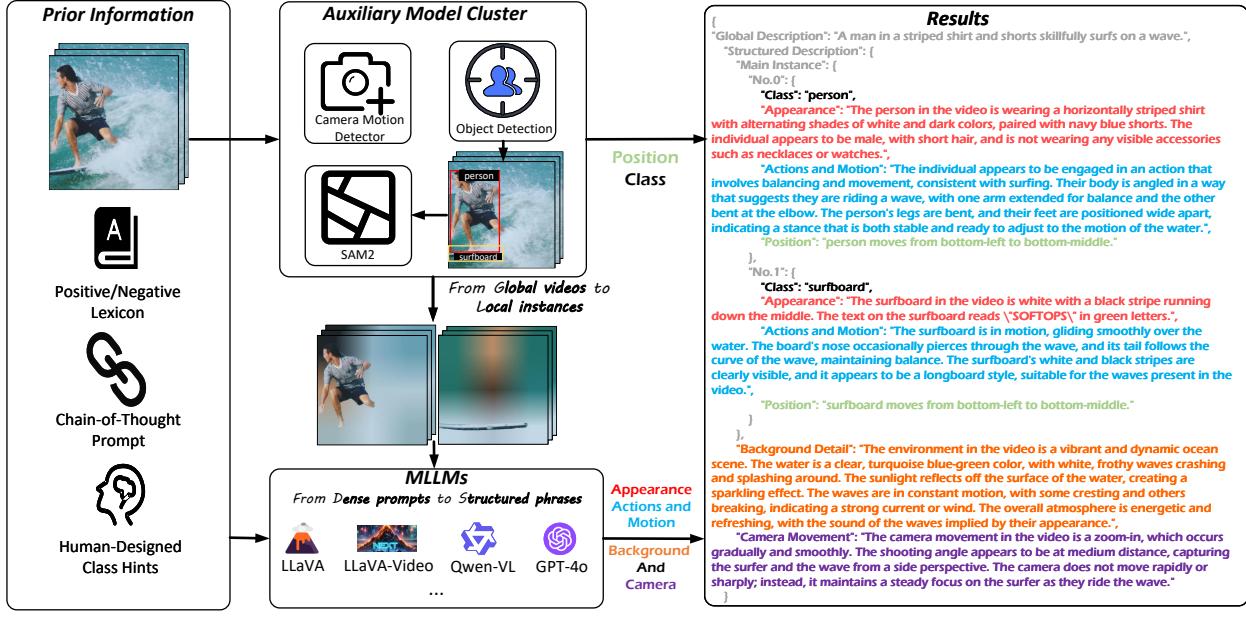


Figure 2. **Overview of InstanceCap pipeline.** Details of “from dense prompts to structured phrases” design are shown in Figure 3.

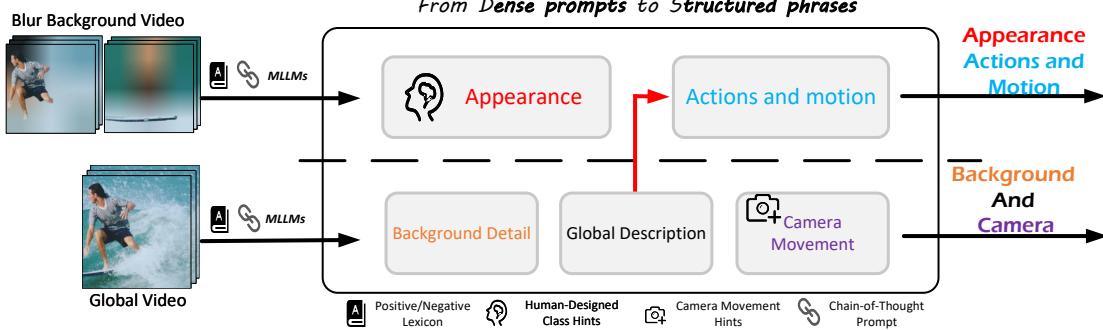


Figure 3. Details on “from dense prompts to structured phrases” design. We propose an improved CoT pipeline with carefully designed information interactions (red arrow), which facilitates MLLMs to accurately capture instances with precise descriptions on attributes.

based video reception methods shows that simple prompts, like “Please provide a detailed description of this video” or Chain-of-Thought prompts “Let’s think step by step... First, please note... Finally, summarize the video content...” fail to capture precise instance details. Additional experiments reveal that MLLMs can effectively annotate these details when given fine-grained prompts, such as “Please note if the characters have any accessories” or “Please observe whether there are spots on the bananas”. To enhance the details of instances, we developed **Human-designed Class Hints**, crafting specific prompts for about 80 detectable categories using our auxiliary models cluster. Specifically, we present the “person” class prompt here: “Please focus primarily on the person’s facial expressions, attire, age, gender, and race in the video and give a detailed description. Please mention if there are any necklaces, watches, hat or other decoration; otherwise, there’s no need to bring them up.” Besides, we also developed a curated **Positive/Negative Lexicon** to guide MLLMs in generating more aesthetically refined captions. More details can be found in our supplementary material.

**3.2. InstanceVid**

**Data collection.** InstanceVid is curated via refining a subset from the high-aesthetic, high-consistency videos from OpenVid-1M [16]. To showcase our method’s high-fidelity labeling of instance details and motion, we selected video samples that included at least one instance exhibiting high motion intensity during dataset filtering.

**Statistical analysis of InstanceVid.** Figure 4 illustrates the statistical characteristics of InstanceVid across two main dimensions: video scenes, and temporal durations. Our data collection emphasizes videos with distinct instances while ensuring a balanced representation of outdoor scenes to prevent biases from an overemphasis on

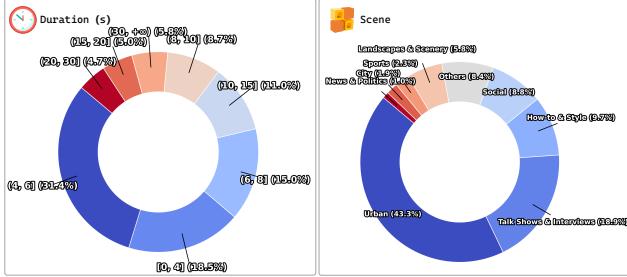


Figure 4. InstanceVid provides structured captions for videos in open-domain scenarios, featuring diverse instance, expansive scenes, precise and instance-aware captions, and video-generation-friendly durations.

instance-focused content. We achieve detailed descriptions capturing human movements, physical appearances, and documentation of common objects and animals. Besides, InstanceVid focuses on short-duration videos (2-10 seconds) for two main reasons. First, OpenVid-1M segments longer sequences to eliminate excessive scene transitions. Second, most of the current open-source T2V models are optimized for video generation within this duration range.

### 3.3. InstanceEnhancer

When the caption distribution of training data differs from that of inference text, it may result in poor instruction-following performance or even problematic outputs. This issue is evident in T2V generation, particularly when long captions are used for training but short captions for inference, leading to subpar results. Since users typically prefer short captions, it is essential to enhance short caption effectively to better align with our proposed instance-aware structured caption during training.

As shown in Figure 5, we introduce a tuning-free approach called InstanceEnhancer that achieves this by strictly limiting the generated formats to match the caption corresponding to the training input we used. Our method differs from existing tuning-free caption enhancement approaches, such as those presented in RPG [29]. Instead of directly enhancing short captions, which we found can introduce inconsistencies between multiple instances’ actions and their environmental context in video generation, we employ a two-stage enhancement strategy. In Stage A, short prompts are expanded into detailed long prompts. Stage B(I)&(II) uses both expanded and original captions to segment and enhance specific instances, preserving contextual coherence while ensuring precise instance identification. Due to the space limitation, more details of our enhance pipeline can be found in the supplementary material.

