InteractiveAvatar: Real-Time Streaming Video Generation for Consistent and Intent-Aware Avatars

Quanyue Song[1,2*†], Yishan He[2†], Yanfei Zhang[2], Shihao Cheng[2,3*], Zhixiang He[2], Zhizhi Guo[2✉], Chi Zhang[4], Xuelong Li[4], and Caigui Jiang[1✉]

[1] State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[2] China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd., China
[3] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China
[4] Institute of Artificial Intelligence (TeleAI), China Telecom, China

Abstract. Recent diffusion-based models have enabled realistic audio-driven avatar generation in real-time streaming. However, existing approaches struggle to maintain visual temporal consistency and fail to explicitly perceive user intent in complex interactive streaming scenarios. To address these challenges, we propose InteractiveAvatar, a real-time infinite-streaming video generation framework that supports visually consistent avatar video generation and intent-aware interactions. With autoregressive distillation, InteractiveAvatar achieves real-time streaming generation of human avatars over arbitrarily long durations. For visual consistency, we introduce a Long-Short Visual Memory (LSVM) mechanism that flexibly compresses historical visual information into compact tokens, preserving both short-range coherence and long-term consistency. To generate avatars with speeches and actions aligned with user intent, we propose a Reasoning-Reaction Module (RRM), which incorporates a State-Cycling strategy and a Cache-Switching mechanism. Extensive experimental results over diverse scenarios demonstrate that our method achieves state-of-the-art visual consistency in long-duration generation, while enabling complex user-avatar interaction in real time.

Keywords: Real-Time Video Generation · Audio-Driven Avatar · Diffusion Model

1 Introduction

With the development of diffusion-based models for audio-driven avatar generation, it has become possible to synthesize videos with accurate lip synchroniza-

* Work done during an internship at China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd.
† Equal contribution.
✉ Corresponding author.

tion, highly realistic and visually appealing appearances [1, 4, 16, 34]. Furthermore, recent advances [20, 26, 42] show promising results in real-time streaming diffusion generation of human avatars. However, existing approaches [14, 26, 33, 38, 46] show limitations in user-avatar interaction and visual consistency in streaming generation. These methods are typically restricted to simple interactive scenarios, such as speaking or listening synchronously with the audio, which is limited in understanding the user intent and maintaining consistency in more complicated scenarios. These limitations highlight two fundamental obstacles that continue to hinder the development of truly realistic and interactive human avatars.

The first challenge lies in maintaining temporal consistency during long-duration video generation. In interactive scenarios with sustained user engagement, visual content evolves continuously, resulting in substantial scene variations and increased complexity. Most existing real-time streaming methods [14, 26] employ causal attention, limiting the model's receptive field to a few previously generated chunks along with reference image. When the generated content gradually diverges from this initial reference, inconsistencies tend to emerge between adjacent chunks, causing significant temporal drift and the degradation of long-range video consistency.

The second challenge concerns the user-avatar interaction. Beyond generating verbal responses, a realistic avatar is expected to perform actions that are semantically aligned with the user's intent. For instance, when a user asks "Can you check what time it is on the watch?", the avatar should glance at a watch while verbally providing the time. Existing approaches [33, 38], however, predominantly follow an audio-driven paradigm, using response speech generated by large language models to animate lip movements and synthesizing limited actions or gestures based on coarse audio-motion correlations learned from training data. This design overlooks explicit modeling of user intent, making it difficult to generate fine-grained and intent-aware actions such as checking time using a watch, thereby limiting the realism of interactive behavior.

To address these challenges, we propose InteractiveAvatar (see Fig. 1). InteractiveAvatar introduces a real-time interactive generation paradigm, in which the avatar can respond to users by not only producing speech, but also performing intent-aligned actions and adaptively updating the surrounding scene, while maintaining strong visual consistency over long-duration video generation. Specifically, our framework incorporates a Long-Short Visual Memory (LSVM) mechanism and a Reasoning-Reaction Module (RRM) into the real-time streaming diffusion generation.

The LSVM mechanism maintains temporal consistency in streaming generation by jointly modeling short-term and long-term visual information. The short-term memory preserves recently generated frames to ensure coherent transitions and local continuity, while the long-term memory retains representative visual states that capture critical information throughout the entire generation process. We propose a Dynamic Key-Frame Selection strategy, which flexibly transfers visually important content from the short-term memory to the long-

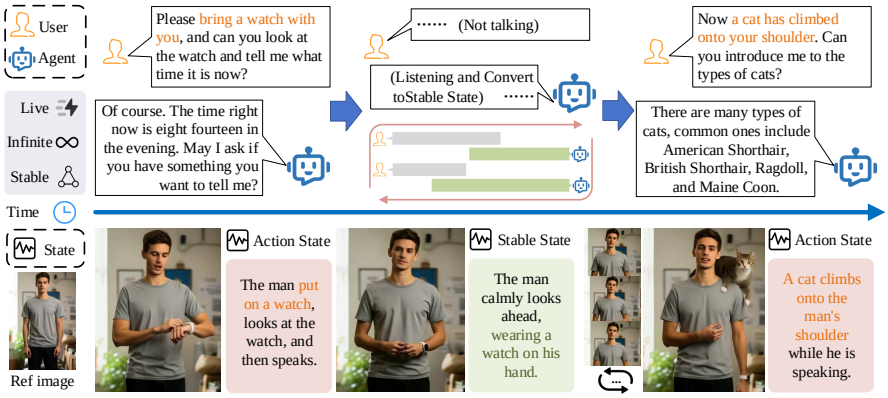

Fig. 1: We propose InteractiveAvatar, a real-time streaming audio-driven avatar generation framework that enables intent-aware interaction. InteractiveAvatar interprets user intent to generate contextually relevant actions throughout the dialogue while maintaining long-range visual consistency.

term memory during inference, enabling the model to preserve essential global information and prevent temporal drift over long-duration generation.

