SAX: Advancing Video Diffusion Models for Sequential Action Execution

Haoyu Wang[1], **Baorui Ma**[2], **Donglin Di**[2], **Shiliang Zhang**[1,*]
[1] State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University
[2]Li Auto
haoyu.wang@stu.pku.edu.cn, mabaorui2014@gmail.com,
donglin.ddl@gmail.com, slzhang.jdl@pku.edu.cn

Abstract

Video diffusion models have recently achieved remarkable visual fidelity. However, when handling complex instructions with multi-action sequential prompts, existing models frequently suffer from severe action omission or state freezing. We attribute this bottleneck to two fundamental flaws: i) semantic entanglement, where appearance and motion concepts are heavily intertwined within the text embedding space; and ii) temporal attention leakage, wherein the global nature of cross-attention along the temporal dimension biases the model towards the most visually salient action tokens throughout the entire video. To break this bottleneck, we propose SAX, a novel diffusion transformer framework designed for the synergistic generation of appearance and dynamic actions. First, we introduce a dual-stream attention mechanism that explicitly decouples dynamic action interactions from the static appearance generation process. Besides, to establish a precise temporal-action mapping, we design the Layer-adaptive Action Choreographer. By integrating a shared Global Temporal Table and an Action Ordinal Table, this module dynamically generates a temporal-action alignment map for each layer. Furthermore, to overcome the implicit convergence difficulties inherent in temporal-action alignment optimization, we innovatively employ a Vision-Language Model as a temporal-semantic supervisor. Through distribution distillation, we directly inject the VLM's powerful temporal-action grounding priors into the learning of the alignment maps. Extensive experiments on action generation and prediction demonstrate that SAX substantially outperforms existing models across core metrics.

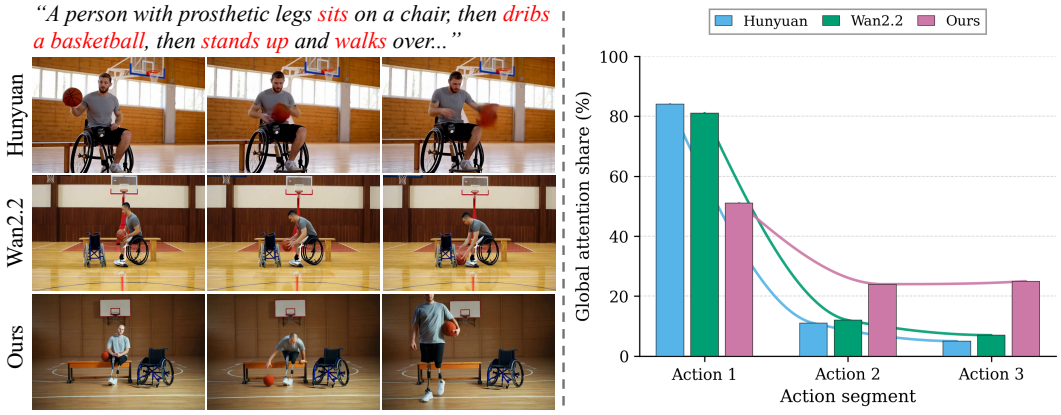
1 Introduction

Video generation and prediction based on diffusion models [1, 2, 3, 4, 5, 6] has recently garnered substantial research interest due to its broad application potentials and unprecedented achievements in visual fidelity. However, a core challenge lies in accurately executing complex user instructions, particularly for multi-action sequential prompts, *e.g.* "A person with prosthetic legs sits on a chair, then dribs a basketball, then stands up and walks over", while preserving the appearance consistency of the generated videos. While existing models excel at depicting static scenes or single short-term motions, they frequently struggle to maintain temporal and semantic coherence across multiple sequential actions [7], often suffering from severe action omission or state freezing during the dynamic generation process, as shown in Fig. 1(a).

To understand the root causes of these failures, we delve into the internal mechanisms of current Diffusion Transformers (DiTs) [8], and attribute this temporal-semantic degradation primarily to

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Visual comparison with existing models

(b) Analysis of global attention share rate

Figure 1: (a) Visual results of existing video diffusion models in multi-action execution scenarios. (b) Statistical attention distribution across different models, revealing a systematic bias where the primary action hoards the vast majority of cross-attention capacity, suppressing subsequent dynamics.

two fundamental bottlenecks. First, semantic entanglement arises within the text feature space. State-of-the-art text encoders [9, 10] compress both static appearance attributes (*e.g.* "prosthetic legs") and dynamic motion concepts (*e.g.* "sits", "walks") into a shared, entangled embedding space. During the text-to-video interaction, the inherently more stable appearance features tend to dominate the representation, thereby severely diluting the dynamic semantics required to drive sequential actions. More critically, existing architectures suffer from temporal attention leakage within the standard cross-attention mechanism [11, 12, 13]. By default, the visual tokens of every generated frame globally query the entire text prompt. Devoid of explicit temporal constraints, this unconstrained routing inevitably triggers an attention collapse phenomenon: all frames across the timeline disproportionately fixate on the most visually salient action tokens. To quantitatively validate this hypothesis, we conduct a probing experiment with fifty multi-action prompts on existing models including HunyuanVideo [3], and Wan2.2 [4]. As illustrated in Figure 1(b), the temporal cross-attention map of current models exhibits severe leakage: across all generated frames, the model disproportionately fixates on the most visually salient action token (*e.g.* Action 1), maintaining a persistently high attention weight throughout the video. In contrast, subsequent actions receive suppressed. Consequently, the temporal progression of the model is effectively paralyzed, hindering its ability to transition to subsequent motions prescribed by the prompt.

