NTIRE 2024 Quality Assessment of AI-Generated Content Challenge

Xiaohong Liu* Xiongkuo Min* Guangtao Zhai* Chunyi Li* Tengchuan Kou*
Wei Sun* Haoning Wu* Yixuan Gao* Yuqin Cao* Zicheng Zhang* Xiele Wu*
Radu Timofte * Fei Peng Huiyuan Fu Anlong Ming Chuanming Wang
Huadong Ma Shuai He Zifei Dou Shu Chen Huacong Zhang Haiyi Xie
Chengwei Wang Baoying Chen Jishen Zeng Jianquan Yang Weigang Wang
Xi Fang Xiaoxin Lv Jun Yan Tianwu Zhi Yabin Zhang Yaohui Li Yang Li
Jingwen Xu Jianzhao Liu Yiting Liao Junlin Li Zihao Yu Fengbin Guan
Yiting Lu Xin Li Hossein Motamednia S. Farhad Hosseini-Benvidi
Ahmad Mahmoudi-Aznaveh Azadeh Mansouri Ganzorig Gankhuyag Kihwan Yoon
Yifang Xu Haotian Fan Fangyuan Kong Shiling Zhao Weifeng Dong
Haibing Yin Li Zhu Zhiling Wang Bingchen Huang Avinab Saha
Sandeep Mishra Shashank Gupta Rajesh Sureddi Oindrila Saha Luigi Celona
Simone Bianco Paolo Napoletano Raimondo Schettini Junfeng Yang Jing Fu
Wei Zhang Wenzhi Cao Limei Liu Han Peng Weijun Yuan Zhan Li
Yihang Cheng Yifan Deng Haohui Li Bowen Qu Yao Li Shuqing Luo
Shunzhou Wang Wei Gao Zihao Lu Marcos V. Conde Radu Timofte
Xinrui Wang Zhibo Chen Ruling Liao Yan Ye Qiulin Wang Bing Li
Zhaokun Zhou Miao Geng Rui Chen Xin Tao Xiaoyu Liang Shangkun Sun
Xingyuan Ma Jiaze Li Mengduo Yang Haoran Xu Jie Zhou Shiding Zhu
Bohan Yu Pengfei Chen Xinrui Xu Jiabin Shen Zhichao Duan Erfan Asadi
Jiahe Liu Qi Yan Youran Qu Xiaohui Zeng Lele Wang Renjie Liao

Abstract

This paper reports on the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge, which will be held in conjunction with the New Trends in Image Restoration and Enhancement Workshop (NTIRE) at CVPR 2024. This challenge is to address a major challenge in the field of image and video processing, namely, Image Quality Assessment (IQA) and Video Quality Assessment (VQA) for AI-Generated Content (AIGC). The challenge is divided into the image track and the video track. The image track uses the AIGIQA-20K, which contains 20,000 AI-Generated Images (AIGIs) generated by 15 popular generative models. The image track has a total of 318 registered participants. A total of 1,646 submissions are received in the development phase, and 221 submissions are received in the test phase. Finally, 16 participating teams submitted their models and fact sheets.

The video track uses the T2VQA-DB, which contains 10,000 AI-Generated Videos (AIGVs) generated by 9 popular Text-to-Video (T2V) models. A total of 196 participants have registered in the video track. A total of 991 submissions are received in the development phase, and 185 submissions are received in the test phase. Finally, 12 participating teams submitted their models and fact sheets. Some methods have achieved better results than baseline methods, and the winning methods in both tracks have demonstrated superior prediction performance on AIGC.

1. Introduction

With the fast development of generative models, AI-Generated Content (AIGC) has become popular in daily lives. Among them, AI-Generated Images (AIGIs) and AI-Generated Videos (AIGVs) are two of the most common media. However, the quality of AIGIs and AIGVs can be varied due to the differences in performance of various

*The organizers of the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge.
The NTIRE 2024 website: https://cvlai.net/ntire/2024/.

models. Therefore, it is significant to propose efficient Image Quality Assessment (IQA) and Vmage Quality Assessment (VQA) methods to accurately predict the quality of generated images and videos

This NTIRE 2024 Quality Assessment of AI-Generated Content Challenge aims to promote the development of the I/VQA methods for generated images and videos to guide the improvement and enhancement of the performance of generative models, thereby improving the quality of experience of AIGC. The challenge is divided into the image track and the video track. In the image track, we use the AIGIQA-20K [45], which contains 20,000 AIGIs generated by 15 Text-to-Image (T2I) models. 21 subjects are invited to produce accurate Mean Opinion Scores (MOSs). The video track uses the T2VQA-DB [44], in which 9 Text-to-Vido (T2V) models are used to generate 10,000 videos. The MOSs are obtained from 27 subjects.

This is the first time that a quality assessment of AIGC challenge has been held at the NTIRE workshop. The challenge has a total of 514 registered participants, 318 in the image track and 196 in the video track. A total of 2,637 submissions were received in the development phase, while 406 prediction results were submitted during the final testing phase. Finally, 16 valid participating teams in the image track and 12 valid participating teams in the video track submitted their final models and fact sheets. They have provided detailed introductions to their I/VQA methods for AIGIs and AIGVs. We provide the detailed results of the challenge in Section 4 and Section 5. We hope that this challenge can promote the development of I/VQA methods for image and video generation.

This challenge is one of the NTIRE 2024 Workshop [1] series of challenges on: dense and non-homogeneous dehazing, night photography rendering, blind compressed image enhancement, shadow removal, efficient super-resolution, image super-resolution ($\times 4$), light field image super-resolution, stereo image super-resolution, HR depth from images of specular and transparent surfaces, bracketing image restoration and enhancement, portrait quality assessment, Restore Any Image Model (RAIM) in the wild, raW image super-resolution, short-form UGC Video quality assessment, low light enhancement, and raw burst alignment and ISP challenge.

2. Related Work

2.1. AIGI dataset

Several AIGI datasets have been proposed in recent years. Benefiting from the successful Stable Diffusion [77], DiffusionDB [98] is the first large-scale text-to-image prompt dataset, containing 14 million images generated by Stable Diffusion using prompts and hyperparameters specified by real users. HPS [110] collects 98,807 generated images from the Stable Foundation Discord channel, along with 25,205 human choices. ImageReward [112] proposes a dataset containing 137k prompt-image pairs sampled from DiffusionDB. Each pair has 3 MOSs from overall rating, image-text alignment, and fidelity. Pick-A-Pic [41] contains over 500,000 examples and 35,000 distinct prompts, Each example contains a prompt, two generated images, and a label for which image is preferred. AGIQA-1K [123], AGIQA-3K [47], and AIGCIQA2023 [93] contain 1,080, 2,982, and 2,400 images respectively. AGIN [4] collects 6,049 images and conducts a large-scale subjective study to collect human opinions on the overall naturalness. In this challenge, we use the AIGIQA-20K [45], including 20,000 images generated by 15 popular T2I models, along with the MOSs collected from 21 subjects.

2.2. AIGV dataset

Compared with AIGI datasets, the number of proposed AIGV datasets is small. Chivileva *et al*. [5] proposes a dataset with 1,005 videos generated by 5 T2V models. 24 users are involved in the subjective study. EvalCrafter [57] builds a dataset using 500 prompts and 5 T2V models, resulting in 2,500 videos in total. However, only 3 users are involved in the subjective study. Similarly, FETV [58] uses 619 prompts, 4 T2V models, and 3 users for annotation as well. VBench [36] has a larger scale with in total of $\sim$1,7k prompts and 4 T2V models. In the video track, we use the T2VQA-DB [44]. The dataset has 10,000 videos generated by 9 different T2V models. 27 subjects are invited to collect the MOSs.

2.3. IQA model

Traditional IQA models focus on distortions like noises, blurriness, semantic contents, etc. DBCNN [120] handles both synthetic and authentic distortions by training two CNN networks. StairIQA [88] proposes a staircase structure to hierarchically integrate the information from low-level to high-level. LIQE [121] proposes a general and automated multitask learning scheme to exploit auxiliary knowledge from IQA, scene classification, and distortion type identification. In the meantime, several IQA models designed for AIGIs have been proposed. HPS [110] and PickScore [41] are CLIP-based [74] models to imitate human preference on generated images. ImageReward [112] uses a BLIP-based [49] architecture to predict the image quality.

In recent years, researchers have been paying attention to using the ability of Large Multi-modality Models (LMMs) to solve IQA tasks. Q-bench [105] first investigates the performance of LMMs in evaluating visual quality. [106–108, 125] further introduce the training procedure to utilize LMMs for IQA tasks. Q-Refine [46] is a quality-awarded

[1] https://cvlai.net/ntire/2024/

refiner to guide the refining process in T2I models. The development of IQA models not only provides more accurate predictions on AIGIs quality but also benefits the development of image generation models.

2.4. VQA model

The traditional VQA models are usually designed for user-generated videos or a certain attribute of videos. [9, 19, 43, 56, 124, 126]. For example, SimpleVQA [87] trains an end-to-end spatial feature extraction network to directly learn quality-aware spatial features from video frames, and extracts motion features to measure temporally related distortions at the same time to predict video quality. FAST-VQA [100] proposes the "fragments" sampling strategies and the Fragment Attention Network (FANet) to accommodate fragments as inputs. DOVER [104] evaluates the quality of videos from the technical and aesthetic perspectives respectively. Q-Align [107] can also address the VQA task by relying on the ability of multi-modal large models.

There are several works targeting the VQA tasks of AIGVs. VBench [36] and EvalCrafter [57] build benchmarks for AIGVs by designing multi-dimensional metrics. MaxVQA [103] and FETV [58] propose separate metrics for the assessment of video-text alignment and video fidelity, while T2VQA [44] handles the features from the two dimensions as a whole. We believe the development of the VQA model for AIGV will certainly benefit the generation of high-quality videos.

3. NTIRE 2024 Quality Assessment of AI-Generated Content Challenge

We organize the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge in order to promote the development of objective I/VQA methods for AIGIs and AIGVs. The main goal of the challenge is to predict the perceptual quality of the generated images and videos. Details about the challenge are as follows:

3.1. Overview

The challenge has two tracks, *i.e*. image track and video track. The task is to predict the perceptual quality of a generated image/video based on a set of prior examples of images/videos and their perceptual quality labels. The challenge uses the AIGIQA-20K [45] and the T2VQA-DB [44] dataset and splits them into the training, validation, and testing sets. As the final result, the participants in the challenge are asked to submit predicted scores for the given testing set.

3.2. Datasets

In the image track, we use the AIGIQA-20K [45] for training, validating, and testing. The dataset contains 20,000 images generated by 15 T2I models, which are: DALLE 2 [76], DALLE 3 [76], Dreamlike [11], IF [6], LCM Pixart [66], LCM SD1.5 [66], LCM SDXL [66], Midjourney v5.2 [31], Pixart α [3], Playground v2 [71], SD1.4 [78], SD1.5 [78], SDXL [79], SDXL Turbo [82], and SSD1B [27]. Concretely, 20,000 prompts are selected from DiffusionDB [98]. For Dreamlike, Pixart α, Playground v2, SD1.4, SD1.5, SDXL, and SSD1B [3, 11, 27, 71, 78, 79], each model generates 2,000 images for their strong generalize ability. LCM Pixart, LCM SD1.5, LCM SDXL, and SDXL Turbo [66, 82] are assigned 1,000 images each, and DALLE2, DALLE3, IF, Midjourney v5.2 [6, 31, 76] are 500 images each.

In the video track, we use the T2VQA-DB [44]. The dataset contains 10,000 generated videos from: Text2Video-Zero [40], AnimateDiff [26], Tune-a-video [109], VidRD [24], VideoFusion [67], ModelScope [95], LVDM [30], Show-1 [118], and LaVie [97]. 1,000 prompts are selected from WebVid-10M [1], a large-scale text-video dataset. Each model generates one video for each prompt. For Tune-a-video [109], two different pre-trained weights are used. The video resolution is unified to 512×512, and the video length is 4s.

