

# Who is a Better Talker: Subjective and Objective Quality Assessment for AI-Generated Talking Heads

Yingjie Zhou<sup>1,2</sup> Jiezhong Cao<sup>\*3</sup> Zicheng Zhang<sup>1,2</sup> Farong Wen<sup>1,2</sup>  
 Yanwei Jiang<sup>1,2</sup> Jun Jia<sup>1,2</sup> Xiaohong Liu<sup>\*1,2</sup> Xiongkuo Min<sup>\*1,2</sup> Guangtao Zhai<sup>\*1,2</sup>  
<sup>1</sup> Shanghai Jiao Tong University <sup>2</sup> PengCheng Laboratory <sup>3</sup> Harvard Medical School

## Abstract

Speech-driven methods for portraits are figuratively known as “Talkers” because of their capability to synthesize speaking mouth shapes and facial movements. Especially with the rapid development of the Text-to-Image (T2I) models, AI-Generated Talking Heads (AGTHs) have gradually become an emerging digital human media. However, challenges persist regarding the quality of these talkers and AGTHs they generate, and comprehensive studies addressing these issues remain limited. To address this gap, this paper **presents the largest AGTH quality assessment dataset THQA-10K** to date, which selects 12 prominent T2I models and 14 advanced talkers to generate AGTHs for 14 prompts. After excluding instances where AGTH generation is unsuccessful, the THQA-10K dataset contains 10,457 AGTHs. Then, volunteers are recruited to subjectively rate the AGTHs and give the corresponding distortion categories. In our analysis for subjective experimental results, we evaluate the performance of talkers in terms of generalizability and quality, and also expose the distortions of existing AGTHs. Finally, **an objective quality assessment method based on the first frame, Y-T slice and tone-lip consistency is proposed**. Experimental results show that this method can achieve state-of-the-art (SOTA) performance in AGTH quality assessment. The work is released at <https://github.com/zyj-2000/Talker>.

## 1. Introduction

Digital humans represent an emerging digital media technology focused on generating realistic representations of virtual characters endowed with human-like characters [67]. Currently, the majority of high-quality digital humans are predominantly crafted and manipulated by skilled designers, necessitating extensive expertise and experience. In particular, the design process is cumbersome and time-consuming in terms of character modeling and facial an-

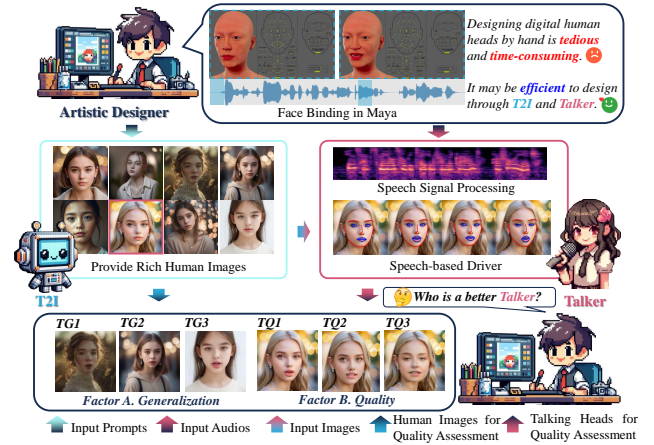


Figure 1. Manual approach to digital human head design versus Text-to-Image and Talker-based approaches.

imation. This method of manual design obviously suffers from low efficiency and high cost, hindering the popularization and promotion of digital humans. Fortunately, advancements in artificial intelligence (AI) have significantly facilitated the design of digital humans. On one hand, various types of text-to-image (T2I) models [1–3, 9–11, 25, 31, 32, 34, 35] have been able to generate various types of character images, allowing for a more diverse appearance of digital humans. On the other hand, a variety of speech-driven methods [6, 7, 23, 33, 42, 43, 46, 47, 49–51, 63, 64], which can be viewed as “Talkers,” have been developed to achieve the effect of Talking Head (TH). Although these talkers have improved the efficiency of digital human design, they inevitably face a variety of quality problems, which adversely impact the user experience. Therefore, it is imperative to evaluate these Talkers to provide objective and reliable reference metrics for various generative methods, thereby fostering the ongoing development of the digital human domain and enhancing user experience with AI-Generated Talking Head (AGTH) videos. More specifically, considering the AGTH generation process, two aspects should be considered to assess the quality of the talkers. A) **Generalization**: A better talker should produce high-quality AGTHs across a wide range of portrait images.

\*Corresponding authors.

Table 1. The comparison of digital human databases and THQA-10K.

Database	Modal	Scale	Methods	Distortions	Description
DHH-QA [55]	Mesh + UV	1,540	Zhang et al. [55], Zhou et al. [69]	7 model distortions	Scanned Real Human Heads
DDHQA [54]	Mesh + UV	800	Zhang et al. [56], Chen et al. [5]	7 model distortions and 2 motion distortions	Dynamic 3D Digital Human
SJTU-H3D [53]	Mesh + UV	1,120	Zhang et al. [53]	7 model distortions	Static 3D Digital Humans
6G-DTQA [60]	Mesh + UV	400	Zhang et al. [60]	3 model distortions and 2 stream media issues	Dynamic 3D Digital Human
THQA-3D [73]	Mesh + UV	1,000	Zhou et al. [73]	5 stream media issues	Scanned Real Human Heads
CDHQA [77]	Video	254	None	3 generative distortions	Interactive Digital Human
THQA [72]	Video + Audio	800	None	9 generative distortions	AI-Generated Talking Heads
AHQA [78]	Video	1200	Zhou et al. [78]	4 generative distortions	Animated Humans
ReLI-QA [75]	Image	840	None	4 relighted methods	Relighted Human Heads
MEMO-Bench [71]	Image	7,145	None	Sentimental Error	Emotional Human Heads
<b>THQA-10K (Ours)</b>	<b>Image + Audio</b>	<b>10,457</b>	<b>None</b>	<b>10 generative distortions</b>	<b>AI-Generated Talking Heads</b>

