

# NTIRE 2025 XGC Quality Assessment Challenge: Methods and Results

Xiaohong Liu*	Xionghuo Min*	Qiang Hu*	Xiaoyun Zhang*	Jie Guo*
Guangtao Zhai*	Shushi Wang*	Yingjie Zhou*	Lu Liu*	Jingxin Li*
Farong Wen*	Li Xu*	Yanwei Jiang*	Xilei Zhu*	Chunyi Li*
Huiyu Duan*	Xiele Wu*	Yixuan Gao*	Yuqin Cao*	Jun Jia*
Jiezhong Cao*	Radu Timofte *	Baojun Li	Jiamian Huang	Dan Luo
Weixia Zhang	Bingkun Zheng	Junlin Chen	Ruikai Zhou	Meiya Chen
Hao Jiang	Xiantao Li	Yuxiang Jiang	Jun Tang	Yimeng Zhao
Zelu Qi	Chaoyang Zhang	Fei Zhao	Ping Shi	Lingzhi Fu
Shuai He	Rongyu Zhang	Jiarong He	Zongyao Hu	Wei Luo
Fengbin Guan	Yiting Lu	Xin Li	Zhibo Chen	Mengjing Su
Tuo Chen	Chunxiao Li	Shuaiyu Zhao	Jiaxin Wen	Chuyi Lin
Ningxin Chu	Jing Wan	Yu Zhou	Baoying Chen	Jishen Zeng
Xianjin Liu	Xin Chen	Lanzhi Zhou	Hangyu Li	You Han
Zhenjie Liu	Jianzhang Lu	Jialin Gui	Renjie Lu	Shangfei Wang
Jingyu Lin	Quanjian Song	Jiancheng Huang	Yufeng Yang	Changwei Wang
Shupeng Zhong	Yang Yang	Lihuo He	Jia Liu	Yuting Xing
		Yuchun Jin		

## Abstract

This paper reports on the NTIRE 2025 XGC Quality Assessment Challenge, which will be held in conjunction with the New Trends in Image Restoration and Enhancement Workshop (NTIRE) at CVPR 2025. This challenge is to address a major challenge in the field of video and talking head processing. The challenge is divided into three tracks, including user generated video, AI generated video and talking head.

The user-generated video track uses the FineVD-GC, which contains 6,284 user generated videos. The user-generated video track has a total of 125 registered participants. A total of 242 submissions are received in the development phase, and 136 submissions are received in the test phase. Finally, 5 participating teams submitted their models and fact sheets.

The AI generated video track uses the Q-Eval-Video, which contains 34,029 AI-Generated Videos (AIGVs) generated by 11 popular Text-to-Video (T2V) models. A total of 133 participants have registered in this track. A total of 396 submissions are received in the development phase, and 226 submissions are received in the test phase. Finally, 6 par-

ticipating teams submitted their models and fact sheets.

The talking head track uses the THQA-NTIRE, which contains 12,247 2D and 3D talking heads. A total of 89 participants have registered in this track. A total of 225 submissions are received in the development phase, and 118 submissions are received in the test phase. Finally, 8 participating teams submitted their models and fact sheets.

Each participating team in every track has proposed a method that outperforms the baseline, which has contributed to the development of fields in three tracks.

## 1. Introduction

With the rapid development of video generation technologies, User-Generated Videos (UGVs), AI-Generated Videos (AIGVs), and Talking Head have become widely used in various applications. However, the quality of these videos can vary significantly due to differences in capture conditions, generation models, and animation techniques. Therefore, it is crucial to develop effective Video Quality Assessment (VQA) methods to accurately evaluate the visual quality of UGVs, AIGVs, and Talking Head, ensuring better user experience and reliable performance in real-world scenarios. A robust quality assessment framework can help identify distortions, enhance generation techniques, and op-

\*The organizers of the NTIRE 2025 XGC Quality Challenge.

timize models for improved visual fidelity and realism.

This NTIRE 2025 XGC Quality Assessment Challenge aims to promote the development of the quality assessment methods for videos and talking heads to guide the improvement and enhancement of the video capture, compression, and processing techniques and performance of generative models. The challenge is divided into three tracks, including user generated video track, AI generated video track and talking head track. In the user generated video track, we use the FineVD-GC [45], which contains 6,284 user generated videos. 120 subjects are invited to produce accurate Mean Opinion Scores (MOSs). The AI generated video track uses the Q-Eval-Video [101], in which 11 popular Text-to-Video (T2V) models are used to generate 34,029 videos. And the Sample & Scrutinize strategy was employed during this dataset annotation process to make sure the quality and accuracy of the dataset. The talking head track uses the THQA-NTIRE [117, 118], which contains 12,247 2D and 3D talking heads.

This challenge has a total of 347 registered participants, 125 in the user generated video track, 133 in the AI generated video track and 89 in the talking head track. A total of 863 submissions were received in the development phase, while 480 prediction results were submitted during the final testing phase. Finally, 5 valid participating teams in the user generated video track, 6 valid participating teams in the AI generated video track and 9 valid participating teams in the talking head track submitted their final models and fact sheets. They have provided detailed introductions to their quality assessment methods. We provide the detailed results of the challenge in Section 4 and Section 5. We hope that this challenge can promote the development of quality assessment in video and talking head.

This challenge is one of the NTIRE 2025<sup>1</sup> Workshop associated challenges on: ambient lighting normalization [73], reflection removal in the wild [90], shadow removal [72], event-based image deblurring [66], image denoising [67], XGC quality assessment [51], UGC video enhancement [65], night photography rendering [23], image super-resolution (x4) [12], real-world face restoration [13], efficient super-resolution [63], HR depth estimation [95], efficient burst HDR and restoration [44], cross-domain few-shot object detection [25], short-form UGC video quality assessment and enhancement [47, 48], text to image generation model quality assessment [33], day and night rain-drop removal for dual-focused images [46], video quality assessment for video conferencing [38], low light image enhancement [53], light field super-resolution [80], restore any image model (RAIM) in the wild [50], raw restoration and super-resolution [16] and raw reconstruction from RGB on smartphones [17].

