AIGC-VQA: A Holistic Perception Metric for AIGC Video Quality Assessment

Yiting Lu[1], Xin Li[1], Bingchen Li[1], Zihao Yu[1],
Fengbin Guan[1], Xinrui Wang[1], Ruling Liao[2],Yan Ye[2], Zhibo Chen[1(†)]
[1]University of Science and Technology of China, [2]Alibaba Group
{luyt31415,lixin666,lbc31415926}@mail.ustc.edu.cn,
{yuzihao, guanfb,wxrui_18264819595}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn
{ruling.lrl, yan.ye}@alibaba-inc.com

Abstract

With the development of generative models, such as the diffusion model, and auto-regressive model, AI-generated content (AIGC) is experiencing an explosive growth. Moreover, existing quality metrics extracted from fixed pre-trained models struggle to align accurately with human perception. There is an urgent need for an adaptive metric capable of gauging the multiple critical factors (i.e., technical quality, aesthetic quality, and video-text alignment) related to quality within AIGC videos, to provide quality assessment and guide optimization of generative models. In this work, we propose a holistic metric for AIGC video quality assessment, termed AIGC-VQA, which contains three functional branches for the cooperation on technical, aesthetic, and video-text alignment aspects in AIGC videos. Specifically, to efficiently transfer the knowledge of image-text alignment to the video-text alignment, we introduce the spatial-temporal adapter to exploit the pre-trained prior from a large-scale image-text model and achieve the temporal knowledge adaptation. Besides, we propose a divide-and-conquer training strategy for progressive cooperation on multiple branches. Due to the holistic perception ability, our proposed AIGC-VQA obtains state-of-the-art results on the T2VQA-DB dataset.

1. Introduction

Recent years have witnessed the remarkable advancement of AI-generated content, with several generative models [4–6, 39, 40, 49] demonstrating significant potential in assisting human creativity. As the representative AIGC task, text-to-video models [5, 6, 10, 49] have shown their ability to produce inventive content based on textual descriptions. These AI-generated outputs span a wide range of scenarios, from simple animations to complex, lifelike scenes [49]. However, as shown in Fig. 2, AIGC often suffer from counterfactual textures or features that do not align with human's understanding of the world (i.e., the unnatural wave ripples and twisted, terrifying faces), leading to quality degradation distinct from those found in natural content or user-generated content (UGC).

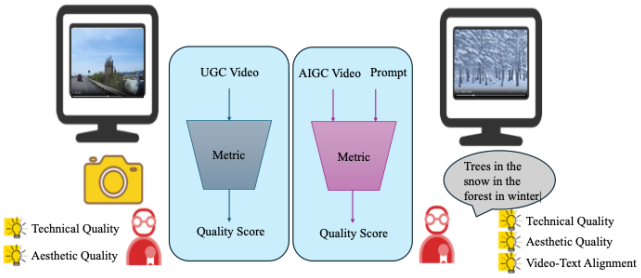

Figure 1. The difference of UGC video quality assessment metric and AIGC video quality assessment.

There are some metrics especially designed for text-to-video generative models, which are based on distribution distance or extracted features from fixed models. However, they are not suitable for the complex contents generated by the rapidly developed and incremental models due to the lack of flexibility [60] , and have shown a large gap with human perception. Meanwhile, some benchmarks for AIGC evaluation [3, 8, 26, 27] have been proposed to compare the AIGC videos generated from multiple text-to-video models by assessing various dimensions. For instance,VBench [8] divide the quality of AIGC videos into many fine-grained dimensions from a hierarchical perspective through multiple offline tools. However, to build a benchmark for comparing existing generated models, large-scale human annotations at a fine-grained level are labor-intensive for extensive subjective evaluation.

As shown in Fig. 1 and Fig. 2, in contrast to user-generated content (UGC) videos, AI-generated content

† Corresponding authors.

(AIGC) videos present two main challenges. The first challenge is the inconsistency of generated artifacts in AIGC videos with authentic or synthetic distortions commonly found in natural images. The second challenge is to evaluate the alignment between the generated frame and the textual descriptions provided by the user. To overcome these challenges, there is a need for a holistic perception metric specifically designed for assessing the quality of text-guided generated videos.