The RRM enhances the realism of user-avatar interaction by leveraging a large language model for intent understanding and state management. Given user input, it infers the underlying intent and generates corresponding verbal responses and action instructions to guide avatar synthesis. To ensure a natural interaction flow, we design a State-Cycling strategy that enables smooth transitions between listening and execution states. Additionally, memory augmentation is incorporated to maintain consistency between avatar behavior and scene evolution. We further introduce a Cache-Switching mechanism to accelerate action execution and scene transitions, thereby reducing perceived latency during interaction.

By integrating these two components with streaming generation approach, InteractiveAvatar enables real-time synthesis of highly consistent and interactive human avatars. We evaluate InteractiveAvatar across multiple user-instruction scenarios to assess its responsiveness and generation quality. Both qualitative and quantitative results demonstrate that InteractiveAvatar achieves real-time avatar generation while maintaining strong visual consistency over long-duration video sequences, as well as supporting realistic and complex user-avatar interactions.

Contributions of our InteractiveAvatar are listed below:

- We propose InteractiveAvatar, a novel real-time streaming audio-driven avatar generation framework that supports long-duration video synthesis with strong visual consistency and intent-aware user-avatar interaction.
- We introduce a Long-Short Visual Memory mechanism that flexibly preserves historical visual representations, enabling the model to maintain at-

tention to past content and significantly improving visual coherence in long-duration generation.
- We design a Reasoning-Reaction Module equipped with a State-Cycling strategy and a Cache-Switching mechanism, enabling intent-aware interaction and efficient command execution with reduced latency.

2 Related Work

2.1 Audio-Driven Avatar Generation

Audio-driven avatar generation aims to synthesize realistic human videos conditioned on speech while preserving identity and motion consistency. Early methods, such as Wav2Lip [23], focus primarily on accurate lip synchronization but fail to model head and body movements. Later two-staged animators [40, 46] improve controllability on expressions and head movements by first predicting latent motion representations from audio and further rendering avatar expressions and head motions using these latent representations, yet these methods are still constrained by predefined motion priors and lack of natural body movements. More recently, diffusion-based frameworks, particularly DiT-style architectures [2,3,7,10,16,21,27,34], have significantly improved audio-driven avatar generation in visual fidelity and naturalness. However, their iterative denoising process incurs substantial computational overhead, hindering real-time streaming and long-term consistency in interactive scenarios. Moreover, most methods rely solely on audio signals, limiting their ability to model high-level user intent.

2.2 Streaming Video Generation

Typical bidirectional attention-based video diffusion generators [2,31] suffer from low computational efficiency, limited video duration, and poor temporal stability. To address these problems, recent works [14, 26, 38] distill bidirectional diffusion transformers into causal autoregressive architectures for streaming inference. For example, CausVid [42] introduces block-causal attention with distribution matching distillation. And Self-Forcing [13] further mitigate train-test mismatches by conditioning on previously generated frames during training and further improve long-horizon stability. Other approaches [8, 41] explore rolling-window denoising, attention sinks, or KV-recache strategies to reduce temporal drift and error accumulation. Despite these advances, existing methods mainly rely on a limited number of previously generated chunks and focus mainly on local temporal context. This limitation prevents effective modeling of long-range information across the generation horizon, resulting in degraded long-range consistency and temporal drift.

2.3 Interactive Avatar Generation

Interactive avatar generation aims to create avatars capable of engaging in natural and responsive user interactions. Early methods [11,23,49] rely on generating

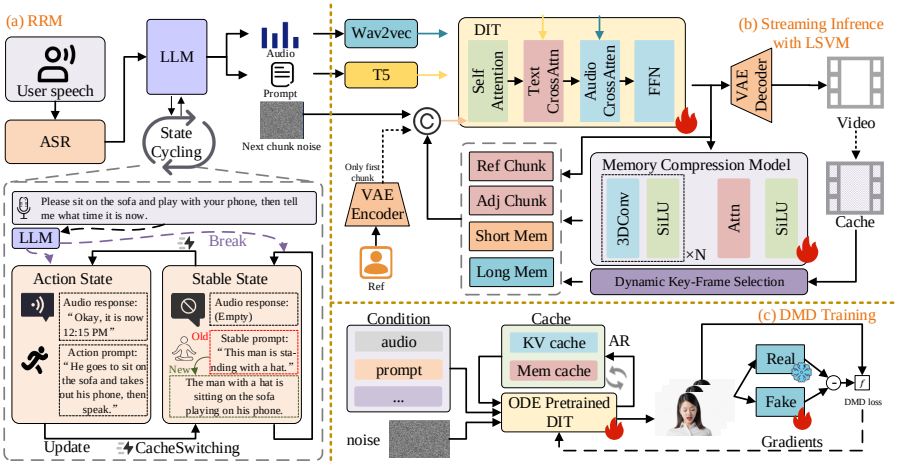

Fig. 2: Overview of InteractiveAvatar, which consists of (a) The Reasoning-Reaction Module (RRM) performs intent-aware interaction with user; (b) Streaming Inference with Long-Short Visual Memory (LSVM) mechanism to enhance the visual consistency; and (c) DMD training for real-time streaming generation.

actions conditioned from audio cues, where a speech model drives lip movements and simple gestures. The generated results are largely depend on acoustic cues, with synchronized but simple gestures. Recent works [9, 20] on interactive avatars have explored audio-driven streaming diffusion to enable responsive user-avatar interactions. However, these methods lack multi-turn conversational memory or action-level reasoning. More recent studies [22, 33, 35] have incorporated limited "listening states" or predefined actions to simulate interaction, yet they cannot maintain long-term context or produce intent-aware behaviors. In contrast, our approach integrates a human-imitating reasoning module with memory-enhanced diffusion-based video synthesis, enables real-time, context-aware, and temporally consistent interactive avatar behavior, bridging the gap between simple reactive systems and fully intent-aware digital humans.

