This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ActAvatar: Temporally-Aware Precise Action Control for Talking Avatars

Ziqiao Peng[1*] Yi Chen[2] Yifeng Ma[2] Guozhen Zhang[2] Zhiyao Sun[2]
Zixiang Zhou[2] Youliang Zhang[2] Zhengguang Zhou[2] Zhaoxin Fan[1]
Hongyan Liu[3‡] Yuan Zhou[2†] Qinglin Lu[2‡] Jun He[1‡]
[1]Renmin University of China [2]Tencent Hunyuan [3]Tsinghua University
https://ziqiaopeng.github.io/ActAvatar/

Figure 1. ActAvatar generates talking avatars with precise, temporally-aligned actions across diverse scenarios and identities. Through structured text prompts, our method controls what actions to perform and when to perform them, while maintaining accurate lip synchronization with the audio.

Abstract

Despite significant advances in talking avatar generation, existing methods face critical challenges: insufficient text-following capability for diverse actions, lack of temporal alignment between actions and audio content, and dependency on additional control signals such as pose skeletons. We present ActAvatar, a framework that achieves phase-level precision in action control through textual guidance by capturing both action semantics and temporal context. Our approach introduces three core innovations: (1) Phase-Aware Cross-Attention (PACA), which decomposes prompts into a global base block and temporally-anchored phase blocks, enabling the model to concentrate on phase-relevant tokens for precise temporal-semantic alignment; (2) Progressive Audio-Visual Alignment, which aligns modality influence with the hierarchical feature learning process—early layers prioritize text for establishing action structure while deeper layers emphasize audio for refining lip movements, preventing modality interference; (3) A two-stage training strategy that first estab-

*Work done during an internship at Tencent Hunyuan.
†Project Leader.
‡Corresponding Author.

lishes robust audio-visual correspondence on diverse data, then injects action control through fine-tuning on structured annotations, maintaining both audio-visual alignment and the model's text-following capabilities. Extensive experiments demonstrate that ActAvatar significantly outperforms state-of-the-art methods in both action control and visual quality.

1. Introduction

Talking avatar generation has become increasingly important for applications ranging from virtual assistants [32] and digital entertainment to online education [56] and telepresence systems [35]. Recent advances in diffusion models [17, 42] have significantly improved the visual quality of generated talking avatars [7, 10, 21, 36, 45, 48, 58].

However, existing methods [21, 33, 40, 59] face three critical limitations. First, while current models can generate plausible hand movements [64], they struggle to accurately execute specific actions described in prompts due to treating the entire prompt uniformly [18], where action-related descriptions compete with scene descriptions for attention. Second, actions may appear at arbitrary moments rather than synchronizing with semantically relevant speech segments [58]. This temporal drift arises because standard conditioning mechanisms lack explicit temporal structure, causing attention to diffuse uniformly across time. Third, many methods resort to explicit control modalities such as pose skeleton sequences [8, 29], which increase pipeline complexity and limit the ability to generate novel actions.

These limitations reveal a fundamental challenge: establishing explicit correspondences between language semantics (what actions to perform), temporal windows (when to perform them), and audio cues (how they relate to speech). Moreover, text-driven action generation and audio-driven lip synchronization represent competing objectives that can interfere during generation. When both modalities exert strong influence simultaneously, the model struggles to balance conflicting signals, often resulting in degraded action quality or compromised lip-sync accuracy [47]. In addition, fine-tuning pre-trained models on domain-specific data to improve audio-visual alignment often leads to catastrophic forgetting, where the model's original text-following capabilities are weakened or lost entirely.

To address these challenges, we present **ActAvatar**, a framework that achieves temporally-aware, precise action control for talking avatar generation through textual guidance. Our key insight is threefold: (1) structured prompt organization with temporal anchors enables learned phase-conditioned attention dynamics for temporal-semantic alignment; (2) progressive modality influence prevents interference between text-driven action generation and audio-driven lip synchronization; (3) staged training preserves multiple capabilities by decoupling.

Our approach introduces three synergistic technical innovations. We propose **Phase-Aware Cross-Attention (PACA)**, which decomposes prompts into hierarchically structured phases with explicit temporal grounding. By organizing textual descriptions into a global base block and phase-specific blocks with temporal anchors, PACA enables the model to concentrate attention on temporally-relevant tokens during corresponding time windows.

Second, we develop **Progressive Audio-Visual Alignment**, which addresses the interference between text-driven action generation and audio-driven lip synchronization by aligning modality influence with the hierarchical feature learning process. Early transformer layers prioritize text conditioning to establish overall action structure. As generation progresses to deeper layers, audio emphasis gradually increases, allowing refinement of lip movements after the primary action framework has been determined. This progressive strategy prevents modality interference while ensuring both accurate action generation and precise lip sync.

Third, we propose a **two-stage training strategy** that addresses the capability preservation challenge through task decomposition. We first establish robust audio-visual alignment in Stage 1, then introduce temporal action control in Stage 2. This staged approach enables the model to learn action control as a compositional extension of existing capabilities rather than through destructive parameter updates.

