

StreamHOI: Interaction-aware Temporal Memory Adaptation for Streaming HOI Video Generation

Zejing Rao, Haoxian Zhang, Xiaoqiang Liu, Yiping Meng, Guoxin Zhang, Pengfei Wan, Fan Tang, Tong-Yee Lee, *Senior Member, IEEE*

Abstract—Existing human–object interaction (HOI) video generation methods are largely limited to offline short-video generation with complex driving conditions, making them unsuitable for real-time interactive applications. We present *StreamHOI*, a low-latency streaming framework for long-duration HOI video generation. Instead of converting heavily conditioned HOI pipelines into streaming systems, we study how an image-to-video streaming generator should organize historical memory to preserve interactions under bounded latency. We find that the standard sink-local memory design faces a trade-off in streaming HOI generation, and different transformer blocks show different historical-memory preferences for HOI regions and surrounding regions. To match memory composition with block behavior, *StreamHOI* performs offline HOI-aware block profiling and applies bias-guided memory-specialized training to adapt the generator to block-specific memory layouts. We further introduce a memory distance scaling module to strengthen long-range access to early interaction states. Extensive comparisons with both long-video baselines and recent HOI generation methods demonstrate that *StreamHOI* achieves strong interaction plausibility, object fidelity, human quality and efficiency, reaching 17.6 FPS with 0.75s first-chunk latency.

Index Terms—Human–object interaction, streaming video generation, long video generation, diffusion transformers, memory.

1 INTRODUCTION

HUMAN–OBJECT INTERACTION (HOI) has received increasing attention in computer vision and graphics [1]–[3], with applications in virtual humans, embodied agents, online object manipulation, and interactive content creation [4], [5]. Existing HOI video generation methods [3], [6]–[8] improve controllability by conditioning on structured HOI cues. However, most of them follow an offline, fixed-length generation paradigm and mainly produce short clips. Such a paradigm fails to meet the needs of real-time virtual human interaction and online content creation, which require immediate response and continuous long-video generation.

As illustrated in Fig. 1, a straightforward way to generate long HOI videos is to chain multiple short HOI clips, using the ending frames of one clip to initialize the next, as in InteractAvatar-long [9]. This increases video duration, but still inherits the high offline generation cost of bidirectional HOI models and remains prone to interaction collapse in long-form generation. Streaming video generation [10]–[12] provides a different paradigm by generating videos chunk by chunk with nearly constant computation. Such streaming generators rely on cached historical memory to condition each newly generated chunk. However, most short-video HOI methods depend on offline structured conditions, such as pose sequences, object trajectories, depth maps, or inter-

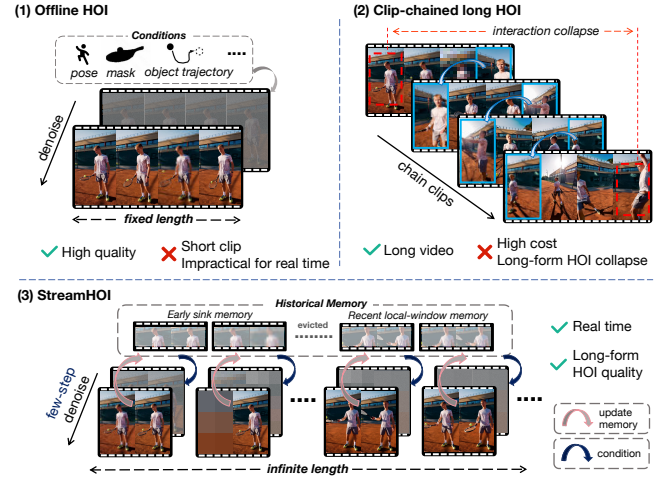


Fig. 1: Conceptual comparison of HOI video generation paradigms. (1) Offline methods are restricted to fixed-length clips. (2) Clip-chained approaches extend duration but suffer from high computational cost and interaction collapse. (3) The proposed *StreamHOI* framework enables causal, low-latency streaming generation while maintaining robust long-horizon interaction consistency.

action layouts, which are difficult to obtain causally during streaming generation. Instead of converting heavily conditioned HOI methods into streaming pipelines, we focus on a more fundamental question: *how should an image-to-video streaming generator organize historical memory to preserve human–object interactions under bounded latency?*

A common memory design for long streaming generation is to keep several early frames as sink memory together

- This work was conducted during the author’s internship at Kling AI Research.
- Corresponding author: Fan Tang, tfan.108@gmail.com
- Z. Rao and F. Tang are with University of Chinese Academy of Sciences, Beijing, China.
- H. Zhang, X. Liu, Y. Meng, G. Zhang and P. Wan are with Kling AI Research, Beijing, China.
- T.-Y. Lee is with National Cheng Kung University, Tainan, Taiwan.

with a rolling local window [13]. Under a fixed memory budget, however, these two memory types create a trade-off in streaming HOI generation. Larger sink memory preserves early appearance and interaction states, improving subject and object consistency and hand-object relations, but it reduces recent context and may cause action jumps and motion discontinuity. Larger local memory improves short-term motion propagation, but it makes the model rely more on recent autoregressive frames and weakens long-term interaction preservation. In Sec. 3, we provide a detailed quantitative analysis of this trade-off. We further observe that different Transformer blocks do not require the same memory composition: some blocks focus more on the subject, hands, object, and contact regions, while others focus more on surrounding scene and object-motion context. This motivates block-wise memory organization for streaming HOI generation.

In this study, we introduce *StreamHOI*, a streaming HOI video generation framework that preserves long-duration human-object interactions through a simple yet effective block-wise organization of historical memory at the Transformer-block level. It first performs *offline HOI-aware block profiling* to identify HOI-biased and surrounding-biased blocks according to their spatial historical-memory bias. Empirically, this profiling captures a stable block-wise memory tendency rather than sample-specific attention fluctuations. To match memory composition with block behavior, we propose *bias-guided memory-specialized training* (B-MST), which assigns block-specific sink-local memory compositions according to the profiled bias: HOI-biased blocks receive more sink memory for consistency-oriented use of reliable early interaction states, while surrounding-biased blocks receive more local-window memory for continuity-oriented use of recent motion and scene context around the human-object interaction. We further introduce a *memory distance scaling* module, which shortens the perceived temporal distance between the current chunk and memory tokens in temporal RoPE, strengthening the access of HOI-biased blocks to distant interaction states.

Our contributions are summarized as follows:

- We identify a sink-local memory trade-off in streaming HOI generation and reveal block-wise historical-memory preferences through offline HOI-aware block profiling, showing that different transformer blocks exhibit HOI-biased and surrounding-biased behaviors.
- We propose bias-guided memory-specialized training, which trains the generator with block-specialized sink-local memory layouts and a memory distance scaling module, enabling different blocks to specialize in long-term interaction consistency and short-term motion continuity under a fixed memory budget.
- We validate StreamHOI through quantitative and qualitative comparisons, long-horizon evaluation, and ablation studies, demonstrating strong performance for streaming HOI video generation.

2 RELATED WORK

2.1 Human-Object Interaction Video Generation

With the progress of controllable human video generation [14]–[16], realistic human-object interaction (HOI) generation has become an important research direction. Human-object interaction video generation aims to synthesize videos in which humans and objects follow physically and semantically plausible relationships. Editing-based methods preserve interactions in real videos by replacing humans or objects, such as MIMO [17] and AnimateAnyone2 [16] for human substitution, and HOI-Swap [7], ReHoLD [18], and iDiT-HOI [19] for object replacement. Generation-based methods improve interaction controllability through explicit conditions such as human poses, object trajectories, depth maps, and bounding boxes. For example, SViMo [20] and HarmoHOI [21] explore video-motion co-generation for egocentric hand-object interactions. AnchorCrafter [3] uses multi-condition guidance to animate human-object interactions, DreamActor-H1 [22] retrieves motion templates and adjusts object layouts from reference images, and HunyuanVideo-HOMA [6] explores sparse user-editable conditions for controllable HOI generation. Recent works such as GeoHOI [8] further improve object awareness and geometric interaction control, while InteractAvatar-long [9] extends HOI generation to longer videos by chaining multiple generated clips.

However, existing HOI video generation methods are still mainly built on offline generation pipelines. Real-time generation and long-term interaction-state preservation in HOI video generation therefore remain underexplored. In contrast, *StreamHOI* targets streaming HOI video generation, aiming to preserve human-object interaction states over long-duration generation while maintaining low-latency inference.

2.2 Streaming and Long Video Generation

Streaming video generation aims to convert offline video diffusion models into causal generators that produce videos chunk by chunk with low latency. Early causal video diffusion models are commonly trained with Teacher Forcing [23] or Diffusion Forcing [24], [25], where the model predicts future frames conditioned on clean or noisy history frames. To further reduce sampling cost, Consistency Distillation [26], [27] and Score Distillation [28], [29] have been used to distill multi-step diffusion models into few-step generators. A common instantiation of Score Distillation is Distribution Matching Distillation (DMD) [30], which minimizes the KL divergence between the student and data distributions by following its gradient. CausVid [10] distills a bidirectional video DiT into a causal student through ODE initialization and asymmetric DMD. Self-Forcing [11] reduces the train-test gap through self-rollout training, while Causal-Forcing [12] improves causal distillation by using an autoregressive teacher for causal ODE initialization.

Long video generation further requires preserving visual and semantic consistency beyond the short training horizon of video diffusion models. Recent methods improve long-horizon stability from different perspectives. FramePack [31] compresses historical frames into a fixed-length context to reduce drift in next-frame prediction.