3 Method

As illustrated in Figure 2, given an avatar reference image and user speech input, InteractiveAvatar aims to generate a real-time video that animates the avatar to naturally respond to the user. The speech signal is first transcribed into text via an automatic speech recognition (ASR) system. The transcribed text is then processed by the Reasoning-Reaction Module (RRM) to produce intent-aware verbal and action instructions (Sec. 3.3). Conditioned on these outputs, a streaming video generation model with Long-Short Visual Memory (LSVM) mechanism synthesizes temporally coherent video frames in real time (Sec. 3.2). Finally, the model is trained using self-forcing for autoregressive distillation to

enable stable, efficient generation (Sec. 3.4), producing intent-aware, temporally consistent avatar videos.

3.1 Preliminaries

Diffusion Transformer We build our method upon a Diffusion Transformer (DiT). A pre-trained VAE encodes input data x into latent representations $z = E(x)$. During forward diffusion, Gaussian noise is progressively added to obtain $z_t = \sqrt{\alpha_t} z + \sqrt{1-\alpha_t}\epsilon$. The DiT model $\epsilon_\theta(z_t, t, c)$ predicts the injected noise at timestep t conditioned on signal c. Training minimizes the mean squared error between predicted and true noise:

$$\mathcal{L} = \mathbb{E}_{t, z_t, c, \epsilon}\left[\|\epsilon_\theta(z_t, t, c) - \epsilon\|_2^2\right] \tag{1}$$

We adopt Wan2.2 5B [31] as the base model, which employs a causal 3D VAE for spatiotemporal representation learning [18, 19, 44]. Text conditioning is obtained via T5 [24] embeddings and injected through cross-attention, while timestep information is incorporated via learned modulation parameters.

Distribution Matching Distillation Distribution Matching Distillation (DMD) distills a pre-trained teacher diffusion model into a few-step student generator by aligning their intermediate noisy distributions [42]. Let $G_\theta(z)$ be the student generator with $\hat{x} = G_\theta(z)$ and $z \sim \mathcal{N}(0, I)$. Denote by $p_{\theta,t}(x_t)$ and $p_{\text{data},t}(x_t)$ the student-induced and teacher-induced distributions at time step t, respectively. DMD minimizes the reverse KL divergence:

$$\mathcal{L}_{\text{DMD}} = \mathbb{E}_t\left[D_{\text{KL}}\big(p_{\theta,t} \,\|\, p_{\text{data},t}\big)\right]. \tag{2}$$

Its gradient can be written as:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t,z}\left[\left(s_{\text{real}}(x_t, t) - s_{\text{fake},\phi}(x_t, t)\right)^\top \frac{\partial G_\theta(z)}{\partial \theta}\right] \tag{3}$$

where $x_t = \Psi(\hat{x}, t)$, and s_{real} and $s_{\text{fake},\phi}$ denote the teacher and student score functions. Training alternates between updating $s_{\text{fake},\phi}$ and optimizing G_θ. For multi-step distillation, a student rollout is first constructed, and a random intermediate state is used to stabilize training. In our approach, we adopt DMD to distill the diffusion model into a few-step generator, enabling real-time video generation.

3.2 Long-Short Visual Memory

Preserving key information from previous frames is essential for temporal coherence and visual consistency in long-duration video generation. However, storing all historical frames is computationally expensive and impractical for streaming generation. Inspired by [45] on memory compression and efficient context

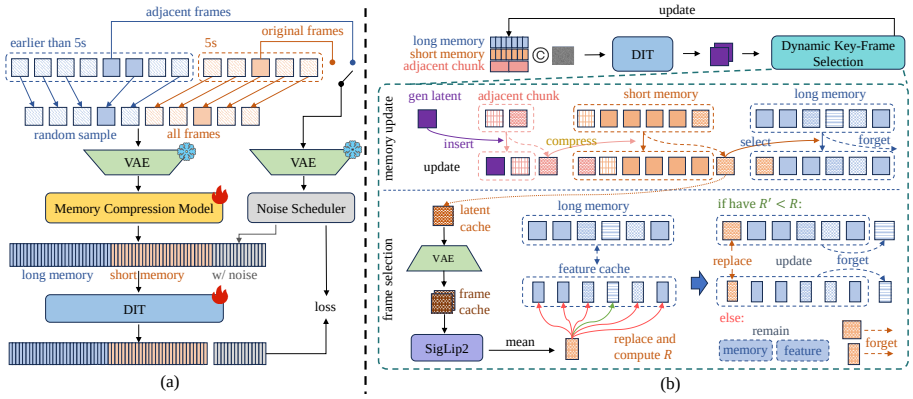

Fig. 3: LSVM Mechanism.(a) During training, long-term memory frames are randomly sampled, while short-term memory retains all recent frames. (b) During inference, Dynamic Key-Frame Selection adaptively updates memory to retain critical visual information.

modeling, we propose Long-Short Visual Memory (LSVM) mechanism (Fig. 3). To effectively capture both global appearance characteristics and recent visual dynamics, we decompose the memory into long-term memory and short-term memory. The short-term memory maintains dense representations of recently generated frames to ensure local temporal coherence, while the long-term memory stores compact representations of globally salient visual states to stabilize overall appearance consistency. A Key-Frame Selection strategy is introduced to determine when short-term memory entries should be promoted to long-term memory, as well as when outdated long-term memory elements should be discarded, enabling adaptive memory updating during streaming generation.