21 subjects are invited to rate the generated images in AIGIQA-20K [45], and 27 subjects for the videos in T2VQA-DB [44]. After normalizing and averaging the subjective opinion scores, the mean opinion score (MOS) of each image/video can be obtained. Furthermore, we randomly split the AIGIQA-20K into a training set, a validation set, and a testing set according to the ratio of $7:1:2$. The same split is conducted to the T2VQA-DB. The numbers of generated images in the training set, validation set, and testing set are $14,000$, $2,000$, and $4,000$, respectively. For the video track, the numbers are $7,000$, $1,000$, and $2,000$.

3.3. Evaluation protocol

In both tracks, the main scores are utilized to determine the rankings of participating teams. We ignore the sign and calculate the average of Spearman Rank-order Correlation Coefficient (SRCC) and Person Linear Correlation Coefficient (PLCC) as the main score:

$$\text{Main Score} = (|\text{SRCC}| + |\text{PLCC}|)/2. \quad (1)$$

SRCC measures the prediction monotonicity, while PLCC measures the prediction accuracy. Better VQA methods should have larger SRCC and PLCC values. Before calculating PLCC index, we perform the third-order polynomial nonlinear regression. By combining SRCC and PLCC, the main scores can comprehensively measure the performance of participating methods.

Table 1. Quantitative results for the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge: Track 1 Image.

Rank	Team	Leader	Main Score	SRCC	PLCC
1	pengfei	Fei Peng	0.9175	0.9076	0.9274
2	MediaSecurity_SYSU&Alibaba	Huacong Zhang	0.9169	0.9076	0.9262
3	geniuswwg	Weigang Wang	0.9157	0.9051	0.9264
4	Yag	Zihao Yu	0.9138	0.9009	0.9268
5	QA-FTE	Tianwu Zhi	0.9091	0.8982	0.9201
6	HUTB-IQALab	Junfeng Yang	0.9087	0.8957	0.9218
7	IQ Analyzers	Avinab Saha	0.9065	0.8912	0.9217
8	PKUMMCAL	Haohui Li	0.9044	0.8933	0.9155
9	BDVQAGroup	Yifang Xu	0.9023	0.8926	0.9119
10	JNU_620	Weijun Yuan	0.8835	0.8746	0.8923
11	MT-AIGCQA	Li Zhu	0.8736	0.8589	0.8883
12	IVL	Luigi Celona	0.8715	0.8486	0.8944
13	CVLab	Zihao Lu	0.8657	0.8522	0.8792
14	z6	Ganzorig Gankhuyag	0.8628	0.8472	0.8785
15	Oblivion	Shiling Zhao	0.8613	0.8751	0.8476
16	IVP-Lab	Hossein Motamednia	0.8595	0.8429	0.8762
Baseline	StairIQA [88]		0.637	0.6179	0.6561
	DBCNN [120]		0.8228	0.7914	0.8542
	LIQE [121]		0.8543	0.8652	0.8433

Table 2. Quantitative results for the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge: Track 2 Video.

Rank	Team	Leader	Main Score	SRCC	PLCC
1	IMCL-DAMO	Yiting Lu	0.8385	0.8322	0.8448
2	Kwai-kaa	Qiulin Wang	0.824	0.8154	0.8326
3	SQL	Wei Gao	0.8232	0.8148	0.8316
4	musicbeer	Xiaoxin Lv	0.8231	0.8144	0.8318
5	finnbingo	Xingyuan Ma	0.8211	0.8131	0.829
6	PromptSync	Jiaze Li	0.8178	0.8102	0.8254
7	QA-FTE	Tianwu Zhi	0.8128	0.805	0.8207
8	MediaSecurity_SYSU&Alibaba	Baoying Chen	0.8124	0.8021	0.8226
9	IPPL-VQA	Pengfei Chen	0.8003	0.7939	0.8066
10	IVP-Lab	Hossein Motamednia	0.7944	0.7852	0.8035
11	Oblivion	Weifeng Dong	0.7869	0.7773	0.7965
12	UBC DSL Team	Jiahe Liu	0.7531	0.7431	0.7632
Baseline	SimpleVQA [87]		0.6602	0.6489	0.6714
	FAST-VQA [100]		0.7197	0.7156	0.7238
	DOVER [104]		0.7698	0.7616	0.7779

3.4. Challenge phases

Both tracks consist of two phases: the developing phase and the testing phase. In the developing phase, the participants can access the generated images/videos of the training set and the corresponding prompts and MOSs. Participants can be familiar with dataset structure and develop their methods. We also release the generated images and videos of the validation set with the corresponding prompts but without corresponding MOSs. Participants can utilize their methods to predict the quality scores of the validation set and upload the results to the server. The participants can receive immediate feedback and analyze the effectiveness of their methods on the validation set. The validation leaderboard is available. In the testing phase, the participants can access the images and videos of the testing set with the corresponding prompts but without corresponding MOSs. Participants need to upload the final predicted scores of the testing set before the challenge deadline. Each participating team needs to submit a source code/executable and a fact sheet, which is a detailed description file of the proposed method and the corresponding team information. The final results

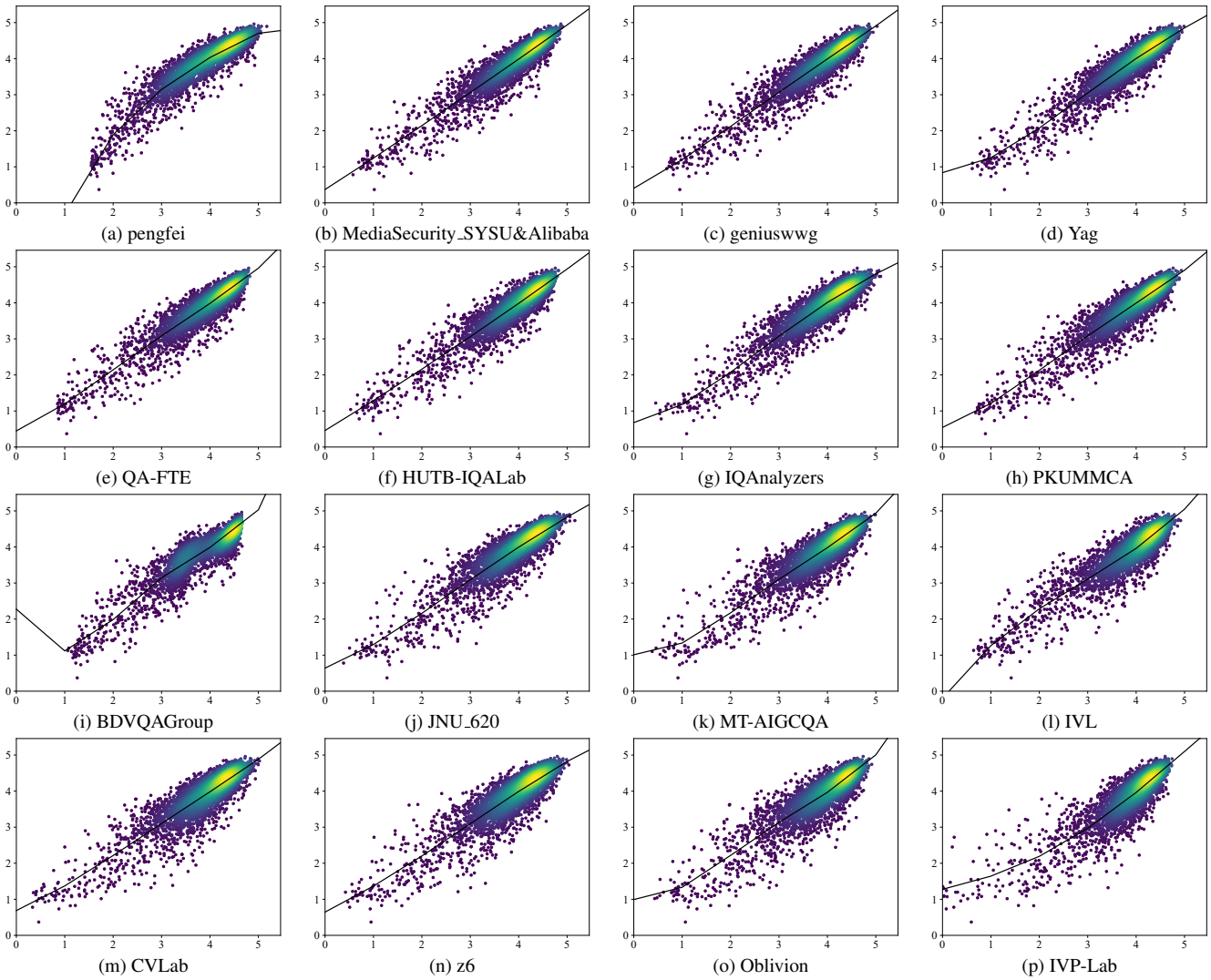

Figure 1. Scatter plots of the predicted scores vs. MOSs in the image track. The curves are obtained by a four-order polynomial nonlinear fitting.

are then sent to the participants.

4. Challenge Results

16 teams in the image track and 12 teams in the video track have submitted their final codes/executables and fact sheets. Table 1 and Table 2 summarize the main results and important information of the 28 valid teams. The methods of these teams are briefly introduced in Section 5 and the team members are listed in Appendix B.

4.1. Baselines

We compare the performance of submitted methods with several state-of-the-art I/VQA methods on the testing set, including StairIQA [47], DBCNN [120], and LIQE [121] for the image track and SimpleVQA [87], FAST-VQA [100], and DOVER [104] for the video track.

4.2. Result analysis

The main results of 28 teams' methods and the baseline methods are shown in Table 1 and Table 2. It can be seen that in both tracks, the results of the baseline methods are not ideal in the testing set of the two datasets, while most of the submitted methods have achieved better results. It means that these methods are closer to human visual perception when used to evaluate the generated images and videos. In the image track, 9 teams achieve a main score higher than 0.9, and 4 teams are higher than 0.91. In the video track, 9 teams achieve a main score higher than 0.8, 5 teams higher than 0.82, and the championship team is higher than 0.83. In the meantime, the top-ranked teams only have a

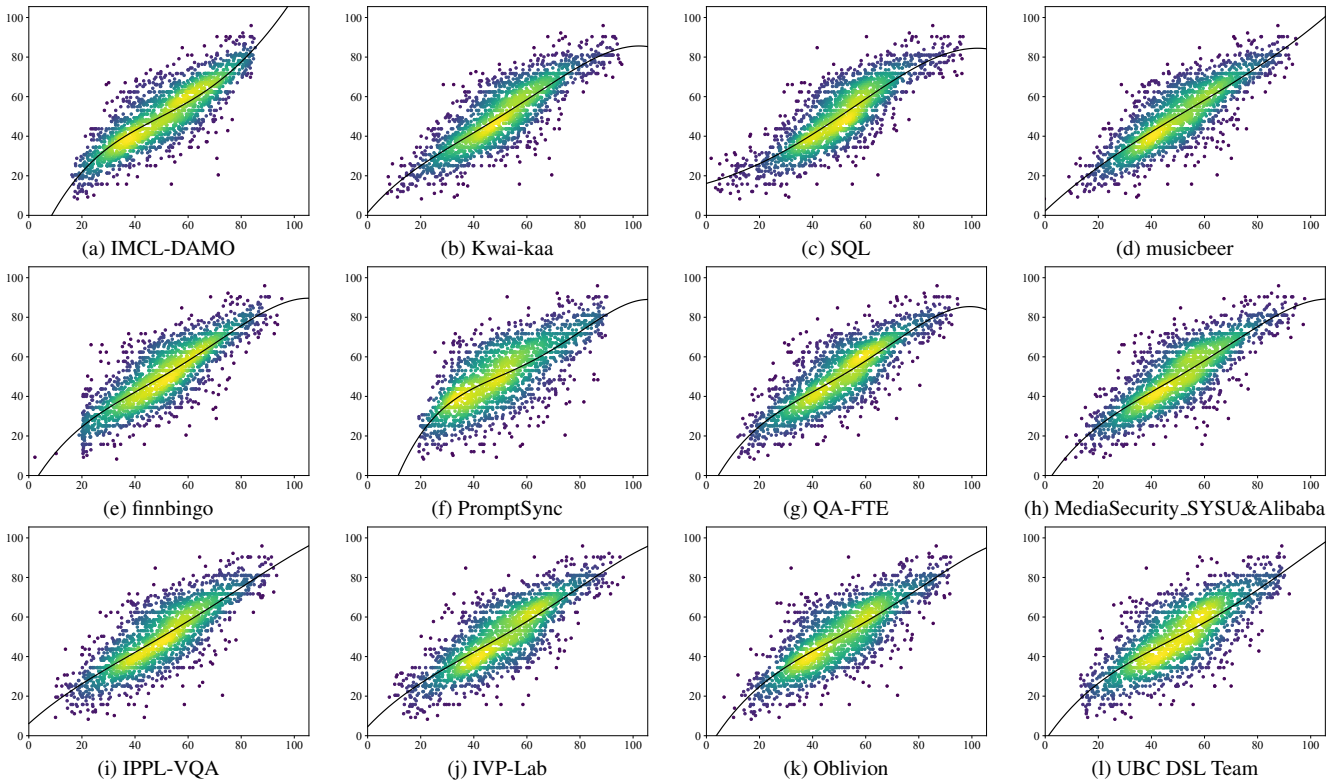

Figure 2. Scatter plots of the predicted scores vs. MOSs in the video track. The curves are obtained by a four-order polynomial nonlinear fitting.

small difference in the main score. Figure 1 and Figure 2 show scatter plots of predicted scores versus MOSs for the 28 teams' methods on the testing set. The curves are obtained by a four-order polynomial nonlinear fitting. We can observe that the predicted scores obtained by the top team methods have higher correlations with the MOSs. They can not only meet the need to predict quality scores for generated images/videos but also contribute to improving the performance of image/video generation methods.