**B) Quality:** A better talker must generate superior AGTHs for identical images and speech inputs.

Unfortunately, there has been limited research addressing these quality concerns and proposing credible solutions. To tackle this challenge, this paper first introduces the largest and most comprehensive AGTH Quality Assessment dataset named THQA-10K. This dataset encompasses a diverse selection of character materials, considering various ages and genders, and includes 14 prompts for character image generation. Each prompt is paired with five tailored speech sentences serving as driving audio. Furthermore, the dataset employs 12 leading T2I models and 14 talkers to generate AGTHs, resulting in a total of 10,457 instances within THQA-10K. Subsequently, volunteers are recruited to conduct subjective evaluations of the AGTHs, focusing on both distortion categories and visual quality scores. The findings reveal 10 distinct types of distortions among the AGTHs, alongside significant variations in quality depending on the talkers utilized. This highlights the critical need for quality assessments in this field. Leveraging the THQA-10K dataset and subjective ratings, we propose FSCD, an objective quality assessment method for AGTH based on the first frame, Y-T Slice [38] and tone-lip consistency. Experimental results validate the efficacy and superiority of this method, offering reliable objective metrics for the continued advancement of AGTH technologies. The principal contributions of this paper are as follows:

- The THQA-10K dataset, comprising 10,457 AGTHs, has been constructed. To our knowledge, this dataset represents the largest collection created for the purpose of AGTH quality assessment, incorporating 12 prominent T2I models and 14 speech-driven methods.
- An objective quality assessment algorithm, referred to as FSCD, has been designed. This method integrates quality features from the first frame, Y-T slice, and tone-lip consistency to deliver an effective and robust objective evaluation of AGTH quality.
- The proposed THQA-10K dataset is representative and comprehensive to advance the field of digital human design. The designed FSCD method can achieve state-of-the-art (SOTA) objective quality assessment performance, which provides a reliable quality indicator for the field.

## 2. Related Works

### 2.1. Talker: Talking Head Driven Methods

Head-driven animation has been a significant area of research within the field of computer animation, primarily because the human head encompasses a wealth of facial details, expressive movements, and identity information. Traditional methods typically utilize computer-aided animation software, such as Maya<sup>1</sup> or Blender<sup>2</sup>. These approaches require the binding of facial bones to the character model and the establishment of corresponding controllers to manipulate facial movements. Subsequently, keyframe animation techniques are employed, where appropriate keyframes are set for each controller to create various mouth shapes corresponding to different phonemes. This process is notably time-consuming and often necessitates extensive debugging. To address these challenges, the current standard solution in film and television production involves the use of facial capture sensors that track the key points of real THs [12, 21, 24, 37]. The captured data is then imported into computer-aided animation software for further refinement and design. However, high-precision facial capture sensors are prohibitively expensive.

In recent years, advancements in AI have opened new avenues for the production of THs. The emergence of T2I models has not only simplified character image design but has also introduced speech-driven methodologies that directly apply speech to face images for AGTH generation. This approach alleviates a series of cumbersome processes and reduces equipment costs significantly. Speech-driven methods can be further classified into image-based [6, 13, 23, 43, 43, 49] and video-based [7, 33, 42, 46, 47, 50, 51, 63] techniques, depending on the input modality. Nonetheless, the absence of human design and oversight raises concerns regarding the quality of AGTHs, which is the central focus of this paper.

### 2.2. Digital Human Quality Assessment

With the development of digital human quality assessment of digital humans has become an emerging research component. As shown in Table 1, many relevant datasets and ob-

<sup>1</sup><https://www.autodesk.com/products/maya/>

<sup>2</sup><https://www.blender.org/>





Table 2. Details of T2I models employed for generation.

Type	Label	T2I Model	Year	Output Resolution
Closed source	DL3	Dalle3 [3]	2023	1,024×1,024
	MJ6	MidjourneyV6 [25]	2023	1,024×1,024
	IDG	Ideagram [1]	2024	1,024×1,024
	SD2	Stable Diffusion 2.1 [35]	2022	512×512
Open source	SD1	Stable Diffusion 1.5 [35]	2022	512×512
	SDX	Stable Diffusion XL [32]	2023	1,024×1,024
	FCS	Fooocus [11]	2023	1,024×1,024
	KDS	Kandinsky [2]	2023	1,024×1,024
	ODE	OpenDalleV1.1 [31]	2023	1,024×1,024
	PTS	Proteus [34]	2024	1,024×1,024
	FLU	FLUX.1 [10]	2024	1,024×1,024
	SD3	Stable Diffusion 3 [9]	2024	1,024×1,024

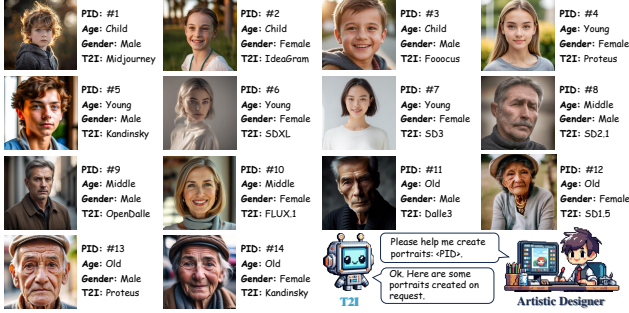


Figure 3. Generated portraits of different PIDs.

Details regarding the selected T2I models are presented in Table 2, with a subset of the generated portraits illustrated in Fig. 3.

Another critical generative model is represented by speech-driven methods, commonly referred as talkers. To ensure that the constructed dataset encompasses a wide range of existing mainstream driving methods, 14 talkers are utilized to target the generated character images for AGTH generation. Table 3 provides details of the selected talkers, including their respective output resolutions. It is important to note that variations in output resolution exist and the selected driving methods are totally implemented using source code provided by the original authors. Additionally, for the video-based driving methods, the input video consists of a repeated driving image with a duration set to one second and a frame rate of 25 frames per second. Ultimately, a total of 10,457 AGTHs are successfully generated for 14 prompts and their corresponding 70 speeches, culminating in the creation of the THQA-10K dataset.

