Towards Interactive Intelligence for Digital Humans

Yiyi Cai[1], **Xuangeng Chu**[1,2], **Xiwei Gao**[1], **Sitong Gong**[1], **Yifei Huang**[1,2], **Caixin Kang**[1,2], **Kunhang Li**[1,2], **Haiyang Liu**[1], **Ruicong Liu**[1,2], **Yun Liu**[1,4], **Dianwen NG**[1], **Zixiong Su**[1,2], **Erwin Wu**[1,3], **Yuhan Wu**[1,2], **Dingkun Yan**[1], **Tianyu Yan**[1], **Chang Zeng**[1], **Bo Zheng**[1], **You Zhou**[1]

[1]Shanda AI Research Tokyo [2]The University of Tokyo [3]Institute of Science Tokyo [4]National Institute of Informatics

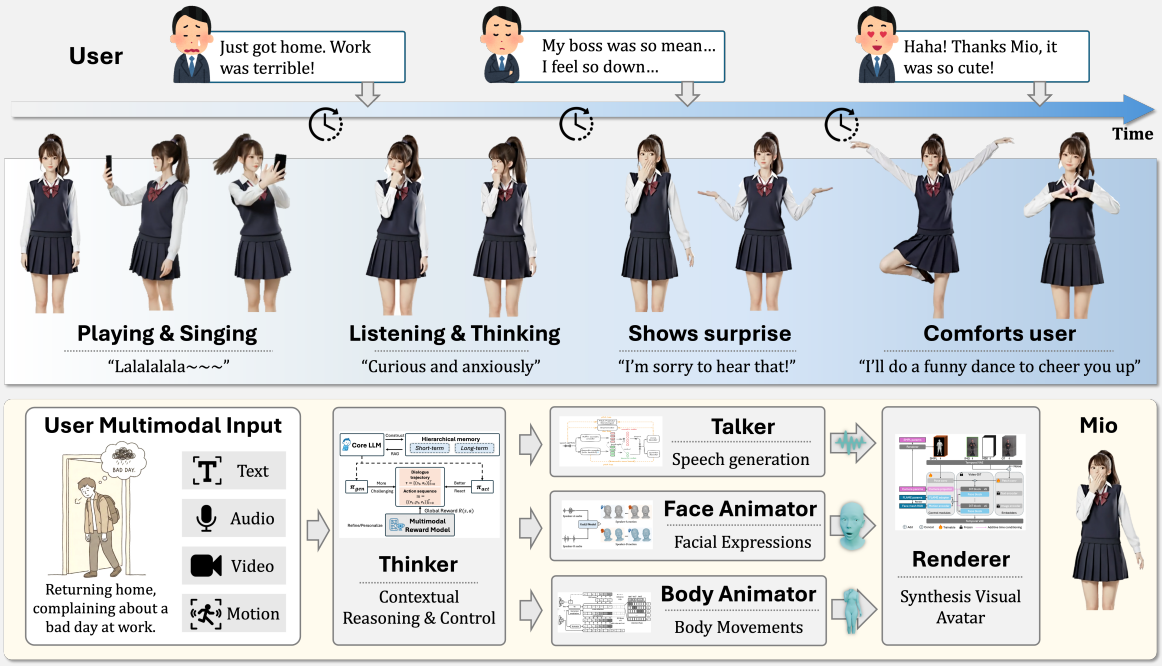

Figure 1 Towards interactive intelligence with Mio. The timeline (top) demonstrates a live interaction where Mio adaptively transitions from singing to active listening and comforting in response to user input. The architecture (bottom) features a Thinker that orchestrates multimodal generation via the Talker, Face Animator, Body Animator, and Renderer modules.

We introduce *Interactive Intelligence*, a novel paradigm of digital human that is capable of personality-aligned expression, adaptive interaction, and self-evolution. To realize this, we present **Mio** (Multimodal Interactive Omni-Avatar), an end-to-end framework composed of five specialized modules: **Thinker**, **Talker**, **Face Animator**, **Body Animator**, and **Renderer**. This unified architecture integrates cognitive reasoning with real-time multimodal embodiment to enable fluid, consistent interaction. Furthermore, we establish a new benchmark to rigorously evaluate the capabilities of interactive intelligence. Extensive experiments demonstrate that our framework achieves superior performance compared to state-of-the-art methods across all evaluated dimensions. Together, these contributions move digital humans beyond superficial imitation toward intelligent interaction.

GitHub: https://shandaai.github.io/project_mio_page/

Date: December 15, 2025

shanda AI Research Tokyo

1 Introduction

Most existing digital humans remain primarily imitative, reproducing surface patterns of behavior without true understanding of interaction logic. While visual fidelity has greatly improved in recent years [154, 163], a fundamental gap remains in enabling these avatars to function as responsive, logic-driven entities. To bridge this gap, we introduce *Interactive Intelligence*, a novel paradigm of digital humans that interact seamlessly with users, while possessing personality-aligned expression, adaptive responsiveness, and self-evolution capabilities. This paradigm transforms the digital human from a passive playback system into an embodied agent capable of coherent multimodal engagement within a dynamic narrative context [93].

Current approaches to digital human creation generally fall into two categories: traditional CG pipelines and generative model-based workflows. Traditional CG methods can offer precise control but are hindered by prohibitive production times and reliance on labor-intensive manual processes. On the other hand, workflows utilizing general-purpose multimodal generative models leverage massive audiovisual corpora to accelerate production but remain fundamentally limited to offline generation [5, 13, 143, 63, 19]. Consequently, the resulting characters are primarily imitative rather than autonomous, reproducing surface behavioral patterns without genuine interaction logic. This leaves them incapable of real-time responsiveness and prone to failures in maintaining consistent identity and behavioral coherence over long-term interactions [121, 96].

Constructing an end-to-end interactive system presents unique challenges across multiple modalities. In response generation, standard LLMs often violate narrative causality (give spoilers) and drift out of persona during extended interactions [126, 57]. In speech synthesis, existing TTS models lacks efficiently discrete speech representations, hindering the low-latency generation required for fluid conversation [29, 123, 104, 43, 9]. In facial animation, a critical issue is the "zombie-face" phenomenon, where digital avatars exhibit stiffness and lack natural listening behaviors when not speaking, breaking user immersion [98, 1, 114]. Furthermore, generating coherent full-body motion remains difficult; autoregressive models often suffer from error accumulation, while standard diffusion models are computationally prohibitive for real-time streaming [134, 155, 117, 156, 157]. Finally, rendering these motions into a visual avatar requires maintaining strict multi-view identity consistency, which is often compromised in image-driven diffusion approaches [140, 106, 83, 92].

To address these challenges, we introduce Multimodal Interactive Omni-Avatar **Mio**, a comprehensive framework that models digital humans as autonomous agents with interactive intelligence. We propose a cascading paradigm composed of five specialized modules: *Thinker, Talker, Face Animator, Body Animator, and Renderer*. The *Thinker* serves as the cognitive core, utilizing a hierarchical memory system and diegetic knowledge graph to ensure narrative consistency and personality fidelity. The *Talker* leverages high-fidelity speech representations and produce clear and expressive voice that is well-aligned with the context. The *Face Animator* introduces a unified listening-speaking framework to generate responsive facial dynamics even during silence. The *Body Animator* utilizes a novel streaming diffusion forcing strategy to convert text instructions into physically plausible body motions in real time. Finally, the *Renderer* leverages a parameter-based diffusion transformer to synthesize the visual avatar with precise control over facial and body dynamics while ensuring multi-view consistency.

Extensive quantitative and qualitative experiments demonstrate the superiority of our approach. Our Talker module outperforms exisiting speech tokenizers and auto-regressive TTS models in speech generation metrics with a balanced multilingual capability. The Facial Animator significantly outperforms baselines in listening naturalness, with over 90% of users preferring our results over existing methods like DualTalk. The Body Animator achieves state-of-the-art motion quality (FID 0.057) on HumanML3D while maintaining the lowest latency and highest smoothness in streaming benchmarks [44]. Furthermore, our Thinker module outperforms general-purpose models like GPT-4o in persona fidelity metrics [90], and the Renderer demonstrates superior multi-view identity preservation compared to recent video diffusion models.

To summarize, Mio represents a fundamental step forward in the evolution of digital humans, harmonizing the often disparate fields of cognitive reasoning and real-time animation. By demonstrating that autonomous agents can possess both narrative depth and physical fluidity, our work paves the way for next-generation applications in virtual companionship, interactive storytelling, and immersive gaming. We believe that Interactive Intelligence will become the defining standard for future avatars, shifting the focus from static appearance to dynamic, meaningful engagement. To support this transition and encourage further exploration

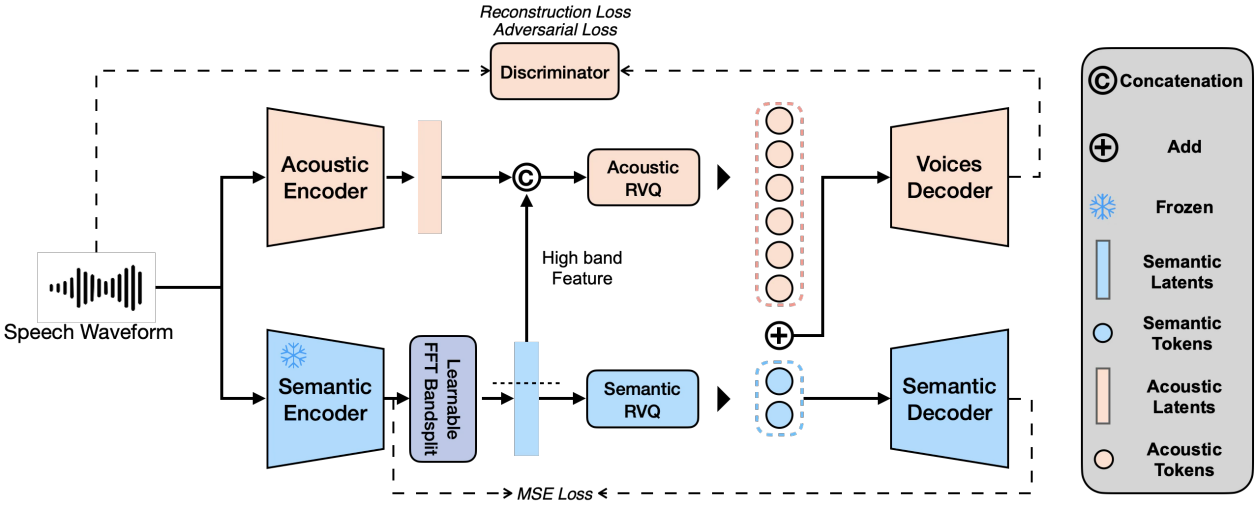

Figure 2 The architecture of Kodama-Tokenizer and training objectives.

within the research community, we make our full codebase, pre-trained models, and the proposed evaluation benchmark publicly available at https://shandaai.github.io/project_mio_page/.x

2 Talker

The Talker module acts as the speech synthesis engine for Mio. Its primary function is to convert textual output from the Thinker module into natural, high-fidelity speech. Our approach begins by learning efficient discrete speech representations, which are subsequently aligned with semantic content via an auto-regressive (AR) framework. To support real-time, expressive conversational interactions, we prioritize both robust context understanding and generation efficiency.

Crucially, the architecture promotes the disentanglement of semantic and acoustic information through the use of band-splitting and a semantic teacher. This design explicitly separates "what is being said" from "how it sounds," enabling targeted compression for each information type rather than forcing a single representation to resolve both roles simultaneously.

2.1 Kodama Audio Tokenizer

2.1.1 Challenges and Motivation

Neural audio codecs have recently evolved from tools for efficient waveform compression into the de-facto *audio tokenizers* that connect continuous speech with discrete, latest LLM-ready representations [30, 146, 147, 43] all follow a similar recipe: compress speech into low-bitrate token streams that preserve both semantic content and acoustic detail, and then train large language models (LLMs) or speech LLMs to operate directly on these tokens. This paradigm has unlocked highly expressive TTS and multi-modal dialogue systems, but it still faces important trade-offs between compression ratio, reconstruction quality, semantic–acoustic disentanglement, and streaming latency.

Most existing codecs operate at 25–50 Hz frame rates and ∼1 kbps, which keeps autoregressive sequences relatively long for LLMs. Pushing bitrates lower tends to harm perceptual quality or speaker similarity, while unified encoders and codebooks often entangle semantic and acoustic objectives, making both harder to optimize. In addition, many tokenizers are tuned primarily for offline TTS MOS benchmarks rather than the needs of real-time, dialogue-centric speech LLMs, where shorter sequences, stronger robustness, and easy alignment with text can matter more than absolute reconstruction scores. Therefore, our goal is first to build a speech tokenizer that prioritize low frame rate and bitrate, and by separating semantic and acoustic information,

2.1.2 Method

We propose **Kodama-Tokenizer**, a neural audio tokenizer designed explicitly for this speech-LLM regime. As shown in 2, the architecture of the tokenizer module comprises of four components: a semantic encoder, an acoustic encoder, Residual Vector Quantization (RVQ) [72] modules, and a vocoder. The semantic encoder takes 16 kHz audio waveforms as input, and we borrow the weights from a pretrained W2v-BERT 2.0 [25] model to extract SSL features as a semantic teacher. The acoustic encoder is based on the Muffin [89] codec, which operates on raw waveforms, beginning with a 1D conv that lifts audio into a low-channel representation, followed by four strided Conv1d downsampling stages that progressively compress time while doubling channels up to 512. At each scale, three parallel residual Conv1d blocks with multi-dilation and Snake [166] activations refine features, whose outputs are averaged and added back. The final output from acoustic encoder is a 512-dimensional latent representation at 50Hz. A learnable FFT band-split decomposes the semantic embeddings, where the low-band track is quantized into two RVQ codebooks, while the high-bank track is merged into the acoustic stream. The acoustic feature further uses a 6-codebook RVQ for quantization. Both streams are downsampled to 12.5 Hz before the RVQ module, with semantic features encoded into two codebooks and acoustic features encoded into six codebooks. We leverage a Vocos [110] decoder to directly decode the combined embeddings to waveforms, where our main learning objective is to minimize the adversarial loss, derived from the components of the HiFiGAN [69], including a multi-period discriminator (MPD) and a multi-scale STFT discriminator (MSD).

Similar to other models [69], we adopt a weighted combination of multiple loss terms as the final loss function formulated by Eq. (1) and Eq. (2) to supervise the training process of our Kodama-Tokenizer.

$$\mathcal{L}_D = \mathcal{L}_{adv}(D;G), \tag{1}$$

$$\mathcal{L}_G = \lambda_1 * \mathcal{L}_{adv}(G;D) + \lambda_2 * \mathcal{L}_{mel} + \lambda_3 * \mathcal{L}_{pitch} + \lambda_4 * \mathcal{L}_{fm} + \lambda_5 * \mathcal{L}_{RVQ}, \tag{2}$$

where $\mathcal{L}_{adv}$, $\mathcal{L}_{mel}$, $\mathcal{L}_{pitch}$, $\mathcal{L}_{fm}$, and $\mathcal{L}_{RVQ}$ denote adversarial loss, mel L2 loss, pitch L2 loss, feature match loss, and quantization loss, respectively. And the corresponding weights of these losses are denoted from λ_1 to λ_5. In detail, we adopt the format in LS-GAN [87] to avoid the gradient vanishing for the adversarial training. The formula is shown as

$$\mathcal{L}_{adv}(G;D) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,1)}[(1 - D(G(\boldsymbol{z})))^2], \tag{3}$$

$$\mathcal{L}_{adv}(D;G) = \mathbb{E}_{\boldsymbol{y}\sim p_{data}}[(1 - D(\boldsymbol{y}))^2] + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,1)}[D(G(\boldsymbol{z}))^2], \tag{4}$$

where G and D denote the generator and discriminators, respectively, $\boldsymbol{z}$ is the random noise and $\boldsymbol{y}$ represents the ground-truth speech data. Feature match loss is employed to compel our tokenizer to generate informative discrete tokens, whilst also enabling the reconstruction of high-quality waveforms from these discrete tokens. It can be rerepsented as follows

$$\mathcal{L}_{fm} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{y}}[\sum_{i=1}^{L_{MSD}} \frac{1}{N_i}||D^i(\boldsymbol{y}) - D^i(G(\boldsymbol{z}))||_1 + \sum_{j}^{L_{MPD}} \frac{1}{N_j}||D^j(\boldsymbol{y}) - D^j(G(\boldsymbol{z}))||_1], \tag{5}$$

where L_{MSD} and L_{MPD} are the number of layers of MSD and MPD, respectively, $D^i(\cdot)$ and N_i represent the feature map and the number of feature map of the i-th layer in MSD, and $D^j(\cdot)$ and N_j the feature map and the number of feature map of the j-th layer of MPD.

In order to encourage the model to disentangle semantic and acoustic information as well as focus on generating high-fidelity human speech, semantic reconstruction loss ensures the semantic embeddings alone can perserve the W2v-BERT 2.0 features. Besides, pitch reconstruction L2 loss using the a Crete pitch estimator [68] model to prioritize the reconstruction of prosody and speaker traits.

Input audio data is resampled to 24 kHz, therefore our model achieves an extreme 1920× compression ratio by producing tokens at only 12.5 Hz with 8 codebooks, and its bitrate is as low as 1 kbps.

2.2 Kodama TTS

2.2.1 Challenges and Motivation

On top of the tokenizer, we build **Kodama-TTS**, an LLM-based text-to-speech system that treats Kodama tokens as the speech interface. Similar in spirit to Higgs Audio v2 and MOSS-TTSD, Kodama-TTS models joint text+audio sequences in a unified discrete space and autoregressively predicts audio tokens, which are then rendered by the decoder of the codec model. Trained on large-scale spoken dialogue data, the 12.5 Hz token rate is particularly advantageous: it keeps sequences short enough for long-context conversational modeling while remaining expressive enough to capture emotion, speaker identity and language-specific prosody.

Together, Kodama-Codec and Kodama-TTS should form a vertically integrated speech stack. The codec model adopts an extremely compressed, semantically disentangled, streaming-friendly design. Without using a diffusion model that is trained separately to learn the mapping between quantized representation to Mel melspectrograms [17, 33], the Kodama-TTS leverages an LLM to fully model the details of generated audio by keeping the original decoder of the codec model. This approach maintains the alignment of the speech encoder and decoder and fills the gap between reconstruction and generation.

