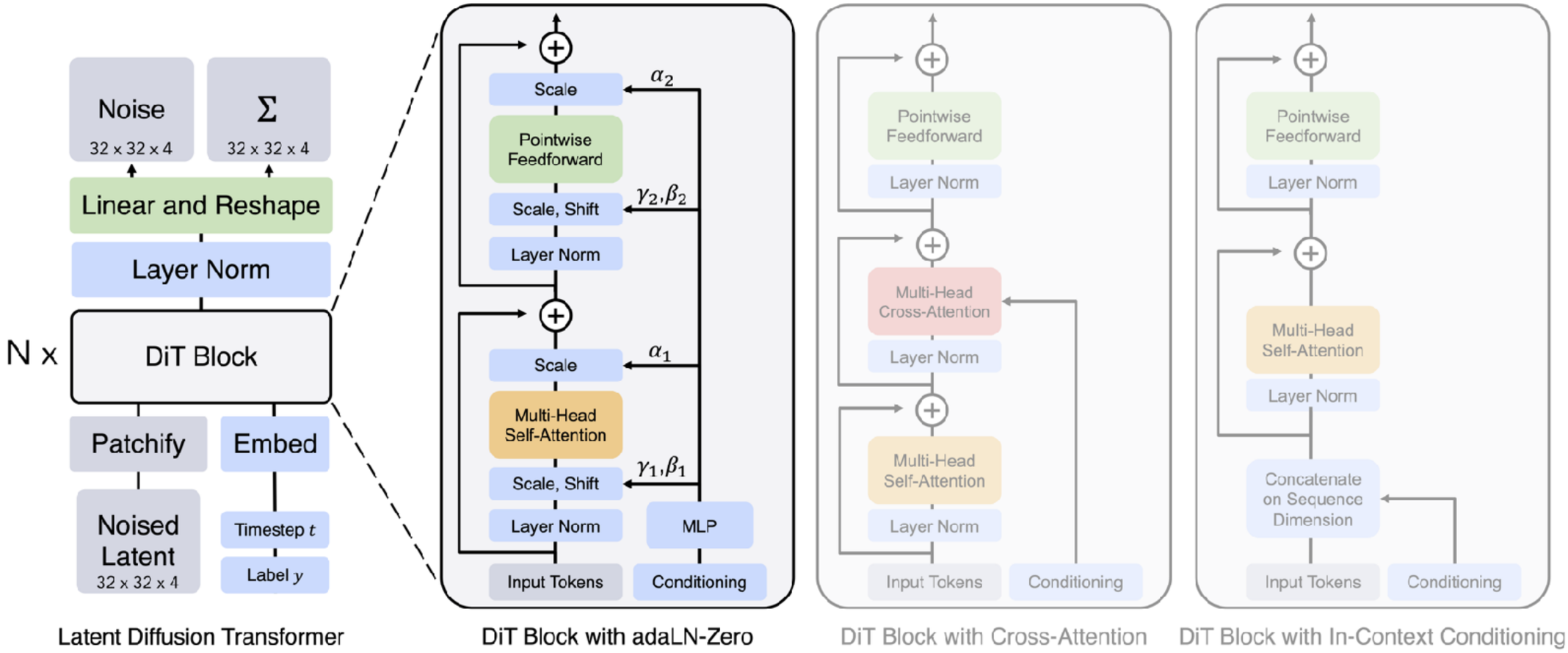
Streaming Video Generation as a Memory Problem

From Contextual to Parametric Memory

MA SHUAI

Diffusion based video generation

Limited generation duration



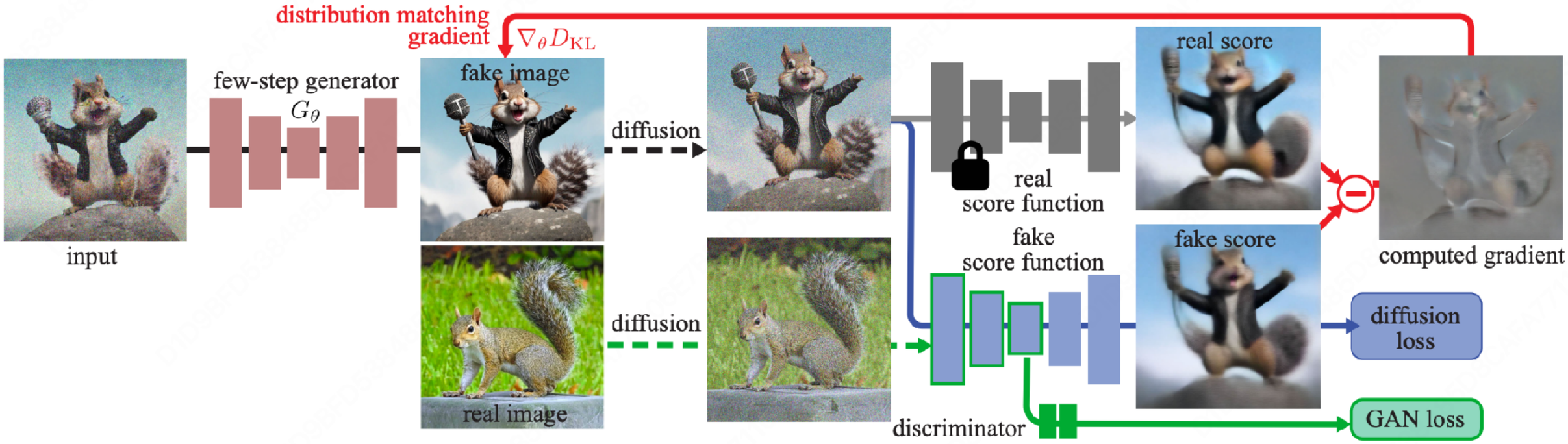
Latent Diffusion Transformer | DiT Block with adaLN-Zero | DiT Block with Cross-Attention | DiT Block with In-Context Conditioning

