# AIGVE-MACS: Unified Multi-Aspect Commenting and Scoring Model for AI-Generated Video Evaluation

Anonymous CVPR submission

Paper ID

## Abstract

*The rapid advancement of AI-generated video models has created a pressing need for robust and interpretable evaluation frameworks. Existing metrics are limited to producing numerical scores without explanatory comments, resulting in low interpretability and human evaluation alignment. To address those challenges, we introduce* AIGVE-MACS*, a unified model for AI-Generated Video Evaluation(AIGVE), which can provide not only numerical scores but also multi-aspect language comment feedbacks in evaluating these generated videos. Central to our approach is* AIGVE-BENCH 2*, a large-scale benchmark comprising 2,500 AI-generated videos and 22,500 human-annotated detailed comments and numerical scores across nine critical evaluation aspects. Leveraging* AIGVE-BENCH 2*,* AIGVE-MACS *incorporates recent Vision-Language Models with a novel token-wise weighted loss and a dynamic frame sampling strategy to better align with human evaluators. Comprehensive experiments across supervised and zero-shot benchmarks demonstrate that* AIGVE-MACS *achieves state-of-the-art performance in both scoring correlation and comment quality, significantly outperforming prior baselines including GPT-4o and VideoScore. In addition, we further showcase a multi-agent refinement framework where feedback from* AIGVE-MACS *drives iterative improvements in video generation, leading to* $53.5\%$ *quality enhancement. This work establishes a new paradigm for comprehensive, human-aligned evaluation of AI-generated videos.*

## 1. Introduction

With the rapid advancement of video generation models such as Sora [35], HunyuanVideo [17], and Mochi-1 [38], AI-generated videos are becoming increasingly photorealistic and temporally coherent, significantly narrowing the gap between synthetic and real-world visual content. Following provided textual instruction prompts, AI-generated videos are gaining widespread adoption across domains such as entertainment, advertising, education, and virtual reality, offering cost-effective solutions and enabling more personalized content experiences for both creators and consumers [41, 51].

Despite significant progress, AI-generated videos still face persistent challenges such as limited spatial resolution, object distortions, and misalignment with user instructions [11, 22, 42]. These issues highlight the growing importance of rigorous evaluation, which plays a critical role in guiding the development and refinement of video generation models. However, research on evaluation methodologies has lagged behind the rapid advances in video generation itself, leaving a gap in systematic, interpretable, and comprehensive evaluation techniques [28].

Existing evaluation methods for AI-generated videos exhibit several key limitations. First, they often rely on outdated numerical metrics such as FVD [39] and IS [4], which are insufficient for capturing the nuanced qualities of modern AI-generated content. Effective AI-Generated Video Evaluation (AIGVE) must account not only for intrinsic video quality but also for alignment with user-provided instructions [28, 46]. Therefore, evaluating video quality with one single scalar metric is inherently limiting. While recent work has made progress by decomposing quality into aspects such as technical quality, motion dynamics, and text-to-video alignment [3, 7, 11], many approaches still rely on aggregating legacy metrics not originally designed for this task [14, 29]. These metrics often overlap in what they measure or fail to address key dimensions altogether, resulting in evaluation pipelines that are fragmented, difficult to interpret, and lacking comprehensive coverage.

Second, although recent studies have explored the use of unified Vision Language Models (VLMs) for evaluating AI-generated videos, they often fail to fully leverage the generative strengths of these models. This shortcoming stems largely from the difficulty of reliably prompting or finetuning VLMs to produce accurate, human-aligned evaluative outputs. Rather than generating natural language assessments directly, many existing approaches either attach a scoring head to predict numerical values from the