<sup>1</sup><https://www.cvlai.net/ntire/2025/>

## 2. Related Work

### 2.1. User Generated VQA Dataset

Over the years, researchers have developed various video quality assessment (VQA) datasets to analyze human visual perception characteristics. Initial datasets primarily examined synthetic degradations, employing restricted original content and manually simulated degradation patterns [18, 60, 74]. With the rise of user-generated content (UGC) platforms, contemporary research has shifted toward creating VQA databases that capture genuine quality issues encountered in practical scenarios. Multiple studies [22, 29, 35, 36] have specifically addressed real-world quality deterioration occurring during content capture or natural viewing environments. Other comprehensive datasets [49, 79, 105, 122] have incorporated both simulated and authentic distortion types to broaden research scope. While existing UGC collections predominantly source materials from conventional platforms like YouTube, emerging datasets like KVQ [57] specifically target short-format video content. Our proposed FineVD-GC expands this landscape by encompassing diverse video formats including on-demand streaming, conventional UGC, and short-form media. Unlike existing databases providing singular quality ratings, FineVD-GC’s multi-dimensional annotations enable broader practical implementations through detailed quality characterization.

### 2.2. AI Generated VQA Dataset

Compared with user generated video quality assessment datasets, the number of proposed AI generated video (AIGV) datasets is small. Chivileva *et al.* [15] proposes a dataset with 1,005 videos generated by 5 T2V models. 24 users are involved in the subjective study. EvalCrafter [54] 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 [55] uses 619 prompts, 4 T2V models, and 3 users for annotation as well. VBench [37] has a larger scale with in total of  $\sim 1.7k$  prompts and 4 T2V models. Continuing with the exploration of AIGV quality assessment, the T2VQA-DB [42] emerges as a significant addition to the landscape. The dataset has 10,000 videos generated by 9 different T2V models. 27 subjects are invited to collect the MOSs. In this track, we use the latest dataset, Q-Eval-Video [101], which contains approximately 34,000 videos generated by 11 different models. Meanwhile, the Sample & Scrutinize strategy was employed during the dataset annotation process.

### 2.3. VQA Model

The traditional VQA models are usually designed for user-generated videos or a certain attribute of videos. [19, 26, 32, 40, 41, 52, 98, 103, 106]. For example, SimpleVQA [68] 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 [83] proposes the “fragments” sampling strategies and the Fragment Attention Network (FANet) to accommodate fragments as inputs. Light-VQA [112] and Light-VQA+ [20] provide methods for assessing the quality of videos enhanced in low-light conditions. DOVER [86] evaluates the quality of videos from the technical and aesthetic perspectives respectively. Q-Align [87] can also address the VQA task by relying on the ability of multi-modal large models [88, 104, 107]. VQA<sup>2</sup> [39] further explores the approach of utilizing multi-modal large models for video quality assessment through visual question answering.

There are several works targeting the VQA tasks of AIGVs. VBench [37], EvalCrafter [54] and Q-Bench-Video [100] build benchmarks for AIGVs by designing multi-dimensional metrics. MaxVQA [85] and FETV [55] propose separate metrics for the assessment of video-text alignment and video fidelity, while T2VQA [42] handles the features from the two dimensions as a whole. The newly emerged model, Q-Eval-Score [101] explores the use of Multimodal Large Language Models (MLLMs) for assessing the quality of AIGV. For the VQA tasks of AIGV, further research is still needed. We believe the development of these models for AIGV will certainly benefit the generation of high-quality videos.

## 2.4. Talking Head

Talking Heads is an emerging form of human-centered media, distinguished by the integration of realistic facial and vocal features [30, 113]. The conventional approach to designing Talking Heads predominantly relies on facial capture technology, wherein designers utilize 3D software to map facial bones and fine-tune facial details for a specific digital persona, based on the captured facial data [61, 93]. While this manual technique can yield high-quality Talking Heads, the substantial costs associated with the required equipment, coupled with the complexity of the operation, significantly constrain the efficiency of the design process.

To address the challenges associated with Talking Head design, a range of AI-based methods have been developed. These methods can be categorized based on the type of data used to generate Talking Heads, distinguishing between generative 2D [31, 76, 82, 89, 94] and generative 3D Talking Heads [21, 27, 34, 78, 92, 123]. Furthermore, generation techniques can be classified into vision-driven [2, 7, 8, 64, 71] and speech-driven [14, 59, 62, 77, 96, 99, 111, 115] approaches, depending on the fundamental principles behind their generation. Given current prominence of Talking Head generation as an active area of research, it is anticipated that more effective methods will continue to emerge.

However, existing quality assessments for Talking Heads

are often limited to subjective evaluations and traditional objective quality metrics, such as PSNR and SSIM [81]. While these approaches provide some insights into the quality of Talking Heads, they have notable limitations. Subjective assessments are typically time-consuming and not conducive to large-scale quantitative analysis, while objective metrics like PSNR and SSIM [81] fail to capture human visual experiences and are inadequate for evaluating generative Talking Heads due to the absence of reference data. Therefore, the development of a more accurate and reliable objective quality assessment framework for Talking Heads is crucial to advancing the field of Talking Head generation.

## 2.5. Digital Human Quality Assessment

With the rapid advancement of digital human technology, the quality of digital humans has garnered significant attention. To explore this issue in greater depth, Zhang *et al.* have developed several datasets, including DHHQA [102], DDH-QA [108], and SJTU-H3D [103], focusing on captured 3D digital humans. These datasets provide rich data for assessing the quality of static heads, dynamic full-body digital humans, and static full-body digital humans. Additionally, they have designed full-reference [102], reduced-reference [106], and no-reference evaluation methods for these datasets, incorporating siamese networks [102], multi-task learning [119], and multi-modal information fusion [11, 109] techniques. These approaches not only offer reliable assessment frameworks for various types of digital humans but also account for different applicability scenarios. Furthermore, to investigate potential quality degradation during communication transmission, Zhou *et al.* and Zhang *et al.* have conducted user experience quality assessments for 3D talking heads and 3D talking digital humans, respectively. They first established the THQA-3D [118] and 6G-DTQA [110] datasets and proposed corresponding objective evaluation algorithms that integrate channel parameters, visual features, and audio features. Despite these advancements, existing datasets for digital human quality assessment are often constrained by limited data size and insufficient diversity of digital human models, which in turn restricts the generalizability of assessment algorithms.

In recent years, the rapid growth of generative AI has enabled more efficient solutions for designing and acquiring digital humans [116, 120, 121]. In response, Zhou *et al.* developed the first THQA dataset [117] for speech-driven Talking Heads. This dataset includes 800 Talking Heads generated by applying eight representative speech-driven algorithms to 20 images. While this dataset introduces the Talking Head Quality Assessment challenge, it unfortunately does not provide a reliable quality assessment framework. To address this gap, the present work seeks to establish a comprehensive evaluation scheme for this emerging media by engaging experts in discussions on the develop-

ment of an appropriate assessment methodology.

