

SAX: Advancing Video Diffusion Models for Sequential Action Execution

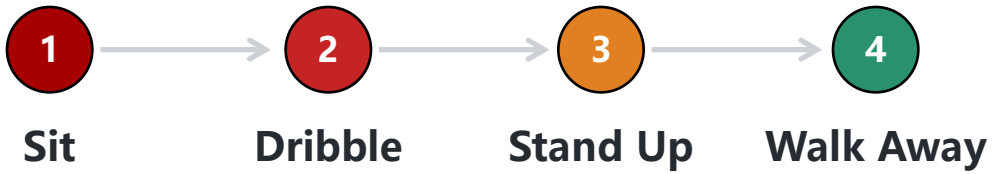
Motivation

北京大学 PEKING UNIVERSITY

Definition

Complex instructions require multiple actions to be completed sequentially in one video while maintaining a consistent subject appearance.

Example: A prosthetic-leg user sits → Dribble → Stand Up → Walk Away



Ideal Action Timeline

Failures of Existing Models

Action Omission
Later steps are not executed

State Freezing
Motion remains in an early state

Temporal Misalignment
Action order does not match the instruction

Research Objective: Achieve high-fidelity appearance and accurate, coherent execution of multi-action sequences

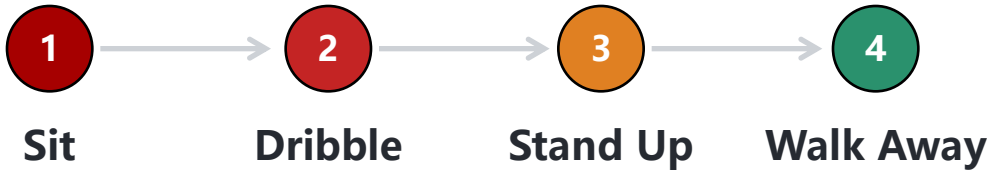
Applications: Scripted video generation · instructional and operational demonstrations · robot action prediction

Motivation

北京大学 PEKING UNIVERSITY

Definition **Complex instructions require multiple actions to be completed sequentially in one video while maintaining a consistent subject appearance.**

Example: A prosthetic-leg user sits → Dribble → Stand Up → Walk Away



Ideal Action Timeline



(a) Visual comparison with existing models

Research Objective: Achieve high-fidelity appearance and accurate, coherent execution of multi-action sequences

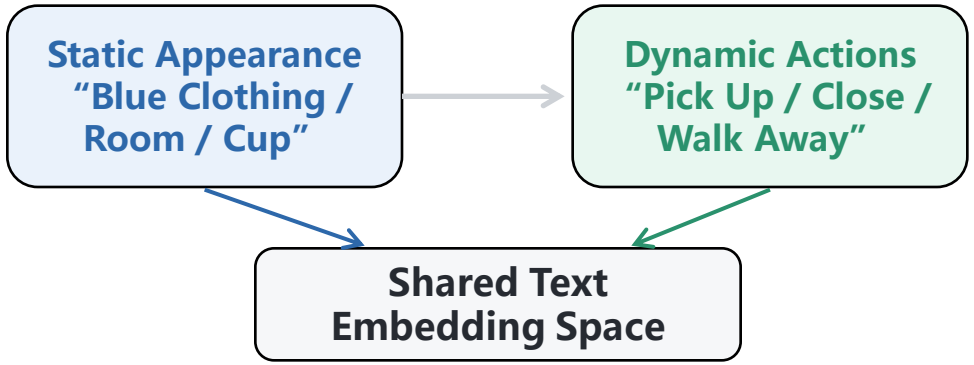
Applications: Scripted video generation · instructional and operational demonstrations · robot action prediction

