WHO IS A BETTER ARCHITECT: SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT FOR TALKING HUMAN INTERACTION PIPELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive digital humans are commonly built as cascades of specialized components, known as learnware (LW), offering lower response latency and smaller deployed models relative to omni models. Existing quality assessment, however, focuses primarily on individual learnware and lacks a system-level framework for evaluating the overall interaction. We integrate four learnware central to digital human interaction—a large language model (LLM), text-to-speech (TTS) synthesis, text-to-image (T2I) character generation, and a talking-human (TH) generator, or Talker—into the LT3 pipeline. To study interaction quality from the perspective of learnware, we introduce THQA-LW, a dataset of 10,000 single-turn interaction videos, together with a subjective protocol that separately scores each learnware and the overall interaction experience. We then analyze how learnware quality relates to overall interaction quality under controlled generation conditions. Finally, we propose LT3-FAMT, a five-task quality predictor that combines Qwen2.5-Omni Thinker with complementary multimodal feature extractors. LT3-FAMT achieves state-of-the-art performance across multiple benchmarks, supporting automatic assessment and diagnosis of quality bottlenecks in interactive digital humans.

1 INTRODUCTION

Digital humans increasingly mediate interactions in education, healthcare, and information services, where the quality of their responses shapes the user experience Guo et al. (2024); Chattopadhyay et al. (2020); Kim et al. (2022). Most interactive digital-human systems are cascades of specialized components, or learnware (LW) Zhou (2016), that may include automatic speech recognition, language generation, speech synthesis, character generation, talking-human generation, and virtual try-on. A weakness in any learnware can substantially diminish the final interaction, making the selection and coordination of learnware a system-level task for the architect Kim et al. (2018); Wang et al. (2026b). We focus on four widely used types of learnware—a large language model (LLM), text-to-speech (TTS) synthesis, text-to-image (T2I) character generation, and a talking-human generator (Talker)—and organize them into the LT3 pipeline shown in Fig. 1. Given a user's question, the LLM first generates an answer, which TTS converts into audio. T2I generates a reference image from an avatar prompt, and the resulting audio and image jointly drive the Talker to synthesize the interaction video. Traditional talking-human quality assessment (THQA) concentrates on the final video Zhou et al. (2024b;c; 2025a); Su et al. (2025), whereas the user's experience reflects the quality of learnware throughout this pipeline. Assessing each stage alongside the overall interaction can therefore identify quality bottlenecks and guide the architect's choices across the system.

To study this problem from both subjective and objective perspectives, we construct THQA-LW, a large-scale dataset of 10,000 single-turn interaction videos with traceable generation provenance. We conduct controlled subjective experiments with multiple participants, collecting ratings for the output of each learnware and for the overall interaction experience. These experiments reveal ten distortion types, which we define for diagnostic analysis. We further analyze how the quality of different learnware relates to overall interaction quality under controlled generation conditions. For efficient objective assessment, we propose LT3-FAMT, a multitask multimodal quality assessment framework. Extensive experiments show that LT3-FAMT outperforms existing quality assessment

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

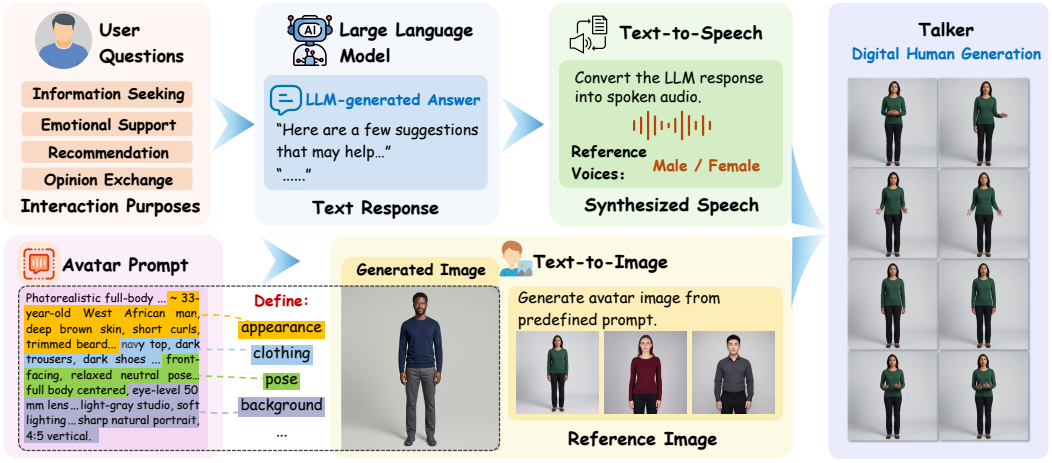

Figure 1: The LT3 pipeline for interactive digital humans. The LLM generates an answer to a user's question, TTS converts the answer into audio, and T2I generates a reference image from an avatar prompt. The audio and reference image jointly drive the Talker to produce the interaction video.

methods across multiple datasets and aligns closely with human audiovisual judgments. By identifying quality bottlenecks across the pipeline, LT3-FAMT guides improvements to interactive digital-human systems and helps architects deliver a better user experience. In summary, our contributions are threefold:

- We select four representative learnware in interactive digital-human systems—LLM, TTS, T2I, and Talker—and integrate them into the LT3 pipeline. This pipeline provides an end-to-end setting for studying how the successive stages of response generation shape interaction quality.
- We construct THQA-LW, a large-scale dataset of 10,000 interaction videos with subjective ratings for each learnware and the overall experience. The dataset extends talking-human quality assessment from the final video to the full interaction pipeline and supports an analysis of how learnware quality relates to overall quality.
- The proposed LT3-FAMT is a multitask multimodal framework for predicting learnware-level and overall interaction quality. Its predictions consistently surpass those of existing assessment methods across multiple datasets and show strong agreement with human audiovisual perception.

2 Related Work

2.1 Interactive Digital-Human Architectures

An interactive digital-human architecture coordinates LW modules that transform user input into an audiovisual response and, where needed, refine the character's appearance or the generated video Fu et al. (2008); Schreer et al. (2008). Such architectures may include automatic speech recognition (ASR), LLM, TTS, T2I, Talker, virtual try-on, and video super-resolution Kim et al. (2018); Wang et al. (2026b); Han et al. (2018); Chan et al. (2021). Representative approaches include connectionist temporal classification and sequence-to-sequence decoding for ASR Graves et al. (2006); Chan et al. (2016), autoregressive Transformers for LLMs Vaswani et al. (2017), acoustic modeling with vocoders for TTS Shen et al. (2018), diffusion-based synthesis for T2I Rombach et al. (2022); Podell et al. (2023), garment warping for virtual try-on Han et al. (2018), and motion-compensated reconstruction for video super-resolution Chan et al. (2021). Talking-human generation includes speech-driven Talkers, from lip-synchronization methods such as Wav2Lip to systems based on 3D motion representations, neural rendering, and diffusion, as well as prompt-driven audiovisual generators, or Speakers, that synthesize speech and video from text Prajwal et al. (2020); Zhang et al. (2023a); Ma et al. (2023); Wang et al. (2026b).

Each LW module introduces distinct quality risks. ASR errors can alter the interpreted request; an LLM can produce an irrelevant or inaccurate answer; TTS can introduce unclear speech or unnatural

Table 1: Comparison of digital-human quality assessment databases.