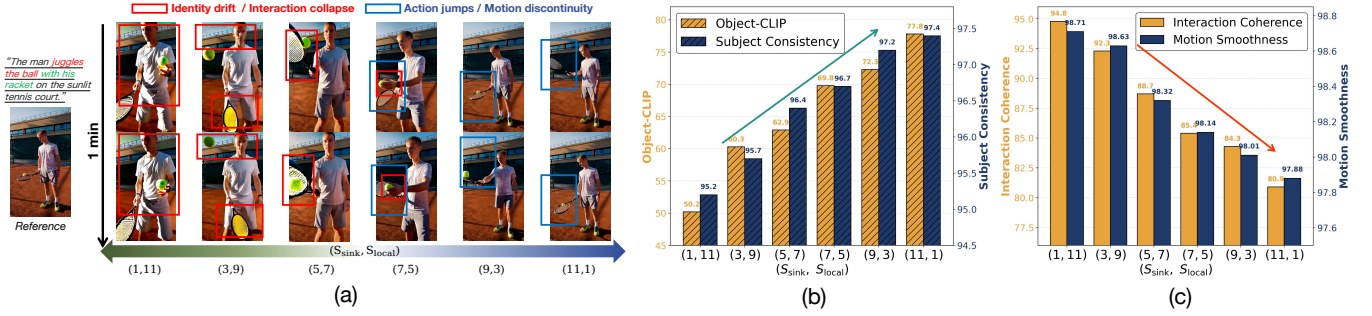


Fig. 2: Illustration of the sink-local memory trade-off in streaming HOI generation under a fixed memory budget. (a) Qualitative impact of varying sink-to-local memory ratios, demonstrating long-term identity drift (insufficient sink memory) versus motion discontinuity (insufficient local memory). (b) Quantitative evaluation on object-clip and subject consistency. (c) Quantitative evaluation on interaction coherence and motion smoothness.

LongLive [13], Rolling Forcing [32], and Self-Forcing++ [33] extend streaming generation through long-context training, self-rollback, rolling generation, or anti-drift strategies. Other methods treat historical context as memory. Context as Memory [34], Memory Forcing [35], MemFlow [36], and Context Forcing [37] retrieve, compress, or organize historical memories to improve scene-level consistency. Another line of work modifies positional or memory mechanisms for longer extrapolation. Infinity-RoPE [38] reformulates temporal RoPE for infinite-horizon rollout, while Deep Forcing [39] uses deep sink and participative compression to stabilize long videos without training. Recent heterogeneous KV methods, such as Head Forcing [40], further profile attention heads and assign head-specific cache strategies for long-range generation.

These methods are closely related to *StreamHOI*, but differ in memory adaptation. Deep Forcing and Infinity-RoPE mainly rely on training-free, discrete RoPE recalibration or cache re-indexing, which may introduce boundary discontinuities when cached temporal segments are reassigned. In contrast, our MDS continuously reshapes temporal RoPE geometry within HOI-biased blocks. Heterogeneous KV methods such as Head Forcing use predefined head-level cache policies, whereas *StreamHOI* learns interaction-aware block-level memory specialization to preserve fine-grained HOI states under bounded latency.

3 PRELIMINARY AND MOTIVATION

Before presenting our method, we analyze how temporal memory affects streaming HOI generation under a bounded memory budget. We first study the sink-local trade-off in the standard memory design, and then examine whether different Transformer blocks retrieve historical memory in the same way. These analyses reveal that HOI generation requires not only a proper global sink-local balance, but also block-wise memory organization.

Why does the standard memory organization fail in HOI generation? Existing streaming generators usually adopt a fixed sink-local memory, where sink frames preserve early appearance and interaction states, and local frames provide recent motion context. To examine this design in HOI generation, we instantiate it with our base streaming generator distilled on human-object interaction data, whose

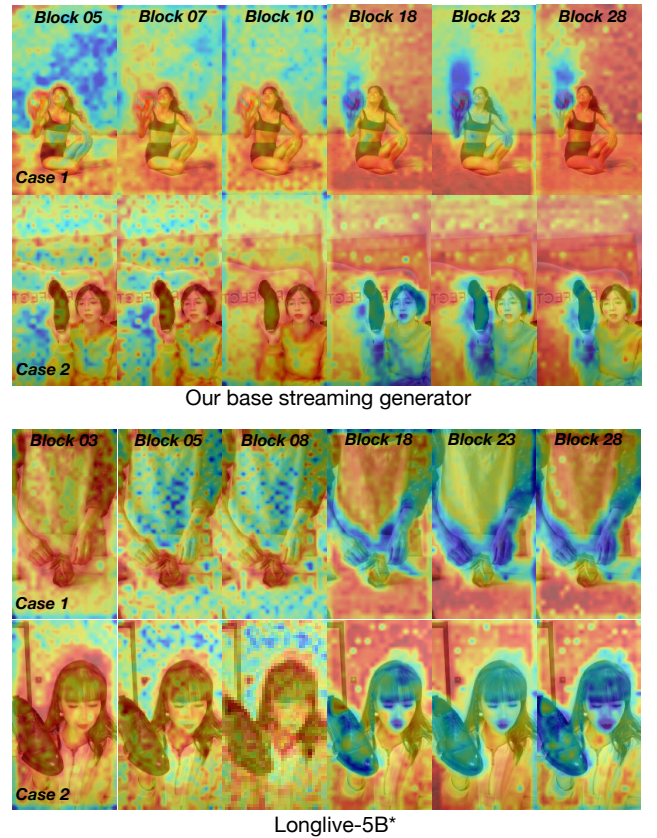


Fig. 3: Block-wise historical-memory attention across two base streaming generators. Within each generator, spatial specialization is stable across diverse HOI examples; across generators, the exact block distribution differs, while shallower blocks are generally more HOI-biased and deeper blocks more surrounding-biased.

construction details are provided in Sec. 5.1.1 and evaluate it on a diagnostic subset of the test set described in Sec. 5.1.2. Under a fixed total memory budget, we vary the sink-local allocation from ($S_{\text{sink}} = 1, S_{\text{local}} = 11$) to ($S_{\text{sink}} = 11, S_{\text{local}} = 1$) over 19k inferred frames.

As shown in Fig. 2(a), small sink memory leads to long-term identity drift and interaction collapse, while small local

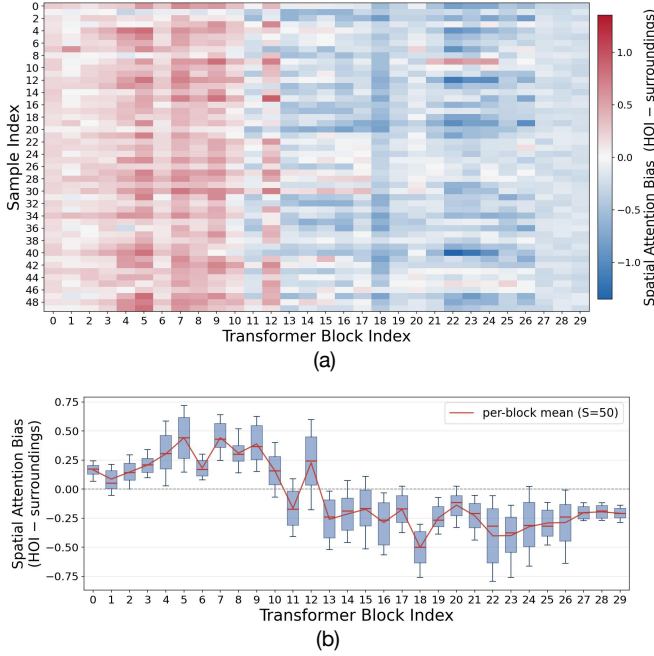


Fig. 4: Cross-sample stability of the block-wise spatial attention bias, computed from 50 randomly sampled HOI examples using the same base streaming generator. (a) Sample-wise heatmap; (b) Per-block box plot showing the 25%–75% IQR with the median in red and whiskers extending to the 10th and 90th percentiles.

memory causes action jumps and motion discontinuities. The quantitative results in Fig. 2(b,c) confirm this trade-off: increasing sink memory improves Object-CLIP and Subject Consistency, but reducing the local window degrades Interaction Coherence and Motion Smoothness. Here, Interaction Coherence is a VLM-based metric for temporal interaction coherence, with details provided in the supplementary material. These results show that, under bounded memory, the standard sink-local organization cannot jointly maintain long-term HOI stability and short-term motion continuity.

Do all Transformer blocks need the same memory composition? To examine whether the sink-local trade-off can be reduced at the block level, we analyze historical-memory attention in two base models: our base streaming generator and LongLive-5B*, our reproduction of LongLive on the Wan2.2-TI2V-5B backbone. We randomly sample 50 HOI examples from HOI-Gen-1M [41] and live e-commerce clips. For each block, we visualize the historical-memory attention over the human–object interaction region and the surrounding region. As shown in Fig. 3, shallower blocks attend more to the subject, hands, interacted object, and contact area, whereas deeper blocks attend more to the surrounding region. We refer to these two tendencies as HOI-biased and surrounding-biased behaviors, respectively.

We further quantify this observation using the profiling procedure later detailed in Sec. 4.2. Fig. 4(a) shows the sample-block heatmap of spatial attention bias. Most blocks keep a consistent bias direction across samples, forming column-wise red or blue patterns in the heatmap. Fig. 4(b) reports the corresponding per-block distributions; the com-

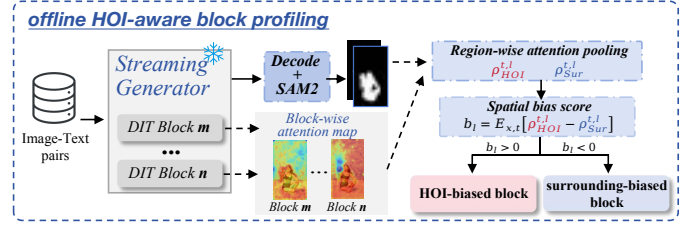


Fig. 5: Pipeline of offline HOI-aware block profiling. We utilize SAM2 to pool historical-memory attention across spatial regions, classifying each Transformer block as either HOI-biased or surrounding-biased based on its attention preference.

pact interquartile ranges further suggest that the bias is a stable block-level property rather than a sample-specific artifact. Our subsequent cross-dataset evaluation on the unseen GeoHOI and HOMA test sets (Sec. 5.2.2) further supports the stability of this block-wise memory behavior beyond the sampled observation set.