Motivation

北京大学 PEKING UNIVERSITY

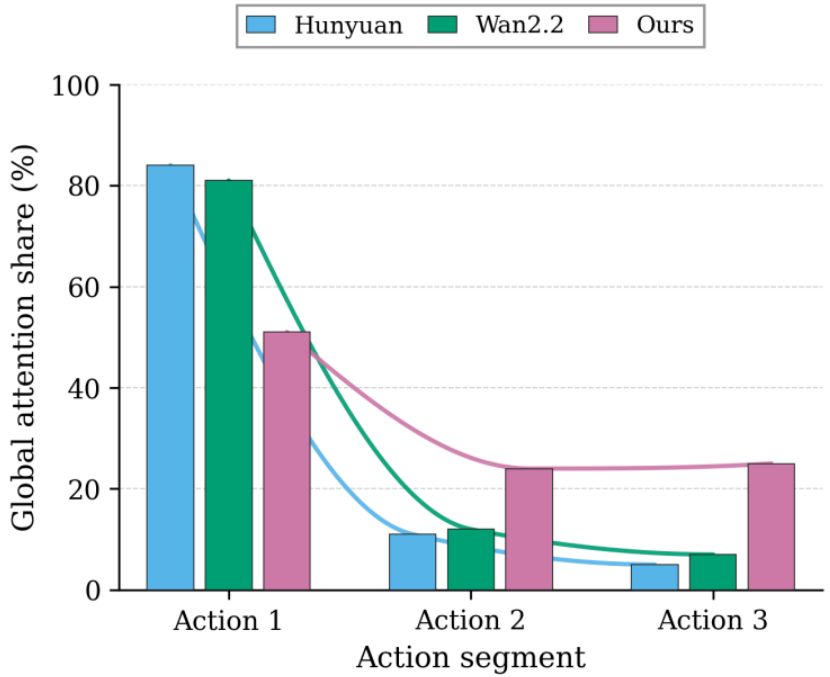
Bottleneck **From the internal mechanisms of DiT, the problem can be attributed to two types of entanglement and leakage.**

① Semantic Entanglement



Stable appearance features tend to dominate, diluting action semantics

② Temporal Attention Leakage



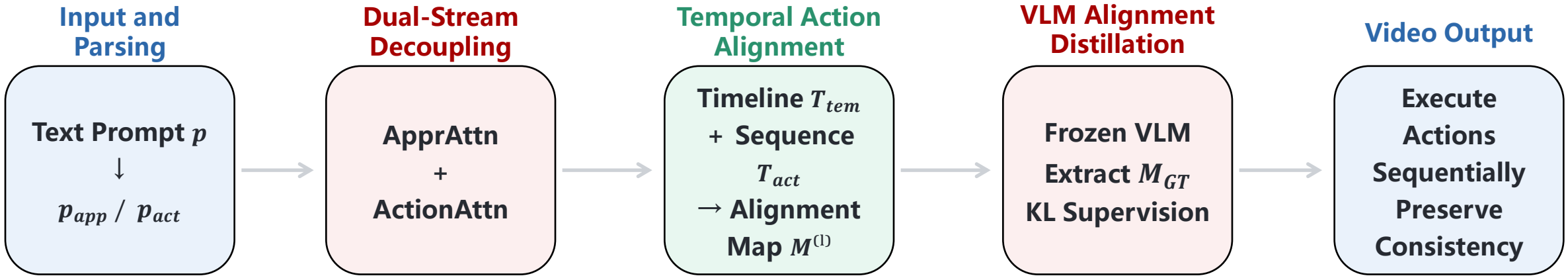
Early actions dominate; later ones are suppressed.

Solution: Decouple appearance-action modeling and explicitly constrain temporal-action alignment.

Methodology

北京大学 PEKING UNIVERSITY

Framework A Collaborative Video Diffusion Framework for Sequential Action Execution



Dual-Stream Modeling

Appearance and action interact independently, avoiding semantic interference before fusion

Explicit Alignment

Generate temporal-action mappings layer by layer to suppress cross-temporal leakage