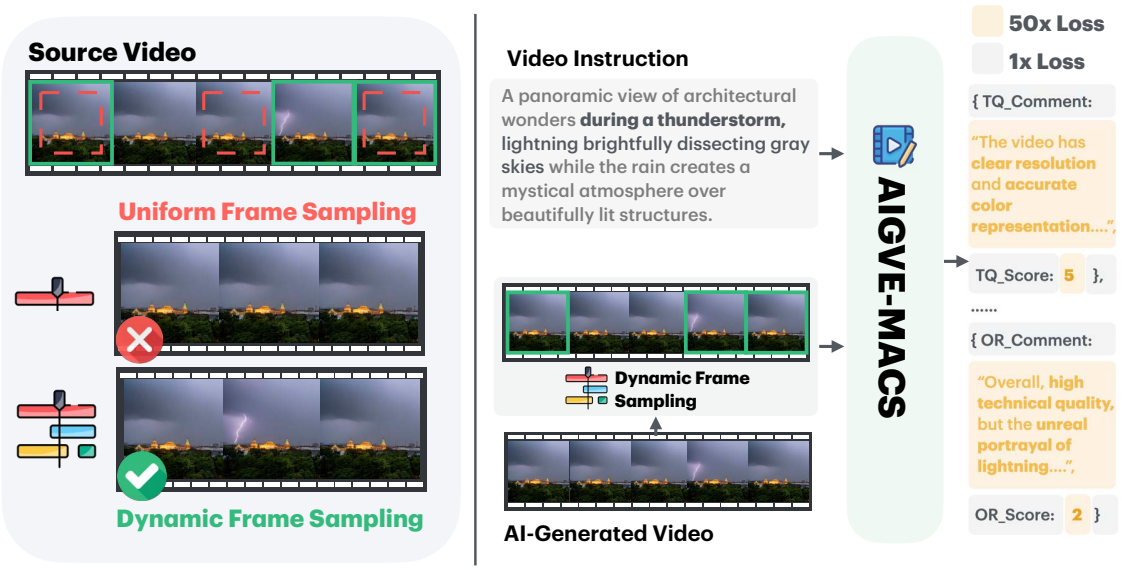


Figure 1. AIGVE-MACS Pipeline. The left side of the diagram illustrates our Dynamic Frame Sampling strategy, which selects key moments based on content variation to better capture temporal dynamics. The right side highlights the Token-Wise Weighted Loss, which emphasizes score and comment tokens to improve alignment with human evaluations. TQ and OR refer to Technical Quality and Overall, respectively. Tokens in the yellow box are assigned a loss weight of 50, while those in the gray box are assigned a loss weight of 1.

VLM's hidden states, thereby bypassing its language generation capabilities [10], or reformulate evaluation as a series of binary yes/no questions [7, 49]. These strategies reduce the VLM to a classification tool, limiting its ability to provide nuanced, context-aware evaluations. Furthermore, the question-based approach relies heavily on handcrafted prompts, which struggle to capture the complexity and diversity of real-world video quality dimensions. Besides, due to the constrained context length of current VLMs, it is impossible to input the full token sequence of an entire video. Instead, existing methods typically sample a fixed number of frames uniformly across the video, which can overlook brief but critical dynamic changes.

Third, due to the absence of datasets that pair aspect-wise scores with explanatory comments, existing video evaluation approaches primarily focus on producing numerical scores [8, 10, 19, 32] while neglecting the generation of human-like comments that offer richer, more actionable feedback. However, we argue that such comments are actually essential for understanding the specific strengths and weaknesses of AI-generated videos. By providing detailed qualitative insights, explanatory comments can guide model improvement more effectively than scalar scores alone, which will also be demonstrated with showcases to be provided in this paper.

To address the aforementioned challenges, we first introduce and release the AIGVE-BENCH 2, a large-scale, human-annotated benchmark comprising 500 diverse prompts, 2,500 videos generated by five state-of-the-art video generation models, and 22,500 high-quality human score and comment annotations spanning nine carefully designed evaluation aspects. The prompts, videos, scores, and comments in AIGVE-BENCH 2 are diligently constructed and rigorously validated to ensure comprehensive coverage of diverse video generation scenarios, making the dataset both robust and representative of real-world applications.

Building on this benchmark, we propose to finetune Qwen2.5-VL-7B [2] and present AIGVE-MACS, a unified evaluation model that provides both numerical score and natural language comment feedback evaluations across the nine critical aspects as covered in the AIGVE-BENCH 2. The overall pipeline of the proposed AIGVE-MACS is also illustrated in Figure 1. By introducing a dynamic frame sampling strategy as well as a weighted loss on both comment and score tokens, AIGVE-MACS achieves accurate numerical predictions while fully leveraging the generative capabilities of VLMs to generate language comment to justify those scores. This joint modeling approach enhances the transparency, interpretability, and practical utility of automated video evaluation. Extensive experiments and ablation studies show that AIGVE-MACS achieves state-of-the-art performance in both supervised and zero-shot settings on AIGVE-BENCH 2 and multiple evaluation benchmarks, significantly outperforming existing methods in both numerical scoring and comment generation.

What's more, to further demonstrate the practical utility of our model, we introduce a multi-agent iterative refinement framework that leverages the generated scores and

comments to progressively improve video quality. Experimental results indicate that the quality of generated videos can be enhanced by 53.5% through this iterative process. This framework underscores the real-world applicability of AIGVE-MACS, illustrating how continuous feedback loops can drive meaningful improvements in AI-generated content. To the best of our knowledge, this is the first work to jointly produce aspect-wise scores and natural language comments for AI-generated videos, providing a more holistic evaluation framework that aligns closely with human judgments.

In summary, our contributions are as follows:

- We construct AIGVE-BENCH 2, the first large-scale AIGVE benchmark covering nine evaluation aspects, with 22,500 human-annotated numerical scores and explanatory comments for 2,500 videos generated by several SOTA video generation models.
- We propose AIGVE-MACS, a unified AIGVE model that jointly predicts aspect-wise scores and comments. Trained with dynamic frame sampling and weighted losses, AIGVE-MACS achieves state-of-the-art performance on AIGVE-BENCH 2 and other benchmarks in both supervised and zero-shot settings, aligning well with human judgments.
- We explore practical applications of our model through a multi-agent iterative refinement framework that uses generated scores and comments to enhance video quality.