Database	Year	Modality	Scale	Typical Evaluation Methods	Distortion Types	LW-level Analysis	Description
DHH-QA Zhang et al. (2023d)	2023	Mesh + UV	1,540	Zhang *et al.* Zhang et al. (2023d), Zhou *et al.* Zhou et al. (2023)	7	×	Scanned Real Human Heads
DDHQA Zhang et al. (2023c)	2023	Mesh + UV	800	Zhang *et al.* Zhang et al. (2023e), Chen *et al.* Chen et al. (2023)	9	×	Dynamic 3D Digital Human
6G-DTQA Zhang et al. (2024)	2024	Mesh + UV	400	Zhang *et al.* Zhang et al. (2024)	5	×	Dynamic 3D Digital Human
THQA-3D Zhou et al. (2024c)	2024	Mesh + UV	1,000	Zhou *et al.* Zhou et al. (2024c)	5	×	Scanned Real Human Heads
SJTU-H3D Zhang et al. (2023b)	2025	Mesh + UV	1,120	Zhang *et al.* Zhang et al. (2023b)	7	×	Static 3D Digital Humans
ReLI-QA Zhou et al. (2024d)	2024	Image	840	Wen *et al.* Wen et al. (2025)	4	×	Relighted Human Heads
THQA Zhou et al. (2024b)	2024	Video + Audio	800	MI3S Zhou et al. (2025d), Xu et al. (2025b)	9	×	AI-Generated Talking Heads
THQA-10K Zhou et al. (2025a)	2025	Video + Audio	10,457	FSCD Zhou et al. (2025a)	10	×	AI-Generated Talking Heads
THQA-NTIRE Liu et al. (2025)	2025	Video + Audio	12,257	Su *et al.* Su et al. (2025)	15	×	2D and 3D Talking Heads
AHQA Zhou et al. (2025c)	2025	Video	1,200	VIP-QA Zhou et al. (2025c)	4	×	Animated Humans
MEMO-Bench Zhou et al. (2024a)	2025	Image	7,145	None	1	×	Emotional Human Heads
CDHQA Zhou et al. (2025b)	2025	Video	254	None	3	×	Interactive Digital Human
THQA-MT Zhou et al. (2025e)	2025	Video + Audio	5,492	EvalTalker Zhou et al. (2025e)	12	×	Multi-Subject Talking Humans
THQA-LW (Ours)	**2026**	**Video + Audio**	**10,000**	**LT3-FAMT (Ours)**	**10**	✓	**Interaction-Oriented Talking Humans**

prosody; and T2I can produce an unsuitable reference appearance Kim et al. (2018); Lin et al. (2021); Shen et al. (2018); Li et al. (2024). Virtual try-on may distort clothing, while video super-resolution may introduce spatial artifacts or temporal instability Han et al. (2018); Chan et al. (2021). At the final generation stage, Talkers can exhibit lip–speech mismatch, identity drift, stiff motion, flicker, and audiovisual inconsistency Zhou et al. (2024b; 2025a). These degradations affect different aspects of the interaction and may propagate through subsequent stages Kim et al. (2018); Wang et al. (2026b). We therefore study four core and widely applicable LW modules—LLM, TTS, T2I, and Talker—within the LT3 pipeline, where their differing effects on overall interaction quality can be assessed.

2.2 Digital Human Quality Assessment

Digital-human quality databases span scanned and dynamic 3D humans, relighted portraits, talking heads, animated humans, and interactive digital humans, as summarized in Table 1. DHH-QA Zhang et al. (2023d), DDHQA Zhang et al. (2023c), 6G-DTQA Zhang et al. (2024), THQA-3D Zhou et al. (2024c), and SJTU-H3D Zhang et al. (2023b) focus on 3D assets. Video-oriented databases such as THQA Zhou et al. (2024b), THQA-10K Zhou et al. (2025a), and THQA-NTIRE Liu et al. (2025) emphasize generated talking humans, while CDHQA Zhou et al. (2025b) considers interactive digital humans. Across these settings, annotations primarily describe the quality of the assessed asset or final media. They provide limited evidence about how upstream LW modules shape the delivered interaction.

Objective methods likewise concentrate on the assessed image or video. BRISQUE estimates spatial degradation from image statistics Mittal et al. (2012a), FAST-VQA models video quality Wu et al. (2022), and SyncNet-derived LSE-D and LSE-C measure speech–lip alignment Chung & Zisserman (2017). Talking-human-specific methods combine additional perceptual cues: FSCD Zhou et al. (2025a) uses the first frame, a Y–T slice, and tone–lip consistency; MI3S Zhou et al. (2025d) integrates image quality, aesthetics, identity consistency, and lip synchronization; Xu *et al.* Xu et al. (2025b) model facial morphology and expression coordination; Su *et al.* Su et al. (2025) fuse spatial, temporal, and audio features; and EvalTalker Zhou et al. (2025e) adds body, identity, and multimodal synchronization features for multi-subject videos. These methods can assess properties of generated media, but a decline in overall interaction quality cannot be reliably attributed to an upstream LW module when its output is not evaluated. THQA-LW extends quality assessment to the full interaction by providing subjective judgments for the LLM answer, TTS speech, T2I reference image, Talker video, and overall experience. The resulting multidimensional ratings support analysis of how each LW module relates to the final judgment. LT3-FAMT uses multimodal evidence to predict these quality dimensions jointly, enabling more specific diagnosis than a single overall Mean Opinion Score (MOS). Such diagnosis can guide the selection and refinement of LW modules, improving generated interactions and supporting digital-human applications across domains.

3 Database Construction

3.1 LT3 Pipeline

The LT3 pipeline connects four core LW modules through two paths that converge at video generation. An LLM generates an answer to a user question, and TTS converts the answer into speech Vaswani et al. (2017); Shen et al. (2018). In parallel, T2I generates a reference portrait from a predefined avatar prompt Rombach et al. (2022). The synthesized speech and reference portrait

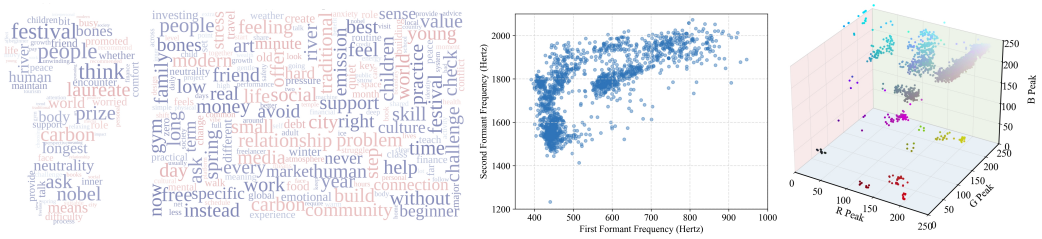

Figure 2: Analysis of multimodal data diversity in the LT3 generation design, including word clouds of interaction questions and LLM-generated answers, the distribution of the first and second formant frequencies of synthesized speech, and the three-dimensional distribution of RGB peak values in generated interaction videos.

Table 2: Models used in the LT3 pipeline.

