

---

# UVE: Are MLLMs Unified Evaluators for AI-Generated Videos?

---

Yuanxin Liu<sup>§</sup>   Rui Zhu<sup>‡</sup>   Shuhuai Ren<sup>§</sup>   Jiacong Wang<sup>¶</sup>  
Haoyuan Guo<sup>‡</sup>   Xu Sun<sup>§</sup>   Lu Jiang<sup>‡</sup>

<sup>§</sup> State Key Laboratory of Multimedia Information Processing,  
School of Computer Science, Peking University

<sup>‡</sup> ByteDance Seed

<sup>¶</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences  
liuyuanxin@stu.pku.edu.cn  
guohaoyuan@bytedance.com   xusun@pku.edu.cn

## Abstract

With the rapid growth of video generative models (VGMs), it is essential to develop reliable and comprehensive automatic metrics for AI-generated videos (AIGVs). Existing methods either use off-the-shelf models optimized for other tasks or rely on human assessment data to train specialized evaluators. These approaches are constrained to specific evaluation aspects and are difficult to scale with the increasing demands for finer-grained and more comprehensive evaluations. To address this issue, this work investigates the feasibility of using multimodal large language models (MLLMs) as a unified evaluator for AIGVs, leveraging their strong visual perception and language understanding capabilities. To evaluate the performance of automatic metrics in unified AIGV evaluation, we introduce a benchmark called UVE-Bench. UVE-Bench collects videos generated by state-of-the-art VGMs and provides pairwise human preference annotations across 15 evaluation aspects. Using UVE-Bench, we extensively evaluate 18 MLLMs. Our empirical results suggest that while advanced MLLMs (e.g., Qwen2VL-72B and InternVL2.5-78B) still lag behind human evaluators, they demonstrate promising ability in unified AIGV evaluation, significantly surpassing existing specialized evaluation methods. Additionally, we conduct an in-depth analysis of key design choices that impact the performance of MLLM-driven evaluators, offering valuable insights for future research on AIGV evaluation. 📄 **Code:** <https://github.com/bytedance/UVE>, 🗃️ **Data:** <https://huggingface.co/datasets/lyx97/UVE-Bench>.

## 1 Introduction

Video generative models (VGMs) have rapidly evolved in recent years, greatly aiding visual content creation in numerous fields. However, even current state-of-the-art (SOTA) VGMs [3, 46, 25, 27] still suffer from issues like incorrect subject structure, unnatural motion, and imperfect alignment with the text prompt. Consequently, it is crucial to develop automatic metrics that can effectively identify these imperfections in modern AI-generated videos (AIGVs), so as to facilitate the advancement of VGMs and enhance the application of AIGVs in real-world scenarios.

Existing AIGV evaluation metrics fall into two categories. The first employs off-the-shelf models (e.g., CLIP [47] and DINO [5]) and designs heuristic rules to assess various specific aspects of AIGVs [21, 40]. The second category involves collecting human assessments for specific aspects of interests and training models to imitate these assessments [17, 2, 75, 61, 56]. Such specifically trained automatic evaluators have shown better correlation with human evaluations compared to the

off-the-shelf models. However, with the rapid development of VGMs and AIGV applications, there is an urgent need for finer-grained and more comprehensive evaluations. Continuously collecting human assessments and training models to accommodate emerging evaluation aspects and frequently changing standards is cost-intensive and difficult to scale.

Unlike existing automatic metrics, we humans can assess any aspect of AIGVs as long as a proper guideline is provided. This ability stems from our robust visual perception and language understanding. Leveraging the knowledge learned from vast amounts of visual and language data, the latest multimodal large language models (MLLMs) [58, 44, 9] also exhibit strong ability in joint vision-language understanding. This progress naturally raises a question: **“Can MLLMs be utilized as a unified AIGV evaluator like humans?”**

To address this question, we propose an approach to unify AIGV evaluations by prompting pre-trained MLLMs and mapping their outputs to evaluation results (§2). This method enables zero-shot evaluation of any aspect of AIGV by simply modifying the prompts. It supports both single video ratings and video pair comparisons. While using MLLMs for AIGV evaluation is a straightforward concept and has been explored in previous works [17, 2, 56], these studies are limited by evaluating only a few aspects and relying on human-annotated ratings to train the MLLMs.

To evaluate the capability of MLLMs as unified AIGV evaluators, we require a benchmark that (1) encompasses a broad range of AIGV aspects, (2) provides accurate human evaluations as references, and (3) includes AIGVs that highlight the weaknesses of state-of-the-art VGMs. Since no existing AIGV dataset meets all these criteria, we introduce a new benchmark called **UVE-Bench** (short for **Unified Video Evaluation**; see §3). UVE-Bench has three key features: **First**, it covers a wide range of 15 evaluation aspects. **Second**, it provides human annotation in the form of pairwise video preference, avoiding the inconsistent standards of absolute ratings between humans [62], and can be used to evaluate both single video ratings and video pair comparisons. **Third**, the videos in UVE-Bench are generated by the latest VGMs (e.g., MovieGenVideo [46] and HunyuanVideo [25]), challenging the evaluators to identify the weaknesses of such videos.

Based on UVE-Bench, we conduct an extensive evaluation of 18 MLLMs. Our results indicate that unified evaluator powered by advanced MLLMs, such as Qwen2-VL-72B [58] and InternVL2.5-78B [9], indeed show promising abilities in evaluating various AIGV aspects, outperforming existing approaches that focus on specific aspects. However, there remains a significant gap between these MLLMs and human evaluators, particularly in aspects requiring a fine-grained understanding of temporal dynamics in videos. Additionally, we perform a series of analytical studies on the design choices that impact the performance of our MLLM-driven AIGV evaluation framework, providing valuable insights for future research in the field.

The contributions of this work are: (1) We introduce a unified approach to evaluate any aspect of AIGV using pre-trained MLLMs. (2) We propose UVE-Bench, a comprehensive benchmark to assess the capability of unified AIGV evaluation. (3) We conduct in-depth analysis on the pros and cons of MLLMs in unified AIGV evaluation and the key design choices that impact their performance.

## 2 Unifying AIGV Evaluation with MLLMs

As shown in Fig. 1, the unified evaluator is designed to handle all aspects of AIGV evaluation, eliminating the need for specialized models tailored to individual aspects. In this section, we describe how this framework can be applied to both single video rating and video pair comparison.

### 2.1 Problem Formulation

**Single Video Rating.** Given a generated video  $\mathcal{V}$ , a textual evaluation guideline  $\mathcal{G}_a$  and the text-to-video (T2V) prompt  $\mathcal{T}$  used to generate the video, the objective of single video rating is to predict a numerical score  $\mathcal{S}$  measuring the quality of  $\mathcal{V}$  based on the aspect  $a$  specified by  $\mathcal{G}_a$ .

**Video Pair Comparison.** Given a pair of generated videos  $\mathcal{V}_1, \mathcal{V}_2$ , the corresponding T2V prompts  $\mathcal{T}_1, \mathcal{T}_2$  used to generate them, and a textual evaluation guideline  $\mathcal{G}_a$ , the objective of video pair comparison is to make a choice  $\mathcal{C}$  from one of the four options:  $\mathbf{O} = \{\text{“}\mathcal{V}_1\text{ better”}, \text{“}\mathcal{V}_2\text{ better”}, \text{“same good”}, \text{“same bad”}\}$ , according to the aspect  $a$  described by  $\mathcal{G}_a$ .

Table 1: MLLM prompting templates for unified AIGV evaluation.

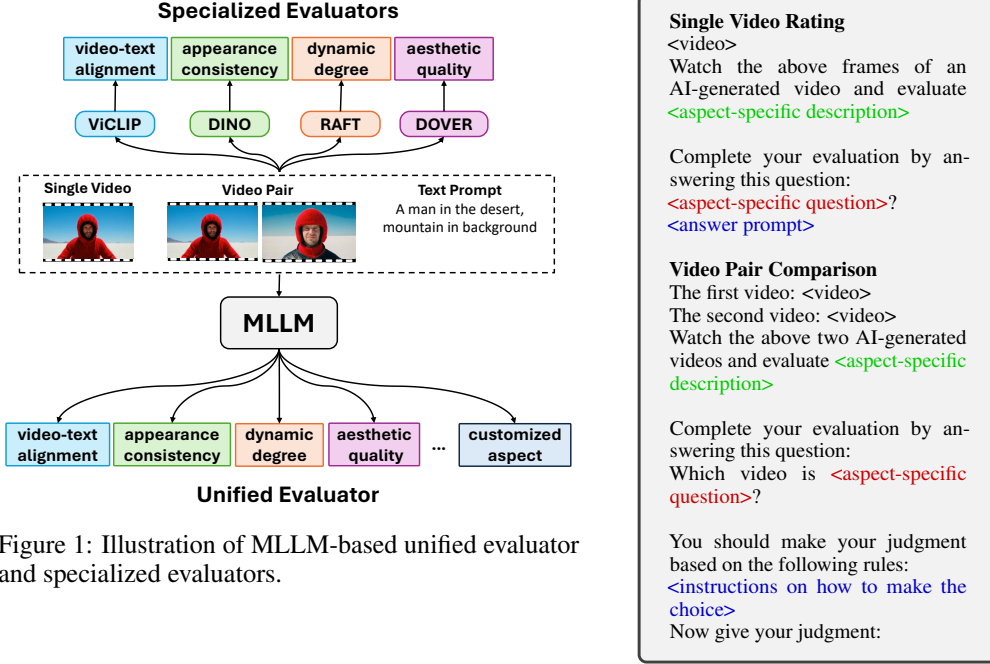


Figure 1: Illustration of MLLM-based unified evaluator and specialized evaluators.

## 2.2 Method

**Single Video Rating.** Research on image quality evaluation [76, 35] has demonstrated that, given an appropriate prompt, the generative likelihood of certain tokens (e.g., *yes/no* or *good/poor*) by MLLMs can be used as an effective indicator of image visual quality and image-text alignment. Motivated by this, we formulate single video rating in a unified manner as:

$$\mathcal{S} = \frac{P_{\theta}(t_{\text{pos}}|\mathcal{V}, \mathcal{T}, \mathcal{G}_a)}{P_{\theta}(t_{\text{pos}}|\mathcal{V}, \mathcal{T}, \mathcal{G}_a) + P_{\theta}(t_{\text{neg}}|\mathcal{V}, \mathcal{T}, \mathcal{G}_a)} \quad (1)$$

where  $t_{\text{pos}}$  and  $t_{\text{neg}}$  are predefined positive/negative scoring tokens.  $P_{\theta}(t)$  is the probability of generating token  $t$  using MLLM  $\theta$ .  $\mathcal{G}_a$  ends with a question related to aspect  $a$  and an answer prompt encouraging the model to generate the scoring tokens (e.g., *Please directly answer yes or no:*).

**Video Pair Comparison.** A straightforward way to perform video pair comparison is directly feeding the video pair into MLLMs, as the latest MLLMs [58, 29, 9] already support interleaved vision-language input with multiple videos. Specifically, we feed the video pair, along with the corresponding T2V prompts if needed, into the MLLM and prompt it to make a choice from  $\mathbf{O}$ :

$$\mathcal{C} = f_{\theta}(\mathcal{V}_1, \mathcal{V}_2, \mathcal{T}_1, \mathcal{T}_2, \mathcal{G}_a) \in \mathbf{O} \quad (2)$$

Tab. 1 illustrates the prompting templates for single video rating and video pair comparison, which can be adapted to any evaluation aspect by modifying the aspect-specific prompt. The specific prompts being used for different aspects are provided in the supplementary material.

## 3 UVE-Bench

As illustrated in Fig. 2, UVE-Bench is designed to assess AIGV automatic evaluation methods. This section details the process of video collection (§3.1), the evaluation aspects covered by UVE-Bench (§3.2), the criterion for assessing the performance of automatic AIGV evaluators using human preferences (§3.3), the data annotation procedure (§3.4) and a comparison with existing AIGV datasets (§3.5).

