This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3572000

Who is a Better Imitator: Subjective and Objective Quality Assessment of Animated Humans

Yingjie Zhou, Zicheng Zhang, Jun Jia, Yanwei Jiang, Xiaohong Liu, Xiongkuo Min, *Member, IEEE*, Guangtao Zhai, *Fellow, IEEE*

***Abstract*—Animated human (AH) have gained popularity due to their vivid appearance and smooth, natural movements. Various animation methods based on artificial intelligence (AI) have been introduced, which are viewed as "Imitators," offering new solutions for designing AHs. However, the effectiveness of these AI-generated AHs varies significantly across different categories and within the same category, leading to visual distortions that adversely affect the viewer's experience. Consequently, it is essential to evaluate the quality of AHs to provide reliable and objective indicators for their further development and to ensure the delivery of higher-quality AH videos to users. In this paper, the first Animated Human Quality Assessment (AHQA) dataset is constructed by selecting 6 advanced and popular imitators and 10 common actions to animate 20 AI-generated characters. The constructed dataset integrates different genders and age groups of character images, and two types of poses, standing and sitting, are selected, highlighting the comprehensiveness and diversity of the AHQA dataset. Subjective experiments reveal significant differences in the quality of AHs produced by different imitators. Finally, we propose a quality assessment method, VIP-QA, incorporating Video quality, Identity consistency, and Posture similarity for the AHQA dataset. Experimental results show that VIP-QA significantly outperforms existing assessment methods on multiple datasets by about 5%, more closely approximates human visual perception, and provides a valid objective metric for assessing imitators. All the work in this paper has been released at https://github.com/zyj-2000/Imitator.**

***Index Terms*—Quality assessment database, video quality assessment, digital human, computer animation, AIGC.**

I. Introduction

THE emergence of the metaverse concept has positioned virtual digital human (DH) design as a cutting-edge and rapidly evolving technology within digital media [1], [2]. Broadly defined, DH technology encompasses any media content that replicates human appearance and behaviors. This technological capability has enabled digital humans to be utilized across a diverse range of fields, including education, medicine, and industry. Despite its potential, as shown in the Fig. 1, DH design remains a complex and time-intensive process. This complexity arises from the dual challenge of creating a DH image that meets audience expectations and the even more demanding task of generating realistic animations. Presently, many high-quality DH animations are still crafted by designers using specialized animation software. However, advancements in artificial intelligence (AI) have introduced new possibilities in the field of animated human (AH) design. AI can not only generate virtual DH images within seconds but also replicate the movements of characters from input videos to create continuous AH videos [3]–[5]. Although several methods have been proposed to achieve these "imitator" animations, the results often fall short of expectations. This can be attributed to the nascent stage of research in this field and underdevelopment of quality assessment methods.

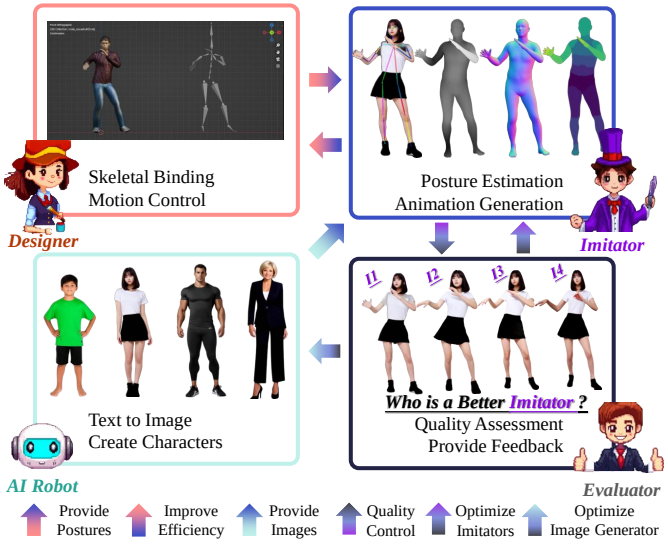


Fig. 1. Current state of animated human generation and the importance of quality assessment.

To address the gap in quality assessment metrics within this field and to support its ongoing development, this paper undertakes a comprehensive evaluation of the subjective and objective quality of AHs. Specifically, this research introduces the first Animated Human Quality Assessment (AHQA) dataset. The AHQA dataset is created by collecting 20 full-body portraits, which are then animated to mimic 10 common actions by 6 imitators. Consequently, the dataset comprises a total of 1,200 AH videos. To assess the subjective quality of these animations, multiple participants are invited to provide ratings. Analysis of the participants' feedback reveals significant quality differences across the various imitators, underscoring the critical importance of quality assessment in AH development. Additionally, this paper proposes a novel

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).

Corresponding author: Guangtao Zhai (zhaiguangtao@sjtu.edu.cn).

Yingjie Zhou, Zicheng Zhang, Jun Jia, Yanwei Jiang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China and PengCheng Laboratory, Shenzhen, China. E-mail:{zyj2000, zzc1998, jiajun0302, jiang-yan-wei, xiaohongliu, minxiongkuo, zhaiguangtao} @sjtu.edu.cn.

TABLE I
COMPARISON WITH EXISTING DIGITAL HUMAN QUALITY ASSESSMENT DATABASES.

Database	Content Form	Scale	Distortions	Description
DHH-QA [6]	3D Mesh	1,540	Noise, JPEG, Downsampling, Quantization	Scanned Real Human Heads
DDH-QA [7]	Mesh Sequence	800	Model and Motion Distortions	Dynamic 3D Digital Humans
SJTU-H3D [8]	3D Mesh	1,120	Noise, JPEG, Downsampling, Quantization	Static 3D Digital Humans
6G-DHQA [9]	Mesh Sequence	400	JPEG, Downsampling, Quantization, Stall, Rebuffer	Digital Twins
THQA [10]	2D Video	800	Image Quality, Lip-sound Consistency, Overall Naturalness	Speech-driven Talking Heads
THQA-3D [11]	Mesh Sequence	1,000	Quantization, Stall, Rebuffer, Conversion, Synchronization	Scanned Real Talking Heads
ReLI-QA [12]	2D Image	840	AI-Generated Distortions	Relighted Human Heads
MEMO-Bench [13]	2D Image	7,145	Sentimental Error	Emotional Human Heads
AHQA (Proposed)	**2D Video**	**1,200**	**Character, Video, Motion Distortions and Hallucination**	**Animated Human Videos**

evaluation algorithm, VIP-QA, which integrates Video quality, Identity consistency, and Pose similarity. Specifically, video quality is used to perceive the blurring and jittering of the AH videos, and identity consistency is used to measure character fidelity. To address the challenge of measuring the motion accuracy of AH videos at different frame rates and resolutions, the Motion-p2point method is proposed. Experimental results demonstrate that VIP-QA outperforms existing image and video quality assessment methods in terms of assessment accuracy, consistency, robustness, and generalization on both AHQA and DDH-QA [7] datasets. Moreover, VIP-QA proves effective for evaluating the quality of both 2D generative AHs and 3D manually-modeled animated digital humans, providing a robust objective quality metric for AHs. Overall, the main novelties and contributions of this paper are as follows:

- The **first** large-scale dataset targeting quality assessment of animated human (AH) is created. It contains 1,200 AH videos generated from 20 character images and 10 common actions through 6 imitators.
- A posture similarity metric Motion-p2point combining pose estimation and affine transformation is proposed to efficiently measure the imitation accuracy of AH videos.
- VIP-QA, which integrates video quality, identity consistency, and pose alignment, is proposed for quality assessment of AHs and achieves about **5%** performance lead and the best generalization performance.

II. RELATED WORK

A. Imitators: Human Animation Methods

Human animation is a significant branch within the fields of computer graphics and computer animation, focusing on the creation of lifelike and expressive virtual characters and their movements [14]–[16]. In the early stages, character animations were primarily created using keyframe-based design. However, this interpolation approach often fails to produce natural-looking movements for certain actions. As technology advanced, researchers introduced forward kinematics (FK) and inverse kinematics (IK) controllers, which conceptualize character animation as the movement of the character's skeletal structure. In recent years, the emergence of diffusion modeling and video generation has reignited interest in the intelligent design of character-specific animation, leading to the development of several notable approaches for the automated generation of character animations. Xu $et\ al.$ introduced MagicAnimate [17], a diffusion-based framework designed to improve temporal consistency and fidelity in human image animation. Zhu $et\ al.$ proposed Champ [18], a methodology that employs a 3D human parametric model to achieve shape alignment and motion guidance in human image animation. Wang $et\ al.$ developed VividPose [19], an end-to-end pipeline leveraging Stable Video Diffusion (SVD) [20] to address issues of appearance degradation and temporal inconsistencies. Additionally, Wang $et\ al.$ presented UniAnimate [21], a framework for the efficient generation of long-term human videos without the need for an external reference model. Hu introduced Animate Anyone [22], which integrates SVD with Pose Guider modules to enhance control over character movement. Zhang $et\ al.$ proposed MimicMotion [23], which incorporates confidence-aware pose guidance to increase the accuracy of pose information. While these studies collectively demonstrate significant advancements in human image animation techniques, they still lack robust metrics to reliably evaluate the quality and effectiveness of the generated animations.