Model	LLM	TTS	T2I	Talker
Model 1	GPT-5.5 OpenAI (2026a)	Confucius4-TTS Wang et al. (2026a)	GPT-Image-2 OpenAI (2026b)	InfiniteTalk Yang et al. (2025)
Model 2	Gemini 3.5 Flash Google (2026b)	Qwen3-TTS Hu et al. (2026)	Gemini 3.1 Flash Image Google (2026a)	HunyuanVideo-Avatar Chen et al. (2025)
Model 3	Claude Fable 5 Anthropic (2026)	CosyVoice Du et al. (2024)	Qwen-Image-2.0 Zhao et al. (2026)	OmniAvatar Gan et al. (2025)
Model 4	Seed2.1 Pro ByteDance Seed Team (2026a)	MOSS-TTS Gong et al. (2026)	Doubao Seedream 5.0 Lite ByteDance Seed Team (2026b)	SoulX-LiveAct Zhen et al. (2026)
Model 5	Qwen 3.7-Max Alibaba Cloud (2026a)	GLM-TTS Cui et al. (2025)	Wan2.7 Image Alibaba Cloud (2026b)	LongCat-Video-Avatar-1.5 Team et al. (2026)

then jointly drive the Talker to produce the interaction video Chen et al. (2025); Gan et al. (2025). We select five models for each LW module, as listed in Table 2. The generation design covers four representative interaction purposes: information seeking, emotional support, recommendation, and opinion exchange. The interaction questions address a broad range of topics and elicit semantically and lexically diverse answers from different LLMs. Multiple TTS models and reference voices provide substantial phonetic and acoustic variation, while different portrait prompts and T2I models produce diverse character appearances. These complementary variations provide textual, acoustic, and visual diversity across the LT3 pipeline, as illustrated by the question and answer word clouds, synthesized speech formant distribution, and generated video RGB peak distribution in Fig. 2.

3.2 THQA-LW Construction

We begin with 16 single-turn user questions, distributed equally across the four interaction purposes. Each of the five LLMs produces one answer per question, yielding $16 \times 5 = 80$ question–answer pairs. Each of the five TTS models synthesizes every answer using four reference voices, comprising two male and two female voices. This produces $80 \times 5 \times 4 = 1{,}600$ source utterances. Separately, each of the five T2I models generates one character image from each of four predefined portrait prompts, yielding $5 \times 4 = 20$ reference images. The utterances and images permit $1{,}600 \times 20 = 32{,}000$ distinct audio–image pairs. We select 2,000 pairs through constrained random sampling that balances coverage across questions, models, voices, and images without repeating an audio–image combination. This selection retains variation among upstream conditions while keeping video generation and subjective evaluation manageable. Applying all five Talkers to the same selected pairs produces $2{,}000 \times 5 = 10{,}000$ interaction videos, which constitute THQA-LW. The matched design yields five videos for each selected audio–image pair and supports comparisons among Talkers under shared upstream inputs. Generation records link every video to its question and interaction purpose, answer, source audio, reference image, model identities, reference voice, and generation configuration.

3.3 Subjective Experiment and Data Processing

We conduct a controlled subjective experiment with 40 participants to evaluate the LLM, TTS, T2I, Talker, and overall interaction quality separately. The experiment follows ITU-R BT.500-13 ITU Radiocommunication Sector (2012) and adopts the ACR protocol with randomized presentation and concealed model identities. During the overall interaction evaluation, participants also identify ten diagnostic distortion categories. Following observer screening, the valid ratings are normalized within each quality dimension and averaged to obtain MOSs. The resulting five MOSs form the multidimensional quality label of each interaction video, while distortion labels are determined by majority voting. Complete details of the evaluation criteria, experimental procedure, rating normalization, and label aggregation are provided in Appendix A.

3.4 Mean Opinion Score Analysis

We examine the MOS distributions and the effects of Talker choice and video duration on interaction quality. Fig. 3 presents the quality distributions across the five dimensions and the duration-dependent performance of the five Talkers. The results yield four key observations: **1)** The MOS distribution for each dimension follows a Gaussian pattern, while its spread differs across dimensions. Talker and overall interaction scores range from below 1 to above 4, whereas LLM, TTS, and T2I scores concentrate around 3–4. The broad distributions of the final-video and overall interaction scores indicate that THQA-LW spans diverse perceptual quality levels and provides representative samples for end-to-end interaction assessment; **2)** Mean quality varies substantially across Talkers, with SoulX-LiveAct achieving the highest mean MOS in the model-wise comparison. Because all Talkers process the same selected audio–image pairs, these differences demonstrate that the final generation module strongly shapes the delivered interaction under matched upstream conditions; **3)** Existing Talkers do not fully satisfy the demands of long-form video generation in digital-human interactions. In particular, HunyuanVideo-Avatar cannot generate minute-scale videos. Output-duration limits can terminate an interaction video before the complete answer is delivered, resulting in an incomplete response to the user's question and weakening the overall interaction experience; **4)** As illustrated by the lower panel of Fig. 3, all five Talkers perform better on second-scale videos than on minute-scale videos. In longer streaming generations, errors accumulate and pronounced artifacts and noise often emerge near the end of the response. Short clips therefore capture only part of a Talker's quality profile, making sustained temporal stability essential to interaction quality assessment.

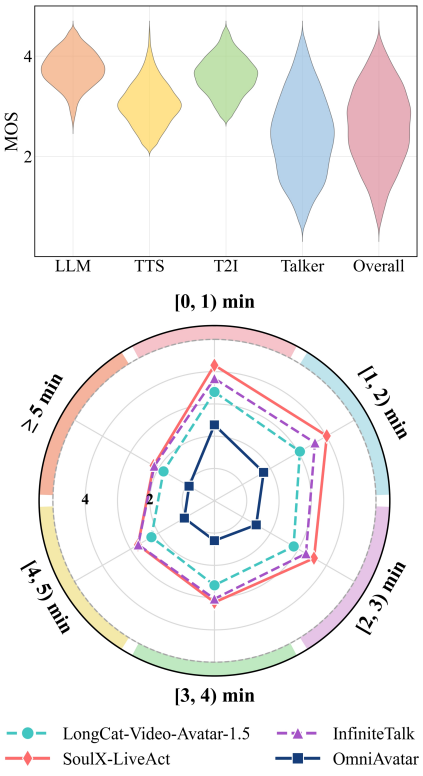

Figure 3: MOS analysis of THQA-LW. The upper panel shows the MOS distributions across the five quality dimensions, while the lower panel compares the MOSs of the Talkers on videos of different durations.

3.5 Distortion Analysis

We analyze the occurrence and co-occurrence of the ten distortion categories recorded during overall interaction quality evaluation. Fig. 4 summarizes their overall and Talker-specific prevalence and presents representative cases of compound quality degradation. The results provide three additional insights: **1)** Nine categories concern Talker output, while one concerns the speech generated by TTS. The high prevalence of Talker-related distortions makes video generation a prominent diagnostic target, while speech quality remains an integral part of the overall interaction experience; **2)** Lip–speech mismatch and unnatural body motion weaken human likeness by disrupting the coordination of speech, facial behavior, and full-body movement. These failures can combine with visual and temporal degradations, diminishing the coherence of the delivered response even when its upstream content and speech are satisfactory; **3)** Most videos exhibit multiple distortion types. Their co-occurrence shows that perceived quality frequently reflects compound failures, supporting an assessment framework that considers multiple quality dimensions rather than a single isolated defect.