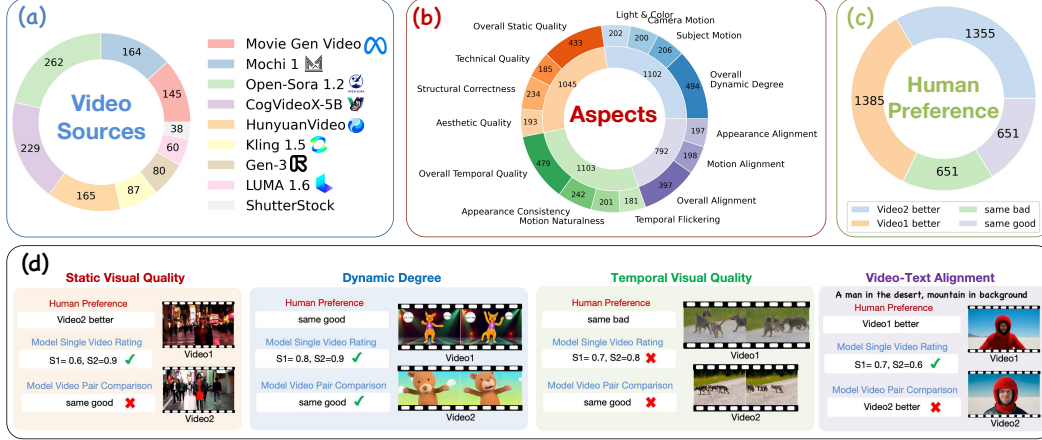


Figure 2: Overview of UVE-Bench. (a) The distribution of video sources. (b) The distribution of data example over 15 fine-grained AIGV evaluation aspects. (c) The distribution of human preference over the four categories. (d) Data examples illustrating how to evaluate both single video rating and video pair comparison using the human preference annotations. More examples can be found in Appendix D.6.

Table 2: Information of the videos in UVE-Bench.

Video Source	Prompt Source	Resolution	Duration	FPS	Release Date	Open Source
Gen-3 [50]	VideoGen-Eval	1280×768	5.3	24	2024.06	✗
Kling 1.5 [27]	VideoGen-Eval	1920×1080	5.1	30	2024.09	✗
LUMA 1.6 [42]	VideoGen-Eval	1360×752	5.0	24	2024.09	✗
Movie Gen Video [46]	Movie Gen	848×480	5.0	24	2024.10	✗
Open-Sora 1.2 [79]	VideoGen-Eval	1280×720	4.3	24	2024.06	✓
	Movie Gen	848×480	5.3			
	Shutterstock					
CogVideoX-5B [66]	VideoGen-Eval	720×480	6.1	8	2024.08	✓
	Movie Gen					
	Shutterstock					
HunyuanVideo [25]	Movie Gen	960×544	5.4	24	2024.12	✓
	Shutterstock					
Mochi 1 [54]	Movie Gen	848×480	5.4	30	2024.10	✓
	Shutterstock					
Shutterstock	Shutterstock	600×316 898×506 596×336	5.0	24, 25 30, 60	-	-

### 3.1 Video Collection

The videos in UVE-Bench are created using eight of the latest video generative models via text-to-video generation. To ensure diverse video content, we select text prompts from three sources: (1) VideoGen-Eval [70] provides over 700 prompts, targeting various application scenarios (e.g., animation and advertisement) and key capabilities (e.g., text alignment and motion diversity) of video generation. (2) Movie Gen Video Bench [46] offers 1,003 prompts covering various categories of video content (e.g., human activity, animals, and scenery). (3) We also modify a small portion of video captions from the Shutterstock<sup>1</sup> platform to generate videos with different degrees of *Light & Color* change (an evaluation aspect that will be discussed in §3.2).

For VideoGen-Eval, we directly collect videos generated by Gen-3, Kling 1.5, LUMA 1.6, Open-Sora 1.2, and CogVideoX-5B, as released by the authors of Zeng et al. [70]. For Movie Gen Video Bench and Shutterstock, we generate videos using Open-Sora 1.2, CogVideoX-5B, Mochi 1 and HunyuanVideo. For the Movie Gen Video model, we directly collect the generated videos released by Polyak et al. [46] and select the first five seconds of the videos. Additionally, we collect some non-AIGV videos related to *Light & Color* from Shutterstock. As shown in Tab. 2, the videos have a duration of around 5.0 seconds, with resolution ranging from  $596 \times 336$  (360p level) to  $1280 \times 720$  (720p level) and  $1920 \times 1080$  (1080p level).

<sup>1</sup><https://www.shutterstock.com/>



### 3.2 Evaluation Aspects

Inspired by previous work in AIGV evaluation [21, 39, 40], we identify four fundamental evaluation aspects to build our benchmark: *Static Quality*, *Temporal Quality*, *Dynamic Degree* and *Video-Text Alignment*, each of them is further divided into several subaspects.

**Static Quality.** This aspect evaluates the visual quality of individual video frames. It is broken down into four subaspects: (1) *Aesthetic Quality*, which assesses the aesthetic elements, including frame layout, lighting, and color harmony. (2) *Technical Quality*, which examines the presence of unwanted noise, blur, and distortion. (3) *Structural Correctness*, which checks for abnormal subject structures that contradict common sense (e.g., a person with three eyes). (4) *Overall Static Quality*, which jointly considers the above three subaspects.

**Temporal Quality.** This aspect assesses visual quality from a temporal perspective, focusing on four subaspects: (1) *Appearance Consistency*, which evaluates whether the appearance and identity of subjects or backgrounds remain consistent across frames. (2) *Temporal Flickering*, which identifies undesirable flickering and jittering that degrade visual quality. (3) *Motion Naturalness*, which determines if subject motions and interactions appear natural and adhere to physical laws. (4) *Overall Temporal Quality*, which comprehensively assesses the temporal visual quality by integrating these three subaspects.

**Dynamic Degree.** This aspect evaluates the degree of dynamic in the generated video. It includes four subaspects: (1) *Subject Motion*, which focuses on the motion degree of subjects. (2) *Camera Motion*, which assesses the motion degree of the camera. (3) *Light & Color*, which considers changes in lighting conditions and color. (4) *Overall Dynamic Degree*, which combines these three subaspects to provide a comprehensive measure of the video’s dynamic degree.

**Video-Text Alignment.** This aspect evaluates how well the generated video aligns with the given text prompt, divided into three subaspects: (1) *Appearance Alignment*, which focuses on the alignment of subject or scene appearance with the text. (2) *Motion Alignment*, which assesses the alignment of subject or camera motions with the text. (3) *Overall Alignment*, which considers how faithfully the video content reflects the entire text prompt.

### 3.3 Evaluation Criterion

As shown in Fig. 2, each example in UVE-Bench consists of three elements: a pair of videos  $\{\mathcal{V}_1, \mathcal{V}_2\}$ , the evaluation aspect  $a$  and corresponding human preference annotation  $\mathcal{P} \in \mathbf{O} = \{\text{“}\mathcal{V}_1\text{ better”}, \text{“}\mathcal{V}_2\text{ better”}, \text{“same good”}, \text{“same bad”}\}$ .

For single video rating, the automatic evaluator is required to predict numerical rating scores  $\mathcal{S}_1, \mathcal{S}_2 \in [0, 1]$  for each video, respectively. Then, we assess the correctness of rating according to the following criteria:

$$\mathcal{A}^{\text{single}} = \begin{cases} \mathbf{1}(\mathcal{S}_1 > \mathcal{S}_2) & \text{if } \mathcal{P} = \text{“}\mathcal{V}_1\text{ better”} \\ \mathbf{1}(\mathcal{S}_1 < \mathcal{S}_2) & \text{elif } \mathcal{P} = \text{“}\mathcal{V}_2\text{ better”} \\ f_c(\mathcal{S}_1|\beta) \cdot f_c(\mathcal{S}_2|\beta) & \text{elif } \mathcal{P} = \text{“same good”} \\ f'_c(\mathcal{S}_1|\alpha) \cdot f'_c(\mathcal{S}_2|\alpha) & \text{elif } \mathcal{P} = \text{“same bad”} \end{cases} \quad (3)$$

Here,  $\mathbf{1}(\cdot) \in \{0, 1\}$  is a binary indicator function and  $\alpha < \beta$  are the thresholds for “bad” and “good”, respectively.  $f_c(\mathcal{S}|\beta)$  is a piecewise function that equals to 1 when  $\mathcal{S} \in [\beta, 1]$ , indicating that the model evaluation aligns with the human judgment of “same good”. When  $\mathcal{S} \in [0, \beta)$ ,  $f_c(\mathcal{S}|\beta)$  exponentially decays from 1 to 0, indicating a gradual deviation from “same good”. Similarly,  $f'_c(\mathcal{S}|\alpha)$  is constructed in a reversed manner to assess whether  $\mathcal{S}$  agrees with the human judgment of “same bad”. Details of  $f_c(\mathcal{S}|\alpha)$  and  $f'_c(\mathcal{S}|\beta)$  can be found in Appendix C.2.

When it comes to video pair comparison, the automatic evaluator is asked to make a choice  $\mathcal{C}$  from  $\mathbf{O}$  in the same way as humans. Therefore, we directly adopt accuracy as the evaluation criteria:  $\mathcal{A}^{\text{pair}} = \mathbf{1}(\mathcal{C} = \mathcal{P})$ .

Table 3: Comparison between UVE-Bench and existing AIGV datasets. “Single” denotes single video rating. “Pair” denotes video pair comparison. “Understanding” denotes general video quality understanding via open-ended question answering.

Dataset	Aspects	Eval Task	Latest VGM
FETV [39]	9	Single	ZeroScope (2023.06)
VBench [21]	16	Pair	VideoCrafter1 (2023.10)
EvalCrafter [40]	4	Single	Gen-2 (2023.12)
T2VQA-DB [26]	1	Single	AnimateDiff (2023.12)
VideoFeedback [17]	5	Single	SORA (2024.02)
LGVQ [75]	3	Single	Gen-2 (2023.12)
TVGE [63]	2	Single	Gen-2 (2023.12)
GAIA [43]	3	Single	Mora (2024.03)
Q-Bench-Video [74]	4	Understanding	SORA (2024.02)
AIGV-Assessor [56]	4	Single&Pair	SORA (2024.02)
UVE-Bench (Ours)	15	Single&Pair	HunyuanVideo (2024.12)

### 3.4 Data Annotation and Quality Review

With the collected videos, the authors of this paper, who have rich research experience in VGMs and MLLMs and are proficient in English, annotate the evaluation aspects and pairwise video preferences. Specifically, videos generated by the same text prompt are presented in a pairwise manner to the annotators, who are asked to determine a preference choice from **O** and identify the aspect that influenced their decision. To minimize subjectivity, annotators are instructed to only annotate video pairs for which they could confidently make a preference judgment, and the remaining video pairs are discarded. Following this procedure, we obtained 4,042 pairwise preference annotations and retained 1,230 individual videos. The distributions of the videos and preference annotations are illustrated in Fig. 2.

To verify the quality of annotations, we randomly sample a subset of the dataset, consisting of 50 samples for each subaspect and assign a different group of three annotators to label pairwise video preferences. Including the original annotations, the inter-annotator agreement, as measured by Fleiss’ Kappa, is 0.803. This indicates a high level of agreement among the human annotators.

### 3.5 Comparison with Existing AIGV Datasets

Tab. 3 compares UVE-Bench with existing AIGV datasets across three dimensions. **First**, current datasets typically focus on a few basic aspects (such as video-text alignment, static and temporal visual quality), making them not suitable for assessing MLLMs as unified evaluators across various aspects. **Second**, existing datasets predominantly provide human annotations in the form of single-video rating, which not only precludes video pair comparison analysis but also introduces greater subjectivity compared to pairwise preference annotation [62, 76]. **Third**, the VGMs employed in existing datasets are relatively outdated, failing to represent the current SOTA performance of AIGVs. Identifying the weaknesses of such videos is both important and challenging for automatic AIGV evaluators. UVE-Bench addresses these limitations and provides a more comprehensive and accurate assessment for automatic AIGV evaluation models.