5. Challenge Methods

5.1. Image Track

5.1.1 pengfei

Team pengfei [70] wins the championship in the image track. Their method enhances LIQE [121] by considering the correlation between prompts and generated images in the AIGC task, as shown in the Figure 3. To represent this correlation, they design corresponding textual templates such as "a how image matching the prompt", where "how" corresponds to five different adverbs: "badly", "poorly", "fairly", "well", and "perfectly". The textual templates are fed into the text encoder of the CLIP model [74] to obtain text features. Subsequently, the images are input into the image encoder of CLIP to obtain image features. CLIP's capability to compute the correlation between two features allows them to derive five levels of correlation between images and prompts. These levels of correlation are then weighted and summed using a softmax function to derive the final score. Additionally, considering the variation in image sizes, they not only feed image chunks into the image encoder but also include a resized version of the original image to learn global semantics.

Furthermore, in AIGI quality assessment, differences in image ratings compared to traditional image assessment ratings are observed. Traditional image assessment involves intuitive sorting based on factors like blurriness, distortion, and contamination. However, AIGI quality assessment is heavily influenced by prompts, making it difficult to effectively rank images between different prompts. Therefore, the ordinal relationship between two different image-text pairs is complex and challenging to directly learn. Traditional ranking loss is ineffective in this scenario. Instead, using L1 loss to directly fit score values and indirectly learn the order yields better SRCC scores.

Additionally, it is observed that the distribution of image ratings varies among images generated by different models. Failure to consider the model generating the images

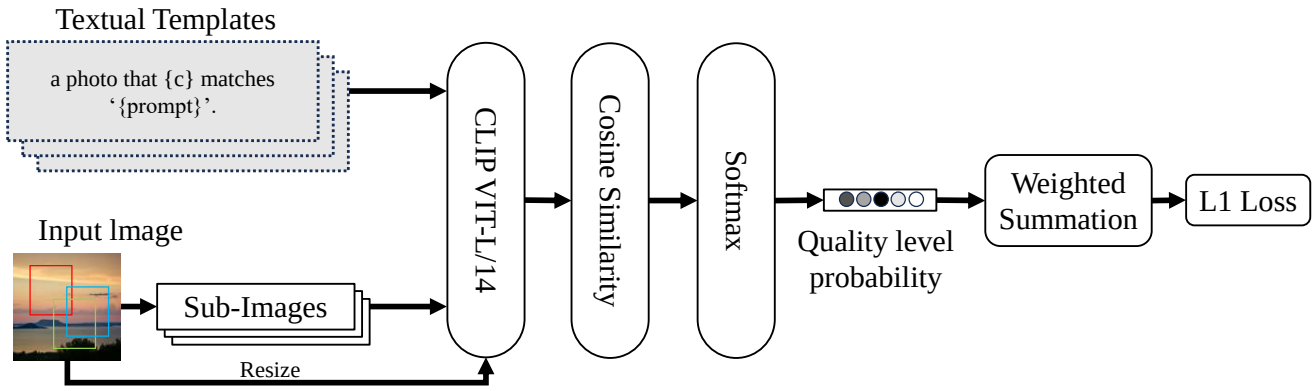

Figure 3. Overview diagram of the proposed method of team pengfei.

may lead to insufficient learning of the data distribution. However, considering the model generating the images may cause overfitting due to the uneven distribution of model categories. Therefore, two models are trained to integrate the results. One model undergoes normal training, while the other model has the name of the generating model added to the prompt to learn different distributions. The final inference results of the two models are multiplied and assembled.

5.1.2 MediaSecurity_SYSU&Alibaba

Team MediaSecurity_SYSU&Alibaba wins second place in the image track. They use both single-modal and multiple multi-modal networks for learning. The single-modal model used the backbone network of EVA02 large [14], while the multi-modal model used EVA02 large, ConvNeXt [63], and ConvNeXt v2 [99] as the image branches of the backbone network. They were combined with the text branches of the backbone network, such as Bert$_{base}$ [8], RoBERTa$_{base}$ [59], and DeBERTaV3$_{base}$ [29], to learn and evaluate scores. The training is done either in the 1st or 2nd fold of the 10-fold training, and finetuned across the entire training set. They use a total of 20 models, and the detailed combinations are shown below.

- weight=32/3200, convnext_large_in22k [63] + bert-base-uncased [8] (max_length=64), VisualBert [52] (trained in fold 0, num_fold=10)

- weight=48/3200, convnext_large_in22k [63] + bert-base-uncased [8] (max_length=64), VisualBert [52] (fine-tuned in all training set)

- weight=32/3200, convnext_xlarge_in22kk [63] + bert-base-uncased (max_length=32) [8], VisualBert [52] (trained in fold 0, num_fold=10)

- weight=48/3200, convnext_xlarge_in22k [63] + bert-base-uncased [8] (max_length=32), VisualBert [52] (fine-tuned in all training set)

- weight=32/3200, convnext_large_in22k [63] + deberta-v3-base [29] (max_length=64), concatenate features (trained in fold 0, num_fold=10)

- weight=48/3200, convnext_large_in22k [63] + deberta-v3-base [29] (max_length=64), concatenate features (fine-tuned in all training set)

- weight=32/3200, convnext_large_in22k [63] + roberta-base [59] (max_length=64), concatenate features (trained in fold 0, num_fold=10)

- weight=48/3200, convnext_large_in22k [63] + roberta-base [59] (max_length=64), concatenate features (finetuned in all training set)

- weight=32/3200, convnext_xlarge_in22k [63] + deb erta-v3-base [29] (max_length=64), concatenate features (trained in fold 0, num_fold=10)

- weight=48/3200, convnext_xlarge_in22k [63] + deberta -v3-base [29] (max_length=64), concatenate features (fine-tuned in all training set)

- weight=50/3200, convnextv2_huge.fcmae_ft_in22k_in1 k_512 [99] + bert-base-uncased [8] (max_length=75), concatenate features (trained in fold 0, num_fold=10)

- weight=75/3200, convnextv2_huge.fcmae_ft_in22k_in1 k_512 [99] + bert-base-uncased [8] (max_length=75), concatenate features (finetuned in all training set)

- weight=560/3200, eva02_large_patch14_448.mi m_m38m_ft_in22k_in1k [14] + deberta-v3-large [29] (max_length=75) + FC + LeakyReLU, concatenate features (trained in fold 0, num_fold=10)

- weight=1040/3200, eva02_large_patch14_448.mi m_m38m_ft_in22k_in1k [14] + deberta-v3-large [29] (max_length=75) + FC + LeakyReLU, concatenate features (finetuned in all training set)

- weight=275/3200, eva02_large_patch14_448.mim_m38 m_ft_in22k [14] + bert-base-uncased [8] (max_length=75), concatenate features (trained in fold 2, num_fold=10)

- weight=325/3200, eva02_large_patch14_448.mim_m38 m_ft_in22k [14] + bert-base-uncased [8] (max_length=75),

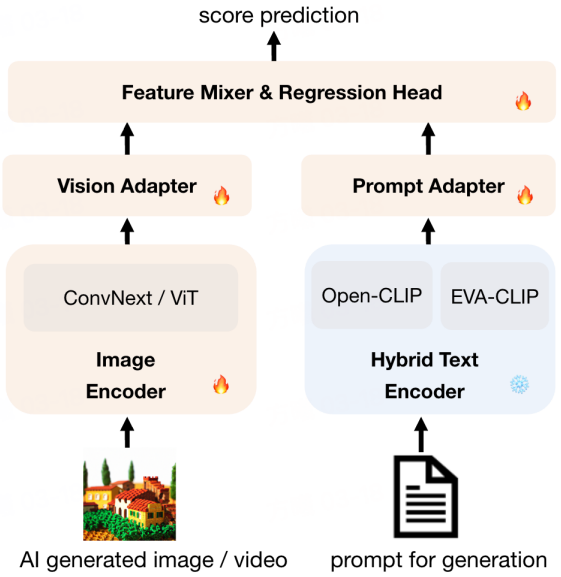

Figure 4. Overview of team geniuswwg proposed method.

concatenate features (finetuned in all training set)

- weight=50/3200, eva02_large_patch14_448.mim_m38m_ft_in22k [14] + bert-base-uncased [8] (max_length=75) + decoder∗6, concatenate features (trained in fold 2, num_fold=10)

- weight=87.5/3200, eva02_large_patch14_448.mim_m38m_ft_in22k [14] + bert-base-uncased (max_length=75) [8] + decoder∗6, concatenate features (finetuned in all training set)

- weight=50/3200, eva02_large_patch14_448.mim_m38m_ft_in22k [14] + roberta-base (max_length=75) [59], concatenate features (trained in fold 2, num_fold=10)

- weight=87.5/3200, eva02_large_patch14_448.mim_m38m_ft_in22k [14] + roberta-base [59] (max_length=75), concatenate features (finetuned in all training set)

5.1.3 geniuswwg

Team geniuswwg [13] wins third place in the image track. They propose an innovative approach to assess the quality of AIGC by treating it as a regression task under specified prompt conditions. It employs a dual-source CLIP [74] text encoder to interpret prompts, combining their features for a nuanced understanding. Image features are extracted using models like ConvNeXt [63], pre-trained on ImageNet [7], and adapted for interaction with the text features and vision features. A feature mixer module then blends the text and video features, using dot product and concatenation to model their correlation and conditional relationship. The final quality score is predicted by a two-layer Multi-Layer Perception (MLP). To enhance generalization, the system applies light data augmentations like flips and brightness adjustments. The ensemble method further refines the assessment by blending predictions from three diverse models, normalized for consistency, and averaged to produce a robust quality evaluation. Figure 4 shows the overview of their method.

Concretely, they use the frozen CLIP text encoder to encode the prompt, whose pre-trained weights are brought from two different open-source CLIP implements (*i.e.* Open-CLIP [74] and EVA-CLIP [86]) and pre-trained on different datasets (*e.g.* DFN-5B [12], LAION-2B [83], DataComp-1B [17] and WebLI). They build a hybrid prompt encoder by simply concatenating the output features from these two different CLIP text encoders. After obtaining the text features for prompts, they use a trainable dense layer as a prompt adapter, to align features to make better interaction with image features.

They simply use ConvNeXt-Small [63] (or ViT [10] and any other backbones with ImageNet [7] pre-trained weights) as the vision backbones. Similarly, they use a trainable dense layer as a vision adapter.

After the adapted prompt feature and vision features have been obtained, they use a module named feature mixer to make these two features interact with each other. They propose to use the features of prompts as a condition. They introduce two types of lightweight feature mixers: dot product and concatenation. The dot product can more effectively model the correlation between the generated images and the prompt. The use of concatenation is more akin to treating the prompt as a conditional factor. They employ these two types of feature mixers with different experts and ultimately utilize them for model blending. After obtaining the fused features, they employ a two-layer MLP as the prediction head to regress the final quality score.