Motivated by these critical observations, we present SAX, a novel synergistic DiT framework designed to break this temporal-semantic bottleneck. Our core philosophy is to preserve the exceptional appearance-rendering capabilities of the foundational video model while equipping it with an effective action generation engine. Specifically, SAX achieves this through two pivotal architectural interventions. First, we propose the dual-stream appearance-semantic decoupling [14, 15]. To prevent motion semantics from being overshadowed by appearance features, we introduce a parallel action attention within each Transformer block. While the original cross-attention mechanism focuses exclusively on static appearance prompts, the parsed dynamic action instructions are routed through this dedicated pathway. This explicit isolation ensures no interference between static and dynamic features before their synergistic fusion. Moreover, we introduce the explicit temporal-action alignment within action attention to alleviate temporal attention leakage. We abandon the unreliable paradigm of letting the model blindly infer temporal correspondences. Instead, we propose a Layer-adaptive Action Choreographer. By integrating the output from the shared Global Temporal Table and an Action Ordinal Table, the lightweight network dynamically predicts a temporal-action alignment map at each layer. By directly injecting this temporal-action alignment map as a deterministic prior into the attention logits, we enforce a strict temporal constraint, *i.e.* visual tokens at a specific video frame are compelled to attend only to their chronologically corresponding action tokens.

While the proposed architectural design provides a rigorous structural constraint, optimizing this dynamic temporal-action alignment map from scratch remains a formidable challenge. Existing work [7] shows that relying solely on the vanilla reconstruction-based diffusion objective often leads

to an overemphasis on appearance fidelity, where the precise time-to-action correspondences are never fully realized. To elegantly bypass this optimization bottleneck, we leverage the pre-trained knowledge of large-scale Video-Language Models (VLMs) [16]. Having been comprehensively trained on massive text-video datasets, these VLMs naturally encapsulate profound temporal grounding capabilities; their internal attention matrices intrinsically encode fine-grained knowledge of exactly which action unfolds at which specific second. Capitalizing on this insight, we introduce an innovative Temporal-Action Alignment Distillation paradigm. Specifically, we repurpose a frozen VLM as a temporal-semantic supervisor, extracting its precise attention distributions to serve as soft-label ground truths. By optimizing the Kullback-Leibler (KL) divergence, we directly distill these rich localization priors into our Layer-adaptive Action Choreographer. This strategic knowledge transfer elegantly resolves the ambiguity of unguided temporal-action alignment, resulting in highly efficient training convergence and exceptionally precise multi-action execution.

In summary, our key contributions lie in three aspects: i) We propose SAX, a novel Diffusion Transformer framework tailored for complex multi-action video generation and prediction. By systematically integrating a dual-stream decoupling mechanism and a Layer-adaptive Action Choreographer, SAX alleviate temporal attention leakage and achieves explicit alignment between sequential instructions and visual motions. ii) We introduce a effective Temporal-Action Alignment Distillation paradigm. By leveraging a frozen VLM as a temporal-semantic supervisor, we distill its precise cross-modal attention priors into our network. This strategy elegantly overcomes the implicit convergence bottlenecks traditionally associated with optimizing dynamic Temporal-Action alignment maps from scratch. iii) Extensive experiments across challenging multi-action video generation and prediction tasks demonstrate the superiority of our approach.

2 Related Work

2.1 Video Generation

Early diffusion-based video generation models [12, 17, 18, 19, 20] predominantly rely on UNet architectures, which typically inflate pre-trained 2D image diffusion models [11, 21] by inserting temporal attention or 3D convolutions to capture motion dynamics. For instance, Stable Video Diffusion (SVD) [12] and VideoCrafter introduce robust temporal layers and spatial-temporal cross-attention mechanisms to enable high-fidelity video synthesis and robust text-to-video generation. These UNet-based approaches established a strong foundation for generating visually coherent short video clips by learning temporal priors from large-scale video datasets. Recently, the paradigm has significantly shifted towards Diffusion Transformer (DiT) architectures [8, 22, 23], which exhibit superior scalability and a stronger capacity for processing complex spatio-temporal tokens. Building upon this foundation, models such as CogVideoX [1, 2] employ 3D causal VAEs [24] and expert transformer blocks to enhance text-video alignment and motion smoothness. Furthermore, DiT-based models like HunyuanVideo [3] and Wan [4] have pushed the boundaries of resolution, duration, and visual quality by leveraging large-scale unified spatio-temporal attention mechanisms and refined latent space representations.

However, despite achieving remarkable visual fidelity, the aforementioned methods frequently struggle when processing complex instructions that contain multi-action sequence prompts. Specifically, they often suffer from severe action omission or state freezing during generation. To address these challenges, we propose a novel diffusion transformer framework designed to synergistically generate appearance and dynamic motion. By introducing a dual-stream attention mechanism to explicitly decouple dynamic action interactions from static appearance generation, alongside a layer-adaptive action orchestrator and VLM-distilled temporal semantic supervision, our method establishes precise temporal-action mappings and overcomes the multi-action generation bottleneck of existing models.