From Short Clips to Causal Rollout

AR based video generation

DMD

Distribution Matching Distillation



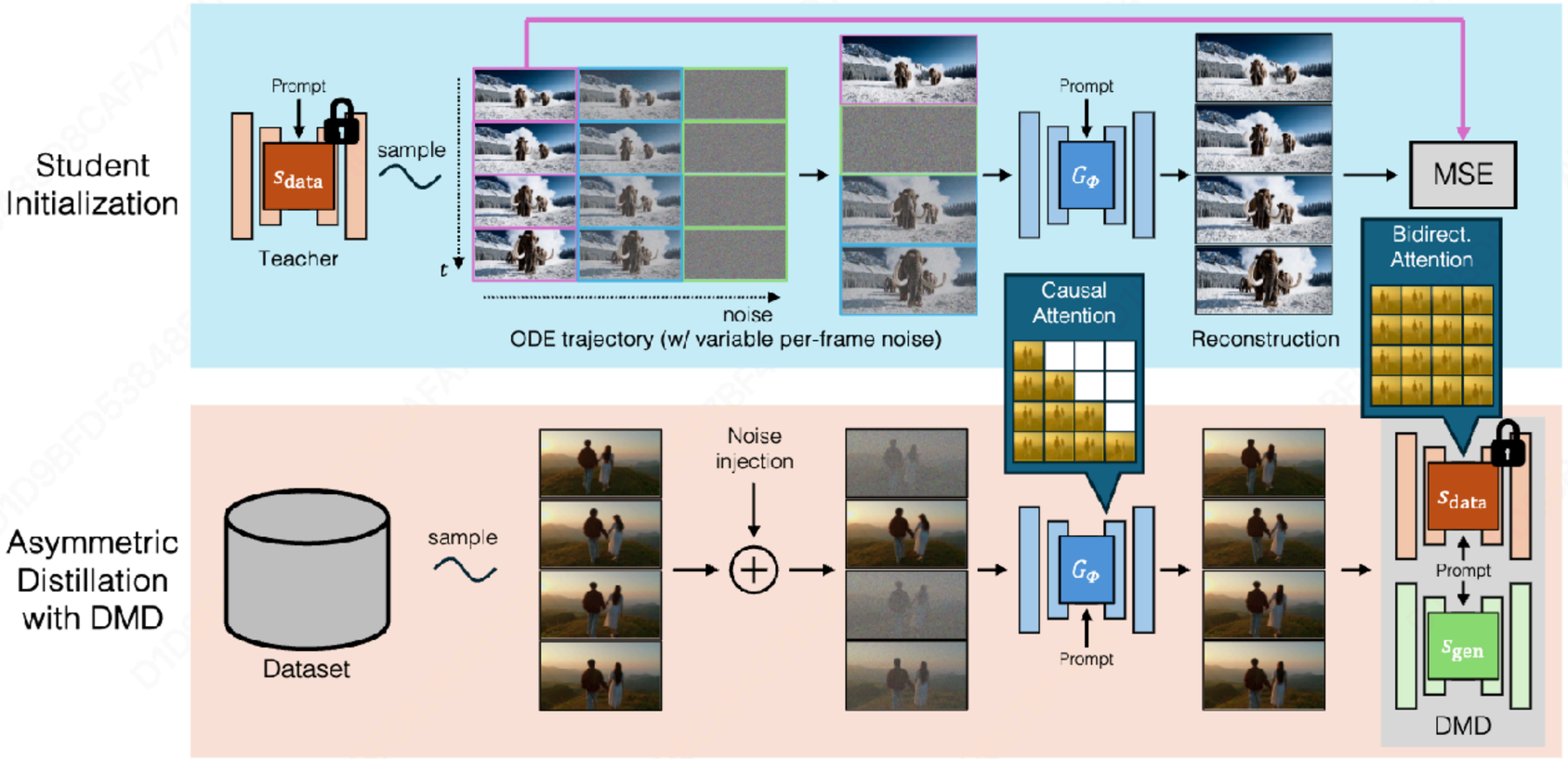
One-step Diffusion with Distribution Matching Distillation CVPR 2024

Improved Distribution Matching Distillation for Fast Image Synthesis NeurIPS 2024 Oral

