AIGV-Assessor: Benchmarking and Evaluating the Perceptual Quality of Text-to-Video Generation with LMM

Jiarui Wang[1], Huiyu Duan[1,2], Guangtao Zhai[1,2], Juntong Wang[1], Xiongkuo Min[1]*
[1]Institute of Image Communication and Network Engineering,
[2] MoE Key Lab of Artificial Intelligence, AI Institute,
Shanghai Jiao Tong University, Shanghai, China

Abstract

*The rapid advancement of large multimodal models (LMMs) has led to the rapid expansion of artificial intelligence generated videos (AIGVs), which highlights the pressing need for effective video quality assessment (VQA) models designed specifically for AIGVs. Current VQA models generally fall short in accurately assessing the perceptual quality of AIGVs due to the presence of unique distortions, such as unrealistic objects, unnatural movements, or inconsistent visual elements. To address this challenge, we first present **AIGVQA-DB**, a large-scale dataset comprising 36,576 AIGVs generated by 15 advanced text-to-video models using 1,048 diverse prompts. With these AIGVs, a systematic annotation pipeline including scoring and ranking processes is devised, which collects 370k expert ratings to date. Based on AIGVQA-DB, we further introduce **AIGV-Assessor**, a novel VQA model that leverages spatiotemporal features and LMM frameworks to capture the intricate quality attributes of AIGVs, thereby accurately predicting precise video quality scores and video pair preferences. Through comprehensive experiments on both AIGVQA-DB and existing AIGV databases, AIGV-Assessor demonstrates state-of-the-art performance, significantly surpassing existing scoring or evaluation methods in terms of multiple perceptual quality dimensions. The dataset and code are released at https://github.com/IntMeGroup/AIGV-Assessor.*

1. Introduction

Text-to-video generative models [12, 27, 44, 64, 73], including auto-regressive [23, 81] and diffusion-based [12, 27, 55] approaches, have experienced rapid advancements in recent years with the explosion of large multimodal models (LMMs). Given appropriate text prompts, these models can generate high-fidelity and semantically-aligned videos, commonly referred to as AI-generated videos (AIGVs), which have significantly facilitated the content creation in various domains, including entertainment, art, design, and advertising, *etc* [11, 13, 43]. Despite the significant progress, current AIGVs are still far from satisfactory. Unlike natural videos, which are usually affected by low-level distortions, such as noise, blur, low-light, *etc*, AIGVs generally suffer from degradations such as unrealistic objects, unnatural movements, inconsistent visual elements, and misalignment with text descriptions [25, 31, 43, 65, 79, 84, 85].

The unique distortions in AIGVs also bring challenges to the video evaluation. Traditional video quality assessment (VQA) methods [10, 18, 33, 35, 57, 70, 71] mainly focus on evaluating the quality of professionally-generated content (PGC) and user-generated content (UGC), thus struggling to address the specific distortions associated with AIGVs, such as spatial artifacts, temporal inconsistencies, and misalignment between generated content and text prompts. For evaluation of AIGVs, some metrics such as Inception Score (IS) [52] and Fréchet Video Distance (FVD) [61] have been widely used, which are computed over distributions of videos and may not reflect the human preference for an individual video. Moreover, these metrics mainly evaluate the fidelity of videos, while failing to assess the text-video correspondence. Vision-language pre-training models, such as CLIPScore [22], BLIPScore [37], and AestheticScore [53] are frequently employed to evaluate the alignment between generated videos and their text prompts. However, these models mainly consider the text-video alignment at the image level, while ignoring the dynamic diversity and motion consistency of visual elements that are crucial to the video-viewing experience.

In this paper, to facilitate the development of more comprehensive and precise metrics for evaluating AI-generated videos, we present **AIGVQA-DB**, a large-scale VQA dataset, including 36,576 AIGVs generated by 15 advanced text-to-video models using 1,048 diverse prompts. An

*Corresponding Author.
This work was supported in part by the National Natural Science Foundation of China under Grant 62271312, 62132006, 62401365, 62225112, and in part by STCSM under Grant 22DZ2229005.

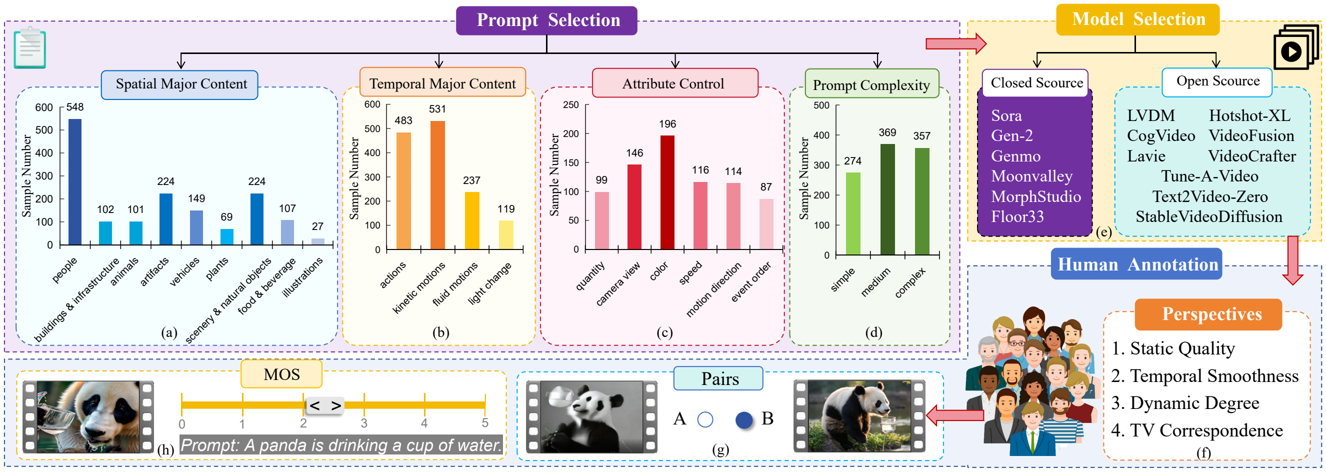


Figure 1. An overview of the AIGVQA-DB construction pipeline, illustrating the generation and the subjective evaluation procedures for the AIGVs in the database. (a) Prompt categorization according to the spatial major content. (b) Prompt categorization according to the temporal descriptions. (c) Prompt categorization according to the attribute control. (d) Prompt categorization according to the prompt complexity. (e) The 15 generative models used in the database. (f) Four visual quality evaluation perspectives, including static quality, temporal smoothness, dynamic degree, and text-video correspondence. (g) and (h) demonstrates the pair comparison and preference scoring processes, respectively.

overview of the dataset construction pipeline is shown in Figure 1. The prompts are collected from existing open-domain text-video datasets [7, 8, 38, 43, 68, 76] or manually-written, which can be categorized based on four orthogonal aspects respectively, as shown in Figure 1(a)-(d). Based on the AIGVs, we collect 370k expert ratings comprising both mean opinion scores (MOSs) and pair-wise comparisons, which are evaluated from four dimensions, including: (1) static quality, (2) temporal smoothness, (3) dynamic degree, and (4) text-video correspondence. Equipped with the dataset, we propose **AIGV-Assessor**, a large multimodal model-based (LMM-based) VQA method for AIGVs, which reformulates the quality regression task into an interactive question-and-answer (Q&A) framework and leverages the powerful multimodal representation capabilities of LMMs to provide accurate and robust quality assessments. AIGV-Assessor not only classifies videos into different quality levels through natural language output, but also generates precise quality scores through regression, thus enhancing the interpretability and usability of VQA results. Moreover, AIGV-Assessor also excels in pairwise video comparisons, enabling nuanced assessments that are closer to human preferences. Extensive experimental results demonstrate that AIGV-Assessor outperforms existing text-to-video scoring methods in terms of multiple dimensions relevant to human preference.

