NEXTAVATAR: NEXT-STATE PLANNING FOR STREAMING AVATAR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming avatar generation synthesizes video chunk by chunk, conditioned on visual history and text instructions. Each chunk typically covers about one second, often too short to complete an action, so a single instruction spans multiple chunks during training and inference. Coherent action progression therefore requires the generator to account for what the avatar has already done. However, we observe that streaming generators often stagnate before executing an action or repeat it after completion. To investigate these failures, we analyze the attention mechanisms that incorporate visual history and text instructions into generation. We identify **history-state confusion**: attention to both conditions remains largely unchanged across distinct execution states under the same instruction, failing to adequately distinguish action progress and leading to stagnation and repetition. To address this problem, we propose **NextAvatar**. We use a frozen pretrained multimodal large language model to interpret the history state and active instruction, then provide an explicit next-state visual plan through state-conditioned attention. To reduce appearance deviations in the planned state, we introduce Reference-Grounded State Anchoring, which refines its visual features through asymmetric attention to fixed initial-reference features. We further introduce the **Multi-Instruction Responsive Avatar Benchmark (MIRA-Bench)**, comprising 2,000 expression–action intervals to evaluate Human Fidelity, Instruction Responsiveness, Video Quality, and Action Progression. With a 1.3B generator backbone, NextAvatar achieves the strongest Instruction Responsiveness among the evaluated methods and reduces combined stagnation and repetition from the best external baseline's 52.5% to 30.5%, while maintaining competitive Human Fidelity and Video Quality. A demo page is available at `https://anonymous.4open.science/w/nextavatar_site/`.

1 INTRODUCTION

Avatar video generation enables digital humans for virtual communication, online presentation, and interactive assistance. To support sustained interaction, streaming methods generate video chunk by chunk using preceding visual context (Yin et al., 2025; Huang et al., 2025b; Zhu et al., 2026; Zhang et al., 2025; Yuan et al., 2026). Building on this approach, streaming avatar systems have focused on real-time synthesis, identity preservation, and speech-driven lip synchronization (Huang et al., 2025c; Zhen et al., 2026; Shen et al., 2025; Zhang et al., 2026). Beyond these capabilities, accurate instruction control over facial expressions and body actions is essential for aligning avatar behavior with user intent and supporting natural, realistic interactions.

Instruction control is complicated by the temporal granularity of streaming generation. A chunk typically covers about one second, often too short to complete a full action. Consequently, instructions may span diverse numbers of chunks, and coherent action progression requires the generator to account for what the avatar has already done. Figure 1(a) illustrates four possible continuations under the same instruction. Before the gesture begins, the avatar should perform it, whereas remaining still constitutes stagnation. After the gesture is completed, the avatar should lower its hand and remain still, whereas performing the gesture again constitutes repetition. We call the action progress reflected in the preceding video the *history state*, and the intended behavior in the upcoming chunk the *next state*. Our pilot study on randomly generated examples (Figure 1(b)) reveals substantial stagnation and repetition. These failures occur in NextAvatar-Base, a streaming avatar generator trained by us (Appendix C), as well as SoulX-FlashTalk (Shen et al., 2025) and LongCat-Avatar 1.5 (Cai et al., 2026).

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

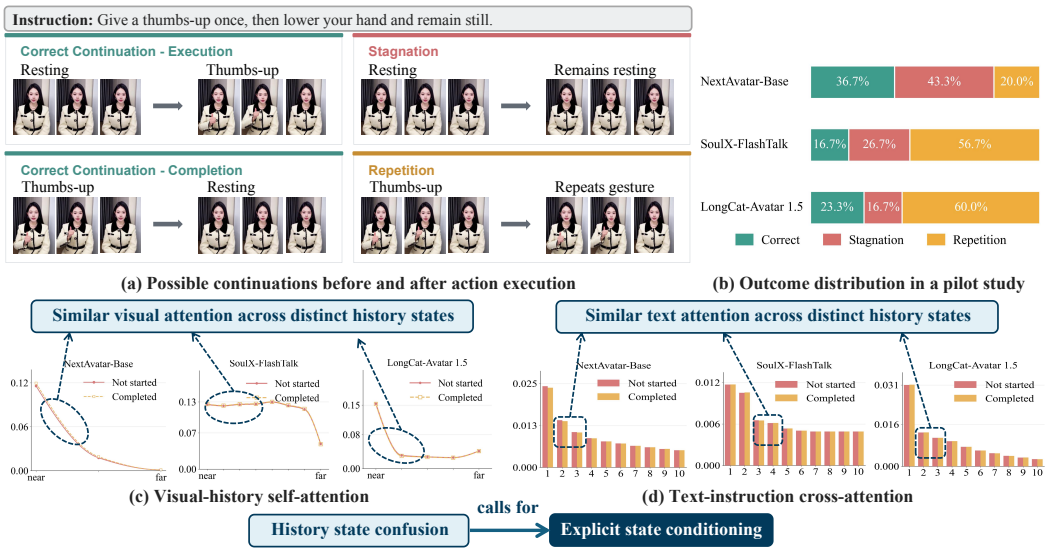

Figure 1: **History-state confusion in streaming avatar generation.** (a) Possible continuations before and after action execution: correct execution or completion, versus stagnation or repetition. (b) Outcome distribution for three streaming avatar models in a pilot study on randomly generated examples. (c) Visual-history self-attention is allocated similarly across temporal positions despite different history states. (d) Text-instruction cross-attention is likewise similar across the two states; the ten shared instruction tokens are ranked under the not-started condition.

We compare paired runs with the same instruction and different history states to examine how these models attend to visual history and text. Across models, self-attention is allocated similarly across visual-history positions in the not-started and completed states (Figure 1(c)). Cross-attention over instruction tokens likewise changes little between the two states (Figure 1(d)). These attention patterns suggest that attention to the visual history and instruction does not adequately reflect changes in action progress. We term this **history-state confusion**: the current attention mechanisms respond similarly to histories with different action progress. This calls for explicit state conditioning based on visual history and the instruction.

To address history-state confusion, we propose **NextAvatar**, which maps the history state and active instruction to an explicit next-state condition through **Next-State Planning (NSP)**. Determining this condition requires interpreting action progress in the visual history alongside the instruction. Drawing on multimodal video understanding (Wei et al., 2025; Yang et al., 2026) and visual semantic planning (Huang et al., 2025a; Liu et al., 2026), NSP uses a frozen pretrained multimodal large language model (MLLM) to predict a visual representation of the appropriate next state. The generator receives this representation through state-conditioned attention to guide its continuation. In practice, we observe that predicted states can inherit appearance deviations from the generated history, which may accumulate over successive chunks. We therefore introduce **Reference-Grounded State Anchoring**, which refines the predicted state using fixed initial-reference features before it conditions generation. This provides an identity anchor alongside the generator's native reference input.

To better evaluate the instruction-driven action progression problem identified above, along with common challenges in streaming video generation, we introduce the **Multi-Instruction Responsive Avatar Benchmark (MIRA-Bench)**. It assesses instruction following, human fidelity, and visual quality under sequential expression and action instructions, with human evaluation explicitly measuring action repetition and stagnation. Our main contributions are summarized as follows:

- We identify **history-state confusion** through attention analysis: streaming generators attend similarly to visual history and instruction tokens across distinct execution states, failing to adequately distinguish action progress and leading to stagnation and repetition.
- We propose **NextAvatar**, which uses multimodal next-state planning to guide action progression and reference-grounded state anchoring to reduce appearance deviations in the predicted state.