## 2. Related Works

### 2.1. AI-Generated Video Evaluation

The field of AI-Generated Video Evaluation (AIGVE) is in its early stages and includes even more challenges than video quality assessment. Previous research proposes that evaluating AI-generated videos requires alignment with both human perception and instructions to ensure high-quality video generation that meets creators' intentions and viewers' expectations [28]. Alignment with human perception focuses on evaluating video quality through traditional metrics like resolution and clarity, while ensuring consistency with physical world properties such as realistic textures and adherence to physical laws. Meanwhile, alignment with human instructions emphasizes how well videos mirror the detailed scenarios, actions, and narratives specified in text descriptions, ensuring the generated content fulfills creators' creative and communicative objectives. These aspects have evolved separately, with perception alignment progressing from statistical approaches [4, 39] to advanced architectures like DOVER [45] for aesthetic assessment and BVQI [44] for CLIP-based evaluation, while instruction alignment developed methods like TIFA [13] and CLIP-Score [12] for evaluating text-visual consistency.

Recent comprehensive frameworks attempt to unify these aspects. VBench [14] evaluates across 16 dimensions, and EvalCrafter [29] assesses four key aspects including visual quality and text-video alignment. VIDEOSCORE [10] leverages large language models for holistic assessment. However, these methods either rely on collections of traditional metrics or provide single scores without interpretable feedback, highlighting the need for more sophisticated evaluation approaches.

## 3. AIGVE-BENCH 2

Our AIGVE-BENCH 2 extends the previous benchmark dataset AIGVE-BENCH [46], which contains multi-aspect numerical scores for AI-generated videos across nine evaluation dimensions. We introduce a rigorous comment process pipeline to enrich AIGVE-BENCH with human-written, aspect-specific justifications. This pipeline ensures that each textual explanation is directly aligned with the corresponding multi-aspect numerical evaluation scores, offering a more comprehensive and interpretable understanding of video quality. The remainder of this section will first present he key characteristics of of AIGVE-BENCH and then detail our comment processing pipeline.

### 3.1. Previous Benchmark: AIGVE-BENCH

AIGVE-BENCH [46] is a comprehensive AIGVE benchmark dataset contains 2,500 AI-generated videos created from 500 diverse prompts, covering a broad range of real-world scenes and interactions. Each prompt falls into one of two categories: *global view*, which focuses on environmental and wide-angle scenes with natural dynamics such as weather changes, camera motion, and *close shot*, which captures fine-grained interactions between humans, animals, plants, or objects. Videos are generated by five state-of-the-art models: Pyramid [15], CogVideoX [47], Genmo [38], Hunyuan [17], and Sora [35], with varying resolutions and frame rates, and are standardized to a uniform duration of 5 seconds to ensure sufficient temporal depth and comparability.

Each video is evaluated by expert raters across nine distinct evaluation aspects: Technical Quality, Dynamics, Consistency, Physics, Element Presence, Element Quality, Action/Interaction Presence, Action/Interaction Quality, and Overall.

### 3.2. Comment Processing Pipeline

Although justification comments are collected in parallel with the scoring process, they present several challenges that limit their direct usability. First, since the comments are written by multiple annotators, they exhibit inconsistencies in writing style, terminology, and level of detail, introducing noise and increasing the complexity of model training. Second, comments may not fully align with scores. As shown in Figure 2, minor issues may be emphasized without noting

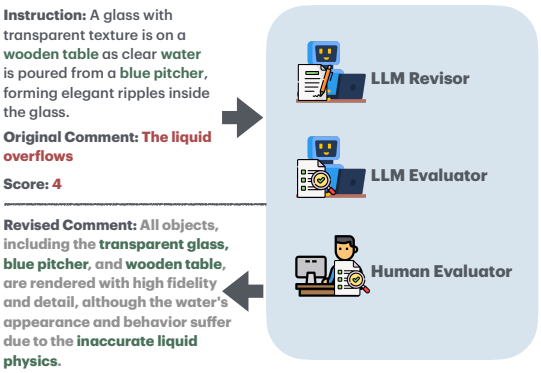


Figure 2. Overview of the comment processing pipeline for AIGVE-BENCH 2. The pipeline includes a revisor, validator, and human evaluator to ensure high-quality, consistent, and objective comments.

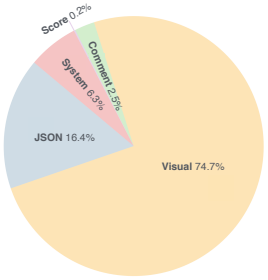


Figure 3. Averaged ratios of different types of input tokens in the input sequence. The five types of tokens are: (1) system prompt tokens, (2) visual tokens, (3) JSON structure tokens, (4) comment tokens, and (5) numerical score tokens.

strengths, leading to insufficient justification for high scores and reducing their value for training or explanation.

To address these challenges, we design a robust comment processing pipeline that revises the original comments to ensure consistency, objectivity, and alignment with the corresponding scores. As shown in Figure 2, we first employ a Large Language Model (LLM) as a revisor to refine and extend the original comments, conditioned on the video instructions and associated scores. The LLM is prompted to produce comments with a consistent and objective writing style, providing a comprehensive explanation that covers both the strengths and weaknesses relevant to the evaluation aspect.

Next, the revised comment is passed to a second LLM, which we refer to as the evaluator. This model checks whether the revised comment introduces any content that are not presented in the original comment or instruction, mitigating the risk of hallucination. Finally, a human evaluator manually reviews and, if necessary, rewrites the revised comments to ensure clarity, factual accuracy, and faithful reflection of the corresponding scores. Through this meticulous process, we extend each aspect-level score in AIGVE-BENCH with a high-quality justification comment, resulting in a total of 22,500 comments with an average length of 267 words.