### 3. NTIRE 2025 XGC Quality Assessment Challenge

We organize the NTIRE 2025 XGC Quality Assessment Challenge, including user generated video quality assessment, AI generated video quality assessment and talking head quality assessment, in order to promote the development of objective quality assessment methods. The main goal of the challenge is to predict the perceptual quality of videos and talking heads. Details about the challenge are as follows:

#### 3.1. Overview

The challenge has three tracks, *i.e.* user generated video track, AI generated video track and talking head track. The task is to predict the perceptual quality of video and talking head based on a set of prior examples and their perceptual quality labels. The challenge uses FineVD-GC [22], the Q-Eval-Video [101] and THQA[117, 118] 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 user-generated video track, we use a new dataset called “Fine-grained Video Database - Generated Content” (FineVD-GC) [22], which comprises a total of 6,284 web-crawled UGC videos sourced from YouTube, TikTok, and Bilibili. Each video is randomly clipped into 8-second segments. We initially employ FastVQA [83] to assess the video quality. Based on the distribution of the quality scores, we uniformly sample videos that span a wide range of categories and exhibit a diverse spectrum of quality. These videos are subsequently manually filtered to ensure a comprehensive representation of various distortion types. 120 subjects are invited to rate the videos in FineVD-GC. After normalizing and averaging the subjective opinion scores, the mean opinion score (MOS) of each video can be obtained. Furthermore, we randomly split the FineVD-GC into a training set, a validation set, and a testing set according to the ratio of 4 : 1 : 1. The numbers of videos in the training set, validation set, and testing set are 4, 190, 1, 047, and 1, 047, respectively.

In the AI generated video track, we use the Q-Eval-Video [101]. The dataset contains 34,000 generated videos from: CogVideoX [91], GEN-2 [28], GEN-3 [1], Latte [58], Kling [70], Dreamina [9], Luma [3], PixVerse [4], Pika [43], SVD [6] and Vidu [5]. These videos were generated using approximately 4,700 prompts sampled from VBench, EvalCrafter, T2VCompench, and VideoFeedback. Every video resolution is unified to  $512 \times 512$ , and the video length is 2s.

In the Talking Head track, we utilize the THQA-NTIRE dataset for training, validation, and testing. This dataset integrates and extends the existing THQA [117] and THQA-3D [118] datasets, comprising a total of 12,247 Talking Heads. Specifically, it includes 11,247 generative 2D Talking Heads and 1,000 3D Talking Heads, providing a comprehensive dataset for the development of a unified Talking Head quality assessment framework. All Talking Heads in the dataset contain audio information and exhibit a diverse range of resolutions and durations, thereby posing increased challenges for accurate and robust quality assessment.

#### 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 quality assessment 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.

#### 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 are then sent to the participants.

Table 1. Quantitative results for the NTIRE 2025 XGC Quality Assessment Challenge: Track 1 User Generated Video. *Color*, *Noise*, *Artifact*, *Blur*, *Temporal*, and *Overall* indicate the main scores for each dimension.

Rank	Team	Leader	Color	Noise	Artifact	Blur	Temporal	Overall	Main Score	SRCC	PLCC
1	SLCV	Baojun Li	0.8898	0.8411	0.8805	0.9101	0.8216	0.8954	0.8731	0.8724	0.8738
2	SJTU-MOE-AI	Weixia Zhang	0.8734	0.8327	0.8706	0.8880	0.8237	0.8836	0.8620	0.8591	0.8649
3	MiVQA	Ruikai Zhou	0.8655	0.8136	0.8467	0.8695	0.8055	0.8636	0.8440	0.8386	0.8494
4	XGC-Go	Xiantao Li	0.8512	0.7906	0.8353	0.8575	0.7623	0.8521	0.8248	0.8222	0.8273
5	FoodVQA	Yimeng Zhao	0.8390	0.7773	0.8239	0.8527	0.7633	0.8415	0.8162	0.8125	0.8199
Baseline	FastVQA [84]		0.7982	0.7476	0.7929	0.7988	0.7325	0.8038	0.7789	0.7740	0.7837

Table 2. Quantitative results for the NTIRE 2025 Quality Assessment of AI-Generated Content Challenge: Track 2 AI Generated Video.

Rank	Team	Leader	Main Score	SRCC	PLCC
1	SLCV	Baojun Li	0.6645	0.6621	0.6669
2	CUC-IMC	Zelu Qi	0.6310	0.6080	0.6539
3	opdai	Lingzhi Fu	0.5903	0.5854	0.5952
4	Magnolia	Zongyao Hu	0.5889	0.5933	0.5844
5	AIGC VQA	Wei Luo	0.5606	0.5485	0.5727
6	SJTU-MOE-AI	Bingkun Zheng	0.5463	0.5530	0.5396
Baseline	Q-Eval-Score [101]		0.4741	0.4861	0.4642
	DOVER [86]		0.5055	0.5057	0.5054
	T2VQA [42]		0.5161	0.5161	0.5160

Table 3. Quantitative results for the NTIRE 2025 XGC Quality Assessment: Track 3 Talking Head.

Rank	Team	Leader	Main Score	SRCC	PLCC
1	QA Team	Mengjing Su	0.8244	0.8036	0.8453
2	MediaForensics	Baoying Chen	0.8236	0.8024	0.8448
3	AutoHome AIGC	Xin Chen	0.8046	0.7864	0.8229
4	USTC-AC	Zhenjie Liu	0.8044	0.7813	0.8275
5	SJTU-MOE-AI	Junlin Chen	0.8003	0.7797	0.8209
6	FocusQ	Donghao Zhou	0.7921	0.7708	0.8135
7	NJUST-KMG	Shupeng Zhong	0.7896	0.7599	0.8193
8	XIDIAN-VQATeam	Lihuo He	0.7872	0.7730	0.8015
Baseline	SimpleVQA [69]		0.7862	0.7662	0.8062

## 4. Challenge Results

5 teams in the user generated video track, 6 teams in the AI generated video track and 8 teams in the talking head track have submitted their final codes/executables and fact sheets. Table 1, Table 2 and Tabel 3 summarize the main results and important information of the 19 valid teams. Detailed information about all participating teams and their algorithms can be found in the supplementary materials.