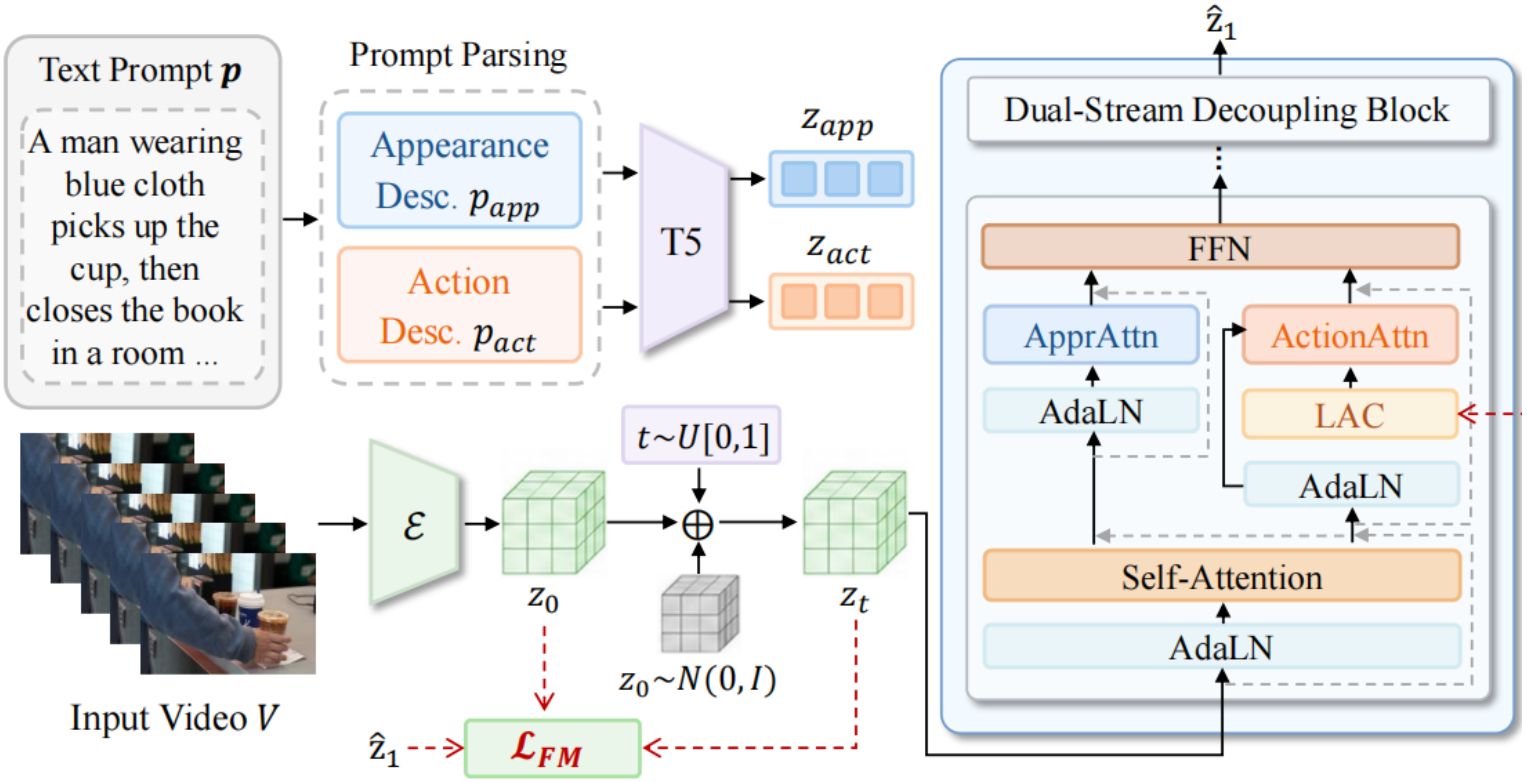
Alignment Distillation

Use VLM temporal localization priors to stabilize alignment optimization

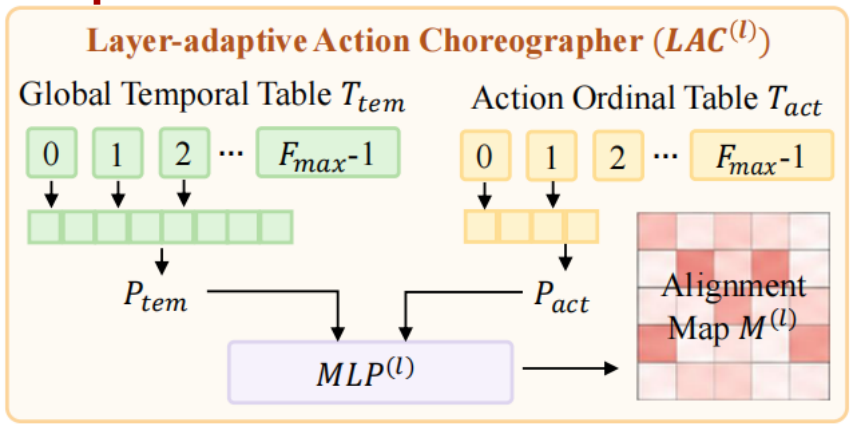
$$h_out^{(l)} = FFN\left[\, h^{(l)} + ApprAttn(h^{(l)}, z_{app}) + ActionAttn(h^{(l)}, z_{act}, M^{(l)}) \,\right]$$