## 4. AIGVE-MACS

To build the AIGVE-MACS framework, we propose to fine-tune the VLMs with our AIGVE-BENCH 2, enabling VLMs to generate human-aligned evaluation scores and explanatory comments. Specifically, given an AI-generated video, we provide the model with the video instruction and video content. The VLM is trained to generate a JSON-formatted output containing both the aspect-specific score and the corresponding natural language comment.

Formally, given an AI-generated video $v$, the video's instruction $i$, the VLM is trained to generate a structured output:

$$y = \text{VLM}(i, v|\theta), \tag{1}$$

where $y = \{C, S\}$ denotes the JSON-formatted output containing aspect-specific evaluation comments $C$ and scores $S$, $\theta$ is the learnable parameter of the VLM.

However, as discussed in Section 1, existing methods are facing two major challenges. First, they struggle to effectively leverage the native language generation capabilities of VLMs to produce human-aligned evaluation scores and comments. Second, the commonly adopted uniform frame sampling strategy may overlook brief yet critical dynamics, leading to incorrect evaluation. To address these issues, we first introduce a token-wise weighted loss in Section 4.1, which encourages the model to focus more on generating informative comments and accurate scores by highlighting these tokens during training. Additionally, in Section 4.2, we propose a dynamic frame sampling strategy to better capture key moments in the video.

### 4.1. Token-wise Weighted Loss

To better understand the challenge of leveraging language generation ability, we analyze the input tokens used by VLMs in our task. Specifically, there are five types of input tokens: (1) system prompt tokens that describe the task, (2) visual tokens that encode the video content, (3) JSON structure tokens that maintain the output format, (4) comment tokens that contain user comments, and (5) score tokens that indicate the evaluation scores. Figure 3 illustrates the averaged ratios of each type of input token in the input. We observe that the system prompt tokens, viusal tokens, and JSON structure tokens, which are not directly related to the task, occupy a significant portion of the input, while the task-related comment tokens and score tokens are relatively sparse. This unbalanced distribution of input tokens will lead to the VLMs focusing more on the system prompt,

visual content, and output format, while neglecting the user comments and scores.

To address this issue, we propose a simple yet effective token-wise weighted loss that encourages the model to focus on accurately generate comment and score tokens. During tokenization, we identify comment and score tokens and assign them a higher weight $\alpha$ in the loss function. The weighted loss is defined as follows:

$$\mathcal{L} = -\sum_{t=1}^{T} w_t \cdot \log p(y_t \mid y_{<t}, i, v), \tag{2}$$

where $T$ is the total number of tokens in the output sequence, $y_t$ is the target token at position $t$, $x$ is the input sequence, and $p(y_t \mid y_{<t}, i, v)$ is the predicted probability of token $y_t$. The token weight $w_t$ is defined as:

$$w_t = \begin{cases} \alpha, & \text{if } y_t \text{ is a comment or score token,} \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

Besides, inspired by reasoning schema, we put the comment content prior to the numerical score in training data to encouage the model to build stronger connections between the comment and the score to enhance the comment-score alignment. The process is shown on the right part of Figure 1.

### 4.2. Dynamic Frame Sampling

To avoid missing the key moments in the video, we propose a dynamic frame sampling strategy that selects frames based on content variation instead of uniform sampling. As shown at the bottom-left of Figure 1, this approach allows the evaluation model to focus on the most informative frames, leading to more accurate and aspect-aware assessments.

To select representative frames based on content changes, we compute a frame-wise difference score as follows:

$$\Delta(f^t, f^{t-1}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbb{I}\left(\left|f_{i,j}^t - f_{i,j}^{t-1}\right| > 0\right) \tag{4}$$

where $f^t \in \mathbb{R}^{H \times W}$ denotes the grayscale frame with width $W$ and height $H$ at time $t$, and $\mathbb{I}(\cdot)$ is the indicator function. A frame $f^t$ is selected if it satisfies both:

$$\Delta(f^t, f^{t-1}) > \theta \quad \text{and} \quad t - t_{\text{last}} \geq \gamma, \tag{5}$$

where $\theta$ is a content change threshold and $\gamma$ is the minimum frame gap to prevent redundant selection. If the number of selected frames exceeds the target number $N$, we uniformly subsample them. If no frames meet the criteria, we fallback to uniform sampling.

## 5. Experiment

### 5.1. Benchmark Datasets and Evaluation Metrics

We evaluate AIGVE-MACS on four benchmarks: a supervised dataset, AIGVE-BENCH 2-TEST, and three zero-shot datasets—VIDEOFEEDBACK-Test [10], GenAI-Bench [23], and VBench [14], using the same test samples as VideoScore [10]. For AIGVE-BENCH 2-TEST, we report Spearman's $\rho$ for score prediction and use ROUGE-1, ROUGE-L [25], UniEval-Fact [52], BERTScore [50], and G-Eval [30] for comment evaluation. VIDEOFEEDBACK-Test uses Spearman's $\rho$ over five aspects; GenAI-Bench and VBench reports pairwise accuracy inferred from predicted scores.