## 4 Experiments

### 4.1 Experimental Settings

**Evaluated Systems.** We conduct zero-shot evaluations on 18 MLLMs, including 15 open-sourced models along with three proprietary models: GPT-4o [45], Seed1.5-VL [15] and Gemini2.5-Flash [14]. Additionally, we assess five evaluation systems specialized in particular aspects: VBench [21], VideoScore [17], UMTScore [39], VIDEOCON-PHYSICS [2] and DOVER [60]. Details of these automatic evaluators are presented in Appendix C.1. To establish a human baseline, we engage three annotators to perform video pair comparisons on a subset of UVE-Bench, evaluating 50 samples for each subaspect.

Table 4: Performance of single video rating measured by  $\mathcal{A}^{\text{single}}$  in Eq. 3. The best and second-best results are highlighted with yellow and light yellow, respectively. Abbreviations: SM (subject motion), CM (camera motion), LC (light&color), TQ (technical quality), SC (structural correctness), AQ (aesthetic quality), AC (appearance consistency), MN (motion naturalness), TF (temporal flickering), MA (motion alignment), AA (appearance alignment).

Method	Model Size	Overall Dynamic	SM	CM	LC	Overall Static	TQ	SC	AQ	Overall Temporal	AC	MN	TF	Overall Alignment	MA	AA	AVG
Random	-	48.8	49.5	48.9	49.8	49.2	47.6	49.0	48.1	49.0	48.6	48.2	46.2	48.4	48.6	48.3	48.6
<b>Specialized Evaluators</b>																	
VideoScore-v1.1	8B	57.4	-	-	-	40.1	30.4	47.7	41.9	-	43.1	36.1	-	38.9	-	-	-
VBench	-	<b>87.8</b>	-	-	-	-	62.5	-	75.6	-	54.0	50.2	-	54.4	-	-	-
UMTScore	-	-	-	-	-	-	-	-	-	-	-	-	-	66.4	-	-	-
VIDEON-CON-PHYSICS	7B	-	-	-	-	-	-	-	-	-	-	53.4	-	68.7	-	-	-
DOVER	58M	-	-	-	-	-	69.2	-	80.3	-	-	-	-	-	-	-	-
<b>Unified Evaluators</b>																	
Video-LLaVA	7B	52.4	74.8	63.4	47.6	47.8	66.1	44.5	63.4	44.1	51.3	55.5	48.6	59.4	54.9	66.8	54.5
LongVA-DPO	7B	59.8	76.2	69.9	57.5	62.4	74.9	56.2	72.0	56.0	47.3	40.7	54.3	68.7	63.9	71.6	61.6
ShareGPT4Video	8B	77.5	81.0	77.1	82.5	59.4	68.7	54.1	69.4	54.4	48.1	42.6	60.9	62.7	54.3	70.3	63.9
VideoLLaMA2.1	7B	72.1	80.5	67.1	77.2	61.4	78.7	47.5	72.6	46.1	50.9	50.2	61.8	72.6	65.7	77.7	64.4
mPLUG-Owl3	7B	78.6	84.8	77.8	79.9	76.0	83.1	55.6	80.0	59.8	59.5	42.0	72.0	80.4	75.2	87.6	72.6
VideoChat2-Mistral	7B	83.1	92.2	89.6	74.9	68.1	76.2	53.8	74.1	58.7	58.3	52.8	85.6	75.6	78.0	80.9	72.6
MiniCPM-V-2.6	8B	81.4	86.3	80.3	88.8	70.9	75.0	52.9	80.6	61.1	59.4	51.7	70.9	82.1	74.3	90.8	73.4
LLaVA-OneVision	7B	81.0	87.6	83.2	84.9	70.7	78.2	50.6	83.1	62.9	60.4	41.5	85.9	79.3	66.9	86.7	73.0
LLaVA-OneVision	72B	82.3	87.8	78.4	88.2	71.3	77.6	60.2	81.5	64.6	61.9	39.9	86.8	84.4	71.7	93.5	75.0
LLaVA-Video	7B	80.4	85.5	80.9	81.5	66.2	74.2	49.5	75.7	58.3	58.2	39.3	82.3	80.5	69.9	90.0	71.0
LLaVA-Video	72B	82.8	86.1	82.9	86.9	70.2	80.0	55.4	77.7	60.1	59.2	40.4	83.5	84.8	73.7	94.6	74.0
Qwen2-VL	7B	84.6	89.7	<b>94.2</b>	79.7	64.6	67.3	50.7	70.6	51.1	51.3	48.0	62.7	85.4	78.7	92.2	70.9
Qwen2-VL	72B	86.5	<b>92.6</b>	92.7	86.0	70.6	76.9	60.2	83.4	52.5	58.0	48.9	71.0	<b>89.0</b>	<b>81.8</b>	95.0	75.4
InternVL-2.5-MPO	8B	81.3	86.1	80.4	88.0	68.1	77.9	53.6	77.5	60.9	54.9	50.9	72.5	80.6	73.2	90.5	72.6
InternVL-2.5-MPO	78B	84.3	86.6	82.4	<b>91.6</b>	72.8	<b>84.0</b>	<b>67.2</b>	82.3	61.5	65.2	<b>61.8</b>	<b>88.4</b>	87.4	79.3	<b>95.5</b>	78.2
GPT-4o	-	79.0	84.3	74.9	81.8	74.0	81.6	65.2	84.2	70.0	<b>79.8</b>	54.0	58.6	80.8	77.2	89.6	75.7
Seed1.5-VL	20B Act.	83.2	91.2	83.7	89.5	<b>82.4</b>	82.9	66.8	<b>88.8</b>	<b>70.5</b>	78.6	59.5	70.1	84.2	79.0	94.2	<b>80.0</b>

Table 5: Performance of video pair comparison measured by accuracy. The best and second-best results are highlighted with yellow and light yellow, respectively.

Method	Model Size	Overall Dynamic	SM	CM	LC	Overall Static	TQ	SC	AQ	Overall Temporal	AC	MN	TF	Overall Alignment	MA	AA	AVG
Random	-	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
<b>Unified Evaluators</b>																	
LLaVA-OneVision	7B	44.2	52.6	40.0	38.5	35.8	35.2	25.7	51.0	37.4	23.9	29.1	52.0	43.1	39.7	43.4	38.6
LLaVA-OneVision	72B	36.5	55.1	40.0	53.8	37.0	51.4	22.2	54.1	25.6	43.3	34.5	45.7	63.0	58.9	69.7	44.3
LLaVA-Video	7B	37.0	52.6	37.5	44.9	44.4	43.8	38.9	62.2	33.7	28.9	31.1	57.5	43.1	39.7	46.9	41.1
LLaVA-Video	72B	42.5	57.7	32.5	56.4	41.6	55.2	31.2	60.2	32.6	41.1	27.7	55.9	60.3	53.0	66.2	46.1
Qwen2-VL	7B	46.4	56.4	43.8	39.7	42.8	38.1	20.8	54.1	29.8	24.4	29.7	42.5	54.9	50.3	57.9	41.1
Qwen2-VL	72B	51.4	<b>66.7</b>	<b>71.2</b>	53.8	47.7	<b>62.9</b>	19.4	60.2	41.0	40.0	31.8	40.2	<b>69.7</b>	<b>62.9</b>	<b>78.6</b>	51.6
InternVL-2.5-MPO	8B	42.5	51.3	33.8	43.6	38.3	31.4	40.3	55.1	35.1	32.2	27.0	44.9	53.5	50.3	53.8	41.8
InternVL-2.5-MPO	78B	43.1	57.7	41.2	<b>61.5</b>	54.7	<b>62.9</b>	27.1	<b>71.4</b>	45.2	<b>47.8</b>	33.8	<b>70.9</b>	67.7	55.6	76.6	53.7
GPT-4o	-	42.0	48.7	38.8	59.0	53.5	54.3	41.0	<b>71.4</b>	44.9	38.9	31.8	59.8	58.9	57.0	61.4	50.2
Seed1.5-VL	20B Act.	52.5	56.4	47.5	<b>61.5</b>	51.0	52.4	<b>44.4</b>	62.2	48.6	36.7	35.8	65.4	56.9	53.6	61.4	51.6
Gemini2.5-Flash	-	<b>55.8</b>	60.3	45.0	59.0	<b>55.6</b>	50.5	43.1	64.3	<b>50.8</b>	41.7	<b>38.5</b>	60.6	65.0	62.3	66.2	<b>54.6</b>
Human	-	87.3	85.3	87.3	85.3	90.0	92.0	88.0	91.3	88.0	90.0	78.7	92.0	88.0	84.7	92.0	88.0

**Single Video Rating.** We sample 16 frames per video for our evaluation, except for Video-LLaVA, which has an 8-frame limitation. Unless otherwise specified, we adopt *yes/no* as the default scoring tokens for single video rating.

**Video Pair Comparison.** In this evaluation mode, we utilize 12 frames per video. As shown in Fig. 2 (c), the original UVE-Bench contains more “ $\mathcal{V}_{1/2}$  better” compared to “same good/bad”. While this imbalance does not compromise the fairness of single video ratings (since  $\mathcal{A}^{\text{single}}$  handles “better” and “same” independently) it introduces a bias toward “ $\mathcal{V}_{1/2}$  better” when computing four-way selection accuracy for video pair comparisons. To mitigate this bias, we create a subset of 2,411 samples, with a more balanced distribution across the four preference categories, in evaluating video pair comparison.

## 4.2 Performance of MLLMs as Unified Evaluator

**Single Video Rating.** Tab. 4 presents the results of MLLMs in single video rating, from which we can obtain the following findings: (1) SOTA MLLMs demonstrate strong capabilities in assessing *Dynamic Degree* and *Video-Text Alignment*, with Qwen2-VL-72B achieving over 80  $\mathcal{A}^{\text{single}}$  score across all eight subaspects. (2) current MLLMs show limitations in evaluating *Temporal Quality* aspects, which require a nuanced understanding of video temporal dynamics. Additionally, their performance in *Motion Alignment* lags behind that of *Appearance Alignment*. These observations echo with previous research [41, 52] which reveal the limitation of MLLMs in fine-grained video

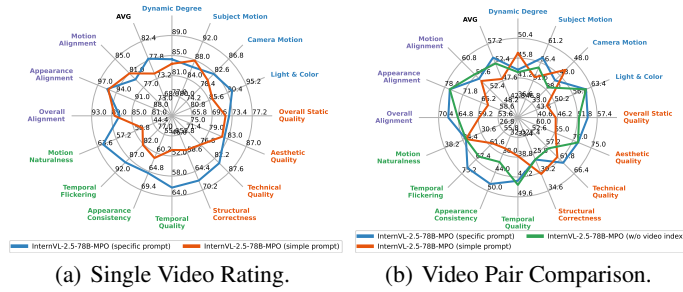


Figure 3: Results of different prompting strategies.

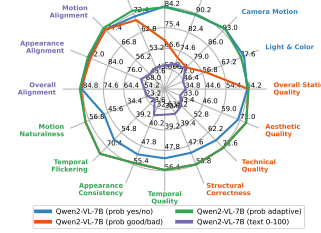


Figure 4: Results of different scoring strategies with Qwen2-VL-7B for single video rating.

temporal understanding. Interestingly, Video-LLaVA and VideoChat2-Mistral, despite their lower average  $\mathcal{A}^{\text{single}}$ , perform well in assessing motion naturalness, surpassing the 72B Qwen2-VL and LLaVA-OneVision. We attribute this to their use of native video encoders, rather than the frame-by-frame processing approach used by most MLLMs. (3) While advanced MLLMs can effectively assess *Technical Quality* and *Aesthetic Quality*, they still struggle to identify incorrect subject structures (e.g., human hands with six fingers). (4) MLLMs’ performance in single video rating strongly correlates with their general multimodal understanding capability: Larger models, such as Qwen2-VL-72B, InternVL2.5-78B-MPO, GPT-4o and Seed1.5-VL demonstrate superior performance compared to the smaller ones at 7B scale. (5) The unified evaluators consistently outperform or match specialized methods across all evaluation aspects. This indicates that utilizing MLLMs for unified AIGV evaluation is a promising direction, which showcases both versatility and effectiveness.