2.2 Video Prediction

Beyond standard text-to-video synthesis, forecasting future video states given an initial visual context, often referred to as video prediction [25, 26, 5, 6, 27, 28, 29, 30], has emerged as a crucial frontier for developing comprehensive physical world models. Recent advancements have begun integrating autoregressive paradigms with diffusion processes to tackle this challenge. For instance, VideoGPT [6] introduces a next clip diffusion framework that treats video clips as a new form of visual language. By autoregressively denoising future clips conditioned on clean historical frames, it

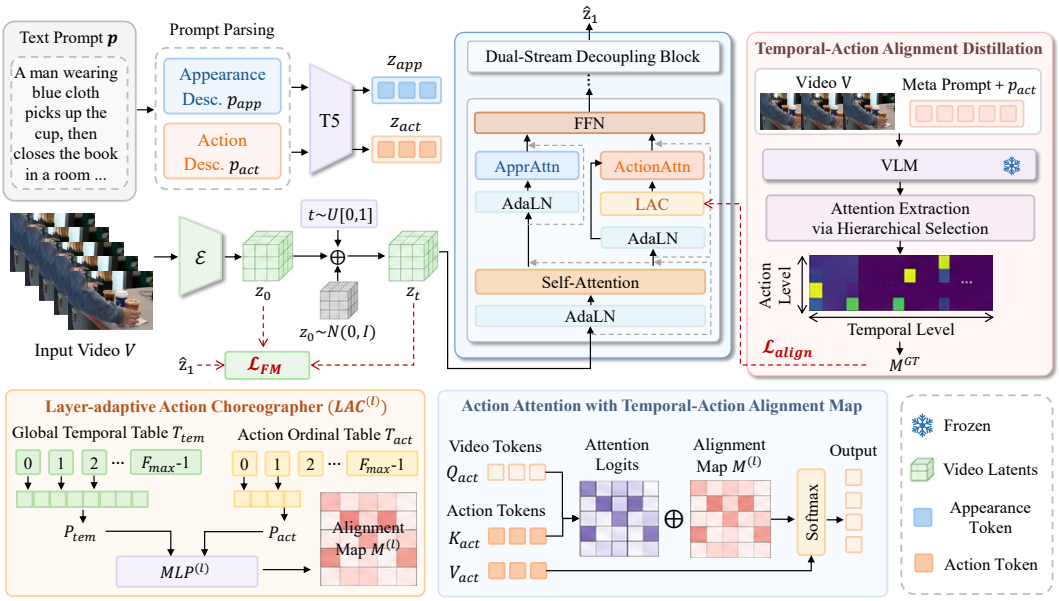


Figure 2: Overview of the proposed method, which employ a diffusion transformer framework for jointly modeling appearance and dynamic motion in complex video generation and prediction.

effectively bridges language modeling principles with continuous video generation, demonstrating strong capabilities in long-term prediction and physical world simulation. Furthermore, the scope of video prediction has expanded from raw frame continuation to instruction-driven visual reasoning. This shift is exemplified by the video-next-event prediction paradigm [30, 5], which demands dynamic video responses to procedural or predictive queries rather than simple text answers. To achieve this, recent methods such as the Video-as-Answer [5] framework employ a joint reinforcement learning strategy to orchestrate a vision-language model and a video diffusion model as a unified system. By co-optimizing these models via shared rewards, it ensures that the predicted future events are logically deduced from the multimodal context and rendered with high visual and semantic consistency.

However, because these video prediction models either utilize existing video generation models as their foundational backbone [5, 30] or adopt highly similar model architectures and training paradigms [27, 26], they inevitably inherit the same limitations. Consequently, when tasked with predicting complex dynamic actions or executing multi-step visual reasoning, they frequently suffer from identical issues of severe action omission and state freezing. This shared bottleneck across both video generation and prediction tasks underscores the critical need for a more dedicated framework capable of explicitly disentangling and aligning complex spatio-temporal dynamics.

3 Methodology

In this section, we present the proposed SAX, a novel Diffusion Transformer (DiT) framework tailored for multi-action video generation, as shown in Fig. 2. We first formulate the background of DiT-based video generation and provide an overview of our framework (Sec. 3.1). Subsequently, we detail the Layer-adaptive Action Choreographer for explicit temporal-action alignment (Sec. 3.2), followed by the Temporal-Action Alignment Distillation paradigm via a Vision-Language Model (VLM) supervisor (Sec. 3.3).

3.1 Overview

Preliminaries on Video Generation. The remarkable empirical success and intrinsic scalability of Diffusion Transformers [8] have positioned them at the forefront of contemporary video generation research. Aligning with established methodologies [2, 4, 3], we employ a DiT framework as our core generative engine, systematically extending its architecture to facilitate the precise execution of multi-action sequences.

Formally, a training instance comprises a textual prompt p and an associated video sequence V. We utilize a pre-trained T5 text encoder [9] to derive the contextualized embeddings $z_p \in \mathbb{R}^{B\times L\times D}$, where B, L, and D define the batch size, sequence length, and embedding dimensionality, respectively. Concurrently, the visual input V is projected into a lower-dimensional manifold via a VAE encoder $\mathcal{E}$, yielding the pristine data latent $z_1 \in \mathbb{R}^{B\times F\times C\times H\times W}$, distributed across temporal (F), channel (C), and spatial (H, W) axes. Under the continuous-time flow matching paradigm, we construct a probability flow connecting a tractable prior distribution, typically standard Gaussian noise $z_0 \sim \mathcal{N}(0, \mathbf{I})$, to the target data distribution z_1. Specifically, we define a linear interpolation schedule to formulate the intermediate state at continuous time $t \in [0, 1]$ as $z_t = tz_1 + (1 - t)z_0$. Taking both the conditioning signal z_p and the intermediate latent z_t as inputs, the network $\mathcal{M}_\theta$ is optimized to estimate the velocity vector field governing this transformation. The training objective focuses on matching the predicted velocity with the ground-truth trajectory derivative $z_1 - z_0$, formally defined as:

$$\mathcal{L}_{FM} = \mathbb{E}_{z_0,z_1,p,t} \left\| \mathcal{M}_\theta(z_t, z_p, t) - (z_1 - z_0) \right\|_2^2, \tag{1}$$

where $t \sim \mathcal{U}[0, 1]$ is a uniformly sampled time step. During inference, we sample an initial noise tensor $z_0 \sim \mathcal{N}(0, \mathbf{I})$ and simulate the empirical probability flow ordinary differential equation (ODE) from $t = 0$ to $t = 1$ using numerical ODE solvers, continuously conditioned on the text embeddings z_p. The resulting integrated latent z_1 is subsequently mapped back to the pixel space by the VAE decoder $\mathcal{D}$ to generate the final output video $V^{'}$.