The main contributions of this paper are summarized as follows:

- We construct AIGVQA-DB, a large-scale dataset comprising 36,576 AI-generated videos annotated with MOS scores and pairwise comparisons. Compared with existing benchmarks, AIGVQA-DB provides a more comprehensive assessment of the capabilities of text-to-video models from multiple perspectives.

Table 1. An overview of popular text-to-video (T2V) and image-to-video (I2V) generation models. † Representative variable.

Model	Year	Mode	Resolution	Frames	Open
CogVideo [23]	22.05	T2V	480×480	32	✓
Make-a-Video [55]	22.09	T2V	256×256	16	✓
LVDM [21]	22.11	T2V	256×256	16	✓
Tune-A-Video [73]	22.12	T2V	512×512	8	✓
VideoFusion [44]	23.03	T2V	128×128	16	✓
Text2Video-Zero [27]	23.03	T2V	512×512	8	✓
ModelScope [64]	23.03	T2V	256×256	16	✓
Lavie [67]	23.09	T2V	512×320	16	✓
VideoCrafter [12]	23.10	T2V, I2V	1024×576	16	✓
Hotshot-XL [1]	23.10	T2V	672×384	8	✓
StableVideoDiffusion [9]	23.11	I2V	576×1024	14	✓
AnimateDiff [20]	23.12	T2V, I2V	384×256	20	✓
Floor33 [2]	23.08	T2V,I2V	1024×640	16	–
Genmo [3]	23.10	T2V, I2V	2048×1536	60	–
Gen-2 [4]	23.12	T2V, I2V	1408×768	96	–
MoonValley [5]	24.01	T2V, I2V	1184×672	200†	–
MorphStudio [6]	24.01	T2V, I2V	1920×1080	72	–
Sora [7]	24.02	T2V, I2V	1920×1080	600†	–

- Based on AIGVQA-DB, we evaluate and benchmark 15 representative text-to-video models, and reveal their strengths and weaknesses from four crucial preference dimensions, *i.e.,* static quality, temporal smoothness, dynamic degree, and text-to-video correspondence.
- We present a novel LMM-based VQA model for AIGVs, termed AIGV-Assessor, which integrates both spatial and temporal visual features as well as prompt features into a LMM to give quality levels, predict quality scores, and conduct quality comparisions.
- Thorough analysis of our AIGV-Assessor is provided and extensive experiments on our proposed AIGVQA-DB and other AIGV quality assessment datasets have shown the effectiveness and applicability of AIGV-Assessor.

2. Related Work

2.1. Text-to-video Generation

Recent advancements in text-to-video generative models have substantially broadened video creation and modifica-

Table 2. Summary of existing text-to-image and text-to-video evaluation datasets.

Dataset Types	Name	Numbers	Prompts	Models	Annotators	Dimensions	MOSs / Pairs	Annotation
AIGIQA	AGIQA-3k [34]	2,982	180	6	21	2	5,964	MOS
	AIGCIQA2023 [63]	2,400	100	6	28	3	7,200	MOS
	RichHF-18k [39]	17,760	17,760	3	3	4	71,040	MOS
	HPS [75]	98,807	25,205	1	2,659	1	25,205	Pairs
	Pick-a-Pic [29]	-	37,523	3	4,375	1	584,247	Pairs
AIGVQA	MQT [15]	1,005	201	5	24	2	2,010	MOS
	EvalCrafter [42]	2,500	700	5	7	4	1,024	MOS
	FETV [43]	2,476	619	4	3	3	7,428	MOS
	LGVQ [84]	2,808	468	6	20	3	8,424	MOS
	T2VQA-DB [31]	10,000	1,000	9	27	1	10,000	MOS
	GAIA [13]	9,180	510	18	54	3	27,540	MOS
	AIGVQA-DB (Ours)	**36,576**	**1,048**	**15**	**120**	**4**	**122,304**	**MOS and Pairs**

tion possibilities. As shown in Table 1, these models exhibit distinct characteristics and capacities, including modes, resolution, and total frames. CogVideo [23] is an early text-to-video (T2V) model capable of generating short videos based on CogView2 [16]. Make-a-video [55] adds effective spatial-temporal modules on a diffusion-based text-to-image (T2I) model (*i.e.*, DALLE-2 [50]). VideoFusion [44] also leverages the DALLE-2 and presents a decomposed diffusion process. LVDM [21], Text2Video-Zero [27], Tune-A-Video [73], and ModelScope [64] are models that inherit the success of Stable Diffusion (SD) [51] for video generation. Lavie [67] extends the original transformer block in SD to a spatio-temporal transformer. Hotshot-XL [1] introduces personalized video generation. Beyond these laboratory-driven advancements, the video generation landscape has also been enriched by a series of commercial products. Notable among them are Floor33 [2], Gen-2 [4], Genmo [3], MoonValley [5], MorphStudio [6], and Sora [7], which have gained substantial attention in both academia and industry, demonstrating the widespread application potential of AI-assisted video creation.

2.2. Text-to-video Evaluation

The establishment of the AI-generated image quality assessment (AIGIQA) dataset is relatively well-developed, including both mean opinion scores (MOSs) for absolute quality evaluations, and pairwise comparisons for relative quality judgments. Recent developments in text-to-video generation models have also spurred the creation of various AI-generated video quality assessment (AIGVQA) datasets, addressing different aspects of the T2V generation challenge, as shown in Table 2. MQT [15] consists of 1,005 videos generated by 5 models using 201 prompts. EvalCrafter [42] and FETV [43] extend the scale of the videos, prompts, and evaluation dimensions. LGVQ [84] increases the number of annotators, providing more reliable MOSs. T2VQA-DB [31] consists of 10,000 videos from 1,000 prompts representing a significant improvement in scale. GAIA [13] collects 9,180 videos focusing on action quality assessment in AIGVs, but falls short in addressing the consistency between the generated visuals and their textual prompts. Most existing VQA datasets predominantly rely on MOS, an absolute scoring method, which suffers from the same drawback: absolute scores alone may cause ambiguity and overlook subtle quality differences. In contrast, our AIGVQA-DB includes both MOSs and pairwise comparisons, addressing the limitations of current works by providing fine-grained preference feedbacks.