### 4.1. Baselines

We compare the performance of submitted methods with several quality assessment methods on the testing set, including FastVQA [84], Q-Eval-Score [101], DOVER [86], T2VQA [42] and SimpleVQA [69] for these three tracks.

### 4.2. Result Analysis

The main results of 19 teams’ methods and the baseline methods are shown in Table 1, Table 2 and Table 3. It



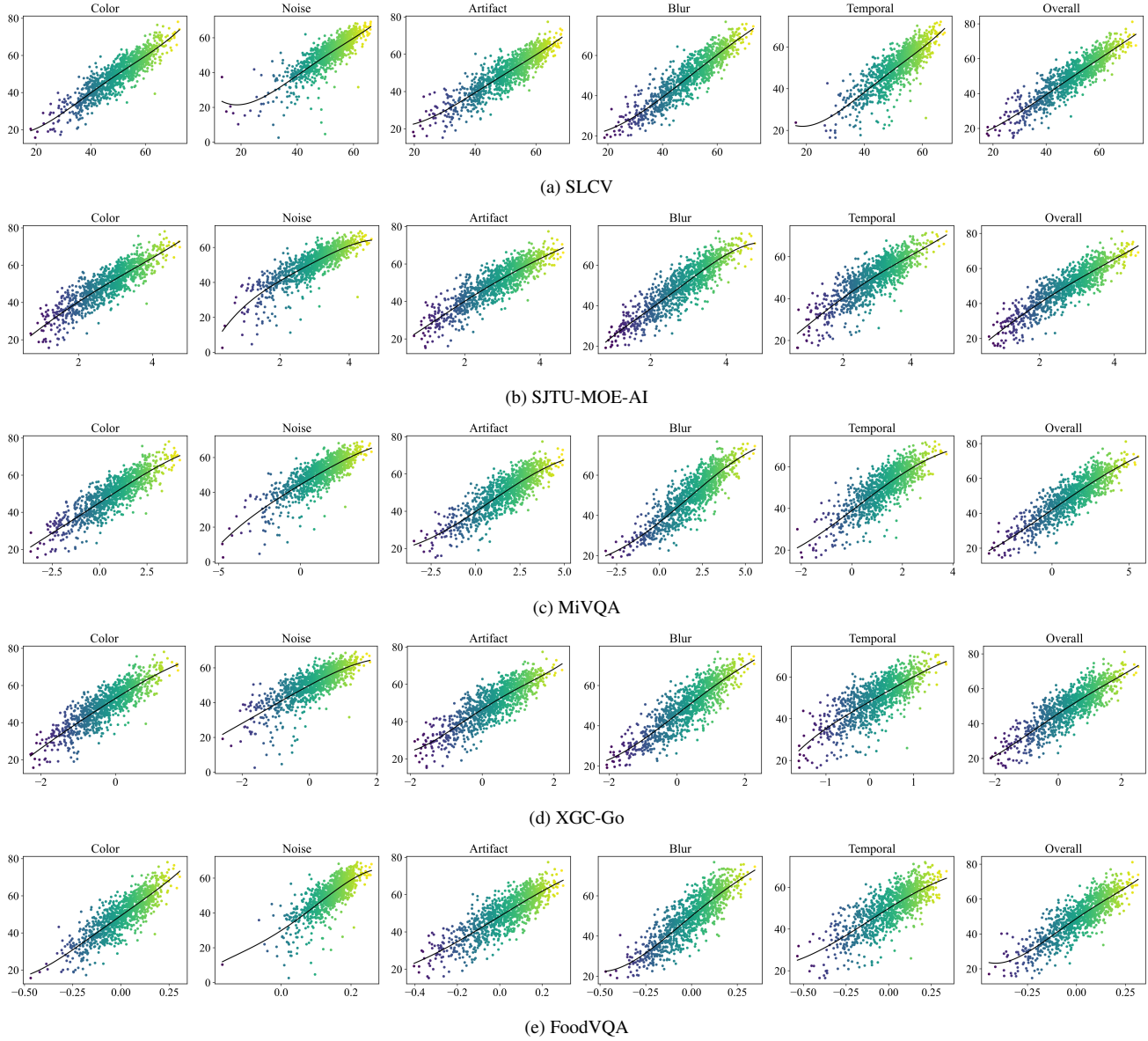


Figure 1. Scatter plots of the predicted scores vs. MOSs in the user-generated video track. The curves are obtained by a four-order polynomial nonlinear fitting.

can be seen that in three tracks, the results of the baseline methods are not all ideal in the testing set of three 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 content. In the user generated video track, 5 teams all achieve a main score higher than 0.8, and 2 teams are higher than 0.85. In the AI generated video track, 6 teams achieve a main score higher than 0.5, 2 teams higher than 0.6, and the championship team is higher than 0.65. In the talking head track, 7 teams achieve a main score higher than baseline, and 5

teams higher than 0.8. In the meantime, the top-ranked teams only have a small difference in the main score. Figures 1 and 2 show scatter plots of predicted scores versus MOSs for the 10 teams' methods on the testing set. The curves are obtained by polynomial nonlinear fitting. We can observe that the predicted scores obtained by the top team methods have higher correlations with the MOSs. In track 3, Figure 3 more intuitively shows the performance of the 8 teams' methods. These results demonstrate the effectiveness of the submitted methods in improving quality assessment across all tracks, highlighting their potential for