AR based video generation

CausVid

Distill from Bidirectional to Causal

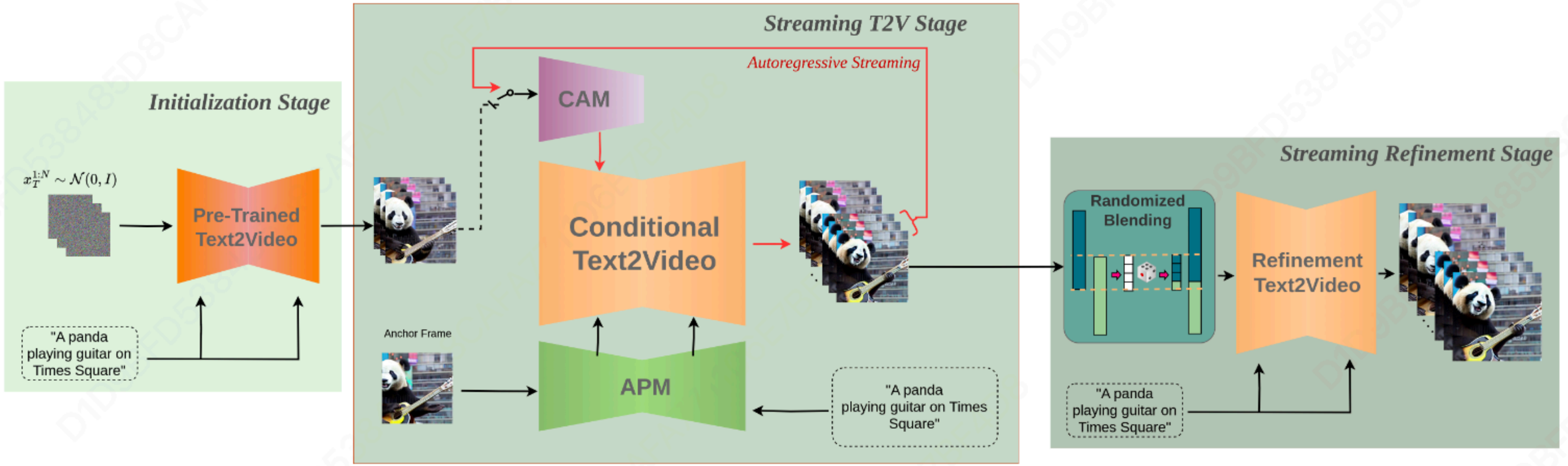


From Slow Bidirectional to Fast Autoregressive Video Diffusion Models CVPR 2025

History frames as context

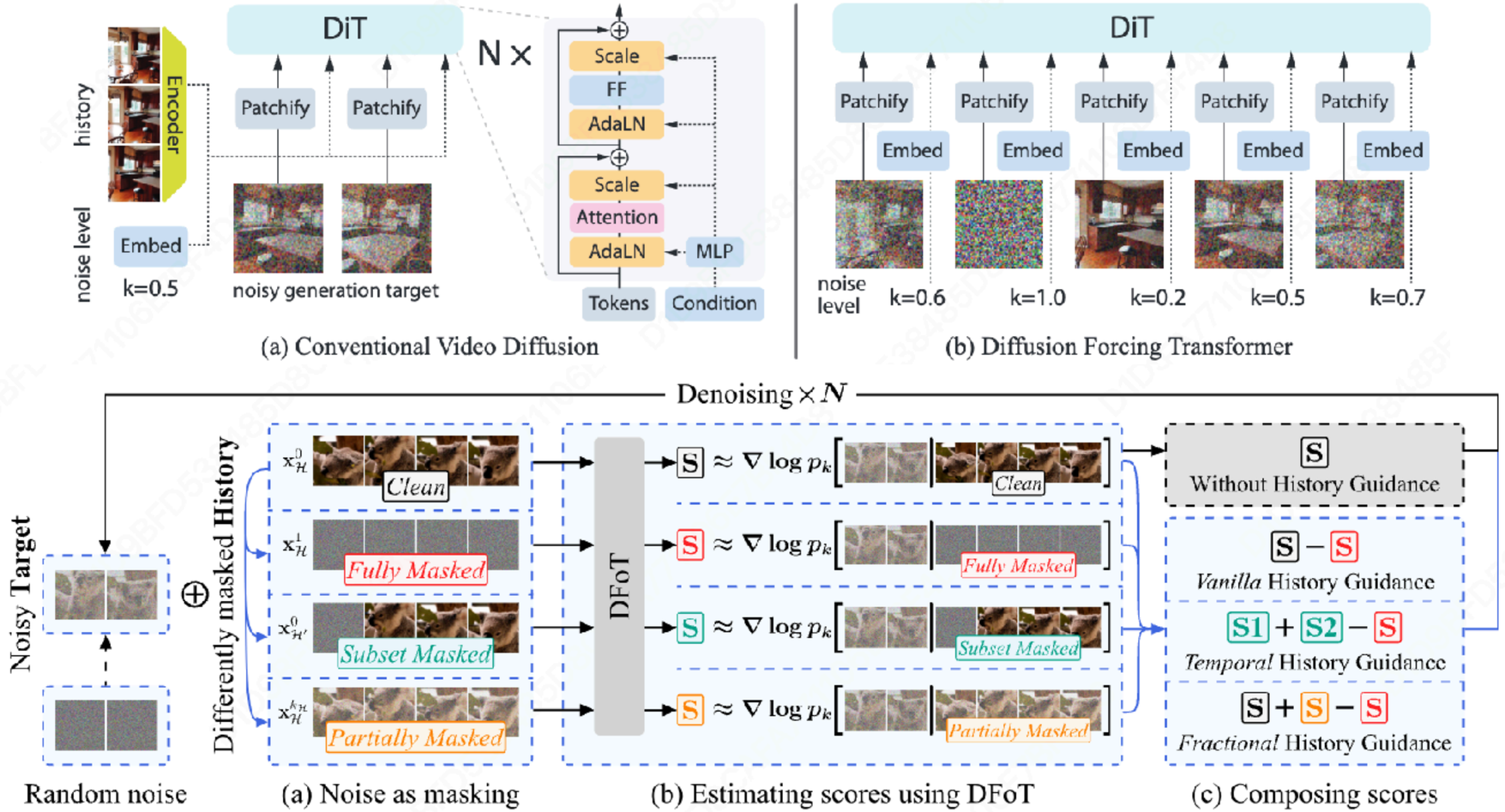
Streaming T2V

Dual-timescale Memory



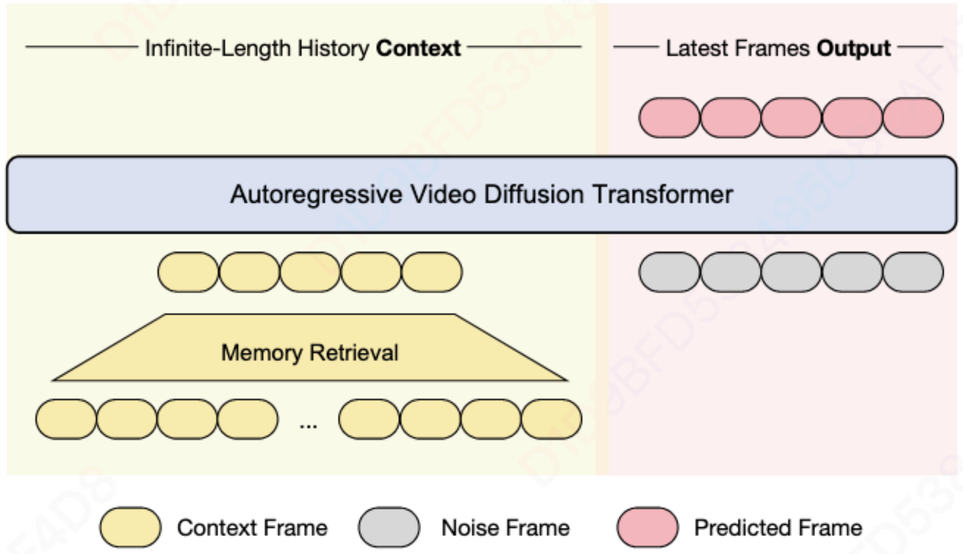