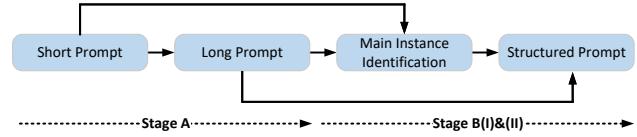


Figure 5. High-level overview of InstanceEnhancer, illustrating the data flow and the partitioning of stages. For a detailed implementation, refer to the supplemental materials, which provide an in-depth description of the enhancer pipeline design and the inter-dependencies between the stages.



Figure 6. Comparison on reconstruction-via-recaption between InstanceCap and MiraData. Corresponding 3DVAE scores are also indicated. Similar semantics shared between InstanceCap and GT are indicated by red circles and lines.

## 4. Experiments

### 4.1. Experimental setup

**Video reconstruction with recaptions.** To comprehensively evaluate InstanceCap, we conducted a series of experiments, benchmarking against state-of-the-art methods including Panda-70M [4], ShareGPT4Video [3], and MiraData [9]. To this end, we carefully selected 100 video clips from OpenVid-1M [16] and Animal Kingdom [17]. For each video, we generated one caption using various caption models, which were then input into the advanced T2V model CogvideoX-5b [30] for video generation. We calculated the differences between the generated videos and the ground truth videos to evaluate each caption model’s performance, where smaller visual differences indicate more accurate captions and higher fidelity.

We introduced several metrics to evaluate the video reconstruction performance: 1)  $3DVAE_{score}$ : Using 3DVAE from CogVideoX [30] as the backbone, we extract hidden-space representations from both the original videos and their recaption-reconstructed counterparts. These representations quantify the perceptual distance between them. 2)  $CLIP_{SenbySen}$ : To handle CLIP’s 77-token processing limit, we segment long captions into individual sentences

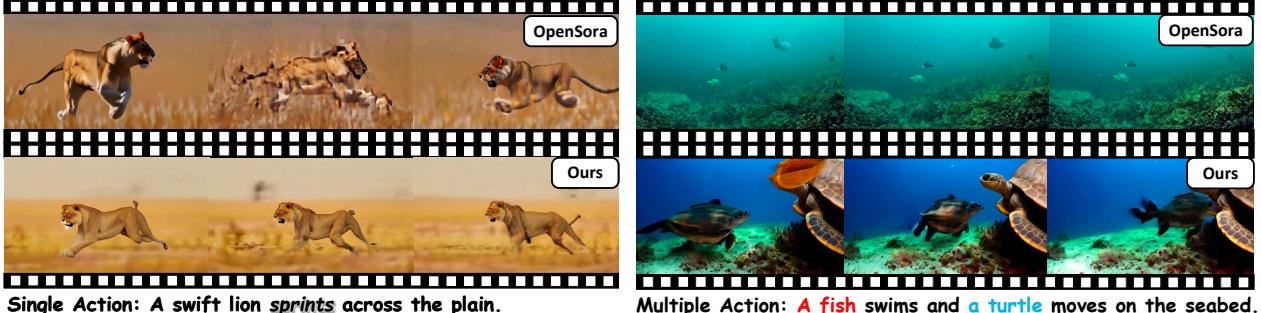


Figure 7. Visual comparison of InstanceCap and Opensora on Single and Multiple Action Score. In terms of the dynamic degree of video generation, we show better consistency and enhanced multi-instance dynamic generation effect.

and compute CLIP [19] similarity between each sentence and every original video frame. The final score is obtained by first averaging the similarity scores of each sentence across all frames, then averaging these sentence-level scores for a comprehensive result. 3) Human Evaluation: We conducted a user study with a panel of evaluators to assess caption quality across two aspects: Instance Detail (ID) and Hallucination Scores (HS).

**T2V generation.** To thoroughly evaluate the T2V generation performance of our InstanceCap, we utilize the InstanceVid dataset to finetune the state-of-the-art DiT-based T2V generation model Open-Sora [35]. In our evaluation, we compare with Open-Sora, CogVideoX-5b [30], Pyramid-Flow [8], and Open-Sora-Plan [11]. To enable fine-grained, instance-level assessment, we construct a highly challenging evaluation benchmark called Inseval, inspired by recent advancements in T2I and T2V evaluation [5, 6, 13, 14, 27]. Specifically, we curate a diverse evaluation dataset of over 200 carefully crafted instance-level prompt-answer pairs, covering both single-object and multi-object scenarios systematically across five key dimensions: Action, Color, Shape, Texture, and Detail. Motivated by the previous evaluation benchmarks [6], we implement a CoT reasoning framework for generating structured QA responses to ensure objective and consistent evaluation, allowing us to derive instance-level evaluation scores that align closely with human perception and preferences. This approach provides a more nuanced and reliable assessment of instance-level generation quality.

## 4.2. Comparison with SOTA caption models

**Qualitative evaluation.** CogVideoX-5b [30] is a latent text-to-video generation model known for its capability to generate realistic, long-duration videos based on text prompts. With the integration of our InstanceCap into CogVideoX-5b, as substantiated by Figure 6, the model exhibits a notable enhancement in the video reconstruction capacity. This demonstrates that our instance-aware structured captions retain more of the original video’s

information, leading to higher fidelity. For instance, our InstanceCap can retain information such as “glasses”, “grey sweater”, and “relative position of two people”, whereas MiraData [9] almost completely loses these important details. A similar conclusion can be drawn from Figure 1. These results underscore the significant improvements achieved by our InstanceCap, resulting in high-quality reconstruction characterized by rich detail and high fidelity between our captions and the original videos.

**Quantitative evaluation.** Table 1 presents quality comparisons between our InstanceCap and other caption methods across two metrics. Based on the results, we make the following observations: 1) Our method delivers comparable or superior quality to the four baselines, demonstrating its ability to enhance fidelity between videos and captions. This strong alignment with human perceptual judgments and preferences is evident in Figures 1 and 6. 2) Our method consistently excels across all metrics for captions under 200 words, an acceptable length for most T2V models, illustrating its generalizability. Figure 8 compares MiraData’s captions with our instance-aware structured ones. We randomly selected videos from an open-domain dataset, and a panel of evaluators assessed caption quality using standardized criteria. Results indicate that our captions offer significantly richer and more accurate descriptions while reducing hallucination artifacts compared to MiraData’s output.

## 4.3. Text-to-video generation

**Qualitative evaluation.** Figures 7 and 9 provide visual comparisons of the T2V generation results. It can be observed that the infusion of our InstanceVid dataset into Open-Sora [35] serves to further enhance its video synthesis capabilities across four fundamental dimensions (Action, Color, Shape, and Texture). These four different aspects correspond to information in our instance-aware structured captions such as “Actions and Motion”, “Appearance”, etc. For instance, our model accurately generates the “sprints” action of the lion in Figure 7, as opposed to Open-

Captioning Methods	$3DVAE_{score} \downarrow$	$CLIP_{SenbySen} \uparrow$	Avg. Length
Panda-70M	140.25	0.1956	13 words
ShareGPT4Video	141.00	0.2132	191 words
LLaVA-Video-72B	139.88	0.2060	102 words
MiraData(GPT-4o)	<u>137.50</u>	<b>0.2156</b>	263 words
<b>InstanceCap(Ours)</b>	<b>134.25</b>	<u>0.2133</u>	157 words

Table 1. Quantitative comparisons on reconstruction-via-recaption results. The best results are marked in **bold**, and the second-best are underlined. As a reference, CogVideoX-5b accepts 226 text tokens, with any excess being truncated.