3. Database Construction and Analysis

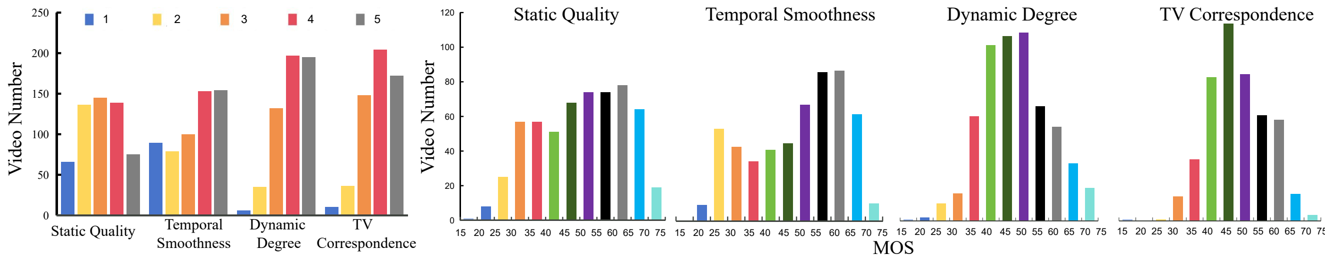
3.1. Data Collection

Prompt Scources and Categorization. Prompts of the AIGVQA-DB are primarily sourced from existing open-domain text-video pair datasets, including InternVid [68], MSRVTT [76], WebVid [8], TGIF [38], FETV [43] and Sora website [7]. We also manually craft prompts describing highly unusual scenarios to test the generalization ability of the generation models. As shown in Figure 1(a)-(d), we follow the categorization principles from FETV [43] to organize each prompt based on the "spatial major content", "temporal major content", "attribute control", and "prompt complexity".

Text-to-Video Generation. We utilize 15 latest text-to-video generative models to create AI-generated videos as shown in Figure 1(e). We leverage open-source website APIs and code with default weights for these models to produce AIGVs. For the construction of the MOS subset, we collect 48 videos from the Sora Website [7], along with their corresponding text prompts. Using these prompts, we generate additional videos using 11 different generative models. This process results in a total of 576 videos (12 generative models $\times$ 48 prompts). In addition to the MOS subset, we construct the pair-comparison subset using 1,000 diverse prompts, and 12 generative models including 8 open-sourced and 4 close-sourced are employed for text-to-video generation. Specifically, for each prompt, we generate four distinct videos for each open-source generative model and one video for each closed-source generative model. This process yields a total of 36,000 videos. More details of the database can be found in the *supplementary material*.

3.2. Subjective Experiment Setup and Procedure

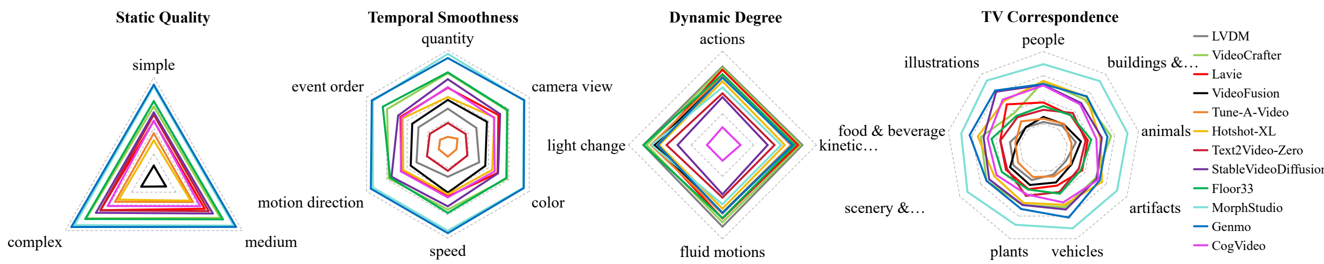
Due to the unique and unnatural characteristics of AI-generated videos and the varying target video spaces dictated by different text prompts, relying solely on a single score, such as "quality", to represent human visual preferences is insufficient. In this paper, we propose to measure



(a) Distribution of Raw Scores

(b) Distribution of Mean Opinion Scores (MOSs)

Figure 2. Video score distribution from the four perspectives including static quality, temporal smoothness, dynamic degree, and t2v correspondence. (a) Distribution of raw scores. (b) Distribution of Mean Opinion Scores (MOSs)



(a) (b) (c) (d)

Figure 3. Comparison of averaged win rates of different generation models across different categories. (a) Results across prompt complexity. (b) Results across attribute control. (c) Results across temporal major contents. (d) Results across spatial major contents.

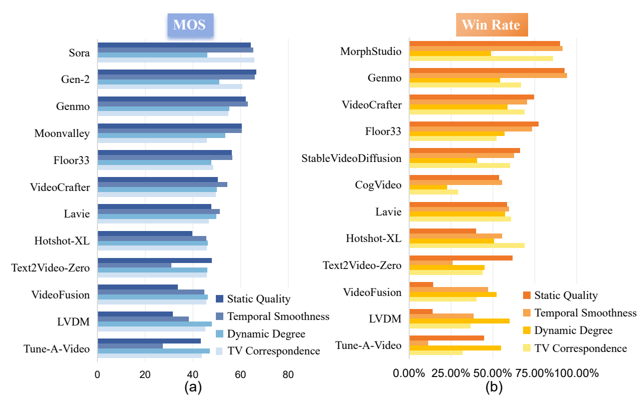


Figure 4. (a) Comparison of text-to-video generation models regarding the MOS in terms of four dimensions sorted bottom-up by their averaged MOS. (b) Comparison of text-to-video generation models regarding the win rate in terms of four dimensions sorted bottom-up by their averaged win rate.

the human visual preferences of AIGVs from four perspectives. **Static quality** assesses the clarity, sharpness, color accuracy, and overall aesthetic appeal of the frames when viewed as standalone images. **Temporal smoothness** evaluates the temporal coherence of video frames and the absence of temporal artifacts such as flickering or jittering. **Dynamic degree** evaluates the extent to which the video incorporates large motions and dynamic scenes, which contributes to the overall liveliness and engagement measurement of the content. **Text-video (TV) correspondence** assesses how accurately the video content reflects the details, themes, and actions described in the prompt, ensuring that the generated video effectively translates the text input into a visual narrative. Each of these four visual perception perspectives is related but distinct, offering a comprehensive evaluation for AIGVs. To evaluate the quality of the videos in the AIGVQA-DB, we conduct subjective experiments adhering to the guidelines outlined in ITU-R BT.500-14 [17, 54]. For the MOS annotation type, we use a 1-5 Likert-scale judgment to score the videos. For the pairs annotation type, participants are presented with pairs of videos and asked to choose the one they prefer, providing a direct comparison method for evaluating relative video quality. The videos are displayed using an interface designed with Python Tkinter, as illustrated in Figure 1(g)-(h). A total of 120 graduate students participate in the experiment.