History frames as context

History guided video diffusion



History frames as context

Context as memory



(a) **Context learning**: long video generation with Memory Retrivel

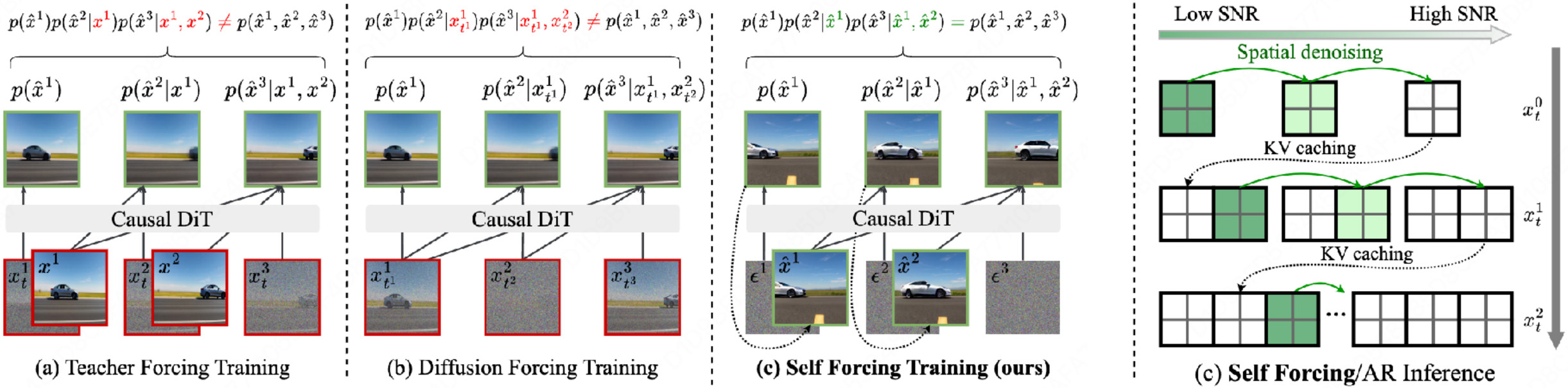
(b) **Memory Retrieval**: select High-overlap Context to condition the generation of Predicted Frame.