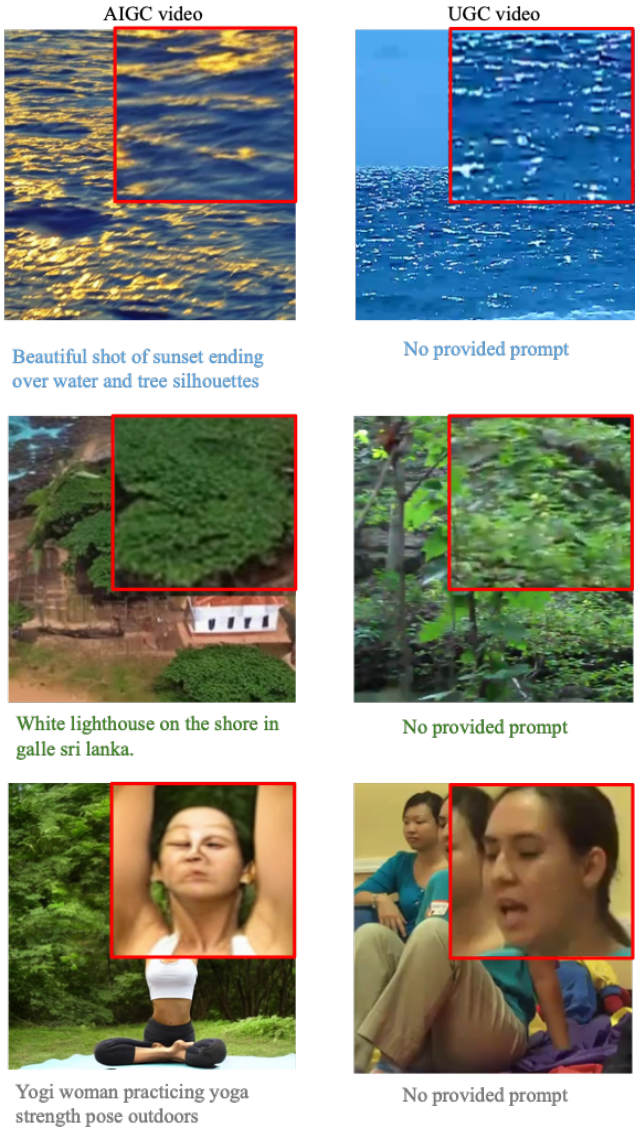

Figure 2. The difference of spatial distortion in AIGC videos and UGC videos.

From existing works [15, 60, 66], we can find that the key aspects of AIGC video quality assessment usually consist of three parts: technical quality, aesthetic quality, and video-text alignment score. Typically, technical quality [51] measures the perception of distortions existing in inter-frame spatial content and cross-frame temporal consistency. Aesthetic quality [9, 52, 60] can capture artistic factors perceived by humans from an aesthetic perspective, which is more biased to the compositional layout, colorfulness, and non-toxic content. Video-text alignment [13, 38] is important for the text-to-video generation model, which can capture the mismatch between video content and textual prompt.

In this work, to tackle the multiple aspects of quality assessment for AIGC videos, we propose a holistic perception metric, termed AIGC-VQA, to assess the quality of AIGC videos. Our proposed AIGC-VQA contains three functional branches for cooperation in technical, aesthetic, and alignment aspects. We employ a pre-trained, disentangled framework, Dover [52], to separately evaluate technical and aesthetic quality. This framework feeds the fragment-sampled input into the technical branch, while the aesthetic branch is leveraged to capture aesthetic features through resized inputs. For video-text alignment, two points should be fulfilled: the first is utilizing a vision-language model that concurrently harnesses image-text knowledge with both semantic and quality awareness, enabling efficient adaptation for downstream tasks. The second is to efficiently adapt the image-text model for discerning temporal information. For the first aspect, we introduce the BLIP [17] pretrained after semantic-aware image-text pretraining and quality-aware image-text pretraining. For the second aspect, the spatial-temporal adapter is inserted between the visual encoder and text encoder in BLIP, transferring the knowledge of image-text alignment to video-text alignment.

To maximize the effectiveness of each branch, we propose a divide-and-conquer training strategy for progressively optimizing our proposed AIGC-VQA. Initially, we optimize features that do not require textual information, such as technical and aesthetic quality. Subsequently, we fine-tune partial parameters of the vision-language model and spatial-temporal adapter for the alignment score. Finally, the three branches are ensembled and a few parameters are fine-tuned to optimize cooperative learning for multiple aspects.

Finally, the contributions of this paper are summarized as follows:

- We propose a holistic perception metric termed AIGC-VQA to assess the quality of AIGC video through a more comprehensive perspective, containing technical, aesthetic, and alignment aspects.
- We aim to progressively optimize our AIGC-VQA through a divide-and-conquer training strategy, which can enhance cooperation among multiple aspects for AIGC video quality assessment.
- Experiment results of our proposed AIGC-VQA achieve superior performance compared with other methods.

2. Related Work

2.1. AIGC IQA

AIGC image quality assessment (AIGC IQA) is required to assess the perceptual quality and text-image alignment [15, 48, 60, 63, 65], which is different from natural IQA [19, 22–24, 30, 31, 33, 36, 41, 45, 46, 67, 69]. Accordingly, the mainstream approaches to AIGC-IQA can be categorized into two types: fixed-model based methods [35, 38, 42] and fine-tuned-based methods [11, 60].

The first category, fixed-model based methods, consistently employ pre-trained models to assess perceptual quality or image-text alignment. Perceptual quality is often evaluated using metrics such as the Fréchet Inception Distance (FID) [35] and Kernel Inception Distance (KID) [1], which measure the divergence between the distribution of AIGC images and that of natural images. For text-image alignment, pre-trained vision language models like CLIP [18, 37, 38] are commonly used to compute the cosine similarity between text and image features. However, these metrics frequently fail to align with human perception or to measure alignment accurately due to their lack of adaptability.