3.3. Subjective Data Processing

In order to obtain the MOS for an AIGV, we linearly scale the raw ratings to the range $[0, 100]$ as follows:

$$z_{ij} = \frac{r_{ij} - \mu_{ij}}{\sigma_i}, \quad z'_{ij} = \frac{100(z_{ij} + 3)}{6},$$

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} r_{ij}, \quad \sigma_i = \sqrt{\frac{1}{N_i - 1} \sum_{j=1}^{N_i} (r_{ij} - \mu_{ij})^2}$$

where r_{ij} is the raw ratings given by the i-th subject to the j-th video. N_i is the number of videos judged by subject i. Next, the MOS of the video j is computed by averaging the rescaled z-scores as follows:

$$MOS_j = \frac{1}{M} \sum_{i=1}^{M} z'_{ij}$$

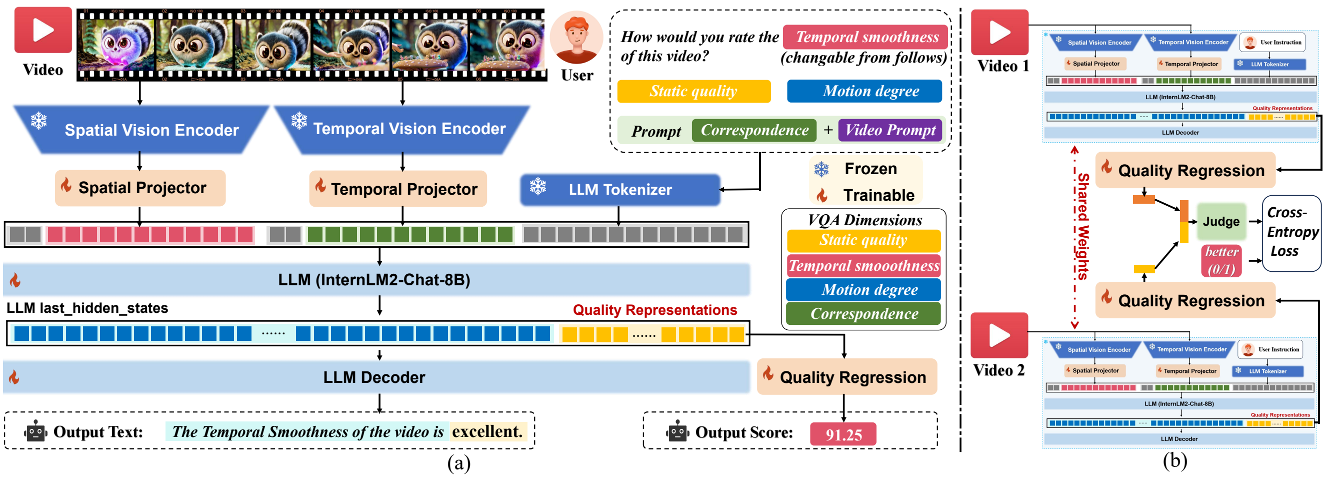


Figure 5. The framework of AIGV-Assessor: (a) AIGV-Assessor takes AI-generated video frames as input and outputs both text-based quality levels and numerical quality scores. The system begins with the extraction of spatiotemporal features using two vision encoders, which are then passed through spatial and temporal projection modules to generate aligned visual tokens into language space. The LLM decoder produces text-based feedback describing the video quality level for four evaluation dimensions, respectively. Simultaneously, the last-hidden-states from the LLM are used to perform quality regression that outputs final quality scores in terms of four dimensions. (b) AIGV-Assessor is fine-tuned on pairwise comparison, further allowing the model to output the evaluation comparison between two videos.

where MOS_j indicates the MOS for the j-th AIGV, M is the number of subjects, and z'_{ij} are the rescaled z-scores.

For the pairs annotation type, given a text prompt p_i, and 12 video generation models labeled $\{A, B, C, ..., L\}$, we generate videos using each model, forming a group of videos $G_{i,j} = \{V_{i,A,j}, V_{i,B,j}, V_{i,C,j}, ..., V_{i,L,j}\}$. For each prompt p_i, we generate four different videos randomly for each of the eight open-source generative models and one video for each of the four closed-source generative models, resulting in a group of 36 videos $\{G_{i,A,1}, G_{i,A,2}, G_{i,A,3}, G_{i,A,4}, G_{i,B,1}, ...G_{i,L,1}\}$. For each group, we create all possible pairwise combinations, resulting in C_{36}^2 pairs: (V_{A1}, V_{B1}), (V_{A1}, V_{B2}), (V_{A1}, V_{B3}), (V_{A1}, V_{B4}), (V_{A1}, V_{C1}), ... , (V_{K1}, V_{L1}). In the AIGVQA-DB construction pipeline, a prompt suite of 1000 prompts results in 630,000 ($1000 \times C_{36}^2$) pairwise video comparisons. From this extensive dataset, we randomly sample 30,000 pairs for evaluation from four perspectives. Each pair is judged by three annotators, and the final decision of the better video in each pair is determined by the majority vote. Finally, we obtain a total of 46,080 reliable score ratings (20 annotators $\times$ 4 perspectives $\times$ 576 videos) and 360,000 pair ratings (3 annotators $\times$ 4 perspectives $\times$ 30,000 pairs).

3.4. AIGV Analysis from Four Perspectives

As shown in Figure 2, the videos in the AIGVQA-DB cover a wide range of perceptual quality. We further analyze the win rates of various generation models across categories in Figure 3, revealing the strengths and weaknesses of each T2V model. As shown in Figure 3(a), the performances of T2V models rank uniform for different prompt complexity items in terms of static quality, which manifests current T2V model rank consistently for different prompts, likely due to shared architectures like diffusion-based systems, with common strengths and limitations in handling complex prompts. As shown in Figure 3(b), in terms of attribute control, StableVideoDiffusion [9] excels in managing quantity over event order, as it first generates static images before animating them, preserving the original event sequence. As shown in Figure 3(d), in terms of spatial content, most videos featuring "plants" and "people" show poor T2V correspondence. More comparison and analysis can be found in the *supplementary material.* We also launch comparisons among text-to-video generation models regarding the MOS and pairwise win rates shown in Figure 4. Notably, models such as LVDM [21] demonstrate exceptional performance in handling dynamic content, but exhibit relatively lower performance in temporal smoothness. Sora [7] and MorphStudio [6] perform well in static quality and temporal smoothness while lagging in dynamic degree. Additionally, closed-source models exhibit much better performance compared to open-source models.

4. Proposed Method

4.1. Model Structure

Spatial and Temporal Vision Encoder. As shown in Figure 5(a), the model leverages two different types of encoders to capture the spatial and temporal characteristics of the video: (1) 2D Encoder: A pre-trained 2D vision transformer (InternViT [69]) is used to process individual video frames. (2) 3D Encoder: A 3D network, *i.e.,* SlowFast [19], is employed to extract temporal features by processing sequences of video frames.

Spatiotemporal Projection Module. Once the spatial and temporal features are extracted, they are projected into a

Table 3. Performance comparisons of the state-of-the-art quality evaluation methods on the AIGVQA-DB from four perspectives. The best performance results are marked in **RED** and the second-best performance results are marked in **BLUE**.