B. Digital Human Quality Assessment

Digital human quality assessment (DHQA) has become a critical field alongside the growing prominence and popularity of DHs [24]. This field not only evaluates the quality of existing DHs but also plays a crucial role in guiding and optimizing the design of DH communication systems, ultimately enhancing the user's experience. Consequently, a range of research efforts has been dedicated to this area. As outlined in Table I, various DHQA datasets have been developed. The establishment of these datasets has spurred further scholarly activity, leading to the development of quality assessment methods that advance the field. For instance, Zhang $et\ al.$ introduced a twin network architecture for full-reference (FR) digital face quality assessment [6], providing an effective benchmark for algorithms related to 3D DH head compression and sampling. Additionally, they addressed the quality assessment of dynamic DHs by leveraging geometric features, such as curvature and the distribution of dihedral angles, in their subsequent work [25], [26]. Recently, they pioneered a reduced-reference (RR) algorithm in DHQA [27], combining mesh and mapping features to offer a novel solution. Zhou $et\ al.$ employed a multi-task learning approach to enhance the performance of DHQA methods by incorporating a subtask focused on 3D distortion classification [28]. They also

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3572000

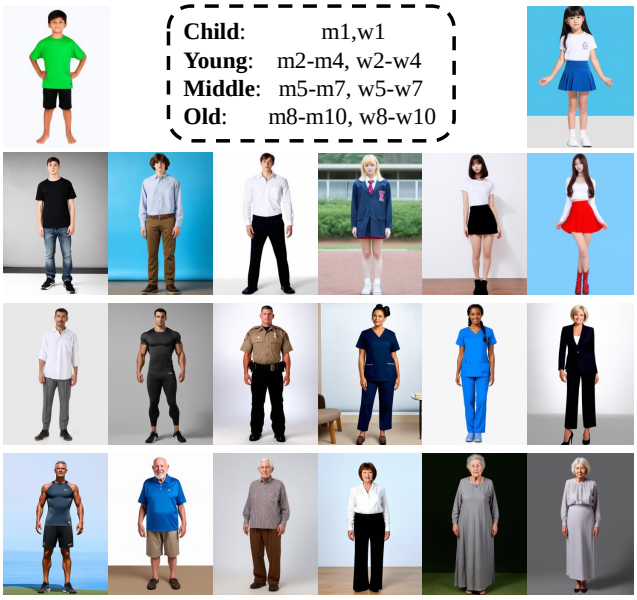


Fig. 2. All selected full-body portraits.

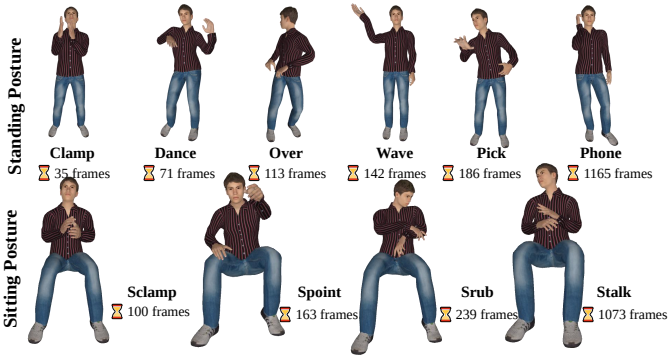


Fig. 3. Illustration of the prepared action material.

TABLE II
DETAILS OF THE IMITATORS EMPLOYED.

Label	Methods	Posture Estimation	Frame Rate	Output Resolution
UNI	UniAnimate	DWPose	8fps	512×768
MIM	Mimic Motion	DWPose	15fps	576×1024
MUS	Musepose	DWPose	15fps	512×512
MAG	MagicAnimate	Densepose	25fps	512×512
ANI	Animate Anyone	DWPose	30fps	512×784
CHA	Champ	DWPose	24fps	768×1024

concentrated on the quality assessment of talking heads, identifying various distortions inherent in current speech-driven talking head algorithms through both subjective and objective experiments [10]. However, there has been limited attention to the quality assessment of AHs generated by AI. To address this gap, this paper proposes the establishment of the AHQA dataset and conducts an in-depth exploration of effective and feasible subjective and objective quality assessment methods based on this dataset.

III. Database Construction

A. Material Preparation

To ensure the representativeness of the proposed AHQA dataset, the selection process for the original character portraits considered a diverse range of ages and genders. To mitigate potential copyright issues related to character portraits, Stable Diffusion 3 [29] is employed to generate full-body images of the characters. After a careful selection process, 20 full-body portraits that meet the necessary criteria are chosen to construct the AHQA dataset. As illustrated in the Fig. 2, the selected images include 10 male and 10 female characters, encompassing a broad spectrum of attributes such as age, posture, and clothing. It is important to note that in modern animation design and film production, cutout and segmentation are commonly used in post-production video editing in order to place the character against any background. It can be seen that the background of the selected image is not an important factor in practical applications compared to the character image, so we just keep the original background of the generated image for simplification. For the selection and design of motions, 10 common actions are developed. Specifically, skeletal animations are created using Maya for the male character provided in the DDH-QA dataset. These dynamic 3D models are then rendered into 2D animations using the Arnold renderer, serving as the action material. As depicted in Fig. 3, the rendered actions include six standing and four sitting motions. To facilitate the subsequent evaluation of the mimicry capabilities of various imitators over short, moderate, and long durations, the selected actions varied in duration, ranging from 1 to 45 seconds.

B. Animated Human Generation

As shown in Table II , six state-of-the-art (SOTA) imitators are selected for evaluation: UniAnimate [21], MimicMotion [23], MusePose [30] , MagicAnimate [17], Animate Anyone [22], and Champ [18]. Among them, pose estimation employed by imitators contains two common methods DWPose [31] and Densepose [32]. Each imitator is used to mimic 10 prepared motions for each of the 20 AI-generated full-body characters, resulting in a total of 1,200 AH videos. Notably, these imitators are implemented using the original source code provided by the authors, with consistent parameter settings maintained across the board. The exception is the Animate Anyone method, which utilizes a replication of Moore's thread code. Different imitators contribute to differences in frame rate and output resolution.

C. Distortion Effects

To provide a more intuitive visualization of the distortion effects in AH videos, representative distorted frames are selected, as illustrated in Fig. 4. Additionally, to exemplify distortion effects in AH videos temporally, an X-T Slice analysis [33] is employed to characterize variations in the legs, as shown in Fig. 6. Specifically, the X-T slice analysis [33] employed is a two-dimensional spatio-temporal representation of videos, formed by fixing horizontal rows of pixels (X-axis) and stacking their values over time (T-axis). Upon examining the distortion effects in Figs. 4 and 6, the prevalent distortion issues can be classified into four primary

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3572000

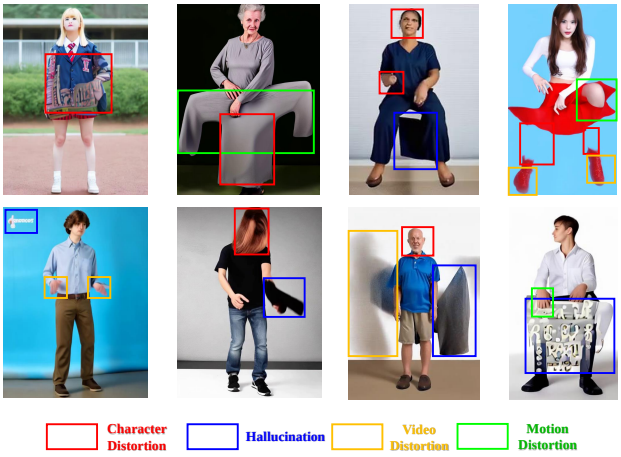


Fig. 4. Visualization of typical distortion frames.