Overview of the SAX Framework. In vanilla cross-attention mechanisms, visual tokens across all frames globally query the unified text embedding z_p, which inevitably leads to semantic entanglement and temporal attention leakage. To address this, we parse the prompt p into a static appearance description p_{app} and a sequence of dynamic action instructions $p_{act} = \{a_1, a_2, \ldots, a_N\}$, yielding respective embeddings z_{app} and z_{act}.

Building upon the vanilla DiT block, we propose a dual-stream appearance-semantic decoupling architecture. To prevent action semantics from being overshadowed by appearance features, we introduce a parallel action attention module within each Transformer block. While the original cross-attention mechanism focuses exclusively on the static appearance prompt z_{app}, the parsed dynamic action instructions z_{act} are routed through this dedicated pathway. This explicit isolation ensures no interference between static and dynamic features prior to their synergistic fusion. Formally, for the l-th DiT block, the hidden video states $h^{(l)}$ are updated as:

$$h_{out}^{(l)} = \text{FFN}\left(h^{(l)} + \text{ApprAttn}(h^{(l)}, z_{app}) + \text{ActionAttn}(h^{(l)}, z_{act}, M^{(l)})\right), \tag{2}$$

where ApprAttn represents the appearance attention module, ApprAttn represents the action attention module, and $M^{(l)} \in \mathbb{R}^{(F\times H\times W)\times N}$ serves as a deterministic temporal-action alignment map specifically tailored for the l-th layer.

To construct $M^{(l)}$, we design the Layer-adaptive Action Choreographer (LAC). By integrating the temporal embeddings P_{tem} from a shared Global Temporal Table (representing frame indices) and the action embeddings P_{act} from a shared Action Ordinal Table (representing the sequential order of actions), a lightweight network dynamically predicts the alignment map. This map acts as a spatial-temporal bias injected into the attention logits of the ActionAttn module, mathematically compelling specific visual frames to attend only to their chronologically corresponding action tokens.

Finally, since optimizing the dynamic alignment map $M^{(l)}$ from scratch solely via $\mathcal{L}_{FM}$ leads to implicit convergence difficulties, we introduce the Temporal-Action Alignment Distillation paradigm. We leverage a frozen VLM as a temporal-semantic supervisor. By extracting its precise cross-modal attention distributions as soft-label ground truths, we optimize the Action Choreographer via a Kullback-Leibler (KL) divergence loss, ensuring highly efficient convergence and accurate spatiotemporal grounding. Details of the Layer-adaptive Action Choreographer and this distillation paradigm are elaborated in Sec. 3.2 and Sec. 3.3, respectively.

3.2 Layer-adaptive Action Choreographer

To thoroughly eradicate temporal attention leakage and address the semantic entanglement highlighted in Sec. 1, we introduce the Layer-adaptive Action Choreographer. This module is embedded within the action attention mechanism, which operates in parallel with the appearance attention block. It

explicitly regulates the temporal execution of multiple actions through a dynamically computed temporal-action alignment map.

3.2.1 Global Temporal and Action Ordinal Tables

To construct a deterministic mapping between physical video time and the logical action sequence, we define two globally shared, learnable embedding tables across all DiT blocks: i) Global Temporal Table $\mathcal{T}_{tem} \in \mathbb{R}^{F_{max} \times D}$ that encodes the physical frame indices. For a video latent with F frames, we extract $P_{tem} = \mathcal{T}_{tem}(t_{idx}) \in \mathbb{R}^{F \times D}$, where $t_{idx} \in \{0, 1, \ldots, F-1\}$. ii) Action Ordinal Table $\mathcal{T}_{act} \in \mathbb{R}^{L_{max} \times D}$ that encodes the sequence order of the valid action tokens. Supposing the parsed action prompt z_{act} contains N valid tokens, we obtain $P_{act} = \mathcal{T}_{act}(n_{idx}) \in \mathbb{R}^{N \times D}$, where $n_{idx} \in \{0, 1, \ldots, N-1\}$.

A naive approach to temporal-action alignment would be directly optimizing a static layer-wise map parameter $\mathbf{M} \in \mathbb{R}^{F \times N}$. However, we deliberately abandon this in favor of our global tables coupled with lightweight MLPs. The motivations are twofold: i) Flexibility for Variable Lengths: The number of valid action tokens (N) dynamically fluctuates across different user prompts. A static map with fixed dimensions inherently fails to accommodate variable-length instructions seamlessly. ii) Generalization: A static bias map acts as a mere lookup table, essentially memorizing dataset biases (*e.g.*, blindly memorizing that "the 10-th frame typically corresponds to the 1st action"). It lacks a fundamental understanding of continuous time and sequential progression. In contrast, by learning the topological mappings of P_{tem} and P_{act} through an MLP, the model captures the generalized relative rules, *e.g. "as the action ordinal increases, its corresponding temporal activation peak should smoothly shift backward."* Armed with this inductive bias, even if the model is primarily trained on sequences of two actions, it can effortlessly generalize to execute more actions during inference, adaptively distributing temporal attention without catastrophic failure.

3.2.2 Layer-adaptive Temporal-Action Alignment Map

Within the l-th DiT block, we first spatially expand the temporal canvas P_{tem} to match the spatial resolution of the visual tokens. Since all spatial pixels (h, w) within the same frame t share identical temporal characteristics, P_{tem} is broadcasted to $P_{t_expand} \in \mathbb{R}^{(F \times H \times W) \times D}$.