- We introduce **MIRA-Bench** to evaluate Human Fidelity, Instruction Responsiveness, Video Quality, and Action Progression. NextAvatar improves responsiveness and reduces stagnation and repetition while maintaining competitive fidelity and quality.

2 Related Work

2.1 Streaming Video Generation

Streaming video generation extends video sequences causally while addressing rollout stability and computational costs. Existing methods mainly advance two complementary directions: causal training and distillation reduce the train–test mismatch of autoregressive rollout (Yin et al., 2025; Huang et al., 2025b; Zhu et al., 2026), while compact history representations, memory management, and efficient inference improve streaming efficiency (Zhang et al., 2025; Kodaira et al., 2026; Liu et al., 2025; Yang et al., 2025b; Yuan et al., 2026; Cai et al., 2025). These advances support long and interactive video generation.

2.2 Avatar Video Generation

Existing avatar generation research has largely focused on real-time generation, identity consistency, and speech control. Real-time generation reduces synthesis latency to support interactive animation (Guo et al., 2024; Zhang et al., 2024; Huang et al., 2025c). Identity consistency relies on reference conditioning and temporal memory to preserve character appearance during extended generation (Lin et al., 2025; Chen et al., 2025; Zhen et al., 2026; Gan et al., 2026). Speech control aligns lip movements with audio to produce lip-synchronized avatar videos (Corona et al., 2025; Tu et al., 2025; Yang et al., 2025a; Shen et al., 2025; Zhang et al., 2026). Beyond these goals, accurate instruction control over facial expressions and body actions is essential for aligning avatar behavior with user intent throughout an interaction. In streaming generation, prior methods organize conditions across chunks: TempAct decomposes a complex prompt into simpler prompts, each spanning a fixed number of chunks, while StreamChar plans speech conditions at the chunk level. NextAvatar instead plans a visual state for each chunk under the same instruction, using the video history to infer action progress. Detailed comparisons are provided in Appendix B.

3 Method

3.1 Overview

Given an initial reference image x_0 and streaming instructions, we generate successive video chunks conditioned on the active instruction p_k and available history $H_{<k}$:

$$y_k \sim p_\theta(y_k \mid x_0, p_k, H_{<k}). \tag{1}$$

An instruction may span multiple chunks, so the appropriate *next state* depends on the action progress recorded in the visual history, namely the *history state*. When attention responds similarly to histories at different stages of execution, this *history-state confusion* can lead to an incorrect next state, causing stagnation before execution or repetition after completion.

NextAvatar addresses history-state confusion by explicitly planning the next state from the visual history and active instruction. Next-State Planning (NSP) uses an MLLM to interpret the history state and predict the intended continuation. Reference-Grounded State Anchoring then refines this prediction using the initial reference to constrain appearance deviations inherited from generated history. A state adapter converts the anchored prediction into generator conditions, which are introduced through state-conditioned attention. Each generated chunk updates the history for subsequent planning. Figure 2 illustrates the overall framework.

3.2 Next-State Planning

Under a persistent instruction, visual-history self-attention and instruction cross-attention can show similar attention allocations across histories with different action progress. This limited sensitivity to action progress is the history-state confusion identified in Figure 1. To address this limitation, we

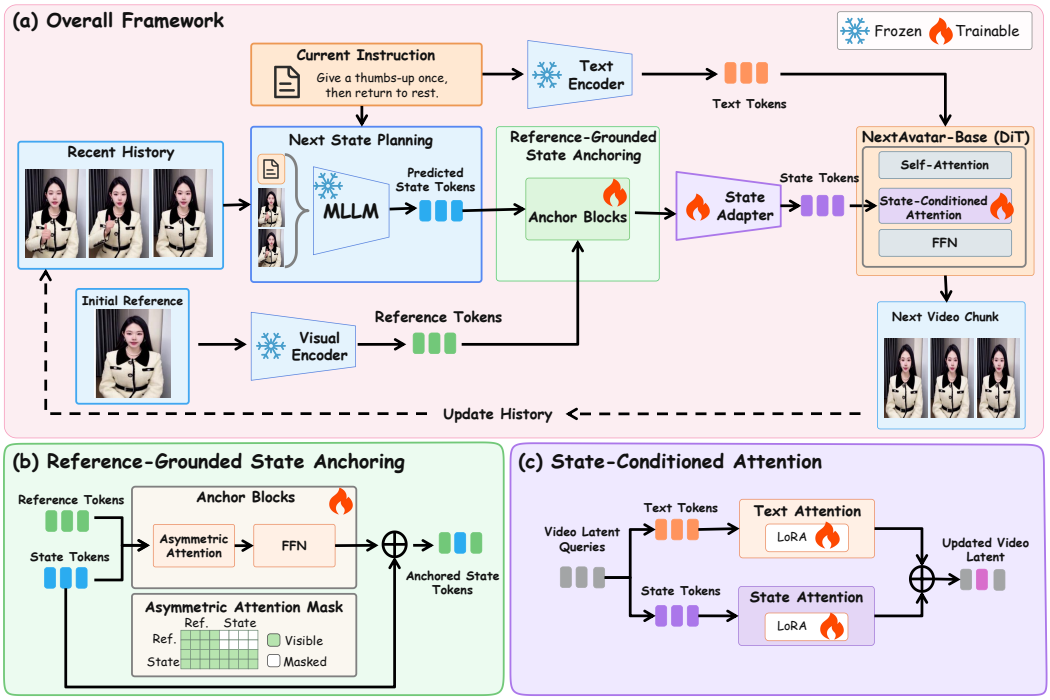

Figure 2: Overview of NextAvatar. (a) NSP predicts the next state from history and instruction; reference-grounded anchoring and a state adapter produce generator conditions. Generated chunks update the history. (b) State tokens attend to both reference and state tokens, while reference tokens attend only to reference tokens. A gated residual produces the anchored state. (c) Separate text and state attention responses are combined to condition generation.

introduce **Next-State Planning (NSP)**. NSP interprets action progress from the visual history under the active instruction and predicts a representation of the appropriate next state for explicit conditioning. We implement NSP with a frozen pretrained multimodal large language model (MLLM) to jointly process visual observations and language semantics.

Let $\widehat{H}_k$ denote the visual history supplied to NSP before generating y_k. Given this history and the active instruction p_k, the frozen MLLM predicts hidden features h_k for the next state. The frozen visual encoder E_v, a Vision Transformer (ViT) (Dosovitskiy et al., 2021), defines the visual token space. A separate frozen pretrained visual prediction module D_v converts the MLLM prediction into this space, yielding the initial next-state tokens s_k:

$$h_k = P_\phi(\widehat{H}_k, p_k), \qquad s_k = D_v(h_k) \in \mathbb{R}^{N_p \times d_p}, \tag{2}$$

where N_p and d_p denote the number and dimension of state tokens. These tokens form a visual plan for the upcoming chunk, specifying how the avatar should proceed.