### 3.3. Data Statistics and Subjective Experiment

To assess the generalization capability of various T2I models and talkers, we conduct a statistical analysis of the number of portraits and AGTHs successfully generated by these generative models, as illustrated in Fig. 4. From Fig. 4(a), it is evident that the two T2I models, SD1 and SD2, demonstrate limited effectiveness in generating portraits across a range of prompts. However, for IDG, its suboptimal performance can be attributed to the constraints posed by prompt-sensitive vocabulary. Additionally, Fig. 4(b) highlights the

Table 3. Details of talkers employed for generation.

Type	Label	Methods	Year	Head Motion	Output Resolution
Image-based	MI	MakelIttalk [64]	2020	✓	256×256
	AH	Auido2Head [43]	2021	✓	256×256
	ST	Sadtalker [49]	2023	✓	512×512
	DT	Dreamtalk [23]	2023	✓	256×256
	ET	EAT [13]	2023	✓	256×256
	EM	EchoMimic [6]	2024	✓	512×512
Video-based	WL	Wav2Lip [33]	2020	×	1,024×1,024
	VR	Video-Retalking [7]	2022	×	1,024×1,024
	SH	StyleHeat [47]	2022	×	1,024×1,024
	DN	DINet [51]	2023	×	1,024×1,024
	IL	IP-LAP [63]	2023	×	1,024×1,024
	TL	TalkLip [42]	2023	×	1,024×1,024
	MT	MuseTalk [50]	2024	×	1,024×1,024
	EG	EmoGen [46]	2024	×	1,024×1,024

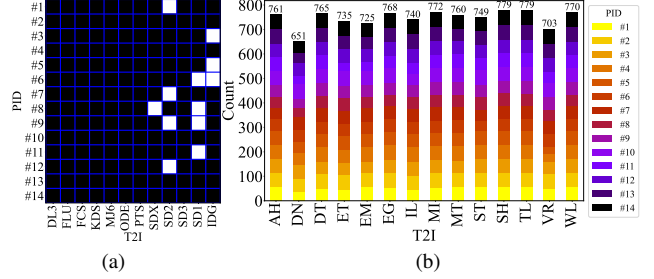


Figure 4. Visualization of the number of successful generation. (a) Portraits that can be successfully generated. The black block indicates success while the white one indicates failure. (b) Number of AGTHs that can be successfully generated.

generalization performance of different talkers. Notably, the SH and TL talkers exhibit the strongest generalization abilities. Remarkably, the disparity in the total number of AGTHs generated between the highest and lowest performing talkers reaches 128, underscoring a significant variation in generalization performance across the different talkers.

To investigate the quality of AGTHs and their distortions in detail, we recruit 13 male and 12 female participants to conduct subjective quality assessments of all AGTHs in the THQA-10K dataset. The subjective evaluation is carried out in a well-controlled laboratory in accordance with the guidelines outlined in ITU-R BT.500-13 [4]. Participants can view AGTHs on an iMac monitor with a resolution of  $4,096 \times 2,304$ . Given that the AGTHs include audio components, a wired headset is utilized to ensure low-latency and high-quality audio transmission, while also minimizing potential interference between participants due to the audio output. The AGTHs are organized into 100 phases, with each phase comprising at most 120 AGTHs. To mitigate visual fatigue and discomfort associated with prolonged viewing, all participants are required to take a 15-minute break after completing each phase. Furthermore, each participant is limited to a maximum of 6 assessment phases per day.

### 3.4. Data Processing

At the end of the subjective experiment, we receive a total of  $261,425 = 25 \times 10,457$  subjective ratings. Each rating can be described as  $\{s_{ij}, D_{ij}\}$ , where  $s_{ij}$  and  $D_{ij}$  are the quality rating and labeled distortion of the  $j$ -th AGTH by the

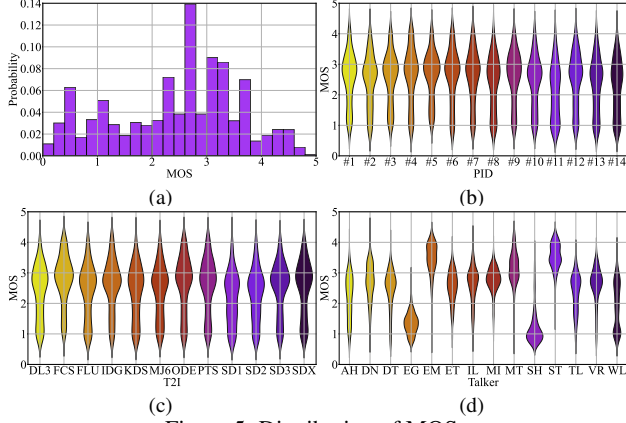


Figure 5. Distribution of MOSs.

$i$ -th subject. In particular,  $D_{ij}$  is a ten-dimensional 0-1 distortion vector, with each dimension denoting a corresponding distortion type. According to existing works [26, 53–55, 59, 62, 65, 66, 68, 70, 74–76],  $s_{ij}$  is processed as z-scores according to the following equation:

$$z_{ij} = \frac{s_{ij} - \mu_i}{\sigma_i}, \quad (1)$$

where  $\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} s_{ij}$ ,  $\sigma_i = \sqrt{\frac{1}{N_i-1} \sum_{j=1}^{N_i} (s_{ij} - \mu_i)^2}$ , and  $N_i$  represents the total number of AGTHs evaluated by subject  $i$ . Following the rejection procedure outlined in [4], ratings from unreliable subjects are excluded. The remaining z-scores  $z_{ij}$  are linearly rescaled to the range [0, 5]. Finally, the mean opinion scores (MOSs) for the  $j$ -th AGTH are computed by averaging the rescaled z-scores. For the distortion vector  $D_{ij}$ , the summation operation is employed to count the distortion type of the  $j$ -th AGTH:

$$D_j = \sum_{i=1}^{N_j} D_{ij}, \quad (2)$$

where  $N_j$  denotes the number of subjects who classified distortions of the  $j$ -th AGTH. A threshold vector  $T_j$ , defined as  $N_j/2$  across all ten dimensions, is utilized to derive the final distortions  $\mathcal{D}_j$ :

$$\mathcal{D}_j = u(D_j - T_j), \quad (3)$$

where  $u(\cdot)$  denotes the step function. The whole process can be interpreted as for each category of distortion for each AGTH, more than half of the subjects need to believe that the distortion exists before it is officially acknowledged.