Motivation for block-wise memory organization. The above observation suggests that the sink-local trade-off is not only a global memory-budget problem, but also a block-wise memory-matching problem. In the standard fixed design, all Transformer blocks use the same memory composition, regardless of whether they mainly retrieve historical information for the interaction region or the surrounding region. This uniform design can be suboptimal for HOI generation. For HOI-biased blocks, early sink frames provide important cues for preserving the object identity, subject identity, and hand-object relation. Reducing sink memory in these blocks may weaken long-term interaction states. For surrounding-biased blocks, recent local frames provide more useful cues for short-term visual changes and motion continuity. Allocating too much sink memory to these blocks may reduce the local context needed for smooth temporal propagation. Therefore, the memory composition should be adapted according to the block-wise spatial bias, rather than being shared uniformly across all blocks. This motivates our interaction-aware temporal memory adaptation in Sec. 4, where historical memory is organized differently for different Transformer blocks under the same total memory budget.

4 METHOD

4.1 Overview

Starting from a few-step streaming generator with a uniform sink-local memory design, StreamHOI adapts historical memory organization at the Transformer-block level for long-duration HOI generation under a fixed memory budget.

The pipeline includes offline HOI-aware block profiling and bias-guided memory-specialized training (B-MST). Offline HOI-aware block profiling (Fig. 5) identifies the block-wise spatial bias in historical-memory usage. To match memory composition with block behavior, B-MST (Fig. 6) adapts the generator to a block-specific sink-local memory layout and learns memory distance scaling for long-

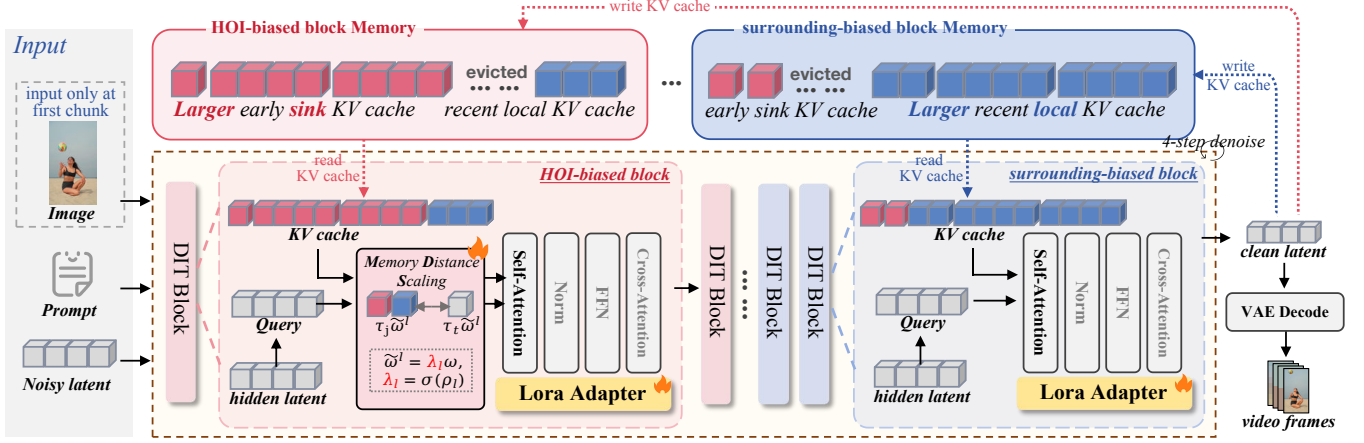


Fig. 6: Pipeline of bias-guided memory-specialized training (B-MST). To resolve the sink-local memory trade-off in streaming HOI generation, B-MST specializes memory composition based on the profiled spatial bias: HOI-biased blocks receive more sink memory, coupled with a memory distance scaling (MDS) module, for consistency-oriented use of reliable early interaction states, while surrounding-biased blocks receive more local memory for continuity-oriented use of recent context.

duration HOI generation. The following subsections describe these two components in detail.

4.2 Offline HOI-aware block profiling

Before B-MST, we perform offline HOI-aware block profiling to categorize Transformer blocks according to their historical-memory attention patterns. Some blocks retrieve historical memory more strongly from the HOI regions, including the human, object and contact area, while others rely more on the surrounding region. This provides a direct criterion for block classification: measuring the relative historical-memory attention assigned to the HOI region and the surrounding region.

For offline profiling, we decode latent chunks into RGB frames on a small profiling set and use SAM2 [42] to obtain the masks of the human subject and the interacted object. The HOI region is defined as the union of the human and object masks, slightly dilated to include local contact areas, and then downsampled to the latent-token resolution. The surrounding region is defined as its complement. These masks are only used during offline profiling and are not required during streaming inference.

For block l at chunk t , let $\rho_{\text{hoi}}^{t,l}$ and $\rho_{\text{sur}}^{t,l}$ denote the average historical-memory attention mass assigned to the HOI region and the surrounding region, respectively. The block-wise spatial bias is computed over the profiling videos and chunks:

$$b_l = \frac{1}{|\mathcal{D}_{\text{prof}}||\mathcal{T}_{\text{prof}}|} \sum_{x \in \mathcal{D}_{\text{prof}}} \sum_{t \in \mathcal{T}_{\text{prof}}} \left(\rho_{\text{hoi}}^{t,l}(x) - \rho_{\text{sur}}^{t,l}(x) \right), \quad (1)$$

where $\mathcal{D}_{\text{prof}}$ denotes the profiling set and $\mathcal{T}_{\text{prof}}$ denotes the profiled chunks. A larger b_l means that block l uses historical memory more strongly for the interaction region, while a smaller b_l indicates stronger reliance on the surrounding region.

The block type is then determined by

$$\text{Type}(l) = \begin{cases} \text{HOI-biased,} & b_l > 0, \\ \text{surrounding-biased,} & \text{otherwise.} \end{cases} \quad (2)$$

A positive bias means that the block assigns more historical-memory attention to the HOI region than to the surrounding region, and is therefore treated as HOI-biased. Otherwise, the block is treated as surrounding-biased. **Since the block-wise spatial bias remains stable across interaction scenarios within the same base generator, a small profiling set is sufficient to determine the block types.** Implementation details are provided in the supplementary material.

4.3 Bias-guided memory-specialized training

After obtaining the HOI-aware block types through offline profiling, we train the streaming generator with a block-specialized memory layout. The goal is to align the temporal memory composition of each Transformer block with its spatial memory preference. HOI-biased blocks are encouraged to use more reliable early states for preserving interaction consistency, while surrounding-biased blocks are encouraged to use richer recent context for maintaining local continuity. We refer to this second-stage training strategy as *bias-guided memory-specialized training* (B-MST).

Let B denote the total memory budget for each Transformer block. In the uniform streaming generator, every block uses the same sink-local allocation (s_0, w_0) , where $B = s_0 + w_0$. B-MST instead assigns a block-specific allocation (s_l, w_l) according to the spatial bias score b_l obtained in Sec. 4.2. We first normalize the block-wise bias as

$$\hat{b}_l = \frac{b_l}{\max_k |b_k| + \epsilon}, \quad (3)$$

where ϵ is a small constant for numerical stability. The sink size of block l is then defined as

$$s_l = \text{Clip} \left(\text{Round} \left(s_0 + B \hat{b}_l \right), s_{\min}, B - w_{\min} \right), \quad (4)$$

and the local-window size is

$$w_l = B - s_l. \quad (5)$$

Here, $s_{\min}$ and $w_{\min}$ ensure that each block retains a minimum number of sink and local-window frames. This

allocation preserves the same total memory budget for every block. A positive $\hat{b}_l$ increases the sink size for HOI-biased blocks, allowing them to retain more early latent frames that contain stable object appearance, contact relations, and manipulation states. A negative $\hat{b}_l$ increases the local-window size for surrounding-biased blocks, allowing them to condition on more recent latent frames for local motion and scene continuity.

At chunk t , the block-specific KV memory is constructed as

$$\begin{cases} \mathcal{M}_l^t = \text{Concat}(\mathcal{M}_{l,\text{sink}}^t, \mathcal{M}_{l,\text{local}}^t), \\ \mathcal{M}_{l,\text{sink}}^t = \text{KV}_l(\{z_0, \dots, z_{s_l-1}\}), \\ \mathcal{M}_{l,\text{local}}^t = \text{KV}_l(\{z_{tm-w_l}, \dots, z_{tm-1}\}). \end{cases} \quad (6)$$

Thus, B-MST changes only the memory composition of each block, without increasing the total KV-cache budget.

Memory distance scaling. Although sink memory preserves early interaction states, its key tokens are temporally far from the current chunk during long streaming generation. Under a standard temporal RoPE, the relative rotation between a query at temporal position τ_t and a key at τ_j is proportional to $(\tau_t - \tau_j)\omega$, where ω is the temporal RoPE frequency. Because $\tau_t - \tau_j$ is large for sink tokens, the resulting rapid phase variation can weaken their query-key correlation and reduce the contribution of early interaction states. To alleviate this issue, B-MST introduces a learnable memory distance scaling mechanism for HOI-biased blocks that acts on the temporal RoPE frequency.

For a HOI-biased block l , we rescale its temporal RoPE frequency by a learnable block-specific factor:

$$\tilde{\omega}^l = \lambda_l \omega, \quad \lambda_l = \sigma(\rho_l), \quad (7)$$

where $\rho_l \in \mathbb{R}$ is a learnable scalar and $\sigma(\cdot)$ is the sigmoid function. For surrounding-biased blocks, we fix $\lambda_l \equiv 1$ so that the original temporal distance is preserved and no additional parameter is introduced.

The phase perturbation induced by this scaling is

$$\Delta\phi_{tj}^l = (1 - \lambda_l)(\tau_t - \tau_j)\omega. \quad (8)$$

This expression shows that MDS is not a hard sink-only re-indexing operation: since the temporal RoPE frequency is rescaled at the block level, all query-key temporal distances in the corresponding HOI-biased block are affected. However, its effect is inherently distance-dependent. Local-window keys have bounded temporal distances to the current chunk, whereas sink keys become increasingly distant as streaming generation proceeds. Therefore, the same scaling factor produces a much stronger phase correction for distant sink tokens than for recent local tokens.