### 5.2. Baseline Methods

To benchmark the performance of AIGVE-MACS, we compare it against a comprehensive suite of existing video evaluation methods, organized into three main categories:

**Feature-based Metrics** These methods evaluate videos by extracting handcrafted or pretrained model features. This category includes traditional quality metrics such as PIQE [40] and BRISQUE [33], perceptual similarity-based scores like SSIM-sim [43], SSIM-dyn [43], MSE(Mean Square Error)-dyn, and DINO-sim [5], as well as text-video alignment metrics such as CLIP-sim [18], CLIP-temp [29], CLIP-Score [12], X-CLIP-Score [31], BLIP-sim [24], and PickScore [16].

**Modeling-based Metrics** This group includes supervised deep learning models trained specifically to predict video quality. We include LightVQA+ [53], GSTVQA [6], and SimpleVQA [9], which are adapted for multi-aspect video assessment.

**VLM-based Metrics** These approaches leverage VLMs to generate scores or conduct evaluations via either prompting or finetuning. Baselines include foundation models such as Qwen2.5-VL [1], VideoLLaMA3-7B [48], GPT-4.1 [36], GPT-4o [34], LLaVA-1.5-7B [26], LLaVA-1.6-7B [27], Gemini-1.5-Flash [37], Gemini-1.5-Pro [37], Idefics1 [20], and Idefics2 [21]. In addition, we consider VLM-based evaluation frameworks like VideoScore [10], and TIFA [13].

### 5.3. Implementation Details

We implement AIGVE-MACS based on the Qwen2.5-VL-7B [2] architecture, initializing from the official checkpoint[1]. During finetuning, the vision encoder is kept frozen,

[1] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

| Method | Technical | Dynamic | Consistency | Physics | Element Pre | Element Qu | Act Pre | Act Qu | Overall | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| Random | -6.71 | 4.28 | -3.55 | -1.62 | -3.03 | 1.42 | -5.16 | 1.44 | -3.19 | -1.79 |
| **Feature-based Metrics** | | | | | | | | | | |
| CLIP-sim | 9.12 | 6.54 | 5.79 | -4.45 | 19.34 | 7.93 | 23.85 | 8.25 | 21.58 | 10.88 |
| BLIP-sim | 12.02 | 10.92 | 12.24 | 3.69 | 22.80 | 9.42 | 17.02 | 10.34 | 19.09 | 13.06 |
| CLIP-temp | 16.46 | 4.26 | 27.69 | 23.51 | 11.07 | 25.12 | -2.62 | 20.20 | 16.63 | 15.81 |
| PickScore | 22.48 | 5.90 | 11.36 | 6.55 | 26.25 | 17.69 | 20.37 | 13.24 | 24.28 | 16.46 |
| **Modeling-based Metrics** | | | | | | | | | | |
| LightVQA+ | -3.68 | -7.58 | -8.08 | -10.73 | -0.16 | -6.33 | 4.77 | -7.40 | -4.70 | -4.88 |
| GSTVQA | 17.92 | 13.40 | 15.97 | 1.23 | -2.65 | 15.91 | 9.81 | 9.68 | 20.71 | 11.33 |
| SimpleVQA | 24.50 | 11.50 | 16.58 | 0.28 | 2.41 | 18.30 | 7.76 | 3.58 | 21.22 | 11.79 |
| **VLM-based Metrics** | | | | | | | | | | |
| Qwen2.5-VL | 8.77 | 4.00 | 1.24 | -6.01 | 9.19 | 10.19 | 18.74 | 0.72 | 9.59 | 6.27 |
| VideoLLaMA3 | 15.94 | 19.44 | 11.70 | 13.21 | -3.13 | 12.27 | 13.61 | -0.69 | 11.58 | 10.44 |
| GPT-4.1 | 36.49 | 5.81 | 26.68 | 19.87 | 27.22 | 28.77 | 32.75 | 20.22 | 29.98 | 25.31 |
| GPT-4o | 34.71 | 7.05 | 18.12 | 20.28 | 23.10 | 30.47 | 36.57 | 31.58 | 38.57 | 26.72 |
| Videoscore | -9.50 | -8.20 | -0.20 | 20.10 | 9.70 | -7.50 | -3.10 | -0.60 | -7.30 | -0.73 |
| VideoPhy | 0.10 | 4.00 | -1.40 | -5.60 | 0.80 | -1.30 | 11.90 | 2.70 | 9.70 | 2.32 |
| TIFA | 17.81 | 8.83 | 9.62 | 3.85 | 16.67 | 12.41 | 17.90 | 5.87 | 17.78 | 12.30 |
| AIGVE-MACS | **40.60** | **57.31** | **61.49** | **64.36** | **40.32** | **40.81** | **44.31** | **60.71** | **59.88** | **52.20** |

Table 1. Scoring Correlation Evaluation Result on AIGVE-BENCH 2. The underlined score represents the best zero-shot model. Element Pre and Element Qu represents Element Presence and Quality. Act Pre and Qu represents Action Presence and Quality.