### 3.5. Mean Opinion Score Analysis

Based on the results of the subjective experiments, we conduct a comprehensive analysis of mean opinion scores (MOSs) and distortion of AGTHs in order to evaluate the various types of talkers. Initially, the overall MOS distribution is plotted as shown in Fig. 5(a). To further analyze the effects of various possible factors on the MOSs,

the violin plots shown in Fig. 5(b-d) provide a more intuitive picture of the relationship between different PIDs, T2Is, talkers and MOSs. By observing Fig. 5, we can draw some valuable conclusions: 1) The majority of AGTHs received quality scores centered around 3, indicating that some talkers are capable of meeting user expectations regarding audiovisual quality. However, it is noteworthy that only a limited number of AGTHs achieved MOSs exceeding 4 points, while a significant proportion of AGTHs received low-quality scores in the range of 0 to 2 points. This suggests considerable potential for improving the audiovisual quality produced by talkers; 2) AGTHs generated from different PIDs and T2Is exhibit a similar distribution of MOSs. This observation indicates that the PIDs and T2Is utilized in constructing the THQA-10K dataset are both representative and universal, thereby facilitating generalization to other PIDs and T2Is. Additionally, it implies that variations in PIDs and T2Is are not the primary determinants affecting the quality of AGTHs; 3) There are significant disparities in the quality distribution of AGTHs produced by different talkers. Notably, two types of talkers, EM and ST, generate the highest quality AGTHs, whereas two other types, EG and SH, yield lower quality outputs.

### 3.6. Distortion Visualization & Analysis

An examination of the AGTHs in the THQA-10K dataset reveals that the 9 distortion types previously identified in the THQA dataset by Zhou *et al.* [72] remain prevalent. Additionally, new distortions have emerged, predominantly characterized by misalignment of facial keypoints. To illustrate the impact of each distortion type and quantify their occurrence, we select representative samples for visualization and cataloged the distortions identified in the subjective assessments. The results are depicted in Fig. 6. From this analysis, several key insights can be derived: 1) A total of 17,191 distortions are identified across 10,457 AGTHs, indicating that distortions are a common issue within existing AGTHs. Notably, individual AGTHs often exhibit a combination of multiple distortion types; 2) Among the various distortion types, blur (BL), noise (NI) and artifacts (AF) remain the predominant issues, suggesting that current talkers still face limitations in these areas; 3) The use of Y-T slices to represent the two distortion types, little lip motion (LLM) and muscle twitch (MT), as demonstrated by Zhou *et al.*, offers a more pronounced indication of the visual effects associated with these distortions. Specifically, LLM manifests as parallel lines or minor fluctuations in the mouth texture within the Y-T slice, while MT is characterized by periodic repetitive textures. The newly identified misaligned keypoints (MK) distortion results in a noticeable displacement of facial features, leading to severe distortion.

To further investigate the frequency of various distortion types across different talkers, we plot a heatmap to visu-

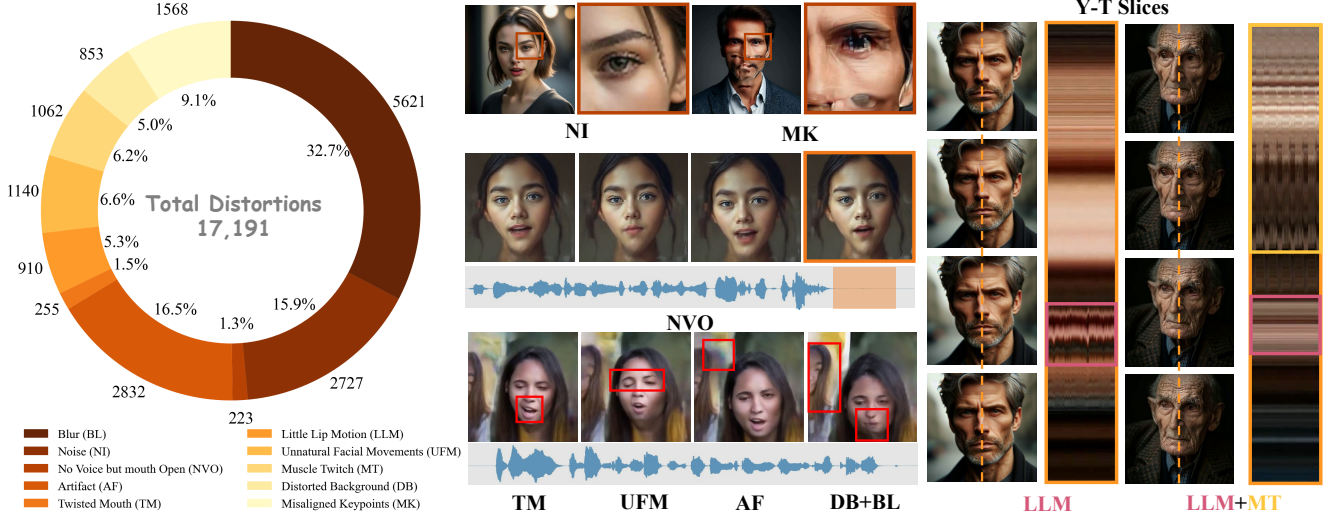


Figure 6. Visualization of distortions. The left side shows how often each distortion occurs, while the right side shows typical cases.