Dimension	Static Quality				Temporal Smoothness				Dynamic Degree				TV Correspondence			
Methods / Metrics	Pair Acc	SRCC	PLCC	KRCC	Pair Acc	SRCC	PLCC	KRCC	Pair Acc	SRCC	PLCC	KRCC	Pair Acc	SRCC	PLCC	KRCC
NIQE [49]	54.32%	0.0867	0.1626	0.0615	52.67%	0.0641	0.1152	0.0451	45.64%	0.1765	0.2448	0.1194	46.99%	0.1771	0.2231	0.1193
QAC [80]	49.96%	0.1022	0.1363	0.0680	54.90%	0.1633	0.2039	0.1105	54.72%	0.0448	0.0427	0.0295	54.48%	0.0303	0.0197	0.2233
BRISQUE [48]	59.98%	0.2909	0.2443	0.1969	55.67%	0.2325	0.1569	0.1553	44.60%	0.1351	0.0959	0.0893	51.02%	0.1294	0.1017	0.0869
BPRI [46]	52.28%	0.2181	0.1723	0.1398	47.26%	0.1766	0.0880	0.1138	46.83%	0.1956	0.1688	0.1329	49.13%	0.1569	0.1548	0.1052
HOSA [77]	61.54%	0.2420	0.2106	0.1643	57.31%	0.2311	0.1757	0.1559	44.97%	0.0755	0.0449	0.0496	52.23%	0.1645	0.1324	0.1097
BMPRI [47]	53.71%	0.1690	0.1481	0.1075	49.31%	0.1434	0.0844	0.0894	45.07%	0.1153	0.0925	0.0777	48.43%	0.1567	0.1500	0.1041
V-Dynamic [25]	51.34%	0.0768	0.0792	0.0494	31.91%	0.3713	0.4871	0.2557	53.11%	0.1466	0.0253	0.0988	46.96%	0.0405	0.0576	0.0223
V-Smoothness [25]	61.63%	0.6748	0.4506	0.4590	76.59%	0.8526	0.8313	0.6533	47.63%	0.2446	0.2328	0.1580	61.28%	0.3188	0.3073	0.2214
CLIPScore [22]	47.09%	0.0731	0.0816	0.0473	46.33%	0.0423	0.0334	0.0271	52.99%	0.0675	0.0835	0.0439	55.62%	0.1519	0.1731	0.1014
BLIPScore [37]	53.24%	0.0492	0.0421	0.0330	53.07%	0.0659	0.0487	0.0437	53.03%	0.1786	0.1904	0.1205	61.53%	0.1813	0.1896	0.1219
AestheticScore [53]	70.24%	0.6713	0.6959	0.4784	54.82%	0.5154	0.4946	0.3484	52.96%	0.2295	0.2322	0.1527	59.64%	0.2381	0.2440	0.1602
ImageReward [78]	56.69%	0.2606	0.2646	0.1749	54.09%	0.2382	0.2305	0.1600	53.90%	0.1840	0.1836	0.1237	63.97%	0.2311	0.2450	0.1568
UMTScore [43]	48.93%	0.0168	0.0199	0.0117	49.93%	0.0302	0.0370	0.0207	52.69%	0.0168	0.0198	0.0117	53.82%	0.0172	0.0065	0.0108
Video-LLaVA [40]	50.90%	0.0384	0.0513	0.0297	50.36%	0.0431	0.0281	0.0347	50.34%	0.1561	0.1436	0.1176	50.54%	0.1364	0.1051	0.1009
Video-ChatGPT [45]	51.20%	0.1242	0.1587	0.0940	50.16%	0.0580	0.0533	0.0453	50.47%	0.0724	0.0436	0.0563	50.07%	0.0357	0.0124	0.0274
LLaVA-NeXT [36]	52.85%	0.1239	0.1625	0.0954	52.41%	0.4021	0.3722	0.3052	51.84%	0.1767	0.1655	0.1328	59.20%	0.4116	0.3428	0.3261
VideoLLaMA2 [14]	52.73%	0.2643	0.3271	0.1928	52.27%	0.3608	0.2450	0.2696	50.78%	0.1900	0.1561	0.1379	54.25%	0.1656	0.1633	0.1210
Qwen2-VL [66]	56.50%	0.4922	0.5291	0.3838	49.12%	0.1681	0.4219	0.1233	52.08%	0.1122	0.1335	0.0849	53.30%	0.3111	0.2775	0.2306
HyperIQA [56]	68.30%	0.7931	0.8093	0.5969	54.65%	0.7426	0.6630	0.5407	53.32%	0.2103	0.2100	0.1384	57.54%	0.6226	0.6250	0.4432
MUSIQ [26]	66.46%	0.7880	0.8044	0.5773	55.16%	0.7199	0.6920	0.5034	52.85%	0.5206	0.4846	0.3521	58.46%	0.4125	0.4093	0.2844
LIQE [83]	63.86%	0.8776	0.8691	0.7008	55.84%	0.7935	0.7720	0.6084	49.02%	0.5303	0.5840	0.3837	55.10%	0.3862	0.3639	0.2640
VSFA [35]	46.43%	0.3365	0.3421	0.2268	50.95%	0.3317	0.3273	0.2202	51.46%	0.1201	0.1362	0.0815	48.07%	0.1024	0.1064	0.0666
BVQA [33]	29.98%	0.4594	0.4701	0.3268	37.65%	0.3704	0.3819	0.2507	**55.08%**	0.4594	0.4701	0.3268	42.32%	0.3720	0.3978	0.2559
simpleVQA [57]	68.12%	0.8355	0.6438	0.8489	54.14%	0.7082	0.7008	0.4978	53.08%	0.4671	0.3160	**0.3994**	58.20%	0.4643	0.5440	0.3163
FAST-VQA [70]	70.64%	0.8738	0.8644	0.6860	**62.93%**	0.9036	0.9134	0.7166	54.34%	0.5603	**0.5703**	0.3895	**65.05%**	**0.6875**	**0.6704**	**0.4978**
DOVER [71]	**72.92%**	**0.8907**	**0.8895**	**0.7004**	58.83%	**0.9063**	**0.9195**	**0.7187**	53.16%	0.5549	0.5489	0.3800	62.35%	0.6783	0.6802	0.4969
Q-Align [72]	71.86%	0.8516	0.8383	0.6641	57.95%	0.8116	0.7025	0.6195	53.71%	**0.5655**	0.5012	0.3950	62.91%	0.5542	0.5647	0.3870
AIGV-Assessor (Ours)	**79.83%**	**0.9162**	**0.9190**	**0.7576**	**76.60%**	**0.9232**	**0.9216**	**0.8038**	**60.30%**	**0.6093**	**0.6082**	**0.4435**	**70.32%**	**0.7500**	**0.7697**	**0.5591**
Improvement	+ 6.9%	+2.7%	+3.0%	+ 5.7%	13.7%	+1.7%	+0.2%	+8.5%	+5.2%	+4.4%	+3.8%	+ 4.4%	+5.3%	+6.3%	+9.9%	+6.13%

shared feature space for alignment with text-based queries. This is done through two projection modules that map the spatial and temporal visual features respectively into the language space. The mapped visual tokens are aligned with text tokens, enabling the model to query the video content in a multimodal fashion.