In summary, our contributions are as follows:

- **Phase-Aware Cross-Attention:** A hierarchical prompt decomposition mechanism with temporal anchors that enables learned phase-conditioned attention dynamics for precise temporal-semantic alignment.
- **Progressive Audio-Visual Alignment:** A depth-aware modality scaling mechanism that aligns text and audio influence with hierarchical feature learning, preventing interference between text-driven action generation and audio-driven lip synchronization.
- **Staged Capability Preservation Training:** A two-stage training paradigm that addresses catastrophic forgetting through task decomposition, enabling compositional capability extension while preserving both audio-visual alignment and text-following capabilities.

2. Related Work

2.1. Video Generation Models

Text-to-video generation [20, 30, 37, 38, 51, 57] has witnessed remarkable progress with the advent of diffusion models. Early works such as Video Diffusion Models [14] and Imagen Video [13] established foundational frameworks by extending image diffusion to the temporal domain through 3D U-Net architectures and factorized spatiotemporal attention. Subsequent methods like AnimateD-

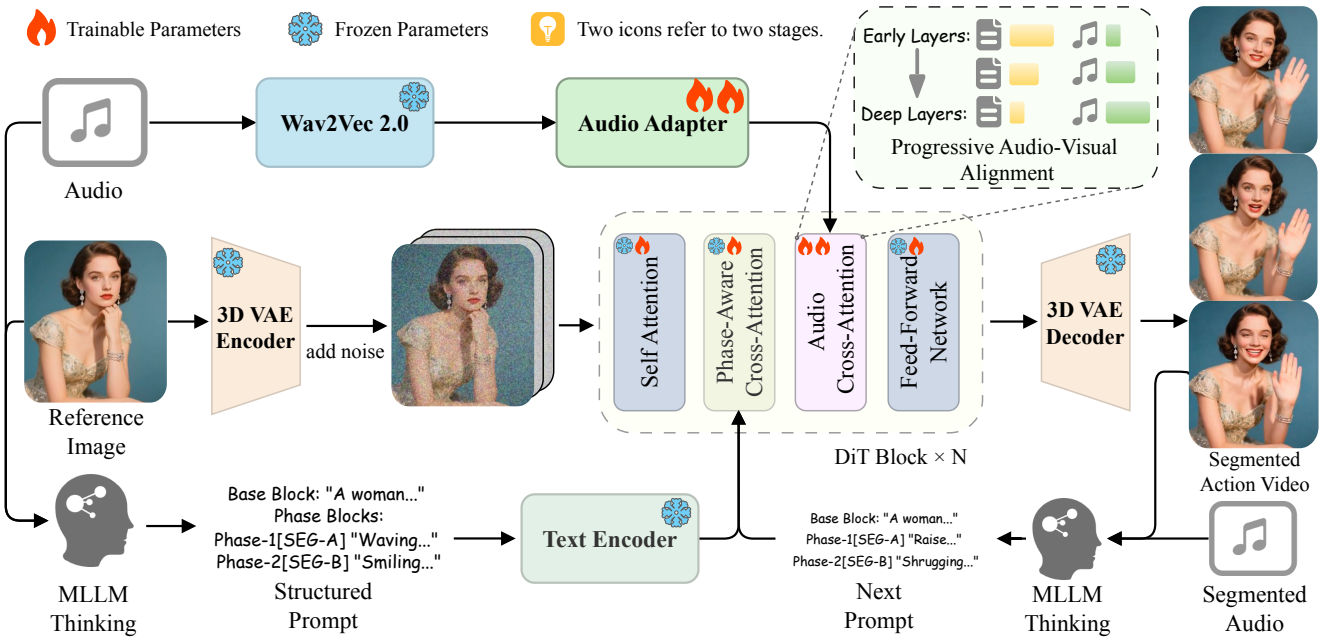

Figure 2. **Overview of ActAvatar.** Given an audio input and reference image, ActAvatar generates temporally-controlled action videos guided by structured prompts automatically generated from MLLM.

iff [12] and Stable Video Diffusion [2] achieved substantial improvements in visual quality and motion coherence by leveraging pre-trained text-to-image models and introducing specialized temporal modeling modules.

Recently, video generation models have shifted towards Transformer-based [41] architectures to achieve better scalability. HunyuanVideo [17] creates a comprehensive training framework that advances efficient model training and inference, Wan [42] builds a complete and open video generation model suite that drives the development of the open-source community, and SkyReels v2 [3] further extends video generation duration to infinite-length generation.

2.2. Talking Avatar Generation

Talking avatar generation [4, 7, 15, 23–28, 31, 34, 40, 43, 46, 50, 53, 54, 62, 63] aims to synthesize realistic human video with synchronized lip movements and natural expressions driven by audio input. EMO [39] pioneered audio-conditioned video diffusion for talking heads, introducing audio cross-attention for lip synchronization. EchoMimic [5]further improved controllability through multi-modal conditioning combining audio with visual references. Methods like Echomimic V2 [29] and Hallo 4 [8] resort to explicit pose guidance through skeleton sequences, introducing additional annotation requirements and limiting the naturalness of language-based interaction. MultiTalk [18] and HunyuanVideo-Avatar [4] extend these approaches to natural motion generation, but suffer from poor text-following capability.