Besides, they use random horizontal flip, slight random resized crop, and slight brightness contrast transformation for augmentation. These augmentations are relatively minor and generally do not affect the subjective quality assessment of the image, but can improve the model's generalization ability.

They blend 3 different models with different vision backbones: ConvNeXt [63], ViT-Transformer [10], and EVA02-Transformer [14]. Firstly, we normalize the predicted scores of each model on the testing set, so that different models have the same mean and variance of prediction. Finally, they blend all the models by averaging.

5.1.4 Yag

Team Yag [116] feeds images sampled through SAMA [60] into Swin Transformer V2 [61] for the image quality prediction component. Swin Transformer V2 is adept at extracting local features of video frames across various levels. By

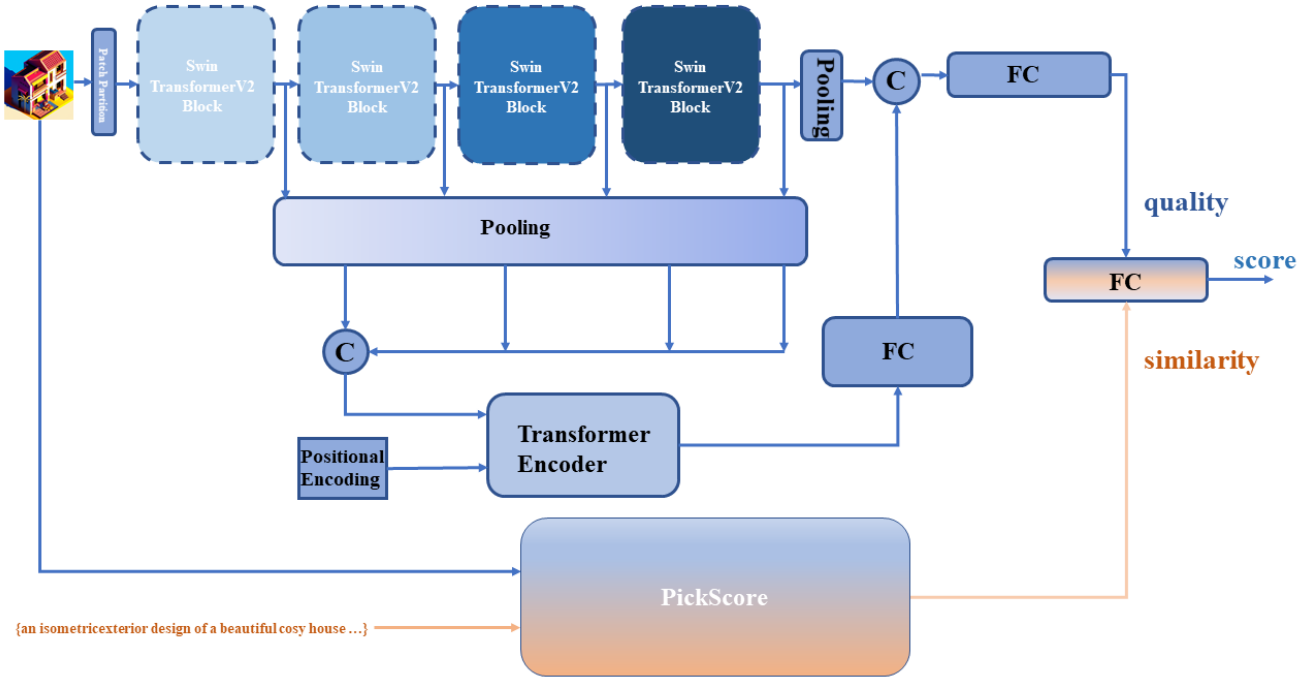

Figure 5. Overview of team Yag proposed method.

normalizing and rescaling the local features via a feature pooling operation, they obtain global frame-level features through the amalgamation of local representations across a multi-scale fusion module, which comprises a series of transformer layers. Subsequently, they concatenate the local and global features along the channel dimension and input them into a linear layer to ascertain the quality score, which is inspired by TReS [22]. This methodology harnesses spatial and temporal information from both global and local perspectives, thereby enhancing the perceptual capability of video quality assessment.

For the image-text similarity prediction component, they employ PickScore [41] to predict the similarity between images and text. They input the results from both the quality prediction and similarity prediction into a fully connected layer to derive the final quality score. The overview of the proposed method is shown in Figure 5.

Besides the provided training data, they also use CLIVE [20], LIVE [84], KonIQ-10K [32], KADID-10K [54], AGIQA-1k [123], AGIQA-3K [47], AIGCIQA2023 [93] and PKU-I2IQA [117] as additional data. The training images are paired-cropped into 256×256 patches for the image quality prediction component and 224×224 patches for the image-text similarity prediction component. They train the model using the Adam optimizer, setting the initial learning rate to $2e^{-5}$ for the Swin Transformer V2 [61], to $2e^{-6}$ for AIGC images, and to $1e^{-6}$ for the entire model.

5.1.5 QA-FTE

Team QA-FTE proposes a vision-language fused video quality evaluator, which is designed for AIGC. Considering the quality of AI-generated images is affected by the consistency of vision and language, they use CLIP [74] as the backbone model. Specifically, they first use the CLIP encoder to extract the vision feature of the image and the language feature of the prompt. Then, a bilinear pooling is used to obtain an interactive feature, which represents the consistency between vision information and language information. Finally, the above features are fused to predict AI-generated image quality scores.

5.1.6 HUTB-IQALab

Team HUTB-IQALab [113] proposes a novel mixture-of-experts boosted visual perception and semantic-aware quality assessment for AI-generated images. Firstly, they design the visual perception network to establish perceptual rules to obtain visual perception features. Secondly, they enhance the diversity of degradation-specific knowledge through the semantic-aware network, generating semantic-aware features. Thirdly, instead of fusing on predicted image scores,

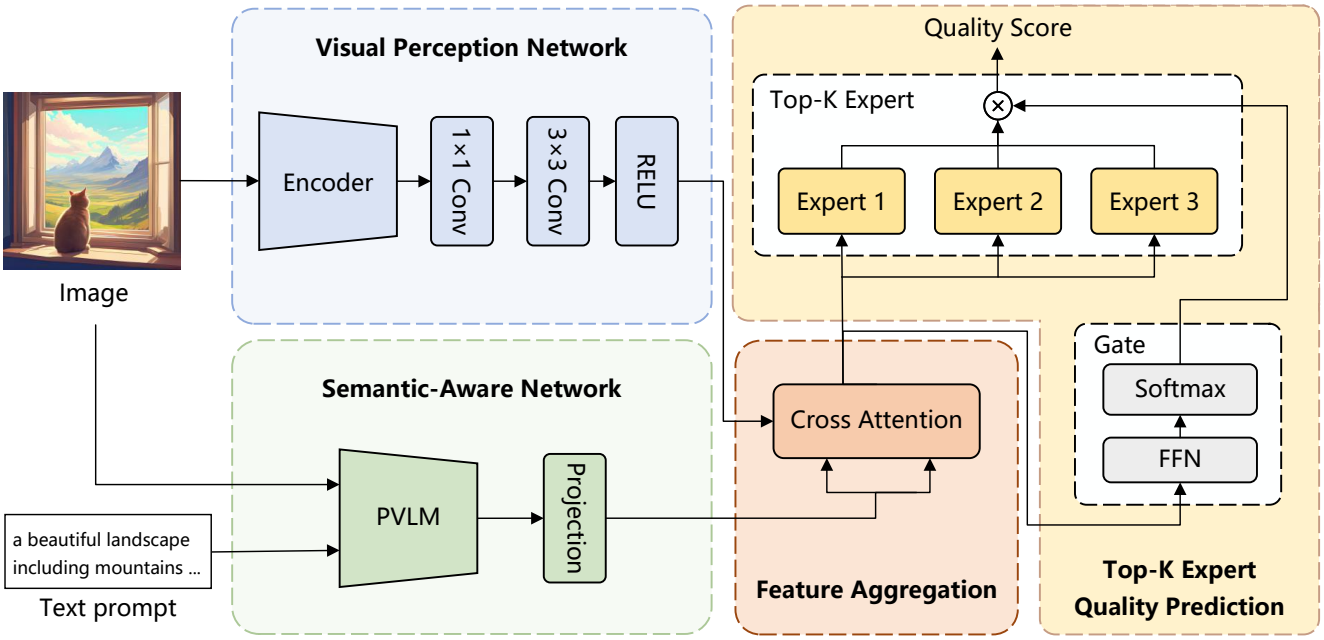

Figure 6. Overview of team HUTB-IQALab proposed method.

they propose to conduct cross-attention on visual perception and semantic-aware features, so that they can obtain comprehensive features and the inherent correlation between these features. Finally, they propose a mixture-of-experts model, involving multiple experts working collaboratively. Each expert is responsible for a specific set of features and outputs a corresponding prediction score. The mixture of multiple experts will ultimately yield a holistic, perceptually-aware score. Figure 6 shows the overview of the proposed method.

5.1.7 IQ Analyzers

IQ Analyzers propose a methodology that adopts a Mixture-of-Experts approach, integrating a broad spectrum of feature types. This includes low-level quality features obtained from Re-IQA [80], text-to-image alignment features via BLIP2 [48] and ImageBind [21], image aesthetics representations from VILA [39], the naturalness attributes of images as judged by DINOv2 [69], and traits from the text-to-image human preference model, ImageReward [112]. The features are combined and utilized to train an ElasticNet model, which maps the aggregated representation to the MOSs.

5.1.8 PKUMMCAL

Team PKUMMCAL believes that compared to the task of Natural Scene Image Quality Assessment (NSIQA), which focuses only on the perceptual quality of images, the quality evaluation of AI-generated images needs to consider the text-image consistency additionally. Therefore, they bring textual information into the model using a text encoder pre-trained in CLIP [74]. In terms of training methods, inspired by the works on natural image quality evaluation based on multi-task learning, they introduce additional tasks, hoping that the model can learn auxiliary knowledge from them. Unlike natural images, the distortion types of AI-generated images are difficult to distinguish directly, so they choose to predict the generative model used for creating AI-generated images as their auxiliary task. At the same time, to utilize the consistency of text and image information, they integrate the fine-tuning method from CLIPIQA [92] into their model. They specifically designed three models, which are ResNet50-based [28], DINOv2-based [69], and ConvNeXt-Based [63]. Besides, they integrate the HyperNet part of HyperIQA [85] into the ResNet50-based model to achieve semantic adaptive evaluation for different images.

The three models all share a similar framework, which is a dual-stream architecture to simultaneously process the image and its associated textual prompt. To fusion the visual and textual information, they propose an attention-based module. The fused feature and the visual feature are both fed to the quality prediction head, while the visual feature is also fed to the generation model classification head. Meanwhile, there is something different in ResNet50-based model architecture. They integrate several effective modules proposed for the NSIQA task, as illustrated.