Subsequently, a layer-specific lightweight $\text{MLP}^{(l)}$ computes the matching score between every visual spatial-temporal coordinate and every action ordinal token, yielding the precise alignment map $M^{(l)}$:

$$M^{(l)} = \text{MLP}^{(l)}\left(P_{t_expand}, P_{act}\right) \in \mathbb{R}^{(F \times H \times W) \times N}. \tag{3}$$

Crucially, $M^{(l)}_{i,j}$ mathematically signifies the explicit degree of attention that the i-th visual token (spanning time and space) is compelled to allocate to the j-th semantic action token.

To integrate this choreographed bias into the cross-attention calculation, we compute the standard attention logits between the visual queries Q_{act} and the action text keys K_{act}. The final modulated attention scores are formulated as:

$$\mathbf{S}_{final} = \frac{Q_{act} K_{act}^{\top}}{\sqrt{d}} + M^{(l)}, \tag{4}$$

where d is the dimension of the attention head. Finally, the Softmax activation is applied to aggregate the value features V_{act}:

$$\text{ActionAttn}_{out} = \text{Softmax}(\mathbf{S}_{final}) \cdot V_{act}. \tag{5}$$

Through this deterministic addition, the Action Choreographer acts as a rigid mask. It alleviates the temporal leakage issue by forcing the attention mechanism to drop visually salient but temporally irrelevant tokens, guaranteeing that the multi-action prompt is executed accurately over the physical video timeline.

3.3 Temporal-Action Alignment Distillation

While the Layer-adaptive Action Choreographer provides a rigorous structure to mitigate temporal leakage, optimizing the dynamic bias map $M^{(l)}$ from scratch solely via the standard diffusion loss $\mathcal{L}_{FM}$ is notoriously unstable, trapping itself in suboptimal local minima without learning the precise

time-to-action correspondences. To alleviate this, we propose the Temporal-Action Alignment Distillation paradigm. We capitalize on the profound temporal grounding capabilities of pre-trained VLMs [16], repurposing them as temporal-semantic supervisors to provide soft-label ground truths for our choreographer.

3.3.1 Hierarchical Extraction Mechanism

To extract the localization prior, we concatenate a task-specific meta-prompt (*"Describe the temporal sequence of actions in the video"*) with the parsed p_{act} as a unified text input to a frozen VLM. During a single forward pass over the corresponding video, the VLM computes a complete causal attention matrix across all layers and heads. Let $\mathbf{A} \in \mathbb{R}^{L\times H\times N_{all}\times N_{all}}$ denote the raw attention weights. We specifically slice the cross-attention matrix $\mathbf{A}_{cross} \in \mathbb{R}^{L\times H\times N\times M}$ that captures the interaction between the target action tokens (indexed by $\mathcal{I}_{act}$, size N) and the visual video tokens (indexed by $\mathcal{I}_{vid}$, size $M = F_v \times H_v \times W_v$).

Modern VLMs consist of dozens of layers and heads, many of which are responsible for internal text-text syntactic reasoning rather than vision-language alignment. To automatically identify the optimal parts that genuinely bridges the temporal semantics, we design a hierarchical extraction mechanism based on two physically-motivated criteria:

i) Layer-level Cross-Modal Engagement: If a layer allocates the vast majority of its attention weights to the video tokens rather than preceding text context or punctuation, it signifies that this layer is specialized in cross-modal feature fetching. We calculate the average video attention ratio for each layer and head:

$$\mathcal{R}_{layer,head} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{A}_{cross}[layer, head, i, j]. \tag{6}$$

We filter out candidate layers where $\mathcal{R}$ falls below a threshold, ensuring the selected layers are actively observing the video.

ii) Head-level Temporal Sharpness: Among the cross-modal layers, an excellent alignment head should not uniformly average all video frames, which equates to global average pooling. Instead, it should act as a spotlight, precisely focusing on the specific spatiotemporal patches where an action unfolds. We quantify this property using information entropy. We first normalize the attention over the video dimension to form a valid distribution $\hat{\mathbf{A}}$:

$$\hat{\mathbf{A}}[i,j] = \frac{\mathbf{A}_{cross}[i,j]}{\sum_{k=1}^{M}\mathbf{A}_{cross}[i,k] + \epsilon}, \tag{7}$$

where ϵ is a small scalar for numerical stability. The entropy H and the subsequent sharpness score $\mathcal{S}_{sharp}$ are defined as:

$$H_{head,i} = -\sum_{j=1}^{M}\hat{\mathbf{A}}[i,j]\log(\hat{\mathbf{A}}[i,j] + \epsilon), \quad \mathcal{S}_{sharp} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{H_{head,i} + \epsilon}\right). \tag{8}$$

The heads that maximize $\mathcal{S}_{sharp}$ within the high-$\mathcal{R}_{video}$ layers are selected as the optimal heads. Their average attention map is transposed and adopted as the target map $\mathbf{M}^{GT} \in \mathbb{R}^{M\times N}$.

3.3.2 Distribution Distillation via KL Divergence

With the optimal map $\mathbf{M}^{GT}$ obtained, we formulate the distillation process as a probability distribution matching task. Let $M^{(l)}$ be the alignment map predicted by our LAC. We convert $M^{(l)}$ into a log-probability space along the action token dimension, yielding $\log \mathbf{P}^{(l)} = \text{LogSoftmax}(M^{(l)})$.

Simultaneously, we ensure the target $\mathbf{M}^{GT}$ is strictly normalized into a valid probability distribution $\mathbf{Q}$ over the action dimension to prevent numerical instability caused by zero-regions:

$$\mathbf{Q}[j,i] = \frac{\mathbf{M}^{GT}[j,i]}{\sum_{k=1}^{N}\mathbf{M}^{GT}[j,k] + \epsilon}. \tag{9}$$

Table 1: Quantitative comparison with existing methods on the action generation task. Lower FVD values indicate better performance, whereas higher values on the other metrics correspond to better results. **Bold** represents the best performance, and underline represents the second-best.