Recent works further optimize this task, such as Kling-Avatar [9] and OmniHuman 1.5 [16]. These methods rely on strong video generation models to achieve relatively good action generation capabilities, but still perform diffusion on global prompts, failing to achieve precise temporal action control. AgentAvatar [44] try to introduce a timeline method, but only generate facial expressions. Our method addresses this limitation by introducing structured temporal prompts that enable phase-level precision in action control through hierarchical prompt decomposition and phase-aware attention mechanisms.

3. Method

3.1. Overview

ActAvatar aims to achieve precise temporal action control in talking avatar generation through structured prompt conditioning. Given an audio sequence $\mathbf{a} \in \mathbb{R}^{T_a \times D_a}$ and a reference image $\mathbf{I}_{\text{ref}} \in \mathbb{R}^{H \times W \times 3}$, our goal is to generate a video $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^{T}$ where the avatar exhibits accurate lip synchronization with the audio and performs specific actions at semantically appropriate temporal windows as described in the structured prompt $\mathbf{P}$. The structured prompt is automatically generated by a Multimodal Large Language Model (MLLM, e.g., Qwen3-Omni [52]) based on the input image and audio content.

Our framework builds upon an image-to-video diffusion backbone and introduces three synergistic components: (1) Phase-Aware Cross-Attention (PACA), which enables temporal-semantic alignment through hierarchical prompt decomposition and learned phase-conditioned attention dy-

namics; (2) Progressive Audio-Visual Alignment, which prevents modality interference by aligning text and audio influence with the hierarchical feature learning process; and (3) Two-Stage Training preserves both audio-visual correspondence and text-following capabilities through task decomposition. Figure 2 illustrates the overall architecture.

3.2. Phase-Aware Cross-Attention

3.2.1. Hierarchical Prompt Decomposition

Standard talking avatar methods condition generation on a single global prompt $\mathbf{P}_{\text{global}}$ that describes the overall scene. This flat representation lacks temporal structure, causing semantic diffusion where action-related information is uniformly distributed across all timesteps. To address this, we introduce hierarchical decomposition that explicitly encodes temporal grounding:

$$\mathbf{P} = \{\mathbf{P}_{\text{base}}, \{\mathbf{P}_k, \mathcal{T}_k\}_{k=1}^{K}\}, \tag{1}$$

where $\mathbf{P}_{\text{base}}$ is a global base block encoding time-invariant scene semantics including identity descriptors, environmental context, affective state, stylistic constraints, and global motion characteristics. Each phase block $\mathbf{P}_k$ describes temporally-localized actions within a designated temporal window $\mathcal{T}_k = [\tau_k^{\text{start}}, \tau_k^{\text{end}}]$ specified in normalized time coordinates. The base block establishes stable scene context, while phase blocks introduce temporal specificity through explicit time anchoring.

For example: *Base*: "A woman in business attire speaking professionally"; *Phase-1 [0-2s]*: "Gestures outward with open palm"; *Phase-2 [2-4s]*: "Points downward to emphasize detail."

3.2.2. Phase-Conditioned Attention Dynamics

Let $\mathbf{x}_f \in \mathbb{R}^{N \times D}$ denote the video latent features at frame index f, where N is the number of spatial tokens and D is the feature dimension. The structured prompt $\mathbf{P}$ is first encoded into a token sequence through a pre-trained text encoder (umT5-XXL), yielding $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^{M} \in \mathbb{R}^{M \times D_c}$.

Phase Position Encoding. To explicitly encode phase membership and enhance phase-conditioned attention, we introduce learnable phase position embeddings. For each token $\mathbf{c}_i$ belonging to phase k, we add a phase-specific positional bias:

$$\mathbf{c}_i' = \mathbf{c}_i + \mathbf{e}_k, \tag{2}$$

where $\mathbf{e}_k \in \mathbb{R}^{D_c}$ is a learnable phase embedding for phase k. The embeddings are zero-initialized to ensure identity behavior at the start of training. These phase embeddings provide an inductive bias that encourages the model to distinguish between base and phase-specific tokens in the attention mechanism.

In standard cross-attention, queries $\mathbf{Q}_f = \mathbf{x}_f \mathbf{W}_Q$, keys $\mathbf{K} = \mathbf{C}'\mathbf{W}_K$, and values $\mathbf{V} = \mathbf{C}'\mathbf{W}_V$ are computed from the phase-augmented token embeddings:

$$\text{Attention}(\mathbf{Q}_f, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}_f \mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V}. \tag{3}$$

Through training on temporally-annotated data, the model learns to concentrate attention on phase-relevant tokens when frame f (normalized to video time $\tau \in [0, T_{\text{video}}]$) falls within the corresponding temporal range $\mathcal{T}_k$, achieving temporal-semantic alignment through the combination of temporal anchors and phase position embeddings.

3.3. Progressive Audio-Visual Alignment

Text-driven action generation and audio-driven lip synchronization represent competing objectives that can interfere during the generation process. To address this, we introduce progressive audio-visual alignment that aligns modality influence with the hierarchical feature learning process in diffusion transformers.