3.6 LW-to-Overall Interaction Quality Attribution

We fit a standardized linear regression to quantify the relationship between overall interaction MOS and the MOSs of the four LW modules. Standardizing all variables places them on a common scale, enabling direct comparison of their regression coefficients. The results show that the Talker coefficient is substantially larger than those of LLM, TTS, and T2I. Holding the other LW scores fixed, a one-standard-deviation increase in Talker MOS corresponds to a 0.83783-standard-deviation



Figure 4: Distortion analysis of THQA-LW. The left side reports the overall prevalence of the ten diagnostic distortion categories and their prevalence across the five Talkers. The right side presents three representative cases of the ten distortion categories.

increase in overall interaction MOS. This result identifies Talker quality as the factor most strongly associated with the final interaction experience in the LT3 pipeline. It further demonstrates that the perceptual quality of the generated video plays a dominant role in users' overall judgments, making the Talker module the primary target for improving the quality of current digital-human interactions. The complete fitted relationship is summarized as follows:

$$\begin{aligned} Z(\mathrm{MOS}_{\mathrm{overall}}) = {} & 0.011941\, Z(\mathrm{MOS}_{\mathrm{LLM}}) \\ & + 0.1897\, Z(\mathrm{MOS}_{\mathrm{TTS}}) + 0.09659\, Z(\mathrm{MOS}_{\mathrm{T2I}}) \\ & + 0.83783\, Z(\mathrm{MOS}_{\mathrm{Talker}})\,. \end{aligned} \tag{1}$$

4 Proposed Method

LT3-FAMT predicts the quality of the four LW outputs and the overall interaction from multimodal evidence across the LT3 pipeline. Its inputs are the generated video V, the complete source TTS waveform A^{src}, the reference image I, and the interaction text $C = (Q, R)$, where Q is the user question and R is the LLM answer. As shown in Fig. 5, frozen encoders extract complementary features, which task-specific attention modules aggregate before independent regression heads produce the five quality scores.

4.1 Segment-Level Multimodal Feature Extraction

We divide V at its presentation timestamps into N non-overlapping segments. Each segment V_i covers an interval $[t_i, t_{i+1})$ with a target duration of 60 seconds and retains its corresponding embedded audio. If the final segment lasts at most one second, we merge it into the preceding segment. LT3-FAMT uses the video's embedded audio and the source TTS waveform for different quality judgments. The embedded audio forms part of each Talker-generated segment, while the complete pre-Talker waveform A^{src} provides direct evidence of TTS quality. A frozen Qwen2.5-Omni Thinker Xu et al. (2025a) encodes each audiovisual segment together with the reference image and interaction text. Jina Sturua et al. (2024) extracts a feature from C, DINOv2 Oquab et al. (2023) from I, WavLM Chen et al. (2022) from A^{src}, and VideoMAE Tong et al. (2022) from each video segment. We concatenate their outputs to obtain a multimodal feature $\mathbf{x}_i$ for each segment. The text, reference-image, and source-audio features are shared across segments, whereas the Qwen2.5-Omni and VideoMAE features vary with the segment. All encoders remain frozen, allowing their features to be cached before training.

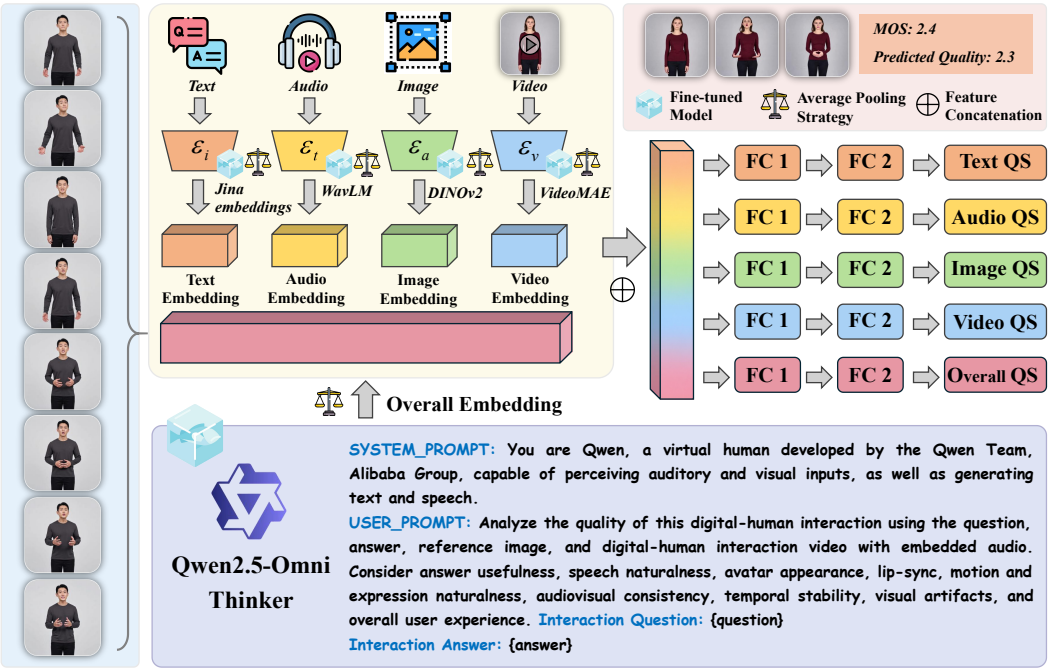

Figure 5: Architecture of LT3-FAMT. Frozen encoders extract features from the interaction text, reference image, source audio, and audiovisual video segments. Task-specific attention modules aggregate the segment features, and independent regression heads predict LW-level and overall interaction quality.

4.2 Task-Specific Attention Pooling and Regression

The five quality targets use different evidence from the interaction. LLM quality concerns the answer, TTS quality concerns the source speech, T2I quality concerns the reference portrait, Talker quality concerns the generated audiovisual response, and overall interaction quality reflects their combined effect. We assign an independent attention module to each task in $\mathcal{K} = \{\mathrm{LLM}, \mathrm{TTS}, \mathrm{T2I}, \mathrm{Talker}, \mathrm{overall}\}$. For task k, a learnable function g_k scores each valid segment feature. Softmax normalization produces attention weights $\alpha_i^{(k)}$, which form the task representation $\mathbf{p}_k$:

$$\alpha_i^{(k)} = \frac{\exp(g_k(\mathbf{x}_i))}{\sum_{j=1}^{N} \exp(g_k(\mathbf{x}_j))}, \qquad \mathbf{p}_k = \sum_{i=1}^{N} \alpha_i^{(k)} \mathbf{x}_i. \tag{2}$$

A validity mask excludes padded segments during batch training. Because the shared text, image, and source-audio features appear in every segment, attention primarily selects among the segment-dependent audiovisual features. An independent regression head maps each $\mathbf{p}_k$ to its quality score. The LLM, TTS, and T2I heads use the MOSs of the corresponding source answer, utterance, and image as targets; the Talker and overall interaction heads use video-level MOSs. We train the five attention modules and regression heads jointly with equally weighted mean squared error losses, while keeping the feature encoders frozen.

5 Experiments

5.1 Experimental Settings

We evaluate LT3-FAMT on five existing digital-human quality assessment datasets—THQA Zhou et al. (2024b), THQA-3D Zhou et al. (2024c), THQA-10K Zhou et al. (2025a), THQA-NTIRE Liu et al. (2025), and THQA-MT Zhou et al. (2025e)—together with the THQA-LW dataset constructed in this work. THQA and THQA-10K evaluate 2D talking heads, THQA-3D evaluates dynamic 3D digital humans, THQA-NTIRE covers hybrid 2D/3D talking heads, THQA-MT evaluates multi-subject Talking Humans. On THQA-LW, we assess LLM, TTS, T2I, Talker, and overall interaction quality, while the external datasets supervise the overall interaction quality branch. We compare

Table 3: Performance on six digital-human quality assessment datasets. Best in **RED**, second in **BLUE**.