Because NSP requires visual history, the first chunk is generated directly from x_0 and p_1 in image-to-video mode. From the second chunk onward, we compute the visual plan from generated history and anchor it to x_0 to constrain inherited appearance deviations.

3.3 Reference-Grounded State Anchoring

The predicted state provides a visual plan for action continuation, but its representation in the visual encoder's feature space also carries appearance information. Appearance deviations inherited from generated history can therefore reach the generator through this additional conditioning path and accumulate across chunks. To constrain these deviations before the state conditions generation, we refine it using fixed features from the initial reference. In the asymmetric attention blocks, state tokens attend to both state and reference tokens to retrieve reference appearance, whereas reference tokens attend only to reference tokens. This prevents history-derived state features from altering the reference stream across blocks, keeping the appearance anchor stable.

We use the frozen visual encoder E_v to extract reference tokens $r_0 = E_v(x_0) \in \mathbb{R}^{N_r \times d_p}$, where N_r is the number of reference tokens. They share the feature space of the predicted state s_k and are reused across chunks.

We concatenate reference and predicted state tokens:

$$u_k = [r_0; s_k]. \tag{3}$$

We apply layer normalization to u_k and project the normalized tokens into queries Q, keys K, and values V. Each is partitioned into reference and state groups, indexed by r and s, respectively. Asymmetric attention is then computed as

$$\begin{aligned} A_{rr} &= \mathrm{softmax}\left(\frac{Q_r K_r^\top}{\sqrt{d}}\right), \\ [A_{sr}, A_{ss}] &= \mathrm{softmax}\left(\frac{Q_s [K_r; K_s]^\top}{\sqrt{d}}\right), \\ O &= \begin{bmatrix} A_{rr} & \mathbf{0} \\ A_{sr} & A_{ss} \end{bmatrix} \begin{bmatrix} V_r \\ V_s \end{bmatrix}, \end{aligned} \tag{4}$$

Here d is the attention dimension. Applying the attention residual and the normalized feed-forward update $\mathcal{B}$ yields

$$[\bar{r}_k; \bar{s}_k] = \mathcal{B}(u_k + O W_{\mathrm{a}}^O), \tag{5}$$

where $\mathcal{B}$ includes a residual connection. We stack two such blocks, then retain only the state tokens. The anchored tokens are mapped by the state adapter into conditions for the video generator.

3.4 State-Conditioned Generation

To condition the generator on the anchored visual plan, we first map $\bar{s}_k$ from the visual encoder's feature space into the generator's conditioning space with a token-wise state adapter:

$$c_k = \mathrm{Adapter}(\bar{s}_k), \qquad c_k \in \mathbb{R}^{N_p \times d_c}, \tag{6}$$

where d_c is the conditioning dimension. The adapter consists of two linear projections with a Gaussian error linear unit (GELU) activation and root mean square (RMS) normalization between them.

We then introduce the adapted state tokens alongside the text condition. Because text and visual state tokens serve different roles, we compute their attention responses separately and scale the state response with a learnable weight to limit interference with the generator's existing capabilities. At a representative generator block, video queries Q_k attend to encoded text e_k and adapted state c_k through shared key and value projections:

$$\begin{aligned} r_k^{\mathrm{text}} &= \mathrm{Attn}(Q_k, e_k W^K, e_k W^V), \\ r_k^{\mathrm{state}} &= \mathrm{Attn}(Q_k, c_k W^K, c_k W^V). \end{aligned} \tag{7}$$

Each branch uses its own softmax normalization. Their responses are combined to produce the generator's cross-attention output:

$$o_{k,\mathrm{g}} = \left(r_k^{\mathrm{text}} + \alpha r_k^{\mathrm{state}}\right) W_{\mathrm{g}}^O, \qquad \alpha = \sigma(\gamma), \tag{8}$$

where γ is a block-specific learned scalar, σ is the sigmoid function, and W_{g}^O is the generator's output projection.

The added cross-attention branch injects the planned next state while retaining the generator's original text, reference, and history inputs. Self-attention continues to use reference and history context for visual continuity, and each generated chunk updates the history for subsequent planning.

3.5 Training Strategy

We train NextAvatar in two stages. The first establishes a history-conditioned generator, NextAvatar-Base; the second adds next-state planning while retaining the flow-matching objective. Training data and implementation details for both stages are provided in Appendix C.

Base training. We train NextAvatar-Base on digital-human videos with temporally aligned expression and action descriptions. Low-rank adaptation (LoRA) modules in the generator's linear layers are optimized together with the multiscale history patch embeddings, while the backbone weights remain frozen. Training uses a flow-matching objective. Let x_k denote the clean video latent of chunk y_k, $\epsilon \sim \mathcal{N}(0, I)$ the noise, and $x_{k,t} = (1-t)x_k + t\epsilon$ the interpolated latent at $t \in [0, 1]$. The generator predicts the velocity $\epsilon - x_k$ under conditioning $\mathcal{C}_k$:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{x_k, \epsilon, t, \mathcal{C}_k} \left[\| v_\theta(x_{k,t}, t; \mathcal{C}_k) - (\epsilon - x_k) \|_2^2 \right]. \tag{9}$$

For base training, $\mathcal{C}_k = (x_0, p_k, H_{<k})$ contains the reference, instruction, and multiscale history.

Joint training. Starting from NextAvatar-Base, we retain the flow-matching objective and augment the conditioning with adapted state tokens: $\mathcal{C}_k = (x_0, p_k, H_{<k}, c_k)$. We jointly optimize state anchoring, the state adapter, conditioning gates, and the generator's cross-attention LoRA parameters. All other base parameters, the MLLM, and its pretrained visual prediction module remain frozen. This stage learns to ground the predicted states in the reference and use them to guide generation.

4 Experiments

We first introduce the benchmark and experimental setup, then present quantitative and qualitative comparisons, component ablations, and an analysis of identity drift and state anchoring. Additional training and inference details, benchmark construction and evaluation protocols, baseline configurations and adaptations, and qualitative results are provided in Appendices C, D, and E.

4.1 Experimental Setup

MIRA-Bench. Streaming avatar research has emphasized real-time synthesis, identity preservation, and speech-driven lip synchronization, leaving instruction-driven action progression across chunks less thoroughly evaluated. We introduce the **Multi-Instruction Responsive Avatar Benchmark (MIRA-Bench)** to assess instruction following, human fidelity, and visual quality under sequential expression and action instructions. It pairs 50 reference images with 200 ten-instruction sequences, yielding 2,000 expression–action intervals with balanced clear and subtle categories.

We measure human fidelity with Human Anatomy, Human Identity, and Human Clothes from VBench-2.0 (Zheng et al., 2025), measuring identity against the input reference; instruction responsiveness with Instruct Follow from Vision Reward (Xu et al., 2026); and video quality with Vision Reward, Aesthetic Quality, and Imaging Quality (Xu et al., 2026; Huang et al., 2024; 2026). We assess action progression through human annotations of repetition and stagnation. From each of the 200 videos, we randomly select one of its ten intervals and evaluate the corresponding interval for every method, yielding 200 annotated intervals per method. Appendix D details benchmark construction and evaluation protocols.