Long-Short Memory Compression We apply a lightweight compression model $\mathcal{C}(\cdot)$ to obtain compact memory tokens. The architecture of $\mathcal{C}(\cdot)$ is similar to [45], consisting of a sequence of convolutional layers followed by attention modules. Given the generated video latent frames z_t, the compact memory tokens can be represented as:

$$m_t = \mathcal{C}(z_t) \tag{4}$$

where the temporal compression ratio is set to 1, ensuring no temporal downsampling during streaming generation. This design avoids temporal merging across latents, which simplifies latent-wise separation and update during streaming generation.

We maintain two memory buffers: a short-term memory $\mathcal{M}_s$ and a long-term memory $\mathcal{M}_l$. The short-term memory stores recent compressed tokens within a fixed window of size K to preserving local continuity:

$$\mathcal{M}_s = \{m_{t-K+1}, \ldots, m_t\} \tag{5}$$

The long-term memory maintains a compact set of globally representative tokens with fixed capacity N:

$$\mathcal{M}_l = \{\tilde{m}_1, \ldots, \tilde{m}_N\} \tag{6}$$

We first concatenate the compressed long-term and short-term memory tokens into a unified history representation $\Phi(H) = \text{Concat}(\mathcal{M}_l, \mathcal{M}_s)$, where $\mathcal{M}_s$ contains the recent K short-term tokens and $\mathcal{M}_l$ contains the N long-term representative tokens. This design enables the model to jointly capture local temporal continuity and global long-range dependencies.

During training, we randomly sample a video segment $H = \{z_1, \ldots, z_T\}$. The last K consecutive frames (we set the length to 5s) of the segment are treated as the short-term memory source Ω_s, while from the preceding history $i < t-K+1$, we randomly sample a fixed number of frame indices Ω_l to construct the long-term memory:

$$\mathcal{M}_l = \{m_i \mid i \in \Omega_l, ; i < t - K + 1\} \tag{7}$$

To train the reconstruction capability, we sample a set of frame indices Ω from the full segment H as diffusion targets. Depending on the temporal location of each sampled frame, we apply different strategies. For sampled indices within the short-term memory, we randomly keep the subset Ω unchanged and mask all remaining frames. For sampled indices within the long-term range, we preserve the frame that is temporally closest to Ω as an anchor frame and mask all remaining frames, as it preserves similar semantic content while providing a slightly different observation for denoising. This design encourages the long-term memory to capture not only pixel-level content but also consistent scene elements.

All masked frames are corrupted using a noise-as-mask strategy. After masking, we clone the clean frames $\{z_i \mid i \in \Omega\}$ as the diffusion targets. The diffusion model is then trained to reconstruct these target frames at arbitrary temporal positions conditioned on the compressed memory representation $\Phi(H)$. This objective can be written as:

$$\mathbb{E}_{H,\Omega,c,\epsilon,t_i} ||(\epsilon - H_\Omega) - G_\theta\big((H_\Omega)_{t_i}, t_i, c, \Phi(H)\big)||_2^2 \tag{8}$$

where H_Ω denotes the selected clean target frames, $\epsilon_i \sim \mathcal{N}(0, I)$ correspond to sampled noise levels and $\Phi(H)$ is the concatenated long-short memory representation.

By randomly sampling reconstruction targets from both short-term and long-term regions, the model is compelled to encode the entire history in a balanced manner, preserving fine-grained details and global visual consistency essential for stable streaming autoregressive generation.

Dynamic Key-Frame Selection To maintain a compact yet semantically representative long-term memory, we introduce a Dynamic Key-Frame Selection (DKFS) strategy during inference.

The short-term memory $\mathcal{M}_s$ is implemented as a fixed-length first-in-first-out (FIFO) queue of size K. At initialization, all entries are filled with the compressed representation of the first frame $\mathcal{M}_s^{(0)} = \{m_1, \ldots, m_1\}$. During streaming generation, each newly generated latent frame z_t is compressed into $m_t = \mathcal{C}(z_t)$ and appended to $\mathcal{M}_s$, while the oldest token is removed $\mathcal{M}_s \leftarrow \mathrm{Push}(\mathcal{M}_s, m_t)$. Whenever a token m_{out} is popped from $\mathcal{M}_s$, the corresponding frame becomes a candidate for long-term memory update.

For each popped frame, we retrieve its corresponding real image frames and feed them into SigLIP2 [28] to obtain semantic feature vectors. The features are averaged to produce a single semantic descriptor:

$$\mathbf{s}_{\text{cand}} = \frac{1}{n} \sum_{j=1}^{n} f_{\text{SigLIP2}}(I_j) \tag{9}$$

where I_j^n are the associated real frames and $f_{\text{SigLIP2}}(\cdot)$ denotes the feature extractor.