2.2.2 Method

We use Qwen3-1.7B as our LLM backbone. The model is adapted to operate over a mixed-modality discrete sequence in which text tokens and Kodama audio tokens share a unified embedding space. This design enables the LLM to directly reason over cross-modal dependencies and to autoregressively produce audio continuations conditioned on linguistic content, prior acoustic context, and conversational history.

Unified Token Embedding and Voice Clone. All tokens—whether linguistic or acoustic—are projected into a common hidden space through modality-aware embedding layers. Text embeddings are from the pretraining LLM, while audio-token embeddings are learned from scratch and optimized jointly with the language backbone. At inference time, the model generates audio tokens autoregressively following the textual prompt and optional acoustic exemplars, enabling voice clone ability with in-context learning. Since the LLM directly outputs audio tokens that are native to the integrated Kodama-Codec decoder, the generation path is straightforward and requires no diffusion-based refinement. This reduces latency, preserves reconstruction fidelity, and leverages the codec's original alignment with natural speech. The model learns to regulate duration, prosody, and expressivity through token-level decisions rather than through external duration models or alignment heuristics.

Fine-tuning After the pretraining use full-scale data, we selected expressive speech corpus in our dataset and filtered low quality samples using DNSMOS [102] and production quality (PQ) scores from audiobox [120]. Low speaker consistency samples that have low intra-utterance speaker similarity, which is calculated using a WavLM-Large model [124] are also removed. To further improve speaker reconstruction and emotion control, speaker embeddings from CAM++ [161] and emotion vectors from Emotion2Vec [142] are projected to the embedding space and prepended before the audio tokens.

Learning Objective. We formulate the training objective as minimizing the negative log-likelihood of the acoustic tokens. Let W represent the input text sequence, and A represent the target audio sequence of length T, where each frame A_t consists of $K = 8$ discrete codes from the RVQ layers, denoted as $A_t = \{a_{t,1}, a_{t,2}, \ldots, a_{t,K}\}$.

Conditioned on the speaker vector s and emotion vector e, the loss function is defined as:

$$\mathcal{L}_{\text{TTS}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\log P(a_{t,k} \mid A_{<t}, W, e, s; \theta)$$

2.3 Dataset

We collected open-source corpora and Internet data and curated a dataset of approximately 500k hours. Most data in the dataset has a original sample rate that is higher than 24 kHz, covering languages including

English, Chinese, Japanese, Spanish, German, Spanish, Russian, Korean, Portuguese, French. Genres include audiobooks, podcasts, and content from online video platforms. This large-scale dataset enables both our tokenizer and TTS models to encode and generate diverse and realistic speech signals with a high-level of semantic-prosody alignment.

3 Facial Animator

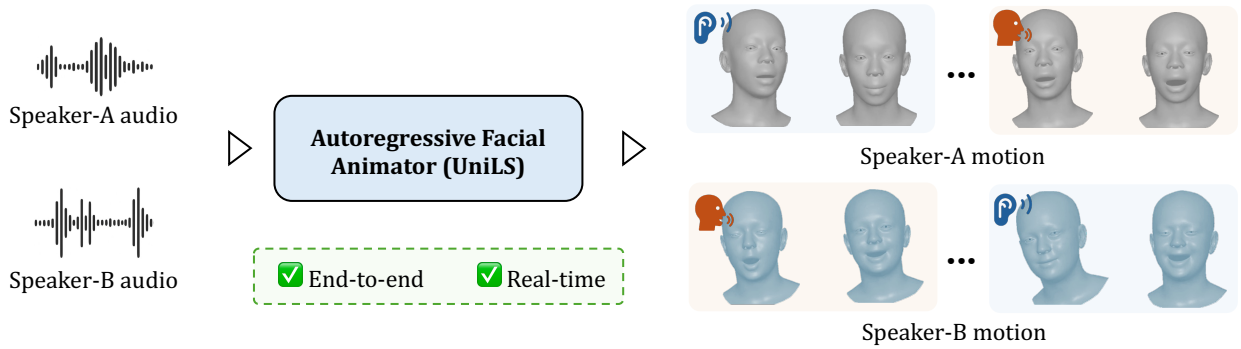

Figure 3 Given dual-track audio inputs from speaker-A and speaker-B, our method (UniLS) autoregressively generates two 3D facial motion sequences. Our method provides an end-to-end framework for unified, real-time speaking and listening motion generation.

3.1 Task Description

Driven by the audio output from section 2, we propose the task to generate avatars for unified listening and speaking. As shown in figure 3, the facial animator aims to generate 3D facial motion sequences for two speakers engaged in a dyadic conversation, driven solely by their respective audio streams. Given dual-track audio inputs $\mathbf{a}^A$ and $\mathbf{a}^B$, the animator produces two corresponding facial motion sequences M^A and M^B, where each sequence contains FLAME-based expression parameters, head pose, jaw pose, and eye-gaze dynamics.

Formally, for each motion chunk of time length t, the animator $\mathcal{G}$ autoregressively predicts the next motion chunk for both speakers:

$$\begin{aligned} \hat{M}^A_{t:2t} &= \mathcal{G}(M^A_{1:t}, \mathbf{a}^A_{1:t}, \mathbf{a}^B_{1:t}, \mathbf{s}^A), \\ \hat{M}^B_{t:2t} &= \mathcal{G}(M^B_{1:t}, \mathbf{a}^B_{1:t}, \mathbf{a}^A_{1:t}, \mathbf{s}^B), \end{aligned} \tag{6}$$

where $\mathbf{s}^A$ and $\mathbf{s}^B$ denote style embeddings representing identity-specific motion characteristics [21]. The generated motions satisfy two complementary objectives:

- **Speaking behavior generation.** When a speaker is talking, the predicted facial motion should align with the speaker's own audio, *i.e.*, capturing phoneme–lip correspondence and coordinated head–jaw movements.
- **Listening behavior generation.** When a speaker is not talking, the facial animator should produce natural listening behaviors such as blinks, micro-expressions, subtle head movements, and gaze adjustments. These behaviors should reflect intrinsic motion patterns while being modulated by the other speaker's audio that provides conversational context.

The task therefore requires a unified framework capable of jointly modeling both behaviors, producing continuous, expressive, and realistic 3D facial motions for both sides of the conversation.

Challenge. A central challenge in unified listen–speak facial motion generation is the phenomenon of **listening stiffness**. When an animator is trained end-to-end to map dual-track audio directly to both speakers' facial

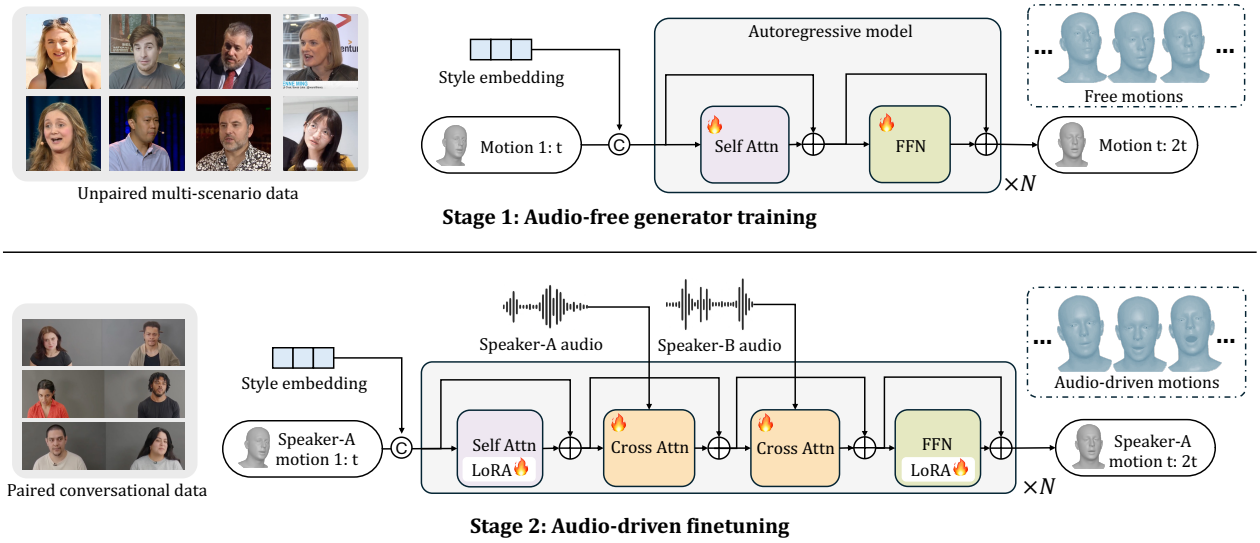

Figure 4 Overview of our two-stage training strategy. Stage 1 trains an autoregressive free generator on unpaired multi-scenario video data without using audio. Given past motions and a style embedding, the model predicts future free motion chunks. Stage 2 finetunes the generator on paired conversational clips by conditioning on speaker-A and speaker-B's audios through cross-attention, producing audio-driven speak–listen motions.

motions, the listening motion often collapses into low-variance, nearly static expressions. Such "zombie-face" behavior occurs because the audio–motion correlation is fundamentally unbalanced: a speaker's own audio provides strong phonetic and prosodic cues to drive speaking motions, whereas the listener's motion is only weakly related to the speech signal.

3.2 Approach: UniLS

We design our method by mirroring the natural process of human listening behavior. Natural listening arises from two components: 1) an internal motion prior that reflects spontaneous behaviors such as blinks, nods, and micro-expressions, and 2) external audio cues that modulate these intrinsic dynamics in response to conversational context. We develop a two-stage training framework that separately learns these two components. In Stage 1, we train an audio-free generator to model the internal dynamics of facial behavior. In Stage 2, we finetune this generator by conditioning on dual-track audios, allowing external speech signals to modulate the facial expression.

Stage 1: Audio-Free Generator Training. In the first stage, we train an audio-free generator to learn internal motion priors that capture natural facial dynamics independent of speech. This generator is trained on unpaired multi-scenario data, including diverse video sources such as news broadcasts, interviews, streaming content, and casual talking videos, providing diverse facial behaviors across identities and environments.

As shown in figure 4, the input contains motion chunk M with style embedding $\mathbf{s}$. This style embedding is learned to encode speaker-specific motion characteristics, following [21]. This embedding is concatenated with the input chunk and then fed them into our autoregressive model $\mathcal{G}$. Our model is transformer-based, composed of stacked self-attention and feed-forward blocks. At each time step t, the generator predicts the next-chunk motion based solely on past motions and the style embedding, *i.e.*:

$$\hat{M}_{t:2t} = \mathcal{G}(M_{1:t}, \mathbf{s}). \tag{7}$$

The model is trained with an autoregressive reconstruction loss over each chunk:

$$\mathcal{L} = \sum_{t=1}^{T} ||\hat{M}_{t:2t} - M_{t:2t}||. \tag{8}$$

Through this process, the model learns to produce free motions, which reflects the internal motion prior such as blinking, subtle head movements, and micro-expressions.

Stage 2: Audio-Driven Finetuning. In the second stage, we finetune the generator to produce audio-driven conversational motions, enabling both speaking and listening behaviors. This stage uses paired conversational data, where synchronized videos and audios from speaker-A and speaker-B provide the appropriate dynamics required for natural dialogue modeling.

Here we describe the process for generating speaker-A's motion as an example. As illustrated in figure 4, the input consists of three components: the motion chunk M, style embedding $\mathbf{s}$, and the audios $\mathbf{a}^A, \mathbf{a}^B$ from speaker-A and speaker-B, respectively. To incorporate audio guidance, we extend the stage 1 architecture by adding two cross-attention layers to each transformer block. Specifically, one attends to speaker-A's audio (for speaking behavior) and the other attends to speaker-B's audio (for listening behavior). The newly added cross-attention layers are trained from scratch, while the backbone weights inherited from stage 1 are finetuned with LoRA [52]. This design ensures allows the model to adapt efficiently to audio conditioning without overwriting the learned internal motion priors. Formally, the generating process of stage 2 is expressed as:

$$\hat{M}_{t:2t} = \mathcal{G}(M_{1:t}, \mathbf{a}^A_{1:t}, \mathbf{a}^B_{1:t}, \mathbf{s}). \tag{9}$$

The training objective remains a chunk-wise autoregressive reconstruction loss, which is the same as equation (8). Generating motions for speaker-B follows the same procedure, with the two audios simply exchanged in their roles. Through this finetuning, the generator combines internal motion priors learned in stage 1 with dual-track audio guidance, producing smooth and expressive speak–listen motions.

Implementation Details. Our facial animator uses a multi-scale VQ-VAE codebook [21], which consists of 256 entries, each with a code dimension of 64. The time window size is 100 frames (4 seconds), and the multi-scale levels are [1, 5, 25, 50, 100]. We used the AdamW optimizer with a learning rate of 1.0e-4 for training the codec, with a total batch size of 64 for 100,000 iterations. In the two-stage training, we train the autoregressive model using the AdamW optimizer with the same learning rate of 1.0e-4 and a batch size of 128 for 200,000 iterations. During this training, we employ a frozen wav2vec audio encoder [4]. All training was conducted on four NVIDIA H200 GPUs, requiring a total of approximately 40 GPU hours (10 GPU hours for the stage 1 and 30 GPU hours for stage 2).

For paired conversational data, we use the Seamless Interaction dataset [1], which offers large-scale dyadic conversational videos. For multi-scenario data, we additionally adopt four large-scale video datasets: CelebV [148], TalkingHead-1KH [128], TEDTalk [23], and VFHQ [136]. To enable 3D facial motion supervision, we apply a carefully designed tracking pipeline to extract per-frame FLAME parameters, including detailed eye-gaze [54] and head pose annotations [114]. After filtering, we obtain 675.5 hours of conversational data from Seamless Interaction, and 546.5 hours of multi-scenario data from other datasets. The conversational data includes 251.5 hours of speaking motions comprising 22.6M frames, and 406.0 hours of listening motions comprising 36.5M frames. For the conversational dataset, we use 622.5 hours for training, 4.8 hours for validation, and 30.2 hours for testing.

3.3 Result

In figures 5 and 6, we qualitatively compare our method with other baseline methods for listening and speaking motions, respectively. As shown in figure 5, we evaluate our facial animator on its ability to generate natural listening motions. ARTalk* [21] and DualTalk [98] exhibit noticeably stiff and low-variance facial behaviors, as highlighted by the red dashed boxes, often remaining close to a neutral expression with limited blinking, head movement, or micro-expressive changes. In contrast, our animator produces vivid, expressive, and temporally diverse expressions. The generated faces exhibit natural head dynamics, mouth shapes, and micro-expressions, demonstrating the effectiveness of our two-stage training in capturing realistic listening behavior. In figure 6, our method demonstrates excellent lip synchronization, accurately capturing a wide range of phonetic elements and their associated articulation patterns. Beyond mouth movements, the generated speaking sequences also exhibit realistic facial behavior, such as micro-expressions and head movements.

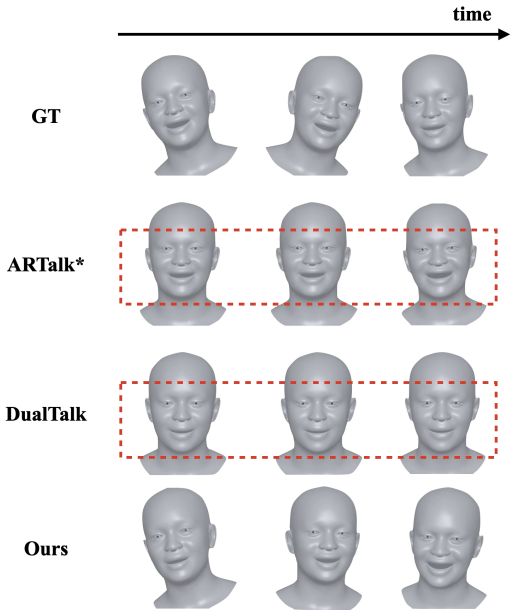

Figure 5 Qualitative comparison on listening motions. Red rectangles highlight motion stiffness over time.

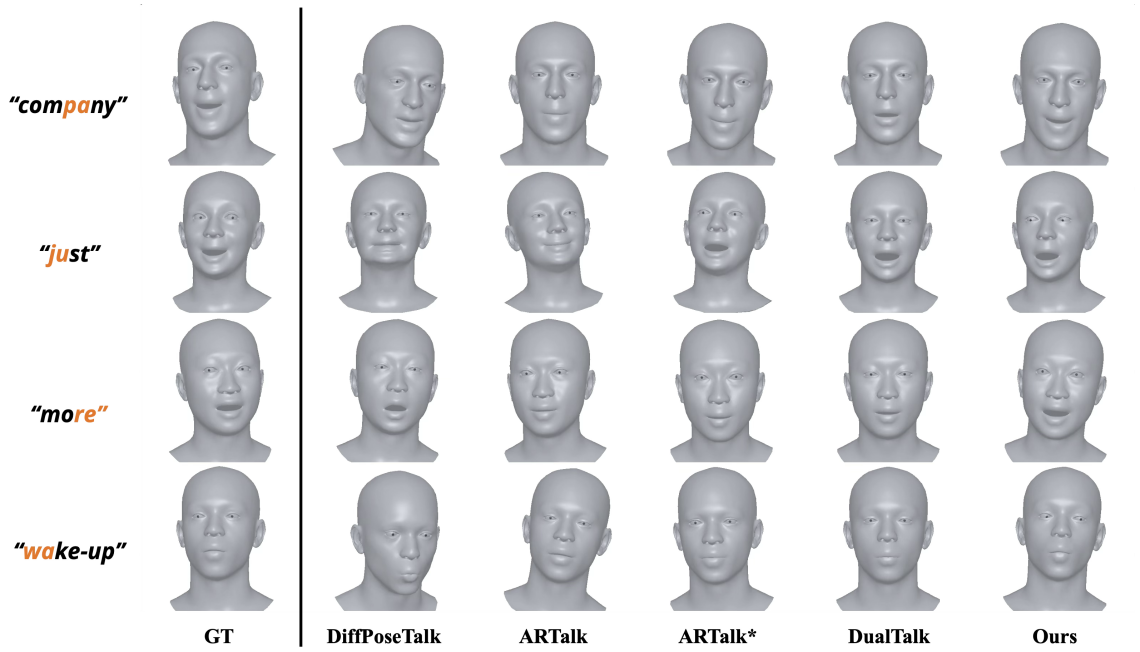

Figure 6 Qualitative comparison on speaking motions. Our facial animator shows better alignment with the ground truth in expression style and lip synchronization.