MDS is intentionally applied only to HOI-biased blocks. These blocks are assigned larger sink memories by B-MST and require stronger access to early interaction anchors, such as contact relations, relative poses, and object states. Surrounding-biased blocks keep the original RoPE frequency, preserving their recent-context behavior for motion and scene continuity.

Training objective. B-MST is conducted as a lightweight second-stage training procedure. We freeze the pretrained streaming generator and optimize only LoRA adapters ϕ inserted into the Transformer blocks, together with the

distance-scaling parameters $\{\rho_l\}$ for HOI-biased blocks. Given the reference human image I_H , text prompt P , latent noise ϵ^t , and block-specific memories $\{\mathcal{M}_l^t\}_{l=0}^{L-1}$, the generator produces the current latent chunk:

$$\hat{Z}^t = G_{\theta,\phi}(I_H, P, \epsilon^t, \{\mathcal{M}_l^t\}_{l=0}^{L-1}), \quad (9)$$

where θ denotes the frozen generator parameters.

We use the DMD distillation objective:

$$\mathcal{L}_{\text{B-MST}} = \mathcal{L}_{\text{DMD}} + \lambda_c \mathcal{L}_{\text{critic}}. \quad (10)$$

Here, $\mathcal{L}_{\text{DMD}}$ updates the LoRA-augmented generator by matching the generated distribution to the real-score model, while $\mathcal{L}_{\text{critic}}$ updates the fake-score critic with a denoising objective on generated samples.

Different from the standard DMD training that rolls out the generator with a uniform sink-local memory for all Transformer blocks, B-MST performs generator rollouts and score evaluations with the bias-guided memory layout, so the distillation gradients are computed under block-specialized historical contexts.

Through B-MST, the streaming generator learns to use historical memory according to the functional preference of each block. HOI-biased blocks are strengthened toward consistency-oriented memory use, while surrounding-biased blocks are strengthened toward continuity-oriented memory use. During inference, we use the profiled block-specific memory layout and the learned distance-scaling coefficients without introducing additional model branches or increasing the total memory budget.

5 EXPERIMENTS

5.1 Experimental Settings

5.1.1 Implementation Details

We use Wan2.2-TI2V-5B as the base video diffusion model and generate videos at a resolution of 704×1280 . We first construct a chunk-wise few-step streaming generator through teacher-forcing causal training, causal ODE distillation, and DMD distillation, and then apply the proposed bias-guided memory-specialized training (B-MST) for interaction-aware temporal memory adaptation. All stages are implemented in a chunk-wise streaming manner, where each chunk contains 4 latent frames. In the teacher-forcing causal training stage, we train an autoregressive diffusion model for 13K steps on 100K video clips. We then use the trained autoregressive model as the teacher to sample 10K causal ODE trajectories, and perform causal ODE distillation for 5K steps to obtain a 4-step generator. In the DMD distillation stage, we adopt a uniform sink-local memory configuration with $s = 4$ sink latent frames and $w = 8$ local-window latent frames, and train the streaming generator for 300 steps. Finally, we apply B-MST for 700 steps under the profiled block-specific memory layout. The LoRA rank and alpha are both set to 256. All training experiments are conducted on 32 80GB GPUs.

5.1.2 Dataset

Training dataset. We train StreamHOI on a 100K video dataset. The dataset consists of 80K internal live e-commerce video clips and 20K HOI clips sampled from HOI-Gen-1M

TABLE 1: Quantitative comparison with representative methods under different temporal settings on the main test set. LC-Avatar denotes Longcat-Video-Avatar 1.5. Our 5B model substantially improves HOI-specific metrics while maintaining high general visual quality, ranking second only to the 13.6B LC-Avatar on most general metrics.

Method		HOI-specific Metrics				General Metrics			
		Obj-CLIP↑	InternVL _O ↑	InternVL _H ↑	InternVL _I ↑	Subject Cons.↑	Background Cons.↑	Motion Smooth.↑	Aesthetic Quality↑
5-second	Causal Forcing-1.3B	0.7104	0.823	0.921	0.836	0.9688	0.9665	0.9821	0.5298
	LongLive-1.3B	0.7261	0.821	0.908	0.842	0.9683	0.9672	0.9815	0.5281
	Causal Forcing-5B*	0.8032	0.869	0.937	0.860	0.9699	0.9724	0.9833	0.5493
	LongLive-5B*	0.7927	0.862	0.994	0.857	0.9742	0.9778	0.9869	0.5490
	Deep Forcing-5B	0.8181	0.885	0.982	0.864	0.9729	0.9780	0.9877	0.5503
	LC-Avatar-13.6B	0.8539	0.899	0.995	0.889	0.9879	0.9893	0.9969	0.5482
	InteractAvatar-10B	0.7688	0.851	0.982	0.884	0.9760	0.9789	0.9893	0.5426
	Ours-5B	0.8621	0.909	0.993	0.901	0.9782	0.9801	0.9934	0.5558
30-second	Causal Forcing-1.3B	0.6033	0.681	0.794	0.704	0.9486	0.9599	0.9812	0.5139
	LongLive-1.3B	0.6109	0.631	0.802	0.715	0.9495	0.9622	0.9808	0.5118
	Causal Forcing-5B*	0.6607	0.782	0.831	0.782	0.9586	0.9681	0.9827	0.5211
	LongLive-5B*	0.6586	0.789	0.824	0.790	0.9701	0.9766	0.9845	0.5234
	Deep Forcing-5B	0.6936	0.805	0.816	0.801	0.9692	0.9773	0.9858	0.5269
	LC-Avatar-13.6B	0.8244	0.863	0.951	0.866	0.9869	0.9891	0.9966	0.5438
	InteractAvatar-10B	0.7492	0.827	0.924	0.822	0.9709	0.9785	0.9872	0.5377
	Ours-5B	0.8496	0.882	0.958	0.874	0.9751	0.9802	0.9923	0.5431
60-second	Causal Forcing-1.3B	0.4930	0.537	0.688	0.613	0.9480	0.9582	0.9796	0.4998
	LongLive-1.3B	0.5164	0.548	0.707	0.626	0.9488	0.9603	0.9821	0.4981
	Causal Forcing-5B*	0.5882	0.687	0.792	0.702	0.9577	0.9669	0.9828	0.5203
	LongLive-5B*	0.5937	0.688	0.808	0.712	0.9640	0.9760	0.9842	0.5193
	Deep Forcing-5B	0.6334	0.703	0.831	0.745	0.9659	0.9731	0.9857	0.5251
	LC-Avatar-13.6B	0.8156	0.842	0.940	0.834	0.9846	0.9880	0.9963	0.5431
	InteractAvatar-10B	0.7344	0.808	0.923	0.791	0.9693	0.9781	0.9878	0.5372
	Ours-5B	0.8298	0.857	0.942	0.868	0.9734	0.9796	0.9902	0.5429

TABLE 2: Comparison with bidirectional HOI methods under the 5-second setting. Uni-DiT denotes UniAnimate-DiT. Our method achieves competitive performance without extra object-information injection.

Method	HOI-specific Metrics				General Metrics			
	Obj-CLIP↑	InternVL _O ↑	InternVL _H ↑	InternVL _I ↑	Subject Cons.↑	Background Cons.↑	Motion Smooth.↑	Aesthetic Quality↑
GeoHOI test set								
Uni-DiT-14B	0.8786	0.818	0.945	0.848	0.9814	0.9858	0.9948	0.5274
VACE-14B	0.8516	0.736	0.962	0.825	0.9831	<u>0.9863</u>	<u>0.9951</u>	0.5391
GeoHOI-5B	0.9015	<u>0.882</u>	<u>0.963</u>	0.898	0.9810	0.9860	0.9923	<u>0.5396</u>
Ours-5B	<u>0.8993</u>	0.914	0.980	<u>0.892</u>	<u>0.9824</u>	0.9879	0.9957	0.5483
HOMA test set								
HUMO-17B	0.8560	0.758	0.926	0.769	<u>0.9792</u>	0.9891	0.9940	0.5100
VACE-14B	0.8572	0.880	0.997	0.842	0.9788	<u>0.9878</u>	<u>0.9900</u>	0.5350
HOMA-13B	0.8803	<u>0.887</u>	0.983	0.885	0.9774	0.9860	0.9810	0.5860
Ours-5B	<u>0.8790</u>	0.905	<u>0.984</u>	0.896	0.9806	0.9869	0.9830	<u>0.5380</u>

[41]. The internal videos contain diverse human demonstrations and object manipulation scenarios in live e-commerce settings, while the HOI-Gen-1M subset provides additional human-object interaction examples with broader interaction categories.

Evaluation dataset. Existing HOI test sets mainly target short offline video generation and cover limited interaction scenarios. We therefore construct a main test set under diverse real-world interaction scenarios, containing 400 HOI cases and resulting in 384k inferred frames. Specifically, it includes 100 live e-commerce videos and 300 videos selected from HOI-Gen-1M, all excluded from training. To characterize its scenario diversity, we group the videos into coarse interaction categories according to their captions, as

summarized in Table 3.

In addition, we use GeoHOI-testset [8] and HOMA-testset [6] for comparison with existing bidirectional HOI generation methods. These two test sets are tied to their corresponding methods and provide standardized evaluation prompts or conditions for short-video HOI generation. Each test set contains 100 examples. Therefore, we use them only for 5-second comparison with bidirectional HOI baselines.

5.1.3 Comparison Methods

We compare StreamHOI with two groups of baselines. For long-video generation methods, we evaluate all methods under the 5-, 30-, and 60-second settings to measure both short-term quality and long-duration stability. For bidirec-



Fig. 7: Qualitative comparison against state-of-the-art long-video generation frameworks. While baseline methods exhibit severe identity drift, object distortion, or temporal inconsistencies over extended contexts, our approach preserves robust subject-object fidelity and coherent interaction dynamics. Please refer to the supplementary materials for the full video sequences.