Methodology

Module **Dual-Stream Attention + Layer-Adaptive Temporal Action Alignment (LAC)**



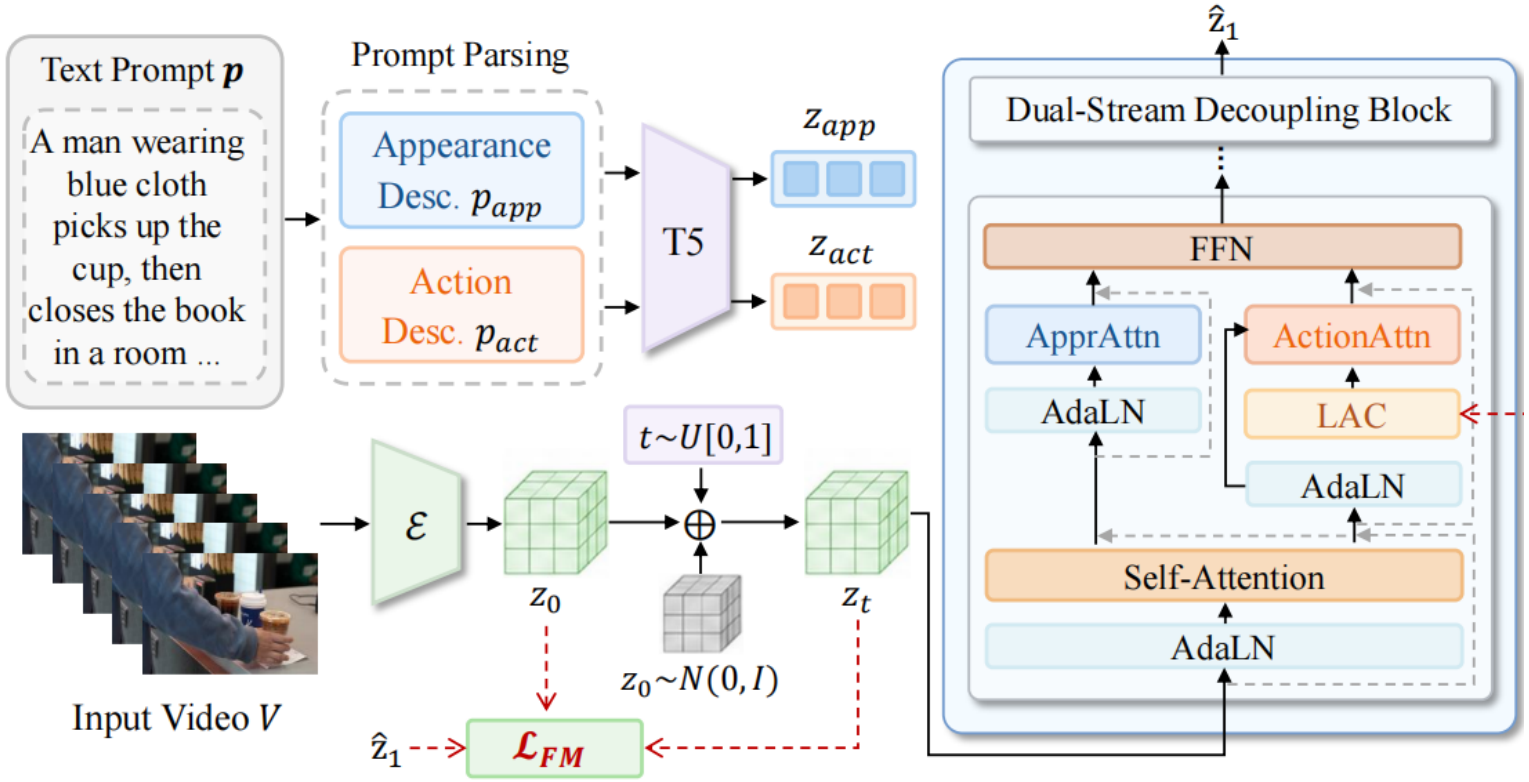
LAC Generates Layer-wise Alignment Maps $M^{(l)}$