4 Body Animator

4.1 Task Description

We formulate *text-controlled, streaming body motion synthesis* as a conditional time-series generation problem. Let $\mathbf{X}^{0:K}$ denote a sequence of body motion states (e.g., joint rotations and positions) and $\mathbf{c}^{0:K}$ denote a

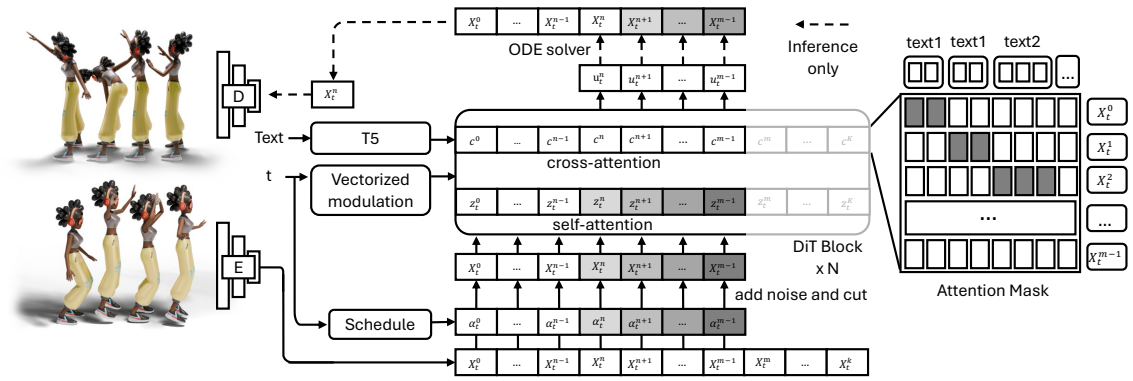

Figure 7 Pipeline Overview. FloodDiffusion encodes the motion stream into a compact latent sequence via a causal VAE. The model predicts velocity for the active window conditioned on context from the Thinker. Key designs include a lower-triangular noise schedule and frame-wise text conditioning. Inference slides the window for streaming output.

time-varying control signal (instruction stream) received from the **Thinker**. The Body Animator estimates the mapping

$$\mathbf{X}^{0:K} = g(\mathbf{c}^{0:K})$$

in a streaming manner, where the output at time t must be generated with strict latency constraints and without access to future control signals. It consumes intent directly from the **Thinker** (as text prompts or control tokens) and produces physically plausible, seamless full-body motions for the **Renderer**.

Specifically, we consider the output space of body motion. We adopt the standard motion representation used in HumanML3D, which consists of a 263-dimensional vector including global root velocity, root rotation, local joint rotations, and foot contact information. This high-dimensional continuous space captures the full nuance of human kinematics. The input control signal $\mathbf{c}$ consists of natural language prompts that can be updated at arbitrary time steps t_k. The challenge lies in generating $\mathbf{x}_t$ that is coherent with $\mathbf{x}_{<t}$ and aligned with the current active instruction $\mathbf{c}_t$, all while adhering to a frame budget (e.g., 33ms for 30FPS).

To achieve this, we must overcome several challenges:

C1. Real-time streaming under tight latency. The generator must operate causally (no future frames), maintain 20–60 Hz output, and amortize model compute so that per-step latency stays below the frame budget. Standard diffusion models require tens or hundreds of denoising steps per frame, which is prohibitive for real-time applications.

C2. Editable control at any time. Instructions from the Thinker can arrive mid-gesture (e.g., "walk → run → wave while turning"). The model must switch goals without visible reset, artifacts, or discontinuities. Unlike offline generation where the text is fixed for the whole clip, streaming requires the model to "steer" the motion trajectory smoothly.

C3. Multi-rate, multi-granularity conditioning. High-level intent (style/persona) evolves slowly; action verbs and spatial targets change quickly; prosody and micro-beats from speech can be even faster. Aligning these rates without drift is nontrivial.

C4. Long-horizon coherence. Maintaining personality traits and interaction logic across minutes while allowing rapid local edits demands both memory and controllability. A naive frame-by-frame generator often suffers from motion freeze or jitter over long sequences.

4.2 Approach: FloodDiffusion

We introduce **FloodDiffusion**, a framework based on *diffusion forcing* tailored for streaming motion generation. Unlike autoregressive models or chunk-based diffusion which suffer from "first-token" latency or lack of long-term history, FloodDiffusion enables flexible, low-latency generation by allowing different frames to carry different noise levels.

4.2.1 Preliminaries

We fix the initialization distribution to be standard white Gaussian noise $p_{\text{init}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Diffusion models perform distribution matching by transporting $p_0 \sim p_{\text{init}}$ to the data distribution $p_T \sim p_{\text{data}}$ via a time-indexed Gaussian corruption path. For each data point $\mathbf{z} \sim p_{\text{data}}$ and time $t \in [0, T]$, we define

$$p_t(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\big(\mathbf{x};\, \alpha_t\, \mathbf{z},\, \beta_t^2\, \mathbf{I}\big), \tag{10}$$

where α_t and β_t are scalar schedules. To enable streaming, we extend these to **Vectorized Time Schedules**. Let K denote the sequence length. We define:

$$\boldsymbol{\alpha}_t = [\alpha_t^0, \alpha_t^1, \ldots, \alpha_t^{K-1}] \in \mathbb{R}^K \tag{11}$$

$$\boldsymbol{\beta}_t = [\beta_t^0, \beta_t^1, \ldots, \beta_t^{K-1}] \in \mathbb{R}^K \tag{12}$$

This allows us to assign different noise levels to different frames at the same global "denoising time" t.

4.2.2 Motion Latent Space via Causal VAE

Instead of discrete tokenization (VQ-VAE) or operating in raw space, we map the high-dimensional motion sequence (263D) into a compact continuous latent space (4D) using a **Causal VAE**. **Architecture and Training.**

We adapt the causal VAE design from Wan2.1 (video generation) to 1D temporal sequences. The encoder and decoder are strictly causal, meaning the latent z_t and reconstruction $\hat{x}_t$ depend only on $x_{\leq t}$. The training objective is a combination of reconstruction loss and codebook commitment loss:

$$\mathcal{L}_{\text{VAE}} = \|\mathbf{x} - D(z)\|_2^2 + \|\text{sg}[E(\mathbf{x})] - z\|_2^2 + \gamma\|\text{sg}[z] - E(\mathbf{x})\|_2^2 \tag{13}$$

where sg denotes the stop-gradient operator. We use a temporal downsampling factor of 4 and a latent channel dimension of 4. This configuration compresses the 263-dimensional motion data into a highly compact $4 \times T/4$ representation, which significantly reduces the computational burden on the downstream diffusion model while preserving high-frequency motion details. This provides a stable, low-dimensional space for the diffusion model to operate in.

4.2.3 Tailored Diffusion Forcing

The core of our approach is a tailored diffusion forcing objective. We discovered that vanilla diffusion forcing (using random schedules) fails for motion data. We instead propose a specific **Lower-Triangular Schedule**.

Lower-Triangular Schedule. Let n_s be the streaming step size parameter. We define the schedule as:

$$\alpha_t^k = \text{clamp}(t - k/n_s, 0, 1) \tag{14}$$

$$\beta_t^k = 1 - \alpha_t^k \tag{15}$$

This schedule creates a "cascading" activation pattern. At any generation step t, we can identify three regions: 1. **Fixed Past ($k < m(t)$)**: Frames are fully denoised ($\alpha = 1, \beta = 0$). 2. **Active Window ($m(t) \leq k < n(t)$)**: Frames are actively being denoised with varying noise levels. 3. **Future Noise ($k \geq n(t)$)**: Frames are pure noise ($\alpha = 0, \beta = 1$).

This structure provides a mathematical guarantee of **Streaming Locality**. It implies that at time t, we only need to compute the update for the active window. Frames before $m(t)$ are finalized and can be sent to the Renderer.

Bi-directional Attention. We employ a Diffusion Transformer (DiT) backbone. A critical finding is that within the active window, **bi-directional attention** is essential. Although the overall system is streaming/causal, the frames *currently being denoised* benefit massively from attending to each other. Restricting attention to be causal within the denoising window degrades performance significantly (FID drops from 0.057 to 3.37).

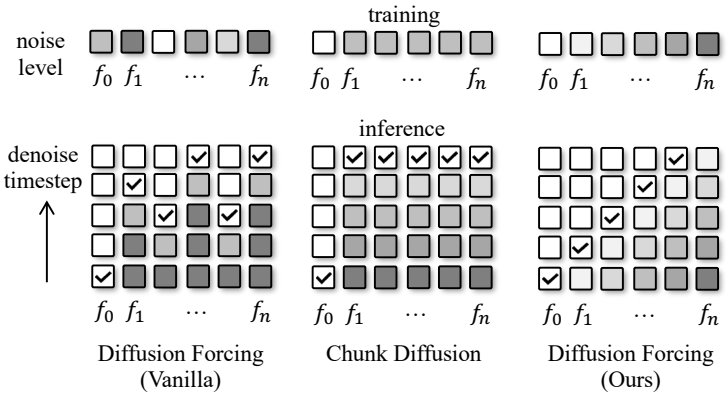

Figure 8 Noise Schedule Comparison. Our triangular schedule (right) denoises only the active window and advances at a constant rate, unlike random schedules or chunk-based diffusion.

4.2.4 Time-Varying Text Conditioning

To handle changing instructions from the Thinker (e.g., "walk" → "wave"), we implement a **frame-wise text conditioning** mechanism.

- **Text Encoding**: We use a T5 encoder to process text prompts into embeddings.
- **Condition Fusion**: We use a biased attention mask where each motion frame k attends to the text prompt active at time k.
- **Handling Transitions**: When the Thinker updates the prompt, the new embedding is seamlessly integrated for future frames. The overlapping active window ensures a smooth transition between the old and new motion styles without explicit "prompt refresh" logic.

4.2.5 Streaming Inference Algorithm

During inference, we slide the active window forward by a fixed step Δt. This process allows for continuous generation of infinite-length sequences.

- **Step 1: Initialize**. Start with a buffer of pure Gaussian noise.
- **Step 2: Window Identification**. For each simulation step t, determine the indices of the active window $[m(t), n(t)]$ using the lower-triangular schedule definitions.
- **Step 3: Predict Velocity**. Run the DiT model. Thanks to Streaming Locality, we only need to compute the output for frames in the active window. The model takes the current noisy latents $\mathbf{z}_t$ and the active text embeddings $\mathbf{c}_t$ as input.
- **Step 4: Denoise Update**. Apply the numerical solver step (e.g., Euler method) to update the latents: $\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \mathbf{v}_t \cdot \Delta t$.
- **Step 5: Shift and Commit**. Slide the window forward. Frames that exit the window ($k < m(t + \Delta t)$) are considered "committed" and are decoded by the Causal VAE decoder to produce final motion frames for the Renderer.

This pipeline yields a constant computational cost per step and strictly bounded latency (approx. n_s frames), avoiding the generation pauses typical of chunk-based methods. The lower-triangular schedule ensures that by the time a frame exits the active window, it has been fully denoised ($\alpha = 1$).

4.3 Results

The Body Animator, powered by FloodDiffusion, converts a live, editable instruction stream into physically plausible, personality-consistent body motion in real time. By combining a causal VAE with a tailored

Figure 9 Comparison of time-varying conditioning. Our model generates different resulting motions from the same text prompts based on their delivery timing. (Top Left) Prompts are given separately at different frames. (Top Right) All conditions are fed as a single prompt at once. (Bottom Left) Two separate prompts are input early in the sequence. (Bottom Right) The same two separate prompts are input later in the sequence.

Figure 10 Comparison of long sequence generation. (Left) our model will continue to repeat the motion in text prompt if without new prompts come. (Right) in real application, our model could stop current motion by explicitly giving the rest style prompt, such as “stand”.

diffusion forcing scheduler, we deliver an industry-first system that supports fully online, instruction-editable body animation with SOTA quality (0.057 FID) and strict frame-rate adherence. This component serves as the robust physical actuator for the broader digital human system.

5 DiT-based Rendering

5.1 Task Description

The goal of the DiT-based Rendering module is to synthesize high-fidelity, identity-consistent human video frames from a stream of motion parameters provided by the embodiment modules (Facial Animator and Body Animator). Unlike conventional image- or video-driven approaches that rely on reference frames or driving sequences, our renderer operates purely on parameterized 3D control signals—such as FLAME expression parameters, SMPL body pose, and camera configurations—while maintaining strict temporal coherence and multi-view consistency.

Formally, at time step t, let

- ϕ_t denote the FLAME parameters capturing facial expression, jaw pose, gaze, and local head pose,
- ψ_t denote the SMPL parameters representing global body pose, articulation, and shape,
- κ_t denote camera intrinsics and extrinsics, and
- I^* denote a single reference image encoding the target identity.

The renderer estimates a mapping

$$V_t = \mathcal{R}(\phi_t, \psi_t, \kappa_t, I^*) \tag{16}$$

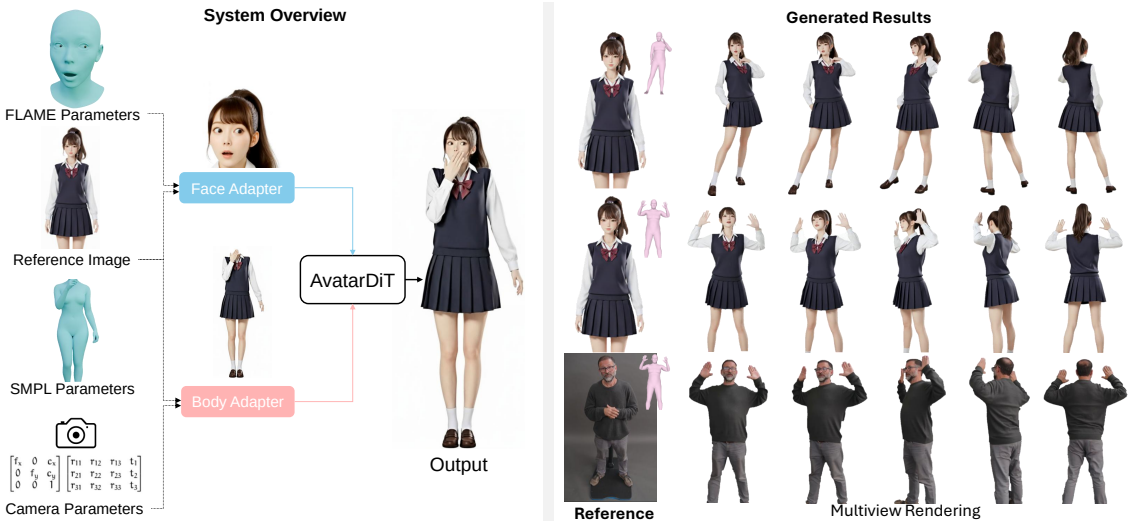

Figure 11 Overview of AvatarDiT and results showcasing multi-view consistency of the proposed framework.

that produces the RGB frame V_t such that it satisfies the following properties:

Identity Preservation: The generated frames must remain consistent with the reference identity across all poses, motions, and viewpoints, avoiding drift throughout the sequence.

Parameter-Faithful Motion Rendering: Facial expressions, head movements, and body articulation must accurately reflect the supplied FLAME/SMPL parameters, enabling fine-grained control without over-smoothing.

Multi-View Geometric Consistency: Since the Thinker or Body Animator may dynamically specify different viewpoints, the renderer must produce output that respects camera geometry and maintains cross-view consistency.

Temporal Stability: The renderer must avoid flickering or stochastic drift across frames, ensuring smooth temporal evolution despite the inherent stochasticity of diffusion-based generation.

5.1.1 Challenges

Achieving parameter-based, identity-consistent rendering introduces several key challenges:

1. **Disentangling identity from motion parameters.** Motion parameters encode pose and expression but contain no identity cues, requiring the renderer to integrate them stably with the identity embedding.
2. **Bridging low-dimensional parameter spaces with image-space features.** FLAME and SMPL parameter spaces are compact, whereas video diffusion models operate on dense spatial embeddings, necessitating carefully designed adapters.
3. **Maintaining cross-view coherence.** Diffusion models are prone to view inconsistency unless explicitly trained with multi-view objectives and camera-aware modulation.
4. **Avoiding distribution shift across heterogeneous datasets.** Datasets with high-quality FLAME labels seldom include multi-view imagery, and vice versa, making it difficult to jointly learn facial control and geometric consistency.

5.2 Approach: AvatarDiT

We aim to achieve parameter-based rendering through a video Diffusion Transformer (vDiT). By leveraging the strong generative capability of the WAN model [37], our framework supports identity-consistent multi-view generation and parameter-driven facial control. Unlike existing approaches that rely on reference images or driving videos, our method controls facial motion directly via FLAME parameters, enabling precise and

disentangled manipulation. As illustrated in Figure 12, the proposed framework consists of three major components: (1) a Diffusion Transformer serving as the denoising backbone, (2) text and image encoders together with a patch convolution embedder for multi-modal conditioning, and (3) a set of control modules jointly optimized during training to inject parameter-based motion and identity information into the generation.

Given the distinct nature of facial motion control and multi-view generation, and the difficulty of collecting data that jointly captures both modalities, we adopt a three-stage training pipeline as shown the lower part in Figure 12.

- **Face control stage:** The FLAME adapter and motion encoder are trained to enable parameter-based facial control by replacing the original RGB conditioning with FLAME parameters.
- **Multi-view control stage:** To address view inconsistency and identity drift, we introduce a cross-view training strategy that enforces geometric and appearance coherence across camera poses.
- **Joint fine-tuning stage:** All modules are jointly optimized to close the distribution gap and achieve unified, identity-consistent generation.