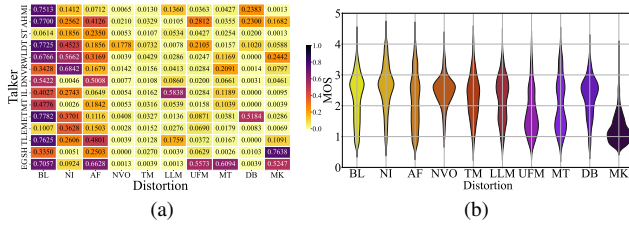


Figure 7. Statistical analysis of different distortions.

ally illustrate the relationship between talkers and distortions. The data presented in Fig. 7(a) reveal another two key findings: 1) Specific distortions tend to be concentrated among one or a few talkers. For instance, LLM predominantly occurs with the IL talker, while distortion from distorted background (DB) is mainly observed with the ET talker, and MK is frequently associated with the SH and EG talkers; 2) Among all talkers included in the comparison, ST and EM exhibit the least frequency of distortion, highlighting their superiority and explaining why these two methods achieve the highest MOSs. Finally, in conjunction with the MOS analysis method of Sec. 3.5, the effect of each type of distortion on the MOSs is plotted as Fig. 7(b). The results indicate that different distortion types have varying effects on human audiovisual perception. Notably, the three most common distortions do not significantly detract from the overall quality of AGTHs, and in some cases, they still yield high MOSs. Conversely, MK has the most pronounced negative impact on AGTH quality, primarily due to the intolerable displacement of facial features, resulting in significant distortion.

## 4. Proposed Method

### 4.1. First Frame & Slice Process

Considering that each AGTH video is generated by a stationary talker, it is reasonable to assert that the video qual-

ity remains relatively stable throughout the duration of the AGTH. Consequently, the initial frame of an AGTH provides a rich set of spatial features suitable for quality assessment. In terms of temporal feature selection, this paper proposes a slice-based temporal feature extraction scheme, as illustrated in Fig. 8. Firstly, we advocate for the use of Y-T slices over X-T slices, as Y-T slices capture more comprehensive facial information due to the inherent symmetry of the face. Additionally, to ensure that the tangent line intersects the mouth, which is a critical aspect of AGTHs, we utilize the image generated by the T2I model as a reference and apply landmark detection to identify the key points  $p_k$  of the mouth within the generated image. Subsequently, we compute the coordinates of the mouth’s centroid  $C_o$ :

$$C_o = \frac{1}{K} \sum_{k=1}^K p_k, \quad (4)$$

where  $K$  denotes the total number of key points on the face. However, this derived center point cannot be directly employed to guide the slicing of corresponding AGTHs, as the videos produced by different talkers exhibit varying resolutions  $R_v$ . Therefore, a scale transformation is applied to determine the location of the mouth centroid in the AGTH videos:

$$C_v = \frac{R_v}{R_o} C_o, \quad (5)$$

where  $R_o$  is the resolution of the generated portrait and  $C_v$  denotes the center of the human face in AGTHs. Following this, the entire AGTH video can be projected as a Y-T slice within the three-dimensional space defined by XY-T coordinates. Ultimately, the Y-T slice is resized to match the resolution of the first frame, facilitating subsequent processing. This entire slicing process underscores that the Y-T slice effectively captures temporal features over the video duration, in contrast to conventional frame extraction methods.



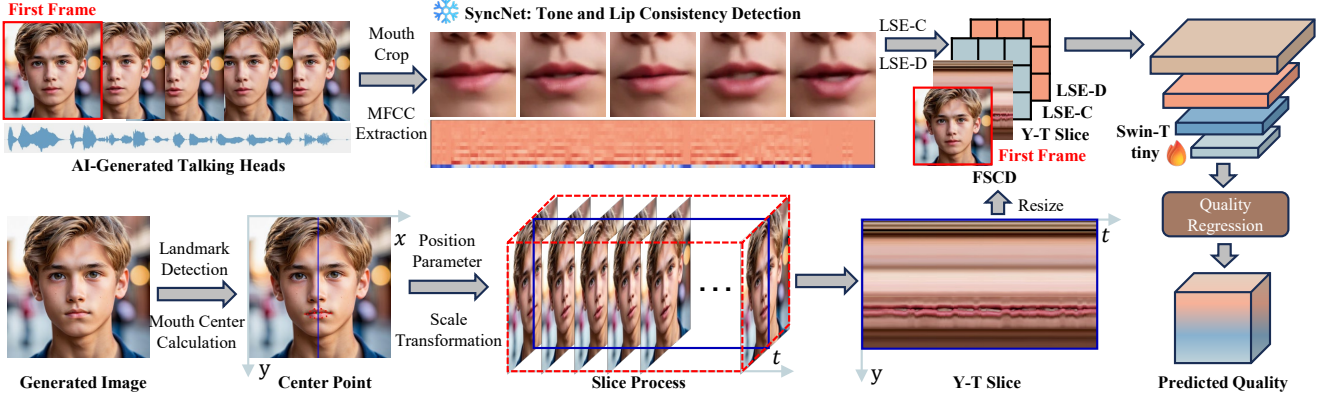


Figure 8. The proposed framework for FSCD. The method mainly consists of four compartmentalized modules: the Y-T slice process, tone-lip consistency detection, backbone feature extraction, and quality regression.

## 4.2. Tone-lip Consistency Detection

AGTH is a media that exhibits a high sensitivity to audio-visual synchronization, particularly in relation to the synchronization between lip movements and speech. To address this sensitivity, we initially perform a cropping of the mouth region in the AGTH and extracted Mel-frequency cepstral coefficients (MFCCs) from the accompanying audio. Subsequently, we employ the classical SyncNet [8] to assess audio-lip consistency, yielding two key outputs: lip sync error confidence (LSE-C) and distance (LSE-D). In contrast to existing methodologies that integrate these two features directly into the final quality regression layer, this study proposes an innovative approach whereby the two LSE features are expanded into a tensor of dimensions corresponding to those of the first frame. This tensor is then utilized as two additional images within the first frame, enabling the backbone network to learn the relative significance of tonal lip consistency autonomously, rather than relying on predetermined weights based on prior knowledge.