Diffusion transformers naturally follow a coarse-to-fine feature learning hierarchy: early layers capture global structure and layout, while deeper layers refine local details and high-frequency features. We leverage this characteristic to prevent modality interference by progressively scaling audio influence across transformer blocks.

For transformer block $\ell \in \{1, \ldots, L\}$, we apply depth-aware scaling to the audio cross-attention residual:

$$\mathbf{x}_\ell \leftarrow \mathbf{x}_\ell + f(\ell) \cdot \mathbf{r}_{\text{audio}}^{\ell}, \tag{4}$$

where the scaling function is:

$$f(\ell) = \left(\frac{\ell}{L}\right)^{\gamma}, \tag{5}$$

with $\gamma > 1$. This creates progressive amplification of audio influence in deeper layers.

The design aligns modality influence with the generation hierarchy. Early layers ($\ell \ll L$, small $f(\ell)$) prioritize text conditioning to establish overall action structure—body pose, hand trajectory, and gesture type. During this phase, audio influence remains minimal, allowing text to dominate the generation of action semantics without interference from audio signals. As generation progresses to deeper layers ($\ell \to L$, larger $f(\ell)$), audio emphasis gradually increases, enabling precise refinement of lip movements and facial articulation after the primary action framework has been determined.

This progressive strategy prevents modality interference by ensuring that text and audio operate in complementary rather than competing regimes: text establishes the action structure in early layers where coarse features dominate,

while audio refines lip articulation in deep layers where high-frequency details emerge.

3.4. Two-Stage Training Strategy

Our two-stage training strategy addresses the poor text-following capability observed in previous talking avatar methods by decoupling audio-visual learning from temporal action control. This staged approach enables the model to first establish robust audio-visual correspondence, then integrate precise temporal action semantics without compromising either capability.

3.4.1. Stage 1: Audio Adapter Training and Extraction

Stage 1 establishes robust audio-visual correspondence by training on diverse and large-scale talking-head videos. We adopt the Flow Matching training paradigm [22]. Given the original latent representation $\mathbf{x}_0$ (data) and pure noise $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$, we construct the optimal transport flow path:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \tag{6}$$

where $t \in [0, 1]$ is the flow time uniformly sampled from the training timestep sequence. The model is trained to predict the velocity field:

$$\mathbf{v}_{\text{target}} = \mathbf{x}_1 - \mathbf{x}_0, \tag{7}$$

which represents the direction from data to noise. The complete Stage 1 loss is:

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}_{\mathbf{x}_0, t, \mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{C}_{\text{brief}}, \mathbf{A}) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right], \tag{8}$$

where $\mathbf{C}_{\text{brief}}$ represents brief text captions (e.g., "A woman speaking"), and $\mathbf{A}$ are audio embeddings from Wav2Vec 2.0 [1]. The audio adapter consists of an audio projection module (mapping audio encodings to frame-aligned tokens) and audio cross-attention layers (fusing audio semantics via frame-wise attention).

Crucially, we freeze the base text-to-video backbone parameters θ_{base} and only train the audio adapter parameters θ_{audio}:

$$\theta_{\text{stage1}} = \{\theta_{\text{audio}}\}, \quad \theta_{\text{base}} \text{ frozen}. \tag{9}$$

3.4.2. Stage 2: Temporally-Aware Action Control

Stage 2 injects temporal action control through structured annotations. We construct a dataset via: (1) applying DWPose [55] to compute motion magnitude and selecting videos with significant movement; (2) using a multimodal large language model to generate hierarchical prompts with base blocks and phase-specific descriptions with temporal anchors. This yields a focused dataset with high-quality structured annotations.

We construct the Stage 2 model by starting with a base image-to-video model and injecting the pretrained audio adapter from Stage 1. The training objective maintains the Flow Matching form:

$$\mathcal{L}_{\text{stage2}} = \mathbb{E}_{\mathbf{x}_0, t, \mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{C}_{\text{PACA}}, \mathbf{A}) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right], \tag{10}$$

where $\mathbf{C}_{\text{PACA}}$ is the hierarchically-structured prompt encoding with phase position embeddings.

In Stage 2, we adopt full fine-tuning that simultaneously optimizes both speech synchronization and action control:

$$\theta_{\text{stage2}} = \{\theta_{\text{base}}, \theta_{\text{audio}}, \theta_{\text{PACA}}\}. \tag{11}$$

The two-stage approach preserves both audio-visual correspondence and text-following capabilities: Stage 1 establishes robust lip synchronization on diverse data with frozen backbone, while Stage 2's full fine-tuning on structured annotations enables precise temporal action control.

4. Experiments

4.1. Experimental Setup

4.1.1. Implementation Details

We implement ActAvatar using PyTorch on 40 NVIDIA H20 GPUs. The backbone is Wan2.2-TI2V-5B with 30 DiT blocks. The audio encoder is Wav2Vec 2.0 and text encoder is umT5-XXL. Stage 1 trains for 20K steps and Stage 2 for 14K steps, both with batch size 40, learning rate 5×10^{-6}, and AdamW optimizer. For Progressive Audio-Visual Alignment, we set $\gamma = 1.5$ in $f(\ell) = (\ell/30)^\gamma$. We generate 125-frame videos (5s at 25 FPS) at 704×1280 resolution using flow-matching sampling with 40 steps and classifier-free guidance scale 5.0 for both text and audio.