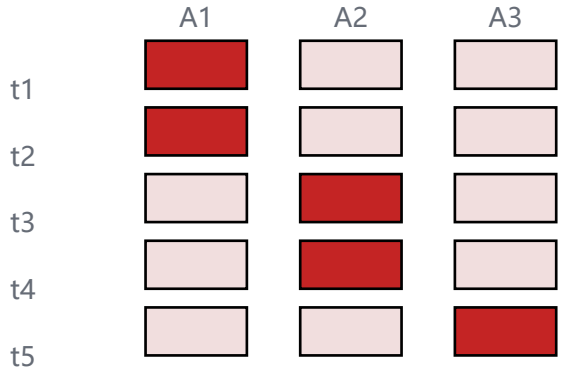
$$ActionAttn_{out} = Softmax(Q_{act}K_{act}^{T}/\sqrt{d} + M^{(l)}) \cdot V_{act}$$

Methodology

Module **Dual-Stream Attention + Layer-Adaptive Temporal Action Alignment (LAC)**



Temporal-Action Alignment Map $M^{(l)}$ (Illustration)

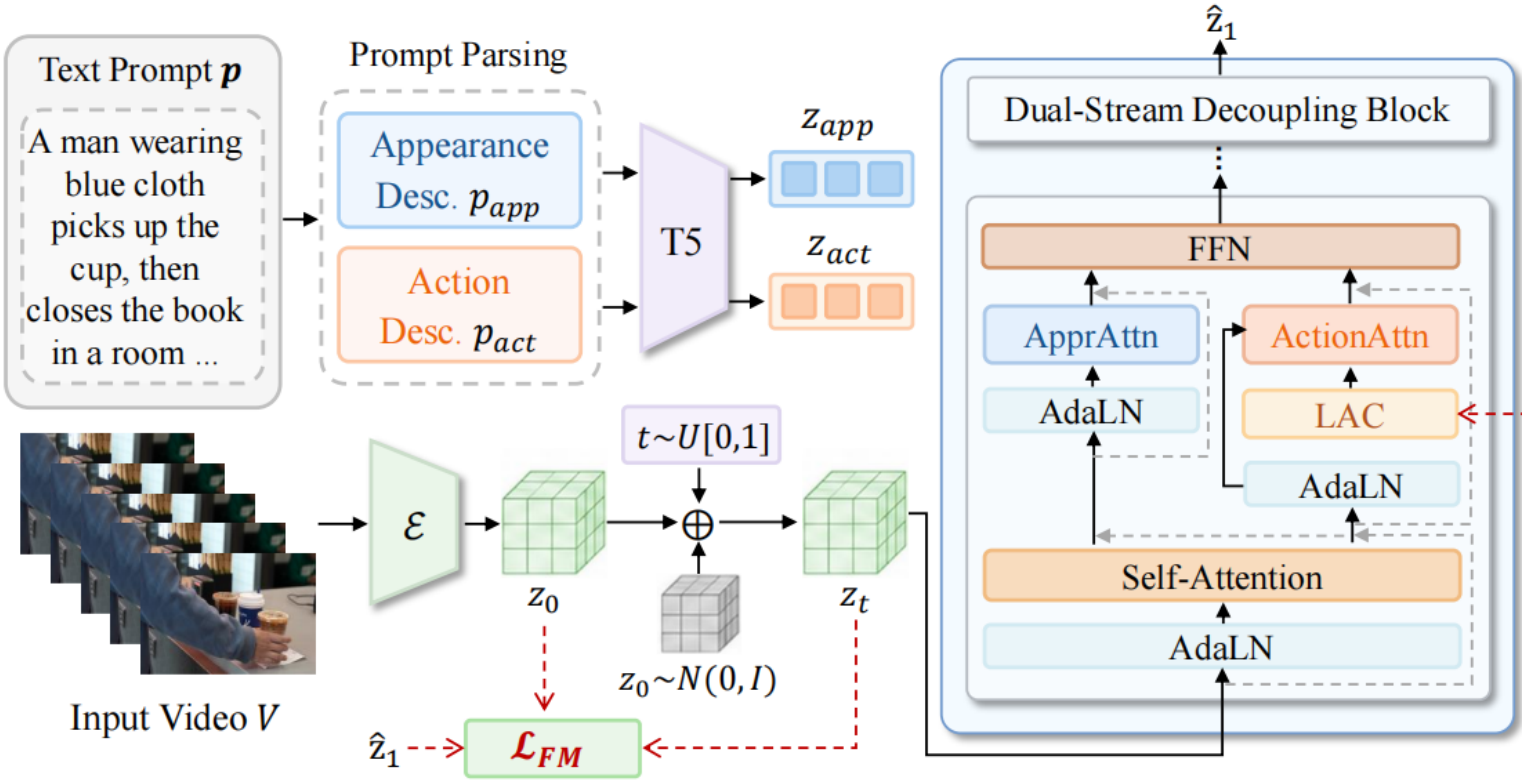