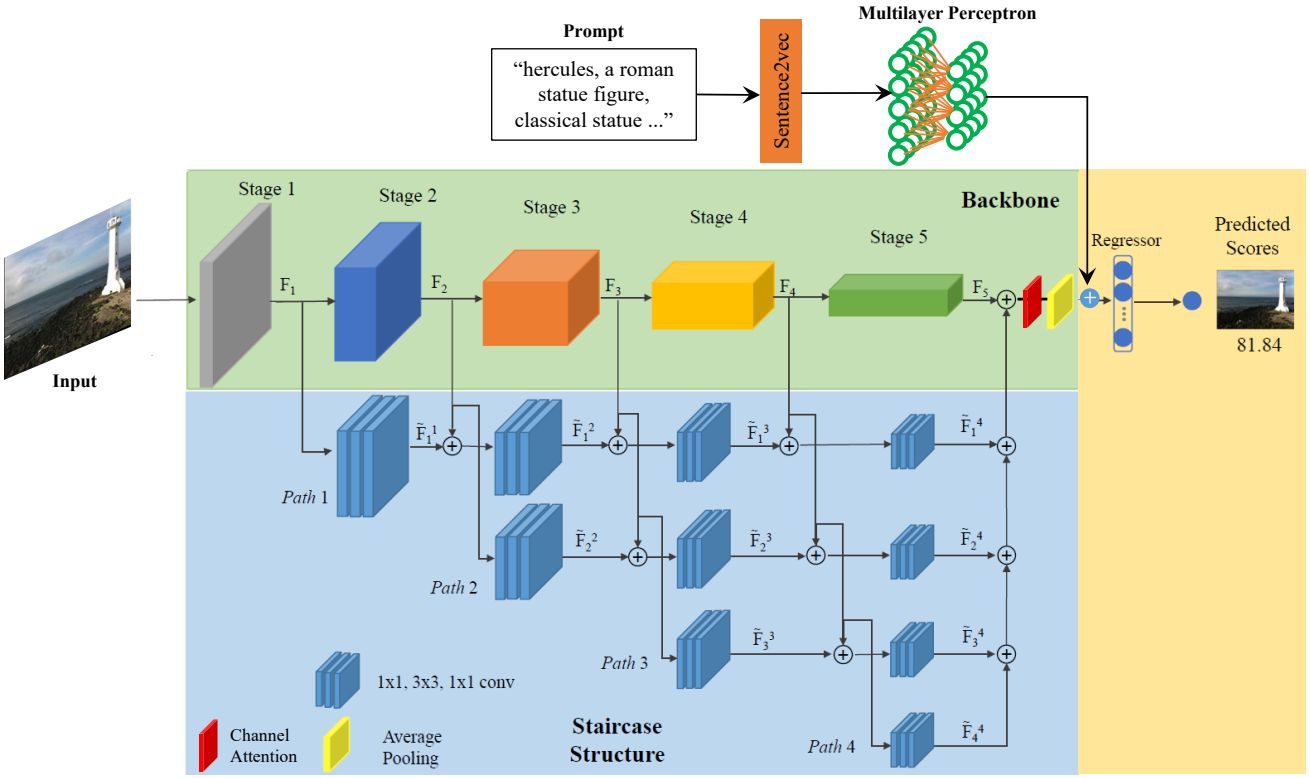

Figure 7. Overview of team JNU_620 proposed method.

5.1.9 BDVQAGroup

Team BDVQAGroup chooses two methods to assess the quality of AIGIs. One is Q-Align [107], which is based on large multi-modality models (LMMs). Q-Align converts MOSs into rating levels and uses classification to teach LMMs with text-defined rating levels instead of scores. During inference, it extracts the close-set probabilities of rating levels and performs a weighted average to obtain the LMM-predicted score. Another is based on MSTRIQ [94], a Swin-Transformer based method. They use several data augmentation methods to increase the training dataset and enhance the robustness of MSTRIQ, which are: 1) Expand the image along its longer side to form a square. 2) Randomly rotate the image at a degree of 90. 3) Resize the image to 448×448. 4) Randomly resize and crop the image at a ratio of 0.7.

They use a Q-Align model pre-trained on KonIQ [32], SPAQ [15], KADID [54], AVA [23], and LSVQ [115], and finetune this model through three strategies. The first model is based on the Q-Align Image Quality Scorer, which is finetuned for 4 epochs on the AGIQA-1K [123] images and then finetuned for another 2 epochs on the provided training images. The second model is based on the Q-Align Image Aesthetic Scorer, which is also finetuned for 2 epochs on the provided training images. The third model is based on the Q-Align Image Quality Scorer, which is finetuned for 2 epochs on the provided training images. The fourth model is an MSTRIQ model pre-trained on TID2013 [72], KonIQ-10k [32] and PIPAL [37], and finetune this model for 150 epochs on the provided training images.

The following methods are implemented to increase model performance: 1) Expand the image along its longer side to form a square. 2) Randomly crop the image 18 times. The crop size is 384×384 for all the 18 patches. 3) They use four models to predict and weight the four prediction results according to the following weights to obtain the final output. That is:

$$output = (0.3 \times model_1 + (0.4 \times model_2 + 0.6 \times model_3) \times 0.7) \times 0.8 + 0.2 \times model_4. \quad (2)$$

5.1.10 JNU_620

Team JNU_620 designs a method based on StairIQA [88], which includes two parts, a staircase network, and an image quality regressor. To make the model pay more attention to the important features, they added channel attention at the end of the staircase network. Moreover, to make full use of the prompts, they leverage the Sentence2Vec technology to convert prompts into sentence vectors. Then, the sen-

tence vectors are transferred into the multi-layer perceptron, which strengthens the representation of the features. After extracting prompt-aware features by the multi-layer perceptron and quality-aware features by the staircase network, they add these features and map them to the quality scores with a regression model. Figure 7 shows the overview of the proposed method.

They use seven models as the backbone of the proposed model, including ShuffleNet [122], MobileNetV2 [81], MobileNetV3 [33], ResNet50 [28], Res2Net50 [18], ResNeXt50 [111], and ResNeSt [119]. The weights of the backbone are initialized by training on ImageNet [7], and other weights are randomly initialized. The proposed model is only trained on the provided image training set. In the training stage, images are resized to 680×680 and randomly cropped with resolutions of 640×640. The Adam method is employed for optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set to 3×10^{-3}. MSE loss is used as the loss function for training. In the testing phase, the seven models are used for the ensemble. By performing the model ensemble, the results produced by multiple models are averaged for better results.

5.1.11 MT-AIGCQA

Similar to VBench [36], team MT-AIGCQA integrates a variety of basic models by fine-tuning or directly testing on the provided dataset to obtain image quality scores in different dimensions. The basic models include improved versions of BLIP [49], CLIP [74], etc. The inspection dimensions include image-text consistency, image quality, etc. Finally, they fuse the scores of the basic models to obtain the final MOS.

Specifically, for the image quality dimension, they use three models to obtain deep differential information, based on ResNet [28], NAS [129], and Swin Transformer [62] respectively. Since there is no separate image quality score in the training set, the MOS is expressed as the target image quality score. For the image-text consistency dimension, they obtaine the image-text consistency score by finetuning BLIP [49] on the training set. Since there is no separate image-text consistency score in the training set, they regard image-text pairs with MOSs exceeding 2.5 (ranging from 0 to 5) as matches, and MOSs lower than 2.5 as mismatches. For the overall image quality dimension, they added an MLP layer to BLIP for the regression of MOS.

In the testing phase, they use the five models trained in the previous stage to obtain the corresponding basic scores respectively, and then use a series of pre-trained models to obtain the corresponding scores of image-text pairs, including CLIP [74], Q-Instruct [106] and ImageReward [112]. Following [35], they obtain the corresponding BLIP score and CLIP score. Finally, for images generated by different models, they use a linear regression model to fit the final output MOS through the scores of the basic models.

5.1.12 IVL

Team IVL exploits BLIP-2 [48] to encode prompt and image, respectively. BLIP-2 consists of a vision encoder, a language model, and a Querying Transformer (Q-Former). The input image is first resized to the resolution of 224×224 pixels and then fed to the model that outputs a feature map of 32×768 features. Spatial features are finally averaged to obtain the 768-dimensional feature vector. The text prompt is first tokenized and then fed to the model which outputs a feature map with shape 12×768. Following [48], they select the first token as representative of the whole text input. The two feature vectors are l2-normalized and concatenated into a 1536-dimensional feature vector. A Support Vector Regression (SVR) machine with a Radial Basis Function (RBF) kernel is used to map the features into the final quality score.

5.1.13 CVLab

Team CVLab proposed model is based on a pre-trained text encoder (CLIP [74]) and an image encoder (ConvNeXt [63]). They use the image encoder to extract features from the images, and then pass those features with the dropout function (ratio 0.3) to a full-connection layer with 1000 input and 512 output. At the same time, they use the text encoder and the tokenized prompts, to extract the text features. Next, they concatenate the image features and text features and provide a vector with 1024 dimensions. They use this concatenated image-text feature to predict the final MOS with another fully connected layer.

5.1.14 z6

Team z6 introduces a network that combines image and text features for assessing the quality of AI-generated images. The network comprises three main components. The initial component is the image feature encoder, inspired by the architecture of the staircase network [88]. For the text feature encoder, they employ a basic transformer network. Finally, the feature fusion component utilizes a concatenation function along with a straightforward 1×1 convolution operation.

5.1.15 Oblivion

Team Oblivion uses a Swin Transformer [62] as the visual backbone, and then they use CLIP [74] to get the text and visual feature relationship. These features are used to enhance visual understanding to evaluate the quality of the image, and a Densenet [34] is used to help it get a high

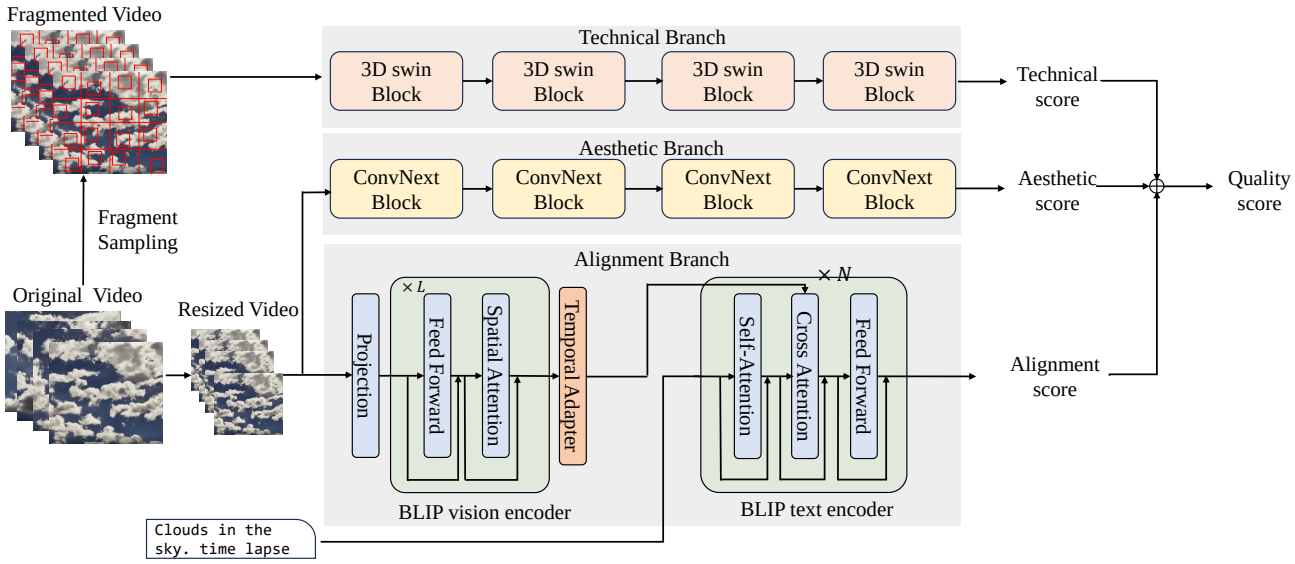

Figure 8. The overview of IMCL-DAMO team proposed versatile video quality evaluator.

visual understanding. They use the Swin-tiny network pre-trained on the Kinetics-400 [38] dataset to initialize the Swin Transformer backbone, and the ViT [10] pre-trained on the Kinetics-400 dataset to initialize the CLIP model.

5.1.16 IVP-Lab

Team IVP-Lab proposes a hybrid model employing both text and image information to estimate the quality of the generated image. Initially, the image and its corresponding text are processed using the models mentioned in [25, 42, 47, 75, 117] to map image and text information into feature vectors. Then, in order to align the text and image feature vectors, multiple mapping layers are considered. In the next stage, two quality values are computed: one measuring the similarity of the image to the text (conceptual_Q) and the other representing solely the image quality independent of the text (Image_Q). Finally, these two values are combined to calculate the final quality score. In fact, the hybrid model assesses the quality of generated images and provides final scores using both text and image information. There are numerous factors that are important in assessing the quality of AI-generated images. These include the image's natural appearance or naturalness, maintaining structural information, and effectively preserving the conceptual relevance to the image's content. In the proposed approach, naturalness and structural information are represented by Image_Q factor, and conceptual relevance is estimated using conceptual_Q.