T2V Model	Single $\uparrow$					Multiple $\uparrow$			Average $\uparrow$
	Action	Color	Shape	Texture	Detail	Action	Color	Texture	
CogVideoX-5B [30]	64%	60%	44%	60%	20%	8%	48%	40%	43.00%
Pyramid-Flow-2B [8]	44%	68%	32%	32%	7%	4%	24%	16%	28.38%
Open-Sora Plan v1.3-2.7B [11]	64%	44%	36%	32%	27%	20%	32%	12%	33.38%
Open-Sora v1.2-1.1B [35]	40%	<u>56%</u>	<u>36%</u>	<u>40%</u>	13%	12%	16%	16%	28.63%
+ InstanceCap(Ours)	<b>56%</b>	<b>60%</b>	<b>40%</b>	<b>48%</b>	<b>27%</b>	<b>16%</b>	<b>32%</b>	<b>24%</b>	<b>37.88%</b>
+ Panda-captioner [4]	40%	48%	28%	40%	20%	8%	20%	12%	27.00%
+ ShareGPT4Video [3]	40%	44%	32%	24%	13%	<b>16%</b>	8%	20%	24.63%
+ LLaVA [16]	52%	52%	28%	28%	20%	12%	28%	16%	29.50%

Table 2. Quantitative comparison between InstanceCap and SOTA video captioning models, all based on the popular T2V model Open-Sora. Additionally, we also compare three powerful T2V models, including CogVideoX-5B, Pyramid-Flow, and Open-Sora Plan. The best results of video captioning methods and Open-Sora are marked in **bold**, and the second-best are underlined.

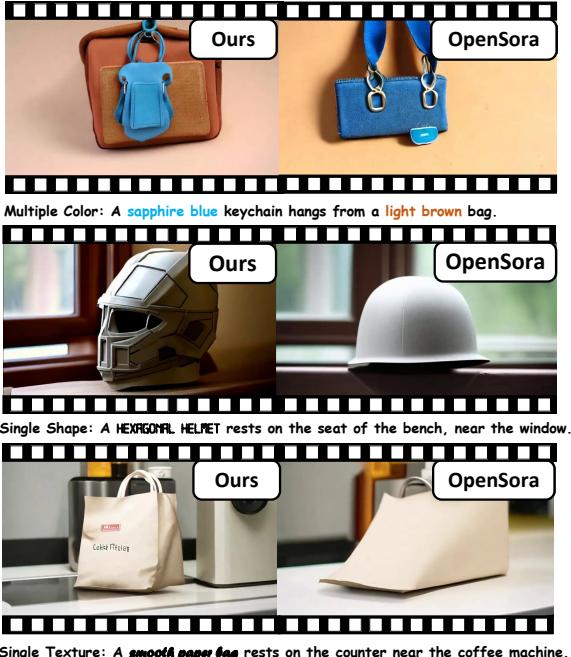


Figure 9. Visual comparison of InstanceCap and OpenSora on *instance-level attributes*. InstanceCap excels in precise instance detail fidelity and instruction-following capabilities, even with complex multi-instance and multi-attribute scenarios.

Sora [35]. In Figure 9, benefiting from our instance-aware caption, our model generates the accurate “light brown

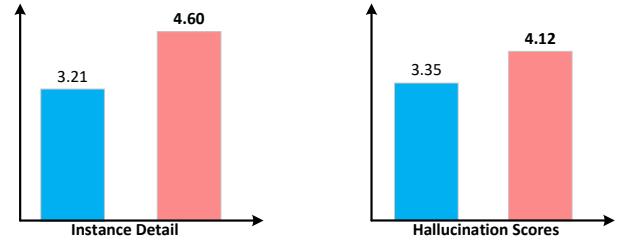


Figure 8. User study on instance detail and hallucination scores. Our instance-aware structured caption shows clear advantages compared to the coarse-structured MiraData [9].

bag” instance described in caption, where Open-Sora [35] completely loses this instance. These results indicate that our InstanceVid can provide accurate and instance-level guiding information for video generation models.

**Quantitative evaluation.** We conduct the quantitative evaluation for InstanceCap using the proposed Inseval metrics in Table 2. We can draw the following conclusions: 1) Fine-tuning with InstanceVid consistently improves all metrics over the base model Open-Sora, demonstrating the effectiveness of InstanceCap. In particular, our Detail score ranks first, justifying the capacity of InstanceCap to capture complex instance detail in video. 2) Compared to other video captioning models finetuned based on Open-Sora, InstanceCap shows clear advantages in video generation tasks. 3) Compared to larger models like CogVideoX or Pyramid-Flow, our approach achieves a higher average metric than Pyramid-Flow, and performs comparably to CogVideoX in several specific metrics like ‘Single-Color/Shape/Detail’ and ‘Multiple-Action’, but with much fewer parameters.

#### 4.4. Ablation Study

**Effects of human design class hints and camera movement hints.** We discuss the impact of incorporating human-designed class hints and camera movement hints on annotation outcomes and provide relevant caption visualizations in Figure 10. These annotations aid MLLMs in focusing more precisely on key elements, resulting in richer and



Figure 10. (a) Ablation study on the effect of camera movement hints on the accuracy of MLLM labeling. (b) Impact of human-designed class hints on the details of instance labeling.



Figure 11. (a) Comparison against the weak visual prompt for reconstruction-via-caption visualization on multi-instance targets. (b) Comparison against color screen backgrounds (red), which may negatively affect MLLM labeling performance.

more accurate annotations.

**Ablations on different video visual prompts.** Comparative results for various video visual prompt methods used in caption generation are shown in Figure 11. As illustrated in (a), weak visual prompts derived from static image techniques, such as red circles, bounding boxes, or selective grayscale manipulation of non-target areas [22, 31], limit MLLMs in distinguishing and describe specific targets in multi-instance scenes, leading to attribute blending and vague annotations across instances. In contrast, our method excels in instance-specific feature extraction, accurately differentiating figures like the coach and players. Figure 11 (b) illustrates strong visual prompts that involve complete occlusion of non-target regions to eliminate MLLM interactions with irrelevant instances. Conventional methods use primary color screens, but this often misguides MLLMs, causing them to incorporate incorrect context in captions. Our designed blur background masking approach, however,

preserves visual consistency with natural scenes, enabling MLLMs to generate accurate and contextually relevant annotations with minimal prompting guidance.

## 5. Conclusions and Limitations

In this paper, we introduce InstanceCap, the first instance-aware structured caption method for text-to-video generation. We design an Auxiliary Models Cluster (AMC) to convert global video into instances, enhancing instance fidelity. we also propose an improved CoT pipeline with MLLMs to refine dense prompts into structured phrases, achieving concise yet precise instance descriptions compared to the previous video caption models. Additionally, based on InstanceCap, we curated InstanceVid dataset for training and InstanceEnhancer during inference, significantly enhancing T2V models’ generation capabilities on instance details and actions.