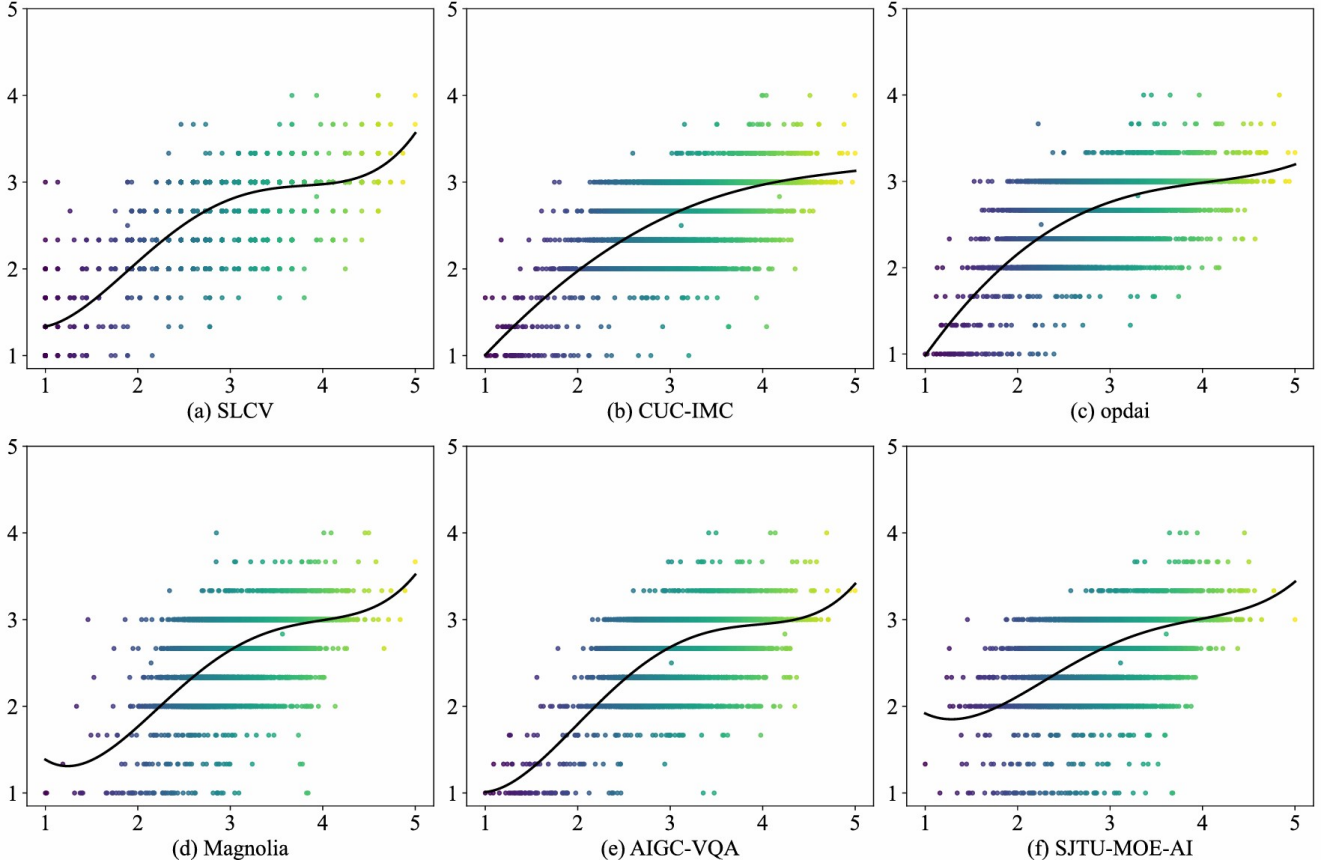


Figure 2. Scatter plots of the predicted scores vs. MOSs in the AI generated video track. The curves are obtained by a four-order polynomial nonlinear fitting.

better alignment with human perception.

## 5. Challenge Winner Methods

### 5.1. User-generated Video Track

Team SLCV wins the championship in the user-generated video track. Unlike conventional approaches that rely on regression or classification for video quality assessment (e.g., LIQE [97], Q-Align [87], Fast-VQA [84], and SimpleVQA [69]), their method leverages a multimodal large language model (MLLM) to estimate video quality. In InternVL 2.5 [10], an effective data filtering process was introduced, leveraging large language model (LLM) scoring to evaluate and remove low-quality samples, thereby improving the overall quality of the training data. Inspired by this capability of InternVL 2.5 to assess data quality using LLM-based scoring, they adopt a multimodal large language model (MLLM) for estimating video quality in our work. Specifically, they directly utilize the InternVL 2.5 model as the MLLM to achieve robust and reliable video quality assessment. To overcome the limitation in the spatial domain, they introduce Spatial Window

Sampling as a data augmentation strategy. Specifically, they employ a sliding window approach that crops the original video frames with a window size set to 3/4 of the video’s longest side. This method effectively triples the amount of training data, thereby enhancing the model’s ability to learn fine-grained spatial features. They employ the LoRA (Low-Rank Adaptation) method to efficiently fine-tune the InternVL 2.5 model, enabling it to perform the six fine-grained quality assessments. During inference, the same data processing strategy used during training is applied to the test videos. Specifically, the model independently predicts quality scores for the three sub-videos generated by the sliding window sampling process. The final prediction is then obtained by averaging the results across these sub-videos. This approach not only ensures robust training but also facilitates accurate and reliable evaluation of fine-grained video quality.

### 5.2. AI Generated Video Track

Team SLCV is the final winner of the AI generated video track. They propose temporal pyramid sampling, to ad-

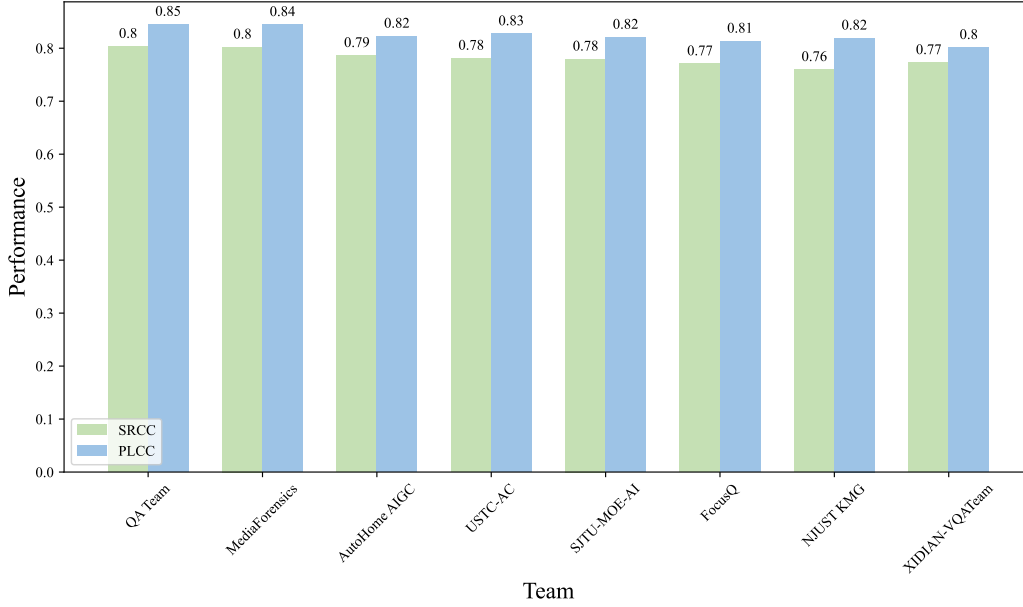


Figure 3. The performance of methods proposed by different teams in Talking head track.

dress the unique challenges posed by AI-generated videos in quality assessment. Unlike user generated video, the quality assessment of these AI generated videos primarily focuses on two core aspects: the smoothness of object motion and the authenticity of the content. To effectively capture these critical metrics, the team design the temporal pyramid sampling method to capture the dynamic characteristics of videos at multiple temporal resolutions. This is achieved by performing multi-scale frame interval sampling at varying frequencies. The original video is sampled at different frame rates and lengths, generating multiple subsets of data with diverse temporal granularities. Each subset is then used to independently train the model, enabling it to learn distinct motion smoothness and content authenticity features at different temporal scales.