Method	FVD	CLIPSIM	Completion Rate	Subject Consistency	Motion Smoothness	Dynamic Degree	Imaging Quality
ModelScope	1088.87	0.2842	0.3056	94.67	96.91	44.33	62.16
VideoCrafter2	874.87	0.2820	0.3080	94.84	98.70	15.34	64.51
LaVie	653.76	0.2816	0.2508	95.10	98.85	11.01	66.48
CogVideoX	603.89	0.2779	0.4710	95.35	97.44	49.13	68.76
HunyuanVideo	577.69	0.2952	0.5134	95.30	98.27	53.66	68.45
Wan 2.2	539.42	0.2925	0.5237	**96.09**	98.20	54.00	68.44
Ours	**515.05**	**0.2977**	**0.6358**	95.53	**98.91**	**70.33**	**69.87**

Table 2: Quantitative comparison with existing methods on the image-to-video generation task. Lower FVD values indicate better performance, whereas higher values on the other metrics correspond to better results. **Bold** represents the best performance, and underline represents the second-best.

Method	FVD	CLIPSIM	Completion Rate	Subject Consistency	Motion Smoothness	Dynamic Degree	Imaging Quality
CogVideoX	574.41	0.2878	0.4936	93.15	96.32	50.01	66.09
HunyuanVideo	344.67	0.2926	0.5793	93.59	97.68	55.27	65.73
Wan 2.2	317.26	0.2933	0.5875	**94.78**	98.35	53.01	65.92
Ours	**238.03**	**0.2959**	**0.6793**	93.43	**98.89**	**56.67**	**66.69**

Finally, we penalize the discrepancy between the student's temporal-action alignment and the VLM's prior using the KL divergence loss:

$$\mathcal{L}_{align} = \frac{1}{M} \sum_{j=1}^{M} \mathcal{D}_{KL}(\mathbf{Q}_j \parallel \mathbf{P}_j^{(l)}). \tag{10}$$

During training, this distillation loss is optimized jointly with the main diffusion objective: $\mathcal{L}_{total} = \mathcal{L}_{FM} + \lambda \mathcal{L}_{align}$, where λ is a balancing coefficient. This white-box distillation strategy elegantly injects dense temporal grounding priors into the DiT, ensuring rapid convergence and exceptional multi-action sequential control.

4 Experiments

4.1 Datasets and Evaluation Metrics

We used the HumanVid [31] and TASTE-Rob [32] datasets for model training and comparison, which contain a large number of videos with multi-action execution. For the evaluation metrics, Following previous work, we adopt Fréchet Video Distance (FVD) [33] and CLIP similarity (CLIPSIM) [10] to measure the model's performance. Besides, we follow the StoryEval benchmark [34] to evaluate the action completion rate in the generated video. Furthermore, we utilize VBench [35] to conduct a more comprehensive assessment of model performance across multiple dimensions, including subject consistency, motion smoothness, motion dynamics, and visual quality.

4.2 Comparison with Existing Methods

We adopt Wan2.1 [4] as the backbone, and more implementation details are provided in Sec. F of the appendix. We first compare our method with existing video diffusion models on the action generation task, including ModelScope [17], VideoCrafter2 [18], LaVie [20], CogVideoX [2], HunyuanVideo [3], and Wan2.2 [4]. We also conduct a user study for more comprehensive comparison, and details are provided in Sec. B. We also compare our method on the image-to-video generation task. The detailed results and analysis are in Sec. 5 of the appendix.



Figure 3: Visual comparison on the action generation task. For clarity, VideoCrafter2 [18] is represented as VC2. CogvideoX [2] uses version 5B, Wan 2.2 [4] uses version A14B.

Quantitative Comparison. Quantitative comparisons with existing methods on action generation task are shown in Tab. 1. Compared with previous models, our method achieves excellent performance in various metrics, especially in terms of completion rate.

Qualitative Comparison. Visual comparisons are presented in Fig. 3. It can be observed that existing methods often lose certain actions in the action sequence instructions, such as "place another one in the pool". In contrast, our method can perform multi-action generation better.

5 Comparison on Image-to-Video Generation Task

We compared our approach with existing methods on the image-to-video generation task. Specifically, given the current observation (in image form) and the future action sequence instructions, the model predicts the visual outcome of the future action. The results, shown in Tab. 2, demonstrate the superiority of our method.

5.1 Ablation Study

Effect of the Layer-adaptive Action Choreographer. The analysis of the Layer-adaptive Action Choreographer's performance is shown in Tab. 3. The first row represents the process of removing the alignment map generation, and the second row represents directly setting a learnable static $M^{(l)}$. As mentioned earlier, the static $M^{(l)}$ fails to accommodate variable-length instructions seamlessly, resulting in minimal performance gain. In contrast, our approach is the most effective.

Effect of the Temporal-Action Alignment Distillation. The analysis of the component's performance is shown in Tab. 4. The first row represents the result after removing the alignment distillation

Table 3: Effect of the Action Choreographer.

How to Obtain Alignment Map	FVD	Rate
✘	532.14	0.5358
Static Alignment Map	530.26	0.5679
Ours	**515.05**	**0.6358**

Table 4: Effect of alignment distillation.

Distillation Forms	FVD	Rate
✘	531.26	0.5513
Using Average Map	530.45	0.5597
Ours	**515.05**	**0.6358**

Table 5: Analysis for the decoupling framework

Appearance-Action Decoupling	FVD	Rate
✘	530.28	0.5519
Ours	**515.05**	**0.6358**

Table 6: Analysis for the $\mathcal{L}_{align}$'s Form

Different Forms	FVD	Rate
MSE	529.34	0.5962
Ours	**515.05**	**0.6358**

process, and the second row represents the result after directly using the average of the attention maps from all layers as $\mathbf{M}^{GT}$. The results demonstrate the effectiveness of the proposed alignment distillation and hierarchical extraction mechanism.