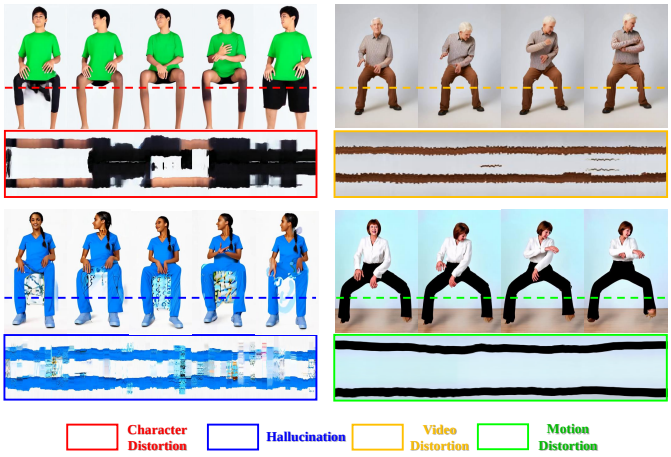


Fig. 6. Visualization of different distortions on X-T slice.

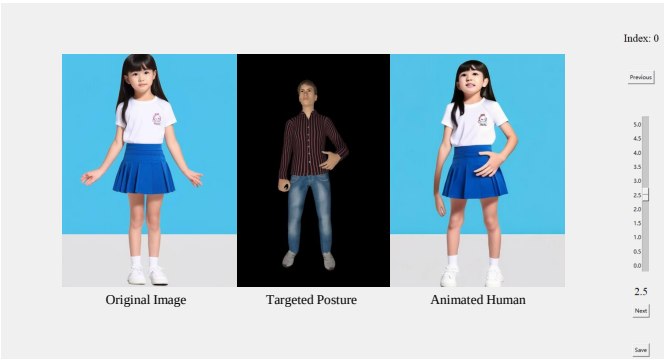


Fig. 5. Software interface used for subjective scoring.

categories: character distortion, illusion, motion distortion, and video distortion. 1) Character distortion includes distortions affecting the character's face, hands, and legs, highlighting the limitations of current mimicry algorithms in preserving the fidelity of character images. 2) Illusion involves the arbitrary introduction of elements that do not originally exist in the character photos, such as sofas or video watermarks. 3) As identified in the research [7], inappropriate molding and range of motion are still evident in this dataset. 4) Video distortion includes typical issues such as blurriness and jittering in the AHQA dataset. These observations indicate that there are still significant quality issues related to the effectiveness of current mimicry algorithms in AH generation. Consequently, it is essential to conduct quality assessments of AH at this stage to better align mimicry algorithms with human intention and visual perception.

D. Subjective Experiment

To accurately and reliably assess the quality of AH videos in the AHQA dataset, a well-controlled in-laboratory subjective quality assessment experiment is conducted following the guidelines of ITU-R BT.500-13 [34]. The experiment involves 20 male and 20 female participants with an age distribution from 18 to 65 years. In addition, all invited subjects have extensive experience in watching short-form videos and thus have rich familiarity with the AH videos to be evaluated. A dual-stimulus protocol based on action cues is designed for the experiment. This approach differs from traditional dual-stimulus subjective scoring experiments by presenting subjects with both the original action-driven video and the corresponding AH video, thereby making the action information more explicit. All video content is shown on an iMac display with a resolution of 4,096 × 2,304 at the frame rate shown in Table II, supporting repeated playback. Participants are asked to provide a final quality rating for each AH video, considering factors such as character appearance, action, and overall video quality. To mitigate the effects of visual fatigue and other factors that could impact the reliability of the ratings due to the experiment's duration, the entire subjective assessment is divided into six phases, each containing 200 AHs. Participants are required to take at least a 30-minute break after completing each phase before proceeding to the next. Each participant is allowed to complete a maximum of two phases per day, with a 30-minute break scheduled before beginning the first phase on each day.

All participants undergo a 30-minute training session before commencing the first day of the subjective experiment. This session includes familiarization with the subjective scoring guidelines and the scoring interface. Specifically, subjective assessments of the AHQA dataset are conducted following the Absolute Category Rating (ACR) protocol. The software interface used for the evaluations is depicted in Fig. 5. During the experiment, participants are simultaneously presented with the original image, the target pose, and the AH video. They are instructed to rate the quality of each AH video on a scale from 0 to 5, with increments of 0.1. The rating scale is defined as follows:

- 0 to 1 points: The visual quality of the AH video is extremely poor, with significant detachment from the original image or posture.
- 1 to 2 points: The video quality of the AH is poor, with noticeable discrepancies in the AH video compared to the original image or posture.
- 2 to 3 points: The video quality of the AH is average, with the AH video generally capturing the action but with some limitations.

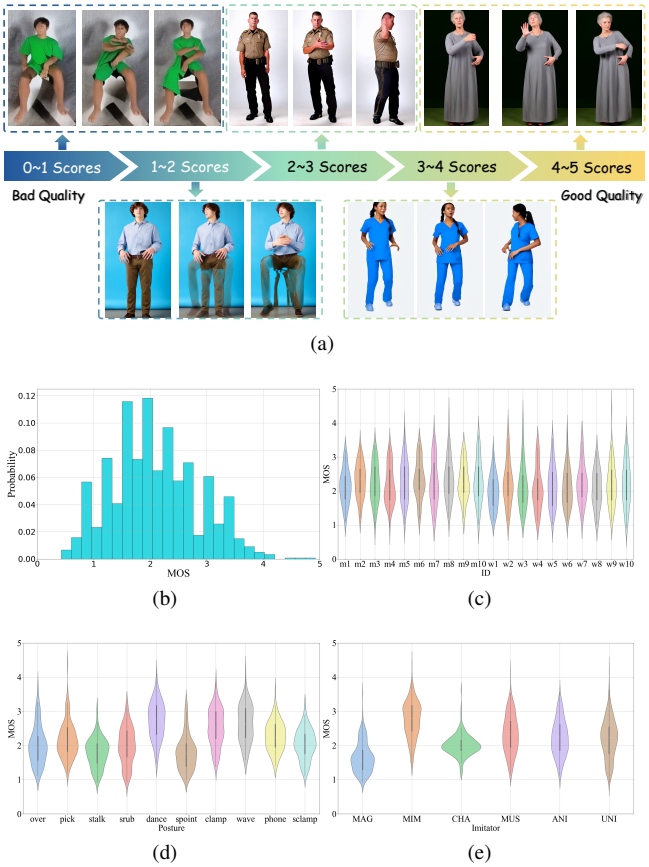


Fig. 7. Distributions of the MOSs. (a) AH samples of different MOSs. (b) MOS Distribution of the AHQA dataset. (c) Distribution of MOS for different portraits. (d) Distribution of MOS for different actions. (e) Distribution of MOS for different Imitators.

- 3 to 4 points: The video quality of the AH is good, and the AH video effectively replicates the action.
- 4 to 5 points: The AH video quality is excellent, characterized by vividness and a high degree of authenticity.

This rating scale facilitates a detailed and structured evaluation of the AH videos' quality. After days, all participants submitted their ratings, resulting in a total of 48,000 subjective ratings. Based on previous work [8], [11], [35], z-scores are computed from the collected subjective ratings. This process can be represented as follows:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \tag{1}$$

where r_{ij} represents the quality rating given by the i-th subject for the j-th AH video, $\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} r_{ij}$ is the mean rating for i-th subject, $\sigma_i = \sqrt{\frac{1}{N_i - 1}\sum_{j=1}^{N_i}(r_{ij} - \mu_i)}$ is the standard deviation of the subject, and N_i denotes the number of AHs evaluated by i-th subject. Quality ratings from unreliable subjects are excluded following the subject rejection procedure recommended in [34]. The obtained z-scores are then linearly rescaled to the range $[0, 5]$. Finally, the Mean Opinion Score (MOS) for the j-th AH is calculated by averaging the rescaled z-scores, which is employed as the j-th AH's quality score.