Baselines. We compare against six external systems: **LongCat-Video-Avatar 1.5** (**LongCat-Avatar 1.5** for short), **SoulX-FlashTalk**, **SoulX-LiveAct**, **Live Avatar**, **TaoMate**, and **AptAvatar**. All methods use the same benchmark inputs and a nominal 60-second generation budget at 24 FPS. Controlled variants share generation settings and, where applicable, training budgets. Baseline configurations are provided in Appendices E.

4.2 Overall Performance

Quantitative comparison. Table 1 compares streaming avatar methods across human fidelity, instruction responsiveness, and video quality. Live Avatar achieves relatively low human fidelity and instruction following scores, while its video quality scores remain competitive. LongCat-Avatar 1.5, SoulX-FlashTalk, and AptAvatar improve human fidelity and achieve strong video quality, but their instruction following scores remain below 0.58. SoulX-LiveAct and TaoMate further raise instruction following to 0.6395 and 0.6245, respectively, yet both still show substantial repetition and stagnation when instructions span multiple chunks. With a 1.3B-parameter generator, NextAvatar achieves the highest instruction following and Vision Reward scores while maintaining competitive aesthetic and imaging quality, with slightly lower identity fidelity than larger baselines.

Table 1: MIRA-Bench comparison. Params denotes backbone parameter counts. Best results are **bold**, and second-best results are underlined.

Method	Params	Human Anatomy ↑	Human Identity ↑	Human Clothes ↑	Instruct Follow ↑	Vision Reward ↑	Aesthetic Quality ↑	Imaging Quality ↑
LongCat-Avatar 1.5	13.6B	0.9680	0.9441	**1.0000**	0.5710	0.9370	0.5429	**0.7479**
SoulX-FlashTalk	14B	0.9602	0.9707	**1.0000**	0.4445	0.9251	**0.5797**	0.7296
SoulX-LiveAct	14B	**0.9691**	0.9650	**1.0000**	0.6395	0.9219	0.5472	0.7425
Live Avatar	14B	0.9129	0.9542	**1.0000**	0.5690	0.9214	0.5533	0.7356
TaoMate	22B	0.9519	0.8102	0.9648	0.6245	0.9281	0.5518	0.7320
AptAvatar	14B	0.9686	**0.9725**	**1.0000**	0.5405	0.9238	0.5619	0.7400
NextAvatar	1.3B	0.9338	0.9323	**1.0000**	**0.6495**	**0.9397**	0.5700	0.7427



(a) Action: The person tilts the head slightly to the right, holds for a moment, then returns to the center.

(b) Action: The person raises one hand and gives a brief salute at the forehead, then lowers the hand and returns to a natural posture.

Figure 3: Qualitative examples on MIRA-Bench illustrating (a) stagnation under a head-tilting instruction and (b) repetition under a brief-salute instruction. NextAvatar demonstrates the intended action progression in both examples.

Qualitative comparison. Fig. 3 illustrates stagnation and repetition under two action instructions. In the left example, the person is instructed to tilt the head slightly to the right and then return to the center. The compared methods show almost no movement, whereas NextAvatar completes the tilt-and-return action. In the right example, the instruction requests a brief salute followed by a return

Table 3: Component ablations. Best results are **bold**, and second-best results are underlined.

Variant	Human Anatomy ↑	Human Identity ↑	Human Clothes ↑	Instruct Follow ↑	Vision Reward ↑	Aesthetic Quality ↑	Imaging Quality ↑
NextAvatar-Base	0.9321	**0.9388**	**1.0000**	0.6035	0.9305	0.5634	0.7399
NextAvatar-Base + NSP (w/o R)	**0.9338**	0.9118	**1.0000**	0.6494	0.9302	0.5623	0.7424
NextAvatar-Base + NSP (w/ R)	0.9292	0.9126	**1.0000**	0.6480	0.9329	0.5614	0.7423
NextAvatar	0.9338	0.9323	**1.0000**	**0.6495**	**0.9397**	**0.5700**	**0.7427**

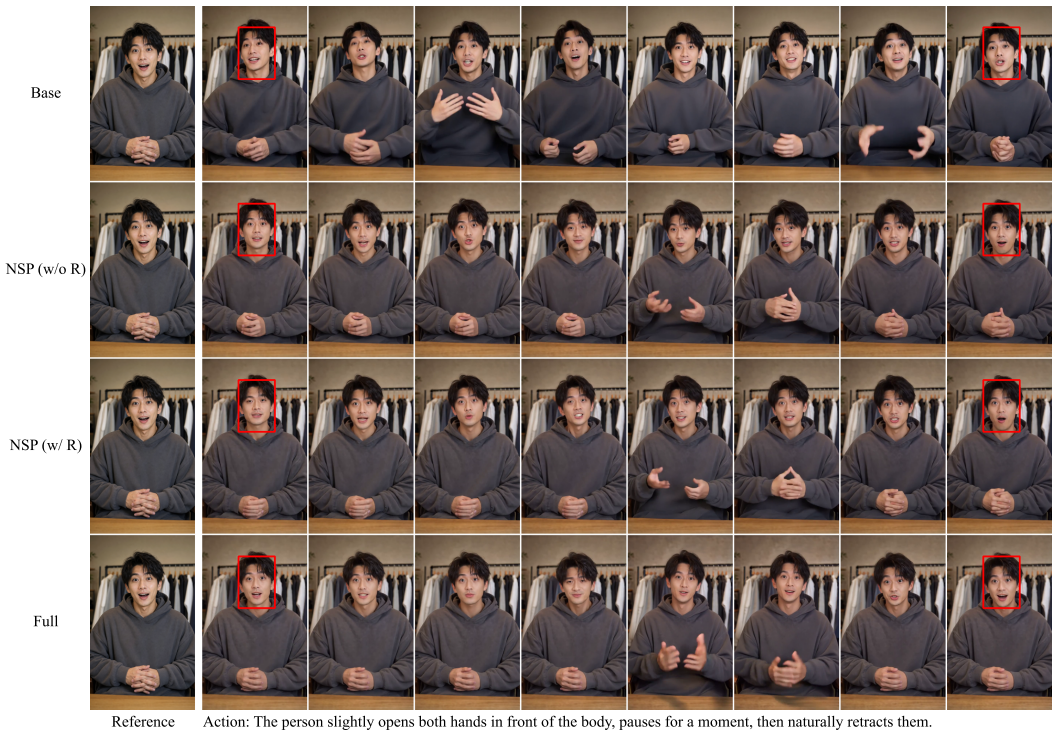

Figure 4: Qualitative ablation example on MIRA-Bench. Base, NSP (w/o R), NSP (w/ R), and Full correspond to NextAvatar-Base, NextAvatar-Base + NSP (w/o R), NextAvatar-Base + NSP (w/ R), and NextAvatar, respectively. NSP prevents repeated execution of the hand action, while reference-grounded state anchoring improves identity consistency with the reference.

to a natural posture. All compared methods repeat the salute, whereas NextAvatar performs it once and returns to a resting posture.

Human evaluation of action progression. As shown in Table 2, NextAvatar has the lowest combined error rate, substantially improving over TaoMate, the strongest external baseline on this measure. It ties for the lowest stagnation rate and achieves the second-lowest repetition rate. SoulX-FlashTalk has less repetition but frequently stagnates, whereas Live Avatar rarely stagnates but often repeats actions. NextAvatar performs well across both failure modes, supporting more reliable action progression under the given instructions.