Towards Stable Rollout

Stable Rollout

Self forcing

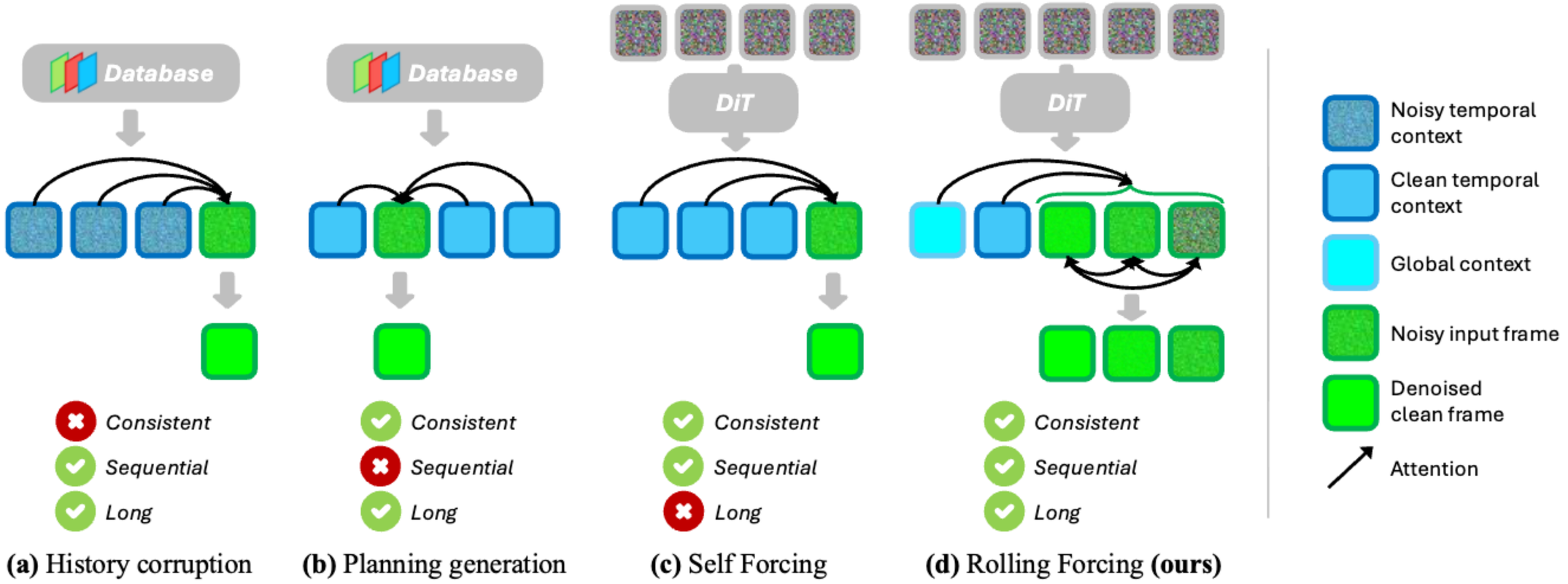
Test-Training gap



Stable Rollout

Rolling forcing

Rolling Window Denoising

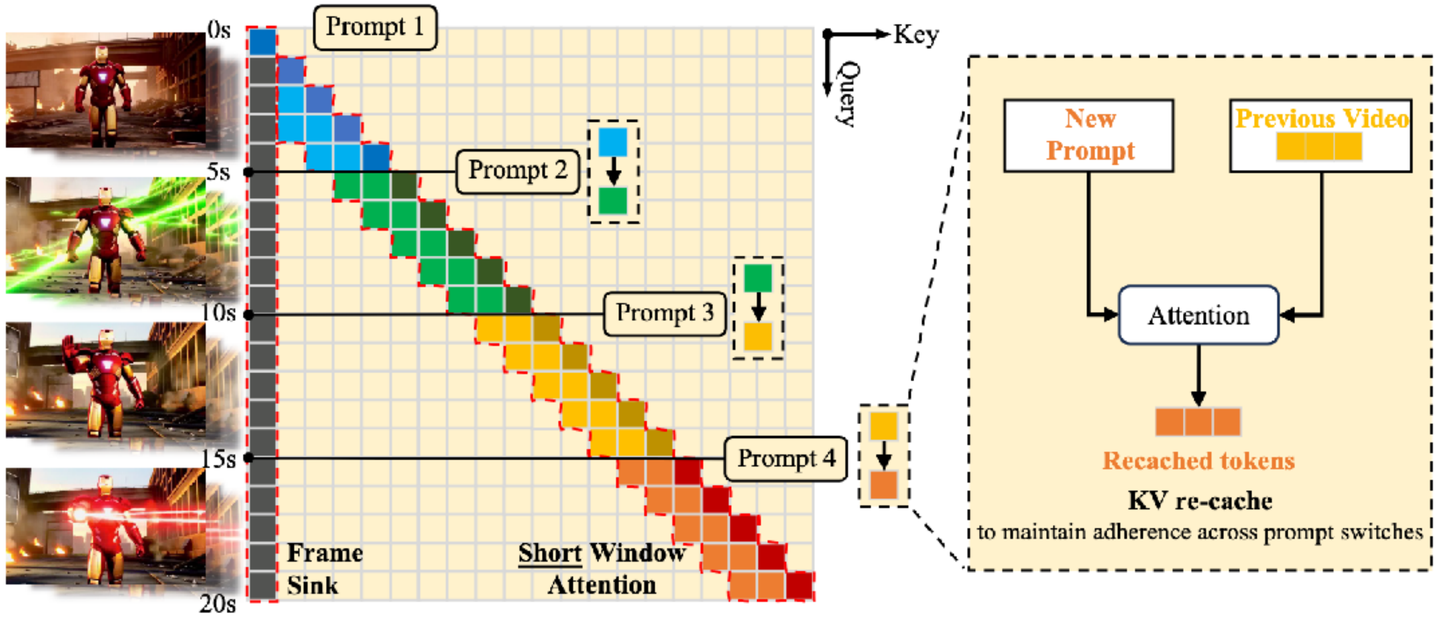


Rolling Forcing: Autoregressive Long Video Diffusion in Real Time ICLR 2026

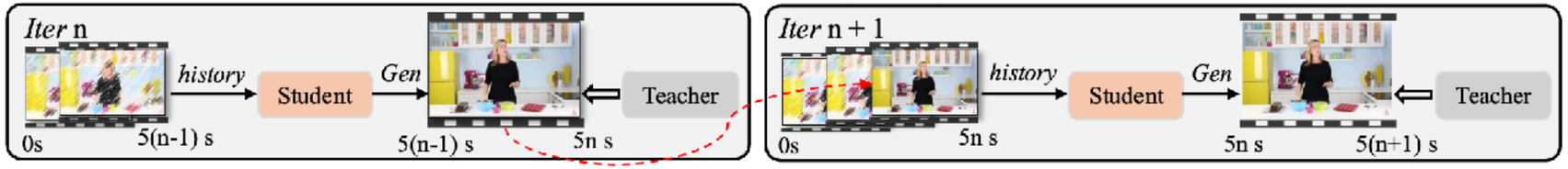