**Limitations.** Since the precision of InstanceCap partly depends on object detection methods, requiring fine-tuning of the detection model for domain-specific instances, and its benefits decrease in instance-free scenes. Furthermore, the scale of InstanceVid limits its use as a large-scale pre-training dataset. Moving forward, we plan to apply InstanceCap to a larger video dataset and train more powerful T2V models to amplify its impact.

## References

- [1] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, 2011. 2
- [2] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320, 2024. 1, 3
- [3] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 1, 3, 5, 7, 2
- [4] 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. *arXiv preprint arXiv:2402.19479*, 2024. 1, 3, 5, 7
- [5] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xi-hui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 6
- [6] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 6, 2
- [7] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 4
- [8] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 1, 3, 6, 7
- [9] 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. *arXiv preprint arXiv:2407.06358*, 2024. 1, 3, 5, 6, 7
- [10] Kuaishou. Kling. <https://kling.kuaishou.com>, 2024. 1, 3
- [11] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 3, 6, 7
- [12] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhui Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*, 2024. 1
- [13] 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, 2024. 6
- [14] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023. 6, 2
- [15] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 1
- [16] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 3, 4, 5, 7
- [17] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19023–19034, 2022. 5
- [18] Pika. Pika 1.0. <https://pika.art>, 2023. 1, 3
- [19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 6, 2
- [20] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3
- [21] Runway. Gen-2. <https://research.runwayml.com/gen2>, 2023. 1
- [22] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11987–11997, 2023. 3, 8
- [23] Jiniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1
- [24] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, 2020. 2

- [25] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. LaVie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1
- [26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023. 3
- [27] Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. Ifadapter: Instance feature control for grounded text-to-image generation, 2024. 6
- [28] 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, 2024. 3
- [29] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multi-modal llms. In *Forty-first International Conference on Machine Learning*, 2024. 5
- [30] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3, 5, 6, 7, 2
- [31] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024. 3, 8
- [32] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *IJCV*, 2024. 1
- [33] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 2
- [34] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 3
- [35] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 3, 6, 7
- [36] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. <https://github.com/hpcaitech/Open-Sora>, 2024. 1
- [37] Luowei Zhou, Nathan Louis, and Jason J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction, 2018. 2
- [38] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023. 3

# InstanceCap: Improving Text-to-Video Generation via Instance-aware Structured Caption

## Supplementary Material

In this supplementary material, we present comprehensive details and analyses across the following sections:

- **Section 1** elucidates our methodology for constructing positive/negative lexical databases, accompanied by their details.
- **Section 2** provides an extensive compilation of human-designed class hints, demonstrating their diverse applications.
- **Section 3** delineates the improved Chain-of-Thought prompting strategies employed in Figure 3, with particular emphasis on their methodological improvements.
- **Section 4** explicates the architectural framework of InstanceEnhancer, supplemented with exemplary prompts utilized in our Large Language Model implementations.
- **Section 5** elaborates a detailed discussion of the principles behind our metric design for video reconstruction, including mathematical formulations and empirical validations.
- **Section 6** demonstrates the prompts used by Inseval in both the inference and evaluation stages.
- **Section 7** presents a evaluation of our methodology across both commercial and open-source models, including experimental results and analytical findings.

### 1. Positive/Negative Lexicon

To enhance the aesthetic quality of generated videos, we carefully collected prompts from various open-source model galleries, extracting adjectives to build a *Positive Lexicon*. Conversely, we manually constructed a *Negative Lexicon*, which was further enriched using the powerful LLM, GPT-4o. Both lexicons were refined through meticulous manual screening. The detailed contents of the Positive/Negative Lexicons are shown in Figure S1.

### 2. Human-designed Class Hints

For the Human-designed Class Hints, we carefully crafted additional prompts for over *eighty* categories, each specifically tailored to its specific characteristics. Below, we present twenty of these categories. The full JSON-formatted hints for all classes, ready for direct use, will be provided in the code we plan to release later.

- **Person:** “Please focus primarily on the person’s facial expressions, attire, age, gender, and race in the video and give description in detail. Please mention if there are any necklaces, watches, hat or other decoration; otherwise, there’s no need to bring them up.”

 **Positive Lexicon**  
Select the appropriate ones of following words in your description: kaleidoscopic, delicate, grand, gentle, soothing, cool, mature, solitary, worn, chaotic, dramatic, cozy, shimmering, desolate, serene, weathered, whispering, loose-fitting, vibrant, tranquil, dimly-lit, purplish, introspective, artfully, sleek, energetic, overcast, brilliant, slender, graceful, picturesque, whimsical, contented, gentle, warm, tender, pastel-colored, elegant.

 **Negative Lexicon**  
Do not use any of the following negative words when describing: dull, rough, harsh, chaotic, cluttered, bleak, uninspired, garish, stiff, unrefined, artificial, heavy, disorderly, grim, rusty, faded, cramped, jarring, obtrusive, awkward, ordinary, harsh, gloomy, cold, rigid, overcrowded, mismatched, messy, uneven, tacky, lifeless, unbalanced, heavy-handed, overbearing, dissonant, grating, oversaturated, unpleasant, rigid, blur.

Figure S1. The detail of Positive/Negative Lexicon

- **Bicycle:** “Please describe the bicycle in terms of color, type, size, condition, and any distinctive marks or decorations. Include details such as the presence of baskets, reflectors, or any branding.”
- **Car:** “Please describe the car by its color, make, model, condition, license plate (if visible), and any distinguishing features such as stickers, dents, or modifications.”
- **Airplane:** “Please describe the airplane by its type (commercial, private, etc.), airline brand, color scheme, size, and any visible markings such as logos or tail numbers.”
- **Bus:** “Please describe the bus by its color, type (public, school, etc.), condition, any branding or advertising on its surface, and the route number or destination if visible.”
- **Train:** “Please describe the train by its type (freight, passenger, high-speed, etc.), color, length, condition, and any visible logos or car numbers.”
- **Truck:** “Please describe the truck by its type (pickup, semi, etc.), color, make, model, any visible logos or branding, and details such as cargo or modifications.”
- **Boat:** “Please describe the boat by its type (sailboat, motorboat, yacht, etc.), size, color, condition, and any identifying features like registration numbers or flags.”
- **Traffic Light:** “Please mention the current state of the traffic light (red, yellow, green), its location, and any additional details like the presence of pedestrian signals.”
- **Fire Hydrant:** “Please describe the fire hydrant by its color, condition, and any notable features such as signs, markings, or proximity to other objects.”
- **Stop Sign:** “Please describe the stop sign’s condition, location, and any visible obstructions or markings on it.”
- **Parking Meter:** “Please describe the parking meter by its condition, type (modern, traditional), and any visible

- information like pricing or operational status.”
- **Bench:** “Please describe the bench by its material, color, condition, and any distinctive features such as inscriptions, decorations, or nearby objects.”
  - **Bird:** “Please describe the bird by its species (if identifiable), color, size, behavior, and any unique markings or features.”
  - **Cat:** “Please describe the cat by its color, breed (if identifiable), size, behavior, and any distinguishing features such as collars or patterns.”
  - **Dog:** “Please describe the dog by its breed (if identifiable), color, size, behavior, and any accessories such as collars or leashes.”
  - **Horse:** “Please describe the horse by its color, breed (if identifiable), size, behavior, and any accessories such as saddles or reins.”
  - **Sheep:** “Please describe the sheep by its color, size, behavior, and any distinguishing features such as markings or tags.”
  - **Cow:** “Please describe the cow by its color, breed (if identifiable), size, behavior, and any distinguishing features such as tags or markings.”
  - **Elephant:** “Please describe the elephant by its size, tusk length, condition, and any unique features such as markings or behavior.”