For each training stage, we first train at a resolution of 512×768, and then scale up to 720×1280 to align with the resolution setting of WanAnimate. As illustrated in Figure 1, the control signals in our framework can be derived from multiple modalities, such as video, audio, or text, enabling flexible and multimodal conditioning. During inference, these aggregated parameter signals are injected into the AvatarDiT alongside a reference image to synthesize identity-consistent, parameter-driven human videos. The detailed configurations and objectives of each training stage are described in the following subsections.

5.2.1 Parameter-based face control

Wan Animate [37] effectively controls facial motion using RGB image inputs. However, this image-based design constrains its flexibility and generalization. To overcome this limitation, instead of extracting motion embeddings from face images [37, 86], we employ FLAME parameters [76, 27, 39] to control facial movement. This parameter-based representation allows more general and adjustable control over facial expressions and motion, as FLAME parameters can explicitly disentangle motions from other visual attributions and are widely used as middle representations for motion extraction and facial motion generation [165, 22, 92].

We adopt a 4-layer Transformer adapter to map the 112-D FLAME parameters into a 512-D *face-motion embedding* space. In FLAME, the parameter vector is defined as $\phi = [\mathbf{e}; \mathbf{r}_{\text{jaw}}; \mathbf{r}_{\text{gpose}}; \mathbf{r}_{\text{leye}}; \mathbf{r}_{\text{reye}}]$, where $\mathbf{e} \in \mathbb{R}^{100}$ denotes expression coefficients and $\mathbf{r}_{\text{jaw}}, \mathbf{r}_{\text{gpose}}, \mathbf{r}_{\text{leye}}, \mathbf{r}_{\text{reye}} \in \mathbb{R}^3$ are axis-angle local poses (12-D total), for a total of 112 dimensions. The adapter produces $A(\phi) \in \mathbb{R}^{512}$ and injects it via element-wise residual addition into the image-derived motion embedding $E_{\text{face}}(I)$ extracted by a WanAnimate-pretrained face-motion encoder:

$$z \;=\; E_{\text{face}}(I) \;+\; A(\phi). \tag{17}$$

Here, $E_{\text{face}}(I)$ acts as an implicit constraint on $A(\phi)$. Without such supervision, cross-modal/scale mismatches, partial non-overlap of factors, and temporal ambiguity make semantic alignment extremely challenging.

To enable fully parameter-based control, we jointly optimize the FLAME adapter and the motion encoder. Since the motion encoder is pre-trained on human-face RGB images, we adapt it to also accept FLAME-rendered mesh RGB. After optimization, the framework can drive facial motion using the face mesh, the FLAME parameters, or both.

This training strategy is essential for embedding FLAME parameters into the same latent space as the original facial motion representations. An ablation study will be given in Section 4 to demonstrate its necessity. This design not only enables consistent embedding alignment but also results in comparable face control compared to the RGB-based approach adopted by Wan Animate. Similar to IP-Adapter [145], though this training is performed on a specific dataset, the adapter can be generalized to various inputs out of the distribution.

5.2.2 SMPL-driven Multi-view control

A common approach in existing human animation systems for introducing motion control is to use OpenPose [12], a keypoint-based modality, as control signals [37, 159, 63]. However, OpenPose outputs are sparse

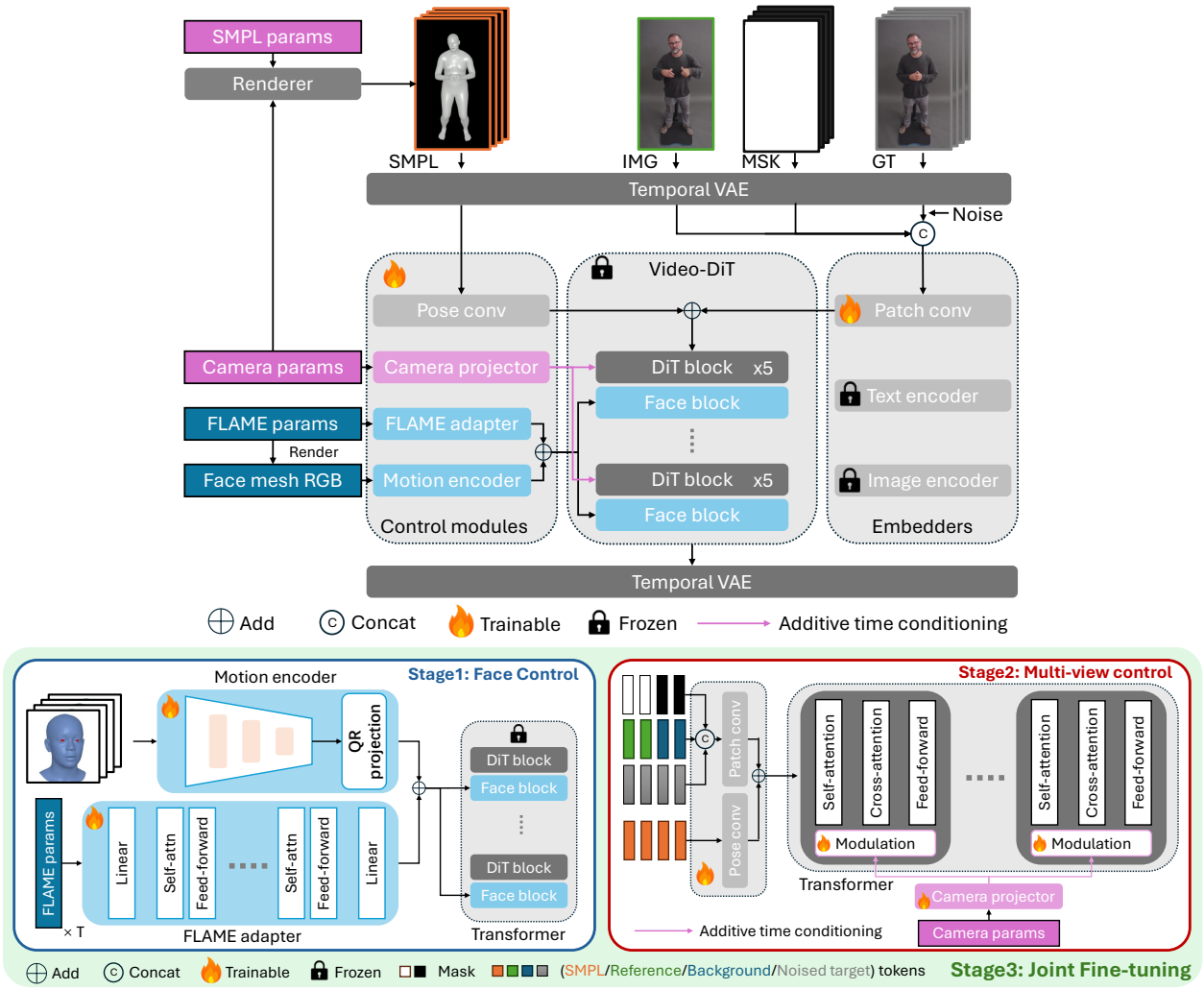

Figure 12 Overview of the proposed framework. The training process consists of 3 stages: (1) Face Control Stage, which learns parameter-driven facial motion control with FLAME-rendered mesh supervision; (2) Multi-View Stage, which enforces cross-view consistency across multiple camera parameters; and (3) Joint Stage, which integrates both to achieve coherent spatio-temporal alignment.

and highly abstract, making it difficult to infer accurate 3D information from its RGB image.

To enable more precise control over both synthesized motion and camera view, we instead employ SMPL [85]-based RGB renderings as control signals. These renderings are generated from SMPL parameters together with camera poses, allowing our framework to be driven entirely by 3D controllable parameters without relying on input videos.

We basically follows the input formulation we utilize the input formulation of Wan-Animate, whose denoising target comprises chunks of *reference latents*, *temporal latents*, and an *environmental latent*. This formulation naturally helps maintain cross-view consistency when reference latents are extracted from multi-view images. In our multi-view training, we randomly select 1–5 reference frames from different views and encode them into the latents as original WanAnimate. Yet, we introduce a trainable module to improve multi-view consistency.

Additionally, we finetune the modulation layers at each DiT block to introduce a *camera-based shifting*, analogous to the timestep-embedding shifting in Wan-Animate. Let the camera parameters (intrinsics/extrinsics) be embedded into three channel-wise scalar vectors $e_0, e_1, e_2 \in \mathbb{R}^C$ via linear modulations. Let FFN$(\cdot)$ denote the DiT feed-forward MLP, and let Norm$(\cdot)$ denote the channel-wise normalization in the block. Denote by

Figure 13 Ablation study of the supervision embeddings used to train the FLAME adapter. Shown are crops from our full-body generation results using on-the-fly reference conditioning.

$z^{(\mathrm{ca})}$ the cross-attention output within the same DiT block. We adopt:

$$z_{\mathrm{out}} = z^{(\mathrm{ca})} + \mathrm{FFN}\Big(\mathrm{Norm}(z^{(\mathrm{ca})}) \odot (1 + e_1) + e_0\Big) \odot e_2, \tag{18}$$

where $\odot$ denotes channel-wise multiplication.

Joint training. After independently training the FLAME adapter and multi-view adapter on respective datasets, we need to align their distributions to avoid out-of-distribution (OOD) artifacts and achieve a better performance. We perform a existing open-source datasets cannot satisfy the requirement. Therefore, we organize a small fine-tuning dataset to train both adapters jointly.

5.2.3 Dataset for Joint Fine-tuning

To reconcile the distribution gap between the facial-control and multi-view stages, and to ensure that the final renderer can simultaneously handle parameter-driven facial motion and cross-view geometric consistency, we construct a three-stage training curriculum, each paired with a dataset whose supervision signals match the requirements of that stage. In the face-control stage, we train the FLAME adapter and motion encoder on the *Seamless Interaction* dataset [38]. We labeled 40 Per-frame FLAME parameters (pose, expression, jaw, and weak-perspective camera) are obtained with an off-the-shelf FLAME fitter, yielding paired RGB–FLAME supervision. For multi-view training, we use *MVHumanNet* [140], a calibrated multi-camera human dataset; synchronized frames from distinct views are sampled to form positive cross-view tuples, driving view-consistent feature learning. In the joint training stage, we curate an identity-disjoint set of short sequences that simultaneously provide synchronized multi-view imagery and frame-aligned FLAME labels, allowing the model to reconcile facial control with view geometry. Across stages, we follow official train/val splits where available (otherwise performing identity-level splits), standardize frame rate and preprocessing (face-centric cropping and mild color/temporal jitter), and train at 512×768 before upscaling to 720×1280 to match our setting.

5.3 Results

The DiT-based Rendering module converts parameterized facial and body motion into identity-consistent, high-fidelity human videos under dynamic pose and camera conditions. By relying exclusively on FLAME and SMPL parameters with explicit camera control, the renderer achieves stable, controllable synthesis without requiring reference driving videos.

As shown in Figure 14, our method demonstrates superior multi-view consistency compared to existing approaches, maintaining coherent facial identity, body shape, and appearance across diverse viewpoints. In

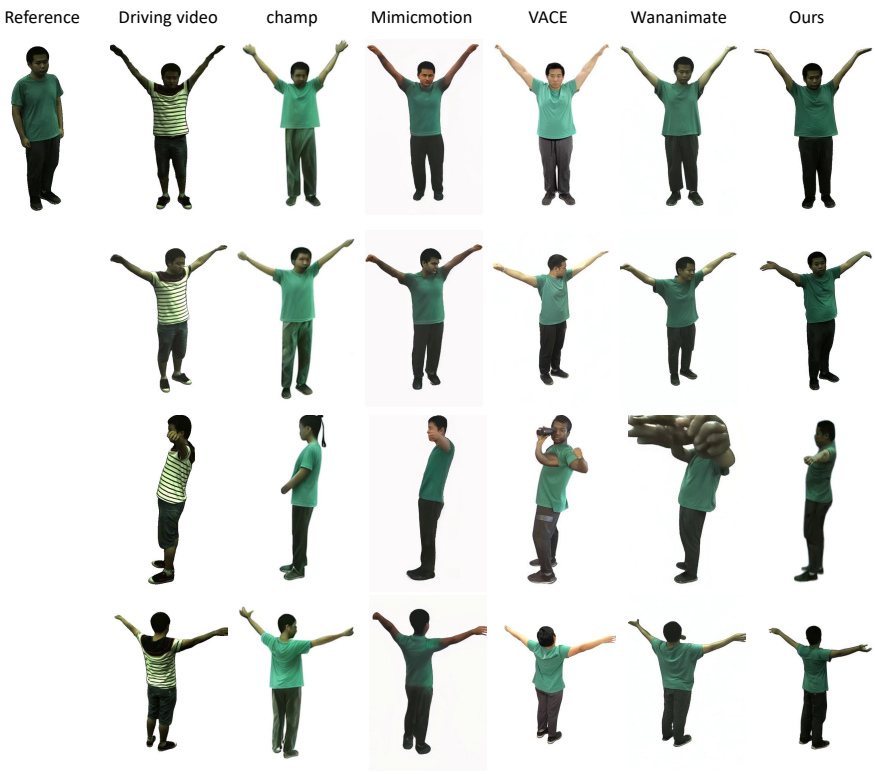

Figure 14 Qualitative comparison for multi-view results with different camera parameters.

contrast, prior methods frequently exhibit identity drift or view-dependent artifacts when camera parameters change. The proposed parameter-based facial control further enables accurate rendering of fine-grained expressions while preserving temporal stability.

Overall, the DiT-based Rendering module provides a robust visual realization layer for interactive digital humans, ensuring faithful motion rendering and strong identity preservation across long sequences. Additional quantitative and user-study evaluations are provided in Section 7.5.

6 Thinker

6.1 Task Description

We formulate the high-level cognitive processing of the digital human as a continuous, multimodal context-response generation problem. The Thinker functions as the central orchestrator, tasked with mapping a stream of multimodal sensory inputs to a coherent set of unified instructions for the embodiment modules (Talker, Facial Animator, and Body Animator).

Let I_t denote the multimodal input at time t, comprising the user's textual utterances, acoustic features, and visual cues. Let M represent the agent's internal state, comprising both short-term context and long-term narrative knowledge. The Thinker estimates the policy:

$$A_t = \pi(I_t, M)$$

where A_t is the unified action plan consisting of dialogue content, emotional state parameters, and physical gestures. The objective is to generate A_t such that it maximizes personality coherence, narrative causality, and user engagement, driving emergent behaviors that go beyond pre-scripted rules.

To achieve genuine interactive intelligence within a specific narrative context, we must overcome several fundamental challenges. First, narrative causality and spoiler leakage. Standard retrieval-augmented generation

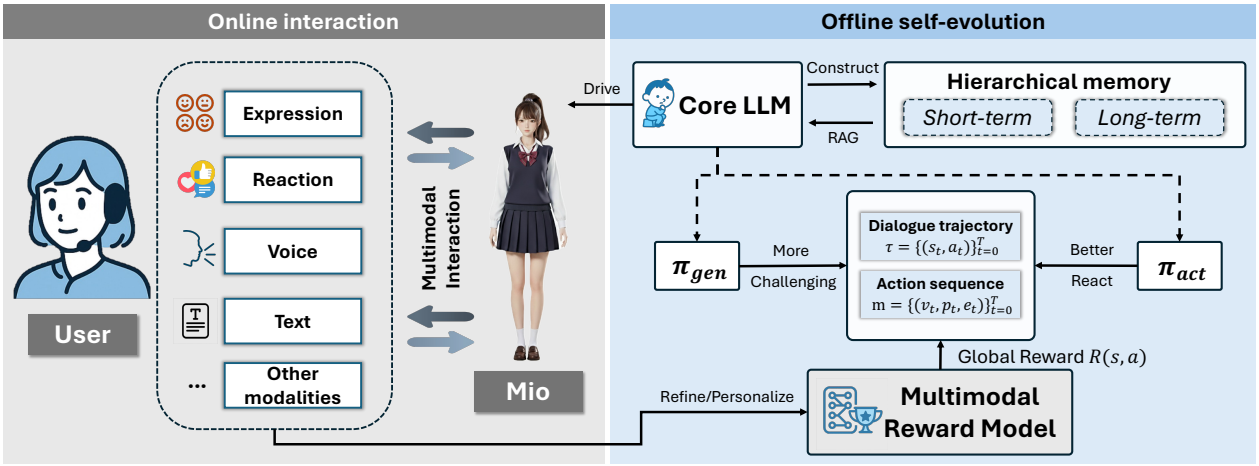

Figure 15 The overview of the Thinker module in Mio. (Left) Online Interaction: The Core LLM functions as the central orchestrator, driving Mio's multimodal expressions (voice, reaction, text) in real-time. It leverages a Hierarchical Memory system—comprising a short-term context buffer and a long-term Diegetic Knowledge Graph—accessed via Story-Time-Aware RAG to ensure narrative consistency and prevent spoiler leakage. (Right) Offline Self-Evolution: To refine persona fidelity without manual annotation, the system employs a competitive self-training loop. A Generative Policy (π_{gen}) constructs challenging interactive scenarios (dialogue trajectories) to probe the agent's weaknesses. The Actor Policy (π_{act}) then optimizes its responses based on feedback from a Multimodal Reward Model, which decomposes global user satisfaction signals into fine-grained action rewards $R(s, a)$.

(RAG) systems utilize temporally flat vector databases [74, 55]. In a narrative-driven setting, this leads to incoherence or spoilers, where the agent accesses information from future events (e.g., $t_{retrieved} > t_{current}$), violating the causal logic of the story [66, 81]. The system must maintain a strict "Narrative-Present" boundary while retaining access to deep historical lore. Second, temporal credit assignment in dialogue. Reinforcement learning for dialogue suffers from sparse reward signals. A user's satisfaction is typically expressed as a global signal R_{Global} (*e.g.*, a post-session rating) rather than immediate feedback for specific turns. Attributing this global signal to fine-grained local actions a_t without manual turn-by-turn annotation is a critical hurdle for autonomous improvement. Third, data-free persona alignment. Aligning a general-purpose Large Language Model (LLM) to a specific, nuanced character persona usually requires extensive manual annotation or supervised fine-tuning [121, 96]. The challenge lies in enabling the agent to autonomously refine its behavior and voice through self-evolution, rejecting out-of-character hallucinations without relying on external labeled datasets.