Stable Rollout

Longlive

Train-long-test-long



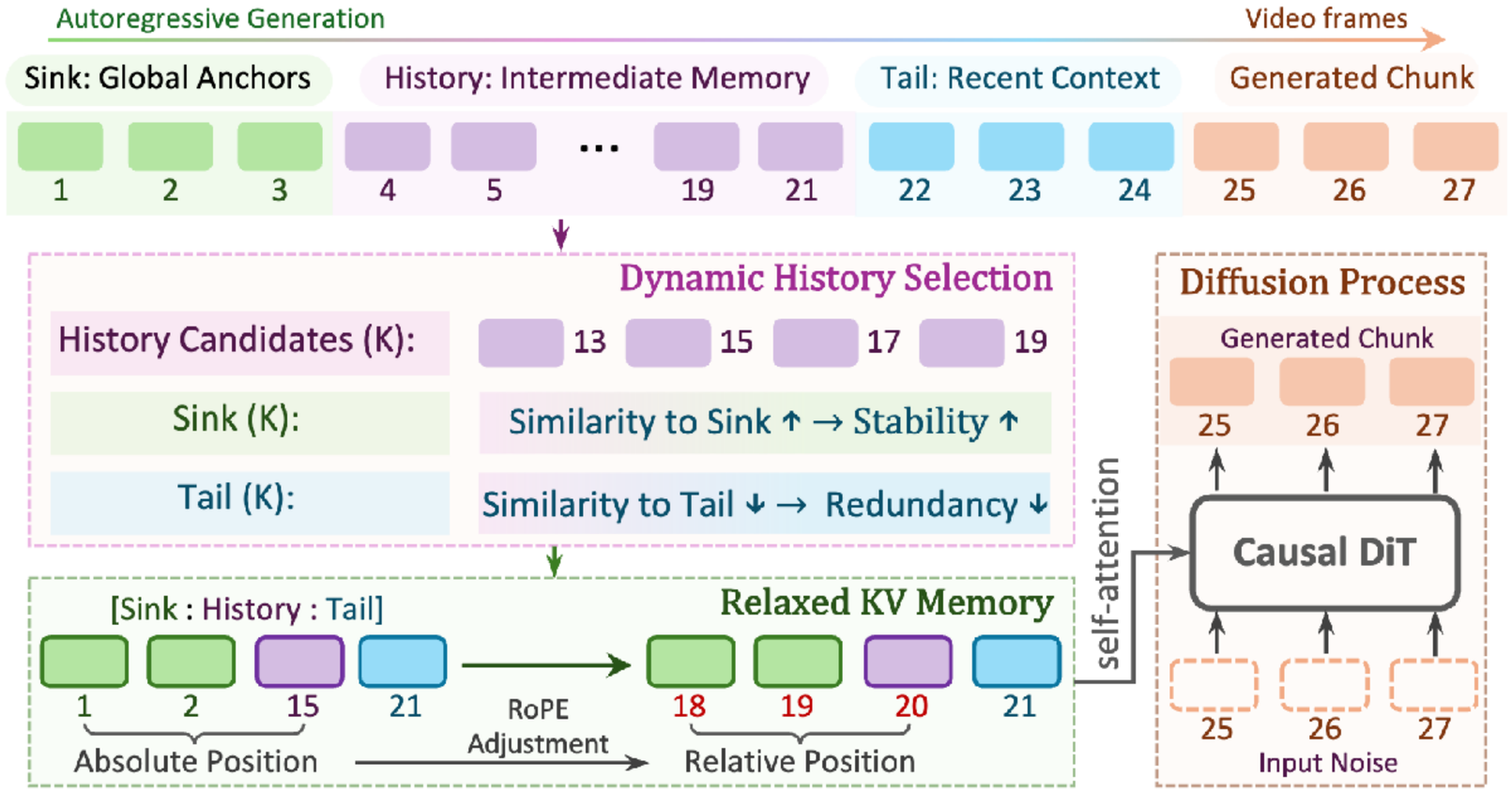
(c) Streaming Long Tuning



Stable Rollout

Relax forcing

Selecting Memory

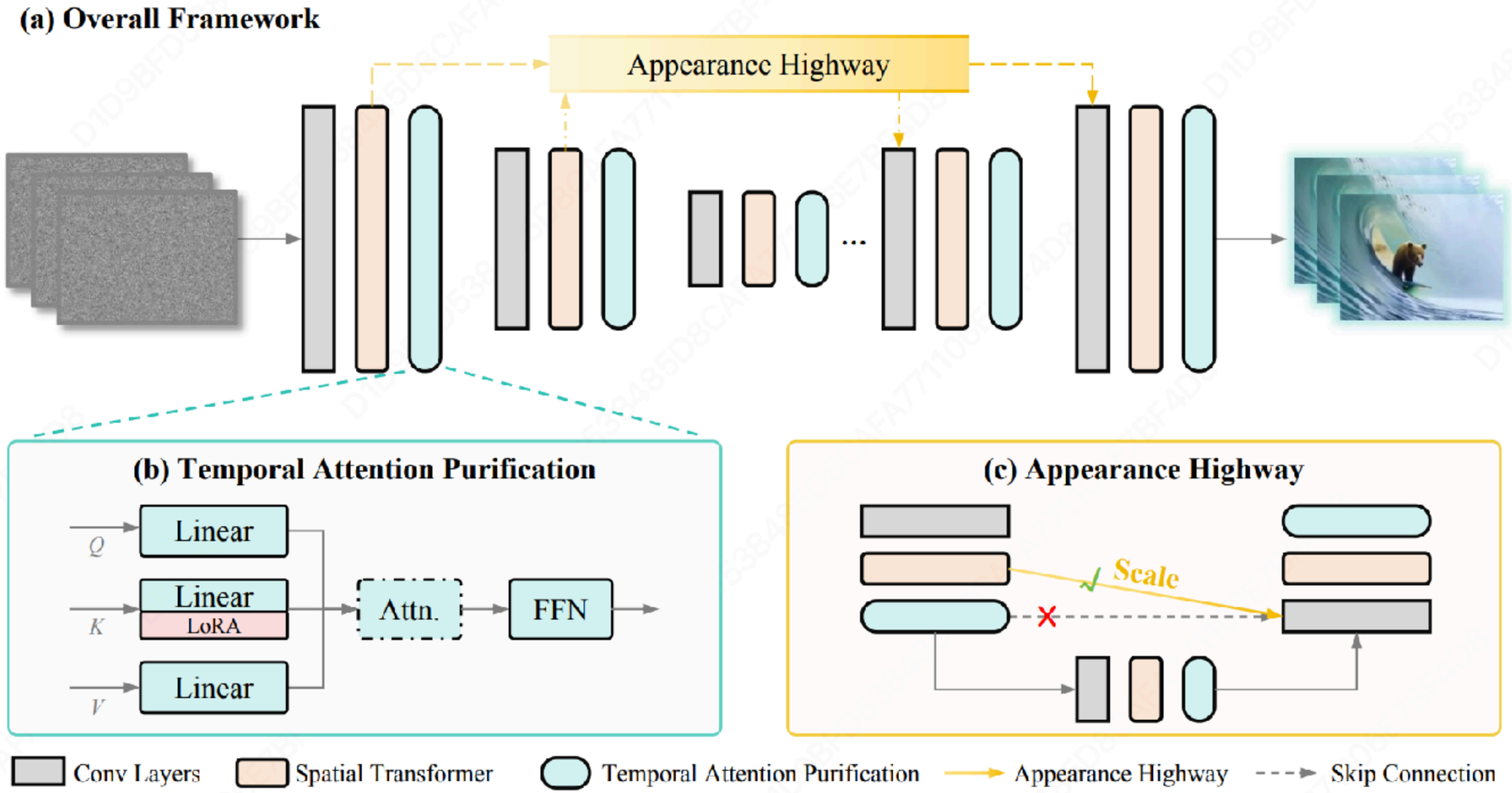


Relax Forcing: Relaxed KV-Memory for Consistent Long Video Generation Arxiv 2026.03