### 5.3. Talking Head Track

The QA Team wins the champion in the Talking Head (TH) track. They proposed a novel NR video quality assessment model based on multimodal feature representations, comprising four modules: spatial feature extraction, temporal feature extraction, audio feature extraction, and audio-visual fusion. Visual distortions are categorized into spatial and motion distortions. The types of visual distortions in videos can be roughly divided into two categories: spatial distortion[114] and motion distortion. First, Talking Head videos are split into clips for spatial and temporal feature extraction. Whole clip is utilized for temporal feature extraction with a fixed pretrained 3D-CNN backbone SlowFast[24]. The first frame of each clip is used for spatial

feature extraction. The spatial feature extraction module utilizes an efficient channel attention module ECA-Net[75], to effectively achieve cross-channel interaction, and then utilize the SwinTransformer-tiny[56] to extract visual features from the first frame.

For audio feature extraction, the audio is aligned with the visual frames, and four techniques—chromagram, CQT, MFCC, and GFCC—are used to extract time-frequency features. These features are stacked into 4 channels and fed into a separable convolution network with frequency, time, and fusion blocks, each consisting of Conv2D layers, BatchNorm, and Maxpool. The frequency and time blocks use  $1 \times m$  and  $n \times 1$  kernels, respectively, to perform spatially separable convolutions, reducing parameters. Temporal information is processed using Bi-LSTM, which captures context from both past and future sequences. Finally, the features are fused into a quality score using fully connected (FC) layers.

Videos are divided into 1-second clips, with 6 clips selected via cyclic sampling. The Swin Transformer extracts spatial features with  $3 \times 224 \times 224$  patches, while SlowFast extracts temporal features from resized  $224 \times 224$  clips.

### Acknowledgments

This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2025 sponsors: ByteDance, Meituan, Kuaishou, and University of Wurzburg (Computer Vision Lab).



## References

- [1] Introducing gen-3 alpha: A new frontier for video generation. In <https://runwayml.com/research/introducing-gen-3-alpha>, 2024. 4
- [2] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Audio-visual face reenactment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5178–5187, 2023. 3
- [3] Luma AI. Dream machine: Ai video generator. In <https://lumalabs.ai/dream-machine>, 2024. 4
- [4] PixVerse AI. Pixverse: Ai video creation platform. In <https://pixverse.ai/>, 2024. 4
- [5] Vidu AI Team. Vidu ai. In <https://www.vidu.studio/zh>, 2024. 4
- [6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, and Adam Letts. Stable video diffusion: Scaling latent video diffusion models to large datasets. In *arXiv preprint:2311.15127*, 2023. 4
- [7] Stella Bounareli, Vasileios Argyriou, and Georgios Tzimiropoulos. Finding directions in gan’s latent space for neural face reenactment. *arXiv preprint arXiv:2202.00046*, 2022. 3
- [8] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Stylemask: Disentangling the style space of stylegan2 for neural face reenactment. In *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2023. 3
- [9] Dreamina by CapCut. Dreamina. In <https://dreamina.capcut.com/>, 2023. 4
- [10] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024. 7
- [11] Shi Chen, Zicheng Zhang, Yingjie Zhou, Wei Sun, and Xiongkuo Min. A no-reference quality assessment metric for dynamic 3d digital human. *Displays*, 80:102540, 2023. 3
- [12] Zheng Chen, Kai Liu, Jue Gong, Jingkai Wang, Lei Sun, Zongwei Wu, Radu Timofte, Yulun Zhang, et al. NTIRE 2025 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [13] Zheng Chen, Jingkai Wang, Kai Liu, Jue Gong, Lei Sun, Zongwei Wu, Radu Timofte, Yulun Zhang, et al. NTIRE 2025 challenge on real-world face restoration: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [14] Kun Cheng et al. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia*, 2022. 3
- [15] Iya Chivileva, Philip Lynch, Tomas E Ward, and Alan F Smeaton. Measuring the quality of text-to-video model outputs: Metrics and dataset. *arXiv preprint arXiv:2309.08009*, 2023. 2
- [16] Marcos Conde, Radu Timofte, et al. NTIRE 2025 challenge on raw image restoration and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [17] Marcos Conde, Radu Timofte, et al. Raw image reconstruction from RGB on smartphones. NTIRE 2025 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [18] Francesca De Simone, Marco Tagliasacchi, Matteo Naccari, Stefano Tubaro, and Touradj Ebrahimi. A h. 264/avc video database for the evaluation of quality metrics. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2430–2433. IEEE, 2010. 2
- [19] 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. 2
- [20] 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*, 2023. 3
- [21] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 14398–14407, 2021. 3
- [22] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, et al. Finevq: Fine-grained user generated content video quality assessment. *arXiv preprint arXiv:2412.19238*, 2024. 2, 4
- [23] Egor Ershov, Sergey Korchagin, Alexei Khalin, Artyom Panshin, Arseniy Terekhin, Ekaterina Zaychenkova, Georgiy Lobarev, Vsevolod Plokhhotnyuk, Denis Abramov, Elisey Zhdanov, Sofia Dorogova, Yasin Mamedov, Nikola