Table 2: Human evaluation on MIRA-Bench. Total Error is the sum of stagnation and repetition rates. Best results are **bold**; second-best are underlined.

Method	Total Error ↓	Stagnation ↓	Repetition ↓
LongCat-Avatar 1.5	54.5%	28.0%	26.5%
SoulX-FlashTalk	78.0%	71.0%	**7.0%**
SoulX-LiveAct	76.0%	27.5%	48.5%
Live Avatar	59.0%	**6.0%**	53.0%
TaoMate	52.5%	24.0%	28.5%
AptAvatar	77.0%	30.0%	47.0%
NextAvatar	**30.5%**	**6.0%**	24.5%

4.3 Ablation Studies

We compare three variants to examine next-state planning and reference-grounded anchoring. **NextAvatar-Base** uses no predicted state conditioning. **NextAvatar-Base + NSP (w/o R)** adds history-conditioned next-state planning without reference input to the MLLM. **NextAvatar-Base + NSP (w/ R)** additionally provides the reference image to the MLLM. Both NSP variants omit reference-grounded state anchoring.

Quantitative analysis. Table 3 examines the contributions of next-state planning and reference-grounded state anchoring. Introducing NSP improves Instruct Follow from 0.6035 to 0.6494, showing the benefit of explicit next-state guidance. At the same time, Human Identity drops from 0.9388 to 0.9118, suggesting that predicted states can carry appearance deviations from generated history into subsequent chunks. Providing the reference image to the frozen MLLM yields Human Identity of 0.9126 and Instruct Follow of 0.6480, suggesting that reference input alone does not prevent identity deviations from propagating through successive state predictions. Reference-grounded state anchoring recovers Human Identity to 0.9323 while retaining the instruction-following gain, supporting its role in limiting appearance drift.

Qualitative analysis. Fig. 4 compares four methods under an instruction to slightly open both hands, pause, and naturally retract them. NextAvatar-Base opens its hands more than once instead of completing a single action sequence. Both NSP variants avoid this repetition, but show visible facial deviations from the reference. Providing the reference image to the frozen MLLM does not resolve these deviations. With reference-grounded state anchoring, NextAvatar retains the single action sequence while better preserving the reference identity. This example illustrates the complementary roles of NSP in action progression and state anchoring in identity preservation.

Identity drift and reference-grounded anchoring. To further examine the identity differences observed in the quantitative and qualitative ablations, Fig. 5 tracks face-region feature similarity to the input reference in the MLLM visual encoder space. We compare the history state before NSP, the predicted state, and the anchored state. As generation proceeds, reference similarity decreases and then fluctuates around a lower level, indicating that appearance drift accumulates before tending to stabilize. The predicted state consistently exhibits lower similarity than the history state, suggesting that NSP can amplify appearance deviations already present in its visual context. Reference-grounded state anchoring increases similarity relative to the predicted state across the displayed sequence. These results further support the role of state anchoring in mitigating the additional appearance drift introduced by next-state planning.

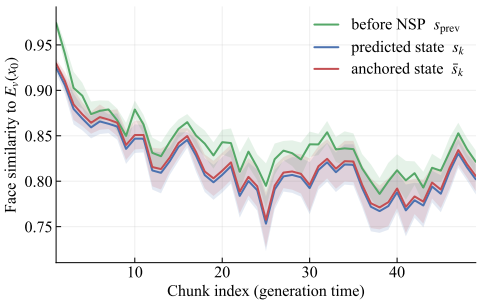

Figure 5: Face-region similarity to the reference in the MLLM visual encoder space, before NSP, after NSP, and after state anchoring.

5 Conclusion

In this paper, we presented NextAvatar, a multimodal next-state planning framework for streaming avatar generation. When an instruction spans multiple chunks, coherent action progression requires understanding what the avatar has already done. However, our analysis identifies history-state confusion: attention to visual history and text instructions changes little across distinct execution states, leading to stagnation and repetition. NextAvatar interprets the history state and active instruction to provide an explicit next-state visual plan, while reference-grounded state anchoring reduces appearance deviations in the plan. Experiments on MIRA-Bench demonstrate improved instruction following and action progression, with the lowest combined stagnation and repetition rate among the evaluated methods. Ablations further show that state anchoring recovers identity fidelity while retaining the instruction-following gain from planning. However, a limited history window can still miss earlier action progress when an instruction spans many chunks, motivating longer-range history modeling for next-state planning.

AI USE STATEMENT

Generative AI tools assisted with code implementation and debugging, experimental planning, data annotation, literature discovery, and manuscript drafting and editing. The authors developed the central research idea, problem formulation, and overall method. AI-generated data were used only for evaluation; all training data were collected from real-world scenes. The authors reviewed, revised, and tested AI-assisted code, verified annotations and synthetic evaluation data, and manually selected and checked all references. They reviewed the entire manuscript and associated artifacts and take full responsibility for the paper's content and results.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 3 describes the NextAvatar architecture, conditioning mechanisms, and training objective. Appendix C provides details of the training data and annotation, base architecture, streaming protocol, and training and inference configurations. Appendix D describes the construction of MIRA-Bench and its evaluation protocols, while Appendix E specifies the baseline conditioning interfaces and generation and evaluation procedures.

REFERENCES

Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report. *CoRR*, abs/2510.22200, 2025. doi: 10.48550/ARXIV.2510.22200. URL https://doi.org/10.48550/arXiv.2510.22200.

Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, Xiaoming Wei, Tianyu Yang, and Yong Zhang. Longcat-video-avatar 1.5 technical report. *CoRR*, abs/2605.26486, 2026. doi: 10.48550/ARXIV.2605.26486. URL https://doi.org/10.48550/arXiv.2605.26486.

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. *CoRR*, abs/2505.20156, 2025. doi: 10.48550/ARXIV.2505.20156. URL https://doi.org/10.48550/arXiv.2505.20156.

Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. VLOGGER: multimodal diffusion for embodied avatar synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 15896–15908. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01482. URL https://openaccess.thecvf.com/content/CVPR2025/html/Corona_VLOGGER_Multimodal_Diffusion_for_Embodied_Avatar_Synthesis_CVPR_2025_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Qijun Gan, Chenwei Zhang, Meiguang Jin, Junfeng Ma, and Qiu Shen. Taomate: Anchor-guided memory bridging evolving and reference states for real-time audio-video digital human generation. *arXiv preprint arXiv:2607.24359*, 2026.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *CoRR*, abs/2407.03168, 2024. doi: 10.48550/ARXIV.2407.03168. URL https://doi.org/10.48550/arXiv.2407.03168.

Lun Huang, You Xie, Hongyi Xu, Tianpei Gu, Chenxu Zhang, Guoxian Song, Zenan Li, Xiaochen Zhao, Linjie Luo, and Guillermo Sapiro. Plan-x: Instruct video generation via semantic planning. *CoRR*, abs/2511.17986, 2025a. doi: 10.48550/ARXIV.2511.17986. URL https://doi.org/10.48550/arXiv.2511.17986.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *CoRR*, abs/2512.04677, 2025c. doi: 10.48550/ARXIV.2512.04677. URL https://doi.org/10.48550/arXiv.2512.04677.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 21807–21818. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 29200–29210. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kodaira_StreamDiT_Real-Time_Streaming_Text-to-Video_Generation_CVPR_2026_paper.html.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 13848–13858. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01285. URL https://doi.org/10.1109/ICCV51701.2025.01285.