4.1.2. Datasets

Training Data. For Stage 1 audio-driven lip synchronization training, we utilize 500K diverse talking head videos from OpenHumanVid [19] and SpeakerVid [60], covering varied speakers, emotions, and speaking styles. For Stage 2, we construct a structured annotation dataset through DWPose-based motion selection and MLLM-based prompt generation, yielding 100K samples with phase-level temporal annotations. The detailed construction pipeline is provided in the supplementary material.

Evaluation Data. We evaluate on two test sets: (1) *HDTF Test Set* [61], containing 100 high-quality talking-head videos (5s each) focusing on lip synchronization quality, as it includes only upper body without hand movements; (2) *Action Bench*, our constructed benchmark consisting of 200 samples with diverse action instructions. Each sample includes a reference image, TTS-synthesized speech, and structured prompts with action annotations.The prompts are MLLM-generated with human verification.

Table 1. Quantitative comparison on HDTF Test Set. Best results in **bold**, second best underlined.

Method	Visual Quality				Lip Sync		Model Info		
	FID↓	FVD↓	IQA↑	ASE↑	Sync-C↑	Sync-D↓	Params	Resolution	Time
Hallo3 [7]	30.424	318.451	3.883	2.488	6.994	8.647	5B	480p	32 min
FantasyTalking [45]	25.615	458.733	3.955	2.663	4.179	11.067	14B	720p	83 min
EchoMimic v3 [28]	43.544	595.805	4.027	2.690	4.555	10.804	1.3B	480p	7 min
HunyuanVideo-Avatar [4]	24.515	322.844	4.054	2.693	7.647	7.564	13B	720p	74 min
MultiTalk [18]	25.578	368.719	3.981	2.570	7.610	7.645	14B	480p	45 min
OmniAvatar [10]	24.398	374.594	4.088	2.664	**7.986**	7.696	14B	480p	36 min
StableAvatar [40]	23.742	323.601	3.984	2.572	5.457	10.151	1.3B	480p	12 min
Wan-S2V [11]	23.850	**299.491**	4.108	2.684	7.462	7.745	14B	720p	68 min
ActAvatar (Ours)	**23.471**	301.064	**4.120**	**2.714**	7.663	**7.545**	5B	720p	16 min

Table 2. Quantitative comparison on Action Bench. Best results in **bold**, second best underlined.

Method	Lip Sync		Visual Quality		Gemini-Based Action Metrics				
	Sync-C↑	Sync-D↓	IQA↑	ASE↑	Hit@ Segment↑	Action Accuracy↑	Temporal Correctness↑	Action Quality↑	Hand Clarity↑
Hallo3 [7]	4.288	10.266	4.462	3.531	0.756	5.299	6.541	6.793	7.844
FantasyTalking [45]	4.209	10.948	4.510	3.560	0.728	4.995	6.435	6.501	8.147
EchoMimic v3 [28]	3.199	11.896	4.301	3.543	0.764	4.717	6.684	5.817	6.115
HunyuanVideo-Avatar [4]	6.251	8.862	4.759	3.725	0.674	3.977	5.491	6.609	8.044
MultiTalk [18]	6.654	8.356	4.446	3.477	0.574	3.490	4.654	5.455	7.584
OmniAvatar [10]	6.765	8.785	4.569	3.610	0.818	5.505	7.032	7.147	8.042
StableAvatar [40]	4.035	11.615	4.760	3.658	0.758	4.802	6.576	6.722	7.768
Wan-S2V [11]	6.473	8.401	4.701	3.647	0.754	4.934	6.465	6.630	8.168
ActAvatar (Ours)	**6.893**	**8.246**	**4.814**	**3.743**	**0.854**	**5.971**	**7.353**	**7.671**	**8.483**

4.1.3. Evaluation Metrics

We evaluate ActAvatar using comprehensive metrics. For lip sync, we use Sync-C and Sync-D from SyncNet [6] (higher Sync-C and lower Sync-D indicate better alignment). For visual quality, we report FID and FVD (lower is better), and use Q-Align [49] for video quality (IQA) and aesthetics (ASE) (higher is better). For action control, we develop a Gemini-based evaluation framework that provides: (1) Action Occurrence (AO): whether the described action appears; (2) Action Accuracy (AA): how well the action matches the description (0-10); (3) Temporal Correctness (TC): whether the action occurs in the specified time window (0-10); (4) Action Quality (AQ): overall naturalness of execution (0-10); (5) Hand Clarity (HC): hand quality (0-10). We report Hit@Segment (H@S) as the percentage of phases where AO=1, and mean scores for AA, TC, AQ, and HC. We run the Gemini evaluation 5 times for each video and report the average scores.

4.2. Quantitative Evaluation

We present comprehensive quantitative comparisons on both HDTF Test Set and Action Bench. Table 1 shows results on HDTF, focusing on lip synchronization and visual quality in natural talking scenarios. And table 2 presents results on Action Bench, evaluating action control capabilities alongside lip-sync and visual quality.