5.2. Video Track

5.2.1 IMCL-DAMO

Team IMCL-DAMO [65] is the final winner of the video track. They propose a versatile video quality evaluator for AI-generated content, which can learn technical quality, aesthetic quality, and text-video alignment from different priors, as shown in Figure 8. Specifically, the input video is pre-processed to handle the disentangled information extraction from the three perspectives (i.e., technical quality, aesthetic, text-video alignment): they utilize the fragments extracted from original videos for technical quality assessment and resize the videos for aesthetic assessment and text-video alignment. Then, these separate inputs pass through multiple branches (technical branch, aesthetic branch, and alignment branch) to obtain the related score for different perspectives. To fuse the scores from different prior, we simply add them. Finally, they use PLCC loss and rank loss for score regression of each branch.

During training, they train the technical branch and aesthetic branch by loading the pre-trained weight from LSVQ [115]. Then the alignment branch is trained with 40% unfixed parameters, loading the pre-trained weight from ImageReward [112]. Note that these datasets are not involved in training with the provided video dataset. Finally, they finetune the technical branch, aesthetic branch, and alignment branch with 85% unfixed parameters for late fusion. During testing, they test their network using videos provided on the official website. A self-ensemble strategy is used during testing, and it brings performance gains of

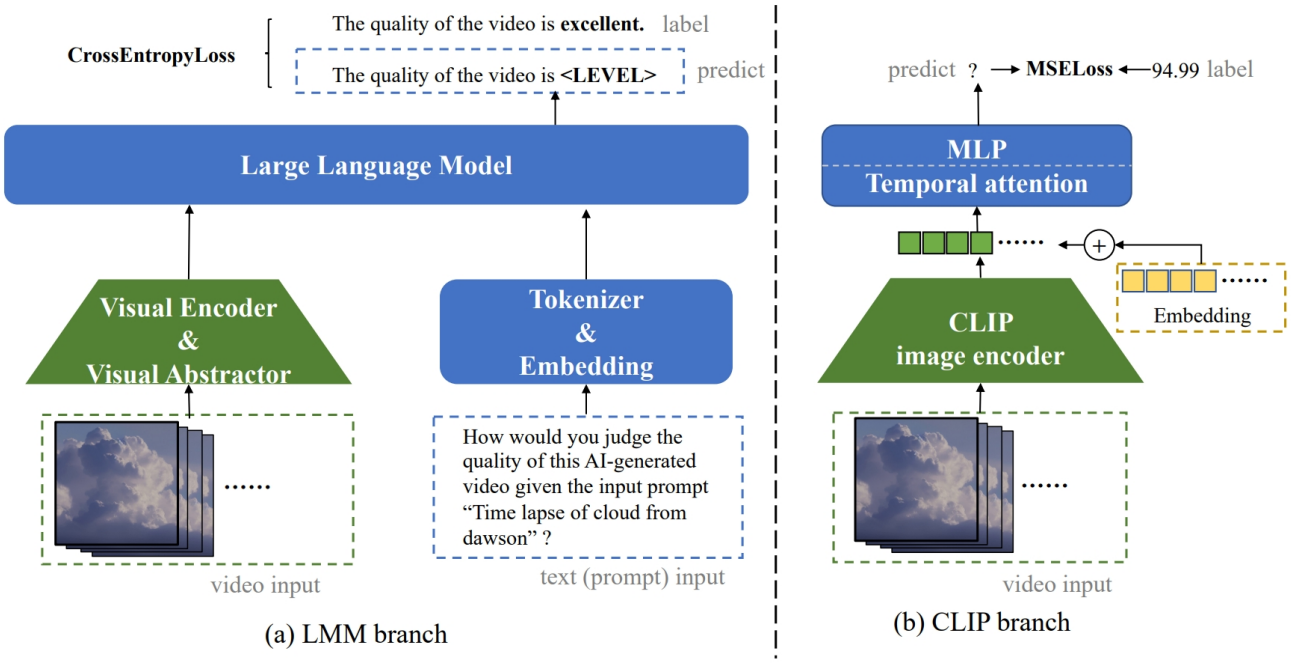

Figure 9. The overview of team Kwai-kaa proposed LMM and CLIP branches.

about 0.008 on PLCC.

In the training phase, the input frames for the aesthetic and text-video alignment branches are resized to 224×224. And "fragments" are sampled for the technical branch like in DOVER [104]. They use the Swin Transformer [62] as the technical branch backbone, the ConvNeXt [63] as the aesthetic branch, and BLIP [49] as the alignment branch. The training process takes 12 hours on 4 V100 GPUs. During testing, it takes 4 seconds for each video including the ensemble strategy.

5.2.2 Kwai-kaa

Team Kwai-kaa wins second place in the video track. They propose to tackle the challenge by leveraging LMMs. They follow a similar design to Q-Align [107], which is based on mPLUG-Owl2 [114], with the exception of the conversation formats. Specifically, they reformulate the conversation for AI-generated video assessment as follows:

#User: `<video>` *How would you judge the quality of this AI-generated video given the input prompt* `<prompt>` *?*

#Assistant: The quality of the video is `<level>`.

In the context of conversation, `<video>` denotes the input video, `<prompt>` denotes the prompt used to generate the input video, and `<level>` denotes the predicted score by LMMs.

However, the exclusive utilization of the Q-Align [107] achieves limited performance due to the quantification of MOS into 5 discrete text-defined levels. This quantization strategy restricts the model's ability to learn more precise quality scores. To complement the Q-Align architecture, they introduce an additional CLIP-based [74] architecture, leveraging it as a robust feature extractor for predicting precise quality scores with MSE constraints. The original CLIP was tailored for images and cannot capture the temporal consistency and interconnectedness between video frames, which significantly influences video quality. To address this, they incorporate attention layers between frames to capture temporal relationships. As the prompt serves as a global abstract of the video, they believe that assessing alignment is sufficient within the LMM branch and, therefore, does not utilize text information within the CLIP branch. By employing diverse architectures and training strategies across different branches, they aim to enhance the variety of information and contribute to improved results. The final score is obtained by averaging the results of each branch. The overview of LMM and CLIP branches is shown in Figure 9.

In particular, the LMM branches are finetuned with the pre-trained One-Align [107] weights, and the CLIP branches are finetuned with the pre-trained clip-vit-large-patch14 weights. They employed two finetuning strategies for the LMM branch: Vision Encoder & Vision Abstractor (VEVA) finetuning and full model finetuning. The pre-trained model of the LMM branch is fine-tuned with an initial learning rate of 2×10^{-5}, gradually decreasing to 0 using a cosine scheduler. Each strategy is trained for 2

epochs. The precise labels are divided into five text-defined levels: <excellent>, <good>, <fair>, <poor>, and <bad>. During training, they utilize 8 Tesla V100 GPUs, with a batch size of 24 for VEVA fine-tuning and 8 for full fine-tuning. The fine-tuning process for the LMM branch takes approximately 1 hour each to complete. The CLIP backbone is trained for 20 epochs, with a learning rate set to 1×10^{-6}. The training process is carried out on 8 Tesla A100 GPUs, with a batch size of 8. It takes approximately 8 hours to complete the training process. The provided generated videos consist of 4-second sequences at 4 frames per second and all frames are provided as input to the models. For videos with 15 frames, they pad the last frame to generate a complete set of 16 frames. All frames are resized to 448×448 and training processes are finished with the AdamW optimizer.

They get the MOS values of LMM branches via the weighted average of the LMM-predicted probabilities for each rating level, which can be denoted as:

$$s = \mathbf{w}^{\mathbf{T}}\mathbf{p} = \sum_{i=1}^{5} w_i \times p_i = \sum_{i=1}^{5} w_i \times \frac{e^{l_i}}{\sum_{j=1}^{5} e_j^l}, \quad (3)$$

where w_i is the logit weight for text level i and l_i is the corresponding logit output. they set the value of w to $[1, 0.75, 0.5, 0.25, 0]$ for text-label <excellent>, <good>, <fair>, <poor>, and <bad>. They rescale the predicted scores of the CLIP branches to $[0, 1]$ by dividing them with a constant factor of 100. To get the final score, they combined the scores from different branches using a weighted approach. Each video takes approximately 1 second to process in a single LMM branch in the inference, while the CLIP branch requires approximately 1.46 seconds per video.

5.2.3 SQL

Team SQL [73] wins third place in the video track. They propose to evaluate the video quality of AIGVs from five dimensions: aesthetic scores, technical scores, video-text consistency, fluency, and temporal consistency, as shown in Figure 10. They refer to aesthetic and technical aspects as visual harmony and refer to fluency and temporal consistency as temporal dynamics. Additionally, they apply model assembling and domain distribution estimation to optimize the model performance. They referred to DOVER [104] for the aesthetic and technical evaluation of the videos. To measure video-text consistency, they apply explicit prompt injection, implicit text guidance, DIFT [89] feature, and caption similarity. They inject the corresponding prompts of the videos into the video features using cross-attention. They also utilized BVQI's implicit text method [102] and jointly optimized the evaluation network

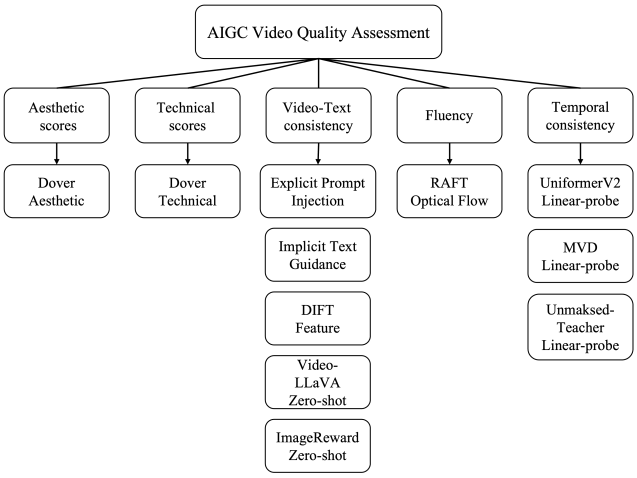

Figure 10. Five dimensions for SQL team proposed method.

using both implicit text and explicit prompts. They further use the diffusion feature to measure the per-frame text-content consistency. Building upon this, they utilize the cross-modal text-video multi-modal large language model, Video-LLaVA [53] to generate additional captions for each video segment. They then calculate the similarity between the generated captions and the given prompts to further optimize the network. As for fluency, they incorporate an optical flow estimation module to measure the flicker degree between keyframes in the videos. After estimating the optical flow between keyframes using the RAFT [90] algorithm, they assess video stability by computing the flow magnitude and warp loss. Regarding temporal consistency, they introduced pre-trained video understanding backbones such as Uniformer-V2 [50], UnmaskedTeacher [51], and MVD [96] to extract robust spatiotemporal features.

In regard to model assembly, they incorporate additional five models into their main model: FAST-VQA [100], Faster-VQA [101], ZoomVQA [127], XCLIP [68], and ImageReward [112]. It's important to note that the integration of these models serves as a finishing touch, and they haven't thoroughly adjusted the weights of the integration. The enhancement they bring is usually found in three decimal places.

Considering the domain distribution differences in the results generated by different models, in supervised learning, they predict not only the score but also which text-to-video model generated the video. This additional prediction aids the model in better understanding video features. Experiments have shown that this significantly enhances the performance of their model.

5.2.4 musicbeer

Team musicbeer and team geniuswwg in the image track are the same teams. They use the same architecture in the video track as shown in Figure 4. Detailed information can refer to Section 5.1.3. By changing the visual input from image to video frames, the model is able to predict the quality of AIGVs.

5.2.5 finnbingo

Team finnbingo and team pengfei in the image track are the same teams. They use the same architecture as introduced in Section 5.1.1. To adapt their methodology for video quality assessment, they treat a video as a collection of N frames, from which they selectively extract n frames at regular intervals to represent the video. They then calculate the average score of these n frames to determine the overall quality score of the video.

5.2.6 PromptSync

The method proposed by team PromptSync is structured into three main components. At the video level, building upon the FAST-VQA [100] framework, they introduce additional feature constraints. They use the CLIP [74] image encoder to obtain image features for all video frames and apply a cross-attention mechanism to detect semantic discontinuities between frames. Furthermore, they project the video features derived from FAST-VQA into the text feature space obtained from the CLIP text encoder, calculating the semantic consistency between the video and the prompt. The final AIGC evaluation score is aggregated from the FAST-VQA video features, video-prompt consistency features, and frame sequence consistency features.