Chenchen Liu, Junyi Chen, Lei Li, Lu Chi, Mingzhen Sun, Zhuoying Li, Yi Fu, Ruoyu Guo, Yiheng Wu, Ge Bai, and Zehuan Yuan. Bernini: Latent semantic planning for video diffusion. *CoRR*, abs/2605.22344, 2026. doi: 10.48550/ARXIV.2605.22344. URL https://doi.org/10.48550/arXiv.2605.22344.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *CoRR*, abs/2509.25161, 2025. doi: 10.48550/ARXIV.2509.25161. URL https://doi.org/10.48550/arXiv.2509.25161.

Le Shen, Qiao Qian, Tan Yu, Ke Zhou, Tianhang Yu, Yu Zhan, Zhenjie Wang, Ming Tao, Shunshun Yin, and Siyuan Liu. Soulx-flashtalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation. *CoRR*, abs/2512.23379, 2025. doi: 10.48550/ARXIV.2512.23379. URL https://doi.org/10.48550/arXiv.2512.23379.

Linrui Tian, Qi Wang, and Bang Zhang. Streamchar: Long-horizon streaming character audio-video generation with decoupled orchestration. *CoRR*, abs/2605.25659, 2026. doi: 10.48550/ARXIV.2605.25659. URL https://doi.org/10.48550/arXiv.2605.25659.

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. *CoRR*, abs/2508.08248, 2025. doi: 10.48550/ARXIV.2508.08248. URL https://doi.org/10.48550/arXiv.2508.08248.

Jing Wang, Xiangxin Zhou, Jiajun Liang, Kaiqi Liu, Wanyun Pang, Zhenyu Xie, Tianyu Pang, and Xiaodan Liang. Tempact: Advancing temporal plausibility in autoregressive video generation via planner-executor RL. *CoRR*, abs/2606.28016, 2026. doi: 10.48550/ARXIV.2606.28016. URL https://doi.org/10.48550/arXiv.2606.28016.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *CoRR*, abs/2510.08377, 2025. doi: 10.48550/ARXIV.2510.08377. URL https://doi.org/10.48550/arXiv.2510.08377.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 11269–11277. AAAI Press, 2026. doi: 10.1609/AAAI.V40I13.38107. URL https://doi.org/10.1609/aaai.v40i13.38107.

Hao Yang, Zhiyu Tan, Jia Gong, Luozheng Qin, Hesen Chen, Xiaomeng Yang, Yuqing Sun, Yuetan Lin, Mengping Yang, and Hao Li. Omni-video 2: Scaling mllm-conditioned diffusion for unified video generation and editing. *CoRR*, abs/2602.08820, 2026. doi: 10.48550/ARXIV.2602.08820. URL https://doi.org/10.48550/arXiv.2602.08820.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *CoRR*, abs/2508.14033, 2025a. doi: 10.48550/ARXIV.2508.14033. URL https://doi.org/10.48550/arXiv.2508.14033.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *CoRR*, abs/2509.22622, 2025b. doi: 10.48550/ARXIV.2509.22622. URL https://doi.org/10.48550/arXiv.2509.22622.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 22963–22974. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02138. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *CoRR*, abs/2603.04379, 2026. doi: 10.48550/ARXIV.2603.04379. URL https://doi.org/10.48550/arXiv.2603.04379.

Hengyuan Zhang, Jingna Sun, Meiguang Jin, and Junfeng Ma. Aptavatar: Fast and vivid long-form audio-driven video generation for production-ready avatars. *CoRR*, abs/2607.24013, 2026. doi: 10.48550/ARXIV.2607.24013. URL https://doi.org/10.48550/arXiv.2607.24013.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/2bde8fef08f7ebe42b584266cbcfc909-Abstract-Conference.html.

Yue Zhang, Zhizhou Zhong, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high-fidelity video dubbing via spatio-temporal sampling. *arXiv preprint arXiv:2410.10122*, 2024.

Dingcheng Zhen, Xu Zheng, Ruixin Zhang, Zhiqi Jiang, Yichao Yan, Ming Tao, and Shunshun Yin. Soulx-liveact: Towards hour-scale real-time human animation with neighbor forcing and convkv memory. *CoRR*, abs/2603.11746, 2026. doi: 10.48550/ARXIV.2603.11746. URL https://doi.org/10.48550/arXiv.2603.11746.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *CoRR*, abs/2503.21755, 2025. doi: 10.48550/ARXIV.2503.21755. URL https://doi.org/10.48550/arXiv.2503.21755.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *CoRR*, abs/2602.02214, 2026. doi: 10.48550/ARXIV.2602.02214. URL https://doi.org/10.48550/arXiv.2602.02214.

A Appendix Overview

This appendix provides supplementary comparisons, implementation details, evaluation protocols, and qualitative results for NextAvatar, organized as follows:

- Section B compares NextAvatar with closely related streaming generation methods, TempAct and StreamChar, focusing on planning targets, granularity, and information flow.
- Section C details the training data, base architecture, streaming protocol, and training and inference settings of NextAvatar-Base and NextAvatar.
- Section D describes MIRA-Bench construction and evaluation protocols for human fidelity, instruction responsiveness, video quality, and action progression.
- Section E introduces the compared baselines and specifies their conditioning adaptations and shared generation and evaluation protocol.
- Section F discusses the limitations of our method and potential directions for future work.

B Comparison of Planning in Streaming Generation

TempAct (Wang et al., 2026) and StreamChar (Tian et al., 2026) are closely related to NextAvatar in their use of planning or orchestration to guide streaming generation. They operate in different settings: TempAct addresses general-purpose video generation, whereas StreamChar targets joint audio-video generation of speaking characters. NextAvatar focuses on instruction-controlled streaming avatar generation, where an action may unfold across multiple chunks under an unchanged instruction. To clarify these connections and differences, Figure A1 compares what each method plans and how information flows into subsequent generation: textual subgoals in TempAct, speech progress in StreamChar, and history-conditioned visual states in NextAvatar.