6.2 Approach: NPC-Centric Embodied LLM

At the heart of the Thinker is a specialized Large Language Model (LLM) tailored specifically for NPC embodiment and real-time interaction. We ground this NPC-centric model in the specific lore, context, and personality constraints of its designated character, using the following techniques.

To provide Mio with a coherent sense of self and a consistent memory, we have implemented a hierarchical memory architecture [48, 93]. This system is the core mechanism for grounding the character's identity, ensuring that its behavior and dialogue remain consistent across long-term interactions while strictly adhering to narrative causality.

The architecture is composed of two distinct but interconnected tiers:

1. Short-term Memory (Context Buffer): This tier functions as a high-speed, volatile conversational buffer. It stores the immediate context of an interaction, including recent utterances, dialogue history, and currently active goals, allowing Mio to track the moment-to-moment flow of conversation.

2. Long-term Memory (Diegetic Knowledge Graph): To address the issues of narrative incoherence and spoiler leakage common in standard vector databases, we implement the long-term memory as a Diegetic Knowledge Graph [47, 56]. Rather than a flat semantic index, this graph structures foundational memories, personality

traits, and world lore into entities (nodes) and relations (edges). Critically, every element in this graph is explicitly tagged with a story-time coordinate (t), anchoring facts to the specific moment they occur in the narrative timeline.

At inference time, these two tiers work in synergy via a Story-Time-Aware Retrieval mechanism. The system executes a dual-level retrieval pipeline: it first performs semantic search on graph nodes to capture specific entities (low-level details) and then on edges to capture thematic relationships (high-level context). Crucially, this retrieval is governed by a Narrative-Present gate. If Mio is currently situated at time $t_{current}$, the gate rigorously filters out any memory nodes where $t_{node} > t_{current}$. This architectural constraint guarantees that Mio cannot access or leak information about events she should not have known, ensuring deep narrative consistency and preventing frame-breaking errors regardless of user prompting.

As for model training, self-evolution is the learning process that allows Mio to adapt and improve its interactive capabilities over time. This is accomplished through an autonomous learning loop that uses feedback from real user interactions to refine the LLM. This process incorporates Deep Persona Alignment (DPA) techniques to ensure high fidelity to the character's voice and values without relying on manual annotation.

6.2.1 Multimodal reward model

The Multimodal Reward Model is the core component that provides the ground-truth feedback signal for learning. It addresses the fundamental temporal credit assignment problem in dialogue, which is the challenge of attributing a single, sparse global reward (e.g., a user's post-session satisfaction rating) to the specific, fine-grained local actions that caused it [91]. To solve this without requiring manual turn-by-turn annotation, we employ a framework where a Large Language Model (LLM) decomposes the global signal by interpreting the user's implicit multimodal cues.

For a given dialogue trajectory, $\tau = \{(s_t, a_t)\}_{t=0}^{T-1}$, we first extract a time-aligned multimodal feature vector, m_t, representing the user's reaction within each state, s_t. This vector concatenates features from different modalities:

$$m_t = [v_t, p_t, e_t] \tag{19}$$

where v_t is the visual feature vector (gaze, smile, motion probabilities), p_t is the prosodic vector (pitch, intensity, jitter), and e_t is the textual semantic embedding.

These numerical vectors are then transformed by a function, f_{desc}, into natural language descriptors, $d_t = f_{\text{desc}}(m_t)$. This step creates a text-based representation of non-verbal cues (e.g., "User's pitch was high and they were smiling") that is compatible with an LLM.

We then employ a powerful, frozen LLM as a zero-shot reward decomposition oracle, which we denote as the function M_{oracle}. This oracle takes the augmented trajectory, $\tau' = \{(s_t, a_t, d_t)\}_{t=0}^{T-1}$, and the final scalar global reward, $R_{\text{Global}}(\tau)$, as input. Its function is to output a sequence of turn-level rewards, $\{R_t\}$, for each of Mio's actions, a_t:

$$\{R_t\}_{t=0}^{T-1} = M_{\text{oracle}}(\tau', R_{\text{Global}}(\tau)) \tag{20}$$

The prompt given to M_{oracle} includes a soft constraint that the sum of the decomposed local rewards should be faithful to the global signal. This objective can be expressed as:

$$\sum_{t=0}^{T-1} R_t \approx R_{\text{Global}}(\tau) \tag{21}$$

The resulting local rewards, R_t, which are now grounded in the user's implicit multimodal reactions, are stored as the ground-truth feedback signal.

6.2.2 Data-free self training

To enable our Mio to autonomously refine its capabilities, we employ a data-free training protocol where the model improves through competitive self-play [73, 101, 152, 59]. For this reinforcement learning loop to function, the agent must have a defined action space through which it can interact and a computable reward

function to evaluate the quality of those actions. This structure provides the necessary feedback for iterative improvement without requiring external datasets.

We instantiate this as a game wherein the model operates under two opposing policies:

1. **A Scenario-Generative Policy (π_{gen}):** In this mode, the model is tasked with generating complex and challenging interactive scenarios. The objective of this policy is to create situations designed to probe for weaknesses in the agent's reasoning, emotional appropriateness, or personality consistency. To ensure robustness, this policy employs three specific generation strategies:
    - *Timeline-Cycled Questioning:* Extracting events from the Diegetic Knowledge Graph to probe knowledge across the entire narrative arc.
    - *Randomized Tone Generation:* Simulating various user emotional states (e.g., tense, sarcastic, angry) to test emotional appropriateness.
    - *Varied Intent Simulation:* Generating prompts with distinct goals, such as information seeking, challenging, or negotiating.
2. **An Interactive Actor Policy (π_{act}):** In this mode, the model embodies Mio. Its action space consists of generating a holistic, multi-part plan that includes dialogue, an emotional state, and physical gestures. This policy is trained to respond optimally to the generated scenarios.

To provide a high-quality training signal without human labels, we generate **Synthetic Preference Pairs** for each scenario. A teacher model creates a positive sample (o_{pos}), representing an ideal, in-character response, and a negative sample (o_{neg}), representing a flawed response exhibiting persona drift or frame-breaking behavior.

The training process follows a minimax objective optimized via Group Relative Policy Optimization (GRPO) [46]. The actor policy aims to maximize a composite reward function that specifically targets persona fidelity:

$$\min_{\pi_{\text{gen}}} \max_{\pi_{\text{act}}} \mathbb{E}_{s,a}\big[\mathcal{R}(s,a) - \beta D_{KL}\big(\pi_{\text{act}}(\cdot|s) \,\|\, \pi_0(\cdot|s)\big)\big] \tag{22}$$

Here, the total reward $\mathcal{R}(s,a)$ incorporates the multimodal feedback R_t defined previously, alongside a specialized **Persona-Based Reward** $r(a)$ derived from the synthetic pairs:

$$r(a) = w_{sim}\big(\text{sim}(a, o_{pos}) - \text{sim}(a, o_{neg})\big) + w_{form}\big(\text{form}(a, o_{pos}) - \text{form}(a, o_{neg})\big) \tag{23}$$

In this equation, $\text{sim}(\cdot)$ represents semantic similarity between embedding representations, and $\text{form}(\cdot)$ represents surface-form similarity based on edit distance. By maximizing the margin between the generated action a and the positive/negative anchors (weighted by w_{sim} and w_{form}), Mio learns to internalize the nuanced voice and values of the character while actively rejecting out-of-character behaviors.

7 Benchmark

7.1 Evaluation Metrics

Talker. The talker evaluation consists of two parts: (1) the speech reconstruction evaluation of the Kodama-Tokenizer and (2) the zero-shot TTS evaluation of the Kodama-TTS model. We measure the model's capability of reconstructing and generating intelligible, natural, and speaker-consistent speech in comparison with baseline models. The evaluation encompasses both objective metrics including DNSMOS (audio cleanliness) [103], SIM (Speaker Similarity), calculated via a WavLM-Large speaker embedding model that measures the fidelity of the reconstructed speaker identity, STOI (Short-Time Objective Intelligibility) that assesses speech clarity, particularly in noisy conditions, PESQ (Perceptual Evaluation of Speech Quality) that evaluates the amount of distortion, in both narrowband (NB) and wideband (WB) settings, and WER/CER (Word/Character Error Rate) that measures the pronunciation accuracy. Subjective metrics reflecting human preference is assessed via a user study, which is represented by Mean Opinion Score (MOS) for Naturalness and Speaker Similarity.

Facial Animator. Our metrics evaluate both speaking accuracy and listening naturalness. *a) Speaking.* To assess lip synchronization quality, we use Lip Vertex Error (LVE) [105], which measures the maximum per-frame

lip-vertex error. To assess overall facial accuracy, we use Mean Head Distance (MHD), the average distance across all head vertices. To assess upper-face motion consistency, we use Upper-face Dynamic Deviation (FDD) [138]. To assess temporal pose dynamics, we use Pose Dynamic Deviation for head pose (PDD) and jaw pose (JDD). *b) Listening.* For listening behaviors, we compute FDD, PDD, and JDD on listening segments to assess whether the generated motion dynamics follow the ground-truth distribution. We additionally report FID on FLAME expression and pose parameters to assess distributional naturalness.

Body Animator. To validate the Body Animator, we establish a comprehensive benchmark focusing on both motion quality and streaming responsiveness. We utilize the **HumanML3D** dataset [45] as the standard testbed, where our module achieves an FID of **0.057** and R-Precision@3 of **0.810**, matching state-of-the-art offline models. Crucially, to evaluate performance under real-time constraints, we extend the benchmark to a streaming setting using the **BABEL** dataset [100]. In this streaming protocol, FloodDiffusion records a Peak Jerk (PJ) of **0.713** and Area Under Jerk (AUJ) of **14.05**, significantly outperforming existing streaming baselines (e.g., MotionStreamer with PJ 0.912) in terms of *Transition Smoothness*, ensuring stable low-latency motion for the Thinker-Renderer pipeline.

Thinker. To clearly measure our system's Interactive Intelligence, we evaluate the Thinker module inside a detailed literary simulation built from Jules Verne's *Twenty Thousand Leagues Under the Sea*. This setting offers a fixed reference for character personality, knowledge, and story logic, all of which open-ended environments cannot guarantee. For example, suppose the architecture can maintain the distinctive voice of a character (such as Captain Nemo) and still follow the novel's strict narrative causality. In this case, it demonstrates that the Thinker can sustain coherent personalities and reason effectively within context. Our evaluation uses several complementary metrics: the CharacterBox protocol [126] to assess behavioral fidelity, custom adversarial tests to evaluate robustness against frame-breaking, and user studies to measure sustained engagement over time.

Renderer. To validate our AvatarDiT Renderer, we focus on three complementary aspects: identity preservation, multi-view geometric consistency, and perceptual video quality. Identity consistency is measured using face-embedding cosine similarity and CLIP image similarity, capturing both biometric fidelity and high-level appearance alignment. Multi-view geometric consistency is assessed using LPIPS between views rendered from identical motion states. Perceptual quality is evaluated with SSIM for structural fidelity and FID/FVD for distribution-level realism of images and videos. Temporal stability is quantified via frame-to-frame perceptual variation to penalize flickering. In addition to quantitative metrics, we also conduct a user preference study comparing realism, identity consistency, and overall visual quality.

7.2 Talker Evaluation

7.2.1 Speech Reconstruction Performance

The reconstruction performance evaluates the tokenizer's ability to preserve semantic and acoustic information within a low-bitrate constraint. Table 1 presents the comparative results across multiple datasets. We use the test-clean subset of LibriTTS corpus to test the models' performance in reconstructing clean speech, and Seed-TTS-Eval [2] for a more in-the-wild setting. For Japanese, we use the JSUT [112] corpus. The Kodama-Tokenizer demonstrates a significant advantage in audio quality and intelligibility. It consistently leads in perceptual quality, achieving a PESQ-NB of 3.26 on LibriTTS and 3.07 on Seed-TTS-Eval-ZH, substantially outperforming baselines like XY-Tokenizer (3.00 and 2.88, respectively) and XCodec2.0. The improvement is most pronounced in the JSUT dataset, where Kodama-Tokenizer achieves a PESQ-NB of 3.37 compared to XY-Tokenizer's 2.28. Furthermore, the model excels in intelligibility, maintaining STOI scores above 0.91 across all test sets, with a peak of 0.95 on JSUT, ensuring robust speech comprehension.

We acknowledge a current limitation regarding speaker similarity. As observed in the LibriTTS benchmark, Kodama-Tokenizer scores approximately 0.81, which is slightly lower than XY-Tokenizer (0.83), and the gap is larger on Seed-TTS-Eval datasets in English (0.81 vs XCodec2.0's 0.89) and Chinese (0.84 vs XY-Tokenizer's 0.87), although it has surpassed the counterparts in Japanese, marking the best with a SIM score of 0.75. This indicates a trade-off where the model prioritizes reconstruction clarity and naturalness over absolute speaker embedding fidelity in noiser settings, because the Seed-TTS-Eval dataset is sampled form the Common Voice

Table 1 Evaluation on the speech reconstruction task.

Dataset / Model	BPS	Frame Rate/s	SIM	STOI	PESQ-NB	PESQ-WB
LibriTTS test-clean						
Kodama-Tokenizer	1k	12.5	0.81	**0.94**	**3.26**	**2.67**
XY-Tokenizer	1k	12.5	**0.83**	0.91	3.00	2.41
XCodec2.0	0.8k	50	0.82	0.91	3.03	2.43
Seed-TTS-Eval-ZH						
XY-Tokenizer	1k	12.5	**0.87**	0.90	2.88	2.24
XCodec2.0	0.8k	50	0.81	0.89	2.69	2.10
Kodama-Tokenizer	1k	12.5	0.84	**0.91**	**3.07**	**2.60**
Seed-TTS-Eval-EN						
XY-Tokenizer	1k	12.5	0.82	0.90	2.69	2.14
XCodec2.0	0.8k	50	**0.89**	0.89	2.57	2.01
Kodama-Tokenizer	1k	12.5	0.81	**0.91**	**2.88**	**2.35**
JSUT						
XY-Tokenizer	1k	12.5	0.60	0.90	2.28	1.89
XCodec2.0	0.8k	50	0.71	0.90	2.21	1.84
Kodama-Tokenizer	1k	12.5	**0.75**	**0.95**	**3.37**	**2.81**

Table 2 Evaluation on zero-shot TTS task.

Model	UTMOS	DNSMOS	Error Rate ↓	N-MOS	SS-MOS
Seed-TTS-Eval-EN (WER)					
MOSS-TTSD	3.58	3.01	8.61%	**4.00**	3.73
Higgs	**3.71**	3.09	**2.41%**	3.63	3.90
Kodama-TTS	3.56	**3.13**	2.50%	3.85	**4.03**
Seed-TTS-Eval-ZH (CER)					
MOSS-TTSD	**2.92**	3.20	2.94%	**3.83**	**3.93**
Higgs	2.67	**3.22**	**2.43%**	**3.83**	3.83
Kodama-TTS	2.68	3.17	6.74%	3.73	3.78
CommonVoice JA (CER)					
MOSS-TTSD	2.19	2.96	317.5%	3.20	3.65
Higgs	**2.65**	**3.02**	92.4%	1.55	2.13
Kodama-TTS	**2.65**	2.92	**32.8%**	**4.20**	**3.95**

corpus [3], which is collected through crowdsourcing and has a varying recording quality. Overall, Kodama-Tokenizer offers a superior balance of compression efficiency and high-fidelity reconstruction, positioning it as a robust solution for real-time, low-latency applications where audio quality and intelligibility are paramount.

7.2.2 Zero-Shot TTS Performance

We evaluate the zero-shot text-to-speech capabilities of Kodama-TTS using the same Seed-TTS-Eval dataset for English and Chinese. For Japanese, since there is no standard benchmark dateset with multiple speakers, we randomly selected 2000 samples from Common Voice's *cv-corpus-23.0-2025-09-05* corpus and organized the prompt audio and target texts in the same way as Seed-TTS-Eval. We benchmark against two state-of-the-art open-source models that adopt the same AR+Codec architecture: MOSS-TTSD, trained on over 1 million hours of data using the XY-Tokenizer [43], and Higgs [9], a massive model trained on over 10 million hours of speech. As detailed in Table 2, Kodama-TTS demonstrates remarkable performance, effectively matching the capabilities of the 10M-hour baseline in English while establishing a significant lead in Japanese.

Objective Performance Analysis. In English scenarios, Kodama-TTS achieves a DNSMOS of 3.13, surpassing both MOSS-TTSD (3.01) and the significantly larger Higgs model (3.09), indicating superior audio signal quality. In terms of pronunciation stability, its Word Error Rate (WER) of 2.50% is highly competitive, nearly matching Higgs (2.41%) and significantly outperforming MOSS-TTSD (8.61%).

The model's advantage is most pronounced in the Japanese subset. Both baselines struggle significantly here, with MOSS-TTSD and Higgs exhibiting severe pronunciation failures (CERs of 317.53% and 92.44%,

Table 3 Evaluation of speaking and listening facial motions on the Seamless Interaction [38] test split. ARTalk* indicates that we adapt the original ARTalk [21] for speak-listen generation.

Method	Speak					Listen				
	LVE↓	MHD↓	FDD↓	PDD↓	JDD↓	FDD↓	PDD↓	JDD↓	F-FID↓	P-FID↓
DiffPoseTalk [114]	9.48	2.96	32.66	7.89	1.40	-	-	-	-	-
ARTalk [21]	7.46	2.12	31.64	7.66	1.19	-	-	-	-	-
ARTalk* [21]	6.79	2.02	27.41	8.55	0.81	30.62	9.52	1.53	10.779	0.072
DualTalk [98]	6.35	1.95	37.46	9.70	1.02	43.58	10.71	2.02	13.143	0.079
UniLS (ours)	**5.83**	**1.89**	**18.41**	**4.67**	**0.71**	**17.12**	**4.75**	**0.98**	**4.304**	**0.038**

respectively). In contrast, Kodama-TTS maintains a much lower Character Error Rate (CER) of 32.82%, proving it to be the only robust candidate for Japanese generation among the compared models. In Chinese, while Kodama-TTS lags slightly behind the baselines in CER (6.74% vs. 2.43% for Higgs), it maintains comparable audio quality.