**Video Pair Comparison.** As shown in Tab. 5, the MLLMs’ performance in video pair comparison exhibits a similar pattern as single video rating: They achieve relative higher accuracy when evaluating *Dynamic Degree* subaspects, *Technical Quality*, *Aesthetic Quality* and *Appearance Alignment*, while showing weakness in assessing *Structural Correctness*, *Motion Naturalness* and *Motion Alignment*. Additionally, the significant performance gap of over 30 points between SOTA MLLMs and human evaluators indicates that current MLLMs have not yet achieved reliable pairwise video quality assessment capabilities.

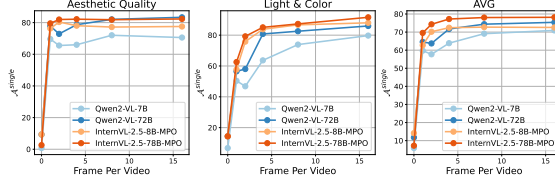
### 4.3 Factors Influencing Unified Evaluator Performance

#### 4.3.1 Impact of Prompting Strategy

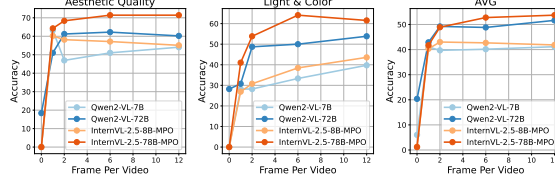
We investigate how different prompting strategies affect model performance by modifying our original prompt templates (Tab. 1). We test two variations: (1) removing the aspect-specific descriptions while maintaining only the question and answer instructions, and (2) for video pair comparisons, eliminating video order indicators (“The first/second video”). The latter modification was motivated by Zhang et al. [77], which emphasizes the importance of specifying relative order in image quality comparisons. The modified prompts can be found in the supplementary material.

**Single Video Rating.** The results in Fig. 3 (a) demonstrate that using simplified prompts (without aspect-specific descriptions) generally led to decreased performance in InternVL-2.5-78B-MPO. The performance decline was particularly pronounced in several categories: *Temporal Quality* subaspects, *Structural Correctness*, *Technical Quality*, and *Light & Color*. However, other aspects maintain relatively stable performance levels. This phenomenon suggests that it is generally beneficial to provide a detailed description of the aspect to evaluate. Results of other models also showcase an advantage of detailed aspect description on the average performance (Appendix D.1).

**Video Pair Comparison.** Similarly, Fig. 3 (b) shows that simplified prompts result in reduced performance for video pair comparisons, reinforcing the value of detailed aspect descriptions in the prompts. Interestingly, contrary to established findings in image quality comparison [77], removing video order indicators does not significantly impact performance. We attribute this robustness to the enhanced temporal order understanding capabilities of recent MLLMs.



(a) Single Video Rating.



(b) Video Pair Comparison.

Figure 5: Results with varying numbers of video frames.

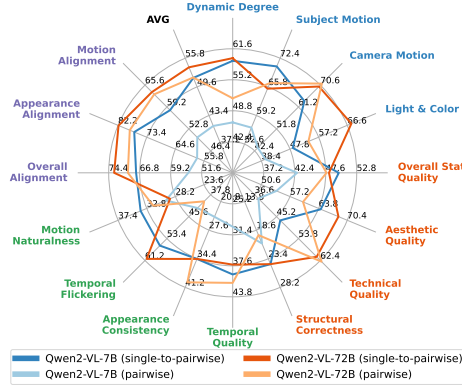


Figure 6: Results of direct video pair comparison versus adapting single video rating to pairwise comparison.

### 4.3.2 Impact of Scoring Strategy

In the previous experiments, we adopt *yes/no* as the default scoring tokens and calculate single video ratings using the *yes* probability according to Eq. 1. Here, we explore three different scoring strategies: (1) using *good/bad* as scoring tokens, (2) an adaptive approach that alternates between *yes/no* and *good/bad* tokens, and (3) prompting the MLLM to directly generate rating scores as texts, ranging from 0 to 100. The detailed prompts are shown in the supplementary material.

The results of Qwen2-VL-7B, presented in Fig. 4, reveal that: (1) *good/bad* demonstrates superior performance for *Temporal Quality* and *Static Quality* assessments, while *yes/no* performs better for *Dynamic Degree* evaluation. (2) The adaptive strategy yields the best overall results for Qwen2-VL-7B. However, for other advanced MLLMs like InternVL-2.5-78B-MPO, the improvement from adaptive scoring is minimal (see Appendix D.2). (3) Directly prompting MLLMs to generate rating scores performs significantly worse than probability-based scoring methods. These findings suggest that using *yes/no* as unified scoring tokens strikes a good balance between simplicity and effectiveness.

### 4.3.3 Adapting Single Ratings for Pairwise Comparison

While our previous experiments involve direct pairwise video comparisons by simultaneously feeding two videos into MLLMs, we now explore an alternative approach. This method first independently rates each video using Eq. 1 and then converts these individual ratings into a four-way selection from set **O**. The detailed conversion methodology is discussed in Appendix C.3.

As shown in Fig. 6, adaptation from single video rating brings substantial improvement to Qwen2-VL-7B compared to directly performing video pair comparison. However, Qwen2-VL-72B shows only minimal improvements under the same approach. Similar patterns were observed with InternVL-2.5-MPO, as shown in Appendix D.4. We hypothesize that smaller-scale models, like the 7B variant, have inherently weaker pairwise comparison capabilities, and adapting from single video rating bypasses this limitation. Based on these findings, we recommend implementing the single-to-pairwise adaptation strategy particularly for MLLMs with limited pairwise comparison capabilities.

### 4.3.4 Impact of Frame Number

Fig. 5 illustrates the impact of varying video frame numbers on model performance. We can see that: (1) In the zero-frame scenario (pure-text input), MLLMs perform poorly as their outputs remain constant regardless of visual information, confirming that language shortcuts cannot be exploited to achieve good performance in UVE-Bench. (2) As the frame count increases, aspects requiring temporal understanding (*Light & Color*) show more consistent improvement compared to aspects focusing on individual frame information (*Aesthetic Quality*). (3) The 72B-scale models demonstrate

Table 6: Automatic and human rating of video generative models. The human prefer rate is computed as the ratio of “better” plus “same good” examples, based on video pair preference annotation.

	Open-Sora 1.2	CogVideoX-5B	Movie Gen Video	HunyuanVideo	Real Videos
<b>Static Quality</b>					
Auto (Qwen2-VL-72B)	63.2	70.0	84.3	84.9	<b>85.7</b>
Human (Prefer Rate)	3.3	21.6	96.4	94.7	-
<b>Temporal Quality</b>					
Auto (Qwen2-VL-72B)	59.7	67.3	74.7	77.1	<b>78.0</b>
Human (Prefer Rate)	1.5	42.4	96.6	86.6	-

more consistent improvement in average accuracy with increasing frame numbers compared to their 7B-scale counterparts. This suggests that larger models possess superior capabilities in processing and integrating information from additional video frames.

#### 4.4 Evaluation of Real versus Generated Videos

To examine the difference when applying MLLMs to evaluate real videos instead of AI-generated videos, we collect 204 real videos from the TempCompass [41] dataset and utilize Qwen2-VL-72B to perform single video rating on two aspects: *Static Quality* and *Temporal Quality*. For comparison, we also evaluate videos generated by Open-Sora 1.2, CogVideoX-5B, Movie Gen Video and HunyuanVideo. As shown in Tab. 6, real videos consistently receive higher scores than generated ones across both quality aspects. Furthermore, the relative ranking of VGMs basically aligns with human perceptual judgments: the more recent models (Movie Gen Video and HunyuanVideo) outperform earlier ones (Open-Sora 1.2 and CogVideoX-5B) in both human and automatic evaluations. However, Qwen2-VL-72B and humans disagree slightly on the rating of Movie Gen Video and HunyuanVideo, highlighting the need for more advanced evaluators capable of finer-grained comparisons.

## 5 Conclusion and Future Work

This study investigates the potential of MLLMs to serve as unified evaluators for AI-generated videos. To answer this question, we introduce the UVE-Bench, a benchmark designed to assess the capability of unified AIGV evaluation. Compared to existing AIGV datasets, UVE-Bench highlights (1) comprehensive evaluation aspects, (2) videos produced by SOTA video generative models and (3) reliable evaluation of both single video rating and video pair comparison. Based on UVE-Bench, we extensively evaluates 18 MLLMs. Our findings reveal that while SOTA MLLMs cannot fully replace human evaluators, they significantly outperform existing specialized evaluation methods, showcasing promising potential as unified AIGV evaluators. Our analytical studies highlight several design choices, including: providing detailed aspect-specific description in the prompts (§4.3.1), using *yes/no* as the scoring tokens for single video rating (§4.3.2) and adapting single video rating for pairwise comparison when using 7B-scale MLLMs (§4.3.3).

Looking ahead, we identify two main directions for future work: **First**, we aim to extend the benchmark to cover broader AIGV evaluation scenarios, including safety, ethics, and bias assessments, as well as more diverse generation settings such as image-to-video and video-to-video tasks. **Second**, we plan to enhance MLLM performance in AIGV evaluation through three complementary directions: (1) **Advancing video encoding techniques** to capture richer temporal information without compromising efficiency, which is essential to the temporally sensitive evaluation aspects. (2) **Automatically curating high- and low-quality video pairs** to train evaluators. This can be achieved by leveraging VGMs at different training stages or intentionally varying prompts or other hyperparameters during inference to generate controlled degradations in videos. (3) **Incorporating visual generative loss** into MLLM training—alongside language generation loss—to strengthen their understanding of real-world visual distributions and underlying physical dynamics.

## Acknowledgments and Disclosure of Funding

We thank all the anonymous reviewers for their constructive comments. This work is supported in part by ByteDance Seed and National Natural Science Foundation of China (No. 62176002). Haoyuan Guo and Xu Sun are the corresponding authors of this paper.

## References

- [1] H. Bansal, Y. Bitton, I. Szpektor, K. Chang, and A. Grover. Videocon: Robust video-language alignment via contrast captions. In *CVPR*, pages 13927–13937. IEEE, 2024.
- [2] H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, C. Jiang, Y. Sun, K.-W. Chang, and A. Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.
- [3] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024.
- [4] Z. Cai, M. Cao, H. Chen, K. Chen, K. Chen, X. Chen, X. Chen, Z. Chen, Z. Chen, P. Chu, X. Dong, H. Duan, Q. Fan, Z. Fei, Y. Gao, J. Ge, C. Gu, Y. Gu, T. Gui, A. Guo, Q. Guo, C. He, Y. Hu, T. Huang, T. Jiang, P. Jiao, Z. Jin, Z. Lei, J. Li, J. Li, L. Li, S. Li, W. Li, Y. Li, H. Liu, J. Liu, J. Hong, K. Liu, K. Liu, X. Liu, C. Lv, H. Lv, K. Lv, L. Ma, R. Ma, Z. Ma, W. Ning, L. Ouyang, J. Qiu, Y. Qu, F. Shang, Y. Shao, D. Song, Z. Song, Z. Sui, P. Sun, Y. Sun, H. Tang, B. Wang, G. Wang, J. Wang, J. Wang, R. Wang, Y. Wang, Z. Wang, X. Wei, Q. Weng, F. Wu, Y. Xiong, X. Zhao, and et al. Internlm2 technical report. *CoRR*, abs/2403.17297, 2024.
- [5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640. IEEE, 2021.
- [6] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.
- [7] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang, L. Yuan, Y. Qiao, D. Lin, F. Zhao, and J. Wang. Sharegpt4video: Improving video understanding and generation with better captions. *ArXiv preprint*, abs/2406.04325, 2024.
- [8] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238, 2023.
- [9] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [10] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, and L. Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024.
- [11] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [12] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen,