The second approach, finetuned-based methods, relies on the construction of a large-scale AIGC quality assessment database. These methods typically finetune widely-used vision-language models by training on the large-scale database for optimization. Pic-a-pic [11] designs a quality metric of CLIP [37] based on a large-scale AIGC IQA database, selecting preferences from pairs of images. It can obtain the ranking of all generated images based on a text prompt. ImageReward [60] constructs a dataset by scoring multiple images generated from the same text prompt, ensuring prompt diversity while considering image-text consistency, image fidelity, and image harmlessness. The quality of images is evaluated by extracting representations through BLIP [17]. Nevertheless, these methods do not provide multi-functional capabilities for assessing AIGC quality across multiple aspects, such as technical quality, aesthetic quality, and text-to-image alignment.

2.2. AIGC VQA

Building upon the advances in text-to-image generation, text-to-video generation has recently seen rapid development with the emergence of models such as SORA. Analogous to AIGC IQA, there are three main streams for AI-generated content video quality assessment (AIGC VQA): fixed-model-based methods [42], offline tools-based methods, and finetuning-based methods.

Within the first stream, the Inflated-3D Convnets (I3D) [2] and Fréchet Video Distance (FVD) [47] are the prevalent metrics for gauging the perceptual quality of AIGC video. CLIPsim [37], commonly employed in AIGC-IQA, assess text-frame alignment within AIGC videos. Specifically for text-video alignment, ViCLIP [50] extends CLIP's capabilities from image to video features. Meanwhile, Chivileva et.al. [3] leverage cycle consistency from the generative and caption models to evaluate text-video alignment and ensemble it with multiple naturalness metrics. In the second stream, several benchmarks [8, 26, 27] have been established for text-to-video generation. Vbench [8] utilizes multiple offline tools acting as individual metrics across multiple fine-grained dimensions. However, these offline tools merely offer perception metrics at different semantic levels, which do not afford the necessary precision for assessing the quality of AIGC videos. As the construction of large-scale video databases [13] progresses, finetuned-based methods [13] have been proposed to regress the quality annotations of AIGC videos through representation learning. Nonetheless, these methods fall short of providing holistic functionalities needed for the comprehensive assessment of AIGC video quality, such as evaluating perceptual distortions across spatial and temporal dimensions, measuring aesthetic preferences, and ensuring accurate text-video alignment for more precise content generation.

Nowadays, video quality assessment for natural video has obtained significant progress due to the development of deep learning [14, 20, 32, 51, 52, 54, 64] and large multi-modal model [56–58]. With the help of deep learning, many metrics for video quality assessment have achieved exceptional performance on common UGC databases [7, 32, 62] through specially designed spatio-temporal networks and video sampling. With the advent of multimodal models [17, 21, 37, 61], advancements can be achieved not only in image and video quality assessment [57] but also in generating interpretable textual descriptions [56, 58], thus better aligning with the human subjective quality reasoning process.

3. Our Proposed Method

The whole framework of AIGC-VQA is shown in Fig. 4. To tackle the multiple aspects in the AIGC video quality assessment: technical quality, aesthetic quality, and video-text alignment score, we propose a holistic perception metric for AIGC Video Quality Assessment (termed as AIGC-VQA), fulfilling the versatile capability for the AIGC-VQA assisted with different functionalities, which is achieved through the collaboration of multi-dimensional branches.

3.1. Technical Quality

To circumvent distortions caused by resizing that could lead to erroneous model judgments, the technical quality aims to measure low-level distortions in localized areas of original

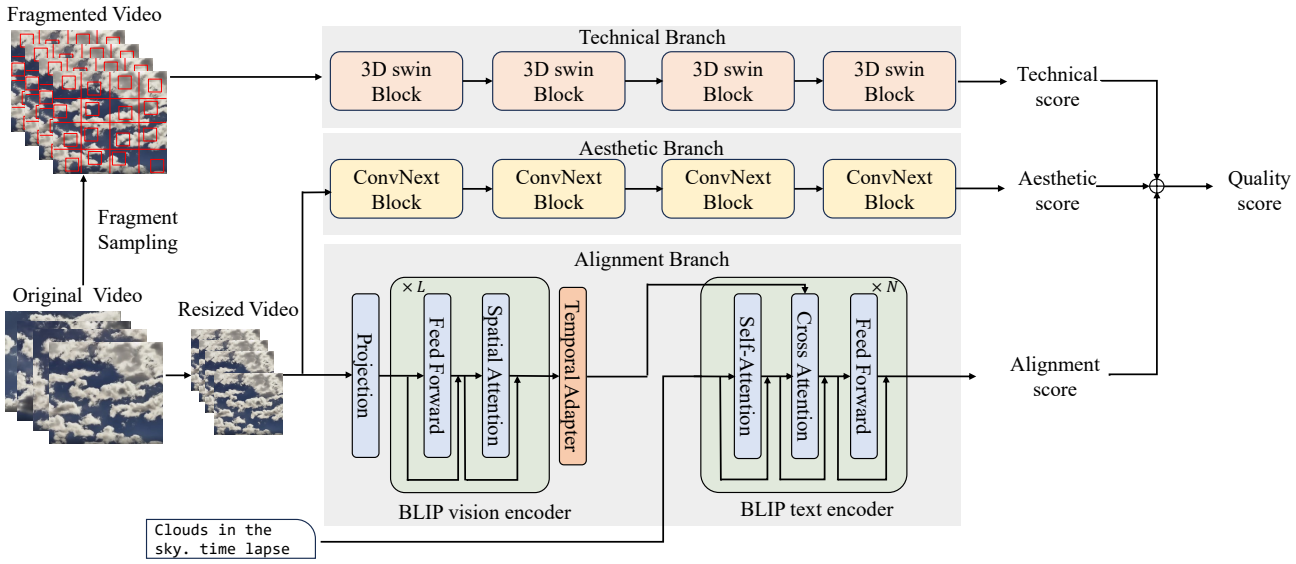

Figure 3. The overview for our proposed AIGC-VQA, which includes three multi-dimensional branches to achieve the holistic perception for AIGC video quality assessment. At the top lies the aesthetic branch, the technical branch occupies the middle layer, and at the bottom is the temporal-adapted alignment branch.

videos, providing the model with texture-related information for guidance.