### 3. Prompt Design of Figure 3

**System prompt.** Referring to ShareGPT4Video [3], we divided the System prompt into three parts. Through extensive tests on challenging samples, including multi-instance, complex scenes, and high-intensity motion, we finalized the system prompt shown in Figure S5. Additionally, temporal metadata extracted using the code provided in Figure S6.

**Prompts of global description, background detail and camera movement.** The global description is derived from a single prompt: “*Please describe this video in one sentence, no more than 20 words.*”. To illustrate the acquisition of camera motion and background details, we provide an example of implementing camera hints with movement cues in Figure S7. A similar approach is used for extracting background details included in our code released later.

**Prompts of structured caption.** In the structured caption section, we use Actions and Motion as examples, with the CoT prompt shown in Figure S8. The acquisition of Appearance and the injection of Human-designed class hints follow a similar approach.

### 4. Design of InstanceEnhancer

In InstanceEnhancer, prompt alignment during inference is achieved through a two-stage process (Figure S2). To pro-

vide more precise instructions to LLMs, we meticulously designed multiple examples as part of the CoT, which are fed into the LLMs. An example of this is shown in Figure S9.

## 5. Evaluation metrics for video reconstruction

**3DVAE score (3DVAE<sub>score</sub>).** The LPIPS score [33] which is widely used to evaluate image reconstruction quality, measures perceptual distance between ground truth (GT) and reconstructed images. We extend this concept for video data by using 3DVAE [30] to extract latent-space video representations from both GT videos and their caption-reconstructed versions. 3DVAE<sub>score</sub> computes the distance between latent representations across spatial and temporal dimensions:

$$d(\mathbf{z}_{\text{GT}}, \mathbf{z}_{\text{rec}}) = \sum_l \sum_t \sum_{h,w} \|w_l \odot (\mathbf{z}_{\text{GT},hwt}^l - \mathbf{z}_{\text{rec},hwt}^l)\|_2^2 \quad (1)$$

where  $\mathbf{z}_{\text{GT},hwt}^l$  and  $\mathbf{z}_{\text{rec},hwt}^l$  represent the latent representations at layer  $l$ , spatial location  $(h,w)$ , and temporal frame  $t$ , with  $w_l$  as the layer-specific weight matrix. We set  $(h,w,t) = (224, 224, 8)$  for evaluation.

To ensure consistency, we use the same video generation model across all captioning methods. Following LIP-IPS methodology, we validate the 3DVAE score by comparing GT videos against various distorted versions. As shown in Tab. S1, the results demonstrate that our score effectively captures perceptual similarities between GT and reconstructed videos.

Distortion type	3DVAE score↓	Setting
Blurring	7.71	GaussianBlur(kernel=(5, 5), sigma=0)
Compression artifacts	11.19	JPEG compression (quality 5-30)
Corruptions	39.80	Random pixel masking (binary mask)
Random noise	49.70	Gaussian noise (mean=0, stddev=25)
Brightness distortion	63.25	Scaling (factor 0.5-1.5)
Spatial shifts	78.94	Random affine shifts ( $\pm 10$ pixels)
T2V models Avg.	134 ~ 145	-
Broken video	149.50	-

Table S1. 3DVAE scores for various distortions and video models, showcasing its effectiveness in capturing perceptual similarities and reconstruction accuracy. The setting column provides details of the experimental setup for each distortion type.

**CLIP score sentence by sentence (CLIP<sub>SenbySen</sub>).** While CLIP [19] is widely used for text-video similarity computation [6, 14], its 77-token limit restricts processing of long texts. To overcome this, we propose CLIP score sentence by sentence (SenbySen), which segments texts into individual sentences and computes CLIP similarity between each sentence and video frame.

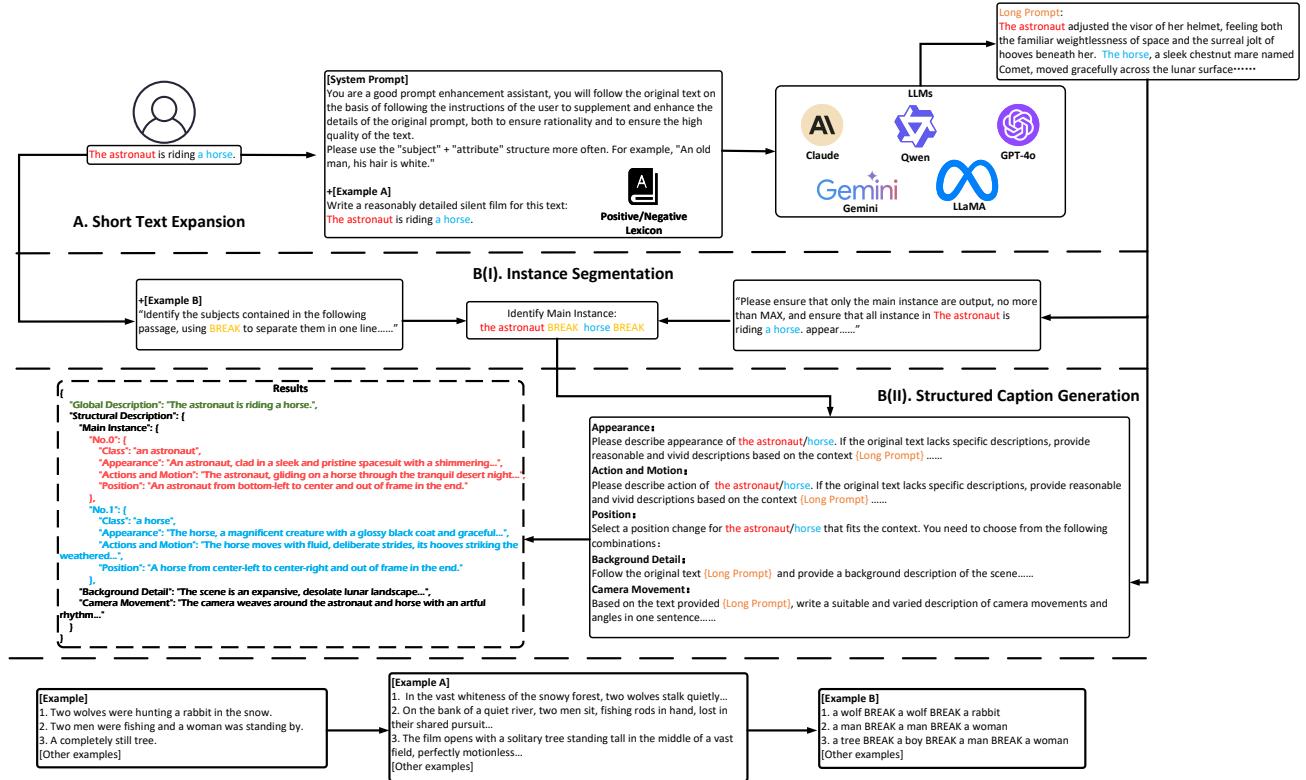


Figure S2. Detailed overview of the InstanceEnhancer pipeline. Example No.1 as shown in Figure S9.