E. Experimental Analysis

After obtaining the MOS for the AHQA dataset, further analyses are conducted to assess the quality of AH videos from a subjective perspective. First of all, in order to visualize the visual effects corresponding to different MOSs, we select representative AHs for visualization as Fig. 7 (a). Besides, the overall distribution of MOSs for the AHQA dataset is visualized in Fig. 7 (b). As shown in Fig. 7 (b), the majority of AH videos fall within the lower-middle quality range, highlighting the current limitations of existing imitators in animation generation. This finding underscores the critical need for robust quality assessment in AH development. Subsequently, Figs. 7 (c-e) are plotted to examine quality variations across different character images, actions, and imitators. The analysis reveals that different characters exhibit similar quality distributions, suggesting that the AHQA dataset is versatile and applicable across various character types. In contrast, action type significantly influences the quality of AH videos. Specifically, AHs involving sitting poses generally exhibit lower quality compared to standing poses, indicating that current imitators struggle with the complexities of sitting movements. Another action with poor quality outcomes is the "Over" gesture, which involves a substantial angular movement of the character's head and appears to be inadequately addressed by existing imitators. Lastly, there are notable differences in the quality of AH videos produced by different imitators. Within the AHQA dataset, the MIM algorithm achieves the best performance, while MAG performs poorly.

IV. Proposed Method

Designing objective quality assessment methods for AH presents significant challenges. One major difficulty arises from the absence of corresponding ground truth for the generated AHs, which prevents algorithms from referencing high-quality examples during the learning process. Additionally, the AH videos generated by various imitators differ in frame rate, resolution, and character size, further complicating the quality assessment process. To address these challenges, we propose the VIP-QA algorithm, as illustrated in Fig. 8, which is tailored for AH quality assessment by considering three key factors: video quality, identity consistency, and posture similarity.

A. Video Quality

Since AH is delivered to users through video, the quality of the video is a critical component of overall AH quality. To comprehensively assess the quality of AH videos, it is essential to consider both spatial and temporal features, as suggested by [36]–[41]. Spatial features are particularly useful for detecting common distortions such as blurriness and artifacts. Given the hierarchical nature of visual perception, we employ multi-scale features extracted from 2D convolutional neural network (CNN) to capture quality perception information across various levels, from low to high. Keyframes for spatial feature extraction are sampled from the AH video at a rate of one frame per second:

$$SF_i = \bigoplus_{i=1}^{L} f_s^i = \bigoplus_{i=1}^{L} f_{avg}^i \oplus \bigoplus_{i=1}^{L} f_{std}^i, \tag{2}$$

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3572000

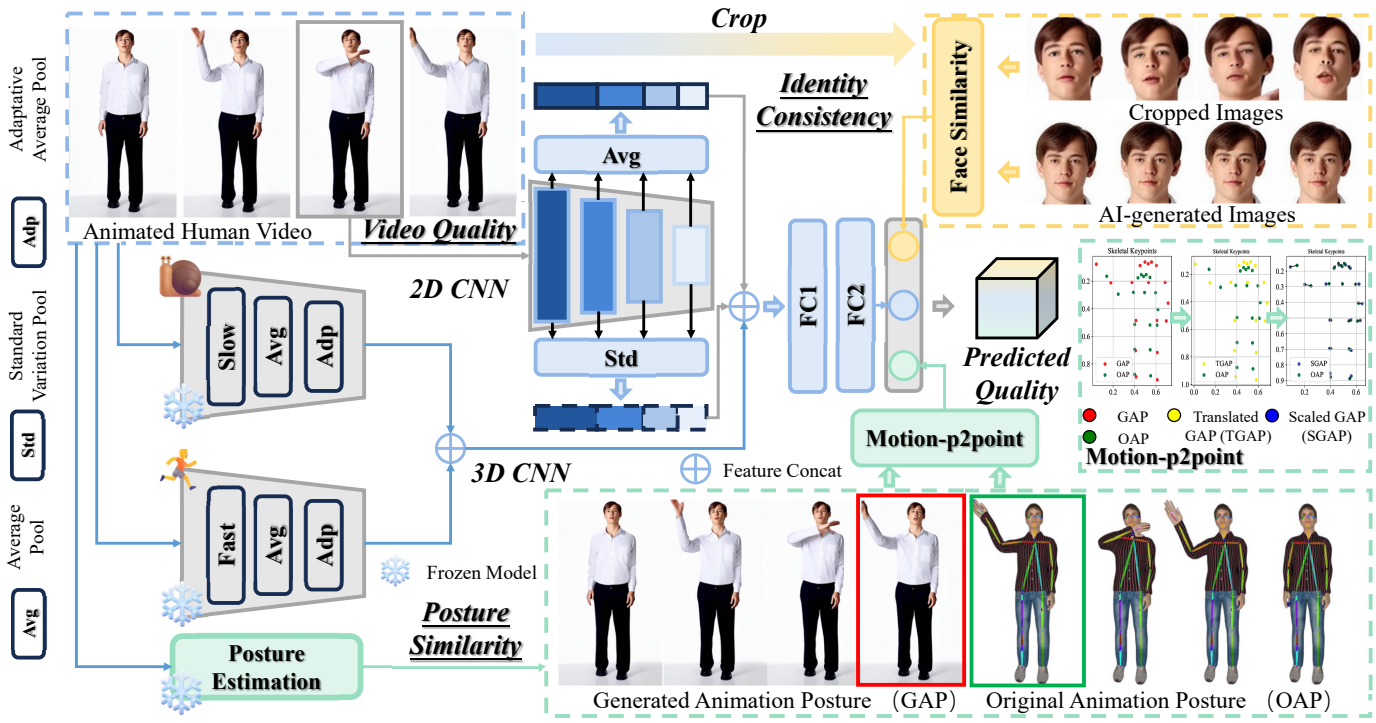


Fig. 8. Framework of proposed VIP-QA.

where $\oplus$ represents the feature concatenation operation, and L denotes the total number of sampled frames for a given AH video. Consequently, the spatial feature SF_i of the i-th keyframe K_i is obtained by concatenating the output features from each layer of the 2D CNN. Specifically, the output of each layer undergoes two operations: average pooling, which yields the average feature f_{avg}, and standard deviation pooling, which produces the standard deviation feature f_{std}. To extract temporal features, we employ a pretrained 3D CNN backbone. Specifically, we use the SlowFast [42], pretrained on the Kinetics dataset [43], within the VIP-QA framework. The features extracted from the Slow and Fast branches are subjected to average pooling and adaptive average pooling, respectively, and then concatenated to produce the temporal feature TF_i. Finally, the spatio-temporal features are fused as F_i, and quality regression is performed to evaluate the video quality:

$$F_i = SF_i \oplus TF_i. \tag{3}$$

Subsequently, two fully connected (FC) layers are utilized to regress the fused features into quality scores. The final quality score is then computed through average pooling:

$$V = \frac{1}{L}\sum_{i=1}^{L} V_i, \tag{4}$$

where V represents the overall quality score for the AH video, and L denotes the total number of keyframes used.

B. Identity Consistency

Unlike general videos, AH is character-specific, meaning that character identity information adds a crucial dimension to quality assessment. Subjectively, a high-quality imitator should preserve the character's physical characteristics as faithfully as possible while driving the character's movements. To quantify the fidelity of character appearance in the evaluation algorithm, InsightFace [44] is employed to measure the disparity between AI-generated images and the video frames extracted from the AH:

$$I_i = SIM_I(I_o, K_i), \tag{5}$$

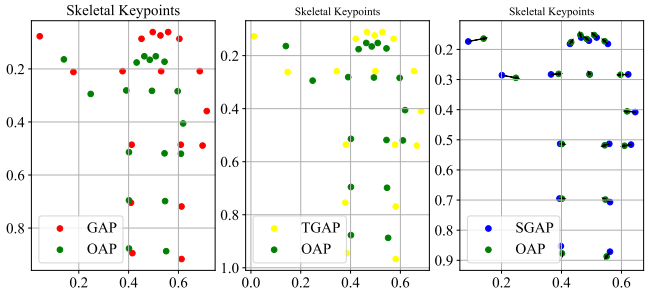
where I_i denotes the identity similarity between the i-th keyframe K_i and the original image I_o and $SIM_I(\cdot)$ denotes the process of difference calculation using InsightFace. Ultimately, the identity consistency I can be computed similarly to Eq. 4.