Performance on HDTF. ActAvatar achieves the best visual quality (FID: 23.471, IQA: 4.120, ASE: 2.714) while maintaining competitive lip synchronization (Sync-C: 7.663, Sync-D: 7.545). Operating at 720p with only 5B parameters, ActAvatar matches or exceeds 14B models running at lower resolutions. The lip-sync performance is on par with the best methods, demonstrating that our two-stage training successfully preserves audio-visual alignment.

Performance on Action Bench. ActAvatar demonstrates substantial advantages in action control. We achieve the highest H@S (0.854), significantly outperforming methods with other methods (OmniAvatar: 0.818, EchoMimic v3: 0.764). Gemini-based metrics show consistent superiority: AA (5.971 vs. 5.505), TC (7.353 vs. 7.032), AQ (7.671 vs. 7.147), and HC (8.483 vs. 8.168). Remarkably, ActAvatar achieves the best lip synchronization on Action Bench (Sync-C: 6.893), demonstrating that PACA enables precise action control without sacrificing audio-visual alignment.

Inference Efficiency. We further evaluate inference speed on a single H20 GPU. ActAvatar generates 5-second videos in 16 minutes, achieving more than 4× speedup over comparable methods (Wan-S2V: 68 min, FantasyTalking: 83 min,

Figure 3. Qualitative comparison with state-of-the-art methods. ActAvatar accurately executes phase-specific actions with correct timing and clear hand articulation, while competing methods show temporal misalignment, vague motions, or degraded hand quality.

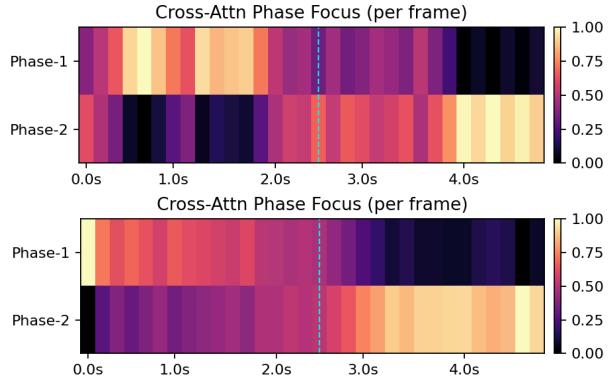

Figure 4. Cross-attention phase focus at layer 5 (top) and layer 20 (bottom). Deeper layers show sharper phase separation.

HunyuanVideo-Avatar: 74 min) while maintaining superior quality. With 8× H20 GPUs, generation time reduces to just 2 minutes per 5-second video. Although lightweight models like EchoMimic v3 (7 min) and StableAvatar (12 min) are faster, they show significantly degraded quality and lip-sync. ActAvatar's 5B model achieves optimal quality-efficiency balance, delivering 14B-level performance at competitive speeds.

4.3. Qualitative Analysis

Attention Visualization. We visualizes the cross-attention distribution across time for different phase blocks at layer 5 (top) and layer 20 (bottom) (Figure 4). For a two-phase prompt, the attention naturally concentrates on Phase-1 tokens during the first half and shifts to Phase-2 tokens during the second half. This phase-conditioned attention pattern emerges, validating that our PACA module enables learned temporal-semantic correspondence.

Comparing layer 5 and layer 20 reveals hierarchical refinement: early layers show coarse phase awareness with some overlap around the boundary, while deeper layers exhibit sharp phase separation with concentrated attention. This aligns with the coarse-to-fine feature learning in transformers and explains the high Temporal Correctness scores in quantitative evaluation.

Visual Quality Comparison. Figure 3 presents side-by-side qualitative comparisons with state-of-the-art methods on two representative examples from Action Bench, each featuring two distinct action phases with structured prompts. ActAvatar successfully executes both action phases with clear temporal separation and natural transitions. In contrast, competing methods exhibit various limitations. Echomimic V3 completely collapses when faced with such large-scale motions. Due to an over-reliance on the reference frame input, Hunyuanvideo-Avatar generates artifacts such as a third hand. Hallo3, StableAvatar, and FantasyTalking have very weak responsiveness to text, producing almost exclusively lip motion. While MultiTalk exhibits some hand movement, the motion does not follow the content of the prompt.

Table 3. User study results.

Method	APA ↑	AQ ↑	HC ↑	LSA ↑	OVQ ↑
EchoMimic v3 [28]	3.43	3.25	3.21	2.45	2.68
HunyuanVideo-Avatar [4]	2.84	3.12	3.68	3.67	3.45
MultiTalk [18]	3.19	3.16	3.95	3.41	3.58
OmniAvatar [10]	3.75	3.92	4.08	3.56	3.61
Wan-S2V [11]	3.67	3.78	4.12	3.68	4.04
ActAvatar (Ours)	**4.03**	**4.15**	**4.22**	**3.89**	**4.18**

Figure 5. Ablation study on PACA. **Top**: Without PACA, the avatar remains static throughout the sequence. **Bottom**: With PACA, the avatar naturally walks forward.