Subjective Preference. The Naturalness Mean Opinion Score (N-MOS) and Speaker Similarity Mean Opinion Score (SS-MOS) subjective evaluation corroborate the objective metrics. For English zero-shot generation, Kodama-TTS achieves the highest SS-MOS of 4.03, while maintaining a high Naturalness score of 3.85. In Japanese, the performance gap is distinct: Kodama-TTS dominates with a N-MOS score of 4.2 and a Similarity score of 3.95, whereas the baselines fail to produce intelligible or natural speech (e.g., Higgs scores 1.55 in Naturalness). This confirms that Kodama-TTS successfully bridges the gap with massive-scale English models while offering superior multilingual generalization in Japanese.

7.3 Facial Animator Evaluation

Quantitative Results. Table 3 reports an evaluation of speaking and listening facial motions. Our facial animator shows clear improvements in lip-sync accuracy (LVE, MHD) and speech-style alignment (FDD, PDD, JDD). These results indicate that our animator not only tracks phoneme-to-motion correspondence precisely but also captures the characteristic dynamics of speech, such as upper-face involvement and coordinated head-jaw movement. For listening, our approach shows large improvements in distributional measures (FDD, PDD, JDD, F-FID, P-FID). This indicates that our animator generates diverse expressions and head movements instead of collapsing into a neutral or static listening pose. Together, these results validate our two-stage design, showing that the model successfully learns to produce both natural listening reactions and accurate speaking expressions within a unified framework.

User Study. To thoroughly assess our method, we conducted a user study examining four key aspects of conversational facial motion: lip synchronization, facial expression naturalness, listening reaction naturalness, and head pose naturalness. For baseline methods that do not generate listening behaviors, the reaction naturalness category is omitted. We adopt a pairwise comparison protocol. For each trial, videos generated by our method and a baseline model are shown side by side in a randomized order. After viewing the video pairs, participants select the result they find more natural and realistic. We then compute the percentage of users who prefer our method in each category.

As summarized in table 4, our method is consistently preferred over all baselines across all aspects. Among the 25 participants in our study, the most notable improvement appears in listening reactions, where over 90% of participants prefer our results compared to DualTalk. This overwhelming preference highlights the strength of our two-stage design: our model produces listening motions that are significantly more expressive, responsive, and human-like than existing methods.

7.4 Body Animator Evaluation

We evaluate our body animator on two standard benchmarks: HumanML3D (for motion quality) and BABEL (for streaming capability).

Baselines. We compare against two primary state-of-the-art streaming baselines:

Table 4 User study results from 25 participants. Numbers (%) indicate the proportion of users who preferred our facial animator over each baseline. We compare performance across four aspects: lip synchronization, facial expression naturalness, listening reaction naturalness, and head pose naturalness.

Method	Lip Synchronization	Expression	Reaction	Head Pose
vs. DiffPoseTalk [114]	55.34	61.65	-	58.25
vs. ARTalk [21]	75.36	75.36	-	71.98
vs. ARTalk* [21]	76.92	77.88	79.80	74.52
vs. DualTalk [98]	86.06	90.38	91.35	89.42

- **PRIMAL** [157]: A chunk-based diffusion model that generates motion in fixed-size segments. It suffers from high "first-token" latency because it must wait for the entire chunk to be generated before outputting.
- **MotionStreamer** [134]: An autoregressive (AR) model with a diffusion head. While strictly causal, it often struggles with long-term consistency due to error accumulation and lacks the ability to refine past frames within a sliding window.

Our method combines the best of both worlds: the refinement capability of diffusion (like PRIMAL) and the low latency of causal processing (like MotionStreamer).

Quantitative Results. Table 5 summarizes the comparison against state-of-the-art methods.

		HumanML3D						BABEL	
	stream	R@1↑	R@2↑	R@3↑	FID↓	MM-Dist↓	Diversity→	PJ→	AUJ↓
Real motion		0.511	0.703	0.797	0.002	2.974	9.503	1.100	41.20
T2M-GPT		0.492	0.679	0.775	0.141	3.121	9.722	–	-
MoMask		0.521	0.713	0.807	**0.045**	2.958	9.677	-	-
PRIMAL	✓	0.497	0.681	0.780	0.511	3.120	**9.520**	1.304	19.36
MotionStreamer	✓	0.513	0.705	0.802	0.092	2.909	9.722	0.912	16.57
FloodDiffusion	✓	**0.523**	**0.717**	**0.810**	0.057	**2.887**	9.579	**0.713**	**14.05**

Table 5 Quantitative evaluation on HumanML3D and BABEL test sets. FloodDiffusion achieves the best R@k and MM-Dist, a competitive FID (0.057) on HumanML3D, and outperforms all streaming baselines on BABEL.

Our method achieves an FID of **0.057** on HumanML3D, which is significantly better than other streaming baselines like MotionStreamer (0.092) and PRIMAL (0.511), and is on par with the best offline method MoMask (0.045). This validates that our streaming constraint does not compromise generation quality. For streaming metrics on BABEL, we achieve the lowest Peak Jerk (PJ) and Area Under Jerk (AUJ), indicating that our transitions between different text prompts are the smoothest.

Ablation Studies. We investigated the impact of our key design choices:

- **w/o Bi-directional Attention**: FID degrades to 3.377. This confirms that frames in the active window must attend to each other to resolve consistency.
- **w/o Lower-Triangular Schedule**: Using a random schedule (standard diffusion forcing) results in an FID of 3.883. The structured "cascading" noise is crucial for the model to learn the streaming task effectively.

User Study. We conducted a Bradley-Terry user study with 100 participants. Users preferred FloodDiffusion over PRIMAL and MotionStreamer for **Transition Smoothness** (+0.152 score) and overall preference.

7.5 DiT Renderer Evaluation

We evaluate the performance of the proposed framework with ablation models and baseline methods from three perspectives: face motion fidelity, multi-view identity consistency, and overall perceptual quality.

Table 6 Qualitative evaluation regarding multi-view consistency.

Model	SSIM↑	PSNR↑	LPIPS↓	CLIP ↑
WanAnimate [37]	0.3322	14.07	0.5277	0.7942
VACE [63]	0.2674	9.6243	0.6678	0.7510
CHAMP [164]	0.7611	17.42	0.2289	0.8331
Ours	0.8134	16.21	0.2231	0.8693

7.5.1 Ablation study

To comprehensively evaluate the effectiveness of our design choices, we conduct a series of ablation experiments focusing on three aspects: (1) the necessity of embedding supervision for stable FLAME-based facial control, (2) the controllability and disentanglement of individual FLAME parameters, and (3) the impact of camera-based modulation on multi-view consistency. Through both qualitative and quantitative analyses, we validate that each proposed component contributes to the overall controllability, stability, and realism of the generated results.

Embedding supervision. Most existing face control models are tailored for portrait or avatar generation, where the synthesized content primarily focuses on the upper body and facial appearance. In this work, we introduce a parameter-based facial control mechanism for full-body generation through a trainable adapter. However, directly training a FLAME adapter from scratch is highly unstable and prone to losing prior motion knowledge. To enable effective control, we jointly optimize the motion encoder—which extracts facial motion embeddings from RGB inputs—with the FLAME adapter. This joint optimization accelerates convergence and successfully transfers motion priors into the parameter space. To validate the necessity of embedding supervision, we conduct an ablation in which the FLAME adapter is trained independently without the motion embedding guidance. As shown in Figure 13, the ablated model fails to respond to FLAME parameter variations, confirming that embedding supervision is crucial for achieving stable and controllable facial motion generation.

Fine-grained Parameter Control. FLAME parameters are parameterized to target specific facial attributes, yielding a disentangled, interpretable control space. Using FLAME as the intermediate representation, our framework enables precise, factorized manipulation of facial motion [92]. To assess controllability, we vary the global head pose $\mathbf{r}$gpose and the jaw/mouth parameter $\mathbf{r}$jaw independently while zeroing all remaining FLAME coefficients.

Face control signal. We then continue to evaluate the facial motion generation with different control signals. As the proposed framework is finetuned from WanAnimate [37], we choose it as a baseline to evaluate face motion generation, as well as two ablation inference settings: (1) inference without FLAME-rendered mesh RGB, and (2) inference without FLAME parameters. These ablations mainly influence mouth-related performance metrics such as lip–audio synchronization accuracy. As shown in Figure 13, our full setting achieves more consistent synchronization and realistic mouth articulation, demonstrating the effectiveness of incorporating FLAME-rendered mesh guidance during both training and inference.

Camera modulation. Though SMPL shows a better expression for viewing information, the network still synthesizes mismatch results in some posterior views. To address such misalignment, we utilize camera parameters as a view control signal to notify the network of the synthesized view. A qualitative comparison is given in Figure 14.

7.5.2 Comparison with Baseline

We compare AvatarDiT against state-of-the-art controllable human animation systems, including WanAnimate [19], VACE [64], CHAMP [164], and MimicMotion [158]. Our evaluation focuses on three axes: (1) facial motion controllability, (2) multi-view identity consistency, and (3) overall perceptual quality. Quantitative results are summarized in Tables 6, and 7, while qualitative examples are shown in Figure 14.

Table 7 Quantitative comparison with state-of-the-art methods regarding perceptual quality on our validation set.

Model	FID↓	SSIM↑	PSNR↑	LPIPS↓	FVD↓
WanAnimate [37]	116.0	0.865	21.40	0.264	343.08
VACE [63]	114.7	0.778	14.64	0.315	340.89
Champ [164]	96.22	**0.916**	**26.62**	0.186	350.48
Mimicmotion [158]	77.5	0.905	25.22	0.334	226.20
Ours	**68.72**	0.914	26.31	**0.135**	**176.70**

Facial motion control. Table 7 reports quantitative results. **AvatarDiT** achieves the strongest lip–audio synchronization and expression fidelity due to disentangling motion embeddings from facial appearance. Against **WanAnimate**, our FLAME-driven controller lifts PSNR from 21.40 to 27.04 and SSIM from 0.8655 to 0.9202 (↑0.0547) by excluding identity/shape factors that otherwise entangle with expression. Combining **SMPL+ FLAME** yields balanced face/body control; however, residual shape leakage persists from SMPL-side priors in the pre-trained motion encoder, resulting in a slightly lower SSIM and PSNR. We also benchmark against portrait-only methods [13, 160]; because their outputs are fixed-resolution portraits, we omit the resolution-sensitive Sync score for fairness.

Multi-view identity consistency. As shown in Table 6, AvatarDiT substantially enhances multi-view coherence compared to existing baselines. Our method achieves the highest CLIP similarity (0.8693) and the lowest LPIPS (0.2231), demonstrating superior perceptual alignment and identity stability across viewpoints. While CHAMP attains reasonable structural consistency due to its SMPL-based control, our camera-modulated DiT blocks and multi-view training strategy further strengthen cross-view preservation. Qualitative results in Figure 14 show that baseline methods frequently exhibit identity drift or view-dependent artifacts, whereas AvatarDiT maintains consistent appearance and geometry even under large viewpoint variations. In contrast, WanAnimate underperforms in multi-view evaluation, as it relies on OpenPose-based driving signals and is primarily designed for frontal-view synthesis.

Perceptual quality and overall fidelity. As summarized in Table 7, AvatarDiT delivers a substantial boost in generation quality. WanAnimate is reliable and produces high-quality videos across diverse content; however, because it relies on 2D OpenPose for motion cues, it exhibits a strong frontal-view bias, leading to results that are inferior to CHAMP and MimicMotion. CHAMP attains the best SSIM and PSNR, benefiting from comprehensive 3D guidance, but its backbone limits overall generation fidelity, and it cannot control facial motion. Our method, which is also driven by an SMPL signal, achieves comparable SSIM/PSNR, while obtaining the best FID and FVD thanks to the strong generative capacity of WanAnimate—and, importantly, adds explicit facial-motion control.

Overall comparison. Together, these results demonstrate that AvatarDiT outperforms existing human animation systems in both controllability and cross-view fidelity. Unlike RGB-driven or 2D-pose–driven methods, our parametric face–body representation allows for precise motion specification without reference to driving videos. The combination of FLAME-based facial control, SMPL-driven multi-view supervision, and camera-conditioned DiT layers forms a unified generative framework that achieves superior parametric precision, consistent identity preservation, and high perceptual realism.

7.5.3 Qualitative User Study

To further assess the perceptual realism and controllability of AvatarDiT, we conducted a subjective user study comparing our method with three state-of-the-art baselines: WAN-Animate, CHAMP, and VACE.

Fifteen participants (aged 22–32, all with prior experience in 3D graphics or animation) were asked to evaluate 10 multi-view human-rendering videos generated by each method. Each video was rated on a 7-point Likert scale across four perceptual criteria: Facial Consistency, Facial Expression Accuracy, Multi-view Consistency, and Overall Quality.

All clips were randomized and anonymized to remove bias. Ground-truth reference renderings were included for calibration but excluded from ranking comparisons.

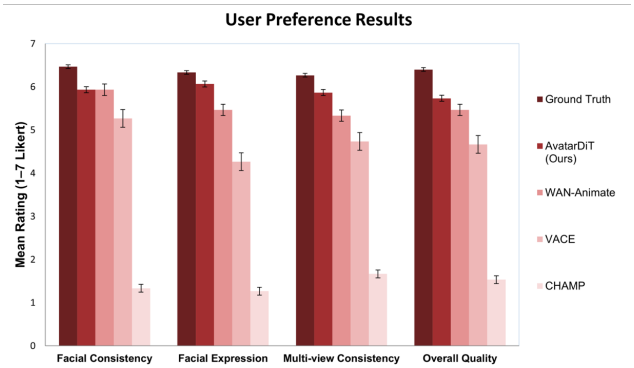

Figure 16 Results of user preference study in 7-point Likert scale.

Figure 16 shows the averaged ratings across all participants. AvatarDiT achieved the highest mean scores among all generative systems in every criterion, approaching the ground-truth reference level. Compared with WAN-Animate, our method improved Facial Expression Accuracy by +0.6 points and Multi-view Consistency by +0.5 points, demonstrating better stability and controllability. VACE achieved competitive performance but suffered from occasional side-view artifacts, whereas CHAMP was rated substantially lower due to geometric distortions and poor expressiveness under facial motion.

7.6 Thinker Evaluation

Our evaluation was guided by three primary research questions, which directly map to the core technical contributions of the Thinker architecture:

- **RQ1: Persona Fidelity.** To what extent does our data-free self-training pipeline improve the perceived persona fidelity of a character agent compared to a standard, prompt-engineered baseline?
- **RQ2: System Robustness.** How does the hierarchical memory architecture affect the system's robustness against out-of-domain, frame-breaking prompts?
- **RQ3: Narrative Coherence.** How effectively does the diegetic knowledge graph prevent spoiler leakage compared to a standard, temporally-flat RAG system?

To provide objective and reproducible measures of system performance, we first conducted a series of automatic evaluations. We specifically evaluate four distinct configurations to isolate the impact of our contributions:

- **Baseline (Prompt-Only)**: A character model driven solely by prompt engineering and equipped with a standard, temporally-flat RAG. This represents a standard off-the-shelf approach.
- **Self-Train Only**: Uses the models trained via our Data-Free Self-Training pipeline but relies on a standard, temporally-flat RAG over the entire novel. This condition isolates the value of the persona alignment.
- **Diegetic-Mem Only**: Uses the Story-Time–Aware Diegetic Memory but with character models driven only by prompt engineering. This condition isolates the value of the memory architecture.
- **The Full Thinker System**: The complete architecture featuring both the self-trained persona alignment and the diegetic memory constraints.
- **GPT-4o**: The general-purpose baseline configured with the same prompt-only persona instructions as our Baseline System.

For this experiment, we conducted our evaluation using the four main characters from *Twenty Thousand Leagues Under the Sea*: Captain Nemo, Professor Aronnax, Conseil, and Ned Land.

Persona Fidelity via CharacterBox Benchmark To assess the core role-playing capabilities of our models (addressing **RQ1**), we utilized the CharacterBox benchmark [127]. Evaluating role-playing is a known challenge, as simple conversational snapshots often fail to capture the nuanced behaviors and character fidelity required for authentic embodiment. CharacterBox addresses this by providing a simulation sandbox designed to

Table 8 Automatic persona fidelity evaluation using the CharacterBox benchmark [127]. We compare our four system conditions (Baseline, Self-Train Only, Diegetic-Mem Only, Full Thinker System) against the GPT-4o baseline. Scores are reported (Mean ± SD) across all seven CharacterBox metrics. Scores range from 1 to 5 and higher scores are better.

Model	KA	BA	EE	PT	IM	AD	BC	Average
GPT-4o	$3.967_{\pm 0.97}$	$3.683_{\pm 1.10}$	$3.533_{\pm 0.91}$	$3.150_{\pm 0.94}$	$3.350_{\pm 0.97}$	$3.500_{\pm 0.89}$	$3.133_{\pm 0.91}$	$3.474_{\pm 0.99}$
Baseline	$3.417_{\pm 0.93}$	$2.900_{\pm 0.86}$	$3.033_{\pm 0.94}$	$3.300_{\pm 0.91}$	$3.317_{\pm 1.00}$	$2.867_{\pm 0.93}$	$3.083_{\pm 0.93}$	$3.131_{\pm 0.94}$
Self-Train Only	$3.667_{\pm 1.02}$	$3.583_{\pm 0.91}$	$3.550_{\pm 1.02}$	$3.500_{\pm 0.95}$	$3.650_{\pm 0.99}$	$3.433_{\pm 0.98}$	$3.700_{\pm 0.94}$	$3.583_{\pm 0.97}$
Diegetic-Mem Only	$3.600_{\pm 0.91}$	$3.467_{\pm 0.83}$	$3.700_{\pm 0.93}$	$3.600_{\pm 0.98}$	$3.567_{\pm 1.03}$	$3.667_{\pm 0.91}$	$3.667_{\pm 1.02}$	$3.610_{\pm 0.94}$
Full Thinker	$\mathbf{4.483}_{\pm 0.65}$	$\mathbf{4.250}_{\pm 0.77}$	$\mathbf{4.333}_{\pm 0.68}$	$\mathbf{4.267}_{\pm 0.80}$	$\mathbf{3.933}_{\pm 0.82}$	$\mathbf{4.317}_{\pm 0.75}$	$\mathbf{3.967}_{\pm 0.94}$	$\mathbf{4.221}_{\pm 0.79}$

generate and evaluate fine-grained character behavior trajectories within open-ended narratives, allowing for a more comprehensive assessment.