C. Posture Similarity

The primary objective of an imitator is to accurately replicate the poses from the input video, making pose similarity between the AH and the input video a critical factor in effective quality assessment. To address this, this paper introduces a similarity measure, Motion-p2point, based on skeletal keypoints. This measure is based on three fundamental assumptions and aims to efficiently handle pose comparisons in videos with different frame rates, resolutions and character sizes:

- **A1**: *The character movements in the original animation video and the corresponding AH video are consistent in their relative positions throughout the entire video.*
- **A2**: *Characters in videos of varying resolutions and sizes are subject only to proportional scaling.*
- **A3**: *The center of each posture in animation is fixed relative to the overall skeletal keypoints.*

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3572000



(a) GAP (b) TGAP (c) SGAP

Fig. 9. Flowchart for the visualization of Motion-p2point. the GAP is aligned with the OAP after translation and scaling, respectively, and the error distance is finally computed.

Based on **A1**, each AH video is uniformly sampled to include five keyframes that capture the action similarity at the beginning, the first half, the middle, the second half, and the end of the video. This approach contrasts with commonly used methods of sampling at equal time intervals or based on a fixed number of frames, offering effective application to AHs with varying frame rates. For each sampled frame, the 2D pose estimation method, DWPose, is employed to record the 18-point joint coordinates. Due to transformations such as scaling and panning applied during the imitator process, directly calculating point-to-point distances is not feasible. To address this, a compensation matrix C is introduced to account for the required scaling and translation adjustments:

$$\begin{aligned}\mathbb{R} &= \begin{pmatrix} \hat{x}_i^1 & \hat{x}_i^2 & \cdots & \hat{x}_i^{18} \\ \hat{y}_i^1 & \hat{y}_i^2 & \cdots & \hat{y}_i^{18} \\ 1 & 1 & \cdots & 1 \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} \alpha & 0 & a \\ 0 & \beta & b \\ 0 & 0 & 1 \end{pmatrix}}_{C} \underbrace{\begin{pmatrix} x_i^1 & x_i^2 & \cdots & x_i^{18} \\ y_i^1 & y_i^2 & \cdots & y_i^{18} \\ 1 & 1 & \cdots & 1 \end{pmatrix}}_{R},\end{aligned} \tag{6}$$

where $(x_i^j, y_i^j, 1)^T$ denote the coordinates of the j-th skeletal keypoint in the i-th frame before transformation. These coordinates are arranged to form the original skeletal matrix R, while the transformed skeletal matrix $\mathbb{R}$ is composed of coordinates $(\hat{x}_i^j, \hat{y}_i^j, 1)^T$. The compensation matrix $C(\alpha, \beta, a, b)$ represents an affine transformation matrix, where α and β determine the scaling factors in the horizontal and vertical directions, respectively, and a and b determine the translations in the horizontal and vertical directions. According to **A2**, the condition $\alpha = \beta$ is applied. Furthermore, based on **A3**, the translations a and b are determined using the center of posture. Specifically, the centers C_{GAP} and C_{OAP} for both Generated Animation Posture (GAP) and Original Animation Posture (OAP) are calculated to inform these translation values as Fig. 9:

$$\begin{aligned} C_{GAP} &= \frac{1}{18}\sum_{j=1}^{18}\begin{pmatrix} x^j \\ y^j \end{pmatrix} = \begin{pmatrix} x_G \\ y_G \end{pmatrix}, \\ C_{OAP} &= \frac{1}{18}\sum_{j=1}^{18}\begin{pmatrix} \bar{x}^j \\ \bar{y}^j \end{pmatrix} = \begin{pmatrix} x_O \\ y_O \end{pmatrix}, \end{aligned} \tag{7}$$

where $(\bar{x}^j, \bar{y}^j)^T$ denotes the coordinates of the j-th skeletal keypoint in the OAP. And, all the $(\bar{x}^j, \bar{y}^j, 1)^T$ are sequentially arranged to form the OAP skeletal matrix $\bar{R}$. By calculating the centers of GAP and OAP, the center offset Δ can be calculated:

$$\begin{aligned} \Delta_x &= x_O - x_G. \\ \Delta_y &= y_O - y_G\,. \end{aligned} \tag{8}$$

Translate the center of the GAP to the center of the OAP, thus obtaining the translated GAP (TGAP). At the same time, C_{GAP} and C_{OAP} obey the transformation relations of the compensation matrix C, so by joining Eq. 6 and Eq. 8, a and b can be determined with respect to the α:

$$\begin{aligned} a(\alpha) &= (1-\alpha)x_G + \Delta_x. \\ b(\alpha) &= (1-\alpha)y_G + \Delta_y\,. \end{aligned} \tag{9}$$

Thus the compensation matrix $C(\alpha, \beta, a, b)$ can be further simplified as $C(\alpha)$. The whole problem of comparing GAP and OAP similarities reduces to finding the optimal scaling factor α that minimizes the distance between pairs of keypoints. This can be formulated as an optimization problem using the following equation:

$$\begin{cases} \min\ D_i(\alpha) = \sum_{j=1}^{18} ||\varphi_i^j, \phi_i^j(\alpha)||_2, \\ s.t. \quad \varphi \in \bar{R}, \phi(\alpha) \in C(\alpha)R \quad , \end{cases} \tag{10}$$

where φ is the column vector in the OAP skeletal matrix $\bar{R}$, and ϕ is the column vector of the scaled GAP (SGAP) skeletal matrix $C(\alpha)R$. The distance between the coordinates of the key points of the bones in the two skeletal matrices is D_i. Obviously, the problem is not complicated to solve by means of derivatives or various heuristic algorithms. It may be assumed that the optimal solution is D_i^*, and the final Motion-p2point metric can be computed by averaging the minimum distances of the extracted frames:

$$P = \frac{1}{5}\sum_{i=1}^{5} D_i^*, \tag{11}$$

where P is the posture similarity computed using Motion-p2point. Since a total of 5 frames are sampled during frame extraction, only the average of 5 frames is computed. This method differs from other pose similarity algorithms by placing more emphasis on pose accuracy and is an important component of imitator performance.

D. Quality Regression & Loss Function

After extracting the three feature aspects, an FC layer is employed to regress V, I, P features into a final predicted AH quality score. During the training process, this predicted AH quality is compared with the actual MOS using the Mean Squared Error (MSE) as the loss function, enabling the algorithm to be gradually optimized for improved performance:

$$L_{MSE} = \frac{1}{n}\sum^{n}(\hat{Q} - Q)^2, \tag{12}$$

where $\hat{Q}$ is the predicted quality scores, Q is the quality labels of the AH, and n is the size of the training batch.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3572000

TABLE III
PERFORMANCE RESULTS ON THE PROPOSED AHQA AND DDH-QA DATABASES, WHERE PSNR AND SSIM ARE FULL-REFERENCE QUALITY ASSESSMENT ALGORITHMS AND ARE THEREFORE NOT APPLICABLE TO THE AHQA DATASET. THE BEST PERFORMANCE FOR EACH METRIC IS LABELED IN **RED** AND THE SECOND ONE IS LABELED IN **BLUE**.

Type	Label	Models	AHQA				DDH-QA			
			SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
IQA	A	PSNR	-	-	-	-	0.4308	0.5458	0.3114	0.9013
	B	SSIM [45]	-	-	-	-	0.5408	0.6057	0.3920	0.8559
	C	BRISQUE [46]	0.0674	0.1836	0.0455	0.7019	0.3664	0.4011	0.2568	1.0067
	D	NIQE [47]	0.1086	0.1562	0.0909	0.7068	0.0923	0.2489	0.0748	1.0418
	E	IL-NIQE [48]	0.1523	0.1742	0.1228	0.7052	0.0604	0.1062	0.0404	1.0718
VQA	F	VIIDEO [49]	0.1663	0.1787	0.1273	0.7053	0.1219	0.1829	0.0732	1.0740
	G	TLVQM [50]	0.3716	0.4394	0.2647	0.6225	0.2515	0.2824	0.1729	1.0480
	H	VIDEVAL [51]	0.3866	0.4317	0.2717	0.6204	0.2218	0.3470	0.1622	1.0246
	I	V-BLIINDS [52]	0.3110	0.4121	0.2187	0.6245	0.4807	0.4936	0.3424	0.9564
	J	RAPIQUE [53]	0.4390	0.4733	0.3153	0.6013	0.1815	0.2368	0.1246	1.0614
	K	SimpVQA [36]	**0.7566**	**0.7584**	**0.5748**	**0.5017**	**0.7444**	**0.7498**	**0.5452**	**0.7228**
	L	VSFA [54]	0.6866	0.6840	0.5107	0.5973	0.5406	0.5708	0.3858	0.9657
	M	FAST-VQA [55]	0.7515	0.7315	0.5710	0.5610	0.5262	0.5382	0.3657	1.0499
	N	BVQA [56]	0.7164	0.7161	0.5366	0.5709	0.6304	0.6396	0.4510	0.7663
	O	**VIP-QA (Proposed)**	**0.8063**	**0.8056**	**0.6216**	**0.4649**	**0.8090**	**0.8053**	**0.6298**	**0.6169**