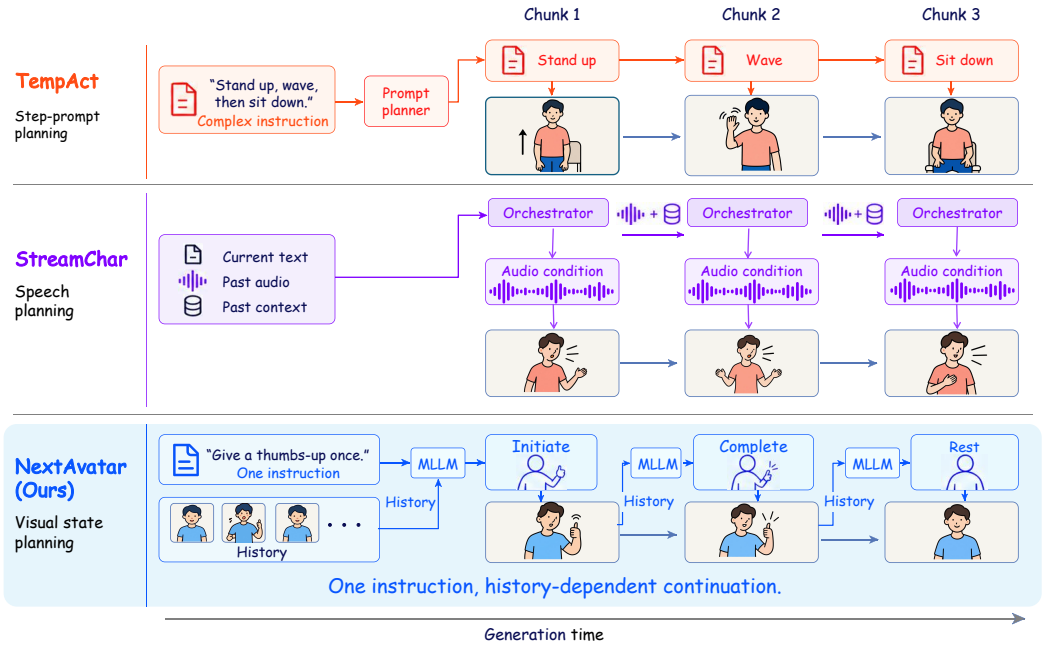

Figure A1: Conceptual comparison of planning in streaming generation across different task settings. Top: TempAct (Wang et al., 2026) assigns textual subgoals to temporal spans for general video generation. Middle: StreamChar (Tian et al., 2026) orchestrates speech to provide audio conditions for joint audio-video generation. Bottom: NextAvatar plans the next visual state from generated history and an unchanged action instruction. The illustrated stages summarize intended behavior, not measured outputs; chunk boundaries are schematic.

TempAct: planning textual subgoals. TempAct (Wang et al., 2026) converts a global instruction into step prompts with temporal spans and jointly optimizes its prompt planner and autoregressive executor through reinforcement learning. The top row illustrates the decomposition of "Stand up, wave, then sit down" into subgoals assigned to successive intervals. The executor uses generated visual history and smoothed adjacent step prompts to improve transitions. Its planning output thus specifies which textual subgoal is active over each span. The key distinction lies in the granularity of planning. TempAct plans at the prompt level, decomposing a multi-action instruction into step prompts and assigning them to temporal spans. NextAvatar plans at the chunk level within a single action: while the action instruction remains unchanged across multiple chunks, it uses generated history to predict the next visual state, guiding the action toward completion.

StreamChar: orchestrating speech. StreamChar (Tian et al., 2026) uses current text, past audio, and historical context to produce frame-aligned audio conditions for a local joint audio-video generator. The middle row illustrates the orchestrator advancing through the text as new audio and context become available, while the local generator synthesizes the corresponding audio and character motion. Its orchestration coordinates speech progress across chunks. NextAvatar instead predicts a visual state that guides the continuation of an instructed action.

NextAvatar: planning the next visual state. The bottom row keeps "Give a thumbs-up once" active across chunks, illustrating the need to update visual guidance even when the instruction remains unchanged. From recent generated history and the current instruction, the MLLM predicts a continuous visual representation for the upcoming chunk. This allows the planned continuation to reflect action execution progress, guiding initiation, completion, and subsequent rest. "Initiate," "Complete," and "Rest" describe the intended visual states; they are not additional prompts or discrete stage labels supplied to the model, and need not align with chunk boundaries. The key distinction is therefore the planning target: NextAvatar makes the next visual state explicit to sup-

Table A1: Phase-specific training configurations.

Setting	Base: low resolution	Base: high resolution	Joint training
Resolution	384×640	576×960	576×960
Initialization	Pretrained Wan	Low-resolution base	NextAvatar-Base
Training steps	25,000	10,000	5,000
Effective batch size	224	16	8
LoRA learning rate	3×10^{-5}	10^{-5}	10^{-5}
State adapter learning rate	–	–	10^{-5}
Anchoring learning rate	–	–	10^{-4}
Gate learning rate	–	–	10^{-2}
LoRA rank	128	128	128

port history-dependent action continuation, complementing the textual subgoal planning and speech orchestration illustrated above.

C NextAvatar-Base and Training Details

C.1 Training Data and Instruction Annotation

We train NextAvatar-Base on Chinese talking-avatar and livestream videos collected from multiple internal sources. Each autoregressive chunk is associated with a Chinese caption describing the scene, subject appearance, facial expression, and body action. A caption may be shared by consecutive chunks within the same annotated segment. Captions are aligned with their corresponding temporal segments and change within a video.

Base adaptation first uses videos at 384×640 resolution and then continues on 576×960 subset. The low- and high-resolution datasets contain approximately 4.56M and 1.06M training chunks, respectively, with 33 video frames per chunk. The high-resolution dataset is reused for joint training.

C.2 Base Architecture and Streaming Protocol

NextAvatar-Base follows the architecture of Helios-Base (Yuan et al., 2026), built on the pretrained Wan2.1-T2V-1.3B diffusion transformer with multi-scale history conditioning and a reference-frame anchor. The video VAE and umT5 text encoder remain frozen. Reference and history tokens interact with current-chunk tokens through self-attention, while text cross-attention conditions only the current chunk.

Each chunk contains 9 latent frames and produces 33 video frames. A rolling buffer retains the most recent 19 history latent frames, grouped into 16, 2, and 1 frames from oldest to newest and encoded with stronger compression for older context. The reference latent is included in the short-history branch. During training, histories come from ground-truth videos, and the first frame of the first chunk serves as the reference x_0. During I2V inference, x_0 is the input image, and the buffer is updated after each generated chunk. The first chunk is generated without preceding video history.

This generator-level reference conditioning is distinct from NextAvatar's reference-grounded state anchoring, which refines predicted state tokens before they condition the generator.

C.3 Training Stages and Optimization

Base training. The architecture and initial adaptation settings of NextAvatar-Base follow Helios-Base (Yuan et al., 2026). We train rank-128 LoRA modules ($\alpha = 128$) in the generator's linear layers together with the multi-scale history patch embeddings, while freezing the backbone, video VAE, and text encoder. Training uses flow matching with conditioning dropout to support text-only, reference-conditioned, and history-conditioned generation. Adaptation comprises $25{,}000$ steps at 384×640 followed by $10{,}000$ steps at 576×960, yielding NextAvatar-Base.

Table A2: MIRA-Bench composition. Clear/subtle categories are assigned independently to expression and action.

Property	Specification
Reference images	50
Reference–instruction assignments	200
Instructions per assignment	10
Total instruction intervals	2,000
Expression–action combinations	Clear–clear, clear–subtle, subtle–clear, subtle–subtle
Intervals per combination	500
Balance within each assignment	5 clear and 5 subtle per factor
Action types	16 clear and 16 subtle
Expression types	12 clear and 12 subtle
Type frequencies	Approximately balanced

Joint training. We initialize the generator from NextAvatar-Base and the next-state planner from Bernini's pretrained Qwen2.5-VL-based model (Liu et al., 2026). The planner remains frozen and predicts visual state tokens from the preceding history chunk and current instruction. The first chunk is generated directly by NextAvatar-Base without planner conditioning. Predicted states and reference features share the planner's visual feature space.