| Method | GenAI-Bench (Acc) | VBench (Acc) | | | | | | VideoFeedback ($\rho$) |
|---|---|---|---|---|---|---|---|---|
| | | Technical Quality | Subject Consistency | Dynamic Degree | Motion Smoothness | Overall Consistency | Average | |
| Random | 37.7 | 44.5 | 42.0 | 37.3 | 40.5 | 44.8 | 41.82 | 0.4 |
| **Feature-based Metrics** | | | | | | | | |
| PIQE | 34.5 | 60.8 | 44.3 | 71.0 | 45.3 | 53.8 | 55.04 | -10.1 |
| BRISQUE | 38.5 | 56.7 | 41.2 | 75.5 | 41.2 | 54.2 | 53.76 | -20.3 |
| CLIP-sim | 34.1 | 47.8 | 46.0 | 34.8 | 44.7 | 44.2 | 43.50 | 8.9 |
| DINO-sim | 31.4 | 49.5 | 51.2 | 24.7 | 55.5 | 41.7 | 44.52 | 7.5 |
| SSIM-sim | 28.4 | 30.7 | 46.2 | 24.5 | 54.2 | 27.2 | 36.56 | 13.4 |
| MSE-dyn | 34.2 | 32.8 | 31.7 | 81.7 | 31.2 | 39.2 | 43.32 | -5.5 |
| SSIM-dyn | 38.5 | 37.5 | 36.3 | 84.2 | 34.7 | 44.5 | 47.44 | -12.9 |
| CLIP-Score | 45.0 | 57.8 | 46.3 | 71.3 | 47.0 | 52.2 | 54.92 | -7.2 |
| X-CLIP-Score | 41.4 | 44.0 | 38.0 | 51.0 | 28.7 | 39.0 | 40.14 | -1.9 |
| **VLM-based Metrics** | | | | | | | | |
| LLaVA-1.5 | 49.9 | 42.7 | 42.3 | 63.8 | 41.3 | 8.8 | 39.78 | 8.5 |
| LLaVA-1.6 | 44.5 | 38.7 | 26.8 | 56.5 | 28.5 | 43.2 | 38.74 | -3.1 |
| Idefics1 / 2 | 34.6 | 20.7 | 22.7 | 54.0 | 27.3 | 33.7 | 31.68 | 6.5 |
| Gemini-1.5-Flash | 67.1 | 52.3 | 49.2 | 64.5 | 45.5 | 49.9 | 52.28 | 20.8 |
| Gemini-1.5-Pro | 60.9 | 56.7 | 43.3 | 65.2 | 43.0 | 56.3 | 52.90 | 16.9 |
| GPT-4o | 52.0 | 59.3 | 49.3 | 46.8 | 42.0 | 60.8 | 51.64 | 23.0 |
| VideoScore | 59.0 | 64.2 | 57.7 | 55.5 | 54.3 | 61.5 | 58.64 | 77.1 |
| AIGVE-MACS (Ours) | **68.5** | **65.5** | **59.0** | **74.5** | **55.8** | **55.6** | **62.08** | **24.6** |

Table 2. Evaluation results on GenAI-Bench, VBench, and VideoFeedback. GenAI and VBench report pairwise accuracy (%), while VideoFeedback uses Spearman's $\rho$. The VBench section spans metrics from Technical Quality to Average. Underlined scores indicate best zero-shot performance.

while all remaining parameters are updated. We set the weight for score and comment to 50. Training is conducted on two NVIDIA A6000 GPUs using a learning rate of 1e-5. We train the model for 3 epochs and apply early stopping based on validation loss. Optimization is performed using AdamW, with a linear learning rate scheduler and 5% warmup steps. The full training process takes approximately 24 hours.

### 5.4. Results

**AIGVE-MACS achieves state-of-the-art alignment with human judgments, significantly outperforming all existing baselines in both score prediction and comment generation.** Table 1 and Table 3 summarize AIGVE-MACS's performance on AIGVE-BENCH 2 compared to a broad range of evaluation baselines. AIGVE-MACS achieves the highest correlation across all nine evaluation aspects, with an average Spearman's $\rho$ of 52.20%, which nearly doubling the performance of the strongest VLM baseline, GPT-4o, and achieving up to 61% improvement over its pretrained

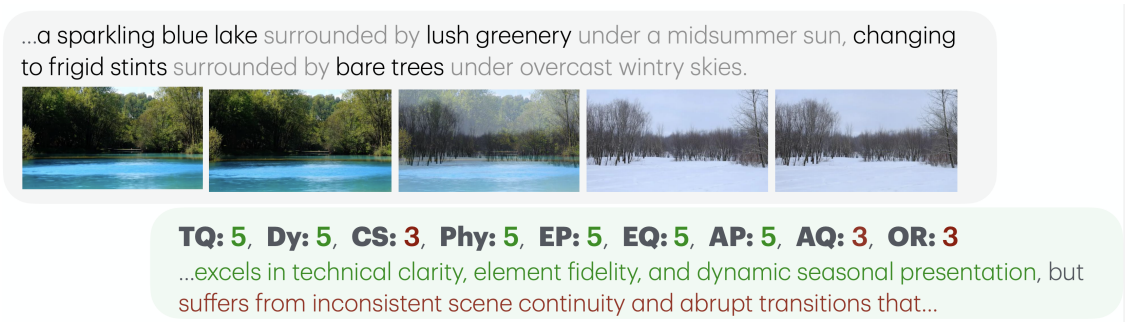


Figure 4. Case Study of AIGVE-MACS. This example showcases AIGVE-MACS's ability to produce human-aligned evaluations and distinguish fine-grained quality differences in AI-generated videos.