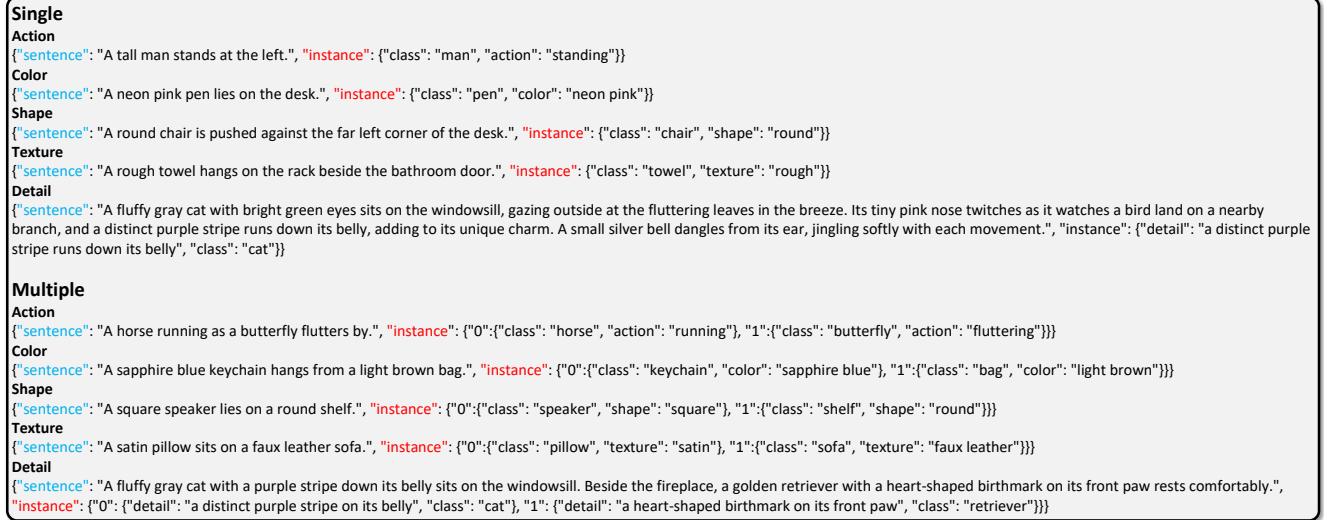


Figure S3. Inference examples of Inseval.

Let  $S = \{s_1, s_2, \dots, s_n\}$  be the sentences from input text and  $V = \{v_1, v_2, \dots, v_t\}$  be the video frames. For a sentence  $s_i$  and frame  $v_j$ , we denote their CLIP similarity as  $\text{CLIP}(s_i, v_j)$ . The comprehensive score is computed as:

$$\text{OverallScore} = \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{t} \sum_{j=1}^t \text{CLIP}(s_i, v_j) \right) \quad (2)$$

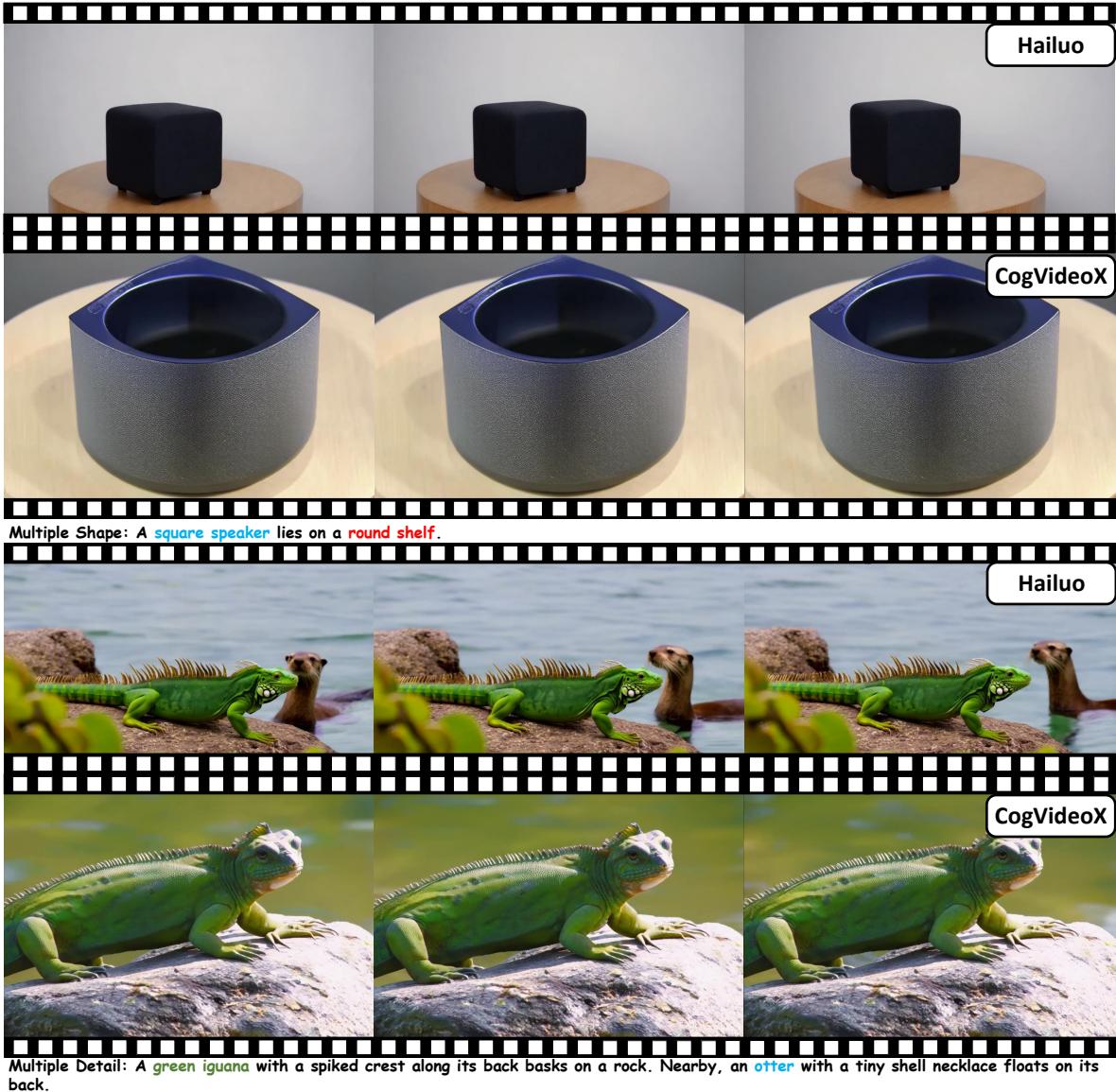


Figure S4. Visualization comparing open-source models and commercial models on prompts with poorer performance.

This approach not only addresses the token limitation but also enhances assessment quality by naturally assigning lower weights to non-specific textual descriptions.

**Human evaluation.** Automated machine-based scoring systems, while offering enhanced objectivity and efficiency, often fail to align with human preferences or fully grasp the nuances of context and meaning in a given task. To ensure a comprehensive and balanced evaluation, we adopted a human-based assessment framework. This evaluation is carried out across several key dimensions, including:

- 1) Instance Detail (**ID**): Evaluate whether the text provides accurate descriptions of the details of the examples in the video.

- 2) Intrinsic Hallucination: Evaluate whether the text hallucinates descriptions of things present in the video.
- 3) Extrinsic Hallucination: Evaluate whether the text introduces content that is not present in the video. For convenience, the latter two have been combined into a single metric called the Hallucination Scores (**HS**) [7]. The specific guidelines and scoring criteria for each metric refers to Table S2.