More Analysis for other Components. We further analyze other components, including the role of the dual-stream decoupling attention mechanism in Tab. 5 and the selection of the alignment constraint form in Tab. 6. Due to page limitations, we conducted a detailed analysis in Sec. C.

6 Conclusion and Future Work

We propose SAX, a novel diffusion transformer framework designed to synergistically generate appearance and dynamic motion for complex video generation and prediction. To address the critical bottlenecks of semantic entanglement and temporal attention leakage, we introduce a dual-stream attention mechanism that explicitly decouples dynamic action interactions from static appearance synthesis. To further establish precise temporal-action mappings, we design a layer-adaptive action orchestrator that integrates a shared global Temporal Table and an action Ordinal Table to dynamically generate layer-wise temporal-action alignment maps. Additionally, to overcome the implicit convergence challenges inherent in optimizing these alignments, we inject the robust temporal-action grounding priors of VLM into the learning process via distribution distillation. While SAX significantly advances the capability of video diffusion models in multi-action execution, we acknowledge a few conservative limitations. Our current framework is primarily tailored for chronological sequential actions. Handling dense, simultaneously overlapping actions with complex spatial interactions remains an open challenge for future spatiotemporal decoupling research.

Acknowledgements

This work was supported in part by National Science and Technology Major Project 2025ZD1601300, in part by Grant 2023-JCJQ-LA-001-088, in part by New Generation Artificial Intelligence-National Science and Technology Major Project 2026ZD0128001, in part by the Natural Science Foundation of China under Grant 62625202.

References

[1] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[2] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

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

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

[5] Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. Video-as-answer: Predict and generate next video event with joint-grpo. *arXiv preprint arXiv:2511.16669*, 2025.

[6] Shaobin Zhuang, Zhipeng Huang, Ying Zhang, Fangyikang Wang, Canmiao Fu, Binxin Yang, Chong Sun, Chen Li, and Yali Wang. Video-gpt via next clip diffusion. *arXiv preprint arXiv:2505.12489*, 2025.

[7] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

[8] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[13] Zhilu Zhang, Haoyu Wang, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Self-supervised high dynamic range imaging with multi-exposure images in dynamic scenes. *arXiv preprint arXiv:2310.01840*, 2023.

[14] Haoyu Wang, Hao Tang, Donglin Di, Zhilu Zhang, Wangmeng Zuo, Feng Gao, Siwei Ma, and Shiliang Zhang. Mosa: Motion-coherent human video generation via structure-appearance decoupling. *arXiv preprint arXiv:2508.17404*, 2025.

[15] Chengfeng Zhao, Jiazhi Shu, Yubo Zhao, Tianyu Huang, Jiahao Lu, Zekai Gu, Chengwei Ren, Zhiyang Dou, Qing Shuai, and Yuan Liu. Comovi: Co-generation of 3d human motions and realistic videos. *arXiv preprint arXiv:2601.10632*, 2026.

[16] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[17] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[18] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024.

[19] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025.

[20] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024.

[21] Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. Mv-vton: Multi-view virtual try-on with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7682–7690, 2025.

[22] Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025.

[23] Yasheng Sun, Zhiliang Xu, Hang Zhou, Jiazhi Guan, Quanwei Yang, Kaisiyuan Wang, Borong Liang, Yingying Li, Haocheng Feng, Jingdong Wang, et al. Cosh-dit: Co-speech gesture video synthesis via hybrid audio-visual diffusion transformers. *arXiv preprint arXiv:2503.09942*, 2025.

[24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.

[26] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3170–3180, 2022.

[27] Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation. *arXiv preprint arXiv:2507.06119*, 2025.

[28] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732*, 2025.

[29] Jingyi Zhang, Ziwei Wang, Haoyu Wang, Jie Zhou, and Jiwen Lu. Anycost network quantization for image super-resolution. *IEEE Transactions on Image Processing*, 33:2279–2292, 2024.

[30] Zhen Xing, Qi Dai, Zejia Weng, Zuxuan Wu, and Yu-Gang Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21243–21253, 2025.

[31] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[32] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27683–27693, 2025.

[33] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[34] Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13629–13638, 2025.

[35] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: reflect these.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: we provide this in appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: No theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we provide it in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be released after it has been organized.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: we provide it in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: not suitable for this work.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: we provide it in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the research conducted in the paper conform, in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: we provide it in appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: no risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: meets requirements.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: no assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: we provide it.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: no risk.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: we provide it in appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix

A More Visual Results

Given the inherent limitations of static frames in depicting continuous sequential dynamics, we highly encourage readers to view the supplementary video (video.mp4). This video provides comprehensive side-by-side comparisons that vividly illustrate the dynamic superiority of our SAX framework. It intuitively demonstrates how our method smoothly executes complex multi-action instructions in strict chronological order, while effectively alleviate the severe action omission and state freezing issues prevalent in current baseline models. We also demonstrate the action generation results in the robotic scene in video.mp4.

B User study

To enable a more comprehensive evaluation, we further conduct a human preference study on the generated videos from different methods. For each text prompt, participants are presented with video samples generated by multiple models and asked to select the most preferred result according to two evaluation criteria: *Action Quality* and *Appearance Quality*. We then compute the preference rate of each method, defined as the proportion of times it is selected as the best-performing result. As shown in Tab. 7, our method achieves the highest preference scores on both Action Quality and Appearance Quality, demonstrating its superior capability in generating realistic appearances while accurately executing complex multi-action sequences.