4.4. User Study

To complement quantitative evaluation, we conduct a user study with 45 participants. Participants evaluate videos generated by ActAvatar and competing methods for the same prompt, rating five dimensions on a 0-5 scale: (1) Action-Prompt Alignment (APA): how well actions match prompt descriptions; (2) Action Quality (AQ): naturalness and expressiveness of movements; (3) Hand Clarity (HC): clarity of hand gestures; (4) Lip Sync Accuracy (LSA): synchronization between lip movements and audio; (5) Overall Video Quality (OVQ): overall visual fidelity and realism. Each participant evaluates 30 videos with randomized order. Table 3 presents the mean scores across all participants.

ActAvatar achieves the highest scores across all dimensions, with particularly strong performance in action-related metrics. For Action-Prompt Alignment (APA: 4.03), ActAvatar substantially outperforms all baselines, confirming that PACA enables perceptually recognizable temporal action control. Hand Clarity (HC: 4.22) is ActAvatar's strongest dimension, validating that the model maintains clear hand articulation during dynamic gestures.

4.5. Ablation Studies

To clarify the contributions of each core component in our framework, we conduct an ablation study targeting three key modules. Table 4 validates the effectiveness of each core component in ActAvatar.

PACA. Adding PACA to the base model substantially improves action control (H@S: 0.725 → 0.829), with corresponding gains in Action Accuracy (AA: 3.91 → 5.78) and Temporal Correctness (TC: 6.47 → 7.48). Figure 5 provides visual compare: without PACA, the avatar remains static, while with PACA, dynamic motion emerges naturally. PACA effectively enables phase-conditioned attention for temporal-semantic alignment.

Table 4. Ablation study of key components on Action Bench.

Configuration	Sync-C↑	H@S↑	AA↑	TC↑	AQ↑	HC↑
Base (Global Prompt)	6.37	0.725	3.91	6.47	6.21	7.68
+ PACA	6.39	0.829	5.78	7.12	7.48	8.36
+ PACA + Progressive Alignment	6.57	0.831	5.75	7.10	7.52	8.47
+ Two-Stage Training (Full)	**6.89**	**0.854**	**5.97**	**7.35**	**7.67**	**8.48**

Progressive Audio Alignment. Adding depth-aware audio scaling improves lip synchronization (Sync-C: 6.39 → 6.57) while maintaining strong action performance (H@S: 0.831). The progressive strategy prevents modality interference by allowing text to dominate action generation in early layers while audio refines lip articulation in deeper layers.

Two-Stage Training. The complete two-stage training strategy provides further improvements across all metrics, achieving the best lip synchronization (Sync-C: 6.89) and action control (H@S: 0.854, AA: 5.97, TC: 7.35). This validates that decoupling audio-visual learning from action control injection is essential for maintaining both capabilities simultaneously.

5. Conclusion

We present ActAvatar, a framework achieving precise temporal action control in talking avatar generation through textual guidance. By introducing Phase-Aware Cross-Attention, Progressive Audio-Visual Alignment, and a two-stage training strategy, ActAvatar addresses fundamental limitations of poor text-following and temporal misalignment in existing methods. Extensive experiments demonstrate that our approach significantly outperforms the SOTA methods in action accuracy, lip synchronization, and visual fidelity. Our work shows that structured textual conditioning can achieve phase-level temporal precision without additional control signals, opening new possibilities for controllable talking avatar generation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62436010, 72572090, 62572474, 62172421, and 62441617, the Beijing Natural Science Foundation under Grant 4254100, the China Postdoctoral Science Foundation under Grant 2024M764093, and the 2023 Outstanding Innovative Talents Cultivation Fund Program of Renmin University of China.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 5

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[3] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 3

[4] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025. 3, 6, 8

[5] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2403–2410, 2025. 3

[6] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 6

[7] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv e-prints*, pages arXiv–2412, 2024. 2, 3, 6

[8] Jiahao Cui, Yan Chen, Mingwang Xu, Hanlin Shang, Yuxuan Chen, Yun Zhan, Zilong Dong, Yao Yao, Jingdong Wang, and Siyu Zhu. Hallo4: High-fidelity dynamic portrait animation via direct preference optimization and temporal motion modulation. *arXiv preprint arXiv:2505.23525*, 2025. 2, 3

[9] Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, et al. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*, 2025. 3

[10] Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025. 2, 6, 8

[11] Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025. 6, 8

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[15] Wentao Hu, Shunkai Li, Ziqiao Peng, Haoxian Zhang, Fan Shi, Xiaoqiang Liu, Pengfei Wan, Di Zhang, and Hui Tian. Ggtalker: Talking head systhesis with generalizable gaussian priors and identity-specific adaptation. *arXiv preprint arXiv:2506.21513*, 2025. 3

[16] Jianwen Jiang, Weihong Zeng, Zerong Zheng, Jiaqi Yang, Chao Liang, Wang Liao, Han Liang, Yuan Zhang, and Mingyuan Gao. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv preprint arXiv:2508.19209*, 2025. 3

[17] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3

[18] Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025. 2, 3, 6, 8

[19] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7752–7762, 2025. 5

[20] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 2

[21] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13847–13858, 2025. 2

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5

[23] Xingpei Ma, Shenneng Huang, Jiaran Cai, Yuansheng Guan, Shen Zheng, Hanfeng Zhao, Qiang Zhang, and Shunsi Zhang. Playmate2: Training-free multi-character audio-driven animation via diffusion transformer with reward feedback. *arXiv preprint arXiv:2510.12089*, 2025. 3

[24] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking

styles. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1896–1904, 2023.