Feature Fusion and Quality Regression. We apply LLM (InternVL2-8B [69]) to combine the visual tokens and user-provided quality prompts to perform the following tasks: (1) Quality level descriptions: the model generates a descriptive quality level evaluation of the input video, such as "The static quality of the video is (bad, poor, fair, good, excellent)." This initial categorization provides a preliminary classification of the video's quality, which is beneficial for subsequent quality regression tasks. By obtaining a rough quality level, the model can more accurately predict numerical scores in later evaluations. (2) Regression score output: the model uses the final hidden states from the LLM to perform a regression task, outputting numerical quality scores for the video from four different dimensions.

4.2. Training and Fine-tuning Strategy

The training process of AIGV-Assessor follows a three-stage approach to ensure high-quality video assessment with quality level prediction, individual quality scoring, and pairwise preference comparison capabilities. This process includes: (1) training the spatial and temporal projectors to align visual and language features, (2) fine-tuning the vision encoder and LLM with LoRA [24], and training the quality regression module to generate accurate quality scores, (3) incorporating pairwise comparison training using the pair-comparison subset with a pairwise loss function for robust video quality comparison.

Spatiotemporal Projector Training. The first stage focuses on training the spatial and temporal projectors to extract meaningful spatiotemporal visual features and map them into the language space. Through this process, the LLM is able to produce the quality level descriptions *i.e.,* bad, poor, fair, good, excellent.

Quality Regression Fine-tuning. Once the model can generate coherent descriptions of video quality level, the second stage focuses on fine-tuning the quality regression module. The goal here is to enable the model to output stable and precise numerical quality scores (MOS-like predictions). The quality regression model takes the last-hidden-state features from LLM as input and generates quality scores from four perspectives. The training objective uses an L1 loss function to minimize the difference between the predicted quality score and the groundtruth MOS.

Pairwise Comparison Fine-tuning. The third stage mainly focuses on integrating the pairwise comparison into the training pipeline. As shown in Figure 5(b), two input video pairs share network weights within the same batch. We design a judge network inspired by LPIPS [82] to determine which video performs better. This network leverages

Table 4. Performance comparisons on LGVQ [84] and FETV [43].

Aspects	Methods	LGVQ			FETV		
		SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
Spatial	MUSIQ [26]	0.669	0.682	0.491	0.722	0.758	0.613
	StairIQA [59]	0.701	0.737	0.521	0.806	0.812	0.643
	CLIP-IQA [62]	0.684	0.709	0.502	0.741	0.767	0.619
	LIQE [83]	0.721	0.752	0.538	0.765	0.799	0.635
	UGVQ [84]	**0.759**	**0.795**	**0.567**	**0.841**	**0.841**	**0.685**
	AIGV-Assessor (Ours)	**0.803**	**0.819**	**0.617**	**0.853**	**0.856**	**0.699**
	Improvement	+ 4.4%	+2.4%	+5.0%	+1.2%	+1.5%	+1.4%
Temporal	VSFA [35]	0.841	0.857	0.643	0.839	0.859	0.705
	SimpleVQA [57]	0.857	0.867	0.659	0.852	0.862	0.726
	FastVQA [70]	0.849	0.843	0.647	0.842	0.847	0.714
	DOVER [71]	0.867	0.878	0.672	0.868	0.881	0.731
	UGVQ [84]	**0.893**	**0.907**	**0.703**	**0.897**	**0.907**	**0.753**
	AIGV-Assessor (Ours)	**0.900**	**0.920**	**0.717**	**0.936**	**0.940**	**0.815**
	Improvement	+0.7%	+1.3%	+1.4%	+3.9%	+3.3%	+6.2%
Alignment	CLIPScore [22]	0.446	0.453	0.301	0.607	0.633	0.498
	BLIPScore [37]	0.455	0.464	0.319	0.616	0.645	0.505
	ImageReward [78]	0.498	0.499	0.344	0.657	0.687	0.519
	PickScore [28]	0.501	0.515	0.353	0.669	0.708	0.533
	HPSv2 [74]	0.504	0.511	0.357	0.686	0.703	0.540
	UGVQ [84]	**0.551**	**0.555**	**0.394**	**0.734**	**0.737**	**0.572**
	AIGV-Assessor (Ours)	**0.577**	**0.578**	**0.411**	**0.753**	**0.746**	**0.585**
	Improvement	+2.6%	+2.3%	+1.7%	+1.9%	+0.9%	+1.3%

learned features and evaluates the perceptual differences between the two videos, allowing more reliable quality assessments in video pair comparison.

Loss Function. In the first stage, the spatial and temporal projectors are trained to align visual and language features using language loss. The second stage refines the vision encoder, LLM, and quality regression module's scoring ability with an L1 loss. The third stage incorporates pairwise comparison training with cross-entropy loss to improve the model's performance on relative quality evaluation.

5. Experiments

5.1. Experiment Settings

Evaluation Datasets and Metrics. Our proposed method is validated on five AIGVQA datasets: AIGVQA-DB, LGVQ [84], FETV [43], T2VQA [31], and GAIA [13]. To evaluate the correlation between the predicted scores and the ground-truth MOSs, we utilize three evaluation criteria: Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Kendall's Rank Correlation Coefficient (KRCC). For pair comparison, we adopt the comparison accuracy as the metric.

Reference Algorithms. To assess the performance of our proposed method, we select state-of-the-art evaluation metrics for comparison, which can be classified into five groups: (1) Handcrafted-based I/VQA models, including: NIQE [49], BRISQUE [48], QAC [80], BMPRI [47], HOSA [77], BPRI [46], HIGRADE [32], *etc.* (2) Action-related evaluation models, including: V-Dynamic [25], V-Smoothness [25] which are proposed in VBench [25]. (3) Vision-language pre-training models, including: CLIPScore [22], BLIPScore [37], AestheticScore [53], ImageReward [78], and UMTScore [43]. (4) LLM-based models, including: Video-LLaVA [40], Video-ChatGPT [45], LLaVA-NeXT [36], VideoLLaMA2 [14], and Qwen2-VL [66]. (5) Deep learning-based I/VQA models, including: HyperIQA [56], MUSIQ [26], LIQE [83], VSFA [35], BVQA [33], SimpleVQA [58], FAST-VQA [70], DOVER [71], and Q-Align [72].

Table 5. Performance comparisons on T2VQA-DB [31].

Aspects	Methods	T2VQA-DB			Sora Testing		
		SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
zero-shot	CLIPScore [22]	0.1047	0.1277	0.0702	0.2116	0.1538	0.1406
	BLIPScore [37]	0.1659	0.1860	0.1112	0.2116	0.1038	0.1515
	ImageReward [78]	0.1875	0.2121	0.1266	0.0992	0.0415	0.0748
	UMTScore [43]	0.0676	0.0721	0.0453	0.2594	0.0840	0.1680
finetuned	SimpleVQA [57]	0.6275	0.6388	0.4466	0.0340	0.2344	0.0237
	BVQA [37]	0.7390	0.7486	0.5487	0.4235	0.2489	0.2635
	FAST-VQA [70]	0.7173	0.7295	0.5303	0.4301	0.2369	0.2939
	DOVER [71]	0.7609	0.7693	0.5704	0.4421	0.2689	0.2757
	T2VQA [31]	**0.7965**	**0.8066**	**0.6058**	**0.6485**	**0.3124**	**0.4874**
	AIGV-Assessor (Ours)	**0.8131**	**0.8222**	**0.6364**	**0.6612**	**0.3318**	**0.5075**
	Improvement	+ 1.7%	+1.6%	+3.1%	+1.3%	+1.9%	+2.0%

Table 6. Performance comparisons on GAIA [13].