Type	Method	THQA				THQA-3D				THQA-10K			
		SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
IQA	BRISQUE Mittal et al. (2012a)	0.4856	0.5970	0.3454	0.8227	0.6749	0.7453	0.5060	0.5717	0.4271	0.4451	0.2993	1.0262
	NIQE Mittal et al. (2012b)	0.0535	0.1643	0.0402	0.9811	0.2243	0.4741	0.1232	0.7707	0.0089	0.0436	0.0051	1.1492
	CPBD Narvekar & Karam (2011)	0.0575	0.0876	0.0376	0.9908	0.2145	0.3136	0.1432	0.8273	0.0553	0.0686	0.0371	1.1476
	IL-NIQE Zhang et al. (2015)	0.0537	0.2160	0.0276	0.9712	0.2293	0.4871	0.1537	0.7600	0.0490	0.0634	0.0286	1.1480
Sync	LSE-C Chung & Zisserman (2017)	0.0056	0.2109	0.0048	0.9723	0.1728	0.2297	0.1355	0.8499	0.0706	0.1634	0.0468	1.1349
	LSE-D Chung & Zisserman (2017)	0.1366	0.2336	0.0855	0.9671	0.0079	0.1054	0.0008	0.8684	0.0580	0.1123	0.0385	1.1431
VQA	VIIDEO Mittal et al. (2015)	0.1777	0.1891	0.1354	0.9595	0.1056	0.2308	0.0721	0.8387	0.1354	0.1782	0.0901	1.1319
	TLVQM Korhonen (2019)	0.0254	0.0355	0.0209	1.0853	0.1887	0.3112	0.1272	0.8240	0.4377	0.4679	0.3070	1.0130
	VIDEVAL Tu et al. (2021a)	0.0317	0.0358	0.0231	1.1916	0.2252	0.3544	0.1556	0.8118	0.3869	0.4147	0.2706	1.0431
	V-BLIINDS Saad et al. (2014)	0.4949	0.6403	0.3533	0.7976	0.5298	0.6412	0.3907	0.6674	0.4740	0.4977	0.3334	0.9941
	RAPIQUE Tu et al. (2021b)	0.1789	0.1908	0.1277	1.0162	0.3748	0.4680	0.2660	0.7643	0.3576	0.3846	0.2490	1.0579
	SimpVQA Sun et al. (2022)	0.6800	0.7592	0.5052	0.6361	0.6321	0.7258	0.4717	0.5983	0.7775	0.8039	0.5931	0.6832
	VSFA Li et al. (2019)	0.7601	0.8106	0.5830	0.5966	0.7463	0.7811	0.5596	0.5726	0.7537	0.7754	0.5726	0.7343
	FAST-VQA Wu et al. (2022)	0.6389	0.7441	0.4677	0.6983	0.7778	0.7984	0.5964	0.5503	0.7351	0.7542	0.5519	0.8026
	BVQA Li et al. (2022)	0.7287	0.7985	0.5549	0.6094	0.7871	0.8298	0.6081	0.5983	0.6335	0.7405	0.4522	0.7634
THQA	MI3S Zhou et al. (2025d)	0.7414	0.8207	0.5658	0.5683	0.7671	0.8135	0.6090	0.4637	0.7789	0.8106	0.6032	0.6522
	Xu *et al.* Xu et al. (2025b)	**0.8167**	**0.8669**	**0.6422**	0.5115	**0.8515**	**0.8617**	**0.6768**	**0.4542**	**0.8267**	**0.8614**	**0.6539**	**0.6070**
	Su *et al.* Su et al. (2025)	0.8123	0.8580	0.6397	0.5127	0.7838	0.8226	0.6371	0.4660	0.7951	0.8266	0.6176	0.6477
	FSCD Zhou et al. (2025a)	0.7812	0.8409	0.5951	**0.5055**	0.8235	0.8505	0.6463	0.4577	0.8066	0.8322	0.6228	0.6333
	LT3-FAMT (Ours)	**0.8310**	**0.8946**	**0.6857**	**0.4893**	**0.8671**	**0.8897**	**0.7096**	**0.4313**	**0.8785**	**0.8863**	**0.7026**	**0.5755**

Type	Method	THQA-NTIRE				THQA-MT				THQA-LW			
		SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
IQA	BRISQUE Mittal et al. (2012a)	0.4474	0.4842	0.3261	0.9321	0.2051	0.2778	0.1387	1.5398	0.2734	0.3955	0.1875	1.4072
	NIQE Mittal et al. (2012b)	0.0454	0.1004	0.0343	1.0231	0.1450	0.3639	0.0944	1.4998	0.4126	0.6624	0.2729	0.9265
	CPBD Narvekar & Karam (2011)	0.0535	0.0745	0.0470	1.1525	0.1127	0.1608	0.0731	1.1634	0.1826	0.2100	0.1241	1.2091
	IL-NIQE Zhang et al. (2015)	0.0598	0.1015	0.0372	1.1097	0.1214	0.4440	0.0731	1.6221	0.1747	0.2874	0.1159	1.1845
Sync	LSE-C Chung & Zisserman (2017)	0.0820	0.1753	0.0666	1.0870	0.3379	0.3887	0.2260	1.3017	0.2530	0.2112	0.1694	1.1305
	LSE-D Chung & Zisserman (2017)	0.0658	0.1349	0.0445	1.1287	0.2912	0.2031	0.1960	1.0157	0.1848	0.3396	0.1234	1.0879
VQA	VIIDEO Mittal et al. (2015)	0.1554	0.1982	0.1091	1.1025	0.0642	0.0691	0.0433	1.1040	0.0819	0.2477	0.0469	1.1982
	TLVQM Korhonen (2019)	0.4026	0.4348	0.2805	1.0280	0.3506	0.4325	0.2462	0.9961	0.3201	0.4071	0.2274	1.3936
	VIDEVAL Tu et al. (2021a)	0.3530	0.3784	0.2641	1.0418	0.4061	0.4586	0.2849	0.9822	0.1099	0.2570	0.0759	1.2477
	V-BLIINDS Saad et al. (2014)	0.4860	0.5329	0.3319	0.9329	0.5102	0.5362	0.3792	0.9371	0.4216	0.4156	0.2947	1.1080
	RAPIQUE Tu et al. (2021b)	0.3392	0.3920	0.2482	1.0395	0.1638	0.1793	0.1120	1.0899	0.3845	0.5464	0.2589	1.2189
	SimpVQA Sun et al. (2022)	0.7508	0.7912	0.5865	0.6740	0.5257	0.5313	0.3730	0.9424	0.5013	0.6226	0.3522	1.1119
	VSFA Li et al. (2019)	0.7586	0.7794	0.5699	0.7254	0.4548	0.4777	0.3190	0.9861	0.4476	0.4531	0.3110	1.2464
	FAST-VQA Wu et al. (2022)	0.7350	0.7449	0.5497	0.7795	0.4926	0.5074	0.3502	1.1043	0.5482	0.7191	0.3833	0.9592
	BVQA Li et al. (2022)	0.6497	0.7485	0.4712	0.7556	0.4639	0.4908	0.3270	1.0703	0.4850	0.6021	0.3476	1.1417
THQA	MI3S Zhou et al. (2025d)	0.7795	0.8205	0.6084	0.6336	0.7509	0.7820	0.6244	0.7233	0.6328	0.6527	0.4652	1.0666
	Xu *et al.* Xu et al. (2025b)	**0.8105**	**0.8676**	**0.6419**	**0.5949**	0.7495	0.7664	0.6058	0.7550	**0.6964**	0.7080	0.5049	0.8790
	Su *et al.* Su et al. (2025)	0.8036	0.8453	0.6285	0.6268	**0.7798**	0.7931	0.6497	0.6991	0.6912	**0.7786**	**0.5496**	**0.7572**
	FSCD Zhou et al. (2025a)	0.8102	0.8314	0.6271	0.6018	0.7769	**0.7944**	**0.6564**	**0.6875**	0.5927	0.6532	0.4715	0.9173
	LT3-FAMT (Ours)	**0.8444**	**0.8842**	**0.6776**	**0.5382**	**0.8450**	**0.8534**	**0.6940**	**0.6514**	**0.7512**	**0.8253**	**0.6155**	**0.6966**