We adopt the powerful 3D-Swin Transformer as our Technical Branch $\mathcal{F}_t(\cdot)$ for quality regression, and its effectiveness has been validated in a series of works [51, 54, 59, 68]. In the context of estimating the technical quality score for a video X in the shape of $T \times H \times W$, we utilize the fragment-input strategy for extracting low-level local features, which is particularly effective in capturing texture-related distortion at the local level. These local features are derived from fragments denoted as $\hat{X} \in \mathbb{R}^{T,N,h,w}$, which are sampled from the original video X. Here, h and w represent the height and width of the mini-patch within the fragments, respectively. And the number of mini-patches in a single frame is given by $N = \frac{H}{h} \times \frac{W}{w}$. Therefore, the technical score map of the AIGC video can be represented as:

$$Q_t = \mathcal{G}_t(\mathcal{F}_t(\hat{X})) \tag{1}$$

Where $\mathcal{G}_t$ denotes the regression head in Technical Branch, and the shape of the technical score map Q_t is $\frac{T}{2} \times \frac{\sqrt{N}h}{8} \times \frac{\sqrt{N}w}{8}$. And the final technical score of Technical Branch is obtained through the average pooling on the score map along the temporal dimension and spatial dimension:

$$q_t = Avg(Q_t) \tag{2}$$

The Technical Branch allows for an efficient extraction of low-level local features, capturing the unique local distortions present in AIGC videos. Consequently, it facilitates the estimation of the technical quality aspect.

3.2. Aesthetic Quality

Aesthetic quality is a sophisticated task as its subjective visual appeal and artistic factors [52]. Aesthetic appeal significantly influences overall quality perception, which contains higher-level semantic attributes compared to technical quality, such as content, composition, lighting, color, and camera trajectory as noted by [53, 55]. And AI-generated content is highly related to aesthetic evaluation, which is consistently integrated into the generation process [16] and serves as a pivotal aspect in benchmarks [43, 60, 66] for text-to-image or text-to-video generation.

To distinguish it from the functionality of technical quality, we follow the setup of Dover [52] for the Aesthetic Branch by utilizing resized videos $\widetilde{X}$ as input. The resize approach for AIGC videos, by altering distortions without changing semantics, allows for the acquisition of distortion-invariant representations, thus better serving the capture of aesthetic quality representation. And following works [52, 68], The 3D ConvNext [28] model can be introduced as Aesthetic Branch $\mathcal{F}_a(\cdot)$ due to its appealing performance and efficiency for representation.

With the same strategy as the Technical Branch, we can obtain the aesthetic score map:

$$Q_a = \mathcal{G}_a(\mathcal{F}_a(\widetilde{X})) \tag{3}$$

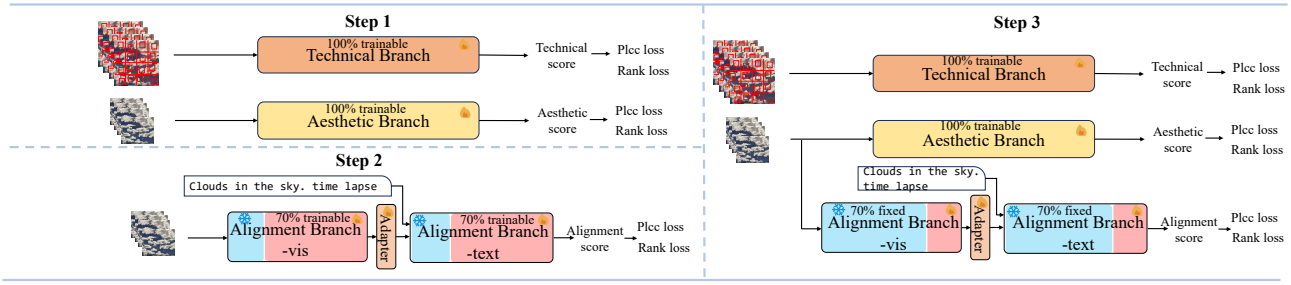

Figure 4. The overview for our training strategy for AIGC-VQA, which includes three steps to progressively adapt the pretrained model to perceive the multiple ascpects of AIGC video quality assessment.

Where $\mathcal{G}_a$ denotes the regression head in the Aesthetic Branch. And the final aesthetic score can be represented as $q_a = Avg(Q_a)$.