- H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.
- [13] P. Gao, R. Zhang, C. Liu, L. Qiu, S. Huang, W. Lin, S. Zhao, S. Geng, Z. Lin, P. Jin, K. Zhang, W. Shao, C. Xu, C. He, J. He, H. Shao, P. Lu, H. Li, and Y. Qiao. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *ArXiv*, abs/2402.05935, 2024. URL <https://api.semanticscholar.org/CorpusID:267547619>.
- [14] Google and DeepMind. Gemini 2.5: Our most intelligent ai model, 2025.
- [15] D. Guo, F. Wu, F. Zhu, F. Leng, G. Shi, H. Chen, H. Fan, J. Wang, J. Jiang, J. Wang, J. Chen, J. Huang, K. Lei, L. Yuan, L. Luo, P. Liu, Q. Ye, R. Qian, S. Yan, S. Zhao, S. Peng, S. Li, S. Yuan, S. Wu, T. Cheng, W. Liu, W. Wang, X. Zeng, X. Liu, X. Qin, X. Ding, X. Xiao, X. Zhang, X. Zhang, X. Xiong, Y. Peng, Y. Chen, Y. Li, Y. Hu, Y. Lin, Y. Hu, Y. Zhang, Y. Wu, Y. Li, Y. Liu, Y. Ling, Y. Qin, Z. Wang, Z. He, A. Zhang, B. Yi, B. Liao, C. Huang, C. Zhang, C. Deng, C. Deng, C. Lin, C. Yuan, C. Li, C. Gou, C. Lou, C. Wei, C. Liu, C. Li, D. Zhu, D. Zhong, F. Li, F. Zhang, G. Wu, G. Li, G. Xiao, H. Lin, H. Yang, H. Wang, H. Ji, H. Hao, H. Shen, H. Li, J. Li, J. Wu, J. Zhu, J. Jiao, J. Feng, J. Chen, J. Duan, J. Liu, J. Zeng, J. Tang, J. Sun, J. Chen, J. Long, J. Feng, J. Zhan, J. Fang, J. Lu, K. Hua, K. Liu, K. Shen, K. Zhang, K. Shen, K. Wang, K. Pan, K. Zhang, K. Li, L. Li, L. Li, L. Shi, L. Han, L. Xiang, L. Chen, L. Chen, L. Li, L. Yan, L. Chi, L. Liu, M. Du, M. Wang, N. Pan, P. Chen, P. Chen, P. Wu, Q. Yuan, Q. Shuai, Q. Tao, R. Zheng, R. Zhang, R. Zhang, R. Wang, R. Yang, R. Zhao, S. Xu, S. Liang, S. Yan, S. Zhong, S. Cao, S. Wu, S. Liu, S. Chang, S. Cai, T. Ao, T. Yang, T. Zhang, W. Zhong, W. Jia, W. Weng, W. Yu, W. Huang, W. Zhu, W. Yang, W. Wang, X. Long, X. Yin, X. Li, X. Zhu, X. Jia, X. Zhang, X. Liu, X. Zhang, X. Yang, X. Luo, X. Chen, X. Zhong, X. Xiao, X. Li, Y. Wu, Y. Wen, Y. Du, Y. Zhang, Y. Ye, Y. Wu, Y. Liu, Y. Yue, Y. Zhou, Y. Yuan, Y. Xu, Y. Yang, Y. Zhang, Y. Fang, Y. Li, Y. Ren, Y. Xiong, Z. Hong, Z. Wang, Z. Sun, Z. Wang, Z. Cai, Z. Zha, Z. An, Z. Zhao, Z. Xu, Z. Chen, Z. Wu, Z. Zheng, Z. Wang, Z. Huang, Z. Zhu, and Z. Song. Seed1.5-vl technical report. *CoRR*, abs/2505.07062, 2024.
- [16] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, L. Fei-Fei, I. Essa, L. Jiang, and J. Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.
- [17] X. He, D. Jiang, G. Zhang, M. Ku, A. Soni, S. Siu, H. Chen, A. Chandra, Z. Jiang, A. Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024.
- [18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- [19] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- [20] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *NeurIPS*, 2022.
- [21] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818. IEEE, 2024.
- [22] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.
- [23] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen. MANTIS: interleaved multi-image instruction tuning. *CoRR*, abs/2405.01483, 2024.
- [24] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. MUSIQ: multi-scale image quality transformer. In *ICCV*, pages 5128–5137. IEEE, 2021.

- [25] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [26] T. Kou, X. Liu, Z. Zhang, C. Li, H. Wu, X. Min, G. Zhai, and N. Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024.
- [27] Kuaishou. Kling ai. <https://klingai.kuaishou.com/>, 2024.
- [28] LAION-AI. aesthetic-predictor, 2022.
- [29] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer. *ArXiv preprint*, abs/2408.03326, 2024.
- [30] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [31] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, L. Wang, and Y. Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv*, abs/2311.17005, 2023. URL <https://api.semanticscholar.org/CorpusID:265466214>.
- [32] X. Li, W. Chu, Y. Wu, W. Yuan, F. Liu, Q. Zhang, F. Li, H. Feng, E. Ding, and J. Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023.
- [33] Z. Li, Z. Zhu, L. Han, Q. Hou, C. Guo, and M. Cheng. AMT: all-pairs multi-field transforms for efficient frame interpolation. In *CVPR*, pages 9801–9810. IEEE, 2023.
- [34] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection. *ArXiv*, abs/2311.10122, 2023. URL <https://api.semanticscholar.org/CorpusID:265281544>.
- [35] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV (9)*, volume 15067 of *Lecture Notes in Computer Science*, pages 366–384. Springer, 2024.
- [36] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. URL <https://api.semanticscholar.org/CorpusID:258179774>.
- [37] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [38] Y. Liu, S. Li, Y. Wu, C. W. Chen, Y. Shan, and X. Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3041, 2022. URL <https://api.semanticscholar.org/CorpusID:247627801>.
- [39] Y. Liu, L. Li, S. Ren, R. Gao, S. Li, S. Chen, X. Sun, and L. Hou. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *NeurIPS*, 2023.
- [40] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, pages 22139–22149. IEEE, 2024.
- [41] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou. TempCompass: Do video LLMs really understand videos? In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 8731–8772, 2024.
- [42] LumaLabs. Dream machine. <https://lumalabs.ai/dream-machine>, 2024.
- [43] G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun, and T. Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

- [44] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL <https://api.semanticscholar.org/CorpusID:266362871>.
- [45] OpenAI. Gpt-4o system card, 2024.
- [46] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.
- [47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
- [48] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- [50] Runway. Gen-3. <https://runwayml.com/research/introducing-gen-3-alpha>, 2024.
- [51] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.
- [52] Z. Shangquan, C. Li, Y. Ding, Y. Zheng, Y. Zhao, T. Fitzgerald, and A. Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*, 2024.
- [53] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.
- [54] G. Team. Mochi 1. <https://github.com/genmoai/models>, 2024.
- [55] Z. Teed and J. Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV* (2), volume 12347 of *Lecture Notes in Computer Science*, pages 402–419. Springer, 2020.
- [56] J. Wang, H. Duan, G. Zhai, J. Wang, and X. Min. Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with Imm. *arXiv preprint arXiv:2411.17221*, 2024.
- [57] J. Wang, B. Wu, H. Jiang, Z. Xun, X. Xiao, H. Guo, and J. Xiao. World to code: Multi-modal data generation via self-instructed compositional captioning and filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4608–4623, 2024.
- [58] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [59] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*. OpenReview.net, 2024.
- [60] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, pages 20087–20097. IEEE, 2023.
- [61] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, Q. Yan, X. Min, G. Zhai, and W. Lin. Q-align: Teaching Imms for visual scoring via discrete text-defined levels. In *ICML*. OpenReview.net, 2024.

- [62] H. Wu, H. Zhu, Z. Zhang, E. Zhang, C. Chen, L. Liao, C. Li, A. Wang, W. Sun, Q. Yan, X. Liu, G. Zhai, S. Wang, and W. Lin. Towards open-ended visual quality comparison. In *ECCV (3)*, volume 15061 of *Lecture Notes in Computer Science*, pages 360–377. Springer, 2024.
- [63] J. Z. Wu, G. Fang, H. Wu, X. Wang, Y. Ge, X. Cun, D. J. Zhang, J. Liu, Y. Gu, R. Zhao, W. Lin, W. Hsu, Y. Shan, and M. Z. Shou. Towards A better metric for text-to-video generation. *CoRR*, abs/2401.07781, 2024.
- [64] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296. IEEE Computer Society, 2016.
- [65] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024.
- [66] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [67] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, Q. Chen, H. Zhou, Z. Zou, H. Zhang, S. Hu, Z. Zheng, J. Zhou, J. Cai, X. Han, G. Zeng, D. Li, Z. Liu, and M. Sun. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800, 2024.
- [68] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *CoRR*, abs/2408.04840, 2024.
- [69] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. Yang, Y. Hao, I. Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023.
- [70] A. Zeng, Y. Yang, W. Chen, and W. Liu. The dawn of video generation: Preliminary explorations with sora-like models. *CoRR*, abs/2410.05227, 2024.
- [71] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11941–11952. IEEE, 2023.
- [72] P. Zhang, K. Zhang, B. Li, G. Zeng, J. Yang, Y. Zhang, Z. Wang, H. Tan, C. Li, and Z. Liu. Long context transfer from language to vision. *ArXiv preprint*, abs/2406.16852, 2024.
- [73] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li. Video instruction tuning with synthetic data. *CoRR*, abs/2410.02713, 2024.
- [74] Z. Zhang, Z. Jia, H. Wu, C. Li, Z. Chen, Y. Zhou, W. Sun, X. Liu, X. Min, W. Lin, et al. Q-bench-video: Benchmarking the video quality understanding of lmms. *arXiv preprint arXiv:2409.20063*, 2024.
- [75] Z. Zhang, X. Li, W. Sun, J. Jia, X. Min, Z. Zhang, C. Li, Z. Chen, P. Wang, Z. Ji, et al. Benchmarking aigc video quality assessment: A dataset and unified model. *arXiv preprint arXiv:2407.21408*, 2024.
- [76] Z. Zhang, H. Wu, E. Zhang, G. Zhai, and W. Lin. Q-bench<sup>++</sup>: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10404–10418, 2024.
- [77] Z. Zhang, Y. Zhou, C. Li, B. Zhao, X. Liu, and G. Zhai. Quality assessment in the era of large models: A survey. *CoRR*, abs/2409.00031, 2024.
- [78] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, J. Liu, W. Wu, J. Keppo, and M. Z. Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *ECCV (56)*, volume 15114 of *Lecture Notes in Computer Science*, pages 273–290. Springer, 2024.

- [79] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.
- [80] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, C. Zhang, Z. Li, W. Liu, and L. Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*. OpenReview.net, 2024.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: See the Abstract and Introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: See Section 5

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: See Section 4.1 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?



Answer: [Yes]

Justification: We provide publicly available code and data repository with clear instruction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Section 4.1 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We set temperature to 0 during generation, and thus there is no random factor.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments are conducted using 80GB GPU.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: This paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Appendix A

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The dataset released by the paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: See the code and data link.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

#### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: See the code and data link.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[No\]](#)

Justification: The human annotations are conducted by the authors.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[No\]](#)

Justification: This study do not pose potential risk to the participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [\[NA\]](#)

Justification: We only use LLMs for paper editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Impact Statement

The evaluation of video generative models (VGMs) currently relies predominantly on human assessment or qualitative analysis. Our research demonstrates that multimodal large language models (MLLMs) could potentially serve as unified evaluators for AI-generated videos. The transition from human to automated evaluation frameworks would have significant dual implications: On the one hand, this shift would substantially reduce evaluation costs, thereby accelerating the development cycle of VGMs. On the other hand, if VGMs are optimized solely based on model feedback, there is a potential risk of divergence between the automated evaluators’ preferences and human values. Therefore, it is essential to carefully scrutinize the correlation between automatic metrics and humans in the future research on AIGV evaluation.

## B Related Work

### B.1 Video Generative Models

The emergence of diffusion models [18, 53] has revolutionized generative applications, spanning from image synthesis [51, 48, 49] to video generation [16, 69, 19, 20]. On the one hand, recent VGMs like SORA [3], Kling [27], and MovieGen [46] have achieved remarkable success in producing highly realistic videos that closely resemble real ones. On the other hand, the condition signal for video generation has been expanded from text (T2V) to RGB images (I2V) [32, 6] and other videos (V2V) [78]. This work specifically focuses on evaluating text-conditioned video generation.

