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

Supplementary Material

8. Details of Benchmark Datasets

To comprehensively evaluate the performance of AIGVE-MACS, we conduct experiments on four benchmark datasets: one supervised dataset, AIGVE-BENCH 2-TEST, and three zero-shot datasets: VIDEOFEEDBACK-Test [10], GenAI-Bench [23], and VBench [14]. For fair comparison, we use the same test samples as those used in VideoScore [10] for the zero-shot settings. The details of each dataset are as follows:

AIGVE-BENCH 2-TEST We perform stratified sampling to select 200 examples from the test split of AIGVE-BENCH 2, ensuring a balanced distribution across the *global-view* and *close-shot* categories, and comprehensive coverage of all motion types and object categories. This approach preserves evaluation efficiency while maintaining diversity and representativeness.

For score evaluation, we compute Spearman's rank correlation coefficient (ρ) between the predicted scores and ground-truth scores across all nine evaluation aspects. To assess the quality of generated comments, we adopt a suite of metrics: ROUGE-1 and ROUGE-L [25] for information coverage, UniEval-Fact [52] for faithfulness, BERTScore [50] for semantic similarity, and G-Eval [30] to assess overall comment quality.

VideoFeedback-Test VideoFeedback [10] decomposes AI-generated video quality into five evaluation aspects: Visual Quality, Temporal Consistency, Dynamic Degree, Text Alignment, and Factual Consistency. We report Spearman's ρ between predicted and human-annotated scores for each aspect.

GenAI-Bench GenAI-Bench [23] is derived from GenAI-Arena, a human preference dataset where users compare pairs of AI-generated videos. We use AIGVE-MACS to predict per-video scores across the nine evaluation aspects, then infer pairwise preferences by comparing average scores. Following the VideoScore protocol, we report pairwise accuracy against human-labeled preferences.

VBench VBench [14] evaluates AI-generated videos using a collection of existing automatic metrics. Following VideoScore, we calculate pairwise accuracy across five aspects: Technical Quality, Subject Consistency, Dynamic Degree, Motion Smoothness, and Overall Consistency.

9. Ablation Study

Method	Technical	Dynamic	Consistency	Physics	Element_Pre	Element_Qu	Act_Pre	Act_Qu	Overall	G-Eval
AIGVE-MACS	**40.60**	**57.31**	**61.49**	**64.36**	**40.32**	**40.81**	**44.31**	**60.71**	**59.88**	**3.42**
-weighted loss	33.15	48.33	41.92	44.81	30.10	38.21	44.10	55.71	48.32	2.97
-dyna sample	37.96	50.76	47.52	40.41	35.36	33.08	38.12	53.24	53.99	3.31
Qwen2.5-VL	8.77	4.00	1.24	-6.01	9.19	10.19	18.74	0.72	9.59	2.37

Table 4. Ablation study of our finetuning strategy.

The ablation results in Table 4 clearly demonstrate the effectiveness of both components in the proposed finetuning strategy—token-wise weighted loss and dynamic frame sampling. Removing the weighted loss leads to a sharp performance drop across nearly all evaluation aspects, most notably in Consistency (-19.57) and Physics (-19.55), suggesting that emphasizing score and comment tokens is crucial for aligning with human judgment. Similarly, removing dynamic sampling substantially hurts aspects sensitive to temporal change, such as Dynamic and Action Quality, confirming its role in capturing meaningful motion cues. Compared to the pretrained Qwen2.5-VL, the full model achieves large gains across the board (e.g., +51.7 in Consistency, +60.4 in Physics), showing that joint modeling of structured outputs with targeted loss design and adaptive input selection is key to high-quality, aspect-aware video evaluation.