The long-term memory $\mathcal{M}_l$ is a fixed-capacity buffer of size N, initialized by repeating the first frame $\mathcal{M}_l^{(0)} = \{\tilde{m}_1, \ldots, \tilde{m}_1\}$. Frames selected for long-term memory are inserted in chronological order, ensuring temporal interpretability. To decide whether a candidate frame should be retained in $\mathcal{M}_l$, we evaluate its contribution to global semantic diversity. Let $\{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ denote the current semantic feature set of the long-term memory. For each memory entry $\mathbf{s}_i$, we compute its mean cosine similarity to all other entries and then compute the global redundancy score:

$$\bar{\rho}_i = \frac{1}{N-1} \sum_{j \neq i} \cos\big(\mathbf{s}_i, \mathbf{s}_j\big), \quad R = \frac{1}{N} \sum_{i=1}^{N} \bar{\rho}_i \tag{10}$$

If replacing a slot in $\mathcal{M}_l$ with the candidate feature $\mathbf{s}_{\text{cand}}$ leads to a lower redundancy score $R' < R$, the candidate frame is retained and inserted into long-term memory. Otherwise, it is forgot.

This strategy encourages the long-term memory to maintain semantically diverse and globally representative key frames. Instead of storing frames uniformly over time, the memory adaptively preserves frames that introduce novel semantic content, thereby maximizing coverage of distinct scenes, identities, and objects across long video streams.

3.3 Reasoning-Reaction Module

To enable interactive and controllable avatar behaviors under streaming generation, we introduce a Reasoning-Reaction Module (RRM), which leverages a Large Language Model (LLM) [39] for intent understanding and state management. The RRM serves as an information reasoning and response center that bridges multimodal user input and diffusion-based visual generation. It consists of two core components: State-Cycling strategy and a Cache-Switching mechanism.

State-Cycling Strategy Under the State-Cycling strategy, when a user provides speech input, the audio is first transcribed into text and then fed into the LLM together with the current stable state. The LLM outputs two states: an action state, which includes the action prompt p^{act} and the response audio a^{resp}, and a stable state, which contains the stable-state prompt p^{stable}. The p^{act} describes the motion to be executed (e.g., standing up or sitting down), the a^{resp} contains the spoken reply, and the p^{stable} represents the static visual condition after the action is completed (e.g., the person is sitting calmly). Formally, given user text x_t and previous stable-state prompt p_{t-1}^{stable}, the LLM produces

$$(p_t^{\text{act}}, a_t^{\text{resp}}, p_t^{\text{stable}}) = \text{LLM}(x_t, \mathcal{S}_{t-1}) \tag{11}$$

During generation, the diffusion model is conditioned on $(p_t^{\text{act}}, a_t^{\text{resp}})$ while the response audio is active, where the audio drives lip synchronization and the action prompt guides body motion and pose transitions. Once the response audio finishes, the audio condition is set to empty and the conditioning prompt is replaced by p_t^{stable}. The system then continues generation using only the stable-state prompt, ensuring that the character remains visually consistent without unnecessary motion drift. The model stays in this stable state until a new user instruction arrives and the LLM produces the next action, forming a cyclic process of reasoning, reaction, and stabilization.

Cache-Switching Mechanism To reduce latency under streaming generation, we introduce a Cache-Switching mechanism for prompt-conditioned key-value (KV) cache. Due to the autoregressive setup, previously adjacent generated chunks are computed based on earlier prompts. When the action prompt switches to a new one p_t^{act}, these cached KV become inconsistent with the updated instruction.

To accelerate adaptation to a new prompt, when the action prompt switches, we re-encode the updated text condition and recompute the prompt-conditioned KV tensors of the previously adjacent chunks using the new prompt. These updated KV cache are then used for subsequent attention. Let $\mathcal{K}_{\text{old}}$ denote the cached latent frame KV tensors under the previous prompt p_{t-1}^{act} and $\mathcal{K}_{\text{new}}$ those recomputed using p_t^{act}. The update is written as

$$\mathcal{K} \leftarrow \text{Replace}(\mathcal{K}_{\text{old}}, \mathcal{K}_{\text{new}}), \tag{12}$$

where only the affected chunks are refreshed. This Cache-Switching mechanism enables efficient realignment with updated action prompts, thereby reducing latency during streaming generation.

3.4 Autoregressive Distillation Adaption

We adapt a pretrained diffusion backbone into a stable and efficient real-time streaming generation model through a four-stage training pipeline. First, we train an image-audio-to-video (AI2V) model based on bidirectional attention on

the large-scale audio-visual data, establishing strong speech-driven video generation capability. Second, we pretrain the proposed memory compression module $\mathcal{C}(\cdot)$ on the video reconstruction task so that it can efficiently and adaptively compress visual tokens. Third, we perform ODE initialization by training the model with block-wise causal attention to approximate the bidirectional teacher's ODE trajectories. The causal attention mask is set as in [41] during training for both acceleration and high fidelity, along with the optimization of the long short memory compression module. Finally, we apply Self-Forcing DMD training to distill the diffusion model, together with the LSVM, into a few-step autoregressive generator by minimizing the reverse KL divergence between their induced noisy distribution, enabling stable rollout with substantially reduced sampling steps for real-time inference.

4 Experiments

4.1 Experimental Setup

Our framework is built upon the Wan 2.2 5B [31] model as the backbone. All training and inference are conducted at the resolution of 576p(e.g. 1024x576 in the aspect ratio of 16:9), generating 3 latent per chunk in the streaming setting. Training is performed on 64 NVIDIA H100 GPUs, with 50K steps in Stage 1, 30K steps in Stage 2, and 20K steps in Stage 3 and Stage 4. We employ Fully Sharded Data Parallel (FSDP) [48] with hybrid sharding to reduce memory consumption while enhancing efficiency. The learning rate is set to 1e-5 for the student branch and 2e-6 for the fake score branch. For inference, to enable real-time streaming interation, we first implement kv caching to avoid recomputing key-values of the generated chunks. We also use the pipeline parallelism [20] for further acceleration by deploying the DiT and VAE model on different GPU devices. With these optimizations, we provide an end-to-end latency breakdown to better understand the system bottleneck. Specifically, the DiT module takes approximately 450 ms, the lightweight VAE decoding requires about 50 ms, the LLM audio response accounts for around 1600 ms, and other overhead contributes roughly 450 ms. Taken together, the overall Time-to-First-Frame(TTFF), is approximately 2.6s.