### B.2 Multimodal Large Language Models

Multimodal large language models [45, 58, 29, 30] integrate LLMs with additional perception modules, enabling them to comprehend information beyond the text modality. With advancements in LLM backbones and the increasing scale of multimodal training, MLLMs are showing promising performance in understanding real-world images [36, 37, 13, 57] and videos [73, 31, 34]. However, understanding AI-generated visual content presents unique challenges compared to real images, and the application of MLLMs in this field is still in its early stages.

### B.3 AI-Generated Video Evaluation

Despite the progress in VGMs, they still struggle to consistently generate high-quality videos, which necessitate a comprehensive and fine-grained evaluation. Initial evaluation approaches [21, 40] employ off-the-shelf models to evaluate specific aspects, such as CLIP [47] for video-text alignment and DINO [5] for frame-to-frame consistency. With the advancement of MLLMs, recent studies have incorporated these models in AIGV evaluation [17, 61, 56, 63, 2]. However, they are constrained to a fixed set of evaluation aspects and typically rely on human annotations to further fine-tune the MLLMs for AIGV evaluation. By contrast, in this work explores using MLLMs as unified evaluators for any AIGV aspect by leveraging their inherent vision-language understanding capabilities, eliminating the need for human annotations.

## C More Details of Experimental Settings

### C.1 Evaluated Systems

**MLLMs.** We conduct zero-shot evaluations on 18 MLLMs, the detailed information of which is summarized in Tab. 7.

**VideoScore-v1.1 [17].** This model is a fine-tuned version of Mantis-Idefics2-8B [23], trained to imitate human rating scores. In their work, He et al. [17] gathered human ratings (on a discrete scale from 1 to 4) for five key evaluation aspects: *Visual Quality*, *Temporal Consistency*, *Dynamic Degree*, *Text-to-Video Alignment*, and *Factual Consistency*. The VideoScore model is trained using a regression approach to predict these human rating scores. According to the definition of these aspects, we associate them with 8 out of the 15 aspects in UVE-Bench, as detailed in Tab. 8.

**VBench [21].** VBench is a benchmark suite that comprehensively evaluates 16 aspects of AIGV using off-the-shelf models and specifically designed rules. As shown in Tab. 8, we identified an overlap of 6 aspects between VBench and our UVE-Bench, and we adopted the VBench metrics to evaluate these 6 aspects in UVE-Bench.

**UMTScore [39].** UMTScore is designed to assess video-text alignment. It builds on the UMT-L/16 model [38], fine-tuned on the MSR-VTT dataset [64] for video-text retrieval. Within UVE-Bench, we employ UMTScore to evaluate the *Overall Alignment* aspect.

**VIDEOCON-PHYSICS [2].** VIDEOCON-PHYSICS evaluates (1) whether a generated video adheres to physical commonsense and (2) whether it matches the given textual prompt. The model is initialized from VIDEOCON [1] and fine-tuned on a dataset of AI-generated videos with binary human annotations covering these two aspects. In UVE-Bench, we use VIDEOCON-PHYSICS to evaluate both *Motion Naturalness* and *Overall Alignment*.

**DOVER [60].** DOVER assesses video quality from both technical and aesthetic perspectives. It is trained on DIVIDE-3k, a dataset of 3,590 real-world videos annotated with mean opinion scores (MOS). We adopt DOVER in UVE-Bench to evaluate *Technical Quality* and *Aesthetic Quality*.

## C.2 Evaluation Criteria for Single Video Rating

We rewrite the evaluation criteria for single video rating in Eq. 3 as follows:

$$\mathcal{A}^{\text{single}} = \begin{cases} 1(\mathcal{S}_1 > \mathcal{S}_2) & \text{if } \mathcal{P} = \text{"}\mathcal{V}_1 \text{ better"}$$

$$\begin{cases} 1(\mathcal{S}_1 < \mathcal{S}_2) & \text{elif } \mathcal{P} = \text{"}\mathcal{V}_2 \text{ better"}$$

$$\begin{cases} f_c(\mathcal{S}_1|\beta) \cdot f_c(\mathcal{S}_2|\beta) & \text{elif } \mathcal{P} = \text{"same good"}$$

$$\begin{cases} f'_c(\mathcal{S}_1|\alpha) \cdot f'_c(\mathcal{S}_2|\alpha) & \text{elif } \mathcal{P} = \text{"same bad"}$$
 \end{cases} \quad (4)

$f_c(\mathcal{S}|\beta)$  and  $f'_c(\mathcal{S}|\alpha)$  exponentially decays from 1 to 0 when the predicted rating score  $\mathcal{S}$  gradually deviates from "same good" and "same bad", respectively:

$$f_c(\mathcal{S}|\beta) = \begin{cases} 1 & \text{if } \beta < \mathcal{S} \leq 1 \\ e^{-s \cdot (\beta - \mathcal{S})} & \text{elif } 0 \leq \mathcal{S} \leq \beta \end{cases} \quad (5)$$

$$f'_c(\mathcal{S}|\alpha) = \begin{cases} 1 & \text{if } 0 \leq \mathcal{S} < \alpha \\ e^{-s \cdot (\mathcal{S} - \alpha)} & \text{elif } \alpha \leq \mathcal{S} \leq 1 \end{cases} \quad (6)$$

where  $s$  is the coefficient that controls the speed of decaying. The graph of  $f_c(\mathcal{S}|\beta)$  and  $f'_c(\mathcal{S}|\alpha)$  are illustrated in Fig. 7. An analysis of the impact of  $\alpha$  and  $\beta$  on the final result is conducted in Appendix D.5.

## C.3 Methodology for Adapting Single Video Rating for Pairwise Comparison

We first independently rate the two videos using Eq. 1 to obtain the rating scores  $\mathcal{S}_1$  and  $\mathcal{S}_2$ . Based on these scores, we then make a choice from  $\mathcal{O}$ :

$$C = \begin{cases} f_1(\{\mathcal{V}_1 \text{ better}, \mathcal{V}_2 \text{ better}\}, \mathcal{S}_1, \mathcal{S}_2) & \text{if } |\mathcal{S}_1 - \mathcal{S}_2| > \tau \\ & \text{or } \alpha < \mathcal{S}_1 < \beta \\ & \text{or } \alpha < \mathcal{S}_2 < \beta \\ f_1(\{\text{same good}, \text{same bad}\}, \mathcal{S}_1, \mathcal{S}_2) & \text{else} \end{cases} \quad (7)$$

where

$$f_1(\{\mathcal{C}_1, \mathcal{C}_2\}, \mathcal{S}_1, \mathcal{S}_2) = \begin{cases} \mathcal{C}_1 & \text{if } \mathcal{S}_1 > \mathcal{S}_2 \\ \mathcal{C}_2 & \text{else} \end{cases} \quad (8)$$

$$f_2(\{\mathcal{C}_1, \mathcal{C}_2\}, \mathcal{S}_1, \mathcal{S}_2) = \begin{cases} \mathcal{C}_1 & \text{if } \mathcal{S}_1 > \beta \text{ and } \mathcal{S}_2 > \beta \\ \mathcal{C}_2 & \text{elif } \mathcal{S}_1 < \alpha \text{ and } \mathcal{S}_2 < \alpha \end{cases} \quad (9)$$

In the above equations,  $\alpha < \beta$  are the thresholds for "bad" and "good" videos, respectively.  $\tau$  is a threshold of the difference between  $\mathcal{S}_1$  and  $\mathcal{S}_2$ , which controls whether we should select from  $\{\text{"}\mathcal{V}_1 \text{ better"}$ ,  $\text{"}\mathcal{V}_2 \text{ better"}\}$  or  $\{\text{"same good"}$ ,  $\text{"same bad"}\}$ . By default, we set  $\alpha = 0.4$ ,  $\beta = 0.8$ ,  $\tau = 0.05$ .



## C.4 Annotation Interface

Fig. 8 presents the annotation interface used for establishing our human baseline. Annotators are shown a pair of videos along with aspect-specific evaluation instructions. Based on these, they indicate their preference by selecting one of four options: {"A is better", "B is better", "same good", "same bad"}.

## D More Experimental Results

### D.1 Effect of Prompting Strategy

Fig. 9 presents the results of different prompting strategy with InternVL-2.5-8B-MPO, Qwen2-VL-7B and Qwen2-VL-72B. Unlike InternVL-2.5-78B-MPO (previously presented in Fig. 3), these three models showed a less pronounced advantage when using detailed aspect-specific descriptions, with the simplified prompt achieving better performance in several aspects. We conjecture that this is because these three models cannot comprehend the detailed description as effectively as InternVL-2.5-78B-MPO. Nevertheless, when considering overall performance, the full prompt consistently maintained a slight improvement over the simplified version across all three models. Consistent with our findings for InternVL-2.5-78B-MPO, removing video order indicator does not result in significant performance change.

### D.2 Effect of Scoring Strategy

In §4.3.2, we analyze the effect of different scoring strategies with Qwen2-VL-7B. Here, we extended the investigation to three additional models: Qwen2-VL-72B, InternVL-2.5-8B-MPO, and InternVL-2.5-78B. The results are presented in Fig. 10. Our findings reveal consistent patterns across all models. Similar to Qwen2-VL-7B, the *good/bad* and *yes/no* scoring strategies demonstrate distinct strengths in different evaluation aspects. Direct score generation through MLLMs consistently yields suboptimal results across all models. However, compared with Qwen2-VL-7B, the advantage of adaptive scoring token is less pronounced in Qwen2-VL-72B, InternVL-2.5-8B-MPO and InternVL-2.5-78B. This suggests that these three models are more robust to the variation of scoring tokens. These findings confirm that using *yes/no* as unified scoring token is simple yet effective.

### D.3 Negative Prompt versus Positive Prompt

By default, we prompt MLLMs to assess whether a video is "good" in a specific aspect. As an alternative, we can also frame the question in a negative form—asking whether a video is "bad" in that aspect. To examine the impact of this change, we modify the evaluation prompts accordingly and invert the positive and negative token assignments when computing single video ratings in Eq. 1.

We then evaluate the Qwen2-VL-7B model using both positive and negative prompts across four aspects. As shown in Table 11, the results reveal clear differences across dimensions. For *Static Quality* and *Temporal Quality*, the performance gap between positive and negative prompts is relatively small. In contrast, for *Dynamic Degree* and *Video-Text Alignment*, positive prompts yield substantially higher performance. These findings suggest that positive prompts provide more stable and reliable evaluations. Therefore, we adopt positive prompts as the default configuration in our experiments.

### D.4 Adapting Single Ratings for Pairwise Comparison

While our previous experiments involve direct pairwise video comparisons by simultaneously feeding two videos into MLLMs, we now explore an alternative approach. This method first independently rates each video using Eq. 1 and then converts these individual ratings into a four-way selection from set  $\mathcal{O}$ . The detailed conversion methodology is discussed in Appendix C.3.

As shown in Fig. 6, adaptation from single video rating brings substantial improvement to Qwen2-VL-7B compared to directly performing video pair comparison. However, Qwen2-VL-72B shows only minimal improvements under the same approach. Similar patterns were observed with InternVL-2.5-MPO, as shown in Appendix D.4. We hypothesize that smaller-scale models, like the 7B variant, have inherently weaker pairwise comparison capabilities, and adapting from single video rating

Table 7: Summary of MLLMs evaluated in the experiments.