3.3. Alignment Score

Video-text alignment score which measures the similarity between provided prompt and generated content is the critical part of quality assessment for AIGC video. The mismatch extent of the generated visual and provided prompt can illustrate the multimodal understanding ability of generation model [12]. Video-text alignment encompasses numerous sub-dimensional factors [60], including attributes, color, objects, background, and motion actions. This necessitates a model capable of comprehending the semantic information in text prompts and capturing the corresponding semantic visual and temporal features in AIGC videos.

Consequently, Alignment Branch must exhibit exceptional multimodal understanding capabilities, aligning features in both the textual space and visual space while also be capable of extracting text-guided visual features. Furthermore, it should perceive temporal characteristics to address the temporal alignment issues.

Based on the first two requirements, we utilize BLIP [17], a powerful vision-language model that has undergone two-stage pre-training, as our foundational framework of Alignment Branch. The initial stage involves pre-training on a web-scale corpus of image-text pairs, focusing on image-text alignment, matching, and image-anchored text captioning to bolster vision-language understanding and generation capacities. The second stage is fine-tuning the part of parameters of BLIP on the ImageReward [60] dataset, which contains 137k pairs of expert comparisons for AIGC image quality assessment. By loading the weights from this two-stages pre-training, Alignment Branch can obtain both knowledge of semantic-aware image-text alignment and quality-ware image-text alignment.

As for the temporal characteristics, we aim to efficiently adapt large pre-trained image-text model for video-text tasks due to the less abundant data for AIGC video quality assessment. To facilitate the knowledge transfer from image-text tasks to video-text tasks, we introduce the spatial-temporal Adapter (ST-Adapter) [34], an efficient and effective mechanism designed to harness the pre-trained knowledge from a large-scale image-text model (i.e., BLIP), thereby enabling enhanced video-text understanding with minimal parameter overhead. Also, the Alignment Branch should capture the semantic-related feature, which is invariant to distortion. Therefore, the resized video $\widetilde{X}$ will be fed into Alignment Branch $\mathcal{H}_a(\cdot)$. And the modeling of video-text alignment can be denoted as:

$$s_a = \mathcal{G}_a(\mathcal{H}_a^{tex}(Tex, f_{st}(\mathcal{H}_a^{vis}(\widetilde{X})))[0, ...]) \quad (4)$$

In which, Tex is referred to the provided text prompt, and $\mathcal{H}_a^{tex}(\cdot)$ and $\mathcal{H}_a^{vis}(\cdot))$ can be denoted as visual-guided text encoder and vision encoder, separately. And we apply the eos token from the output of $\mathcal{H}_a^{tex}(\cdot)$ to be fed into regression head $\mathcal{G}_a$ for video-text alignment. f_{st} represents the ST-Adapter, the equation of ST-Adapter is as follows:

$$f_{st}(r) = r + f(\text{DW3D}(rW_{\text{down}}))W_{\text{up}}, \quad (5)$$

Where DW3D denotes the depth-wise 3D convolution for spatial-temporal representation capture. And W_{down} and W_{up} are weight of downscaling and upscaling. r is the representation after visual encoder. $f(\cdot)$ is the activation function.

3.4. Training Strategy

To maximize the effectiveness of each branch, we propose a divide-and-conquer training strategy. Initially, we focus on enhancing features unrelated to text, such as technical and aesthetic qualities. This stage involves no textual input, refining the purely visual dimensions. Subsequently, we fine-tune the partial parameters of BLIP and ST adapter for video-text alignment. Finally, we synergize the three branches, conducting fine-tuning on a limited set of parameters to foster cooperative learning on the multiple aspects (i.e., technical quality, aesthetic quality, and video-text alignment score).

In particular, we train the Technical Branch and Aesthetic Branch with loading the pre-trained weight from LSVQ [52, 62]. Then the Alignment Branch is trained with 70% unfixed parameters, loading the pretrained weight from ImageReward. Note that these datasets are not involved in training with the AIGC video quality assessment dataset. Finally, we finetuned the Technical Branch, Aesthetic Branch, and Alignment Branch with 30% unfixed parameters for late fusion. After obtaining the final score of AIGC videos by aggregating technical quality, aesthetic quality and video-text alignment score: $s = q_t + q_a + s_a$. For each branch within our architecture, we apply a dual-loss optimization strategy [14], incorporating both Pearson linear correlation coefficient (PLCC) and ranking loss.

$$L_{plcc} = \frac{\left(1 - 1\frac{\sum_{i=1}^{m}(s_i-\bar{s})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(s_i-\bar{s})^2}\sqrt{\sum_{i=1}^{m}(y_i-\bar{y})^2}}\right)}{2} \tag{6}$$

$$L_{\text{rank}} = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\Big(\max\left(0, |y_i - y_j| - e(y_i, y_j)\cdot(s_i - s_j)\right)\Big) \tag{7}$$

where y and $\bar{y}$ represent the quality annotations and the mean value of quality annotations. s and $\bar{s}$ represent the quality predictions and the mean value of quality predictions. This optimization approach refines each branch by aligning each branch's outputs with human perceptual judgments and ensuring that the rank order of the outputs adheres to the expected ordinal characteristics.