10. Details of AIGVE-BENCH 2

10.1. Evaluation Aspect Description

Metric	Description
Technical Quality	Assesses the technical aspects of the video, including whether the resolution is sufficient for object recognition, whether the colors are natural, and whether there is an absence of noise or artifacts.
Dynamic	Measures the extent of pixel changes throughout the video, focusing on significant object or camera movements and changes in environmental factors such as daylight, weather, or seasons.
Consistency	Evaluates whether objects in the video maintain consistent properties, avoiding glitches, flickering, or unexpected changes.
Physics	Determines if the scene adheres to physical laws, ensuring that object behaviors and interactions are realistic and aligned with real-world physics.
Element Presence	Checks if all objects mentioned in the instructions are present in the video. The score is based on the proportion of objects that are correctly included.
Element Quality	Assesses the realism and fidelity of objects in the video, awarding higher scores for detailed, natural, and visually appealing appearances.
Action/Interaction Presence	Evaluates whether all actions and interactions described in the instructions are accurately represented in the video.
Action/Interaction Quality	Measures the naturalness and smoothness of actions and interactions, with higher scores for those that are realistic, lifelike, and seamlessly integrated into the scene.
Overall	Reflects the comprehensive quality of the video based on all metrics, allowing raters to incorporate their subjective preferences into the evaluation.

Table 5. Metrics for Video Generation Evaluation

10.2. AIGVE-BENCH 2 Comment Length Analysis

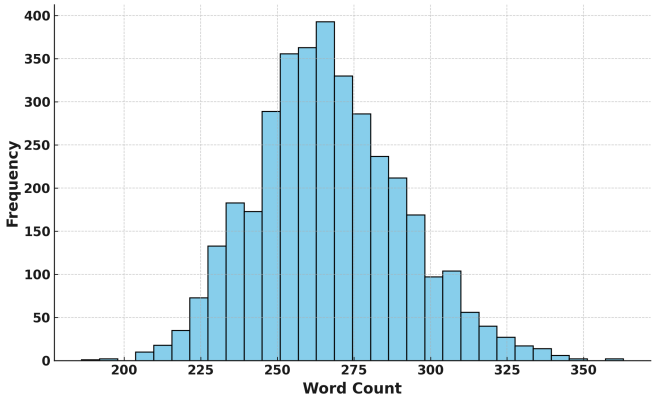


Figure 7. Comment Word Number Distribution of AIGVE-BENCH 2.

The distribution of combined comment word counts indicates that the revised comments are notably rich and thorough, averaging around 267 words per entry. This high word count reflects a strong emphasis on providing contextually rich, multi-aspect evaluations.

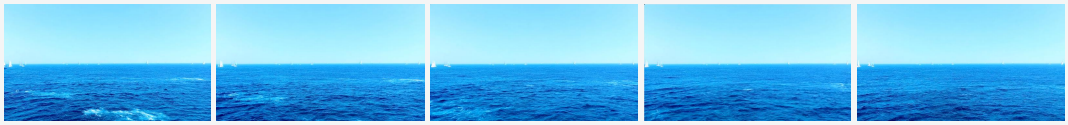
11. More Case Studies

Descending into a metropolis bustling with life and vibrant energy. The rain drenches the busy streets while umbrellas and movement ensue,....

TQ: 2, Dy: 5, CS: 1, Phy: 1, EP: 5, EQ: 1, AP: 5, AQ: 1, OR: 1

The video successfully includes all key elements and actions from the prompt but is significantly limited by low technical quality, inconsistent rendering, and physically...

A vast ocean extending to the horizon on a clear summer day. Segue from morning's calm waves to the sight of sailboats dotting the midday seascape.



TQ: 5, Dy: 1, CS: 5, Phy: 5, EP: 5, EQ: 5, AP: 0, AQ: 0, OR: 3

Overall, the video demonstrates strong technical quality, realism, and consistency, but it lacks the essential time-of-day transition described in the prompt,.....

A young woman in a blue coat strolls briskly along a bustling city street, her expression focused as she navigates through a crowd of people and hears traffic hum.

TQ: 2, Dy: 5, CS: 3, Phy: 5, EP: 5, EQ: 3, AP: 5, AQ: 3, OR: 3

Overall, while the video successfully includes all required elements and..., its impact is diminished by low resolution, inconsistent rendering, and noticeable distortions...

Flowing through a dense forest and a crystal clear river meandering softly beneath towering summer-green trees....

TQ: 5, Dy: 5, CS: 5, Phy: 5, EP: 5, EQ: 5, AP: 5, AQ: 5, OR: 5

The video excels in technical quality, realism, and faithful representation of the prompt, with strong consistency and natural depiction of both elements and actions.

Figure 8. Case Study of AIGVE-MACS.