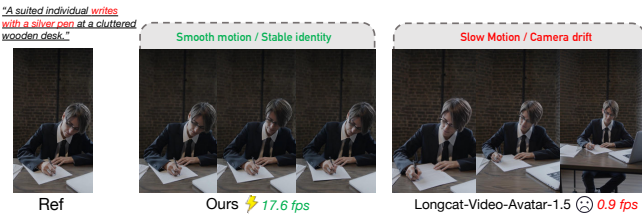


Fig. 8: Further comparison with Longcat-Video-Avatar 1.5 under the 60-second setting. Video results can be found in the supplementary materials.

tional HOI generation methods, we evaluate them under the 5-second setting, since these methods are mainly designed for short offline video generation.

Long-video generation methods. First, we compare with representative streaming video generation methods, including Causal Forcing (frame-wise) [12], LongLive [13], and Deep Forcing [39]. Since Causal Forcing and LongLive only release 1.3B model weights, we reproduce them on

TABLE 3: Interaction category distribution of the main evaluation test set.

Interaction Category	#Videos
Digital device interaction	46
Food preparation	44
Art / craft creation	40
Footwear handling	40
Object demonstration	40
Wearable / grooming adjustment	34
Sports / mobility activity	31
Gesture / explanation	30
Household / plant / pet care	29
Study / tool use	24
Music / dance performance	23
Food / drink serving	19
Total	400

the Wan2.2-TI2V-5B backbone using our training data and closely matched training configurations, denoted as Causal Forcing-5B* and LongLive-5B*. Deep Forcing [39] is a training-free long-video generation method; we apply it

to our baseline model without B-MST for comparison. We also compare with InteractAvatar-long [9], an offline long-video HOI generation method based on sequential clip generation. In addition, we include LongCat-Video-Avatar 1.5 [43], a 13.6B few-step avatar video generation method. Since LongCat-Video-Avatar 1.5 additionally takes audio as input, we use a silent audio track to keep the comparison under the same image-and-text-driven setting. All these methods are evaluated on the main test set.

Bidirectional HOI methods. Second, we compare with bidirectional HOI video generation methods on HOI test sets under the 5-second setting, as these baselines are mainly designed for short video generation. Since StreamHOI only takes a reference image and a text prompt as input, we convert the hand-pose and object-trajectory conditions provided by the datasets into textual descriptions using Qwen2.5-VL [44], and append them to the prompts for our method. Therefore, this experiment focuses on comparing the final HOI generation quality on the same test cases, rather than strict condition-following accuracy. On GeoHOI-testset, we compare with UniAnimate-DiT [45], VACE [46], and GeoHOI [8]. UniAnimate-DiT generates human videos from a person image and a pose sequence based on a Wan-14B backbone. VACE is a general video editing framework built on a 14B video model and supports text-controlled generation and editing. GeoHOI improves HOI video generation by enhancing object awareness and geometric interaction control. On HOMA-testset, we compare with HUMO [47], VACE, and HOMA [6]. HUMO supports video generation conditioned on multiple reference images. HOMA introduces weak motion guidance for HOI human video generation.

5.1.4 Metrics

We evaluate our method using several established metrics. Object-CLIP [3] is used to evaluate the object consistency by computing the CLIP [48] cosine similarity of the objects between the input image and the generated videos within the segmentation. Since traditional metrics struggle to capture fine-grained physical contact and interaction evolution, we further use InternVL3-38B [49] as a structured evaluator. It performs a multi-dimensional binary assessment on sampled frames to evaluate object fidelity, human quality, and interaction plausibility. Details of the metrics are provided in the supplementary materials. Subject Consistency, Background Consistency, Motion smoothness and Aesthetic Quality extracted from Vbench [50] are employed to measure general video quality. We also report FPS and Latency as efficiency metrics in Sec. 5.6.

5.2 Comparison Results

5.2.1 Comparison with long-video generation methods

According to Table 1, our method consistently achieves the best results across nearly all quality metrics under different temporal settings. For HOI-specific evaluation, our method obtains the highest Obj-CLIP scores at 5, 30, and 60 seconds, showing better object appearance preservation than representative baselines. The InternVL-based evaluation further shows that our method better maintains human and object states while producing more reliable hand-object



Fig. 9: Comparison with SOTAs on the GeoHOI-testset under the 5-second setting. Video results can be found in the supplementary materials.

interactions. For general video quality, our method also achieves strong Subject Consistency and Background Consistency, indicating stable foreground subjects and scenes over time. Meanwhile, the competitive Motion Smoothness and Aesthetic Quality scores show that our method maintains smooth motion and high visual quality across different video lengths. Detailed per-category results and additional Qwen3-VL-based evaluations [51] are provided in the supplementary material; **the Qwen3-VL-based results assess cross-evaluator robustness and show broadly consistent performance trends with InternVL3-38B.**

A key advantage of our method is its strong long-duration generation ability. As the generation length increases from 5 seconds to 60 seconds, most streaming baselines show clear performance drops, especially on Obj-CLIP and InternVL scores. This suggests that they gradually lose early object and interaction states during long autoregressive generation. In contrast, our method keeps high scores even at 60 seconds, with only a small decrease. The qualitative results in Fig. 7 show a similar trend. Streaming baselines often suffer from object drift, weak hand-object contact, inconsistent object appearance, and degraded motion after long generation. Our method better maintains human identity, object appearance, background stability, and physically plausible interaction dynamics throughout the video.

We further show qualitative comparisons with LongCat-Video-Avatar 1.5 [43], a 13.6B few-step avatar video generation method. As shown in Fig. 8, LongCat-Video-Avatar 1.5 tends to produce slow motion and shows camera drift over long generation. In contrast, our method maintains more natural motion dynamics, more stable human identity throughout long videos.



Fig. 10: Comparison with SOTAs on the HOMA-testset under the 5-second setting. Video results can be found in the supplementary materials.

5.2.2 Comparison with bidirectional HOI methods

As shown in Table 2, our model achieves highly competitive results compared with recent bidirectional HOI generation methods. On the HOI-specific metrics, our method obtains higher Obj-CLIP than UniAnimate-DiT, VACE, and HUMO, showing stronger object appearance preservation. Without any additional object condition injection, our method still achieves comparable Obj-CLIP performance with GeoHOI and HOMA. More importantly, our method obtains very competitive InternVL_O and InternVL_I scores, indicating that it can effectively maintain object consistency and produce reasonable human-object interactions. The InternVL_H result also shows that our method preserves the human appearance well during generation. In addition, our method achieves strong performance on general video metrics, demonstrating that the proposed streaming framework can jointly support high-quality HOI generation and strong overall video quality. *Crucially, achieving such competitive performance on the completely unseen GeoHOI and HOMA test sets firmly validates the cross-dataset generalization of our offline block profiling design.*

As shown in Fig. 9, we present qualitative comparisons on the GeoHOI-testset. Our method preserves the object appearance well and produces a reasonable interaction process. Even without pose guidance or object-driven control, our method can naturally adjust the strap according to the interaction context. In contrast, UniAnimate-DiT and GeoHOI require additional pose injection, yet their generated results appear less natural, while VACE shows clear color-tone changes in the generated video. As shown in Fig. 10, we further present qualitative results on the HOMA-testset. Our method preserves the face identity well and generates natural contact between the hands and the keyboard. Other methods show more visible artifacts: the generated faces are often distorted, HOMA fails to follow the instruction that the person should look at the keyboard, HUMO changes the

keyboard appearance, and VACE changes the shape of the hat. These qualitative results further show that our method can maintain human identity, object appearance, and interaction plausibility. It is worth noting that our method generates a 5-second video in only 4.6 seconds, which is much faster than existing bidirectional HOI methods.

5.3 Ablation Studies and Discussion

5.3.1 Validation of bias-guided memory-specialized training

We first validate the effect of bias-guided memory-specialized training (B-MST). As shown in Table 4, all variants are evaluated under the same 60-second setting and use the same total memory budget. To verify that the improvement does not come from additional LoRA fine-tuning alone, we further train the uniform-memory baseline for the same 700 steps with the original sink-local layout, denoted as Uniform-Mem FT. This equal-step control brings only marginal changes over the model before B-MST, indicating that the 300-step DMD stage has largely converged under the uniform-memory design. In contrast, the full model substantially improves both HOI-specific and general metrics. These results show that the gains mainly arise from bias-guided memory specialization under a fixed KV-cache budget, rather than from a larger cache or additional LoRA fine-tuning alone.

The qualitative results in Fig. 11 are consistent with the quantitative findings. Without B-MST, the model shows attenuation of early interaction priors during long autoregressive generation, leading to unstable hand-object contact, changed object states, and drifted human identity. In contrast, B-MST aligns each Transformer block with its profiled spatial bias: HOI-biased blocks preserve reliable early interaction cues, while surrounding-biased blocks retain recent context for local visual changes and motion continuity. This block-specific memory behavior helps maintain long-term interaction consistency without sacrificing short-term temporal smoothness.

Block grouping verification. We further examine whether the proposed memory specialization depends on the semantic grouping of blocks. We compare our semantic grouping with two controls: reverse grouping and random grouping. Reverse grouping exchanges the memory assignments of HOI-biased and surrounding-biased blocks, while random grouping keeps the same group sizes but samples block identities randomly. As reported in Table 4, both controls perform worse than our semantic grouping, with reverse grouping causing the largest degradation in long-term HOI consistency. This indicates that non-uniform memory allocation alone is insufficient; the memory behavior needs to match the semantic bias of each block group. The qualitative results in Fig. 13 show the same trend: reverse and random grouping often lead to contact drift, object state changes, or unstable temporal continuity, whereas our grouping better preserves both interaction states and short-term visual consistency.

Generalization validation. To verify that the proposed memory specialization is not tied to a specific backbone, we further apply B-MST to LongLive-5B*. Specifically, we first apply the same offline HOI-aware block profiling strategy

TABLE 4: Ablation study under the 60-second setting. We ablate bias-guided memory-specialized training (B-MST), memory distance scaling (MDS), and the block grouping strategy. Uniform-Mem FT denotes uniform-memory fine-tuning, an equal-step LoRA fine-tuning baseline that uses the same 700 LoRA fine-tuning steps and the same total KV-cache budget as B-MST, while keeping the original uniform sink-local memory layout.