At the segment level, they replace the backbone model with Swin Transformer [62], concatenate text features from the CLIP text encoder, and utilize image features extracted by Swin Transformer and video features from the slowfast [16] model. The entire network is pre-trained on the LSVQ [115] dataset and then finetuned on the competition's training set.

Last, they conduct a frame-level evaluation of video quality. By leveraging the CLIP text encoder, they extract text features and concatenate them with image features obtained from the convent model to perform scoring. By capturing video features from different perspectives, these three components collectively contribute to their comprehensive AI-generated video quality assessment score.

5.2.7 QA-FTE

In the video track, team QA-FTE uses the same method as in the image track (Section 5.1.5). They change the visual input from image to video frames and average the scores of all frames to get an overall quality score for AIGVs.

5.2.8 MediaSecurity_SYSU&Alibaba

Team MediaSecurity_SYSU&Alibaba's solution ensemble consists of four types of models: single-modal model with a single frame, single-modal model with multiple frames, multi-modal model with a single frame, and multi-modal model with multiple frames.

In the single-modal model with a single frame, they utilize the Swin-L [62] pre-trained on ImageNet 22K [7] to predict the quality of a single frame. In the single-modal model with multiple frames, they add NeXtVLAD [55] to the Swin-L model, which is initialized from the single-modal model with a single frame. In the multi-modal model with a single frame, they utilize multiple combinations of image encoder and text encoder, including ConvNeXt-xlarge [63] and Bert-base [8] (max length$=64$) fused by VisualBert [52], Swin large and Bert-base (max length$=64$) fused by VisualBert, ConvNeXt-large and Bert-base (max length$=64$) fused by concatenation, ConvNeXt-large and DeBERTaV3-base [29] (max length$=64$) fused by concatenation, ConvNeXt-xlarge and Bert-base (max length$=32$) fused by concatenation, and Swin large and Bert-base (max length$=64$) fused by concatenation. Finally, in the multi-modal model with multiple frames, they use two combinations, in terms of Swin large and Bert-base (max length$=64$) fused by VisualBert, and ConvNeXt-large and Bert-base (max length$=64$) fused by concatenation, both initialized by weights from the multi-modal model with a single frame. The final score is obtained by the ensemble of all the predictions.

5.2.9 IPPL-VQA

The architecture proposed by IPPL-VQA is composed of text branches and image branches. The input of the text branch is the text description of the image, and text features are extracted by the frozen text encoder of the pre-trained CLIP-B-32 model [74]. There are two ways of sampling the image part (MaxVQA method [103]): 1. Sampling distinct texture details through cropping and splicing fragments; 2. Scaled sampling containing global information. The images of these two sampled branches undergo a frozen CLIP image encoder and two different temporal fusion models respectively. The features of both branches are concatenated and reduced to the width of the textual features with an MLP layer. The inner product of the final text and image features are calculated to get a matching score.

5.2.10 IVP-Lab

The proposed method of team IVP-Lab represents a hybrid model that incorporates both textual and visual data to evaluate the quality of the generated video. The mentioned model is employed to process the video and its related text, mapping the video and textual data into feature vectors. Multiple mapping layers are employed to align the text and video-based feature vectors.

In evaluating the quality of AI-generated videos, several factors should be considered. These include the videos's natural appearance considering both spatial and temporal information and preservation of structural information especially in the spatial domain. Another important factor is the conceptual relevance of the video's content.

In the proposed method, two quality-based feature vectors are computed: one assesses the similarity of the video to the text, while the other evaluates the video quality independently. These two feature vectors are then subjected to an inner product, resulting in a final vector for quality assessment. The resultant quality-based feature vector is fed to the fully connected network to estimate the quality of the AI-generated Videos.

5.2.11 Oblivion

Team Oblivion uses the Video Swin Transformer [64] as the visual backbone, and then they use the CLIP [74] text encoder as the text feature extractor, using text features to enhance visual understanding to evaluate the quality of the video. They use the Swin-tiny network pre-trained on the Kinetics-400 [38] dataset to initialize the Video Swin Transformer backbone, and the ResNet-50 [28] network pre-trained on the Kinetics-400 dataset to initialize the text encoder in the CLIP model.

5.2.12 UBC DSL Team

UBC DSL Team aims to build a video quality assessment model leveraging multi-faceted video representations, taking the visual quality, text prompt, and motion coherence into account. Specifically, they use the off-the-shelf pre-trained video encoder VideoMAE [91] and text encoder CLIP [74] to extract vision and language features. Additionally, they use the Inflated 3D Convnet [2] (I3D) as another video feature extractor, following prior work on generated video quality assessment.

To address temporal inconsistencies in AI-generated data, such as unnatural movements or blurring, they emphasize the importance of motion coherence in video quality evaluation. Leveraging the pre-trained motion tracking network PIPs++ [128], they extract motion features by tracking key points' trajectories in videos. They calculate the velocity and acceleration of these points, underpinning the notion that realistic motions should exhibit consistent acceleration. This approach yields dense motion features, enriching their VQA model's ability to detect and interpret temporal anomalies effectively.

After extracting video and text representations using various encoders, they use a vanilla transformer consisting of an encoder layer only to mix these representations in the token space. They freeze all pre-trained encoders to prevent overfitting. They add an additional global token to improve network capacity and use the global token to read out the final prediction score.

Acknowledgments

We thank Huawei Technology Co., Ltd for sponsoring this NTIRE 2024 challenge and the NTIRE 2024 sponsors: ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab).

A. NTIRE 2024 Organizers

Title:
NTIRE 2024 Quality Assessment of AI-Generated Content Challenge
Members:
Xiaohong Liu[1] *(xiaohongliu@sjtu.edu.cn)*, Xiongkuo Min[1], Guangtao Zhai[1], Chunyi Li[1], Tengchuan Kou[1], Wei Sun[1], Haoning Wu[2], Yixuan Gao[1], Yuqin Cao[1], Zicheng Zhang[1], Xiele Wu[1], Radu Timofte[3,4]
Affiliations:
[1] Shanghai Jiao Tong University, China
[2] Nanyang Technological University, Singapore
[3] ETH Zürich, Switzerland
[4] University of Würzburg, Germany

B. Teams and Affiliations in Image Track

z6

Title:
AI-Generated Image Quality Assessment with Image and Text Feature Mixture Network
Members:
Ganzorig Gankhuyag[1] *(gnzrg25@gmail.com)*, Kihwan Yoon[1]
Affiliations:
[1] Korea Electronics Technology Institute

BDVQAGroup

Title:
AI-Generated Image Quality Assessment Method based on LMMs
Members:
Yifang Xu[1] *(xuyifang.233@bytedance.com)*, Haotian Fan[1],

Fangyuan Kong[1]
Affiliations:
[1] ByteDance

Oblivion

Title:
Text-prompts to enhancement image quality assessment's performance
Members:
Shiling Zhao[1]*(yiyiaiou@163.com)*, Weifeng Dong[1], Haibing Yin[1]
Affiliations:
[1] Hangzhou Dianzi University

MT-AIGCQA

Title:
AIGCQA-Bench: A Comprehensive Benchmark Suite for Text-to-image Generative Models
Members:
Li Zhu[1] *(zhuli09@meituan.com)*, Zhiling Wang[1], Bingchen Huang[1]
Affiliations:
[1] Sankuai

pengfei

Title:
AIGC image quality assessment via image-prompt correspondence
Members:
Fei Peng[1] *(pf0607@bupt.edu.cn)*, Huiyuan Fu[1], Anlong Ming[1], Chuanming Wang[1], Huadong Ma[1], Shuai He[1], Zifei Dou[2], Shu Chen[2]
Affiliations:
[1] Beijing University of Posts and Telecommunications, China
[2] Beijing Xiaomi Mobile Software Co., Ltd.

QA-FTE

Title:
Vision-Language Fused Image Quality Evaluator for AI-Generated Content
Members:
Tianwu Zhi[1] *(zhitianwu@bytedance.com)*, Yabin Zhang[1], Yaohui Li[1], Yang Li[1], Jingwen Xu[1], Jianzhao Liu[1], Yiting Liao[1], Junlin Li[1]
Affiliations:
[1] Bytedance Multimedia Lab

Yag

Title:
AIGCIQA Implemented via Swin Transformer V2 and PickScore
Members:
Zihao Yu[1] *(yuzihao@mail.ustc.edu.cn)*, Fengbin Guan[1], Yiting Lu[1], Xin Li[1]
Affiliations:
[1] University of Science and Technology of China

IVP-Lab

Title:
AI-generated assessment of image quality by combining textual and visual characteristics
Members:
Hossein Motamednia[1] *(h.motamednia@ipm.ir)*, S. Farhad Hosseini-Benvidi[2], Ahmad Mahmoudi-Aznaveh[3] and Azadeh Mansouri[2]
Affiliations:
[1] High Performance Computing Laboratory, School of Computer Science, Institute for Research in Fundamental Sciences, Tehran, Iran
[2] Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran
[3] Cyber Research Institute, Shahid Beheshti University, Tehran, Iran

IQ Analyzers

Title:
Quality Assessment of AI-Generated Content Using Bag of Features Approach
Members:
Avinab Saha[1] *(avinab.saha@utexas.edu)*, Sandeep Mishra[1], Shashank Gupta[1], Rajesh Sureddi[1], Oindrila Saha[2]
Affiliations:
[1] University of Texas at Austin
[2] University of Massachusetts Amher

IVL

Title:
Quality Assessment of AI-Generated Contents through Language-Image Pre-trained Models and Support Vector Regression
Members:
Luigi Celona[1] *(luigi.celona@unimib.it)*, Simone Bianco[1], Paolo Napoletano[1], Raimondo Schettini[1]
Affiliations:
[1] Department of Informatics Systems and Communication,

University of Milano - Bicocca

HUTB-IQALab

Title:
Mixture-of-Experts Boosted Visual Perception and Semantic-Aware Quality Assessment for AI-Generated Images
Members:
Junfeng Yang[1] *(b12100031@hnu.edu.cn)*, Jing Fu[1], Wei Zhang[1], Wenzhi Cao[1], Limei Liu[1], Han Peng[1]
Affiliations:
[1] Xiangjiang Laboratory and Hunan University of Technology and Business

JNU_620

Title:
Prompt-StairIQA
Members:
Weijun Yuan[1] *(yweijun@stu2022.jnu.edu.cn)*, Zhan Li[1], Yihang Cheng[1], Yifan Deng[1]
Affiliations:
[1] Jinan University

MediaSecurity_SYSU&Alibaba

Title:
Single modal and multiple multi-modal networks for learning quality assessment
Members:
Huacong Zhang[1] *(zhanghc8@mail2.sysu.edu.cn)*, Haiyi Xie[1], Chengwei Wang[1], Baoying Chen[2], Jishen Zeng[2], Jianquan Yang[1]
Affiliations:
[1] Sun Yat-sen University
[2] Alibaba Group

geniuswwg

Title:
PCQA: A Strong Baseline for AIGC Quality Assessment Based on Prompt Condition
Members:
Weigang Wang[1] *(geniuswwg@gmail.com)*, Xi Fang[2], Xiaoxin Lv[3], Jun Yan[4]
Affiliations:
[1] Cisco Systems, Inc.
[2] DP Technology, Ltd.
[3] Shopee Pte. Ltd.
[4] Tongji University

PKUMMCAL

Title:
Assessing AI-Generated Image Quality via Multitask Learning
Members:
Haohui Li[1] *(lihaohui@stu.pku.edu.cn)*, Bowen Qu[1], Yao Li[1], Shuqing Luo[1], Shunzhou Wang[1], Wei Gao[1]
Affiliations:
[1] School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University

CVLab

Title:
An Efficient CLIP-based Baseline for Evaluating the Alignment and Quality of AI-Generated Images
Members:
Zihao Lu[1] *(zihao.lu@stud-mail.uni-wuerzburg.de)*, Marcos V. Conde[1], Radu Timofte[1]
Affiliations:
[1] University of Würzburg