Joint training optimizes state anchoring, the state adapter, conditioning gates, and rank-128 cross-attention LoRA in all 30 generator blocks. All remaining parameters are frozen. The anchoring module is a two-layer transformer with hidden dimension 1536, 12 attention heads, and an MLP expansion ratio of 4. The adapter maps visual state features into the generator's conditioning space. Text and state attention share projections but use separate attention normalization. We retain the base flow-matching objective and conditioning-dropout protocol.

Optimization and inference. All training phases use eight NVIDIA A800 GPUs, AdamW, and `bf16` precision, with phase-specific configurations summarized in Table A1. Learning rates remain constant after warmup, where used. High-resolution adaptation and joint training use 500 warmup steps and gradient clipping at 1.0. Inference uses 50 sampling steps, a classifier-free guidance scale of 5.0, and an output frame rate of 24 fps.

D MIRA-Bench Construction and Evaluation Details

D.1 Benchmark Construction

MIRA-Bench evaluates instruction-controlled streaming avatar generation under sequential expression and action instructions. Each benchmark assignment pairs a fixed reference image with ten ordered instructions, each specifying both facial expression and body action. These two factors are independently categorized as clear or subtle, enabling evaluation across different levels of visible change. Table A2 summarizes the benchmark composition.

Clear instructions specify readily observable changes, whereas subtle instructions describe smaller changes such as blinking, gaze shifts, and slight nods. Object interactions and lower-body actions are excluded. Candidate instructions were generated by an LLM in the style of the training descriptions and reviewed by humans for plausibility, duplication, duration feasibility, and conflicting requirements. To reduce initialization ambiguity, clear actions are not paired with reference images already depicting an active pose of the same class. Figure A2 illustrates one assignment. Instructions describe local expression and action changes and, where applicable, a return toward a natural pose before the next interval.

D.2 Human Fidelity

We assess human fidelity using Human Anatomy, Human Identity, and Human Clothes from VBench-2.0 (Zheng et al., 2025), measuring anatomical plausibility, identity retention, and clothing



01	Expression: Clear Action: Subtle	**Expression:** An encouraging smile, with raised mouth corners, relaxed brows and eyes, and a friendly gaze. **Action:** Tilt the head slightly to the right, pause briefly, then return to center.
02	Expression: Subtle Action: Clear	**Expression:** A natural, neutral expression, with relaxed brows, eyes, and lips. **Action:** Raise one hand to give a single clear thumbs-up, hold briefly, then lower it and return to a natural pose.
03	Expression: Clear Action: Clear	**Expression:** A visibly confused expression, with one eyebrow raised, focused eyes, and slightly tightened mouth corners. **Action:** Raise one hand and clearly point toward the right side of the image, hold briefly, then retract it and return to a natural pose.
04	Expression: Subtle Action: Subtle	**Expression:** Brief, mild surprise, with slightly raised eyebrows and slightly widened eyes. **Action:** Lift the chin slightly, pause briefly, then return to its natural height.
05	Expression: Clear Action: Subtle	**Expression:** A clearly surprised expression, with raised eyebrows, widened eyes, and a slightly open mouth. **Action:** Lift the right shoulder slightly, pause briefly, then lower it naturally.
06	Expression: Subtle Action: Clear	**Expression:** A relaxed face, with slightly lowered eyelids and a softened gaze. **Action:** Raise one hand with an open palm clearly facing the camera to signal stop, then lower it.
07	Expression: Clear Action: Clear	**Expression:** A clearly skeptical expression, with one eyebrow noticeably raised, slightly narrowed eyes, and tightened mouth corners. **Action:** Clearly raise one hand as a signal, hold briefly, then lower it and return to a natural pose.
08	Expression: Subtle Action: Subtle	**Expression:** A natural expression, with slightly raised eyebrows and relaxed eyes. **Action:** Lift the left shoulder slightly, pause briefly, then lower it naturally.
09	Expression: Clear Action: Subtle	**Expression:** A clearly serious expression, with slightly lowered brows, level mouth corners, and a steady gaze. **Action:** Make one small emphasis gesture with one hand in front of the chest, then withdraw it naturally.
10	Expression: Subtle Action: Clear	**Expression:** Slightly downturned mouth corners and subtly raised inner eyebrows; the change is restrained but visible. **Action:** Raise both shoulders and spread both hands in a single clear shrug, then relax the shoulders and retract the hands.

Figure A2: Example MIRA-Bench assignment: a fixed reference image paired with ten sequential instructions, each specifying expression and action with independent clear/subtle tags.

consistency, respectively. Human Identity is adapted to use the input reference image as the identity anchor.

D.3 Instruction Responsiveness

For evaluation, each output is divided into ten equal-frame intervals and matched to the instructions in temporal order. We measure Instruct Follow using the text-alignment question of Vision Reward (Xu et al., 2026) and report the proportion of intervals judged compliant with their assigned instructions. This provides an interval-level measure of instruction following.

D.4 Video Quality

We report the overall Vision Reward (Xu et al., 2026) score using 24 uniformly sampled frames and the concatenated instruction sequence, with scores normalized to $[0, 1]$. Aesthetic Quality and Imaging Quality are evaluated using VBench-Long (Huang et al., 2024; 2026) on two-second clips and aggregated at the video level.

D.5 Action Progression

To assess action progression, we randomly select one of the ten equal-frame intervals from each of the 200 videos. The selected interval is the same across methods, yielding 200 evaluated intervals per method. Human annotators review each interval together with its preceding video history and assigned instruction.

E Baseline Protocols

We compare NextAvatar with NextAvatar-Base (Appendix C), our streaming avatar generator built on the Helios architecture (Yuan et al., 2026), and six external avatar systems: LongCat-Video-Avatar 1.5, SoulX-FlashTalk, SoulX-LiveAct, Live Avatar, TaoMate, and AptAvatar. All methods use the same 50 reference images and 200 ten-instruction sequences, with a nominal generation budget of 60 seconds at 24 FPS.



Figure A3: **Repeated actions under long-running instructions.** Each row shows sampled frames from a generated video under one instruction. Although the requested gesture has been performed, the avatar executes a similar gesture again in later chunks. This illustrates a remaining limitation when action progress extends beyond the available history context.

Conditioning interfaces. We use each method's released or documented inference interface. Methods supporting reference images and language instructions receive both inputs directly. For systems requiring audio, we supply silent audio matching the target duration, as MIRA-Bench contains no speech input. Runs that require interface adaptation are explicitly marked.

Generation and evaluation. We retain each method's native chunking and sampling configuration, so actual output durations may differ from the nominal budget. Each output is evaluated over ten ordered intervals following Section D.3. Method-specific adaptations, duration mismatches, and generation failures are recorded to document deviations from the shared protocol.

F Limitations and Future Work

As shown in Fig. A3, NextAvatar can repeat an action after completing it when the same instruction spans many chunks. One possible reason is that the finite history provided to the MLLM does not capture earlier action progress, making it difficult to determine whether the action has already been executed. Future work could provide a longer history context, potentially through compact state memory, to enable a more complete understanding of the history state and more precise next-state planning.