LT3-FAMT with 19 representative objective quality assessment methods from four categories: Image Quality Assessment (IQA), synchronization-based assessment (Sync), Video Quality Assessment (VQA), and specialized Talking Human Quality Assessment (THQA). For fair comparison, all competing methods use their official implementations with default parameter settings.

The task-specific attention modules and independent score heads are trained for 30 epochs with AdamW Loshchilov & Hutter (2017), a learning rate of 10^{-4}, weight decay of 10^{-4}, and cosine annealing Loshchilov & Hutter (2016). Each score head is implemented as a two-layer MLP with GELU activation Hendrycks & Gimpel (2016) and a dropout rate of 0.1. For external benchmarks, Qwen2.5-Omni and VideoMAE encode the final video, and Qwen2.5-Omni processes embedded audio when available. Zero vectors fill unavailable text, reference-image, and source-TTS feature inputs. To ensure robust and reliable evaluation, we adopt five-fold cross-validation with mutually exclusive data splits. We quantify performance using four quality-assessment metrics that are widely adopted in the literature: SRCC, PLCC, KRCC, and RMSE. The correlation metrics, including SRCC, PLCC, and KRCC, range from -1 to 1, where higher values indicate stronger consistency with human perceptual judgments. Lower RMSE indicates more accurate quality prediction.

5.2 Performance Analysis

The quantitative performance of all competing methods across the six datasets is reported in Table 3. The results lead to three key observations: **1)** LT3-FAMT achieves the best performance across all six datasets in terms of SRCC, PLCC, KRCC, and RMSE. These datasets cover conventional 2D talking heads, dynamic 3D digital humans, hybrid 2D/3D content, multi-subject Talking Humans, and the LT3 interaction pipeline, demonstrating the strong generalization ability of LT3-FAMT across diverse digital-human quality assessment settings; **2)** On THQA-LW, LT3-FAMT

Table 4: Ablation results on THQA-LW. *w/o* stands for *without*. Best in **RED**.

Performance	*w/o* Jina	*w/o* WavLM	*w/o* DINOv2	*w/o* VideoMAE	*w/o* Qwen2.5-Omni	**Ours**
SRCC↑	0.6638	0.7126	0.6458	0.7286	0.2903	**0.7512**
PLCC↑	0.6783	0.7932	0.6765	0.7841	0.3163	**0.8253**
KRCC↑	0.5254	0.5784	0.5037	0.5998	0.1859	**0.6155**
RMSE↓	0.8129	0.7532	0.8349	0.7487	1.1936	**0.6966**

achieves an SRCC of 0.7512, exceeding the second-ranked method by approximately 7.9%. Unlike conventional benchmarks that mainly evaluate the final audiovisual output, THQA-LW requires assessment of the generated answer, source speech, reference portrait, Talker output, and their combined effect on the overall interaction quality. LT3-FAMT addresses this multidimensional setting by integrating frozen multimodal representations, task-specific temporal aggregation, and an independent regression head for each quality target; **3)** Existing IQA, VQA, synchronization, and THQA methods exhibit substantially weaker performance on THQA-LW because they primarily characterize isolated visual, temporal, or audiovisual properties of the final media. Their limited performance highlights the distinction between interaction quality assessment and conventional digital-human video quality assessment, confirming the need for LW-aware multimodal assessment throughout the complete interaction pipeline.

5.3 Ablation Experiments

We evaluate the contribution of the five frozen feature extractors using the complete LT3-FAMT and five leave-one-out variants. In each variant, the corresponding feature block is replaced with an all-zero vector while the remaining architecture and training settings are unchanged. Table 4 reports the performance of each variant on overall interaction quality. The ablation results yield two key findings: **1)** Removing any of the five feature extractors degrades the overall performance of LT3-FAMT, demonstrating the effectiveness of its full-modality feature extraction design. The complementary contributions of these encoders show that digital-human interaction quality assessment benefits from jointly modeling audiovisual and contextual evidence when evaluating both the Talker output and the complete interaction experience; **2)** Among the five feature extractors, removing Qwen2.5-Omni causes the largest performance degradation, whereas removing VideoMAE produces the smallest overall effect. Qwen2.5-Omni jointly encodes the generated video, its embedded audio, the reference image, and the interaction text, while the video modality provides its primary quality evidence. VideoMAE supplies additional video-specific representations that complement the Qwen2.5-Omni features, thereby improving the effectiveness and robustness of LT3-FAMT.

6 Conclusion

Interactive digital humans are increasingly used in education, healthcare, information services, media communication, and cultural presentation, making the quality of human–machine interaction central to the user experience. These systems are commonly constructed as cascades of specialized LW modules, each of which can introduce quality degradations that propagate through the pipeline and reduce the overall interaction quality. To address this system-level assessment problem, we select four core and widely used LW modules—LLM, TTS, T2I, and Talker—to construct the LT3 pipeline for studying digital-human interaction quality. We further introduce THQA-LW, a large-scale dataset containing 10,000 single-turn interaction videos. Through controlled subjective experiments involving multiple participants, THQA-LW provides separate quality ratings for each LW output and the overall interaction experience, identifies ten common distortion types, and supports analysis of how different LW modules affect the final interaction quality. Finally, we propose LT3-FAMT, a five-task multimodal quality predictor for efficient objective assessment of the LT3 pipeline. Experimental results show that LT3-FAMT achieves state-of-the-art performance across six digital-human quality assessment datasets and exhibits strong agreement with human perceptual judgments. LT3-FAMT therefore provides an effective tool for automatic quality prediction, LW-aware diagnosis, and systematic optimization of interactive digital-human systems, supporting better interaction experiences and broader applications.

AI USE STATEMENT

Generative AI models are integral to the experimental pipeline and the construction of THQA-LW. As described in Sections 3.1 and 3.2 and Table 2, five large language models were used to generate 80 answers to 16 single-turn questions; five text-to-speech models synthesized 1,600 utterances using four reference voices; five text-to-image models generated 20 avatar reference images from predefined prompts; and five talking-human generators produced the 10,000 interaction videos in THQA-LW. Qwen2.5-Omni was additionally employed as a frozen multimodal feature encoder in LT3-FAMT, as described in Section 4.1. All ground-truth quality labels were collected from human participants rather than generated by AI. The authors reviewed the generated research artifacts and verified the experimental data, analyses, and reported results. The authors take full responsibility for the content, claims, and artifacts presented in this work.