C. Teams and Affiliations in Video Track

IMCL-DAMO

Title:
VC-VQE: Versatile and Comprehensive Video Quality Evaluator for AI-Generated Content
Members:
Yiting Lu[1] *(luyt31415@mail.ustc.edu.cn)*, Xin Li[1], Xinrui Wang[1], Zihao Yu[1], Fengbin Guan[1], Zhibo Chen[1], Ruling Liao[1], Yan Ye[1]
Affiliations:
[1] University of Science and Technology of China

Kwai-kaa

Title:
Leveraging Large Multi-modality Models and CLIP Encoder for AI-Generated Video Assessment
Members:
Qiulin Wang[1] *(wangqiulin@kuaishou.com)*, Bing Li[2], Zhaokun Zhou[3], Miao Geng[1], Rui Chen[1], Xin Tao[1]
Affiliations:
[1] Kuaishou Technology
[2] University of Science and Technology of China
[3] Peking University

SQL

Title:
Exploring AIGC Video Quality: A Focus on Text-Video

Consistency, Visual Harmony, and Temporal Dynamics
Members:
Wei Gao[1] *(gaowei262@pku.edu.cn)*, Xiaoyu Liang[1], Bowen Qu[1], Shangkun Sun[1]
Affiliations:
[1] Peking University

musicbeer

Title:
PCQA: A Strong Baseline for AIGC Quality Assessment Based on Prompt Condition
Members:
Xiaoxin Lv[1] *(musicbeer2017@gmail.com)*, Xi Fang[2], Weigang Wang[3], Jun Yan[4]
Affiliations:
[1] Shopee Pte. Ltd.
[2] DP Technology, Ltd.
[3] Cisco Systems, Inc.
[4] Tongji University

finnbingo

Title:
Video Quality Assessment for AI-Generated Content via Frame-Prompt Correspondence
Members:
Xingyuan Ma[1] *(maxy@bupt.edu.cn)*, Shuai He[1], Anlong Ming[1], Huiyuan Fu[1], Huadong Ma[1], Zifei Dou[2], Shu Chen[2]
Affiliations:
[1] Beijing University of Posts and Telecommunications
[2] Beijing Xiaomi Mobile Software Co., Ltd

PromptSync

Title:
multi-modal AIGC Video Quality Assessment with CLIP and Swin Transformer
Members:
Jiaze Li[1] *(1916444377@qq.com)*, Mengduo Yang[1], Haoran Xu[1], Jie Zhou[1], Shiding Zhu[1], Bohan Yu[1]
Affiliations:
[1] Zhejiang University

QA-FTE

Title:
Vision-Language Fused Video Quality Evaluator for AI-Generated Content
Members:
Tianwu Zhi[1] *(hitianwu@bytedance.com)*, Yabin Zhang[1], Yaohui Li[1], Yang Li[1], Jingwen Xu[1], Jianzhao Liu[1], Yiting Liao[1], Junlin Li[1]
Affiliations:
[1] Bytedance Inc.

MediaSecurity_SYSU&Alibaba

Title:
Integrating 2D and 3D Modalities for Enhanced multi-modal Video Quality Assessment
Members:
Baoying Chen[1] *(chenbaoying.chenba@alibaba-inc.com)*, Jishen Zeng[1], Huacong Zhang[1], Haiyi Xie[2], Chengwei Wang[2], Jianquan Yang[2]
Affiliations:
[1] Alibaba Group
[2] Sun Yat-sen University

IPPL-VQA

Title:
XCS-Net
Members:
Pengfei Chen[1] *(chenpengfei@xidian.edu.cn)*, Xinrui Xu[1], Jiabin Shen[1], Zhichao Duan[1]
Affiliations:
[1] Xidian University

IVP-Lab

Title:
AI-generated assessment of video quality by combining textual and visual characteristics
Members:
Hossein Motamednia[1] *(h.motamednia@ipm.ir)*, S. Farhad Hosseini-Benvidi[2], Erfan Asadi[2], Ahmad Mahmoudi-Aznaveh[3], Azadeh Mansouri[2]
Affiliations:
[1] Institute for Research in Fundamental Sciences
[2] Kharazmi University
[3] Shahid Beheshti University

Oblivion

Title:
Text-prompts to enhancement video quality assessment's performance
Members:
Weifeng Dong[1]*(dongwf@hdu.edu.cn)*, Shiling Zhao[1], Haibing Yin[1]
Affiliations:
[1] Hangzhou Dianzi University

UBC DSL Team

Title:

Enhanced Video Quality Assessment Transformer Based on Motion Feature

Members:

Jiahe Liu[1]*(jiaheliu@ece.ubc.ca)*, Qi Yan[1], Youran Qu[2], Xiaohui Zeng[3], Lele Wang[1], Renjie Liao[1]

Affiliations:

[1] University of British Columbia

[2] Peking University

[3] University of Toronto

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 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. 17

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. 2310.00426, 2023. 3

[4] Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Exploring the naturalness of ai-generated images. *arXiv preprint arXiv:2312.05476*, 2023. 2

[5] 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. 2

[6] DeepFloyd. If-i-xl-v1.0. https://www.deepfloyd.ai, 2023. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 8, 12, 16

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7, 8, 16

[9] Yunlong Dong, Xiaohong Liu, Yixuan Gao, Xunchu Zhou, Tao Tan, and Guangtao Zhai. Light-vqa: A multi-dimensional quality assessment model for low-light video enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1088–1097, 2023. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 8, 13

[11] dreamlike art. dreamlike-photoreal-2.0. https://dreamlike.art, 2023. 3

[12] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 8

[13] Xi Fang, Weigang Wang, Xiaoxin Lv, and Jun Yan. Pcqa: A strong baseline for aigc quality assessment based on prompt condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 8

[14] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 7, 8

[15] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 11

[16] 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*, pages 6202–6211, 2019. 16

[17] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[18] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 12

[19] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. VDPVE: VQA dataset for perceptual video enhancement. *arXiv preprint arXiv:2303.09290*, 2023. 3

[20] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 9

[21] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 10

[22] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1220–1230, 2022. 9

[23] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056,

2018. 11

[24] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 3

[25] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 369–385. Springer, 2020. 13

[26] 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. 3

[27] Yatharth Gupta, Vishnu V. Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. 2401.02677, 2024. 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 10, 12, 17

[29] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. 7, 16

[30] 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. 3

[31] David Holz. Midjourney. https://www.midjourney.com, 2023. 3

[32] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 9, 11

[33] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 12

[34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 12

[35] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 12

[36] 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*, 2024. 2, 3, 12

[37] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 633–651. Springer, 2020. 11

[38] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 13, 17

[39] 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. 10

[40] 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. 3

[41] 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. 2, 9

[42] Hyunsuk Ko, Dae Yeol Lee, Seunghyun Cho, and Alan C Bovik. Quality prediction on deep generative images. *IEEE Transactions on Image Processing*, 29:5964–5979, 2020. 13

[43] Tengchuan Kou, Xiaohong Liu, Wei Sun, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Stablevqa: A deep no-reference quality assessment model for video stability. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1066–1076, 2023. 3

[44] 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

[45] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Haoning Wu, Weixia Zhang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Aigiqa-20k: A large database for ai-generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 2, 3

[46] 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. 2

[47] 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, 5, 9, 13

[48] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 10, 12

[49] 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, 12, 14

[50] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 15

[51] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023. 15

[52] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 7, 16

[53] 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. 15

[54] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 9, 11

[55] Rongcheng Lin, Jing Xiao, and Jianping Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 16

[56] Xiaohong Liu, Radu Timofte, Yunlong Dong, Zhiliang Ma, Haotian Fan, Chunzheng Zhu, Xiongkuo Min, Guangtao Zhai, Ziheng Jia, Mirko Agarla, et al. Ntire 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1551–1569, 2023. 3

[57] 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. 2, 3

[58] 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. 2, 3

[59] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 7, 8

[60] Yongxu Liu, Yinghui Quan, Guoyao Xiao, Aobo Li, and Jinjian Wu. Scaling and masking: A new paradigm of data sampling for image and video quality assessment. *arXiv preprint arXiv:2401.02614*, 2024. 8

[61] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 8, 9

[62] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 12, 14, 16

[63] 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. 7, 8, 10, 12, 14, 16

[64] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 17

[65] Yiting Lu, Xin Li, Bingchen Li, Zihao Yu, Fengbin Guan, Xinrui Wang, Ruling Liao, Yan Ye, and Zhibo Chen. Aigc-vqa: A holistic perception metric for aigc video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 13

[66] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module, 2023. 3

[67] 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*, pages 10209–10218, 2023. 3

[68] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 15

[69] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 10

[70] Fei Peng, Huiyuan Fu, Anlong Ming, Chuanming Wang, Huadong Ma, Shuai He, Zifei Dou, and Shu Chen. Aigc image quality assessment via image-prompt correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 6

[71] PlaygroundAI. playground-v2-1024px-aesthetic. https://playground.com, 2023. 3

[72] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspec-

tives. *Signal processing: Image communication*, 30:57–77, 2015. 11

[73] Bowen Qu, Xiaoyu Liang, Shangkun Sun, and Wei Gao. Exploring aigc video quality: A focus on visual harmony, video-text consistency and domain distribution gap. In *Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 15

[74] 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, 6, 8, 9, 10, 12, 14, 16, 17

[75] 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. 13

[76] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2204.06125, 2022. 3

[77] 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. 2

[78] 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. 3

[79] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. 2207.13038, 2022. 3

[80] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 10

[81] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 12

[82] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023. 3

[83] 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. 8

[84] H Sheikh. Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 9

[85] 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)*, June 2020. 10

[86] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 8

[87] 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*, pages 856–865, 2022. 3, 4, 5

[88] 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. 2, 4, 11, 12

[89] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 15

[90] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 15

[91] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 17

[92] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023. 10

[93] 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. 2, 9

[94] Jing Wang, Haotian Fan, Xiaoxia Hou, Yitian Xu, Tao Li, Xuechao Lu, and Lean Fu. Mstriq: No reference image quality assessment based on swin transformer with multi-stage fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1269–1278, 2022. 11

[95] 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. 3

[96] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6312–6322, 2023. 15

[97] 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. 3

[98] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 2, 3

[99] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 7

[100] 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 *Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 538–554. Springer, 2022. 3, 4, 5, 15, 16

[101] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 15

[102] Haoning Wu, Liang Liao, Jingwen Hou, Chaofeng Chen, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. *arXiv preprint arXiv:2302.13269*, 2023. 15

[103] 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*, pages 1045–1054, 2023. 3, 16

[104] 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 *International Conference on Computer Vision (ICCV)*, 2023. 3, 4, 5, 14, 15

[105] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 2

[106] 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. 2, 12

[107] 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, 11, 14

[108] 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. 2

[109] 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*, pages 7623–7633, 2023. 3

[110] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 2

[111] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 12

[112] 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. 2, 10, 12, 13, 15

[113] Junfeng Yang, Jing Fu, Wei Zhang, Wenzhi Cao, Limei Liu, and Han Peng. Moe-agiqa: Mixture-of-experts boosted visual perception-driven and semantic-aware quality assessment for ai-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 9

[114] 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. 14

[115] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021. 11, 13, 16

[116] 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. 8

[117] Jiquan Yuan, Xinyan Cao, Changjin Li, Fanyi Yang, Jinlong Lin, and Xixin Cao. Pku-i2iqa: An image-to-image quality assessment database for ai generated images. *arXiv preprint arXiv:2311.15556*, 2023. 9, 13

[118] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 3

[119] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2736–2746, 2022. 12

[120] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilin-

ear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020. 2, 4, 5

[121] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 2, 4, 5, 6

[122] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 12

[123] 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, 9, 11

[124] Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Advancing zero-shot digital human quality assessment through text-prompted evaluation. *arXiv preprint arXiv:2307.02808*, 2023. 3

[125] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al. Q-boost: On visual quality assessment ability of low-level multi-modality foundation models. *arXiv preprint arXiv:2312.15300*, 2023. 2

[126] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Kang Fu, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A reduced-reference quality assessment metric for textured mesh digital humans. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2965–2969. IEEE, 2024. 3

[127] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22302–22313, 2023. 15

[128] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 17

[129] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016. 12