## 6. Inseval

**Inference prompts of Inseval.** In implementing Inseval, we designed multiple prompts to test each dimension, as illustrated in Figure S3. To further evaluate the model’s generative capabilities and instruction-following accuracy,

Instance Detail		Hallucination Scores
<b>1</b>	Descriptions are extremely vague, imprecise, or largely inaccurate. Almost no specific details from the video are captured correctly.	<b>1</b> Severe hallucination - Describes many non-existent details, significantly misrepresents what is shown, or introduces extensive irrelevant content with many unrelated topics or external information.
<b>2</b>	Descriptions have major inaccuracies or omit many important details. Only a few basic elements are described correctly.	<b>2</b> Frequent hallucination - Multiple instances of fabricated or misrepresented details and significant extra content introducing information beyond the video scope.
<b>3</b>	Descriptions are moderately accurate but lack precision in some areas. Core details are present but some secondary details are missing or incorrect.	<b>3</b> Occasional hallucination - A few minor instances of fabricated details, misrepresentations, or the addition of extra content not covered in the video.
<b>4</b>	Descriptions are largely accurate and detailed. Most key elements and nuances from the video are captured correctly, with only minor omissions or imprecisions.	<b>4</b> Minimal hallucination - One or two very minor discrepancies or limited introduction of external information.
<b>5</b>	Descriptions are highly precise and comprehensive. All important details from the video are captured accurately, including subtle elements and specific examples.	<b>5</b> No hallucination - All described details accurately reflect what is shown in the video, with no external content added.

Table S2. This table outlines scoring criteria for Instance Detail and Hallucination Scores, integrating intrinsic and extrinsic hallucinations into a unified framework for evaluation.

we deliberately included some “counter-intuitive” shapes in the prompt design.

**Evaluation prompts of Inseval.** For the evaluation, we used a general CoT Q-A pair format (with a slightly different design for the ‘Detail’ dimension, shown in Figure S10 to assess whether the MLLMs successfully matched the generated videos to the corresponding dimensions, as outlined in the specific code. In single-object scenarios, the success rate is calculated as the percentage of correctly matched prompts. In multi-object scenarios, the generation is deemed successful only if all targets meet the requirements. For reproducibility, fixed random seeds are used during generation and evaluation.

In Table 2, the ‘Shape’ and ‘Detail’ dimensions under Multiple category are omitted due to consistently very poor performance across all tested models. Even CogVideoX-5B, the overall best performer, struggles with multi-object tasks in these dimensions, as shown in Figure S4. Two primary error types are observed in Multiple Shape tasks: attribute confusion (**Top** case) and failure to follow multiple target instructions (**Bottom** case), where targets are either missing or rendered incorrectly. Commercial models demonstrate relatively better performance, which we further analyze in Section 7.

## 7. Analysis on Commercial Products vs. Open-source Models

**Prompt processing analysis.** Commercial T2V products excel at processing complex input prompts, effectively handling long-form text in structured formats while preserving semantic coherence. They can seamlessly interpret detailed scene descriptions, character interactions, and sequential events within a single prompt, producing coherent visual narratives, have shown surprising results in many situations.

Open-source T2V models, however, are *unable to directly process long-text structured prompts*, requiring an additional alignment step (Figure S11). This preprocessing can lead to potential information loss and inconsistencies in the final output, restricting the ability to capture nuanced details from the original prompt.

**Information retention capabilities.** Different models exhibit notable differences in information retention (Figure S4). Commercial products (*e.g.*, Hailuo AI) excel in maintaining fidelity between text and visual content, effectively preserving detailed instructions and translating multiple attributes into video sequences. This strength is particularly apparent when our caption contains *complex scenes* that demand temporal consistency and fine-grained details.

Open-source models face challenges in consistently representing instance information (Figure S4), exhibiting variability in detail preservation and limited capability with complex attribute combinations. These shortcomings are particularly evident when processing prompts with multiple interrelated instances or maintaining consistent visual characteristics across temporal sequences.

## System Prompt

### # Character

You are an excellent video frame analyst. Utilizing your incredible attention to detail, you provide clear, sequential descriptions for video frames. You are good at identifying and describing the properties of each target in the video frame, the actions and movement.

### # Skills

#### ## Skill 1: Describing Objects Appearances

- Describe the appearances of instance.
- Determine which parts are colored parts, as the goal of the main description.
- Focus mainly on the color part, the black and white part only as an auxiliary role.
- Highly sensitive to person, describe them in detail, such as the style and color of hat, the style and color of clothes, age, gender, body type, expression, etc.

#### ## Skill 2: Describing Objects' Actions and Behaviors

- Elaborate the action of instance.
- Notice and describe changes in the actions or behaviors.
- Determine which Objects are main instance and give more detailed description.

#### ## Skill 3: Use Fine Words to Describe.

- Select the appropriate ones of following words in your description: kaleidoscopic, delicate, grand, gentle, soothing, cool, mature, solitary, worn, chaotic, dramatic, cozy, shimmering, desolate, serene, weathered, whispering, loose-fitting, vibrant, tranquil, dimly-lit, purplish, introspective, artfully, sleek, energetic, overcast, brilliant, slender, graceful, picturesque, whimsical, contented, gentle, warm, tender, pastel-colored, elegant.
- State facts objectively without using any rhetorical devices such as metaphors or personification.
- Do not use any of the following negative words when describing: dull, rough, harsh, chaotic, cluttered, bleak, uninspired, garish, stiff, unrefined, artificial, heavy, disorderly, grim, rusty, faded, cramped, jarring, obtrusive, awkward, ordinary, harsh, gloomy, cold, rigid, overcrowded, mismatched, messy, uneven, tacky, lifeless, unbalanced, heavy-handed, overbearing, dissonant, grating, oversaturated, unpleasant, rigid, blur.

### # Constraints

- State facts objectively without using any rhetorical devices such as metaphors or personification.
- Exclude sounds-related aspects, given the unavailability of audio signals.
- Descriptions should be fluent and precise, avoiding analyzing and waxing lyrical.
- Descriptions need to be concise, describing only the information that can be determined, without analysis or speculation.
- Do not mention the frame number and timestamp of the current frame.
- The main object will occupy most of the content of the picture, and there may be more than one main object, and there may be no main object in the landscape type of video.
- Only the main object needs to be described in detail, and the other objects only need to be described briefly.
- Strictly follow the format of the structured output, containing all of its elements.

Figure S5. System prompt of InstanceCap.

```

def load_video(video_path, max_frames_num, fps=1, force_sample=False):
    if max_frames_num == 0:
        return np.zeros((1, 336, 336, 3))
    vr = VideoReader(video_path, ctx=cpu(0), num_threads=1)
    total_frame_num = len(vr)
    video_time = total_frame_num / vr.get_avg_fps()
    fps = round(vr.get_avg_fps() / fps)
    frame_idx = [i for i in range(0, len(vr), fps)]
    frame_time = [i / fps for i in frame_idx]
    if len(frame_idx) > max_frames_num or force_sample:
        sample_fps = max_frames_num
        uniform_sampled_frames = np.linspace(0, total_frame_num - 1, sample_fps, dtype=int)
        frame_idx = [uniform_sampled_frames.tolist()]
        frame_time = [i / vr.get_avg_fps() for i in frame_idx]
    frame_time = [f'{i:.2f}' for i in frame_time]
    spare_frames = vr.get_batch(frame_idx, asnumpy())
    return spare_frames, frame_time, video_time

def build_video_prompt(image_processor, video, frame_time, video_time):
    video = image_processor.preprocess(video, return_tensors="pt")["pixel_values"].cuda().bfloat16
    video = [video]
    time_instruction =
        f"The video lasts for {video_time:.2f} seconds, and {[len(video[0])]} frames are uniformly sampled from it. "
        f"These frames are located at {frame_time}. Please answer the following questions related to this video."
    pre_instruction = DEFAULT_IMAGE_TOKEN + f"[time_{_instructon}]"
    return video, pre_instruction

```

Figure S6. Code of getting video temporal metadata.