- Banic, Georgii Perevozchikov, Radu Timofte, et al. NTIRE 2025 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [24] 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. 8
- [25] Yuqian Fu, Xingyu Qiu, Bin Ren Yanwei Fu, Radu Timofte, Nicu Sebe, Ming-Hsuan Yang, Luc Van Gool, et al. NTIRE 2025 challenge on cross-domain few-shot object detection: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [26] 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. 2
- [27] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5609–5619, 2023. 3
- [28] Anastasis Germanidis. Gen-2: Generate novel videos with text, images or video clips. In <https://runwayml.com/research/gen-2>, 2023. 4
- [29] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 28(9):2061–2077, 2017. 2
- [30] Shihan Guo, Jiachen Guo, Han Wang, Haibo Wang, Xiaoling Huang, and Lin Zhang. An efficient ophthalmic disease qa system integrated with knowledge graphs and digital humans. In *2024 7th International Conference on Information Communication and Signal Processing (ICICSP)*, pages 1094–1098. IEEE, 2024. 3
- [31] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10893–10900, 2020. 3
- [32] Jinliang Han, Xiongkuo Min, Jun Jia, Yixuan Gao, Xiaohong Liu, and Guangtao Zhai. Full-reference and no-reference quality assessment for video frame interpolation. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2
- [33] Shuhao Han, Haotian Fan, Fangyuan Kong, Wenjie Liao, Chunle Guo, Chongyi Li, Radu Timofte, et al. NTIRE 2025 challenge on text to image generation model quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [34] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 3
- [35] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *Proceedings of the 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017. 2
- [36] Qiang Hu, Qihan He, Houqiang Zhong, Guo Lu, Xiaoyun Zhang, Guangtao Zhai, and Yanfeng Wang. Varfvv: View-adaptive real-time interactive free-view video streaming with edge computing. *IEEE Journal on Selected Areas in Communications*, pages 1–1, 2025. 2
- [37] 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
- [38] Varun Jain, Zongwei Wu, Quan Zou, Louis Florentin, Henrik Turbell, Sandeep Siddhartha, Radu Timofte, et al. NTIRE 2025 challenge on video quality enhancement for video conferencing: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [39] Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. Vqa<sup>2</sup>: Visual question answering for video quality assessment. In *arXiv preprint arXiv:2503.10078*, 2025. 3
- [40] Tengchuan Kou, Xiaohong Liu, Jun Jia, Wei Sun, 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*, 2023. 2
- [41] 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. 2
- [42] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024. 2, 3, 5
- [43] Pika Labs. Pika: Ai video generation platform. In <https://pika.art/>, 2024. 4
- [44] Sangmin Lee, Eunpil Park, Angel Canelo, Hyunhee Park, Youngjo Kim, Hyungju Chun, Xin Jin, Chongyi Li, Chunle Guo, Radu Timofte, et al. NTIRE 2025 challenge on efficient burst hdr and restoration: Datasets, methods, and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [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. Aigiga-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
- [46] Xin Li, Yeying Jin, Xin Jin, Zongwei Wu, Bingchen Li, Yufei Wang, Wenhan Yang, Yu Li, Zhibo Chen, Bihan Wen, Robby Tan, Radu Timofte, et al. NTIRE 2025 challenge on day and night raindrop removal for dual-focused images: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [47] Xin Li, Xijun Wang, Bingchen Li, Kun Yuan, Yizhen Shao, Suhang Yao, Ming Sun, Chao Zhou, Radu Timofte, and Zhibo Chen. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Kwaisr dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [48] Xin Li, Kun Yuan, Bingchen Li, Fengbin Guan, Yizhen Shao, Zihao Yu, Xijun Wang, Yiting Lu, Wei Luo, Suhang Yao, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [49] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. Ugc-video: Perceptual quality assessment of user-generated videos. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 35–38. IEEE, 2020. 2
- [50] Jie Liang, Radu Timofte, Qiaosi Yi, Zhengqiang Zhang, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Lei Zhang, et al. NTIRE 2025 the 2nd restore any image model (RAIM) in the wild challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [51] Xiaohong Liu, Xiongkuo Min, Qiang Hu, Xiaoyun Zhang, Jie Guo, et al. NTIRE 2025 XGC quality assessment challenge: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [52] 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. 2
- [53] Xiaoning Liu, Zongwei Wu, Florin-Alexandru Vasluianu, Hailong Yan, Bin Ren, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. NTIRE 2025 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [54] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023. 2, 3
- [55] 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
- [56] 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. 8
- [57] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2
- [58] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuanfang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. In *arXiv preprint:2401.03048*, 2024. 4
- [59] Yifeng Ma et al. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023. 3
- [60] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):652–671, 2012. 2
- [61] Antonio Moura, Ingrida Mazonaviciute, João Nunes, and Justinas Grigaravicius. Human lips synchronisation in autodesk maya. In *2007 14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*, pages 365–368. IEEE, 2007. 3
- [62] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, 2020. 3
- [63] Bin Ren, Hang Guo, Lei Sun, Zongwei Wu, Radu Timofte, Yawei Li, et al. The tenth NTIRE 2025 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [64] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13759–13768, 2021. 3
- [65] Nickolay Safonov, Alexey Bryntsev, Andrey Moskalenko, Dmitry Kulikov, Dmitriy Vatin, Radu Timofte, et al. NTIRE 2025 challenge on UGC video enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [66] Lei Sun, Andrea Alfano, Peiqi Duan, Shaolin Su, Kaiwei Wang, Boxin Shi, Radu Timofte, Danda Pani Paudel, Luc Van Gool, et al. NTIRE 2025 challenge on event-based image deblurring: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [67] Lei Sun, Hang Guo, Bin Ren, Luc Van Gool, Radu Timofte, Yawei Li, et al. The tenth ntire 2025 image denoising challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*



- Workshops, 2025. 2
- [68] 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. 2
- [69] 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*, page 856–865, 2022. 5, 7
- [70] Kuaishou Team. Kling ai. In <https://klingai.io/>, 2024. 4
- [71] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1329–1338, 2021. 3
- [72] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Cailian Chen, Zongwei Wu, Radu Timofte, et al. NTIRE 2025 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [73] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Radu Timofte, et al. NTIRE 2025 ambient lighting normalization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [74] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. MCL-JCV: a jnd-based h. 264/avc video quality assessment dataset. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1509–1513. IEEE, 2016. 2
- [75] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539, 2020. 8
- [76] Qiulin Wang, Lu Zhang, and Bo Li. Safa: Structure aware face animation. In *2021 International Conference on 3D Vision (3DV)*, pages 679–688. IEEE, 2021. 3
- [77] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 3
- [78] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hiface: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021. 3
- [79] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13435–13444, 2021. 2
- [80] Yingqian Wang, Zhengyu Liang, Fengyuan Zhang, Lvli Tian, Longguang Wang, Juncheng Li, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2025 challenge on light field image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [81] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3
- [82] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 3
- [83] 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
- [84] 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, 2022. 5, 7
- [85] 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
- [86] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *International Conference on Computer Vision (ICCV)*, 2023. 3, 5
- [87] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Corresponding Authors: Zhai, Guangtao and Lin, Weisi. 3, 7
- [88] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, et al. Towards open-ended visual quality comparison. In *arXiv preprint arXiv:2403.11956*, 2024. 3
- [89] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 603–619, 2018. 3
- [90] Kangning Yang, Jie Cai, Ling Ouyang, Florin-Alexandru Vasluianu, Radu Timofte, Jiaming Ding, Huiming Sun, Lan Fu, Jinlong Li, Chiu Man Ho, Zibo Meng, et al. NTIRE 2025 challenge on single image reflection removal in the wild: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [91] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong,