V. EXPERIMENTS

A. Experiment Setups

A series of experiments are conducted to evaluate the effectiveness of the proposed VIP-QA algorithm. Several typical image quality assessment (IQA) and video quality assessment (VQA) methods are selected as benchmarks. These include RAPIQUE, SimpVQA, VSFA, FAST-VQA, and BVQA, which are based on deep learning, while other methods rely on manual feature extraction. All algorithms are trained and tested on the proposed AHQA dataset and the DDH-QA dataset, following a consistent experimental protocol. Between the selected datasets, The DDH-QA dataset comprises 800 3D dynamic digital human videos, featuring ten distinct actions performed by two characters (one male and one female). For dataset partitioning, both AHQA and DDH-QA datasets are cross-validated using five-fold cross-validation, ensuring no content overlap between the folds. The key difference is that for the AHQA dataset, data division is based on character images, whereas for the DDH-QA dataset, it is based on actions. The performance of each algorithm is recorded as the average across the five test folds. It is important to note that all competing algorithms are implemented using the original source code provided by their respective authors. For the VIP-QA, ResNet50 [57] is used as a 2D CNN, and the Adam optimizer is used [58]. Through careful debugging, we set the learning rate to 3e-5 and the batch size to 8. The experiments are conducted on a server equipped with an RTX 3090 GPU.

B. Experiment Criteria

To quantify the performance of each algorithm, four widely recognized metrics are selected: Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendall Rank Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE). The first three metrics yield values between 0 and 1, with values closer to 1 indicating superior algorithm performance. Conversely, RMSE measures prediction accuracy, where values closer to 0 reflect more accurate predictions by the algorithm.

C. Performance Analysis

The experimental results, presented in Table III, yield several key insights. 1) The existing Image Quality Assessment (IQA) and Video Quality Assessment (VQA) algorithms demonstrate limited effectiveness when applied to the quality assessment of AH videos. This limitation is primarily because most of these algorithms were designed for general images and videos rather than human-centered media like AH. 2) Among all the algorithms evaluated in the experiments, the proposed VIP-QA algorithm consistently outperforms others on both the AHQA and DDH-QA datasets. The VIP-QA shows a significant lead over several representative VQA methods, underscoring its effectiveness. This superior performance can be attributed to the incorporation of identity information and pose similarity measures. 3) Since the proposed AHQA dataset focuses on AI-generated 2D AHs while the DDH-QA dataset focuses on hand-modeled 3D AH videos, the excellent performance of VIP-QA on both datasets suggests that the proposed VIP-QA method is applicable to different types of AHs and is expected to provide a unified evaluation scheme for quality assessment of all types of AHs.

D. Ablation Experiments

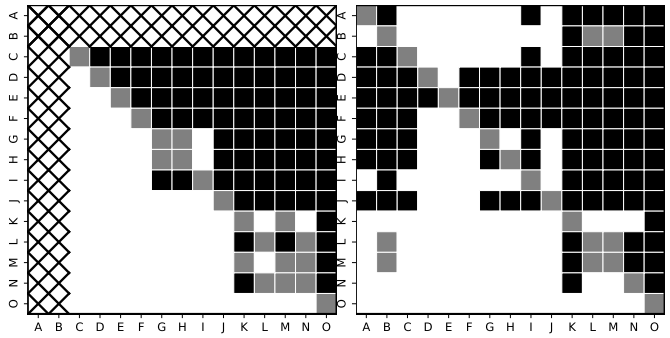
To further evaluate the individual contributions of each feature component, as well as to analyze the parameter count and computational complexity of different model variants, we conduct ablation studies on both the AHQA and DDH-QA datasets. These experiments specifically examine the three types of features integrated within the proposed VIP-QA framework. The results of these experiments are detailed in

TABLE IV
ABLATION STUDY RESULTS IN AHQA AND DDH-QA DATABASES, WHERE V, I, P DENOTE VIDEO QUALITY, IDENTITY CONSISTENCY AND POSTURE SIMILARITY, RESPECTIVELY. THE PARAMETER COUNT AND THE COMPUTATIONAL COMPLEXITY OF THE DIFFERENT MODELS ARE RECORDED. THE BEST PERFORMANCE FOR EACH METRIC IS LABELED IN **RED** AND THE SECOND ONE IS LABELED IN **BLUE**.

Model	Params	FLOPs	AHQA				DDH-QA			
			SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
V	24.72M	526.13G	0.7506	0.7526	0.5745	0.5258	0.7304	0.7311	0.5430	0.6606
I	43.80M	6.32G	0.3447	0.4060	0.2373	0.7076	0.3368	0.3463	0.2356	0.7283
P	54.17M	80.77G	0.3816	0.3943	0.2716	0.7080	0.3576	0.4028	0.2438	0.6953
$V+I$	68.52M	532.45G	0.7686	0.7655	0.5868	0.5136	0.7496	0.7465	0.5585	0.6429
$V+P$	78.89M	606.90G	**0.7753**	**0.7737**	**0.5948**	**0.5056**	**0.7838**	**0.7774**	**0.5889**	**0.6351**
$I+P$	97.97M	87.09G	0.4391	0.4587	0.3109	0.6868	0.4335	0.4529	0.3045	0.6867
$V+I+P$	122.69M	613.22G	**0.8063**	**0.8056**	**0.6216**	**0.4649**	**0.8090**	**0.8053**	**0.6298**	**0.6169**

TABLE V
CROSS-DATABASE EVALUATION, WHERE AHQA→DDH-QA INDICATES THE MODEL IS TRAINED ON THE AHQA DATABASE AND VALIDATED WITH THE DEFAULT TESTING SETUP OF THE DDH-QA DATABASE. THE BEST PERFORMANCE FOR EACH METRIC IS LABELED IN **RED** AND THE SECOND ONE IS LABELED IN **BLUE**.

Model	AHQA→DDH-QA		DDH-QA→AHQA	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BVQA	0.0765	0.0729	**0.3054**	**0.3317**
SimpVQA	0.1991	**0.2564**	0.2028	0.2320
FAST-VQA	**0.2240**	0.2165	0.2632	0.2772
VIP-QA	**0.2844**	**0.3222**	**0.3836**	**0.4052**



(a) AHQA Database (b) DDH-QA Database

Fig. 10. Statistical test results of the proposed VIP-QA method and compared methods on the AHQA and DDH-QA databases. A black/white block means the row method is statistically worse/better than the column one. A gray block means the row method and the column method are statistically indistinguishable. For missing values, the mark ✕ is used. The methods are denoted by the same index as in Table III.

Table IV. From the analysis of Table IV, the following conclusions can be drawn. 1) Each of the three feature types within VIP-QA contributes positively to the overall performance of the algorithm. 2) A comparison between Table III and Table IV shows that the identity consistency and pose similarity metrics proposed in this study outperform existing feature-extraction based assessment schemes. This indicates that these two types of features are a more accurate reflection of AH quality. 3) Among the three feature types, video quality is the most critical, followed by gesture similarity, with identity consistency being the least influential. This aligns with human visual perception, as video-level distortions are more apparent and easily perceived. Gesture information, while more detailed, is also crucial for assessing AH video quality. Lastly, in dynamic videos, the tolerance for variations in character identity is higher. 4) Despite the differences between the AHQA and DDH-QA datasets in terms of the data structure of the AH, the results of the ablation experiments show a similar trend on both datasets, which once again validates the robustness and generalizability of the VIP-QA algorithm.