4.2 Dataset

We curate a large-scale audio-visual dataset consisting of approximately 3 million high-quality clips after filtering [15]. The dataset is composed of three complementary subsets. The first subset focuses on speech-driven talking-head videos, including HDTF [47], VFHQ [37], VoxCeleb2 [5], CelebV-Text [43], and AVSpeech [47], which provide high-resolution facial videos with diverse identities and rich speech content for accurate lip synchronization and fine-grained facial motion modeling. The second subset consists of movie and TV-show data primarily collected from OpenHumanVid [17], capturing complex scenes, expressive performances, and temporal dynamics to improve realism and robustness. The

third subset consists of our proprietary conversational dataset, featuring long-duration speaking segments accompanied by rich and diverse body movements. This subset emphasizes sustained speech, expressive gestures, and natural head-shoulder dynamics, providing strong temporal continuity and interaction cues for streaming generation.

4.3 Metrics

We evaluate our model from two perspectives: video quality and consistency. For video quality, we assess final video's perceptual quality using Q-align (IQA) [36] and aesthetic appeal (ASE). Distribution-level fidelity is measured by FID [12] for frame-wise realism and FVD [30] for overall spatio-temporal coherence. For video consistency, we measure audio-visual synchronization using SynC and SynD [6], capturing the correspondence between lip movements and input audio. Object-level temporal consistency (OBJ) is evaluated with Gemini, while identity preservation across frames (ID) is measured using DINOv3 [25], as avatar videos involve full-body appearance and clothing beyond facial identity. Finally, we use VideoCLIP-XLv2 [32] to assesses the alignment between the generated video and the input text prompt (TV), reflecting the effectiveness of semantic control. In addition, we compare the generation speed (FPS) of different models to evaluate real-time performance and streaming efficiency.

4.4 Results

We compare InteractiveAvatar against current state-of-the-art open-sourced audio-driven avatar generation approaches, including StableAvatar [29], OmniAvatar [10], HYAvatar [2], Hallo3 [7], EchoMimicV3 [21], WanS2V [31] and LiveAvatar [14]. We selected 500 videos as our testset and set up interactive scenarios, ranging from simple short to complex long video interactions, with action switches at designated times. For non-real-time models, inference is performed in batches, referencing the last few frames of the previous batch to construct long videos. Quantitative results on relevant objective metrics are presented in Tab. 1.

In terms of video quality, while it does not achieve the best IQA, ASE, FID and FVD scores, InteractiveAvatar maintains reasonable visual fidelity, indicating its ability to generate avatars with consistent and plausible appearance. Regarding temporal and semantic consistency, InteractiveAvatar demonstrates clear advantages. It achieves the highest OBJ and TV, and competitive ID score, indicating strong object consistency across frames, robust identity preservation, and stable temporal variation throughout the video. Moreover, our model maintains accurate lip synchronization and exhibits improved alignment to user instructions compared with other baselines. These results highlight that InteractiveAvatar can generate videos that are not only visually coherent over time but also faithful to both content and user intent. Moreover, by leveraging the lightweight 5B model, InteractiveAvatar reduces hardware requirements while achieving higher inference speed and faster than all other baselines.

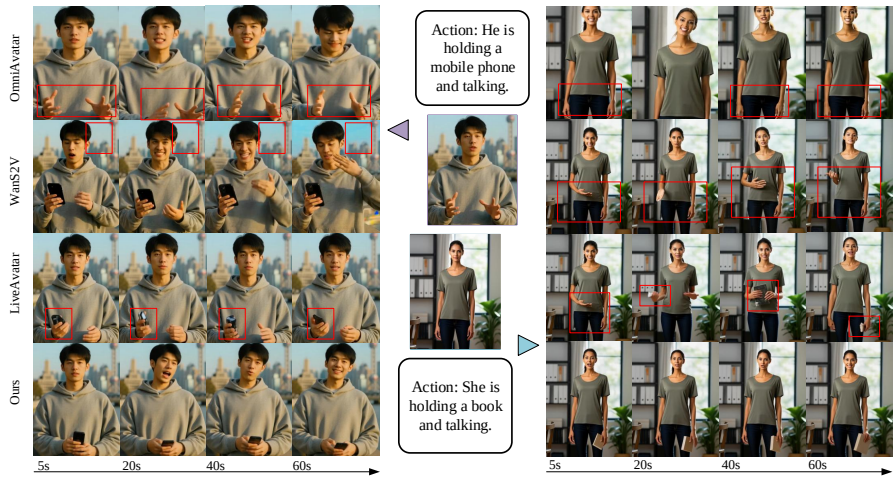

Fig. 4: Qualitative comparisons with state-of-the-art methods. Our method exhibits better visual consistency and following of action instructions.