ETHICS STATEMENT

This study involved 40 participants, comprising 20 male and 20 female observers, who provided perceptual quality ratings in a controlled laboratory experiment. The protocol followed ITU-R BT.500-13. Participants were unaware of model identities, stimulus presentation order was randomized, and a 30-minute training session was provided before formal evaluation. To reduce fatigue, participants rested for 20 minutes between consecutive subsets and completed no more than three subsets per day. No directly identifying participant information is included in THQA-LW. The dataset consists of AI-generated answers, speech, portraits, and talking-human videos. Such content may reflect social biases present in the underlying models and could be misused for impersonation, misinformation, or deceptive media. The dataset and associated models are intended exclusively for research on perceptual quality assessment and diagnostic evaluation. We affirm that this work was carried out in accordance with established standards of academic integrity and with the aim of promoting responsible and socially beneficial technological development. Any application of this research that violates legal requirements, ethical norms, or the rights and interests of individuals is fundamentally inconsistent with the purpose and principles of this work.

REPRODUCIBILITY STATEMENT

We support reproducibility by documenting each stage of the data construction, subjective study, model design, and evaluation. Sections 3.1– 3.2 and Table 2 specify the LT3 pipeline, the models used for each learnware, the combinatorial data-generation process, the constrained sampling procedure, and the provenance recorded for the 10,000 videos in THQA-LW. Section 3.3 and Appendix A describe the subjective evaluation protocol, presentation setup, observer screening, score normalization, MOS aggregation, and distortion-label construction. Sections 4.1–4.2 and Figure 5 provide the model inputs, frozen feature encoders, video segmentation rule, task-specific attention pooling, regression heads, and multitask training objective of LT3-FAMT. Section 5.1 reports the benchmark datasets, optimization hyperparameters, five-fold cross-validation protocol, evaluation metrics, and implementation settings for the compared methods. These specifications support reconstruction and independent evaluation of the proposed dataset and assessment framework.

REFERENCES

Alibaba Cloud. Qwen3.7-Max Model Information, 2026a. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-7-max. Accessed: 2026-09-24.

Alibaba Cloud. Wan2.7 Image Generation and Editing, 2026b. URL https://www.alibabacloud.com/help/en/model-studio/wan-image-generation-and-editing-api-reference. Accessed: 2026-09-24.

Anthropic. Claude Fable 5 and Claude Mythos 5, 2026. URL https://www.anthropic.com/news/claude-fable-5-mythos-5. Accessed: 2026-09-24.

ByteDance Seed Team. Seed2.1 Officially Released: Advancing AI Productivity, 2026a. URL https://seed.bytedance.com/en/blog/seed2-1-officially-released-advancing-ai-productivity. Accessed: 2026-09-24.

ByteDance Seed Team. Seedream 5.0 lite, 2026b. URL https://seed.bytedance.com/zh/blog/deeper-thinking-more-accurate-generation-introducing-seedream-5-0-lite. Accessed: 2026-09-24.

Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4945–4954. IEEE, 2021.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4960–4964. IEEE, 2016.

Debaleena Chattopadhyay, Tengteng Ma, Hasti Sharifi, and Pamela Martyn-Nemeth. Computer-controlled virtual humans in patient-facing systems: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 22(7):e18839, 2020. doi: 10.2196/18839.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

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.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision 2016 Workshops*, pp. 251–263, 2017.

Jiayan Cui, Zhihan Yang, Naihan Li, Jiankun Tian, Xingyu Ma, Yi Zhang, Guangyu Chen, Runxuan Yang, Yuqing Cheng, Yizhi Zhou, et al. Glm-tts technical report. *arXiv preprint arXiv:2512.14291*, 2025.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

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.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, et al. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*, 2026.

Google. Gemini 3.1 Flash Image, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-image. Accessed: 2026-09-24.

Google. Gemini 3.5 Flash, 2026b. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash. Accessed: 2026-09-24.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Peirong Guo, Qi Zhang, Chunwei Tian, Wanli Xue, and Xiaocheng Feng. Digital human techniques for education reform. In *Proceedings of the 2024 7th International Conference on Educational Technology Management*, pp. 173–178, 2024.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7543–7552. IEEE, 2018.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

ITU Radiocommunication Sector. Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-13, International Telecommunication Union, Geneva, Switzerland, January 2012. URL https://www.itu.int/rec/R-REC-BT.500-13-201201-S.

A-Yeong Kim, Hyun-Je Song, and Seong-Bae Park. A two-step neural dialog state tracker for task-oriented dialog processing. *Computational intelligence and neuroscience*, 2018(1):5798684, 2018.

Jihyun Kim, Kun Xu, and Kelly Merrill Jr. Man vs. machine: Human responses to an ai newscaster and the role of social presence. *The Social Science Journal*, pp. 1–13, 2022.

Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.

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.

Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. Aigiqa-20k: A large database for ai-generated image quality assessment. *arXiv preprint arXiv:2404.03407*, 2(3):5, 2024.

Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM International Conference on Multimedia*, pp. 2351–2359, 2019.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. *URL https://arxiv. org/abs/2109.07958*, 1, 2021.

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*, pp. 1389–1402, 2025.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

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, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012b.

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.

Niranjan 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.

OpenAI. Introducing GPT-5.5, 2026a. URL https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-09-24.

OpenAI. GPT-Image-2 Model, 2026b. URL https://developers.openai.com/api/docs/models/gpt-image-2. Accessed: 2026-09-24.

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

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.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, pp. 484–492, 2020.

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*, pp. 10684–10695, June 2022.

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.

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.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*, 2024.

Mengjing Su, Yi Wang, Tuo Chen, Chunxiao Li, Shuaiyu Zhao, Jiaxin Wen, Chuyi Lin, Sitong Liu, Ningxin Chu, and Yu Zhou. Quality assessment for talking head videos via multi-modal feature representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1414–1420, 2025.

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.

Meituan LongCat Team, Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, et al. Longcat-video-avatar 1.5 technical report. *arXiv preprint arXiv:2605.26486*, 2026.

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

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, 2021a.

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, 2021b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Huaxuan Wang, Huimin Wang, Ruiyu Zhang, Yingjie Li, and Yitao Duan. Confucius4-tts: Transcript-free cross-lingual zero-shot tts with a learnable speaker encoder. *arXiv preprint arXiv:2608.11650*, 2026a.

Zhongjian Wang, Peng Zhang, Jinwei Qi, Sheng Xu, Bang Zhang, et al. Omnitalker: One-shot real-time text-driven talking audio-video generation with multimodal style mimicking. *Advances in Neural Information Processing Systems*, 38:21834–21862, 2026b.

Farong Wen, Yingjie Zhou, Zicheng Zhang, Xiaohong Liu, Jia Wang, Jiezhang Cao, Yu Wang, and Guangtao Zhai. A light-aware quality assessment method for relighted human heads based on multi-task learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 0–0. Springer, 2025.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European Conference on Computer Vision*, pp. 538–554. Springer, 2022.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Li Xu, Yingjie Zhou, Sitong Liu, Farong Wen, Yu Zhou, Xiaohong Liu, Jie Guo, Yu Wang, and Jiezhang Cao. Facial quality assessment of digital humans: A dual-branch framework integrating morphological harmony and expressive coordination. *Displays*, pp. 103221, 2025b.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

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.

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*, pp. 8652–8661, 2023a.

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*, 2023b.

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*, pp. 2519–2524, 2023c.

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*, pp. 1–5, 2023d.

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*, pp. 1365–1369, 2023e.