E. Cross-database Evaluation

To assess the universality and generalizability of the proposed VIP-QA method, we conduct cross-database experiments using the established AHQA and DDH-QA datasets. The results of these experiments are presented in Table V. Upon reviewing Table V, several key observations can be made. 1) In all cross-database experiments, the VIP-QA method consistently outperforms the other methods by a significant margin (+6% SRCC), suggesting that it outperforms the other methods in terms of generalization. 2) However, it is noteworthy that all participating methods exhibit relatively poor performance in the cross-database context. This can be attributed to the differing characteristics of the datasets: the AHQA dataset primarily focuses on AI-Generated distortions, while the DDH-QA dataset emphasizes computer-simulated distortions. Therefore, the results presented in Table V suggest a significant variation in the visual effects produced by AH when subjected to generative versus computer-simulated distortions. 3) Interestingly, nearly all objective quality assessment algorithms perform better on the AHQA dataset when trained using the DDH-QA dataset, as opposed to being trained on AHQA dataset and tested on the DDH-QA dataset. Given the 3D AHs in the DDH-QA dataset and the 2D AHs in the AHQA dataset, this result implies that quality assessment algorithms designed for higher-dimensional AH data may still be applicable to lower-dimensional AH data, offering valuable insights for quality assessment in the context of low-dimensional AHs.

F. Statistical Test

In this section, statistical analyses are conducted to further assess the performance of the proposed method. Following the approach outlined in [59], the evaluation involves comparing the differences between the model-predicted quality scores and the subjective human evaluation scores. The null hypothesis posits that the residuals from one quality assessment model

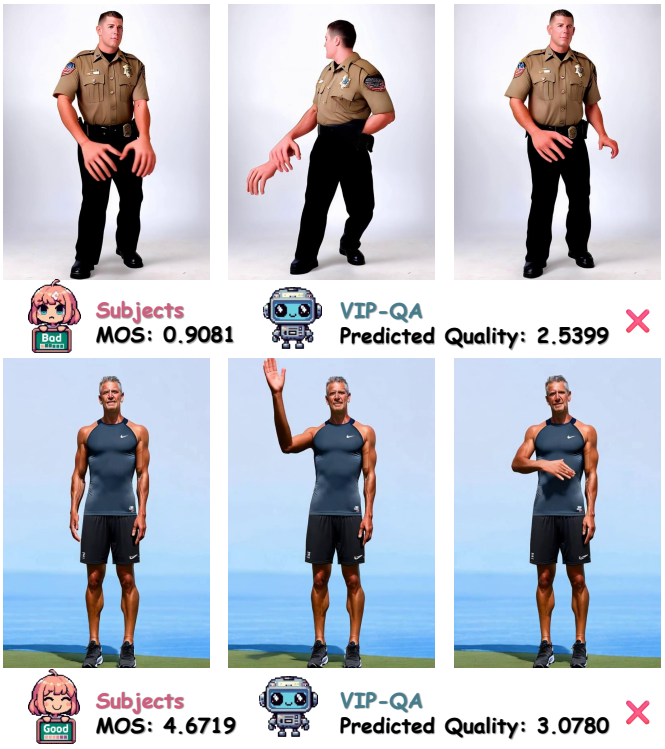


Fig. 11. Illustration of failure cases. From top to bottom are the first and second representative failure cases.

follow the same distribution and, at a 95% confidence level, are statistically indistinguishable from the residuals of another quality assessment model. The analysis includes testing all possible pairs of methods, with the results presented in Fig. 10. Notably, the proposed method demonstrates superior performance relative to all other methods evaluated using both the AHQA and DDH-QA datasets. This observed superiority is statistically significant, highlighting the robustness and reliability of the VIP-QA method in this context.

VI. Discussion

A. Failure Cases

Although Section V presents extensive experimental results that validate the effectiveness, robustness, and generalizability of the proposed VIP-QA framework, certain limitations remain evident in the prediction outcomes. Specifically, while the majority of predicted quality scores demonstrate strong alignment with the Mean Opinion Scores (MOS) derived from subjective evaluations, a small number of cases exhibit significant discrepancies. To provide further insight into these inconsistencies, two representative failure cases are illustrated in Fig. 11. The first example suggests that the VIP-QA model may still encounter difficulties in effectively capturing and interpreting depth-related features, which are critical for accurate quality assessment. The second example, when analyzed in conjunction with the MOS distribution shown in Fig. 7 (b), indicates that the relatively limited representation of high-quality samples within the AHQA dataset may constrain the model's capacity to assign appropriately high scores to genuinely high-quality AH videos.

B. Potential Improvements

In this study, we have conducted a comprehensive subjective and objective quality assessment of animated human (AH) videos for the first time. While the proposed framework demonstrates promising results, several limitations remain. To further advance research in AH quality assessment and its related domains, we outline potential directions for improvement: 1) Given that the visual quality of AH videos is influenced by a complex interplay of multiple perceptual factors, it remains valuable to explore more effective multidimensional subjective evaluation strategies. Developing such protocols not only facilitates the design of more accurate objective quality assessment methods but also enhances our understanding of the perceptual mechanisms underlying human evaluation of AH content from the perspective of the human visual system; 2) Although the proposed VIP-QA framework achieves state-of-the-art performance on current datasets, its posture [31] and facial feature [44] extraction components are still based on classical algorithms, which leaves considerable room for enhancement. Furthermore, the integration of advanced attention mechanisms [60] and adaptive spatio-temporal feature fusion strategies may offer significant opportunities to improve the performance and generalizability of VIP-QA in future work.

VII. Conclusions

Given the increasing prominence of AI animation generation techniques and their potential to supplant manual animation design, this paper addresses the critical need for quality assessment of AI-generated animated human videos. Initially, we introduce the AHQA dataset, the first of its kind for evaluating AI-generated animated human content. This dataset is created by collecting 20 character images and designing 10 common actions to ensure its diversity and comprehensiveness, resulting in a total of 1,200 AH videos produced by six representative animation generation methods. Subjective evaluations of these AH videos revealed that the quality of animations generated by current AI methods remains suboptimal. To address this, we propose VIP-QA, an objective evaluation framework that integrates three key dimensions: video quality, pose similarity, and identity consistency. VIP-QA demonstrates state-of-the-art performance on both the AHQA dataset and the DDH-QA dataset, which assesses hand-designed 3D animated digital humans, underscoring its validity and applicability. This work is intended to serve as a valuable reference and guide for advancing AI animation generation technologies.

References

[1] Y. Zhou, Y. Chen, K. Bi, L. Xiong, and H. Liu, "An implementation of multimodal fusion system for intelligent digital human generation," *arXiv preprint arXiv:2310.20251*, 2023.

[2] Y. Zhou, Z. Zhang, F. Wen, J. Jia, Y. Jiang, X. Liu, X. Min, and G. Zhai, "3dgcqa: A quality assessment database for 3d ai-generated contents," *arXiv preprint arXiv:2409.07236*, 2024.

[3] Z. Zhang, H. Wu, C. Li, Y. Zhou, W. Sun, X. Min, Z. Chen, X. Liu, W. Lin, and G. Zhai, "A-bench: Are lmms masters at evaluating ai-generated images?" *arXiv preprint arXiv:2406.03070*, 2024.

[4] C. Li, T. Kou, Y. Gao, Y. Cao, W. Sun, Z. Zhang, Y. Zhou, Z. Zhang, W. Zhang, H. Wu *et al.*, "Aigiqa-20k: A large database for ai-generated image quality assessment," *arXiv preprint arXiv:2404.03407*, vol. 2, no. 3, p. 5, 2024.

[5] L. Lin, Z. Wang, J. He, W. Chen, Y. Xu, and T. Zhao, "Deep quality assessment of compressed videos: A subjective and objective study," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2616–2626, 2022.

[6] Z. Zhang, Y. Zhou, W. Sun, X. Min, Y. Wu, and G. Zhai, "Perceptual quality assessment for digital human heads," in *ICASSP*. IEEE, 2023, pp. 1–5.

[7] Z. Zhang, Y. Zhou, W. Sun, W. Lu, X. Min, Y. Wang, and G. Zhai, "Ddh-qa: A dynamic digital humans quality assessment database," in *ICME*. IEEE, 2023, pp. 2519–2524.

[8] Z. Zhang, W. Sun, Y. Zhou, H. Wu, C. Li, X. Min, X. Liu, G. Zhai, and W. Lin, "Advancing zero-shot digital human quality assessment through text-prompted evaluation," *arXiv preprint arXiv:2307.02808*, 2023.

[9] Z. Zhang, Y. Zhou, L. Teng, W. Sun, C. Li, X. Min, X.-P. Zhang, and G. Zhai, "Quality-of-experience evaluation for digital twins in 6g network environments," *IEEE Transactions on Broadcasting*, 2024.

[10] Y. Zhou, Z. Zhang, W. Sun, X. Liu, X. Min, Z. Wang, X.-P. Zhang, and G. Zhai, "Thqa: A perceptual quality assessment database for talking heads," *arXiv preprint arXiv:2404.09003*, 2024.