Method	HOI-specific Metrics				General Metrics			
	Obj-CLIP↑	InternVL _O ↑	InternVL _H ↑	InternVL _I ↑	Subject Cons.↑	Background Cons.↑	Motion Smooth.↑	Aesthetic Quality↑
Ours	0.8298	0.857	0.942	0.868	0.9734	0.9796	0.9902	0.5429
w/ fixed MDS	0.8204	0.840	0.923	0.864	0.9730	0.9792	0.9885	0.5366
w/o MDS	0.7519	0.781	0.880	0.774	0.9647	0.9759	0.9883	0.5279
w/o B-MST	0.6028	0.682	0.836	0.701	0.9583	0.9748	0.9851	0.5236
Uniform-Mem FT	0.6093	0.675	0.859	0.688	0.9592	0.9744	0.9855	0.5269
Reverse Grouping	0.5744	0.671	0.808	0.662	0.9579	0.9747	0.9839	0.5231
Random Grouping	0.5939	0.686	0.821	0.724	0.9601	0.9732	0.9852	0.5282

TABLE 5: Generalization of Bias-Guided Memory-Specialized Training.

Method	HOI-specific Metrics				General Metrics			
	Obj-CLIP↑	InternVL _O ↑	InternVL _H ↑	InternVL _I ↑	Subject Cons.↑	Background Cons.↑	Motion Smooth.↑	Aesthetic Quality↑
LongLive-5B*	0.5937	0.688	0.808	0.712	0.9640	0.9731	0.9842	0.5193
+ B-MST	0.8036	0.832	0.926	0.847	0.9694	0.9758	0.9864	0.5281

in Sec. 4.2 to LongLive-5B*, which identifies its HOI-biased and surrounding-biased blocks. The resulting block grouping is shown in Fig. 14(b), and B-MST is then instantiated according to these profiled block types. As reported in Table 5, B-MST consistently improves LongLive-5B* across both HOI-specific and general metrics. The gains on Obj-CLIP and InternVL-based metrics indicate better preservation of object appearance, human states, and interaction relations, while the improvements on general metrics show more stable foreground and background consistency. The qualitative comparison in Fig. 14(a) further shows that the original LongLive-5B* accumulates errors during long generation, leading to unstable hand-object contact and appearance changes. With B-MST, the model better maintains the hand-object interaction and object state across chunks. These results show that B-MST can serve as a general memory adaptation strategy for long-term streaming HOI generation.

5.3.2 Validation of memory distance scaling

We then validate the effect of memory distance scaling module (MDS). MDS adjusts the perceived temporal distance between the current chunk and the sink memory of HOI-biased blocks, making early interaction states easier to access during long streaming generation. As shown in Table 4, removing MDS leads to a clear performance drop compared with the full model. This indicates that memory capacity and temporal distance play complementary roles. Block-specific memory allocation provides HOI-biased blocks with more early interaction tokens, while MDS makes these temporally distant tokens more effective during attention.

Effect of learnable MDS. We also evaluate a variant with manually fixed MDS parameters. These parameters are set according to the offline profiling results in Sec. 4.2, with details provided in the supplementary material. As shown

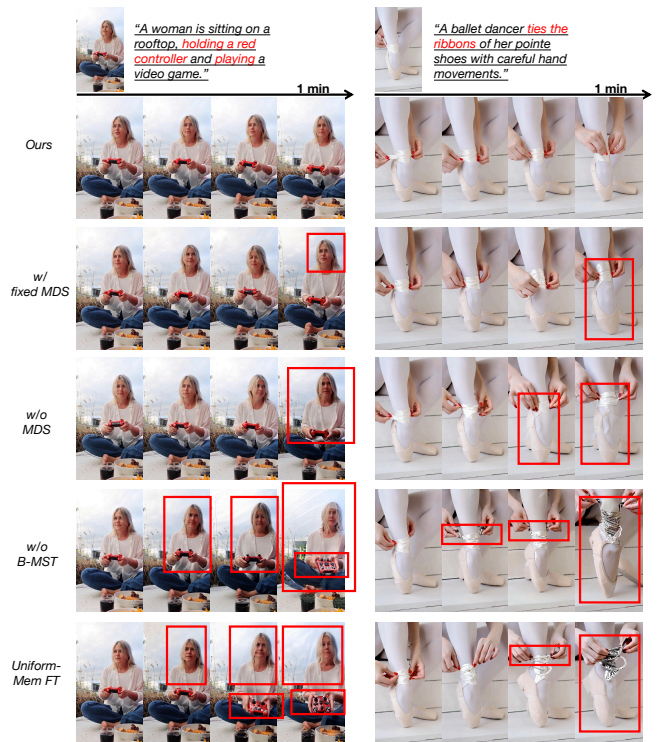


Fig. 11: Visual ablation of B-MST and MDS variants. Removing these components leads to interaction collapse and identity drift, whereas our full method maintains robust long-term interactions.

in Table 4, the fixed MDS parameter variant already achieves strong performance, suggesting that the profiling results provide useful guidance for configuring memory distance scaling module. However, it is still consistently worse than the learnable MDS used in our full model. This shows

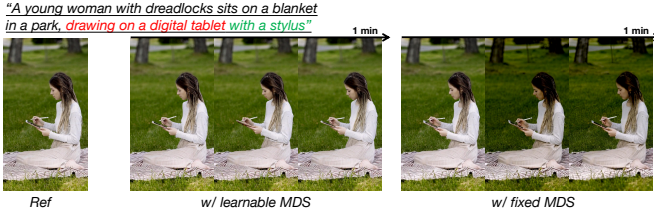


Fig. 12: Mitigation of temporal degradation using learnable MDS. While the fixed MDS configuration suffers from progressive color drift during extended generation, the learnable MDS mechanism effectively preserves color consistency, resulting in improved long-horizon consistency.

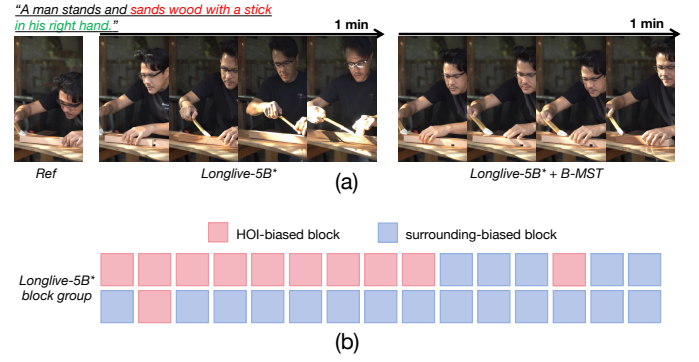


Fig. 14: Generalization of B-MST to LongLive-5B*. (a) Integrating B-MST prevents the interaction drift and object distortion observed in the original LongLive-5B* during long-horizon generation. (b) The profiled block grouping for LongLive-5B*.

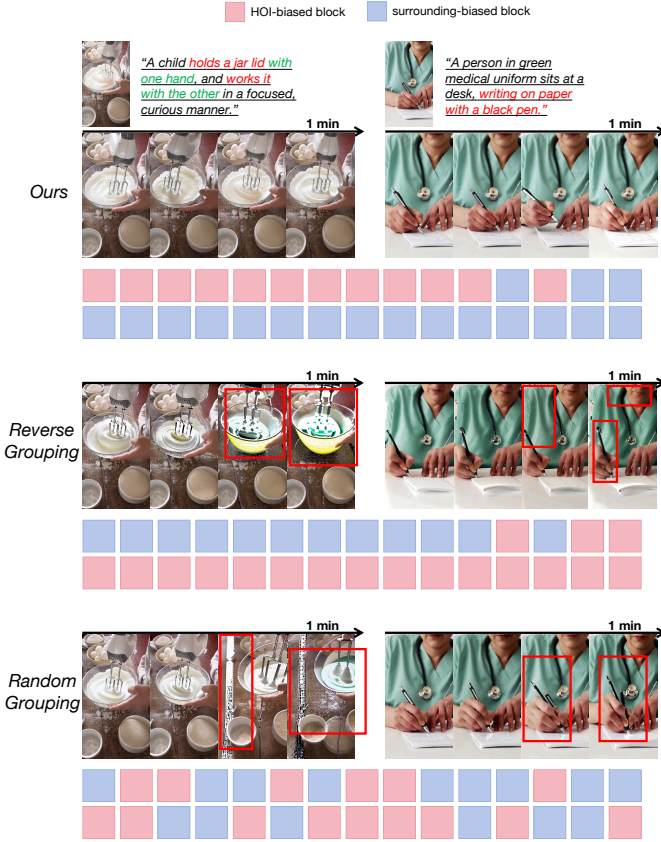


Fig. 13: Ablation of block grouping strategies. Mismatching memory allocation with spatial bias (Reverse/Random) causes contact drift, whereas our semantic grouping preserves stable interactions.

that manually designed scaling can provide a reasonable prior, but learning the scaling coefficients allows different HOI-biased blocks to adapt their access to long-range sink memory more effectively.

Fig. 12 highlights a manifestation of temporal degradation in the fixed MDS variant, characterized by progressive color drift over extended generation horizons. After learning the MDS coefficients, our full model better preserves the color appearance and overall visual consistency.

5.4 User Study

We conducted a user study to assess the perceptual quality of our method relative to long-video methods, and bidirectional HOI generation methods. We adopted an N -way forced-choice protocol. For each image-text pair, participants viewed the videos generated by all competing methods within the corresponding benchmark, presented in randomized order, and selected the single best result according to four criteria: Human Consistency, Object Consistency, Interaction Naturalness, and Overall Video Quality. In total, 28 participants took part in the study, with each participant completing 40 randomized trials across the three benchmarks. Short breaks were provided at regular intervals to reduce fatigue. For each benchmark and criterion, we report the percentage of trials in which each method was selected as the best, referred to as the first-choice preference rate. The results are summarized in Fig. 15.