From Contextual to Parametric Memory

Parametric Memory

Separate Motion from Appearance

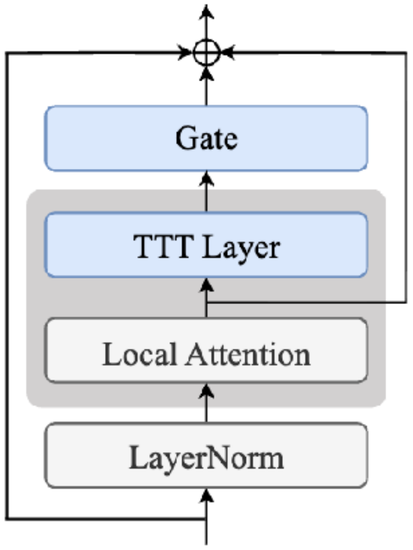


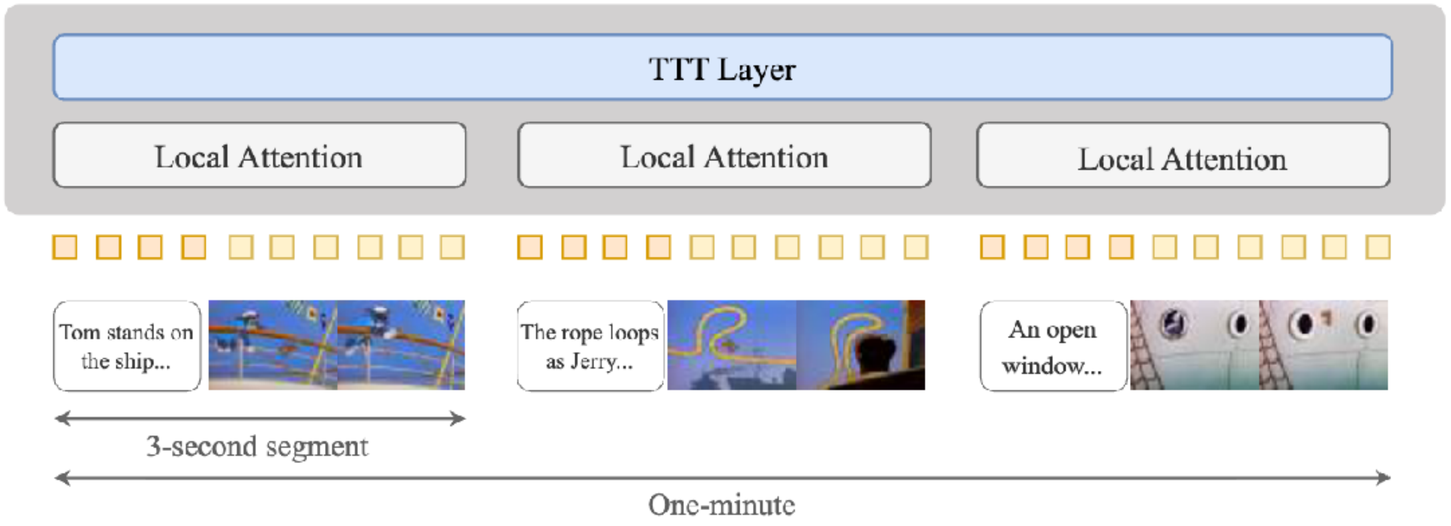
Parametric Memory

One-Minute Video Generation with Test-Time Training

Diffusion based

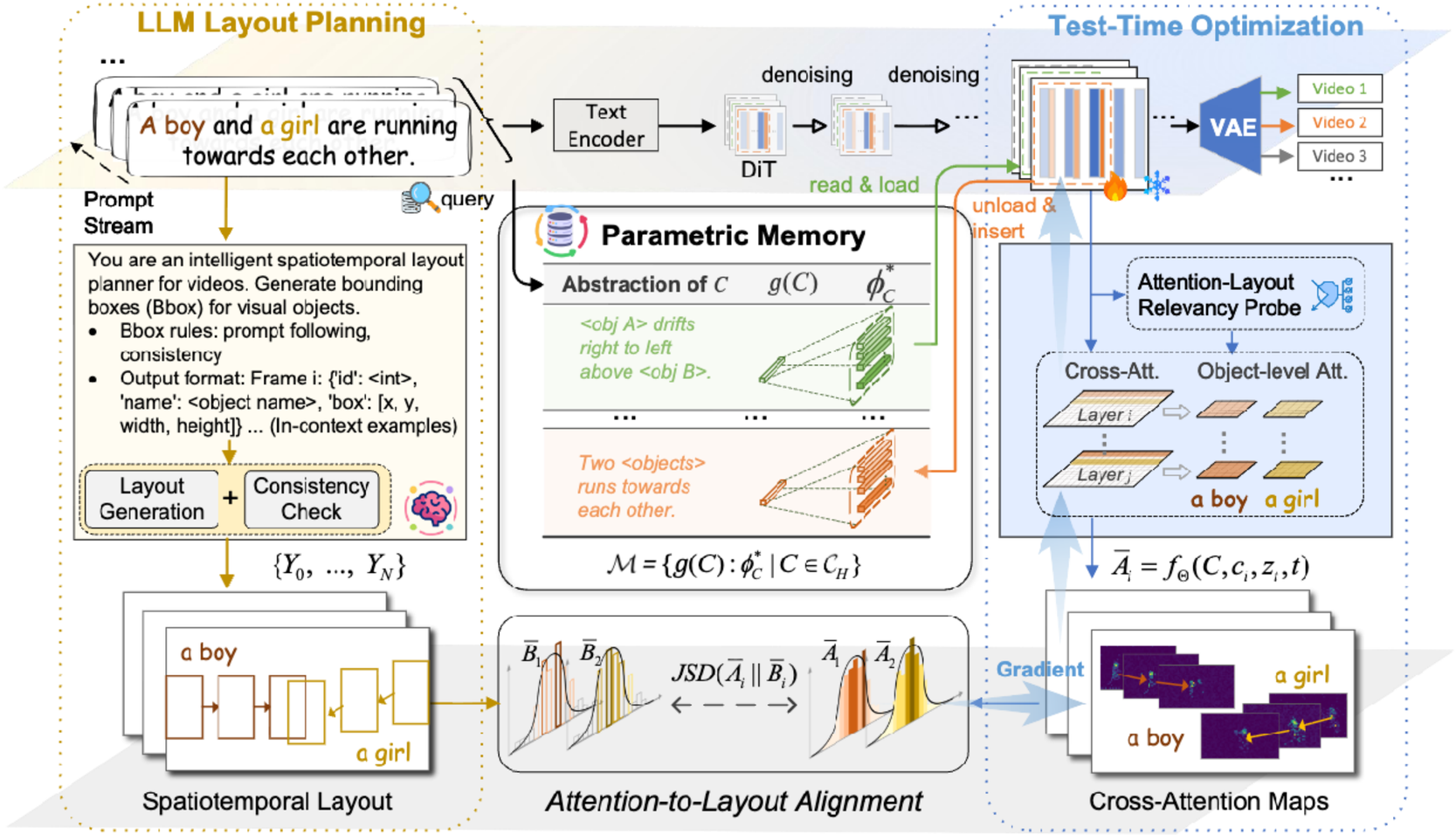
$$\ell(W; x_t) = \|f(\theta_K x_t; W) - \theta_V x_t\|^2$$