[11] Y. Zhou, Z. Zhang, W. Sun, X. Liu, X. Min, and G. Zhai, "Subjective and objective quality-of-experience assessment for 3d talking heads," in *ACM Multimedia 2024*, 2024. [Online]. Available: https://openreview.net/forum?id=Zqi6eZri8Z

[12] Y. Zhou, Z. Zhang, F. Wen, J. Jia, X. Min, J. Wang, and G. Zhai, "Reliqa: A multidimensional quality assessment dataset for relighted human heads," in *IEEE Visual Communications and Image Processing*, 2024.

[13] Y. Zhou, Z. Zhang, J. Cao, J. Jia, Y. Jiang, F. Wen, X. Liu, X. Min, and G. 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.

[14] W. Wang, W. Zhou, J. Bao, and H. Li, "Coherent image animation using spatial-temporal correspondence," *IEEE Transactions on Multimedia*, vol. 25, pp. 3397–3408, 2022.

[15] H. Yu, X. Fan, Y. Hou, W. Pei, H. Ge, X. Yang, D. Zhou, Q. Zhang, and M. Zhang, "Toward realistic 3d human motion prediction with a spatio-temporal cross-transformer approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5707–5720, 2023.

[16] Z. Zhou, G. Zhao, Y. Guo, and M. Pietikainen, "An image-based visual speech animation system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 10, pp. 1420–1432, 2012.

[17] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490.

[18] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, "Champ: Controllable and consistent human image animation with 3d parametric guidance," *arXiv preprint arXiv:2403.14781*, 2024.

[19] Q. Wang, Z. Jiang, C. Xu, J. Zhang, Y. Wang, X. Zhang, Y. Cao, W. Cao, C. Wang, and Y. Fu, "Vividpose: Advancing stable video diffusion for realistic human image animation," *arXiv preprint arXiv:2405.18156*, 2024.

[20] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[21] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang, "Unianimate: Taming unified video diffusion models for consistent human image animation," *arXiv preprint arXiv:2406.01188*, 2024.

[22] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.

[23] Y. Zhang, J. Gu, L.-W. Wang, H. Wang, J. Cheng, Y. Zhu, and F. Zou, "Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance," *arXiv preprint arXiv:2406.19680*, 2024.

[24] W. Sun, W. Zhang, Y. Jiang, H. Wu, Z. Zhang, J. Jia, Y. Zhou, Z. Ji, X. Min, W. Lin *et al.*, "Dual-branch network for portrait image quality assessment," *arXiv preprint arXiv:2405.08555*, 2024.

[25] Z. Zhang, Y. Zhou, W. Sun, X. Min, and G. Zhai, "Geometry-aware video quality assessment for dynamic digital human," in *ICIP*. IEEE, 2023, pp. 1365–1369.

[26] S. Chen, Z. Zhang, Y. Zhou, W. Sun, and X. Min, "A no-reference quality assessment metric for dynamic 3d digital human," *Displays*, vol. 80, p. 102540, 2023.

[27] Z. Zhang, Y. Zhou, C. Li, K. Fu, W. Sun, X. Liu, X. Min, and G. 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)*. IEEE, 2024, pp. 2965–2969.

[28] Y. Zhou, Z. Zhang, W. Sun, X. Min, X. Ma, and G. Zhai, "A no-reference quality assessment method for digital human head," in *ICIP*. IEEE, 2023, pp. 36–40.

[29] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.

[30] Z. Tong, C. Li, Z. Chen, B. Wu, and W. Zhou, "Musepose: a pose-driven image-to-video framework for virtual human generation," *arxiv*, 2024.

[31] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective whole-body pose estimation with two-stages distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4210–4220.

[32] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.

[33] Y. Shan, S. Wang, Z. Zhang, and K. Huang, "An xt slice based method for action recognition," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 1897–1903.

[34] International Telecommunication Union, "Methodology for the subjective assessment of the quality of television pictures," ITU Radiocommunication Sector, Geneva, Switzerland, Recommendation BT.500-11, 2002, recommendation ITU-R BT.500-11. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500

[35] Z. Zhang, W. Sun, Y. Zhou, J. Jia, Z. Zhang, J. Liu, X. Min, and G. Zhai, "Subjective and objective quality assessment for in-the-wild computer graphics images," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 4, pp. 1–22, 2023.

[36] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *ACM MM*, 2022.

[37] C. He, Q. Zheng, R. Zhu, X. Zeng, Y. Fan, and Z. Tu, "Cover: A comprehensive video quality evaluator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5799–5809.

[38] Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A. C. Bovik *et al.*, "Video quality assessment: A comprehensive survey," *arXiv preprint arXiv:2412.04508*, 2024.

[39] S. Mishra, M. Jha, and A. C. Bovik, "Subjective and objective analysis of indian social media video quality," *IEEE Transactions on Image Processing*, 2024.

[40] Q. Zheng, Z. Tu, X. Zeng, A. C. Bovik, and Y. Fan, "A completely blind video quality evaluator," *IEEE Signal Processing Letters*, vol. 29, pp. 2228–2232, 2022.

[41] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, J. Gu, and W. Lin, "Neighbourhood representative sampling for efficient end-to-end video quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 185–15 202, 2023.

[42] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *IEEE/CVF CVPR*, 2019, pp. 6202–6211.

[43] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[44] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[45] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[46] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE TIP*, vol. 21, no. 12, pp. 4695–4708, 2012.

[47] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE SPL*, vol. 20, no. 3, pp. 209–212, 2012.

[48] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE TIP*, vol. 24, no. 8, pp. 2579–2591, 2015.

[49] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE TIP*, vol. 25, no. 1, pp. 289–300, 2015.

[50] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE TIP*, vol. 28, no. 12, pp. 5923–5938, 2019.

[51] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.
[52] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE TIP*, 2014.
[53] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Rapique: Rapid and accurate video quality prediction of user generated content," *IEEE DOAJ*, vol. 2, pp. 425–440, 2021.
[54] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *ACM MM*, 2019, pp. 2351–2359.
[55] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *ECCV*. Springer, 2022, pp. 538–554.
[56] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE TCSVT*, vol. 32, no. 9, pp. 5944–5958, 2022.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF CVPR*, 2016, pp. 770–778.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2014.
[59] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE TIP*, vol. 15, no. 11, pp. 3440–3451, 2006.
[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

Yingjie Zhou received his B.E. degree in electronics and information engineering from China University of Mining and Technology in 2023. He is currently pursuing a PhD degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China. His current research interests include digital human quality assessment and sentiment analysis.

Zicheng Zhang reveiced his B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 2020 and he is currently a PhD canidate at the School of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University. His research interets include large multi-modal models and perceptual quality assessment.

Jun Jia received the B.S. degree in computer science and technology from Hunan University, Changsha, China, in 2018 and received the PhD degree in the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2024. His research interests include computer vision and image processing.

Yanwei Jiang received his B.E. degree from Shanghai Jiao Tong University in 2022, where he is currently pursuing the PhD degree with the Institute of Image Communication and Information Processing. His research interests include image quality assessment and multimedia signal processing.

Xiaohong Liu received the PhD degree in electrical and computer engineering from McMaster University, Canada, in 2021. He is currently an Associate Professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. His research interests lie in computer vision and multimedia. He has published more than 80 academic papers in top journals and conferences such as IEEE TIP, IEEE TMM, CVPR, ICCV, ECCV, etc., and has been awarded the Microsoft Research Asia StarTrack Scholars Program in 2024, Shanghai Pujiang Program in 2022, Chinese Government Award for Outstanding Self-financed Students Abroad in 2021, and Borealis AI Fellowships in 2020. His research has been supported by the Young Scientists Fund of the National Natural Science Foundation of China (NSFC), Young Scientists Fund of Natural Science Foundation of Sichuan Province, and many high-tech companies. He also serves as Associate Editor of the ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM).

Xiongkuo Min (IEEE Member, 2019) received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the PhD degree from Shanghai Jiao Tong University, Shanghai, China, in 2018. From Jan. 2016 to Jan. 2017, he was a visiting student at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Post-Doctoral Fellow with Shanghai Jiao Tong University. His research interests include visual quality assessment, visual attention modeling and perceptual signal processing.

Guangtao Zhai (IEEE Fellow, 2024) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the PhD degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent PhD Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.