Dimension	Subject		Completeness		Interaction	
Methods / Metrics	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
V-Smoothness [25]	0.2402	0.1913	0.1474	0.1625	0.1741	0.1693
V-Dynamic [25]	0.1285	0.0831	0.0903	0.0682	0.1141	0.0758
Action-Score [42]	0.2023	0.1823	0.2867	0.2623	0.2689	0.2432
Flow-Score [42]	0.1471	0.1541	0.0816	0.1273	0.1041	0.1309
CLIPScore [22]	0.3398	0.3330	0.3944	0.3871	0.3875	0.3821
BLIPScore [37]	0.3453	0.3386	0.4174	0.4082	0.4044	0.3994
LLaVAScore [41]	0.3484	0.3436	0.4189	0.4133	0.4077	0.4025
TLVQM [30]	0.5037	0.5137	0.4127	0.4158	0.4079	0.4093
VIDEVAL [60]	0.5237	0.5446	0.4283	0.4375	0.4121	0.4234
VSFA [35]	0.5594	0.5762	0.4940	0.5017	0.4709	0.4811
BVQA [37]	0.5702	0.5888	0.4876	0.4946	0.4761	0.4825
SimpleVQA [58]	0.5920	0.5974	0.4981	0.5078	0.4843	0.4971
FAST-VQA [70]	0.6015	0.6092	0.5157	0.5215	0.5154	0.5216
DOVER [71]	**0.6173**	**0.6301**	**0.5198**	**0.5323**	**0.5164**	**0.5278**
AIGV-Assessor (Ours)	**0.6842**	**0.6897**	**0.6635**	**0.6694**	**0.6329**	**0.6340**
Improvement	+6.7%	+6.0%	+14.4%	+13.7%	+11.65%	+10.6%

Training Settings. Traditional handcrafted models are directly evaluated on the corresponding databases, and the average score of all frames is calculated. For vision-language pre-training and LLM-based models, we load the pre-trained weights for inference. CLIPscore [22], BLIPscore [37], and other vision-language pre-training models are calculated directly as the average cosine similarity between text and each video frame. SimpleVQA [58], BVQA [33], FAST-VQA [70], DOVER [71], and Q-Align [72] are fine-tuned on every test dataset. For deep learning-based IQA and VQA models, all experiments for each method are retrained on each dimension using the same training and testing split as the previous literature at a ratio of 4:1. All results are averaged after ten random splits.

5.2. Results and Analysis

Table 3 presents the pairwise win rates and the score prediction correlation between predicted results and human ground truths. The results indicate that handcrafted-based methods consistently underperform across all four evalu-

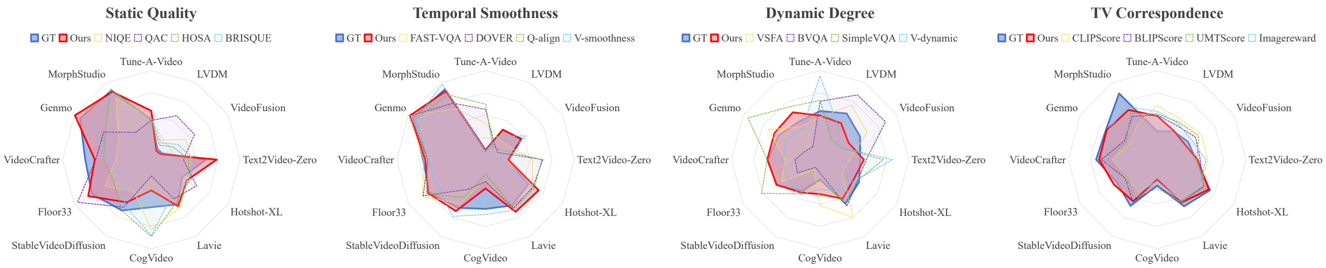


Figure 6. Comparison of win rates of different generation models across four dimensions evaluated by different VQA methods, demonstrating our AIGV-Assessor has better win-rate evaluation ability aligned with Ground Truth (GT).

Table 7. Ablation study of the proposed AIGV-Assessor method.

	Feature & Strategy				Static Quality			Temporal Smoothness			Dynamic Degree			T2V Correspondence		
No.	spatial	temporal	quality level	LoRA finetuning	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
(1)	✔		✔		0.864	0.866	0.726	0.870	0.868	0.727	0.556	0.572	0.432	0.616	0.620	0.492
(2)	✔			✔	0.874	0.876	0.723	0.875	0.876	0.736	0.558	0.573	0.431	0.723	0.734	0.533
(3)	✔		✔	✔	0.887	0.884	0.722	0.881	0.883	0.706	0.562	0.575	0.433	0.739	0.758	0.544
(4)	✔	✔	✔		0.887	0.888	0.753	0.917	0.910	0.796	0.569	0.536	0.438	0.688	0.673	0.557
(5)	✔	✔		✔	0.905	0.908	0.754	0.919	0.917	0.799	0.589	0.587	0.441	0.742	0.763	0.549
(6)	✔	✔	✔	✔	**0.916**	**0.919**	**0.758**	**0.923**	**0.922**	**0.804**	**0.609**	**0.608**	**0.444**	**0.750**	**0.770**	**0.559**

ation perspectives. Vision-language pre-training methods such as CLIPscore [22] and BLIPscore [37] demonstrate moderate performance but are still surpassed by more specialized and fine-tuned VQA models. Specifically, deep learning-based models like FAST-VQA [70] and DOVER [71] achieve more competitive performances after fine-tuning. However, they are still far away from satisfactory. Notably, most VQA models perform better on quality evaluation than on text-video correspondence, as they lack text prompts input used in video generation, making it challenging to extract relation features from the AI-generated videos, which inevitably leads to the performance drop. Finally, the performance exploration of recent LMMs on our database shows that current LMMs are able to produce meaningful evaluations, which can motivate future works to further explore the use of LMMs for AIGV assessment.

The proposed AIGV-Assessor achieves the best performance compared to the competitors for both MOS prediction and pair ranking tasks in terms of all four dimensions. To further validate the effectiveness and generalizability of our proposed model, we also evaluate it on four other AIGVQA datasets [13, 31, 43, 84]. From Tables 4-6, we observe that AIGV-Assessor consistently achieves the best performance across these datasets. As shown in Figure 6, AIGV-Assessor achieves the highest overlap in area with Ground Truth (GT), indicating that AIGV-Assessor can reliably perform T2V model benchmarking, outperforming other assessment models in discerning quality differences in AI-generated videos.