#### Example Prompt of Stage A&B(I)

##### Stage A. Short Text Expansion

# USER

Write a reasonably silent film for this text: Two wolves were hunting a rabbit in the snow. (Short prompt)

# ASSISTANT (human-provided)

In the vast whiteness of the snowy forest, two wolves stalk quietly. One is large and heavyset, its fur a mix of dark grays and black, blending into the shadows of the trees. The other is leaner and quicker, with a pale, almost silver coat that glimmers in the winter light. Their eyes are sharp and focused, and each step is measured—one wolf pacing powerfully, the other darting with agile precision. Ahead of them, a small white rabbit crouches in the snow, nearly invisible but for the black tips of its ears and nose. Suddenly, the rabbit bolts, kicking up a spray of powder. The dark wolf charges forward with brute force, while the silver one sprints with smooth, loping strides. The rabbit's movements are frantic and unpredictable, weaving between tree roots and leaping over snowdrifts. The wolves flank it from both sides, their breaths clouding the air, closing in like shadows. At the last moment, the rabbit dives into a narrow burrow, disappearing into safety. The dark wolf sniffs the ground, frustrated, while the silver wolf paces briefly before the two melt back into the forest, leaving only their crisscrossing prints behind in the snow. (Long prompt)

##### Stage B(I). Instance Segmentation

# USER

Identify the subjects contained in the following passage, using BREAK to separate them in one line:

Long prompt

Please distinguish between instances and scenes, and return only instances, such as: "a man BREAK two cat BREAK a cup..." Please ensure that only the main instance are output, no more than X, and ensure that all objects in Short prompt appear.

Note that the instance you extract must be an entity that can be touched.

When there are multiple targets, you need to give them separately.

# ASSISTANT (human-provided)

a wolf BREAK a wolf BREAK a rabbit

Figure S9. Designed example for LLMs.

**Camera Movement Prompt**

# User

Let's think step by step... Try to separate the camera movement from the video.

If camera\_motion == "Undetermined":  
The motion of the video camera is very complex, can you infer the possible motion of the camera and the shooting Angle (long distance/medium distance/overhead Angle/POV, etc.) from the changes in the video?  
elif camera\_motion == "static":  
Is the camera static or moving? Can you deduce the possible motion of the camera and the shooting Angle (long distance/medium distance/overhead Angle/POV, etc.) in the video?  
else:  
The motion of the camera in this video is (camera\_motion). According to my tips, can you deduce the possible motion of the camera and the shooting Angle (long distance/medium distance/overhead Angle/POV, etc.) in the video?  
Summarize the camera movement and shooting angle, use degree adverbs appropriately(Sharply, rapidly, slowly, etc), try to give description in detail.

Figure S7. Prompt of camera movement.

#### Evaluation Prompt(Single)

##### STEP 1

# USER

Describe this video in one sentence, care about the (dimension\_type).

# ASSISTANT

Answer of STEP 1

##### STEP 2

# USER

Tell me if {instance\_class} is in the video? Note the synonyms of '{instance\_class}'. Your answer can only be YES or NO. Do not output any answer that is not YES or NO.

# ASSISTANT

Answer of STEP 2(if answer is "No", return False)

##### STEP 3

# USER

Based on your previous answer, tell me what is the '(dimension\_type)' of '{instance\_class}' in the video? Be careful to ignore camera movement.

# ASSISTANT

Answer of STEP 3

##### STEP 4 of Others

# USER

Do you think the '(dimension\_type)' of '{instance\_class}' in the video Approximately close to to '{instance\_specific}'?

Your answer can only be YES or NO. Do not output any answer that is not YES or NO.

# ASSISTANT

Answer of STEP 4(Final answer)

##### STEP 4 of Detail

# USER

Is the '{instance\_specific}' of '{instance\_class}' partly reflected in the video? Your answer can only be YES or NO. Do not output any answer that is not YES or NO.

# ASSISTANT

Answer of STEP 4(Final answer)

**Actions and Motion**

STEP 1  
# User

Let's think step by step...

Can you tell what color the background of this video is? Unimportant parts have been intentionally obscured with a mosaic; please focus only on the clearly visible portions.

If I remind you that there's a {target\_object}, can you see it in the video?

This is supplementary information to the full video to aid your description:  
(global\_description)

Can you find the {target\_object} corresponding to this paragraph in the video?

# ASSISTANT

Answer of STEP 1

STEP 2  
# User

Continue ignore mosaic background, if {target\_object} has any movement, answer what you think {target\_object} is doing in video.

Extract the information related to {target\_object} that you are currently describing.

Do not answer any information unrelated to {target\_object}'s state of motion or you're not sure. Observe carefully, use appropriate adjectives, and give description in detail.

Do not answer in bullet points. Do not mention any objects that are not related to {target\_object}.

Do not speculate about the environment in which the {target\_object} is located.

Do not mention anything about a "blurred background." Use a continuous paragraph.

# ASSISTANT

Answer of STEP 2(Final answer)

Figure S8. Prompt of actions and motion.

Figure S10. Evaluation prompts of Inseval.

## Aligning Prompt

### Step 1

#### # User

Let's think step by step...

Read the following JSON and summarize it to continuous text paragraph, ensuring that all main ideas and crucial details are preserved:

{InstanceCap}

What you need to pay attention to is the "Global Description" and "Structural Description" sections. The "Global Description" provides an overall summary of the video, while the "Structural Description" contains detailed information about various aspects of the video, such as main characters, background details, and camera movements. Please focus on the details in the "Structural Description" and combine them with the "Global Description" to create a summary. Select the appropriate ones of following words in your description: kaleidoscopic, delicate, grand, gentle, soothing, cool, mature, solitary, worn, chaotic, dramatic, cozy, shimmering, desolate, serene, weathered, whispering, loose-fitting, vibrant, tranquil, dimly-lit, purplish, introspective, artfully, sleek, energetic, overcast, brilliant, slender, graceful, picturesque, whimsical, contented, gentle, warm, tender, pastel-colored, elegant.

Do not use any of the following negative words when describing: dull, rough, harsh, chaotic, cluttered, bleak, uninspired, garish, stiff, unrefined, artificial, heavy, disorderly, grim, rusty, faded, cramped, jarring, obtrusive, awkward, ordinary, harsh, gloomy, cold, rigid, overcrowded, mismatched, messy, uneven, tacky, lifeless, unbalanced, heavy-handed, overbearing, dissonant, grating, oversaturated, unpleasant, rigid, blur.

### Step 2

#### # User

1. Please use the "subject" + "attribute" + "position" structure more often. For example, "An old man, his hair is white."
2. Please tell me the content of the video directly, don't use "The video shows..." Or other similar forms, you should begin with a direct description of the content, for example:"An old man..."
3. If there are multiple objects, such as people, introduce them with phrases like "A man... and a woman... , and a man...","A car... , and a car...".
4. When you need to describe The background in detail, use "The scene is..." As an opening sentence.
5. Summarize it to approximately 180 words

Figure S11. Aligning prompt used during alignment with the open source model.