## 4.3. Backbone & Quality Regression

The obtained First frame, Y-T Slice, LSE-C, and LSE-D can be collectively viewed as a new image called FSCD. Given the excellent performance achieved by the swin-transformer (Swin-T) [22, 52, 57, 79] in several computer vision tasks, Swin-T is used to extract quality features from FSCDs. During the training phase, the predicted audiovisual quality is compared with the actual MOS using the Mean Squared Error (MSE) as the loss function, facilitating the gradual optimization of the algorithm for enhanced performance:

$$Loss = \frac{1}{n} \sum_{l=1}^n (\hat{Q}_l - Q_l)^2, \quad (6)$$

where  $\hat{Q}_l$  and  $Q_l$  represent the predicted quality and MOS of  $l$ th AGTH, and  $n$  indicates the size of training batch.

## 5. Experiments

### 5.1. Experiment Details & Criteria

To validate the effectiveness of the proposed method, we select 15 quality assessment algorithms applicable to AGTHs for comparison. This selection includes 4 classical image quality assessment (IQA) methods, 2 methods for audio and lip consistency, and 9 video quality assessment (VQA) methods. Among these, RAPIQUE, SimpVQA, VSFA, FAST-VQA, and BVQA are deep learning-based methods, while the remaining methods rely on manually extracted features. All selected methods are tested on THQA-10K, THQA [72] and THQA-3D [73] datasets, utilizing a five-fold cross-validation scheme. The average test results from the five folds are recorded as the performance of each method. Notably, the five-fold data partitioning ensures that there is no content overlap, and all algorithms employed are derived from the source code provided by their authors.

In terms of evaluation criteria, we adopt four commonly used metrics for assessing the performance of objective multimedia quality assessment algorithms: Spearman Rank Correlation Coefficient (SRCC), Kendall's Rank Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE).

### 5.2. Performance Analysis

The performance of the proposed FSCD method, along with other competing methods, on three datasets is presented in Table 4. An analysis of this table yields several key conclusions: 1) FSCD demonstrates optimal performance on all datasets, surpassing the next best algorithm by +2% in SRCC at least. This result strongly supports the effectiveness of the proposed FSCD method for assessing AGTH quality; 2) The methods that achieve suboptimal performance differ among three datasets. In contrast, FSCD consistently achieves optimal performance across all datasets, highlighting the robustness and generalizability of the proposed method; 3) Existing IQA methods, single audio-lip

Table 4. Performance results on the proposed THQA-10K, THQA and THQA-3D databases. Best in **RED**, second in **BLUE**.

Type	Models	THQA-10K				THQA				THQA-3D			
		SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
IQA	BRISQUE [27]	0.4271	0.4451	0.2993	1.0262	0.4856	0.5970	0.3454	0.8227	0.6749	0.7453	0.5060	0.5717
	NIQE [28]	0.0089	0.0436	0.0051	1.1492	0.0535	0.1643	0.0402	0.9811	0.2243	0.4741	0.1232	0.7707
	CPBD [30]	0.0553	0.0686	0.0371	1.1476	0.0575	0.0876	0.0376	0.9908	0.2145	0.3136	0.1432	0.8273
	IL-NIQE [48]	0.0490	0.0634	0.0286	1.1480	0.0537	0.2160	0.0276	0.9712	0.2293	0.4871	0.1537	0.7600
Sync	LSE-C [8]	0.0706	0.1634	0.0468	1.1349	0.0056	0.2109	0.0048	0.9723	0.1728	0.2297	0.1355	0.8499
	LSE-D [8]	0.0580	0.1123	0.0385	1.1431	0.1366	0.2336	0.0855	0.9671	0.0079	0.1054	0.0008	0.8684
VQA	ViDEO [29]	0.1354	0.1782	0.0901	1.1319	0.1777	0.1891	0.1354	0.9595	0.1056	0.2308	0.0721	0.8387
	TLVQM [17]	0.4377	0.4679	0.3070	1.0130	0.0254	0.0355	0.0209	1.0853	0.1887	0.3112	0.1272	0.8240
	VIDEVAL [40]	0.3869	0.4147	0.2706	1.0431	0.0317	0.0358	0.0231	1.1916	0.2252	0.3544	0.1556	0.8118
	V-BLIINDS [36]	0.4740	0.4977	0.3334	0.9941	0.4949	0.6403	0.3533	0.7976	0.5298	0.6412	0.3907	0.6674
	RAPIQUE [41]	0.3576	0.3846	0.2490	1.0579	0.1789	0.1908	0.1277	1.0162	0.3748	0.4680	0.2660	0.7643
	SimpVQA [39]	<b>0.7775</b>	<b>0.8039</b>	<b>0.5931</b>	<b>0.6832</b>	0.6800	0.7592	0.5052	0.6361	0.6321	0.7258	0.4717	0.5983
	VSFA [20]	0.7537	0.7754	0.5726	0.7343	<b>0.7601</b>	<b>0.8106</b>	<b>0.5830</b>	<b>0.5966</b>	0.7463	0.7811	0.5596	0.5726
	FAST-VQA [45]	0.7351	0.7542	0.5519	0.8026	0.6389	0.7441	0.4677	0.6983	0.7778	0.7984	0.5964	<b>0.5503</b>
	BVQA [18]	0.6335	0.7405	0.4522	0.7634	0.7287	0.7985	0.5549	0.6094	<b>0.7871</b>	<b>0.8298</b>	<b>0.6081</b>	0.5983
	<b>FSCD (Ours)</b>	<b>0.8066</b>	<b>0.8322</b>	<b>0.6228</b>	<b>0.6333</b>	<b>0.7812</b>	<b>0.8409</b>	<b>0.5951</b>	<b>0.5055</b>	<b>0.8235</b>	<b>0.8505</b>	<b>0.6463</b>	<b>0.4577</b>

Table 5. Ablation study results on databases, where ‘w/o’ stands for ‘without’. Best in **RED**, second in **BLUE**.