5.3. Ablation Study

We conduct ablation experiments to verify the effectiveness of the main components in our AIGV-Assessor method, including the spatial feature, the temporal feature, the quality level, and the LoRA finetuning strategy. Additionally, we assess how each feature contributes to the performance across different quality dimensions. The results of these experiments are summarized in Table 7. Experiments (1), (2), and (3) validate the effectiveness of the quality regression module and the LoRA finetuning strategy, confirming that fine-tuning and quality regression significantly enhance model performance over only regressing the generated text outputs from the LLM. The addition of temporal features, as seen in Experiments (4), (5), and (6), significantly improves model performance. Experiment (6), which integrates all components, yields the best overall performance, showing that the combination of spatial and temporal features, quality level prediction, and LoRA finetuning provides the most robust and accurate AIGV assessment.

6. Conclusion

In this paper, we study the human visual preference evaluation problem for AIGVs. We first construct AIGVQA-DB, which includes 36,576 videos generated based on 1048 various text-prompts, with the MOSs and pair comparisons evaluated from four perspectives. Our detailed manual evaluations reflect different aspects of human visual preferences on AIGVs and reveal critical insights into the strengths and weaknesses of various text-to-video models. Based on the database, we evaluate the performance of state-of-the-art quality evaluation models and establish a new benchmark, revealing their limitations in measuring the perceptual preference of AIGVs. Finally, we propose AIGV-Assessor, a novel VQA model that leverages the capabilities of LMMs to give quality levels, predict quality scores, and compare preferences from four dimensions. Extensive experiments demonstrate that AIGV-Assessor achieves state-of-the-art performance on both AIGVQA-DB and other AIGVQA benchmarks, validating its robustness in understanding and evaluating the AI-generated videos.

References

[1] Hotshot-XL. https://github.com/hotshotco/hotshot-xl, 2023. 2, 3

[2] Floor33. https://discord.gg/EuB9KT6H, 2023. 2, 3

[3] Gemo. https://www.genmo.ai, 2024. 2, 3

[4] Gen2. https://research.runwayml.com/gen2, 2024. 2, 3

[5] Moonvalley. https://moonvalley.ai, 2024. 2, 3

[6] Morph studio. https://www.morphstudio.com, 2024. 2, 3, 5

[7] Sora. https://openai.com/research/video-generation-models-as-world-simulators, 2024. 2, 3, 5

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1728–1738, 2021. 2, 3

[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 5

[10] Yuqin Cao, Xiongkuo Min, Yixuan Gao, Wei Sun, Weisi Lin, and Guangtao Zhai. Unqa: Unified no-reference quality assessment for audio, image, video, and audio-visual content. *arXiv preprint arXiv:2407.19704*, 2024. 1

[11] Yuqin Cao, Xiongkuo Min, Yixuan Gao, Wei Sun, and Guangtao Zhai. Agav-rater: Adapting large multimodal model for ai-generated audio-visual quality assessment. *arXiv preprint arXiv:2501.18314*, 2025. 1

[12] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1, 2

[13] 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. *arXiv preprint arXiv:2406.06087*, 2024. 1, 3, 7, 8

[14] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6, 7

[15] Iya Chivileva, Philip Lynch, Tomas E Ward, and Alan F Smeaton. Measuring the quality of text-to-video model outputs: Metrics and dataset. *arXiv preprint arXiv:2309.08009*, 2023. 3

[16] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 16890–16902, 2022. 3

[17] Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet. Confusing image quality assessment: Toward better augmented reality experience. *IEEE Transactions on Image Processing (TIP)*, 31: 7206–7221, 2022. 4

[18] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, et al. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 5

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[21] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 3, 5

[22] 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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528, 2021. 1, 6, 7, 8

[23] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1, 2, 3

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 6, 7

[26] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, 2021. 6, 7

[27] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023. 1, 2, 3

[28] Yuval Kirstain, Adam Poliak, Uriel Singer, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 7

[29] 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. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 36652–36663, 2023. 3

[30] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing (TIP)*, 28(12):5923–5938, 2019. 7

[31] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dateset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024. 1, 3, 7, 8

[32] Debarati Kundu, Deepti Ghadiyaram, Alan C Bovik, and Brian L Evans. Large-scale crowdsourced study for tone-mapped hdr pictures. *IEEE Transactions on Image Processing (TIP)*, pages 4725–4740, 2017. 7

[33] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(9):5944–5958, 2022. 1, 6, 7

[34] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2023. 3

[35] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. ACM, 2019. 1, 6, 7

[36] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 6, 7

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022. 1, 6, 7, 8

[38] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, 2016. 2, 3

[39] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19401–19411, 2024. 3

[40] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6, 7

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 7

[42] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023. 3, 7

[43] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 3, 6, 7, 8

[44] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218, 2023. 1, 2, 3

[45] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2024. 6, 7

[46] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, and Chang Wen Chen. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia (TMM)*, pages 2049–2062, 2017. 6, 7

[47] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting (TBC)*, pages 508–517, 2018. 6, 7

[48] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing (TIP)*, pages 4695–4708, 2012. 6, 7

[49] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters (SPL)*, pages 209–212, 2012. 6, 7

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1

[53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Proceedings of the Ad-*

vances in Neural Information Processing Systems (NeurIPS), pages 25278–25294, 2022. 1, 6, 7

[54] BT Series. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, pages 500–13, 2012. 4

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 2, 3

[56] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 7

[57] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia (ACMMM)*, page 856–865, 2022. 1, 6, 7

[58] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 856–865, 2022. 7

[59] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 2023. 7

[60] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing (TIP)*, 30:4449–4464, 2021. 7

[61] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1

[62] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2555–2563, 2023. 7

[63] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CAAI International Conference on Artificial Intelligence (CICAI)*, pages 46–57. Springer, 2023. 3

[64] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2, 3

[65] Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Quality assessment for ai generated images with instruction tuning. *arXiv preprint arXiv:2405.07346*, 2024. 1

[66] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7

[67] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2, 3

[68] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[69] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024. 5, 6

[70] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fastvqa: Efficient end-to-end video quality assessment with fragment sampling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 538–554. Springer, 2022. 1, 6, 7, 8

[71] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1, 6, 7, 8

[72] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 6, 7

[73] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633, 2023. 1, 2, 3

[74] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 7

[75] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023. 3

[76] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[77] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing (TIP)*, pages 4444–4457, 2016. 6, 7

[78] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 6, 7

[79] Zitong Xu, Huiyu Duan, Guangji Ma, Liu Yang, Jiarui Wang, Qingbo Wu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Harmonyiqa: Pioneering benchmark and model for image harmonization quality assessment. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, 2025. 1

[80] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 995–1002, 2013. 6, 7

[81] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 1

[82] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 6

[83] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7

[84] Zhichao Zhang, Xinyue Li, Wei Sun, Jun Jia, Xiongkuo Min, Zicheng Zhang, Chunyi Li, Zijian Chen, Puyi Wang, Zhongpeng Ji, et al. Benchmarking aigc video quality assessment: A dataset and unified model. *arXiv preprint arXiv:2407.21408*, 2024. 1, 3, 7, 8

[85] Tianwei Zhou, Songbai Tan, Wei Zhou, Yu Luo, Yuan-Gen Wang, and Guanghui Yue. Adaptive mixed-scale feature fusion network for blind ai-generated image quality assessment. *IEEE Transactions on Broadcasting (TBC)*, 2024. 1