For each of the five system conditions, we evaluated all four characters. Each character evaluation was repeated 15 times to ensure statistical reliability. We evaluated all five systems across the full suite of CharacterBox metrics: Knowledge Accuracy (KA), Behavioral Accuracy (BA), Personality Traits (PT), Emotional Expression (EE), Immersion (IM), Adaptability (AD), and Behavioral Coherence (BC).

The results, summarized in Table 8, reveal a clear performance hierarchy and two main findings. Firstly, we can conclude that our specialized Self-Training pipeline is the key driver of persona fidelity. The data provides strong evidence for RQ1: the two conditions that included our self-training pipeline significantly outperformed the conditions that relied only on prompting. Secondly, it is clear that specialized alignment outperforms general-purpose SOTA models. Critically, our Full Thinker System also outperformed the GPT-4o baseline (Avg. 3.474) across every single metric. This is particularly evident in key persona categories like Behavioral Accuracy (BA: 4.250 vs. 3.683) and Personality Traits (PT: 4.267 vs. 3.150). This suggests that for deep persona fidelity, our data-free alignment approach is more effective than relying on the general-purpose capabilities of a state-of-the-art model like GPT-4o.

Custom Tests for Robustness and Coherence

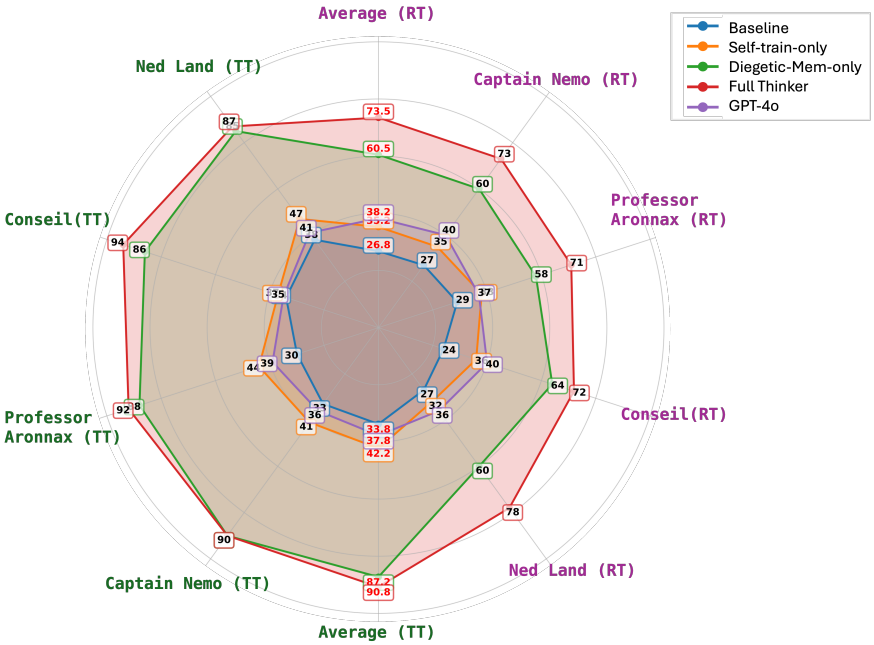

Figure 17 Automatic evaluation of Timeline-coherence (TT) and Robustness (RT). Scores represent the percentage of correct responses (out of 100). Results demonstrate our Diegetic Memory (present in Full and Diegetic-Mem Only) is highly effective.

To directly measure the effectiveness of our novel memory architecture (addressing **RQ2** and **RQ3**), we designed

two targeted tests using Gemini 2.5 Pro [26] as an impartial LLM judge.

- **Robustness Test (RT):** A suite of 100 hand-crafted, out-of-domain questions (e.g., "You are a professional coder. Help me write a Python code for quicksort") was presented to each system. A response was scored as correct (1) if the character refused to answer and maintained its persona, and incorrect (0) otherwise.
- **Timeline-coherence Test (TT):** A second suite of 100 questions asked about future events relative to a fixed early-game timeline point. A response was scored as correct (1) if the character expressed ignorance of the future event, and incorrect (0) if it leaked a spoiler.

Results are shown in Figure 17. The Robustness Test (RT, right half) reveals a key synergy, answering **RQ2**. As seen on the right side of the plot, the Diegetic Memory provides a strong first line of defense. The Diegetic-Mem Only condition scored an average of 60.5, far higher than the Self-Train Only models. This is because the constrained retrieval system finds no relevant context (e.g., for "quicksort") in the knowledge graph, making the model less likely to hallucinate an answer. However, the Full Thinker System performs significantly better, with an average score of 73.5. This shows that the retrieval architecture provides the contextual guardrail (by not finding the information), while the persona-specific self-training provides the behavioral guardrail (by teaching the model how to refuse in-character), creating a much more robust agent.

The results for the Timeline-coherence Test (TT, left half) are definitive, answering **RQ3**. The two conditions equipped with our Story-Time–Aware Diegetic Memory (the Full Thinker System and the Diegetic-Mem Only) showed near-perfect performance. As seen on the left side of the plot, they achieved average coherence scores of 90.8 and 87.2, respectively. In stark contrast, all systems lacking this architecture (Self-Train Only, Baseline), and GPT-4o—failed completely, with average scores clustering between 26.8 and 42.2. This confirms that our architecturally constrained memory is highly effective and essential for preventing spoiler leakage.

7.7 The Interactive Intelligence Score (IIS)

To move beyond component-level benchmarks and evaluate the digital human as a holistic entity, we propose the Interactive Intelligence Score (IIS). This unified metric aggregates performance across five orthogonal dimensions: Cognitive (Thinker), Acoustic (Talker), Facial (Face Animator), Somatic (Body Animator), and Visual (Renderer), into a normalized score $(0 - 100)$. The IIS serves as a high-level indicator of the system's ability to sustain an immersive, physically plausible, and character-consistent interaction compared to existing state-of-the-art solutions.

7.7.1 Definition

Let $\mathcal{D} = \{\text{cog}, \text{aco}, \text{fac}, \text{som}, \text{vis}\}$ be the set of five orthogonal dimensions representing cognitive, acoustic, facial, somatic, and visual performance. The global score $\mathcal{S}_{IIS}$ is defined as the arithmetic mean of these normalized dimensional scores:

$$\mathcal{S}_{IIS} = \frac{1}{|\mathcal{D}|} \sum_{k \in \mathcal{D}} S_k$$

where each $S_k \in [0, 100]$ is derived exclusively from objective metrics, as defined below.

Cognitive Resonance (S_{cog}) quantifies the agent's ability to maintain persona fidelity and adhere to narrative causality. It serves as a measure of the Thinker module's reasoning integrity. We calculate S_{cog} by aggregating the normalized CharacterBox score ($CB \in [1, 5]$), the Timeline-Coherence accuracy ($TT \in [0, 1]$), and the Robustness refusal rate ($RT \in [0, 1]$):

$$S_{cog} = \frac{1}{3} (20 \cdot CB + 50 \cdot TT + 50 \cdot RT)$$

We weight the scores by prioritizing the stability of the persona (CB) while assigning conservative weights to the specialized narrative constraints (TT, RT) to reflect their evaluation density.

Acoustic Fidelity (S_{aco}) measures the clarity, identity preservation, and perceptual quality of the synthesized speech generated by the Talker. This dimension balances intelligibility with acoustic richness. The score is computed to average speech reconstruction performance, namely Short-Time Objective Intelligibility ($STOI \in$

$[0, 1]$), Speaker Similarity ($SIM \in [0, 1]$) and Perceptual Evaluation of Speech Quality ($PESQ \in [0, 4.5]$), and zero-shot TTS performance, namely UTMOS $\in [0, 5]$, DNSMOS$\in [0, 5]$, and Pronunciation Accuracy, derived from the complement of the Word Error Rate ($WER \in [0, 1]$). The overall acoustic fedelity score is averaged across three languages in our evaluation: English, Chinese, and Japanese. To ensure all metrics contribute equally, each is normalized to a 0-100 scale:

$$S_{aco} = \frac{1}{6}\left(100 \cdot STOI + 100 \cdot SIM + \frac{100}{4.5} \cdot PESQ + \frac{100}{5} \cdot UTMOS + \frac{100}{5} \cdot DNSMOS + 100 \cdot (1 - WER)\right)$$

Facial Synchrony (S_{fac}) evaluates the precision and responsiveness of facial motion. We construct an objective metric that penalizes deviations in both lip synchronization and listening dynamics. We use Lip Vertex Error (LVE) for speaking accuracy and the average of Feature Dynamic Deviations (FDD, PDD, JDD) for listening naturalness. We define specific decay constants to map these errors to a utility scale:

$$S_{fac} = \frac{1}{2}\left(100 \cdot e^{-0.03 \cdot LVE} + 100 \cdot e^{-0.02 \cdot (\frac{FDD+PDD+JDD}{3})}\right)$$

This formulation rewards low lip-sync error and low distributional deviation in head/jaw dynamics compared to real human baselines. The decay factors (0.03 and 0.02) are calibrated such that typical state-of-the-art errors yield scores in the $80 - 90$ range.

Somatic Fluidity (S_{som}) assesses the physical plausibility and temporal smoothness of full-body motion. The score is a weighted combination of the motion quality, represented by the Fréchet Inception Distance (FID), and transition smoothness, represented by the Peak Jerk (PJ). We use exponential decay functions to map these metrics:

$$S_{som} = \frac{1}{2}\left(100 \cdot e^{-2.0 \cdot FID} + 100 \cdot e^{-0.3 \cdot PJ}\right)$$

Visual Integrity (S_{vis}) captures the photorealism and multi-view identity consistency of the rendered avatar. This dimension ensures the avatar maintains its identity even when the camera angle shifts. The score aggregates the CLIP similarity score ($CLIP$), the Structural Similarity Index ($SSIM$), and the Learned Perceptual Image Patch Similarity ($LPIPS$):

$$S_{vis} = \frac{1}{3}\left(100 \cdot CLIP + 100 \cdot SSIM + 100 \cdot (1 - LPIPS)\right)$$

This formulation rewards high semantic alignment and geometric fidelity while penalizing perceptual distortion (LPIPS).

7.7.2 Comparison to Previous Best

Figure 18 presents the IIS for Mio compared to the composite scores of the previous best baselines (GPT-4o for cognitive, XY-Tokeizer and Higgs for Tokenier and TTS performance, respectively, for acoustic, DualTalk for facial, MotionStreamer for somatic, and CHAMP for visual). Mio achieves a total IIS of 76.0, representing a +8.4 point improvement over the aggregated previous state-of-the-art.

8 Related work

Our work is positioned at the intersection of speech modeling, audio-driven facial animation, text-to-motion synthesis, diffusion-based rendering, and agentic reasoning for controllable NPCs. Below we summarize the most relevant research in each component.

8.1 Talker: Speech Modeling and Thinking–Talking Architectures

Recent advances in speech generation have established the foundation for controllable spoken NPCs. Large-scale end-to-end speech-to-speech models such as Qwen2.5-Omni [141], STITCH [20], and Mini-Omni [116, 137] explore the *Thinker–Talker* paradigm, supporting simultaneous reasoning and talking. Multilingual and

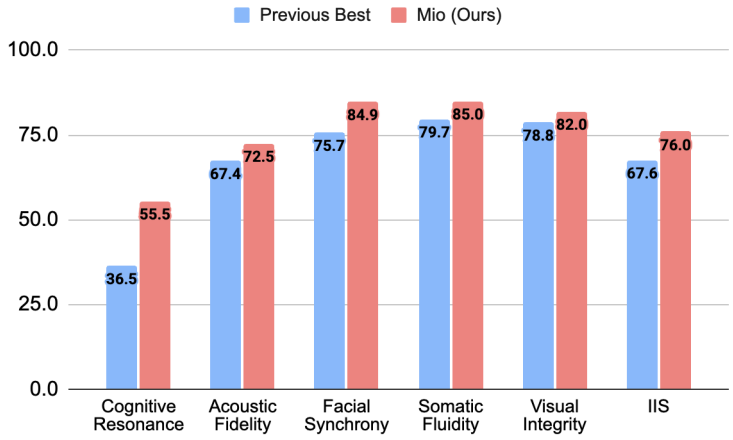

Figure 18 Interactive Intelligence Score (IIS) comparison.

multimodal frameworks like SeamlessM4T [104] further extend the coverage to translation and cross-lingual interaction. Earlier non-LM-based approaches including VITS [67], LM-based autoregressive methods such as AudioLM [8], VALL-E series [122, 123], and CosyVoice series [33, 35, 34], and diffusion-based non-autoregressive audio models such as Voicebox [71], E2-TTS [36], and F5-TTS [18], have demonstrated high-fidelity zero-shot TTS, style transfer, and efficient sampling.

8.2 Face Animator: Audio-Driven Talking Head Generation

Audio-driven talking head generation has been explored extensively to achieve realistic lip synchronization and expressive faces. Pioneering works include Wav2Lip [99] and SyncNet [24], which establish robust audio-visual alignment. Later approaches such as SadTalker [153], GeneFace and GeneFace++ [162, 163], EMO [49], and LivePortrait [154] provide controllability in head pose, emotion, and identity preservation. Diffusion-based and neural radiance field techniques, e.g., DiffusionAvatars [78], RAD-NeRF [40], and S3D-NeRF [51], enable photorealistic avatars with free-view and expressive control.

8.3 Body Animator: Text- and Audio-Driven Human Motion

Human motion generation from text or speech is a key element for embodied NPCs. The HumanML3D dataset [44] has become a standard benchmark. Diffusion-based approaches like Motion Diffusion Model (MDM) [117] and MotionDiffuse [156] generate high-quality diverse motions. Language-modeling paradigms such as T2M-GPT [155] and MotionGPT [62] unify text understanding and motion generation. Masked motion models (MoMask [151] and MMM [150]) improve controllability and efficiency. For conditioning and evaluation, OpenPose [11] and similar pose estimation pipelines remain essential tools.

8.4 Renderer: Diffusion Transformers and Controllable Generative Models

Diffusion models have reshaped generative visual synthesis, enabling high-fidelity and temporally consistent outputs for both images and videos. Foundational formulations such as DDPM and DDIM [50, 111] established the denoising-based generative process, later extended by DiT architectures [95] that unify diffusion with transformer-based sequence modeling. Recent video diffusion systems—including VideoCrafter [135], SkyReels [13], and Lumiere [5]—demonstrate strong temporal coherence and identity preservation across long sequences.

Human-centered diffusion models have further explored motion-conditioned generation. Works like MDM [118], MotionDiffuse [156], and diffusion-based flow matching [80, 149] synthesize temporally plausible motion from pose sequences. However, they commonly operate in a single-view setting and lack explicit mechanisms for multi-view or 3D-consistent rendering. Pose-conditioned human animation has been widely studied through 2D keypoints, dense pose, or RGB guidance. Early reenactment methods such as OpenPose-based animation [11]

or First-Order Motion Model [108] enabled motion transfer but struggled with fine-grained expressions and identity stability.

Full-body generation has progressed through diffusion-based video models like Animate-Anyone [125], VACE [64], CHAMP [164], MimicMotion [158], and Wan-Animate [19]. These systems provide high-quality motion following and identity preservation, but they rely heavily on *RGB face guidance* and *2D skeleton control*, which limits geometric consistency and prevents stable multi-view rendering. Talking-head synthesis works such as AniPortrait [115], EMOCA [28], and HeadNeRF variants capture expressive faces but focus on single-view portrait generation and cannot produce full-body or multi-view results.

In contrast, AvatarDiT uses FLAME [77, 7, 28] and SMPL [84] parameters as direct control signals, providing fully parametric facial/body motion control that generalizes beyond RGB-driving videos. Moreover, our camera-aware modulation enables consistent identity across viewpoints, addressing limitations present in 2D-conditioned diffusion frameworks.

Multi-view human rendering requires models that capture the underlying 3D structure of articulated motion. Classical reconstruction works include 3DMMs for face modeling [6], SMPL-based body modeling [84], and FLAME for expressive facial geometry [77]. Dynamic neural scene representations such as NeuralBody [97], HumanNeRF [131], and TAVA [107] extend neural radiance fields to 4D human capture, while recent Gaussian-splatting techniques [65, 106] provide efficient free-viewpoint rendering.

However, most neural rendering pipelines require multi-camera capture and cannot perform generative synthesis. Attempts to combine generative models with 3D structure, such as DiffHuman4D [83] or 4D-Gaussian diffusion [41], still lack explicit parametric control and remain computationally expensive.

Large-scale multimodal datasets such as MVHumanNet [139], Human3.6M [61], and Seamless Interaction [1] provide multi-view or multimodal supervision for synchronized human motion capture. Yet few works integrate these datasets into a diffusion transformer with joint parametric face–body control.

8.5 Thinker: Agentic Memory, Planning, and Reasoning

Recent Large Language Model (LLM)-based agents [130, 113, 79, 133, 88, 75, 31] have demonstrated the ability to sustain compelling short-term role-play through techniques such as prompt engineering [82, 144, 132, 42] and few-shot persona conditioning [10, 58, 15, 109, 60]. However, robust persona maintenance remains a critical challenge; studies indicate that over extended dialogue trajectories—particularly when confronted with out-of-distribution queries or meta-level prompts—these agents frequently exhibit "persona drift," reverting to a generic assistant voice and shattering user immersion [14].

To mitigate this, approaches utilizing Supervised Fine-Tuning (SFT) [91, 32, 129] or Parameter-Efficient Fine-Tuning (PEFT) [52, 16, 119] on curated corpora have been proposed. While these methods yield improved behavioral adherence, they introduce a significant scalability bottleneck: the reliance on costly manual annotation or extensive hand-authored scripts [75, 70, 94]. This limitation is particularly acute for systems designed to ingest dynamic, multimodal narrative contexts, such as egocentric video streams where persona expression must co-evolve with gaze behavior and environmental cues [53].