Variable action lengths · layer-adaptive · generalizable to more actions

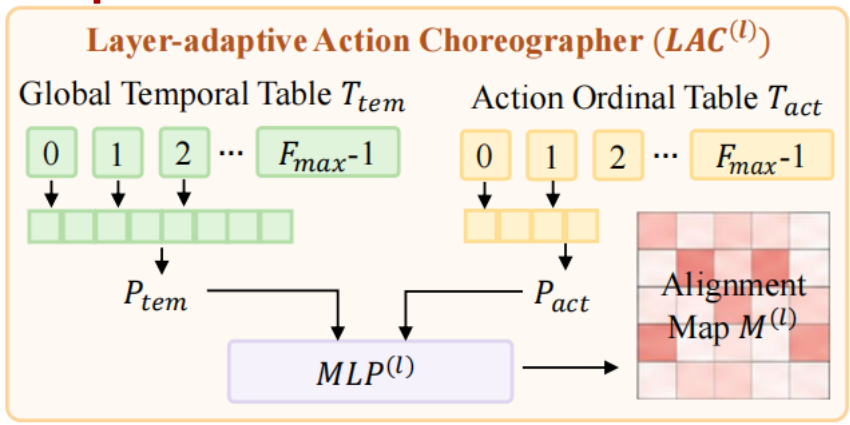
$$ActionAttn_{out} = Softmax(Q_{act}K_{act}^{T}/\sqrt{d} + M^{(l)}) \cdot V_{act}$$

Methodology

Module **Dual-Stream Attention + Layer-Adaptive Temporal Action Alignment (LAC)**