4. Experiment

4.1. AIGC-VQA Databases

T2VQA-DB: The dataset T2VQA-DB [13] encompasses 10,000 videos and 1000 prompts, of which these generated videos come from 10 text-to-video models. All these videos have the shape of 512×512 and 4 fps. And all the text prompts can be divided into 6 categories: nature, human, artificial, animal, object and abstract. The Mean Opinion Score (MOS) of T2VQA-DB ranges from 0 to 100. T2VQA-DB is a relatively large dataset for AIGC video quality assessment compared to works [3, 8, 26, 27]. We divided the data excluding the test part into the random training set and the validation set, and ensured that the category proportions of the generative models in the two sets were consistent. The split strategy was randomly conducted 10 times division, and we finally announced the average result of 10 times.

4.2. Implementation Details

For the technical branch, the input fragments are of size $32 \times 244 \times 244$ with a 1-frame interval, consisting of (7×7) fragments, each of size 32. For the aesthetic branch and alignment branch, The original video is resized into the spatial dimension of 224×224. For all the branches, the number of temporal frames is 16. We adopt two common-used criteria for performance evaluation: Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SROCC). A higher value means a better correlation with human annotations. All experiments are implemented on four 32G V100 GPUs. In the training process, we utilize AdamW optimizer [29] with a learning rate of $3 \times e^{-5}$ and a weight decay of 0.05 for optimization. And batchsize set as 8. Each stage contains 50 epoches for training.

4.3. Experiment Results

To verify the effectiveness of our proposed AIGC-VQA, We select five fixed-model based methods (CLIP-sim [38], BLIP [17], ImageReward [60], ViCLIP [50], UMTScore [27], four UGC model based methods (SimpleVQA [44], BVQA [14], FastVQA [51], Dover [52], and KSVQE [32]), and one finetuned-model based method (T2VQA [13]).

We can see that fixed-model based methods fail to handle the T2VQA-DB dataset since they have less flexibility for quality assessment on complex AIGC video. And UGC model based methods still face performance limitations on AIGC video quality assessment, due to the lack of measurement on video-text alignment. Specifically, our proposed AIGC demonstrates superior performance on T2VQA-DB datasets. Our AIGC-VQA outperforms the second-best method T2VQA with performance gain of 0.0276 and 0.0466 on SROCC and PLCC. It can illustrate that our proposed AIGC-VQA can fulfill the holistic perception ability on multiple key factors for AIGC video quality assessment (i.e., technical quality, aesthetic quality, and video-text alignment).

4.4. The Results in NTRIE2024 AIGC VQA track

The results of NTRIE2024 AIGC VQA track [25] in the testing phase is shown in Table 3. Our proposed AIGC-VQA based on enhancing the holistic perception ability of quality metric for AIGC videos, has achieved the optimal results. It illustrates that our proposed AIGC-VQA achieves the good evaluation ability on AIGC videos.

4.5. Ablation Study

To analyze the effects of the main components in our proposed AIGC-VQA. In this section, we conduct multiple ablation studies to study the effect of the technical branch, aes-

Table 1. Performance of existing SOTA methods and the proposed AIGC-VQA on T2VQA-DB dataset. The best and second-best results are bolded and underlined.

Type	Models	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
fixed	CLIPSim [38]	0.1047	0.1277	0.0702	21.683
	BLIP [17]	0.1659	0.1860	0.1112	18.373
	ImageReward [60]	0.1875	0.2121	0.1266	18.243
	ViCLIP [50]	0.1162	0.1449	0.0781	21.655
	UMTScore [27]	0.0676	0.0721	0.0453	22.559
UGC	SimpleVQA [44]	0.6275	0.6338	0.4466	11.163
	BVQA [14]	0.7390	0.7486	0.5487	15.645
	FAST-VQA [51]	0.7173	0.7295	0.5303	10.595
	DOVER [52]	0.7609	0.7693	0.5704	9.8072
	KSVQE [32]	0.7709	0.7842	0.5936	11.053
finetuned	T2VQA [13]	<u>0.7965</u>	<u>0.8066</u>	<u>0.6058</u>	**9.0221**
Ours	**AIGC-VQA**	**0.8241**	**0.8352**	**0.6474**	9.6617

Table 2. Ablation study on the usage of technical branch, aesthetic branch and alignment branch in AIGC-VQA.

Technical Branch	Aesthetic Branch	Alignment Btanch	SRCC	PLCC	Main Score
✔	✔	✔	**0.8241**	**0.8352**	**0.8296**
✔	✗	✗	0.7465	0.7614	0.7539
✗	✔	✗	0.7322	0.7488	0.7400
✗	✗	✔	0.8096	0.8203	0.814
✔	✔	✗	0.7623	0.7733	0.7678
✔	✗	✔	0.8166	0.8298	0.8231

Table 3. Our proposed AIGC-VQA achieved the optimal performance in NTRIE2024 AIGC VQA track [25], according to the main score.

Team Name	Main Score	Ranking
ours(IMCL-DAMO)	**0.8385**	**1**
Kwai-kaa	0.824	2
SQL	0.8232	3
musicbeer	0.8231	4
finnbingo	0.8211	5
PromptSync	0.8178	6
QA-FTE	0.8128	7
MediaSecurity_SYSU&Alibaba	0.8124	8
IPPL-VQA	0.8003	9
IVP-Lab	0.7944	10
Oblivion	0.7869	11
CUC-IMC	0.7802	12
UBC DSL Team	0.7531	13

thetic branch, and alignment branch on T2VQA-DB dataset in Table. 2. Also, we take a deep step into the different fixed parameter ratios on the alignment branch in Table 4, and the effect of ST-Adapter for image-to-video transfer in Table 5. For the training strategy, we analyze the effect of different stages on Table 6.