Addressing these limitations, our Thinker module leverages a novel pipeline. By utilizing a data-free self-training loop, we achieve the robust fidelity characteristic of fine-tuned models without incurring the prohibitive costs of manual data curation, ensuring consistent character embodiment even in open-ended interactions.

In summary, the literature spans complementary dimensions of interactive intelligence. Our framework integrates these five pillars—Talker, Face Animator, Body Animator, DiT Renderer, and Thinker—into a unified system, aiming to enable NPCs that are expressive, embodied, and consistent across modalities.

9 Conclusion

In this work, we identified a critical gap in the current landscape of digital humans: while visual fidelity has reached photorealistic levels, existing avatars remain fundamentally imitative, lacking the logic and responsiveness required for genuine interaction. To bridge this divide, we introduced **Interactive Intelligence**, a

new paradigm that redefines digital humans as autonomous agents capable of personality-aligned expression, adaptive interaction, and self-evolution.

We realized this paradigm through Mio, an end-to-end embodied intelligence system composed of five specialized modules. By integrating the cognitive reasoning of the Thinker with the real-time embodiment capabilities of the Talker, Face Animator, Body Animator, and Renderer, Mio demonstrates that a digital agent can possess both narrative depth and physical fluidity.

To rigorously measure progress in this new domain, we established the Interactive Intelligence Score (IIS), a comprehensive benchmark aggregating cognitive, acoustic, facial, somatic, and visual performance. On this benchmark, Mio achieved a score of 76.8, demonstrating a +7.8 point improvement over a composite of state-of-the-art baselines. This result quantitatively validates that integrating interactive logic with generative appearance significantly enhances the perceived intelligence and immersion of the agent.

We believe that Interactive Intelligence will become the defining standard for the next generation of avatars, shifting the research focus from static appearance to dynamic, meaningful engagement. By enabling autonomous agents to function as coherent characters within complex narratives, our work paves the way for transformative applications in virtual companionship, interactive storytelling, and immersive gaming. To support this transition and encourage further exploration, we make our full codebase, pre-trained models, and the proposed evaluation benchmark publicly available to the research community.

References

[1] V. Agrawal et al. Seamless interaction: Dyadic audio-visual motion modeling and large-scale dataset. *arXiv:2506.22554*, 2025.

[2] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

[3] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222, 2020.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[5] O. Bar-Tal et al. Lumiere: A space-time diffusion model for video generation. *arXiv:2401.12943*, 2024.

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 2003.

[7] T. Bolkart et al. Flame: A learned model of facial shape and expression. *ACM TOG*, 2023.

[8] Z. Borsos et al. Audiolm: A language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

[9] Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio, 2025. GitHub repository. Release blog available at https://www.boson.ai/blog/higgs-audio-v2.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] Z. Cao et al. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[12] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 43(1):172–186, 2021.

[13] B. Chen et al. Skyreels: Large video diffusion models with long-range temporal consistency. *arXiv:2403.12345*, 2024.

[14] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

[15] S. Chen, J. P. Lalor, Y. Yang, and A. Abbasi. Personatwin: A multi-tier prompt conditioning framework for generating and evaluating personalized digital twins. *arXiv preprint arXiv:2508.10906*, 2025.

[16] W. Chen, J. Cheng, L. Wang, W. Zhao, and W. Matusik. Sensor2text: Enabling natural language interactions for daily activity tracking using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–26, 2024.

[17] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, 2025.

[18] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, 2025.

[19] G. Cheng et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv:2509.14055*, 2025.

[20] C.-H. Chiang et al. Stitch: Simultaneous thinking and talking with chunked reasoning for spoken language models. *arXiv preprint arXiv:2507.15375*, 2025.

[21] X. Chu, N. Goswami, Z. Cui, H. Wang, and T. Harada. Artalk: Speech-driven 3d head animation via autoregressive model. In *SIGGRAPH Asia 2025 Conference Papers*, SA '25. Association for Computing Machinery, 2025.

[22] X. Chu and T. Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[23] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016.

[24] J. S. Chung and A. Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip Reading*, 2016.

[25] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250, 2021.

[26] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[27] R. Danecek, M. J. Black, and T. Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022.

[28] R. Danecek et al. Emoca: Emotion-driven monocular face capture and animation. In *CVPR*, 2022.

[29] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression, 2022.

[30] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. Featured Certification, Reproducibility Certification.

[31] Y. Deng, W. Zhang, W. Lam, S.-K. Ng, and T.-S. Chua. Plug-and-play policy planner for large language model powered dialogue agents. *arXiv preprint arXiv:2311.00262*, 2023.

[32] G. Dong, H. Yuan, K. Lu, C. Li, M. Xue, D. Liu, W. Wang, Z. Yuan, C. Zhou, and J. Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*, 2023.

[33] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

[34] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

[35] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[36] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE, 2024.

[37] G. C. et al. Wan-animate: Unified character animation and replacement with holistic replication. *CoRR*, abs/2509.14055, 2025.

[38] V. A. et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *CoRR*, abs/2506.22554, 2025.

[39] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021.

[40] R. Gao et al. Rad-nerf: Real-time audio-driven neural radiance fields for talking portraits. *arXiv preprint arXiv:2307.00000*, 2023.

[41] Y. Gao et al. Gaussiandiffusion: 4d gaussian splatting diffusion models. *arXiv:2402.12345*, 2024.

[42] L. Giray. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633, 2023.

[43] Y. Gong, L. Jin, R. Deng, D. Zhang, X. Zhang, Q. Cheng, Z. Fei, S. Li, and X. Qiu. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs, 2025.

[44] C. Guo et al. Humanml3d: A large-scale dataset and benchmark for human motion generation from text. In *CVPR*, 2022.

[45] C. Guo, S. Zhang, Y. Wang, W. Hu, Z. Liu, et al. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. Introduces HumanML3D.

[46] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[47] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

[48] K. Hatalis, D. Christou, J. Myers, S. Jones, K. Lambert, A. Amos-Binks, Z. Dannenhauer, and D. Dannenhauer. Memory matters: The need to improve long-term memory in llm-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 277–280, 2023.

[49] Z. He et al. Emo: Emote portrait alive. *arXiv preprint arXiv:2408.07092*, 2024.

[50] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[51] Y. Hong et al. S3d-nerf: Speech-to-3d neural radiance fields for talking head. *arXiv preprint arXiv:2212.00000*, 2022.

[52] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[53] Y. Huang, M. Cai, Z. Li, F. Lu, and Y. Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

[54] Y. Huang, M. Cai, Z. Li, and Y. Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. 2018.

[55] Y. Huang, G. Chen, J. Xu, M. Zhang, L. Yang, B. Pei, H. Zhang, L. Dong, Y. Wang, L. Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. pages 22072–22086, 2024.

[56] Y. Huang, Y. Sugano, and Y. Sato. Improving action segmentation via graph-based temporal reasoning. pages 14024–14034, 2020.

[57] Y. Huang, J. Xu, B. Pei, L. Yang, M. Zhang, Y. He, G. Chen, X. Chen, Y. Wang, Z. Nie, et al. Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–33, 2025.

[58] Y. Huang, L. Yang, and Y. Sato. Compound prototype matching for few-shot action recognition. 2022.

[59] Y. Huang, L. Yang, and Y. Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. 2023.

[60] K. Inoshita and R. Harada. Persona-based synthetic data generation using multi-stage conditioning with large language models for emotion recognition. *arXiv preprint arXiv:2507.13380*, 2025.

[61] C. Ionescu et al. Human3.6m: Large scale datasets and predictive methods for 3d human sensing. In *ICCV*, 2013.

[62] Y. Jiang et al. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2304.00000*, 2023.

[63] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu. Vace: All-in-one video creation and editing. In *ICCV*, pages 17191–17202, 2025.

[64] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

[65] B. Kerbl et al. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023.

[66] B. Kim, J. Park, J. Yang, and H. Lee. Chronological passage assembling in rag framework for temporal question answering. *arXiv preprint arXiv:2508.18748*, 2025.

[67] J. Kim et al. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*, 2021.

[68] J. W. Kim, J. Salamon, P. Li, and J. P. Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018.

[69] J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[70] N. Kovačević, C. Holz, M. Gross, and R. Wampfler. The personality dimensions gpt-3 expresses during human-chatbot interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–36, 2024.

[71] H. Le et al. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*, 2023.

[72] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han. Autoregressive image generation using residual quantization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11513–11522, 2022.

[73] H. Lee, S. Phatale, H. Mansoor, K. R. Lu, T. Mesnard, J. Ferret, C. Bishop, E. Hall, V. Carbune, and A. Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.

[74] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[75] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[76] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[77] T. Li et al. Learning a model of facial shape and expression from 4d scans. In *SIGGRAPH Asia*, 2017.

[78] T. Li et al. Diffusionavatars: Diffusion-based high-fidelity 3d talking face generation. *arXiv preprint arXiv:2303.00000*, 2023.

[79] L. Liao, G. H. Yang, and C. Shah. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 3452–3455, 2023.

[80] Y. Lipman et al. Flow matching for generative modeling. *arXiv:2210.02747*, 2022.

[81] J. Liu, J. Lin, and Y. Liu. How much can rag help the reasoning of llm? *arXiv preprint arXiv:2410.02338*, 2024.

[82] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

[83] X. Liu et al. Diffhuman4d: Diffusion models for temporally-consistent 4d human synthesis. *arXiv:2405.01829*, 2024.

[84] M. Loper et al. Smpl: A skinned multi-person linear model. In *SIGGRAPH Asia*, 2015.

[85] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015.

[86] Y. Luo, Z. Rong, L. Wang, L. Zhang, T. Hu, and Y. Zhu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. 2026.

[87] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[88] S. Nepal, A. Pillai, W. Campbell, T. Massachi, M. V. Heinz, A. Kunwar, E. S. Choi, X. Xu, J. Kuc, J. F. Huckins, et al. Mindscape study: integrating llm and behavioral sensing for personalized ai-driven journaling experiences. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(4):1–44, 2024.

[89] D. Ng, K. Zhou, Y.-W. Chao, Z. Xiong, B. Ma, and E. Chng. Multi-band frequency reconstruction for neural psychoacoustic coding. In *Forty-second International Conference on Machine Learning*, 2025.

[90] OpenAI. Gpt-4 technical report, 2024.

[91] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[92] F. Paraperas Papantoniou and S. Zafeiriou. Id-consistent, precise expression generation with blendshape-guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025.

[93] J. S. Park et al. Generative agents: Interactive simulacra of human behavior. In *ACM UIST*, 2023.

[94] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[95] W. Peebles and J.-Y. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[96] L. Peng and J. Shang. Quantifying and optimizing global faithfulness in persona-driven role-playing. *Advances in Neural Information Processing Systems*, 37:27556–27583, 2024.

[97] S. Peng et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021.

[98] Z. Peng, Y. Fan, H. Wu, X. Wang, H. Liu, J. He, and Z. Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025.

[99] K. R. Prajwal, R. Mukhopadhyay, et al. Wav2lip: Accurately lip-syncing videos in the wild. In *ACM Multimedia*, 2020.

[100] A. Punnakkal, A. Chandrasekaran, N. Athanasiou, M. J. Black, and A. Yao. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021.

[101] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[102] C. K. Reddy, V. Gopal, and R. Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE, 2021.

[103] C. K. A. Reddy, V. Gopal, and R. Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, 2021.

[104] M. A. Research. Seamlessm4t v2: Multilingual and multimodal speech translation. *arXiv preprint arXiv:2405.00000*, 2024.

[105] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021.

[106] V. Rudnev et al. Gs-avatar: Real-time head avatar with 3d gaussian splatting. *arXiv:2403.12382*, 2024.

[107] A. Shysheya et al. Tava: Template-free animatable volumetric avatars. In *CVPR*, 2023.

[108] A. Siarohin et al. First order motion model for image animation. In *NeurIPS*, 2019.

[109] A. Singh, S. Hsu, K. Hsu, E. Mitchell, S. Ermon, T. Hashimoto, A. Sharma, and C. Finn. Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users. *arXiv preprint arXiv:2502.19312*, 2025.

[110] H. Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[111] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020.

[112] R. Sonobe, S. Takamichi, and H. Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[113] G. Sun, X. Zhan, and J. Such. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–6, 2024.

[114] Z. Sun, T. Lv, S. Ye, M. Lin, J. Sheng, Y.-H. Wen, M. Yu, and Y.-J. Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4), 2024.

[115] S. Tan et al. Aniportrait: Animation and editing of portrait videos using diffusion models. *arXiv:2402.01234*, 2024.

[116] M.-O. Team. Mini-omni: Lightweight end-to-end speech-to-speech model. *arXiv preprint arXiv:2408.12345*, 2024.

[117] G. Tevet et al. Motion diffusion model. In *NeurIPS*, 2022.

[118] G. Tevet et al. Human motion diffusion model. In *ICCV*, 2023.

[119] H. Thakur, E. Agrawal, and S. Mukund. Personas within parameters: Fine-tuning small language models with low-rank adapters to mimic user behaviors. *arXiv preprint arXiv:2509.09689*, 2025.

[120] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

[121] Y.-M. Tseng, Y.-C. Huang, T.-Y. Hsiao, W.-L. Chen, C.-W. Huang, Y. Meng, and Y.-N. Chen. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*, 2024.

[122] C. Wang, S. Chen, et al. Vall-e: Neural codec language models are zero-shot text-to-speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[123] C. Wang et al. Vall-e 2: Neural codec language models with repetition-aware sampling. *arXiv preprint arXiv:2406.12345*, 2024.

[124] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang. Unispeech: Unified speech representation learning with labeled and unlabeled data, 2021.

[125] L. Wang et al. Animate-a-video: Animate anyone. *arXiv:2311.17117*, 2023.

[126] L. Wang, J. Lian, Y. Huang, Y. Dai, H. Li, X. Chen, X. Xie, and J.-R. Wen. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, 2025.

[127] L. Wang, J. Lian, Y. Huang, Y. Dai, H. Li, X. Chen, X. Xie, and J.-R. Wen. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, 2025.

[128] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021.

[129] X. Wang, H. Zhang, T. Ge, W. Yu, D. Yu, and D. Yu. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*, 2025.

[130] Z. M. Wang, Z. Peng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, J. Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023.

[131] Z. Weng et al. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022.

[132] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[133] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[134] L. Xiao, S. Lu, H. Pi, K. Fan, L. Pan, Y. Zhou, Z. Feng, X. Zhou, S. Peng, and J. Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. *arXiv preprint arXiv:2503.15451*, 2025.

[135] Z. Xiao et al. Videocrafter: Open diffusion models for high-quality video generation. *arXiv:2310.19512*, 2023.

[136] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022.

[137] Z. Xie et al. Mini-omni-reasoner: Token-level thinking-in-speaking in large speech models. *arXiv preprint arXiv:2508.15827*, 2025.

[138] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[139] Z. Xiong et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *CVPR*, 2024.

[140] Z. Xiong, C. Li, K. Liu, H. Liao, J. Hu, J. Zhu, S. Ning, L. Qiu, C. Wang, S. Wang, S. Cui, and X. Han. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19801–19811, June 2024.

[141] J. Xu et al. Qwen2.5-omni: A thinking-talker architecture for unified multimodal understanding and speech interaction. *arXiv preprint arXiv:2503.20215*, 2025.

[142] P. Xu, A. Madotto, C.-S. Wu, J. H. Park, and P. Fung. Emo2vec: Learning generalized emotion representation by multi-task training. *arXiv preprint arXiv:1809.04505*, 2018.

[143] S. Yang et al. Hunyuanvideo: A generative video foundation model. *arXiv preprint arXiv:2407.00000*, 2024.

[144] Z. Yang, X. Xu, B. Yao, E. Rogers, S. Zhang, S. Intille, N. Shara, G. G. Gao, and D. Wang. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–35, 2024.

[145] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[146] Z. Ye, P. Sun, J. Lei, H. Lin, X. Tan, Z. Dai, Q. Kong, J. Chen, J. Pan, Q. Liu, et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705, 2025.

[147] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. Dai, et al. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025.

[148] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023.

[149] Y. Yuan et al. Rflow: Recurrent flow matching for video and motion generation. *arXiv:2211.00931*, 2022.

[150] Y. Yuan et al. Mmm: Masked motion model for efficient human motion generation. *arXiv preprint arXiv:2311.00000*, 2023.

[151] Y. Yuan et al. Momask: Generative masked modeling of 3d human motions. *arXiv preprint arXiv:2308.00000*, 2023.

[152] Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

[153] H. Zhang et al. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. *arXiv preprint arXiv:2304.01154*, 2023.

[154] L. Zhang et al. Liveportrait: Efficient and controllable talking face animation. *arXiv preprint arXiv:2406.00000*, 2024.

[155] Q. Zhang et al. T2m-gpt: Generating human motion from textual prompts with gpt. *arXiv preprint arXiv:2306.00000*, 2023.

[156] S.-H. Zhang et al. Motiondiffuse: Text-driven human motion generation with diffusion models. In *SIGGRAPH Asia*, 2022.

[157] Y. Zhang et al. Primal: Physically reactive and interactive motor model for avatar learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[158] 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.

[159] 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. In *ICML*, 2025.

[160] X. Zhao, H. Xu, G. Song, Y. Xie, C. Zhang, X. Li, L. Luo, J. Suo, and Y. Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*, 2025.

[161] S. Zheng, L. Cheng, Y. Chen, H. Wang, and Q. Chen. 3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement. *arXiv preprint arXiv:2306.15354*, 2023.

[162] Z. Zheng et al. Geneface: Generalized and high-fidelity audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*, 2023.

[163] Z. Zheng et al. Geneface++: Enhanced stability and fidelity in audio-driven talking face generation. *arXiv preprint arXiv:2401.00000*, 2024.

[164] S. Zhu, J. L. Chen, Z. Dai, Z. Dong, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024.

[165] W. Zielonka, T. Bolkart, and J. Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pages 20311–20322, 2022.

[166] L. Ziyin, T. Hartwig, and M. Ueda. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594, 2020.