Qualitative visualizations further compare our method with OmniAvatar [10], WanS2V [31] and LiveAvatar [14] in Fig. 4. In the two demonstrated scenarios, OmniAvatar fails to follow the action instructions and exhibits abrupt camera angle changes. WanS2V shows noticeable quality degradation over time in the first scenario and fails to follow action instructions in the second. LiveAvatar successfully triggers the specified objects according to the action instructions, but the objects' shapes and colors change rapidly during inference, resulting in low object consistency. In contrast, our method follows the input action instructions while maintaining high consistency for both objects and avatars throughout the interaction.

Table 1: Quantitative comparison with state-of-the-art methods. Best in **bold** and second best underlined. Experiments are conducted using the H100 GPU.

Model	Video Quality				Consistency					Speed
	IQA ↑	ASE ↑	FID ↓	FVD ↓	SynC ↑	SynD ↓	OBJ ↑	ID ↑	TV ↑	FPS ↑
StableAvatar	3.91	3.82	79.7	654.1	3.57	10.27	79.5	4.41	25.34	0.69
OmniAvatar	3.77	3.87	94.1	831.9	**4.94**	**7.87**	82.8	4.38	24.57	0.17
HYAvatar	3.81	**3.93**	**76.5**	**632.6**	4.78	8.11	78.9	4.46	25.61	0.09
Hallo3	3.57	3.29	112.5	1127.6	4.21	9.74	75.1	4.31	24.92	0.28
EchoMimicV3	**3.96**	3.89	85.2	773.9	3.89	10.09	80.3	4.45	25.56	0.81
WanS2V	3.76	3.68	88.4	793.5	4.54	8.95	82.6	4.49	25.14	0.26
LiveAvatar	3.94	3.91	83.9	672.7	4.91	8.17	76.9	**4.53**	25.78	21.94
Ours	3.87	3.89	80.2	701.4	4.86	7.91	**85.2**	4.51	**25.93**	**26.68**

Fig. 5: Qualitative ablation of InteractiveAvatar. Ablation studies show that our Full model maintains the best visual consistency and enables more realistic interactions.

4.5 Ablation Studies

We conduct ablation studies to analyze the impact of each component in our framework by systematically modifying or removing key modules, as summarized in Table 2. We evaluate both consistency-related metrics (OBJ, ID, TV) and inference speed (FPS).

Table 2: Ablation on the LSVM, RRM and DMD.

Method	OBJ ↑	ID ↑	TV ↑	FPS ↑
w/o LongMem	82.6	4.43	<u>25.91</u>	<u>28.92</u>
w/o DKFS	83.1	4.45	25.85	26.83
w/o LSVM	78.4	4.38	25.87	**30.04**
w/o StateCycling	84.1	4.46	25.42	26.68
w/o CacheSwitching	<u>84.5</u>	<u>4.47</u>	25.76	26.75
w/o RRM	83.8	4.46	24.89	26.75
w/o DMD	-	-	-	1.27
Ours	**85.2**	**4.51**	**25.93**	26.68

For the ablation study of the LSVM module, we design a scenario where the avatar wears a watch during interaction. Removing the long-term memory token (w/o LongMem) degrades object and identity consistency, demonstrating the necessity of long-range temporal modeling. Replacing Dynamic Key-Frame Selection with random sampling (w/o DKFS) causes slight distortions in the watch face, highlighting the advantage of informed memory updates. Removing the entire LSVM module (w/o LSVM) leads to a significant drop in OBJ, confirming its importance for visual consistency, though FPS increases due to reduced computation. As shown in Fig. 5, only the full model reliably preserves the watch's shape and color.

For the Reasoning-Reaction Module (RRM), removing the State Cycling strategy and using a fixed action prompt (w/o StateCycling) causes the model to repeatedly execute a single action, reducing interaction realism. Disabling Cache-

Switching (w/o CacheSwitching) increases response latency to prompt changes, which can cause delays or failures in the onset of prompt-related motions, making actions such as picking up and opening a book noticeably slower. Removing the entire RRM (w/o RRM) and reverting to a default prompt prevents the avatar from understanding user intent, resulting in only verbal responses without corresponding actions. These results demonstrate the necessity of dynamic state control and intent-aware reasoning. In contrast, our full model achieves natural interaction with low latency.

Finally, inference speed drops dramatically without DMD distillation (w/o DMD), verifying that DMD is essential for real-time performance. We do not report other metrics for the w/o DMD setting, as the model fails to achieve long video generation. When inference is performed by using the last frame of the previous segment as the first frame of the next segment, the video quality progressively degrades after multiple segments.

5 Conclusion

In this work, we present InteractiveAvatar, a novel real-time streaming diffusion framework that bridges the gap between high-quality avatar generation and realistic, intent-aware human-avatar interaction by jointly addressing two core challenges, long visual consistency and interactive responsiveness. Specifically, the proposed LSVM mechanism effectively mitigates temporal drift during extended streaming generation, preserving both local coherence and global identity consistency. Meanwhile, the RRM empowers the avatar to interpret user intent and generate coordinated speech and actions. In summary, InteractiveAvatar achieves real-time, long-duration, and intent-aware avatar generation, providing a practical foundation for next-generation immersive and intelligent digital humans.

Limitations and Future work Due to the limited memory capacity and keyframe selection strategy, long videos or large motions, especially those that deviate significantly from the reference frame, may degrade as early latents are discarded, or some generated objects may gradually disappear. In addition, objects absent from the first frame may appear abruptly when mentioned in the instruction, which is partly determined by the base model capability. Although prompt design can help smooth object emergence, this remains challenging. In future work, we aim to train end-to-end generative models with larger-scale data and computation to achieve smoother and more natural avatar interaction.

Acknowledgements

We thank our colleagues for their assistance with data processing, and the anonymous reviewers for their suggestive comments. This paper is supported by NSFC under grant No. 62495092, No.62125305 and Natural Science Basic Research Plan in Shaanxi Province of China (No. 2025SYS-SYSZD-023).