| | R1 | RL | BS | UF | GE |
|---|---|---|---|---|---|
| GPT-4o | 18.30 | 15.86 | 74.90 | 40.84 | 2.10 |
| GPT-4.1 | 15.80 | 12.94 | 73.99 | 43.99 | 2.10 |
| QWen2.5-VL | 17.95 | 15.31 | 74.31 | 42.32 | 2.37 |
| VideoLLama | 19.99 | 17.67 | 75.35 | 40.21 | 2.18 |
| AIGVE-MACS | **49.50** | **38.00** | **85.87** | **57.04** | **3.42** |

Table 3. Evaluation Results for Comments. R1, RL, BS, UF, and GE represent ROUGE-1, ROUGE-L, BERTcore, UniEval-Fact, and G-Eval respectively.

backbone, Qwen2.5-VL-7B [2].

In terms of comment generation, as shown in Table 3, AIGVE-MACS also consistently outperforms all baselines across standard automatic metrics. These results highlight a fundamental limitation of current VLMs, which struggle to generate reliable, human-aligned evaluations. In contrast, AIGVE-MACS effectively learns to generate both accurate scores and rich, faithful commentary that aligns closely with human judgment.

**AIGVE-MACS establishes new state-of-the-art performance across all three benchmarks, demonstrating strong zero-shot generalization capabilities for diverse video evaluation tasks.** Table 2 reports zero-shot evaluation results on VideoFeedback [11], GenAI-Bench [23], and VBench [14].

On GenAI-Bench, AIGVE-MACS achieves the highest video preference accuracy, outperforming strong commercial VLMs such as Gemini-1.5-Pro and GPT-4o. On VBench, AIGVE-MACS surpasses VideoScore by 4.24%, while using only one-tenth of the training data.

Notably, on the VideoFeedback dataset, AIGVE-MACS also achieves significant improvements over all open-source and commercial baselines, emerging as the best zero-shot model. By comparing the result in Table 1 and Table 2, while AIGVE-MACS generalizes well to VideoFeedback without training on it, VideoScore, which is trained on VideoFeedback, fails to generalize to our AIGVE-BENCH 2-TEST. This contrast highlights AIGVE-MACS's more robust and transferable evaluation capabilities that are not tied to the idiosyncrasies of any single dataset.

The superior supervised and zero-shot performance of AIGVE-MACS validates the effectiveness of our finetuning strategy, which equips VLMs with the ability to evaluate AI-generated videos using both structured scores and natural language comments. It also underscores the high quality and broad coverage of AIGVE-BENCH 2, enabling the training of generalizable video evaluators that can seamlessly transfer to other benchmarks.

## 6. Multi-Agent Iterative Refinement Framework

Our AIGVE-MACS opens new possibilities for leveraging evaluation scores and comments to guide and optimize the video generation process. To initiate this exploration, we propose a multi-agent iterative refinement framework designed to progressively enhance video quality through feedback-driven revisions.

As shown in Figure 5, our framework consists of three main components: a Video Generator, an Instruction Revisor, and AIGVE-MACS. In each iteration, the Video Generator produces a video based on the current user instruction, and then AIGVE-MACS evaluates the generated video and provides feedback in the form of scores and comments. The Instruction Revisor refines the instruction based on the feedback, aiming to clarify or adjust the requirements for the next iteration. This process continues until the video reaches a satisfactory quality level or a maximum of 4 iterations.

To validate the effectiveness of our multi-agent framework, we use HunyuanVideo [17] as the Video Generator, GPT-4.1 [36] as the Instruction Revisor, and AIGVE-MACS as the evaluator. We conduct experiments on a subset of 50 low-quality videos (overall score $< 3$) sampled from AIGVE-BENCH 2-TEST. The iteration continues until the overall score exceeds 4 or the iteration limit of 4 is reached.

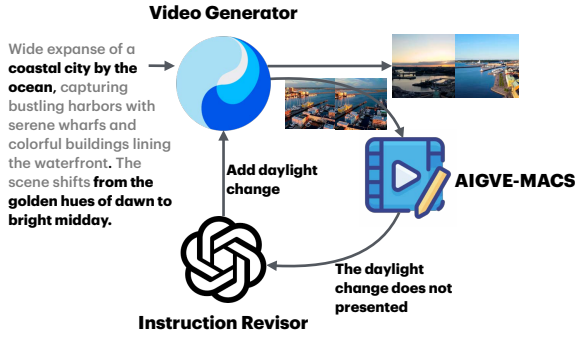


Figure 5. The pipeline of the multi-agent iterative refinement framework.