LAC Generates Layer-wise Alignment Maps $M^{(l)}$



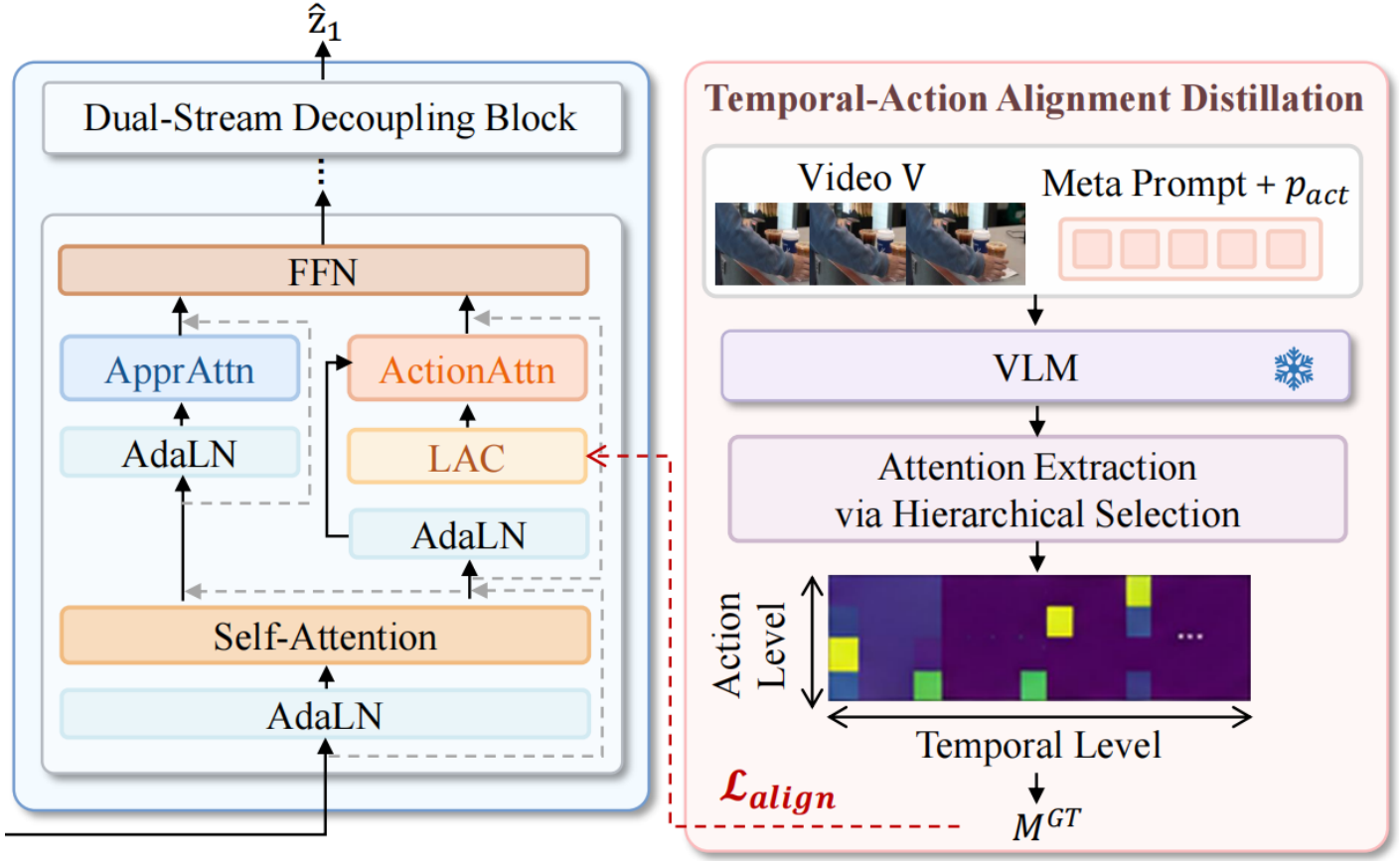
Variable action lengths · layer-adaptive · generalizable to more actions

$$ActionAttn_{out} = Softmax(Q_{act}K_{act}^{T}/\sqrt{d} + M^{(l)}) \cdot V_{act}$$

Methodology

Training

Temporal-Action Alignment Distillation: Guided by VLM Priors



Selection

Relevance + Sharpness

Target Map

Average Selected Heads and Normalize

KL Alignment

$$\mathbf{P}^{(l)} = \mathrm{LogSoftmax}(M^{(l)}).$$

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

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

Instead of forcing the diffusion model to infer action timing from scratch, distill the VLM learned temporal localization capability as auxiliary supervision.

Experimental Results

北京大学 PEKING UNIVERSITY

Comparison SAX Shows Significant Advantages in Action Completion and Dynamics

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
SAX (Ours)	**515.05**	**0.2977**	**0.6358**	**95.53**	**98.91**	**70.33**	**69.87**

FVD 515.05
Best Overall

Action Completion Rate higher than Wan 2.2

Dynamic Degree 70.33
30.2% higher than Wan 2.2

Experimental Results

北京大学 PEKING UNIVERSITY

Comparison **Image-to-Video Generation Task**

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
Hunyuan	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
SAX	**238.03**	**0.2959**	**0.6793**	93.43	**98.89**	**56.67**	**66.69**

Conclusion: SAX Significantly Mitigates Action Omission and State Freezing

Experimental Results

北京大学 PEKING UNIVERSITY

Ablation

Each Core Component Provides Consistent Gains

LAC Alignment Map

How to Obtain Alignment Map	FVD	Rate
✘	532.14	0.5358
Static Alignment Map	530.26	0.5679
Ours	**515.05**	**0.6538**

Distillation

Distillation Forms	FVD	Rate
✘	531.26	0.5513
Using Average Map	530.45	0.5597
Ours	**515.05**	**0.6538**

Decoupling

Appearance-Action Decoupling	FVD	Rate
✘	530.28	0.5519
Ours	**515.05**	**0.6538**

Alignment Loss Form

Different Forms	FVD	Rate
MSE	529.34	0.5962
Ours	**515.05**	**0.6538**



Thanks