References

1. An, H., Hu, W., Huang, S., Huang, S., Li, R., Liang, Y., Shao, J., Song, Y., Wang, Z., Yuan, C., et al.: Ai flow: Perspectives, scenarios, and approaches (2025). arXiv preprint arXiv:2506.12479 (2025)
2. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156 (2025)
3. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2403–2410 (2025)
4. Cheng, S., Zhang, J., Song, Q., Liu, S., Guo, Z., Zhang, X., Zhang, C., Li, X., Tu, Z.: Unison: Harmonizing motion, speech, and sound for human-centric audio-video generation (2026), `https://arxiv.org/abs/2605.08729`
5. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian conference on computer vision. pp. 251–263. Springer (2016)
7. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21086–21095 (2025)
8. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
9. Ding, Y., Hu, X., Guo, Z., Zhang, C., Wang, Y.: Mtvcrafter: 4d motion tokenization for open-world human image animation. arXiv preprint arXiv:2505.10238 (2025)
10. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025)
11. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3497–3506 (2019)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
13. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
14. Huang, Y., Guo, H., Wu, F., Zhang, S., Huang, S., Gan, Q., Liu, L., Zhao, S., Chen, E., Liu, J., et al.: Live avatar: Streaming real-time audio-driven avatar generation with infinite length. arXiv preprint arXiv:2512.04677 (2025)
15. Jiang, W., Zhang, Y., Zheng, S., Liu, S., Yan, S.: Data augmentation in human-centric vision. Vicinagearth **1**(1), 8 (2024)
16. Kong, Z., Gao, F., Zhang, Y., Kang, Z., Wei, X., Cai, X., Chen, G., Luo, W.: Let them talk: Audio-driven multi-person conversational video generation. arXiv preprint arXiv:2505.22647 (2025)
17. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., et al.: Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. arXiv preprint arXiv:2412.00115 (2024)

18. Li, X., Wang, S., Zeng, S., Wu, Y., Yang, Y.: A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth **1**(1), 9 (2024)
19. Li, X.: Positive-incentive noise. IEEE Transactions on Neural Networks and Learning Systems **35**(6), 8708–8714 (2022)
20. Low, C., Wang, W.: Talkingmachines: Real-time audio-driven facetime-style video via autoregressive diffusion models. arXiv preprint arXiv:2506.03099 (2025)
21. Meng, R., Wang, Y., Wu, W., Zheng, R., Li, Y., Ma, C.: Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. arXiv preprint arXiv:2507.03905 (2025)
22. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20395–20405 (2022)
23. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia. pp. 484–492 (2020)
24. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
25. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
26. Sun, Z., Peng, Z., Ma, Y., Chen, Y., Zhou, Z., Zhou, Z., Zhang, G., Zhang, Y., Zhou, Y., Lu, Q., et al.: Streamavatar: Streaming diffusion models for real-time interactive human avatars. arXiv preprint arXiv:2512.22065 (2025)
27. Tian, L., Wang, Q., Zhang, B., Bo, L.: Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: European Conference on Computer Vision. pp. 244–260. Springer (2024)
28. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
29. Tu, S., Pan, Y., Huang, Y., Han, X., Xing, Z., Dai, Q., Luo, C., Wu, Z., Jiang, Y.G.: Stableavatar: Infinite-length audio-driven avatar video generation (2025), https://arxiv.org/abs/2508.08248
30. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
31. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
32. Wang, J., Wang, C., Huang, K., Huang, J., Jin, L.: Videoclip-xl: Advancing long description understanding for video clip models (2024), https://arxiv.org/abs/2410.00741
33. Wang, L., Zhu, Y., Ge, Z., Zheng, Y., Zhang, L., Hu, T., Qin, S., Luo, M., Zhang, J., Chen, X., et al.: Flowact-r1: Towards interactive humanoid video generation. arXiv preprint arXiv:2601.10103 (2026)

34. Wang, M., Wang, Q., Jiang, F., Fan, Y., Zhang, Y., Qi, Y., Zhao, K., Xu, M.: Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9891–9900 (2025)
35. Wang, Y., Fan, Y., Wang, X., Yu, G., Wang, F.: Diffusion-based realistic listening head generation via hybrid motion modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15885–15895 (2025)
36. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023)
37. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022)
38. Xie, Y., Gu, T., Li, Z., Zhang, C., Song, G., Zhao, X., Liang, C., Jiang, J., Xu, H., Luo, L.: X-streamer: Unified human world modeling with audiovisual interaction. arXiv preprint arXiv:2509.21574 (2025)
39. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
40. Xu, S., Chen, G., Guo, Y.X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B.: Vasa-1: Lifelike audio-driven talking faces generated in real time. Advances in Neural Information Processing Systems **37**, 660–684 (2024)
41. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622 (2025)
42. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)
43. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: CelebV-Text: A large-scale facial text-video dataset. In: CVPR (2023)
44. Zhang, H., Huang, S., Guo, Y., Li, X.: Variational positive-incentive noise: How noise benefits models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
45. Zhang, L., Cai, S., Li, M., Zeng, C., Lu, B., Rao, A., Han, S., Wetzstein, G., Agrawala, M.: Pretraining frame preservation in autoregressive video memory compression. arXiv preprint arXiv:2512.23851 (2025)
46. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8652–8661 (2023)
47. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021)
48. Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., et al.: Pytorch fsdp: experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277 (2023)
49. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG) **39**(6), 1–15 (2020)