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.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Dingcheng Zhen, Xu Zheng, Ruixin Zhang, Zhiqi Jiang, Yichao Yan, Ming Tao, and Shunshun Yin. Soulx-liveact: Towards hour-scale real-time human animation with neighbor forcing and convkv memory. *arXiv preprint arXiv:2603.11746*, 2026.

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*, pp. 36–40, 2023.

Yingjie Zhou, Zicheng Zhang, Jiezhang 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*, 2024a.

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*, pp. 15–21, 2024b.

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*, pp. 6033–6042, 2024c.

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*, 2024d.

Yingjie Zhou, Jiezhang Cao, Zicheng Zhang, Farong Wen, Yanwei Jiang, Jun Jia, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Who is a better talker: Subjective and objective quality assessment for ai-generated talking heads. *arXiv preprint arXiv:2507.23343*, 2025a.

Yingjie Zhou, Jing Wan, Sitong Liu, Yinghan Xia, Zhixiang Lu, Farong Wen, Zicheng Zhang, Yu Wang, Yu Zhou, Xiaohong Liu, et al. Cdhqa: A quality assessment database for conversational digital human. In *International Conference on Image and Graphics*, pp. 15–26, 2025b.

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*, 2025c.

Yingjie Zhou, Zicheng Zhang, Sijing Wu, Jun Jia, Yanwei Jiang, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Mi3s: A multimodal large language model assisted quality assessment framework for ai-generated talking heads. *Information Processing & Management*, 63(1):104321, 2025d.

Yingjie Zhou, Xilei Zhu, Siyu Ren, Ziyi Zhao, Ziwen Wang, Farong Wen, Yu Zhou, Jiezhang Cao, Xiongkuo Min, Fengjiao Chen, et al. Evaltalker: Learning to evaluate real-portrait-driven multi-subject talking humans. *arXiv preprint arXiv:2512.01340*, 2025e.

Zhi-Hua Zhou. Learnware: on the future of machine learning. *Frontiers Comput. Sci.*, 10(4):589–590, 2016.

APPENDIX

A Details of the Subjective Experiment and Data Processing **16**

A.1 Experimental Protocol 16

A.2 Subjective Data Processing 17

A DETAILS OF THE SUBJECTIVE EXPERIMENT AND DATA PROCESSING

A.1 EXPERIMENTAL PROTOCOL

The subjective study evaluates the output of each LW module and the overall interaction quality as separate dimensions. Forty participants, comprising 20 male and 20 female observers, take part in a controlled laboratory experiment. The procedure follows ITU-R BT.500-13 ITU Radiocommunication Sector (2012) and uses an Absolute Category Rating (ACR) protocol. Participants are unaware of model identities, and the presentation order is randomized.

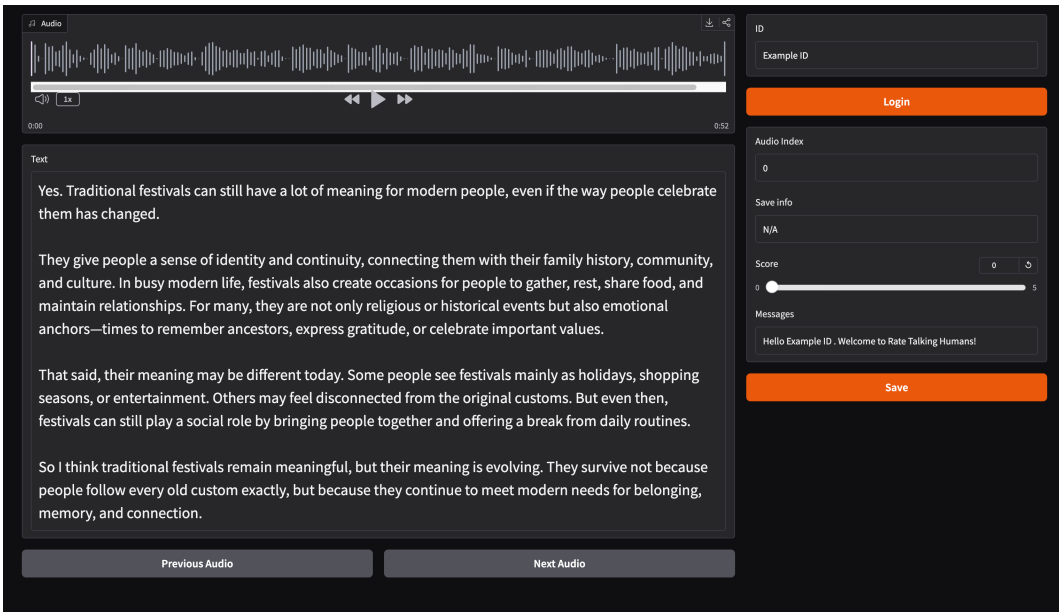

Figure 6: Web-based interface used for subjective evaluation. The interface presents the materials required by the current evaluation dimension and records the observer's quality score. This example shows the layout for an item containing textual and audio content.

All ratings are collected through the unified web-based annotation interface shown in Figure 6. For each evaluation dimension, the interface presents the corresponding materials, including text, audio, images, or videos. The displayed content is updated according to the current item and evaluation dimension. Each evaluation dimension presents the material required for its corresponding judgment. For LLM quality, observers read the question and answer and rate relevance, expression, and expected helpfulness. For TTS quality, they listen to the source utterance and rate clarity, naturalness, prosody, emotional expression, and overall listening quality. For T2I quality, they view the generated portrait and rate its visual quality and naturalness. For Talker quality, they watch the generated video with audio and assess lip synchronization, expression and motion, temporal stability, audiovisual consistency, and visual realism with respect to the source image and speech. For overall interaction quality, they receive the question, answer, and complete audiovisual response and rate the interaction as a whole. During this evaluation, observers also record the presence of ten diagnostic distortion categories.

A.2 SUBJECTIVE DATA PROCESSING

The experiment produces $40 \times 21{,}700 = 868{,}000$ raw scalar ratings. During the overall interaction quality evaluation, participants additionally produce $40 \times 10{,}000 = 400{,}000$ ten-dimensional binary distortion vectors, corresponding to 4,000,000 binary decisions. Following the THQA-10K protocol Zhou et al. (2025a), we normalize the quality ratings separately within each evaluation dimension.

Let $d \in \{\mathrm{LLM}, \mathrm{TTS}, \mathrm{T2I}, \mathrm{Talker}, \mathrm{overall}\}$ denote a quality dimension, and let $q_{ij}^{(d)}$ denote participant i's raw rating of item j in dimension d. We calculate the mean $\mu_i^{(d)}$ and standard deviation $\sigma_i^{(d)}$ of participant i's ratings within that dimension. The standardized rating is

$$z_{ij}^{(d)} = \frac{q_{ij}^{(d)} - \mu_i^{(d)}}{\sigma_i^{(d)}} \,. \tag{3}$$

We apply the observer-screening procedure specified in ITU-R BT.500-13 ITU Radiocommunication Sector (2012) separately to each dimension and exclude ratings from rejected observers. The remaining standardized ratings are linearly rescaled to $[0, 5]$ within each dimension and averaged to obtain MOSs. LLM, TTS, and T2I MOSs are calculated for each source answer, source utterance, and reference image, respectively, and are assigned to the videos generated from those sources. Talker and overall interaction MOSs are calculated separately for each video. These five MOSs form the multidimensional quality label of each interaction video. For every diagnostic distortion category, a video receives a positive label when more than half of its valid observers identify that distortion.