Dimension	THQA-10K				THQA				THQA-3D			
	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
w/o F	0.6510	0.6892	0.4726	0.8497	0.6830	0.7506	0.5035	0.5765	0.7368	0.7844	0.5451	0.5468
w/o S	0.7330	0.7672	0.5509	0.7329	0.7205	0.7968	0.5325	0.5272	0.7506	0.7730	0.5655	0.5593
w/o C	<b>0.7927</b>	<b>0.8169</b>	<b>0.6049</b>	<b>0.6764</b>	<b>0.7660</b>	<b>0.8174</b>	<b>0.5860</b>	<b>0.5086</b>	<b>0.8019</b>	<b>0.8429</b>	<b>0.6160</b>	<b>0.4658</b>
w/o D	0.7610	0.7879	0.5744	0.7221	0.7462	0.8110	0.5579	0.5105	0.7915	0.8359	0.6037	0.4710
<b>FSCD</b>	<b>0.8066</b>	<b>0.8322</b>	<b>0.6228</b>	<b>0.6333</b>	<b>0.7812</b>	<b>0.8409</b>	<b>0.5951</b>	<b>0.5055</b>	<b>0.8235</b>	<b>0.8505</b>	<b>0.6463</b>	<b>0.4577</b>

consistency detection techniques, and VQA methods are constrained in their ability to evaluate AGTH quality. This limitation primarily arises from the inability of these methods to fully leverage the spatio-temporal features and multimodal information inherent in AGTHs. Moreover, compared to the extraction of temporal features through the Y-T slice, conventional approaches often struggle to adequately account for the relationships among frames.

### 5.3. Ablation Experiments

To further evaluate the effectiveness of each component within the FSCD framework, ablation experiments are conducted, with results summarized in Table 5, from which following observations can be drawn: 1) Each component of FSCD contributes positively to the overall performance. This enhancement can be attributed to the fact that the four components address different quality aspects, allowing for a synergistic effect; 2) Regarding the importance of individual components, the quality characteristics provided by the first frame rank highest, followed by the Y-T Slice, LSE-D, and LSE-C. This suggests that spatial features are paramount in influencing the quality of AGTHs, with temporal features and coherence features following in importance; 3) A comparative analysis of the performance results presented in Tables 4 and 5 reveals a noteworthy finding: even when utilizing only the Y-T Slice and the tone and lip coherence features, without the first frame, competitive performance is still achievable on three datasets. This highlights the validity and significance of the Y-T slice as em-

ployed by FSCD in the assessment of AGTH quality.

## 6. Conclusion

As digital human technology continues to advance, various speech-driven methods, commonly referred as “Talkers,” have emerged to enhance the efficiency of digital human face design. To thoroughly investigate the generalization performance and generation quality of different talkers, this study introduces the THQA-10K dataset, comprising a total of 10,457 AI-Generated Talking Head (AGTH) videos. Specifically, 12 Text-to-Image (T2I) methods are employed to generate character portraits, while 14 advanced talkers are utilized to produce AGTHs. Through comprehensive data analysis and subjective experiments conducted on the THQA-10K dataset, we validate both the comprehensiveness and generalizability of the dataset. Additionally, we assess the generalization performance of existing talkers and identify potential quality issues and their distributions. Finally, we propose an objective quality assessment method named FSCD, leveraging the first frame, Y-T slice, and tone-lip consistency. Experimental results substantiate the effectiveness and robustness of FSCD, which is anticipated to inform the ongoing development of talkers.

**Acknowledgements.** This work was supported in part by the Major Key Project of PCL (PCL2023A10-2), National Natural Science Foundation of China (623B2073, 62101326, 62225112, 62271312, 62132006) and STCSM (22DZ2229005).



## References

- [1] Ideagram 2.0. <https://ideogram.ai/>, 2024. 1, 4
- [2] Shakhmatov Arseniy, Razzhigaev Anton, Nikolich Aleksandr, Arkhipkin Vladimir, Pavlov Igor, Kuznetsov Andrey, and Dimitrov Denis. kandinsky 2.1, 2023. 4
- [3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023. 1, 4
- [4] RECOMMENDATION ITU-R BT. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*, 2002. 4, 5
- [5] 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. 2
- [6] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 1, 2, 4
- [7] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *ACM Special Interest Group on Computer Graphics Asia 2022*, pages 1–9, 2022. 1, 2, 4
- [8] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision 2016 Workshops*, pages 251–263, 2017. 3, 7, 8
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 4
- [10] FLUX.1-dev. <https://huggingface.co/black-forest-labs/FLUX.1-dev>, 2024. 4
- [11] Fooocus. <https://github.com/lillyasviel/Fooocus>, 2023. 1, 4
- [12] Yun Fu, Renxiang Li, Thomas S Huang, and Mike Danielsen. Real-time multimodal human–avatar interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(4):467–477, 2008. 2
- [13] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645, 2023. 2, 4
- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3
- [15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3
- [16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 3
- [17] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 8
- [18] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5944–5958, 2022. 8
- [19] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. Aigiq-20k: A large database for ai-generated image quality assessment. *arXiv preprint arXiv:2404.03407*, 2(3):5, 2024. 3
- [20] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM International Conference on Multimedia*, pages 2351–2359, 2019. 8
- [21] Xiaohong Liu, Xiongkuo Min, Qiang Hu, Xiaoyun Zhang, Jie Guo, Guangtao Zhai, Shushi Wang, Yingjie Zhou, Lu Liu, Jingxin Li, et al. Ntire 2025 xgc quality assessment challenge: Methods and results. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1389–1402, 2025. 2
- [22] 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 *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 7
- [23] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023. 1, 2, 4
- [24] Yong-You Ma, Hui Zhang, and Shou-Wei Jiang. Realistic modeling and animation of human body based on scanned data. *Journal of Computer Science and Technology*, 19(4): 529–537, 2004. 2
- [25] Midjourney. <https://www.midjourney.com/home>, 2023. 1, 4
- [26] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *Science China Information Sciences*, 67(11):211301, 2024. 5
- [27] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 8
- [28] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 8
- [29] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015. 8
- [30] Niranjana D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur

- detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011. 3, 8
- [31] OpenDalleV1.1. <https://huggingface.co/dataautogpt3/OpenDalleV1.1>, 2023. 1, 4
- [32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 4
- [33] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Nambodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, pages 484–492, 2020. 1, 2, 4
- [34] ProteusV0.2. <https://huggingface.co/dataautogpt3/ProteusV0.2>, 2024. 1, 4
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 4
- [36] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014. 8
- [37] Oliver Schreer, Roman Englert, Peter Eisert, and Ralf Tanger. Real-time vision and speech driven avatars for multimedia applications. *IEEE Transactions on Multimedia*, 10(3):352–360, 2008. 2
- [38] Yanhu Shan, Shiquan Wang, Zhang Zhang, and Kaiqi Huang. An xt slice based method for action recognition. In *2011 IEEE International Conference on Computer Vision Workshops*, pages 1897–1903, 2011. 2
- [39] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *ACM International Conference on Multimedia*, 2022. 8
- [40] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021. 8
- [41] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. 8
- [42] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023. 1, 2, 4
- [43] 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. 1, 2, 4
- [44] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3
- [45] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fastvqa: Efficient end-to-end video quality assessment with fragment sampling. In *European Conference on Computer Vision*, pages 538–554. Springer, 2022. 8
- [46] Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6368, 2024. 1, 2, 4
- [47] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European Conference on Computer Vision*, pages 85–101, 2022. 1, 2, 4
- [48] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 8
- [49] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 1, 2, 4
- [50] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high quality lip synchronization with latent space inpainting. *arXiv preprint arXiv:2410.10122*, 2024. 2, 4
- [51] 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. 1, 2, 4
- [52] Zicheng Zhang, Wei Sun, Yingjie Zhou, Wei Lu, Yucheng Zhu, Xiongkuo Min, and Guangtao Zhai. Eep-3dqa: Efficient and effective projection-based 3d model quality assessment. In *2023 IEEE international conference on Multimedia and expo (ICME)*, pages 2483–2488. IEEE, 2023. 7
- [53] 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, 5
- [54] 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 *IEEE International Conference on Multimedia and Expo*, pages 2519–2524, 2023. 2
- [55] Zicheng Zhang, Yingjie Zhou, Wei Sun, Xiongkuo Min, Yuzhe Wu, and Guangtao Zhai. Perceptual quality assessment for digital human heads. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023. 2, 5
- [56] Zicheng Zhang, Yingjie Zhou, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Geometry-aware video quality assessment for dynamic digital human. In *IEEE International Conference on Image Processing*, pages 1365–1369, 2023. 2

- [57] Zicheng Zhang, Yingjie Zhou, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Simple baselines for projection-based full-reference and no-reference point cloud quality assessment. *arXiv preprint arXiv:2310.17147*, 2023. 7
- [58] Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are llms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*, 2024. 3
- [59] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 5
- [60] 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*, 2024. 2
- [61] Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, Wenzhe Li, Lu Sun, Yingjie Zhou, Jianbo Zhang, Bowen Yan, Ziheng Jia, Jiahao Xiao, Yuan Tian, Xiangyang Zhu, Kaiwei Zhang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Qi Jia, and Guangtao Zhai. Aibench: Towards trustworthy evaluation under the 45° law. <https://aiben.ch/>, 2025. 3
- [62] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Wei Sun, Xiongkuo Min, Xiaohong Liu, and Guangtao Zhai. Mm-pcqa+: Advancing multi-modal learning for point cloud quality assessment. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(4):1–22, 2025. 5
- [63] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 1, 2, 4
- [64] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeltalk: speaker-aware talking-head animation. *ACM Transactions On Graphics*, 39(6):1–15, 2020. 1, 4
- [65] Yu Zhou, Yanjing Sun, Leida Li, Ke Gu, and Yuming Fang. Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1767–1777, 2021. 5
- [66] 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, 2022. 5
- [67] 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. 1
- [68] Yu Zhou, Weikang Gong, Yanjing Sun, Leida Li, Ke Gu, and Jinjian Wu. Quality assessment for stitched panoramic images via patch registration and bidimensional feature aggregation. *IEEE Transactions on Multimedia*, 2023. 5
- [69] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiongkuo Min, Xianghe Ma, and Guangtao Zhai. A no-reference quality assessment method for digital human head. In *IEEE International Conference on Image Processing*, pages 36–40, 2023. 2
- [70] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Perceptual quality assessment for point clouds: A survey. *ZTE Communications*, 21(4):3, 2023. 5
- [71] 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. 2
- [72] 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 *IEEE International Conference on Image Processing*, pages 15–21, 2024. 2, 3, 5, 7
- [73] 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 *ACM International Conference on Multimedia*, pages 6033–6042, 2024. 2, 7
- [74] 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. *arXiv preprint arXiv:2409.07236*, 2024. 5
- [75] 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 *IEEE Visual Communications and Image Processing*, 2024. 2
- [76] Yingjie Zhou, Jiezhong Cao, Zicheng Zhang, Farong Wen, Yanwei Jiang, Jun Jia, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Q-agent: Quality-driven chain-of-thought image restoration agent through robust multimodal large language model. *arXiv preprint arXiv:2504.07148*, 2025. 5
- [77] Yingjie Zhou, Jing Wan, Sitong Liu, Yinghan Xia, Zhixiang Lu, Farong Wen, Zicheng Zhang, Yu Wang, Yu Zhou, Xiaohong Liu, Xiongkuo Min, Jiezhong Cao, and Guangtao Zhai. Cdhqa: A quality assessment database for conversational digital human. In *International Conference on Image and Graphics*, 2025. 2
- [78] Yingjie Zhou, Zicheng Zhang, Jun Jia, Yanwei Jiang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Who is a better imitator: Subjective and objective quality assessment of animated humans. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2
- [79] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Ct-pcqa: A convolutional neural network and transformer combined method for point cloud quality assessment. *Signal Processing: Image Communication*, page 117371, 2025. 7