The user preferences in Fig. 15 show a consistent advantage for our method on long-video generation. On the **Long-Video Bench**, our method receives the highest first-choice preference rate across all four criteria, ranging from 36.8% to 38.0%. This is substantially higher than the strongest baseline, LC-Avatar 1.5 (24.5%–27.0%), and well above the remaining streaming methods. In particular, Causal Forcing and LongLive at the 1.3B scale are almost never selected as the best result (<0.4%), while their 5B variants and Deep Forcing receive only 6%–13% of the votes. These results indicate that the proposed memory design more reliably preserves human appearance, object identity, and hand-object interaction states during long video generation.

Our method also remains competitive against dedicated bidirectional HOI generation models. On the **GeoHOI Bench**, StreamHOI obtains 33.4%–35.8% first-choice preference across the four criteria, matching the dedicated GeoHOI model (30.9%–34.9%) and outperforming UnianimateDiT (20.8%–22.6%) and VACE (9.1%–12.3%). On the **HOMA Bench**, StreamHOI achieves the highest preference rate for every criterion (40.0%–44.4%), ahead of HOMA (35.0%–40.8%), whereas VACE and HUMO are selected much less frequently. Overall, the study shows that StreamHOI delivers perceptual quality comparable to, and often preferred

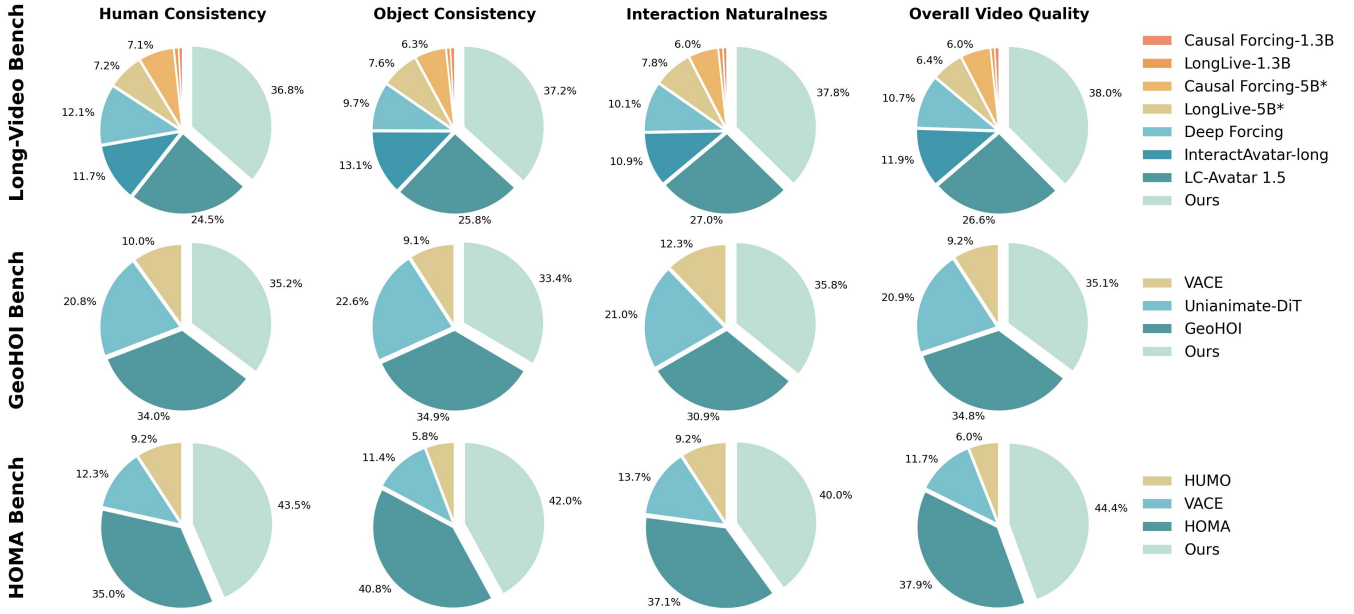


Fig. 15: **User study results.** Pie charts show first-choice preference rates on Long-Video, GeoHOI, and HOMA benchmarks under four perceptual criteria: Human Consistency, Object Consistency, Interaction Naturalness, and Overall Video Quality.

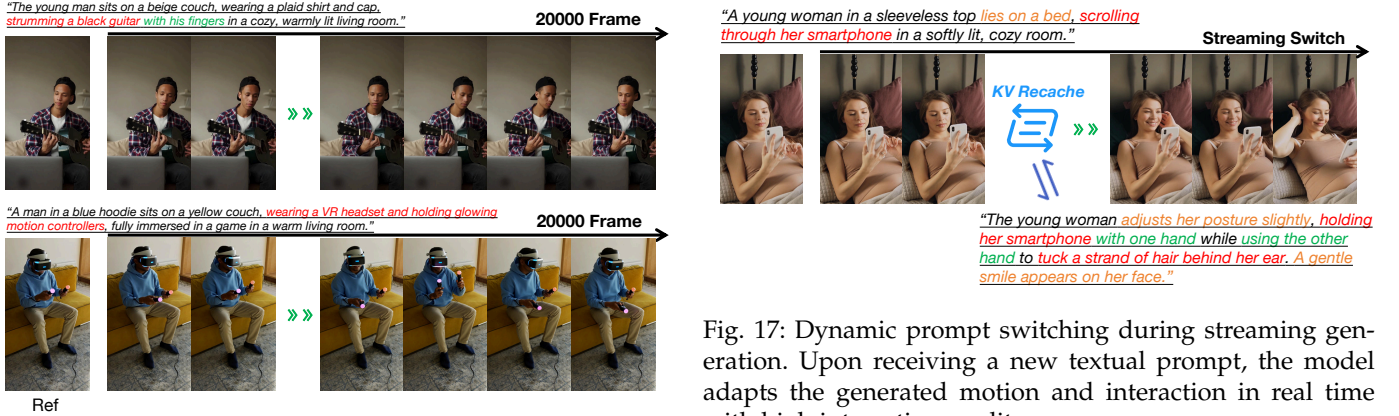


Fig. 16: Qualitative results of extended-horizon generation. StreamHOI maintains stable human-object interactions and structural consistency over a 20,000-frame sequence, showing reduced visual degradation during long autoregressive generation.

over, specialized bidirectional HOI generation methods, while preserving the efficiency and temporal scalability of streaming generation.

5.5 Long-Horizon Evaluation

We further evaluate the long-horizon generation ability of our method. As shown in Fig. 16, our model can generate a 20,000-frame video while maintaining overall visual and interaction consistency. During the whole generation process, the model maintains stable human-object interaction quality, consistent object states, and smooth motion transitions. This result shows that the proposed temporal memory adaptation mechanism can effectively support

Fig. 17: Dynamic prompt switching during streaming generation. Upon receiving a new textual prompt, the model adapts the generated motion and interaction in real time with high interaction quality.

long-duration streaming generation and reduce error accumulation over a very long horizon. Notably, our method can be combined with the KV Recache [13] mechanism to dynamically switch prompts during streaming generation, enabling streaming control of human-object interactions. As illustrated in Fig. 17, the model responds immediately upon receiving the next prompt, accurately following the instructed interaction changes.

5.6 Runtime Analysis

We report the runtime comparison in Table 6. StreamHOI achieves 17.6 FPS with a first-chunk latency of 0.75s. At the same 5B model scale, it matches the throughput of LongLive* and is faster than Causal Forcing* and Deep Forcing, while providing stronger long-duration HOI quality as shown in the previous comparisons. Although the 1.3B LongLive baseline reaches higher FPS, its generation quality is substantially weaker, especially under long temporal settings. Compared with larger avatar or bidirectional

TABLE 6: FPS and latency comparison of different methods.

Method	Model Size	FPS↑	Latency↓
Causal Forcing	1.3B	8.9	0.45
LongLive	1.3B	20.7	0.69
Causal Forcing*	5B	14.4	0.75
LongLive*	5B	17.6	0.75
Deep Forcing	5B	14.3	0.75
LC-Avatar 1.5	13.6B	0.9	87.51
InteractAvatar-Long	10B	0.14	600
UniAnimate-DiT	14B	0.15	549
VACE	14B	0.24	334
HUMO	17B	0.17	466
GeoHOI	5B	0.93	87
HOMA	13B	0.25	329
StreamHOI	5B	17.6	0.75

HOI generation methods, StreamHOI is significantly more efficient: LC-Avatar 1.5 requires 87.51s and InteractAvatar-Long requires 600s, while other bidirectional HOI methods require 87s to 549s for end-to-end generation. These results show that StreamHOI preserves the low-latency advantage of streaming generation while achieving competitive long-duration HOI quality.

6 LIMITATIONS

When the human-object interaction in the reference image is unclear, such as ambiguous contact regions or occluded objects, StreamHOI may generate less stable interactions. In addition, since our training data mainly contains single-person videos, StreamHOI may be less effective in complex multi-person scenarios. We leave these cases for future work.

7 CONCLUSION

We present StreamHOI, a streaming framework for long-duration human-object interaction video generation. By analyzing block-wise spatial attention, we showed that different Transformer blocks have distinct memory preferences for HOI region and surrounding region. Based on this observation, StreamHOI uses bias-guided memory-specialized training to adapt the generator to a block-specific sink-local memory layout, and applies memory distance scaling module to strengthen long-range access to reliable interaction anchors. Experiments on long-video and HOI benchmarks show that StreamHOI improves interaction plausibility, object preservation, temporal stability, and runtime efficiency under bounded memory and low-latency generation. These results suggest that interaction-aware temporal memory adaptation is an effective direction for scalable streaming HOI video generation.