Table 4. Ablation study on the unfixed ratio for optimizing alignment branch in Stage 2.

Ratio	SRCC	PLCC	Main Score
80%	0.8010	0.8200	0.8105
70%	0.8096	0.8203	0.8140
60%	0.8091	0.8232	0.8161
50%	0.8078	0.8212	0.8145
40%	0.8086	0.8223	0.8154

Table 5. Ablation study on the ST-adapter for alignment branch. The term "N/A" denotes the absence of image-to-video transfer application.

Adapter	SRCC	PLCC	Main Score
N/A	0.7910	0.8089	0.7999
ST-Adapter	0.8096	0.8203	0.8140

The effectiveness of Technical Branch. The technical quality aims to measure low-level distortions in localized areas, providing the model with texture-related information for guidance. As depicted in Table 2, we can see that the ensemble for the technical branch and aesthetic branch can boost the performance of the aesthetic branch. The former (the 5^{th} row) can exceed the latter (the 3^{th} row) with a per-

Table 6. Ablation study on the training strategies for optimizing AIGC-VQA.

Stage 1	Stage 2	Stage 3	SRCC	PLCC	Main Score
✔	✗	✗	0.7623	0.7733	0.7678
✗	✔	✗	0.8096	0.8203	0.8140
✔	✔	✗	0.8102	0.8244	0.8173
✔	✔	✔	0.8241	0.8352	0.8296

formance gain of 0.0301 and 0.0245 on SROCC and PLCC. It illustrates that the technical branch has a superior effect on AIGC-VQA.

The effectiveness of Aesthetic Branch. Aesthetic appeal significantly influences overall quality perception, especially for AIGC quality assessment [60]. From the results of Table 2, through the comparison between the 2^{th} row (technical branch) and the 5^{th} row (ensemble of technical branch and aesthetic branch), we can see that introducing of aesthetic branch can bring the performance gain of 0.0153 and 0.0119 on SROCC and PLCC. It demonstrates the necessity of aesthetic ability for AIGC VQA metric.

The effectiveness of Alignment Branch. By comparing the 4^{th} row (Technical Branch) and the 6^{th} row (the ensemble of Technical Branch and Alignment Branch), we can observe that Alignment Branch can boost the performance of Technical Branch according to the performance gain of 0.0701/0.0684 on SROCC and PLCC. The effect of the Alignment Branch is larger than the Technical Branch and Aesthetic Branch, which illustrates that the video-text alignment is crucial for the text-to-video generative model.

The effectiveness of fixed parameter ratio of alignment branch. Also, we analyze the effect of the fixed parameter ratio of the Alignment Branch on AIGC video quality assessment. From the Table. 4, we can observe that the ratio has no significant difference in performance on our Alignment Branch, which shows the robustness of the Alignment Branch during optimization.

The effectiveness of ST-Adapter of alignment branch. For efficient image-video transfer, we delete the ST-Adapter in the Alignment Branch to illustrate the effectiveness of temporal information extraction for video-text alignment. From the Table. 5, we can see that ST-Adapter can outperform the original BLIP through the performance gain of 0.0186 and 0.0114 on SROCC and PLCC.

The effectiveness of training strategy To analyze the effectiveness of our proposed progressive training strategy for optimizing AIGC-VQA, we conduct ablation studies for different stages in Table. 6. In the first phase, leveraging the decoupled pre-training approach of Dover [52], we initialize the Technical and Aesthetic Branches with pretrained weights dedicated to regressing technical and aesthetic quality, respectively. This stage does not require assistance from provided text prompts. In the second phase, we adapt from ImageReward [60]'s pretrained weights to transition from image-text alignment to video-text alignment. In the third phase, we engage in joint training of the Technical Branch, Aesthetic Branch, and Alignment Branch to facilitate cooperation across multiple aspects. From the results of Table. 6, we can summarize two conclusions. Firstly, each stage is essential and contributes to incremental performance improvements. Secondly, joint training in the third stage outperforms the direct ensemble of the results from the first two stages, indicating that joint training effectively fosters collaboration.

5. Conclusion

In this work, we address the multiple factors affecting subjective quality in AIGC videos and propose a holistic perception metric for video quality assessment. It contains multi-dimensional branches tailored for technical quality, aesthetic quality and video-text alignment. Also, we propose a divide-and-conquer training strategy for progressive optimization. And the results show the effectiveness of our proposed AIGC-VQA.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 3

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[3] 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. 1, 3, 6

[4] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 1

[5] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 1

[6] 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

[7] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *QoMEX*, pages 1–6. IEEE, 2017. 3

[8] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023. 1, 3, 6

[9] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10041–10051, 2023. 2

[10] 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*, pages 15954–15964, 2023. 1

[11] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[12] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[13] 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. 2, 3, 6, 7

[14] 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 Trans. Circuits Syst. Video Technol.*, 32(9):5944–5958, 2022. 3, 6, 7