[25] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2(3), 2023.

[26] Yifeng Ma, Jinwei Qi, Chaonan Ji, Peng Zhang, Bang Zhang, Zhidong Deng, and Liefeng Bo. Exploring timeline control for facial motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1940–1950, 2025.

[27] Yifeng Ma, Suzhen Wang, Yu Ding, Bowen Ma, Tangjie Lv, Changjie Fan, Zhipeng Hu, Zhidong Deng, and Xin Yu. Talkclip: Talking head generation with text-guided expressive speaking styles. *IEEE Transactions on Multimedia*, 2025.

[28] Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. *arXiv preprint arXiv:2507.03905*, 2025. 3, 6, 8

[29] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5489–5498, 2025. 2, 3

[30] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 2

[31] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023. 3

[32] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20687–20697, 2023. 2

[33] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024. 2

[34] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025. 3

[35] Ziqiao Peng, Wentao Hu, Junyuan Ma, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Hui Tian, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk++: High-fidelity and efficient synchronized talking heads synthesis using gaussian splatting. *arXiv preprint arXiv:2506.14742*, 2025. 2

[36] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025. 2

[37] Zhefan Rao, Liya Ji, Yazhou Xing, Runtao Liu, Zhaoyang Liu, Jiaxin Xie, Ziqiao Peng, Yingqing He, and Qifeng Chen. Modelgrow: Continual text-to-video pre-training with model expansion and language understanding enhancement. *arXiv preprint arXiv:2412.18966*, 2024. 2

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[39] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024. 3

[40] Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025. 2, 3, 6

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[42] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[43] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 3

[44] Duomin Wang, Bin Dai, Yu Deng, and Baoyuan Wang. Agentavatar: Disentangling planning, driving and rendering for photorealistic avatar agents. *arXiv preprint arXiv:2311.17465*, 2023. 3

[45] Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9891–9900, 2025. 2, 6

[46] Suzhen Wang, Yifeng Ma, Yu Ding, Zhipeng Hu, Changjie Fan, Tangjie Lv, Zhidong Deng, and Xin Yu. Styletalk++: A unified framework for controlling the speaking styles of talking heads. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4331–4347, 2024. 3

[47] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. 2

[48] Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li,

Tingbo Hou, et al. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025. 2

[49] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 6

[50] Haoyu Wu, Ziqiao Peng, Xukun Zhou, Yunfei Cheng, Jun He, Hongyan Liu, and Zhaoxin Fan. Vgg-tex: A vivid geometry-guided facial texture estimation model for high fidelity monocular 3d face reconstruction. *arXiv preprint arXiv:2409.09740*, 2024. 3

[51] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7623–7633, 2023. 2

[52] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 3

[53] Kaixing Yang, Xulong Tang, Ziqiao Peng, Yuxuan Hu, Jun He, and Hongyan Liu. Megadance: Mixture-of-experts architecture for genre-aware 3d dance generation. *arXiv preprint arXiv:2505.17543*, 2025. 3

[54] Kaixing Yang, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Puwei Wang, Jun He, and Hongyan Liu. Flowerdance: Meanflow for efficient and refined 3d dance generation. *arXiv preprint arXiv:2511.21029*, 2025. 3

[55] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 5

[56] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. 2

[57] Guozhen Zhang, Haiguang Wang, Chunyu Wang, Yuan Zhou, Qinglin Lu, and Limin Wang. Arbitrary generative video interpolation. *arXiv preprint arXiv:2510.00578*, 2025. 2

[58] Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025. 2

[59] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023. 2

[60] Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025. 5

[61] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 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*, pages 3661–3670, 2021. 5

[62] Zongzheng Zhang, Jiawen Yang, Ziqiao Peng, Meng Yang, Jianzhu Ma, Lin Cheng, Huazhe Xu, Hang Zhao, and Hao Zhao. Morpheus: A neural-driven animatronic face with hybrid actuation and diverse emotion control. *arXiv preprint arXiv:2507.16645*, 2025. 3

[63] Xukun Zhou, Fengxin Li, Ziqiao Peng, Kejian Wu, Jun He, Biao Qin, Zhaoxin Fan, and Hongyan Liu. Meta-learning empowered meta-face: Personalized speaking style adaptation for audio-driven 3d talking face animation. *arXiv preprint arXiv:2408.09357*, 2024. 3

[64] Xukun Zhou, Fengxin Li, Ming Chen, Yan Zhou, Pengfei Wan, Di Zhang, Yeying Jin, Zhaoxin Fan, Hongyan Liu, and Jun He. Exges: Expressive human motion retrieval and modulation for audio-driven gesture synthesis. *arXiv preprint arXiv:2503.06499*, 2025. 2