REFERENCES

- [1] B. Zuo, Z. Zhao, W. Sun, X. Yuan, Z. Yu, and Y. Wang, "Grasdiff: Grasping generation for hand-object interaction with multimodal guided diffusion," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 9, pp. 5642–5654, 2024.
- [2] Z. Hu, Z. Yin, D. Haeufle, S. Schmitt, and A. Bulling, "Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 11, pp. 7375–7385, 2024.
- [3] Z. Xu, Z. Huang, J. Cao, Y. Zhang, X. Cun, Q. Shuai, Y. Wang, L. Bao, J. Li, and F. Tang, "Anchorcrafter: Animate cyberanchors saling your products via human-object interacting video generation," *arXiv preprint arXiv:2411.17383*, 2024.
- [4] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang *et al.*, "Make-your-video: Customized video generation using textual and structural guidance," *IEEE transactions on visualization and computer graphics*, vol. 31, no. 2, pp. 1526–1541, 2024.
- [5] Y. Zhang, W. Dong, F. Tang, N. Huang, H. Huang, C. Ma, P. Wan, T.-Y. Lee, and C. Xu, "Motioncrafter: Plug-and-play motion guidance for diffusion models," *IEEE Transactions on Visualization and Computer Graphics*, 2025.
- [6] Z. Huang, Z. Zhou, J. Cao, Y. Ma, Y. Chen, Z. Rao, Z. Xu, H. Wang, Q. Lin, Y. Zhou *et al.*, "Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation," *arXiv preprint arXiv:2506.08797*, 2025.
- [7] Z. S. Xue, R. Luo, C. Chen, and K. Grauman, "Hoi-swap: Swapping objects in videos with hand-object interaction awareness," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 77 132–77 164, 2024.
- [8] Z. Xu, Z. Rao, J. Cao, X. Liu, Z. Fang, H. Zhang, S. Tang, and F. Tang, "Geohoi: Geometry-enhanced human-object interaction video generation via hierarchical multi-modal injection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 3739–3748.
- [9] Y. Zhang, Z. Zhou, Z. Yu, Z. Huang, T. Hu, S. Liang, G. Zhang, Z. Peng, S. Li, Y. Chen *et al.*, "Making avatars interact: Towards text-driven human-object interaction for controllable talking avatars," *arXiv preprint arXiv:2602.01538*, 2026.
- [10] T. Yin, Q. Zhang, R. Zhang, W. T. Freeman, F. Durand, E. Shechtman, and X. Huang, "From slow bidirectional to fast autoregressive video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 22 963–22 974.
- [11] X. Huang, Z. Li, G. He, M. Zhou, and E. Shechtman, "Self forcing: Bridging the train-test gap in autoregressive video diffusion," *arXiv preprint arXiv:2506.08009*, 2025.
- [12] H. Zhu, M. Zhao, G. He, H. Su, C. Li, and J. Zhu, "Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation," *arXiv preprint arXiv:2602.02214*, 2026.
- [13] S. Yang, W. Huang, R. Chu, Y. Xiao, Y. Zhao, X. Wang, M. Li, E. Xie, Y. Chen, Y. Lu, and S. H. Y. Chen, "Longlive: Real-time interactive long video generation," 2025.
- [14] Z. Huang, F. Tang, Y. Zhang, X. Cun, J. Cao, J. Li, and T.-Y. Lee, "Make-your-anchor: A diffusion-based 2d avatar generation framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6997–7006.
- [15] Z. Huang, F. Tang, J. Cao, Y. Zhang, X. Cun, Y. Bo, J. Li, and T.-Y. Lee, "Make-your-anchor+: Temporal consistent 2d avatar generation via video diffusion prior," *IEEE Transactions on Visualization & Computer Graphics*, vol. 32, no. 04, pp. 2883–2897, Apr. 2026. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/TVCG.2026.3655478>
- [16] L. Hu, G. Wang, Z. Shen, X. Gao, D. Meng, L. Zhuo, P. Zhang, B. Zhang, and L. Bo, "Animate anyone 2: High-fidelity character image animation with environment affordance," *arXiv preprint arXiv:2502.06145*, 2025.
- [17] Y. Men, Y. Yao, M. Cui, and L. Bo, "Mimo: Controllable character video synthesis with spatial decomposed modeling," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 21 181–21 191.
- [18] Y. Fan, Q. Yang, K. Wang, H. Zhou, Y. Li, H. Feng, E. Ding, Y. Wu, and J. Wang, "Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 17 550–17 560.
- [19] Z. Shen, C. Wu, J. Zhou, C. Zhao, K. Wang, H. Zhou, Y. Li, H. Feng, W. He, and J. Wang, "idit-hoi: Inpainting-based hand object interaction reenactment via video diffusion transformer," *arXiv preprint arXiv:2506.12847*, 2025.
- [20] L. Dang, R. Shao, H. Zhang, W. Min, Y. Liu, and Q. Wu, "Svimo: Synchronized diffusion for video and motion generation in hand-object interaction scenarios," *arXiv preprint arXiv:2506.02444*, 2025.
- [21] L. Dang, J. Li, Z. Li, H. Zhang, L. An, W. Min, Y. Liu, and Q. Wu, "Harmohoi: Harmonizing appearance and 3d motion

- for multi-view hand-object interaction synthesis," 2026. [Online]. Available: <https://arxiv.org/abs/2607.17097>
- [22] L. Wang, Z. Xia, T. Hu, P. Wang, P. Wei, Z. Zheng, M. Zhou, Y. Zhang, and M. Gao, "Dreamactor-h1: High-fidelity human-product demonstration video generation via motion-designed diffusion transformers," *arXiv preprint arXiv:2506.10568*, 2025.
- [23] Y. Jin, Z. Sun, N. Li, K. Xu, H. Jiang, N. Zhuang, Q. Huang, Y. Song, Y. Mu, and Z. Lin, "Pyramidal flow matching for efficient video generative modeling," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 23 378–23 402.
- [24] B. Chen, D. Martí Monsó, Y. Du, M. Simchowitz, R. Tedrake, and V. Sitzmann, "Diffusion forcing: Next-token prediction meets full-sequence diffusion," *Advances in Neural Information Processing Systems*, vol. 37, pp. 24 081–24 125, 2024.
- [25] K. Song, B. Chen, M. Simchowitz, Y. Du, R. Tedrake, and V. Sitzmann, "History-guided video diffusion," *arXiv preprint arXiv:2502.06764*, 2025.
- [26] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," 2023.
- [27] Y. Song and P. Dhariwal, "Improved techniques for training consistency models," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 15 078–15 097.
- [28] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *Advances in neural information processing systems*, vol. 36, pp. 8406–8441, 2023.
- [29] W. Luo, T. Hu, S. Zhang, J. Sun, Z. Li, and Z. Zhang, "Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 525–76 546, 2023.
- [30] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 6613–6623.
- [31] L. Zhang, S. Cai, M. Li, G. Wetzstein, and M. Agrawala, "Frame context packing and drift prevention in next-frame-prediction video diffusion models," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [32] K. Liu, W. Hu, J. Xu, Y. Shan, and S. Lu, "Rolling forcing: Autoregressive long video diffusion in real time," *arXiv preprint arXiv:2509.25161*, 2025.
- [33] J. Cui, J. Wu, M. Li, T. Yang, X. Li, R. Wang, A. Bai, Y. Ban, and C.-J. Hsieh, "Self-forcing++: Towards minute-scale high-quality video generation," in *The Fourteenth International Conference on Learning Representations*.
- [34] J. Yu, J. Bai, Y. Qin, Q. Liu, X. Wang, P. Wan, D. Zhang, and X. Liu, "Context as memory: Scene-consistent interactive long video generation with memory retrieval," *arXiv preprint arXiv:2506.03141*, 2025.
- [35] J. Huang, X. Hu, B. Han, S. Shi, Z. Tian, T. He, and L. Jiang, "Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft," *arXiv preprint arXiv:2510.03198*, 2025.
- [36] S. Ji, X. Chen, S. Yang, X. Tao, P. Wan, and H. Zhao, "Memflow: Flowing adaptive memory for consistent and efficient long video narratives," *arXiv preprint arXiv:2512.14699*, 2025.
- [37] S. Chen, C. Wei, S. Sun, P. Nie, K. Zhou, G. Zhang, M.-H. Yang, and W. Chen, "Context forcing: Consistent autoregressive video generation with long context," *arXiv preprint arXiv:2602.06028*, 2026.
- [38] H. Yesiltepe, T. Meral, A. K. Akan, K. Oktay, and P. Yarnadag, "Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 40 256–40 265.
- [39] J. Yi, W. Jang, P. H. Cho, J. Nam, H. Yoon, and S. Kim, "Deep forcing: Training-free long video generation with deep sink and participative compression," *arXiv preprint arXiv:2512.05081*, 2025.
- [40] J. Tian, Y. Wang, G. Yu, and C. Zhang, "Head forcing: Long autoregressive video generation via head heterogeneity," in *arXiv preprint*, 2026.
- [41] K. Liu, Q. Liu, X. Liu, J. Li, Y. Zhang, J. Luo, X. He, and W. Liu, "Hoigen-1m: A large-scale dataset for human-object interaction video generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 001–24 010.
- [42] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.
- [43] M. L. Team, X. Cai, M. Cheng, F. Gao, Z. Kong, J. Li, L. Li, W. Li, H. Liu, S. Tan *et al.*, "Longcat-video-avatar 1.5 technical report," *arXiv preprint arXiv:2605.26486*, 2026.
- [44] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.
- [45] X. Wang, S. Zhang, L. Tang, Y. Zhang, C. Gao, Y. Wang, and N. Sang, "Unianimate-dit: Human image animation with large-scale video diffusion transformer," *arXiv preprint arXiv:2504.11289*, 2025.
- [46] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu, "Vace: All-in-one video creation and editing," *arXiv preprint arXiv:2503.07598*, 2025.
- [47] L. Chen, T. Ma, J. Liu, B. Li, Z. Chen, L. Liu, X. He, G. Li, Q. He, and Z. Wu, "Humo: Human-centric video generation via collaborative multi-modal conditioning," *arXiv preprint arXiv:2509.08519*, 2025.
- [48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
- [49] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.
- [50] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 21 807–21 818.
- [51] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.