Model	Model Size	LLM	Vision Encoder	Frames per Video
Video-LLaVA [34]	7B	Vicuna-1.5 [11]	LanguageBind [80]	8
LongVA-7B-DPO [72]	7B	Qwen2 [65]	CLIP-ViT-L-336px [47]	16
ShareGPT4Video [7]	8B	LLaMA3 [12]	CLIP-ViT-L-336px [47]	16
VideoLLaMA2.1 [10]	7B	Qwen2 [65]	SigLip-400M [71]	16
mPLUG-Owl3-7B [68]	7B	Qwen2 [65]	SigLip-400M [71]	16
VideoChat2-Mistral [31]	7B	Mistral [22]	UMT-L/16 [38]	16
MiniCPM-V-2.6 [67]	7B	Qwen2 [65]	SigLip-400M [71]	16
LLaVA-OneVision-7B [29]	7B	Qwen2 [65]	SigLip-400M [71]	12,16
LLaVA-OneVision-72B [29]	72B	Qwen2 [65]	SigLip-400M [71]	12,16
LLaVA-Video-7B [73]	7B	Qwen2 [65]	SigLip-400M [71]	12,16
LLaVA-Video-72B [73]	72B	Qwen2 [65]	SigLip-400M [71]	12,16
Qwen2-VL-7B [58]	7B	-	-	12,16
Qwen2-VL-72B [58]	72B	-	-	12,16
InternVL-2.5-8B [9]	8B	InternLM2.5 [4]	InternViT [8]	12,16
InternVL-2.5-78B [9]	78B	InternLM2.5 [4]	InternViT [8]	12,16
GPT-4o-2024-08-06 [45]	-	-	-	12,16
Seed1.5-VL [15]	20B Act.	Seed1.5-LLM	Seed-ViT	12,16
Gemini2.5-Flash [14]	-	-	-	12

bypasses this limitation. Based on these findings, we recommend implementing the single-to-pairwise adaptation strategy particularly for MLLMs with limited pairwise comparison capabilities.

The effectiveness of adapting single video ratings for video pair comparisons using InternVL-2.5 models is illustrated in Fig. 11. The results aligns with our observations with Qwen2-VL (shown in Fig. 6), revealing a model size-dependent pattern. While this adaptation strategy shows no benefit when applied to the 78B-scale model, it yields substantial improvements for the 8B-scale model.

## D.5 Effect of $\alpha$ and $\beta$

We have chosen default values of  $\alpha = 0.4$  and  $\beta = 0.8$  to represent the thresholds for categorizing videos as “bad” and “good”, respectively. To investigate the impact of these two parameters on single video rating evaluation, we conduct a sensitivity analysis. Specifically, we report the performance of four MLLMs across different values of  $\alpha$  and  $\beta$ . As shown in Tab. 9 and Tab. 10, adjusting these thresholds reveals notable trends in performance: (1) Lowering  $\beta$  (i.e., making the criteria for “good” more lenient) leads to increased agreement between MLLMs and humans in the “same good” category, resulting in higher performance. (2) Increasing  $\alpha$  (i.e., making the criteria for “bad” stricter) also improves performance by enhancing alignment in the “same bad” category. However, despite the observed variations in performance as these parameters change, it is important to note that the relative performance of different MLLMs remains stable.

## D.6 Case Studies

Tab. 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 showcase UVE-Bench data examples of different subaspects, along with the evaluation results by InternVL-2.5-8B-MPO and InternVL-2.5-78B-MPO.

Table 8: Association between VideoScore and VBench aspects with UVE-Bench aspects.

Original Aspect	Definition	Evaluation Model	UVE-Bench Aspect
<i>VideoScore</i>			
Visual Quality	<i>the quality of the video in terms of clearness, resolution, brightness, and color</i>	VideoScore-v1.1	Overall Static Quality Aesthetic Quality Technical Quality
Temporal Consistency	<i>the consistency of objects or humans in video</i>	VideoScore-v1.1	Appearance Consistency
Dynamic Degree	<i>the degree of dynamic changes</i>	VideoScore-v1.1	Overall Dynamic Degree
Text-to-Video Alignment	<i>the alignment between the text prompt and the video content</i>	VideoScore-v1.1	Overall Alignment
Factual Consistency	<i>the consistency of the video content with the common-sense and factual knowledge</i>	VideoScore-v1.1	Structural Correctness Motion Naturalness
<i>VBench</i>			
Overall Consistency	<i>overall video-text consistency</i>	ViCLIP [59]	Overall Alignment
Motion Smoothness	<i>whether the motion in the generated video is smooth and follows the physical law of the real world</i>	AMT [33]	Motion Naturalness
Aesthetic Quality	<i>reflect aesthetic aspects such as the layout, the richness and harmony of colors, the photo-realism, naturalness, and artistic quality of the video frames</i>	LAION aesthetic predictor [28]	Aesthetic Quality
Imaging Quality	<i>refers to the distortion (e.g., over-exposure, noise, blur) presented in the generated frames</i>	MUSIQ [24]	Technical Quality
Dynamic Degree	<i>the degree of dynamics (i.e., whether it contains large motions)</i>	RAFT [55]	Overall Dynamic Degree
Subject Consistency	<i>whether the subject appearance remains consistent throughout the whole video</i>	DINO [5]	Appearance Consistency

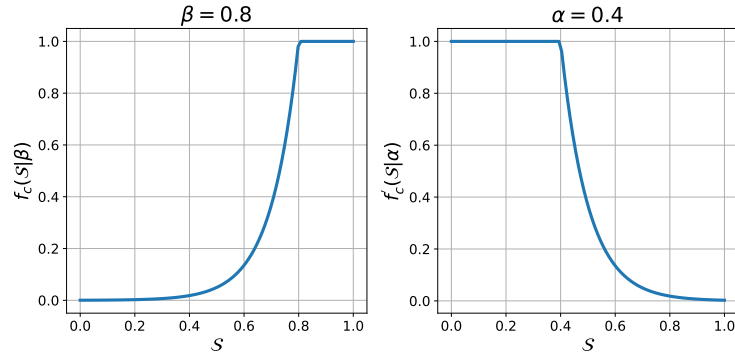


Figure 7: Functions  $f_c(S|\beta)$ ,  $f'_c(S|\alpha)$  used in the evaluation criteria of single video rating.

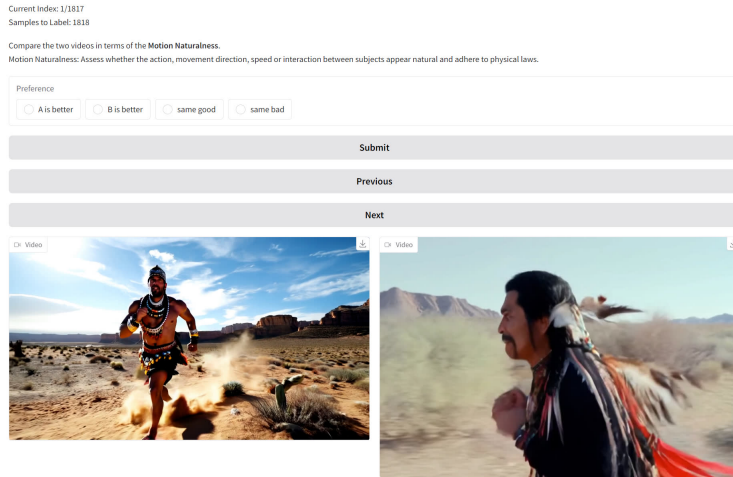
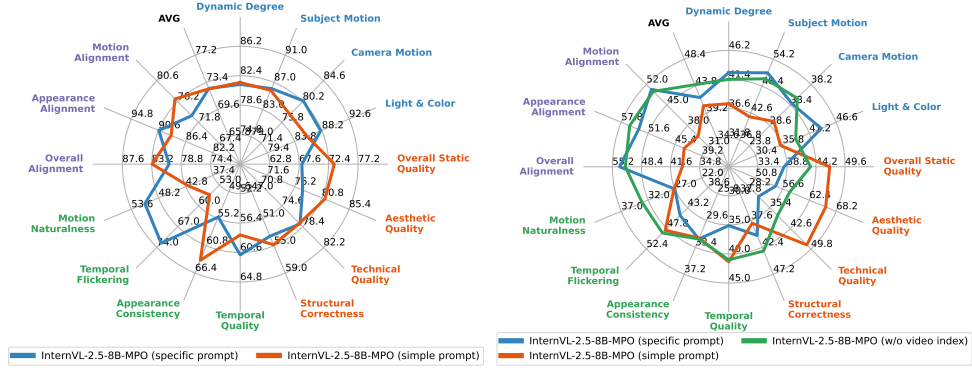
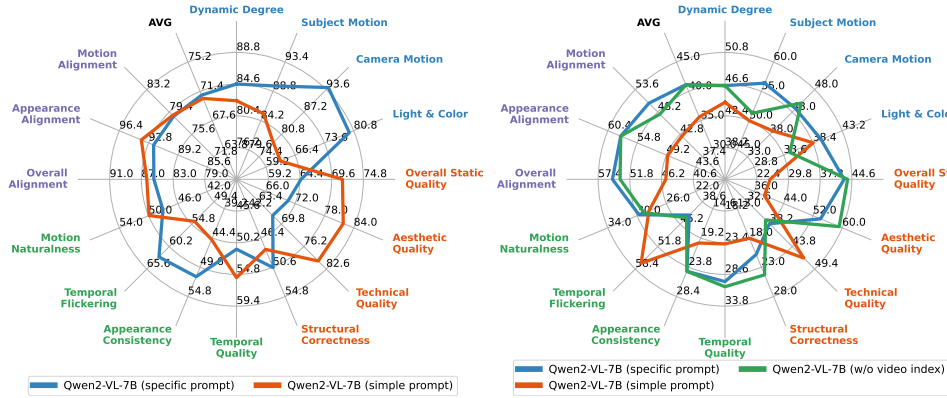


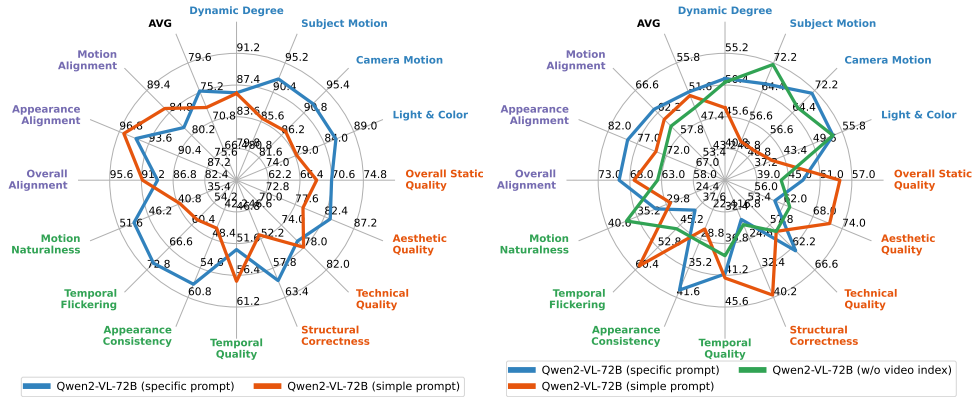
Figure 8: Human baseline annotation interface.



(a) Single Video Rating with InternVL2.5-8B-MPO. (b) Video Pair Comparison with InternVL2.5-8B-MPO.



(c) Single Video Rating with Qwen2-VL-7B. (d) Video Pair Comparison with Qwen2-VL-7B.



(e) Single Video Rating with Qwen2-VL-72B. (f) Video Pair Comparison with Qwen2-VL-72B.

Figure 9: Results of different prompting strategies with InternVL2.5-8B-MPO, Qwen2-VL-7B and Qwen2-VL-72B.

Table 9:  $\mathcal{A}^{single}$  results with varying values of  $\beta$  when  $\alpha$  is set to 0.4.

	$\beta=0.7$	$\beta=0.8$	$\beta=0.9$
Qwen2-VL-2B	61.2	58.3	55.9
LongVA-DPO-7B	64.1	61.6	57.6
Qwen2-VL-7B	74.4	70.9	66.3
Qwen2-VL-72B	77.3	75.4	72.5

Table 10:  $\mathcal{A}^{single}$  results with varying values of  $\alpha$  when  $\beta$  is set to 0.8.