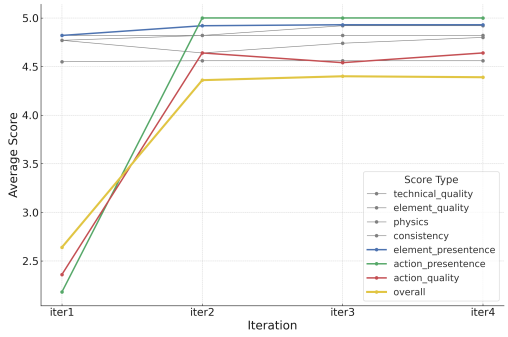


Figure 6. The result of the multi-agent iterative refinement framework.

Figure 6 shows the average scores across refinement iterations. Notably, aspects related to *overall* score and instruction alignment such as element presence, action presence, and action quality, demonstrate significant improvements up to 53%, particularly within the first two iterations.

In contrast, visual aspects like technical quality, element quality, and physics show only minor variation, as they are primarily constrained by the fixed video generator. This suggests that while our framework effectively improves instruction-driven dimensions, visual quality remains bounded by generator capacity.

## 7. Conclusion

We propose AIGVE-MACS, a unified and interpretable evaluation framework that jointly predicts aspect-wise scores and explanatory comments for AI-generated videos. Trained on our human-annotated AIGVE-BENCH 2 dataset, the model achieves state-of-the-art performance in both supervised and zero-shot settings, outperforming strong baselines in score correlation and comment quality across multiple benchmarks. Beyond evaluation, we demonstrate the utility of AIGVE-SCORE in a multi-agent refinement framework, showing its ability to drive meaningful improvements in video generation quality through feedback-driven instruction updates. Our work highlights the value of structured evaluation signals and establishes AIGVE-MACS as a robust and generalizable tool for aligning video generation with human preferences.

## References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023. arXiv:2308.12966 [cs]. 5

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 5, 7

[3] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 1

[4] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 1, 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[6] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1903–1916, 2022. 5

[7] Rui Chen, Lei Sun, Jing Tang, Geng Li, and Xiangxiang Chu. Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. *arXiv preprint arXiv:2504.10358*, 2025. 1, 2

[8] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos, 2024. 2

[9] Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Tongliang Li, Wenhao Huang, and Zhoujun Li. Simplevqa: Multimodal factuality evaluation for multimodal large language models, 2025. 5

[10] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2, 3, 5, 1

[11] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *ArXiv*, abs/2406.15252, 2024. 1, 7

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3, 5

[13] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023. 3, 5

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models, 2023. arXiv:2311.17982 [cs]. 1, 3, 5, 7

[15] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 3

[16] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 5

[17] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 3, 7

[18] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7793–7802, 2024. 5

[19] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-Aligned Dataset and Metric for Text-to-Video Quality Assessment, 2024. arXiv:2403.11956 [cs]. 2

[20] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. 5

[21] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 5

[22] Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. Videorepair: Improving text-to-video generation via misalignment evaluation and localized refinement. *arXiv preprint arXiv:2411.15115*, 2024. 1

[23] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. 5, 7, 1

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 5

[25] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 5, 1

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 5

[27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5

[28] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. A survey of ai-generated video evaluation. *arXiv preprint arXiv:2410.19884*, 2024. 1, 3

[29] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models, 2023. arXiv:2310.11440 [cs]. 1, 3, 5

[30] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. 5, 1

[31] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022. 5

[32] Yibo Miao, Yifan Zhu, Yinpeng Dong, Lijia Yu, Jun Zhu, and Xiao-Shan Gao. T2vsafetybench: Evaluating the safety of text-to-video generative models, 2024. 2

[33] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 5

[34] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 5

[35] OpenAI. Sora. https://openai.com/index/sora/, 2024. 1, 3

[36] OpenAI. Gpt-4.1. https://openai.com/index/gpt-4-1/, 2025. 5, 7

[37] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 5

[38] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024. 1, 3

[39] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges, 2019. 1, 3

[40] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE, 2015. 5

[41] Shreyas Verma and John Leddo. Comparing the effectiveness between human-generated videos and ai-generated videos on learning. 1

[42] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap attention in spatiotemporal diffusions for text-to-video generation. *International Journal of Computer Vision*, pages 1–19, 2025. 1

[43] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 5

[44] Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment, 2023. 3

[45] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives, 2023. arXiv:2211.04894 [cs, eess]. 3

[46] Xinhao Xiang, Xiao Liu, Zizhong Li, Zhuosheng Liu, and Jiawei Zhang. Aigve-tool: Ai-generated video evaluation toolkit with multifaceted benchmark. *arXiv preprint arXiv:2503.14064*, 2025. 1, 3

[47] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[48] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025. 5

[49] Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv preprint arXiv:2412.09645*, 2024. 2

[50] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. 5, 1

[51] Xinyi Zhang, Renyu Zhang, K. Goh, and Chenshuo Sun. The value of ai-generated metadata for ugc platforms: Evidence from a large-scale field experiment. *ArXiv*, abs/2412.18337, 2024. 1

[52] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation, 2022. 5, 1

[53] Xunchu Zhou, Xiaohong Liu, Yunlong Dong, Tengchuan Kou, Yixuan Gao, Zicheng Zhang, Chunyi Li, Haoning Wu, and Guangtao Zhai. Light-vqa+: A video quality assessment model for exposure correction with vision-language guidance, 2024. 5