C More Analysis for other Components

C.1 Analysis for the Dual-Stream Decoupling Framework

Tab. 5 presents an ablation study on the proposed Dual-Stream Decoupling mechanism. The first row corresponds to the baseline configuration, where the original global cross-attention is directly applied without disentangling appearance and motion representations. As discussed in our motivation, existing diffusion-based video generation and prediction models often suffer from semantic entanglement between static appearance attributes and dynamic motion concepts when handling complex multi-action prompts. Such entanglement causes the model to over-focus on dominant visual tokens, leading to action omission, temporal inconsistency, or motion freezing across long action sequences.

In contrast, our Dual-Stream Decoupling mechanism explicitly separates appearance synthesis from action interaction, enabling the model to independently model static visual identity and dynamic temporal evolution. This design effectively alleviates temporal attention leakage and improves the model's ability to preserve coherent multi-action execution throughout the video. As shown in Table 1, introducing the proposed decoupled attention architecture consistently improves both motion-related and visual-quality metrics, demonstrating the effectiveness of our design in achieving accurate temporal-action alignment while maintaining high-fidelity appearance generation.

C.2 Analysis for the Alignment Loss

Tab.6 provides a ablation study on the formulation of the alignment loss. The results demonstrate that the proposed Kullback-Leibler (KL) divergence-based formulation markedly outperforms the conventional Mean Squared Error (MSE) approach. The superiority of KL divergence stems from its ability to treat the map alignment as a distribution-matching problem rather than a point-wise regression task. Since the attention maps extracted from VLM represent a sparse probability distribution over temporal video tokens, the KL Divergence effectively captures the relative importance of action-relevant regions. Conversely, MSE fails to account for the logit-level dynamics of the cross-attention mechanism, often leading to vanishing gradients and blurred temporal boundaries when trying to fit the extremely small absolute values of attention weights.

Table 7: User study with existing video diffusion models.

Method	Action Quality ↑	Appearance Quality ↑
ModelScope [17]	4.35%	1.83%
VC2 [18]	6.64%	3.14%
LaVie [20]	10.47%	11.29%
CogVideoX [2]	12.27%	13.68%
Hunyuan [3]	14.92%	20.85%
Wan 2.2 [4]	15.17%	22.46%
Ours	**36.18%**	**26.75%**

Table 8: Comparison with off-the-shelf baselines on COA metric. Wan2.1* indicates that Wan2.1 is finetuned on our used training sets.

Method	COA ↑
ModelScope [17]	0.3010
VC2 [18]	0.3047
LaVie [20]	0.2496
CogVideoX [2]	0.4690
Hunyuan [3]	0.5094
Wan 2.2 [4]	0.5108
Wan 2.1* [4]	0.5115
Ours	**0.6208**

D Comparison on COA Metric

We introduce the Chronological Order Accuracy (COA) to measure temporal ordering. Following StoryEval, for each generated video and its multi-action prompt, the VLM is queried to answer: i) Did all prompted actions occur. ii) Did they occur in the exact chronological order specified by the prompt, without temporal overlap. A video only scores a Pass if both conditions are strictly met; otherwise, it scores a Fail. The results are presented in the Table R1-1, which will be included in the revised manuscript. As shown in the table, when evaluated on strict chronological ordering, the performance of existing state-of-the-art models drops. In contrast, our SAX framework maintains a higher COA.

E Compare with Wan2.1 finetuned on the HumanVid and TASTE-Rob

We conduct the exact experiment you requested, i.e., Wan2.1 finetuned on the HumanVid and TASTE-Rob datasets with ALL SAX components removed, represented as Wan2.1* . The Table R1-2 presents the disentangled comparison. Please note that the "most-ablated" model you mentioned in Table 2 still retains the Dual-Stream Decoupling mechanism, which contributed to its superiority over the off-the-shelf baseline.

As shown in the table, merely fine-tuning the Wan2.1 on the target datasets provides a marginal improvement. This validates our core motivation stated in the introduction: existing models suffer from semantic entanglement and temporal attention leakage. Simply exposing the model to more multi-action data does not overcome these structural bottlenecks. Moreover, by comparing the Wan2.1 fine-tuned baseline with our full SAX model, we prove that the performance improvement is mainly driven by our architectural designs.

F Implementation Details

We train the model for 10,000 iterations on 8 NVIDIA H20 GPUs with the total batch size of 32. The AdamW [36] optimizer is used with a learning rate of 1e-5. We set the loss weights λ to 1e5, and set the threshold for $\mathcal{R}$ to 0.4.

Table 9: Comparison with Wan2.1 finetuned baseline.

Method	FVD	CLIPSIM	Completion Rate	Subject Consistency	Motion Smoothness	Dynamic Degree	Imaging Quality
Wan 2.1*	536.80	0.2936	0.5242	95.44	98.23	60.17	69.38
Ours	**515.05**	**0.2977**	**0.6358**	**95.53**	**98.91**	**70.33**	**69.87**

G Societal Impacts

Potential Positive Impacts: Advancing controllable, multi-action video generation holds transformative potential for the creative industries. By lowering the barrier to translating complex text scripts into coherent visual narratives, SAX democratizes filmmaking, animation, and game design for independent creators. Furthermore, the ability to generate precise, sequential events can significantly benefit educational and instructional content creation, such as synthesizing step-by-step medical training procedures or DIY tutorials.

Conversely, the enhanced capability to synthesize highly realistic and logically coherent action sequences raises potential risks of misuse. Malicious actors could exploit such technologies to fabricate deceptive events or sophisticated deepfakes, potentially exacerbating the spread of misinformation.

H Declaration of LLM usage

We use Qwen3-VL-32B as the alignment supervisor. Large Language Models (ChatGPT) were used solely to improve the clarity and fluency of the writing. They did not contribute to the research conception, experimental design, data analysis, or interpretation of results.