- Xiaohan Zhang, and Guanyu Feng. Cogvideox: Text-to-video diffusion models with an expert transformer. In *arXiv preprint:2408.06072*, 2024. 4
- [92] Guangming Yao, Yi Yuan, Tianjia Shao, and Kun Zhou. Mesh guided one-shot face reenactment using graph convolutional networks. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1773–1781, 2020. 3
- [93] Mateusz Zajac and Szczepan Paszkiel. Using brain-computer interface technology for modeling 3d objects in blender software. *Journal of Automation Mobile Robotics and Intelligent Systems*, 14, 2020. 3
- [94] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 3
- [95] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, Samuele Salti, Stefano Mattoccia, et al. NTIRE 2025 challenge on hr depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 2
- [96] Wenxuan Zhang et al. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *IEEE/CVF CVPR*, 2023. 3
- [97] 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. 7
- [98] Zicheng Zhang, Yu Fan, Wei Sun, Xiongkuo Min, Xiaohong Liu, Chunyi Li, Haoning Wu, Weisi Lin, Ning Liu, and Guangtao Zhai. Paps-ovqa: Projection-aware patch sampling for omnidirectional video quality assessment. In *IEEE International Symposium on Circuits and Systems*, 2024. 2
- [99] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. *arXiv preprint arXiv:2303.03988*, 2023. 3
- [100] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Q-bench-video: Benchmarking the video quality understanding of llms. In *arXiv preprint arXiv:2409.20063*, 2024. 3
- [101] Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, Xiaohong Liu, and Guangtao Zhai. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In *arXiv preprint arXiv:2503.02357*, 2025. 2, 3, 4, 5
- [102] 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. 3
- [103] 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. 2, 3
- [104] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, et al. Q-boost: On visual quality assessment ability of low-level multi-modality foundation models. In *arXiv preprint arXiv:2312.15300*, 2023. 3
- [105] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. Md-vqa: Multi-dimensional quality assessment for ugc live videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1755, 2023. 2
- [106] 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. 2, 3
- [107] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Yixuan Gao, Zicheng Zhang, Chunyi Li, Haoning Wu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. In *arXiv preprint arXiv:2409.00031*, 2024. 3
- [108] Zicheng Zhang, Yingjie Zhou, Wei Sun, Wei Lu, Xiongkuo Min, Yu Wang, and Guangtao Zhai. Ddh-qa: A dynamic digital humans quality assessment database. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2519–2524. IEEE, 2023. 3
- [109] Zicheng Zhang, Yingjie Zhou, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Geometry-aware video quality assessment for dynamic digital human. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1365–1369. IEEE, 2023. 3
- [110] Zicheng Zhang, Yingjie Zhou, Long Teng, Wei Sun, Chunyi Li, Xiongkuo Min, Xiao-Ping Zhang, and Guangtao Zhai. Quality-of-experience evaluation for digital twins in 6g network environments. *IEEE Transactions on Broadcasting*, 70(3):995–1007, 2024. 3
- [111] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *IEEE/CVF CVPR*, 2023. 3
- [112] Xunchu Zhou, Xiaohong Liu, Yunlong Dong, Tengchuan Kou, Yixuan Gao, Zicheng Zhang, Chunyi Li, Haoning Wu, and Guangtao Zhai. Light-vqa+: A video quality assessment model for exposure correction with vision-language guidance. In *arXiv preprint arXiv:2405.03333*, 2024. 3
- [113] Yingjie Zhou, Yaodong Chen, Kaiyue Bi, Lian Xiong, and Hui Liu. An implementation of multimodal fusion system for intelligent digital human generation. *arXiv preprint arXiv:2310.20251*, 2023. 3
- [114] Yu Zhou, Weikang Gong, Yanjing Sun, Leida Li, Jinjian Wu, and Xinbo Gao. Pyramid feature aggregation for hierarchical quality prediction of stitched panoramic images. *IEEE Transactions on Multimedia*, 25:4177–4186, 2023. 8
- [115] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevar-



- ria, Evangelos Kalogerakis, and Dingzeyu Li. Makeltalk: speaker-aware talking-head animation. *ACM TOG*, 2020. 3
- [116] Yingjie Zhou, Zicheng Zhang, Jiezhong Cao, Jun Jia, Yanwei Jiang, Farong Wen, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Memo-bench: A multiple benchmark for text-to-image and multimodal large language models on human emotion analysis. *arXiv preprint arXiv:2411.11235*, 2024. 3
- [117] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Zhihua Wang, Xiao-Ping Zhang, and Guangtao Zhai. Thqa: A perceptual quality assessment database for talking heads. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 15–21. IEEE, 2024. 2, 3, 4
- [118] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Subjective and objective quality-of-experience assessment for 3d talking heads. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6033–6042, 2024. 2, 3, 4
- [119] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiongkuo Min, Xianghe Ma, and Guangtao Zhai. A no-reference quality assessment method for digital human head. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 36–40. IEEE, 2023. 3
- [120] Yingjie Zhou, Zicheng Zhang, Farong Wen, Jun Jia, Yanwei Jiang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 3dgcqa: A quality assessment database for 3d ai-generated contents. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3
- [121] Yingjie Zhou, Zicheng Zhang, Farong Wen, Jun Jia, Xiongkuo Min, Jia Wang, and Guangtao Zhai. Reli-qa: A multidimensional quality assessment dataset for relighted human heads. In *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2024. 3
- [122] Xilei Zhu, Huiyu Duan, Liu Yang, Yucheng Zhu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Esvqa: Perceptual quality assessment of egocentric spatial videos. *arXiv preprint arXiv:2412.20423*, 2024. 2
- [123] Peiye Zhuang, Liqian Ma, Sanmi Koyejo, and Alexander Schwing. Controllable radiance fields for dynamic face synthesis. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 3