[15] 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*, 2023. 2, 3

[16] Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Q-refine: A perceptual quality refiner for ai-generated image. *arXiv preprint arXiv:2401.01117*, 2024. 4

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 3, 5, 6, 7

[18] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *arXiv preprint arXiv:2309.13625*, 2023. 3

[19] Xin Li, Yiting Lu, and Zhibo Chen. Freqalign: Excavating perception-oriented transferability for blind image quality assessment from a frequency perspective. *IEEE Transactions on Multimedia*, 2023. 3

[20] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[22] Jianzhao Liu, Xin Li, Shukun An, and Zhibo Chen. Source-free unsupervised domain adaptation for blind image quality assessment. *arXiv preprint arXiv:2207.08124*, 2022. 3

[23] Jianzhao Liu, Xin Li, Yanding Peng, Tao Yu, and Zhibo Chen. Swiniqa: Learned swin distance for compressed image quality assessment. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 1795–1799, 2022.

[24] Jianzhao Liu, Wei Zhou, Xin Li, Jiahua Xu, and Zhibo Chen. Liqa: Lifelong blind image quality assessment. *IEEE Transactions on Multimedia*, 2022. 3

[25] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 6, 7

[26] 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. 1, 3, 6

[27] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 6, 7

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[29] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

[30] Yiting Lu, Jun Fu, Xin Li, Wei Zhou, Sen Liu, Xinxin Zhang, Wei Wu, Congfu Jia, Ying Liu, and Zhibo Chen. Rtn: Reinforced transformer network for coronary ct angiography vessel-level image quality assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 644–653. Springer, 2022. 3

[31] Yiting Lu, Xin Li, Jianzhao Liu, and Zhibo Chen. Styleam: Perception-oriented unsupervised domain adaption for non-reference image quality assessment. *arXiv preprint arXiv:2207.14489*, 2022. 3

[32] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kaleidoscope video quality assessment for short-form videos. *arXiv preprint arXiv:2402.07220*, 2024. 3, 6, 7

[33] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Trans. Image Process.*, 31: 4149–4161, 2022. 3

[34] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 5

[35] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 3

[36] G Qin, R Hu, Y Liu, X Zheng, H Liu, X Li, and Y Zhang. Data-efficient image quality assessment with attention-panel decoder. arxiv 2023. *arXiv preprint arXiv:2304.04952*. 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6, 7

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

[40] 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*, pages 10684–10695, 2022. 1

[41] Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5846–5855. IEEE, 2023. 3

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[43] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4

[44] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for UGC videos. In *ACM Multimedia*, pages 856–865. ACM, 2022. 6, 7

[45] 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*, 2023. 3

[46] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018. 3

[47] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 3

[48] 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*, pages 46–57. Springer, 2023. 3

[49] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1

[50] 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. *arXiv preprint arXiv:2307.06942*, 2023. 3, 6, 7

[51] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. FAST-VQA: efficient end-to-end video quality assessment with fragment sampling. In *ECCV (6)*, pages 538–554. Springer, 2022. 2, 3, 4, 6, 7

[52] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Disentangling aesthetic and technical effects for video quality assessment of user generated content. *CoRR*, abs/2211.04894, 2022. 2, 3, 4, 6, 7, 8

[53] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Disentangling aesthetic and technical effects for video quality assessment of user generated content. *arXiv preprint arXiv:2211.04894*, 2(5):6, 2022. 4

[54] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Discovqa: Temporal distortion-content transformers for video quality assessment. *IEEE Trans. Circuits Syst. Video Technol.*, 33(9):4840–4854, 2023. 3, 4

[55] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 1045–1054. ACM, 2023. 4

[56] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. *arXiv preprint arXiv:2311.06783*, 2023. 3

[57] 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. 3

[58] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. *arXiv preprint arXiv:2402.16641*, 2024. 3

[59] Wei Wu, Shuming Hu, Pengxiang Xiao, Sibin Deng, Yilin Li, Ying Chen, and Kai Li. Video quality assessment based on swin transformer with spatio-temporal feature fusion and data augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 1846–1854. IEEE, 2023. 4

[60] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[61] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 3

[62] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan C. Bovik. Patch-vq: 'patching up' the video quality problem. In *CVPR*, pages 14019–14029. Computer Vision Foundation / IEEE, 2021. 3, 6

[63] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, , and Zhibo Chen. Sf-iqa: Quality and similarity integration for ai generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 3

[64] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen. Video quality assessment based on swin transformerv2 and coarse to fine strategy. *arXiv preprint arXiv:2401.08522*, 2024. 3

[65] Jiquan Yuan, Xinyan Cao, Jinming Che, Qinyuan Wang, Sen Liang, Wei Ren, Jinlong Lin, and Xixin Cao. Tier: Text and image encoder-based regression for aigc image quality assessment. *arXiv preprint arXiv:2401.03854*, 2024. 3

[66] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 440–445. IEEE, 2023. 2, 4

[67] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. *CoRR*, abs/2303.00521, 2023. 3

[68] Kai Zhao, Kun Yuan, Ming Sun, and Xing Wen. Zoom-vqa: Patches, frames and clips integration for video quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 1302–1310. IEEE, 2023. 4

[69] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14143–14152, 2020. 3