	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$
Qwen2-VL-2B	56.7	58.3	62.8
LongVA-DPO-7B	60.4	61.6	63.0
Qwen2-VL-7B	69.6	70.9	72.5
Qwen2-VL-72B	74.7	75.4	76.2

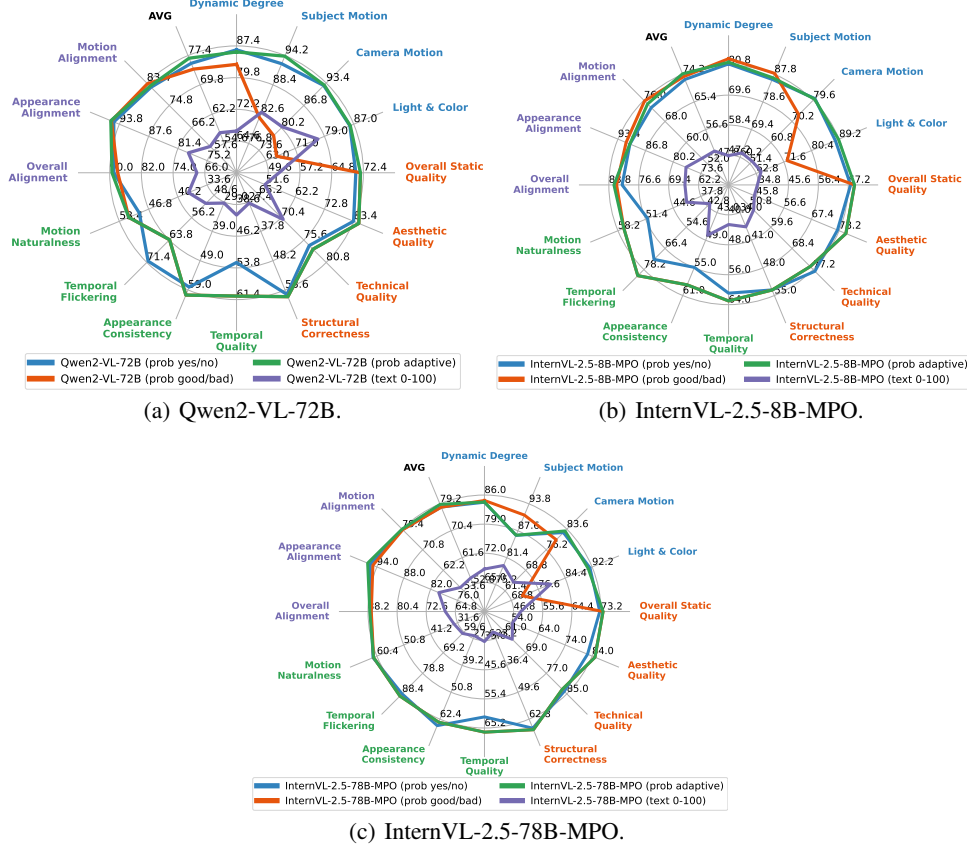


Figure 10: Results of different scoring strategies for single video rating with Qwen2-VL-72B, InternVL-2.5-8B-MPO and InternVL-2.5-78B-MPO.

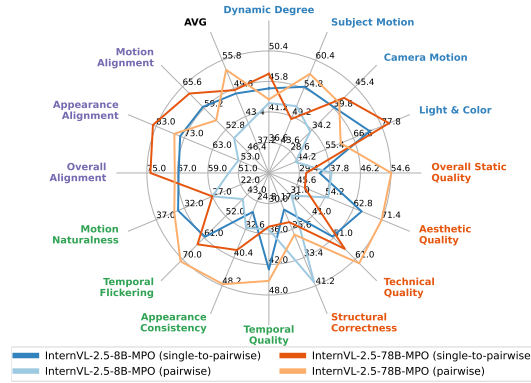


Figure 11: Results of direct video pair comparison versus adapting single video rating to pairwise comparison, using InternVL-2.5-MPO.

Table 11: Comparison of positive and negative prompts in terms of  $\mathcal{A}^{single}$  with Qwen2-VL-7B as the backbone model.

Prompt	Dynamic Degree	Static Quality	Temporal Quality	Video-Text Alignment	AVG
Positive	84.6	64.6	51.1	85.4	70.9
Negative	66.0	62.7	55.0	16.9	50.9

Table 12: Data example of *Subject Motion* and corresponding evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.



<b>Video 1:</b> 	
<b>Video 2:</b> 	
<b>Aspect:</b> Subject Motion (Dynamic Degree) <b>Human Preference:</b> same bad	
<b>Single Video Rating</b> <b>InternVL-2.5-78B</b> - Video 1: 0.076 - Video 2: 0.0028 - $\mathcal{A}_{\text{single}} = 1$	<b>InternVL-2.5-8B</b> - Video 1: 0.085 - Video 2: 0.047 - $\mathcal{A}_{\text{single}} = 1$
<b>Video Pair Comparison</b> <b>InternVL-2.5-78B</b> - Model Preference: same bad - Accuracy = ✓	<b>InternVL-2.5-8B</b> - Model Preference: Video 2 - Accuracy = ✗

Table 13: Data example of *Camera Motion* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

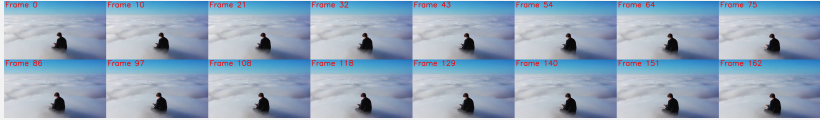



<b>Video 1:</b> 	
<b>Video 2:</b> 	
<b>Aspect:</b> Camera Motion (Dynamic Degree) <b>Human Preference:</b> Video 1 is better	
<b>Single Video Rating</b> <b>InternVL-2.5-78B</b> - Video 1: $1.2 \times 10^{-6}$ - Video 2: $8.8 \times 10^{-8}$ - $\mathcal{A}_{\text{single}} = 1$	<b>InternVL-2.5-8B</b> - Video 1: $1.5 \times 10^{-3}$ - Video 2: $3.4 \times 10^{-4}$ - $\mathcal{A}_{\text{single}} = 1$
<b>Video Pair Comparison</b> <b>InternVL-2.5-78B</b> - Model Preference: same bad - Accuracy = ✗	<b>InternVL-2.5-8B</b> - Model Preference: same bad - Accuracy = ✗



Table 14: Data example of *Light Change* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

**Video 1:**


**Video 2:**


**Aspect:** Light Change (Dynamic Degree)  
**Human Preference:** Video 1 is better


**Single Video Rating**


<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Video 1: 0.531	- Video 1: 0.201
- Video 2: 0.004	- Video 2: 0.047
- $\mathcal{A}^{\text{single}} = 1$	- $\mathcal{A}^{\text{single}} = 1$

**Video Pair Comparison**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Model Preference: Video 1	- Model Preference: Video 1
- Accuracy = ✓	- Accuracy = ✓

Table 15: Data example of *Technical Quality* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

**Video 1:**


**Video 2:**


**Aspect:** Technical Quality (Static Quality)  
**Human Preference:** Video 2 is better

**Single Video Rating**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Video 1: 0.500	- Video 1: 0.651
- Video 2: 0.999	- Video 2: 0.940
- $\mathcal{A}^{\text{single}} = 1$	- $\mathcal{A}^{\text{single}} = 1$

**Video Pair Comparison**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Model Preference: Video 2	- Model Preference: Video 2
- Accuracy = ✓	- Accuracy = ✓



Table 16: Data example of *Aesthetic Quality* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.


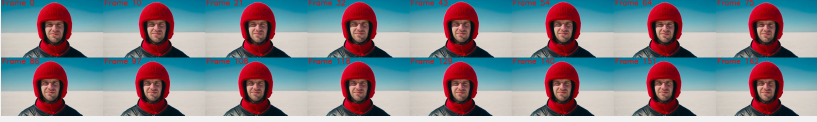
<b>Video 1:</b> 	
<b>Video 2:</b> 	
<b>Aspect:</b> Aesthetic Quality (Static Quality) <b>Human Preference:</b> same good	
<b>Single Video Rating</b>	
<b>InternVL-2.5-78B</b> - Video 1: 0.986 - Video 2: 0.991 - $\mathcal{A}_{\text{single}} = 1$	<b>InternVL-2.5-8B</b> - Video 1: 0.731 - Video 2: 0.755 - $\mathcal{A}_{\text{single}} = 0.319$
<b>Video Pair Comparison</b>	
<b>InternVL-2.5-78B</b> - Model Preference: same good - Accuracy = ✓	<b>InternVL-2.5-8B</b> - Model Preference: Video 2 - Accuracy = ✗

Table 17: Data example of *Structural Correctness* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.



<b>Video 1:</b> 	
<b>Video 2:</b> 	
<b>Aspect:</b> Structural Correctness (Static Quality) <b>Human Preference:</b> same bad	
<b>Single Video Rating</b>	
<b>InternVL-2.5-78B</b> - Video 1: 0.999 - Video 2: 0.971 - $\mathcal{A}_{\text{single}} = 8.6 \times 10^{-6}$	<b>InternVL-2.5-8B</b> - Video 1: 0.971 - Video 2: 0.915 - $\mathcal{A}_{\text{single}} = 1.9 \times 10^{-5}$
<b>Video Pair Comparison</b>	
<b>InternVL-2.5-78B</b> - Model Preference: Video 1 - Accuracy = ✗	<b>InternVL-2.5-8B</b> - Model Preference: Video 1 - Accuracy = ✗

Table 18: Data example of *Appearance Consistency* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

**Video 1:**

**Video 2:**

**Aspect:** Appearance Consistency (Static Quality)  
**Human Preference:** Video 1 is better

**Single Video Rating**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Video 1: 0.990	- Video 1: 0.971
- Video 2: 0.893	- Video 2: 0.980
- $\mathcal{A}^{\text{single}} = 1$	- $\mathcal{A}^{\text{single}} = 0$

**Video Pair Comparison**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Model Preference: Video 1	- Model Preference: Video 2
- Accuracy = ✓	- Accuracy = ✗

Table 19: Data example of *Flickering* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

**Video 1:**

**Video 2:**

**Aspect:** Temporal Flickering (Temporal Quality)  
**Human Preference:** Video 2 is better

**Single Video Rating**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Video 1: 0.852	- Video 1: 0.731
- Video 2: 0.998	- Video 2: 0.905
- $\mathcal{A}^{\text{single}} = 1$	- $\mathcal{A}^{\text{single}} = 1$

**Video Pair Comparison**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Model Preference: Video 2	- Model Preference: Video 2
- Accuracy = ✓	- Accuracy = ✓

Table 20: Data example of *Motion Naturalness* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

**Video 1:**

**Video 2:**

**Aspect:** Motion Naturalness (Temporal Quality)  
**Human Preference:** Video 1 is better

**Single Video Rating**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Video 1: 0.924	- Video 1: 0.731
- Video 2: 0.986	- Video 2: 0.798
- $\mathcal{A}^{\text{single}} = 0$	- $\mathcal{A}^{\text{single}} = 0$

**Video Pair Comparison**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Model Preference: Video 2	- Model Preference: Video 2
- Accuracy = ✗	- Accuracy = ✗

Table 21: Data example of *Appearance Fine* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

**Video 1:**

**Video 2:**

**Aspect:** Appearance Alignment (Video-Text Alignment)  
**Human Preference:** Video 1 is better


**Single Video Rating**


<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Video 1: 0.999	- Video 1: 0.984
- Video 2: 0.009	- Video 2: 0.755
- $\mathcal{A}^{\text{single}} = 1$	- $\mathcal{A}^{\text{single}} = 1$

**Video Pair Comparison**

<b>InternVL-2.5-78B</b>	<b>InternVL-2.5-8B</b>
- Model Preference: Video 1	- Model Preference: Video 1
- Accuracy = ✓	- Accuracy = ✓

Table 22: Data example of *Motion Fine* and evaluations by InternVL-2.5-78B-MPO and InternVL-2.5-8B-MPO.

**Video 1:**


**Video 2:**


**Aspect:** Motion Alignment (Video-Text Alignment)  
**Human Preference:** Video 1 is better

**Single Video Rating**

InternVL-2.5-78B	InternVL-2.5-8B
- Video 1: 0.269	- Video 1: 0.076
- Video 2: 0.245	- Video 2: 0.268
- $\mathcal{A}^{\text{single}} = 1$	- $\mathcal{A}^{\text{single}} = 0$

**Video Pair Comparison**

InternVL-2.5-78B	InternVL-2.5-8B
- Model Preference: same good	- Model Preference: Video 2
- Accuracy = ✗	- Accuracy = ✗