Parametric Memory

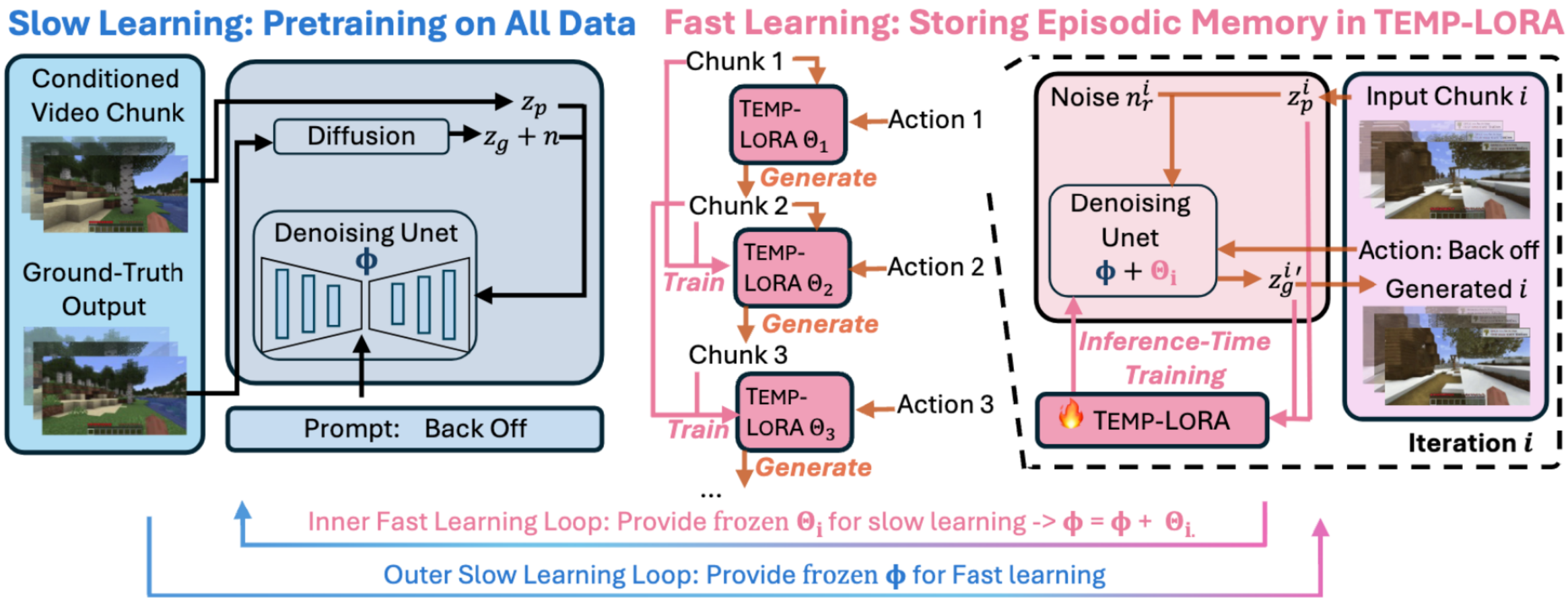
TTOM



TTOM: Test-Time Optimization and Memorization for Compositional Video Generation ICLR2026

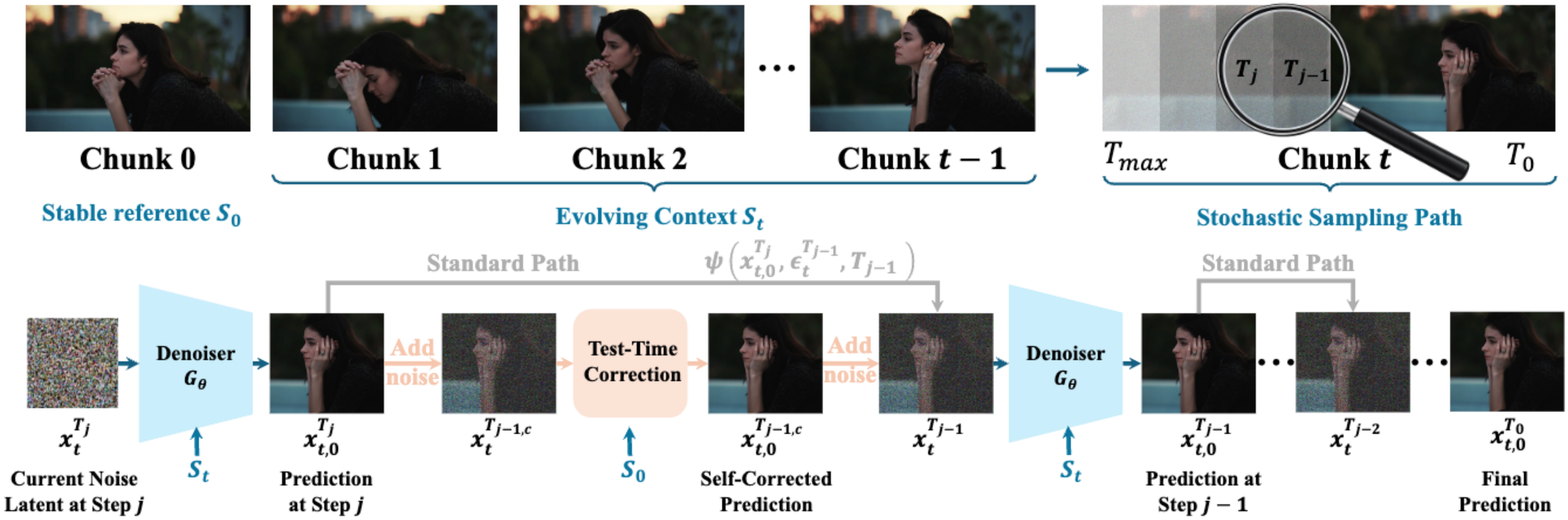
Parametric Memory

SlowFast-VGen



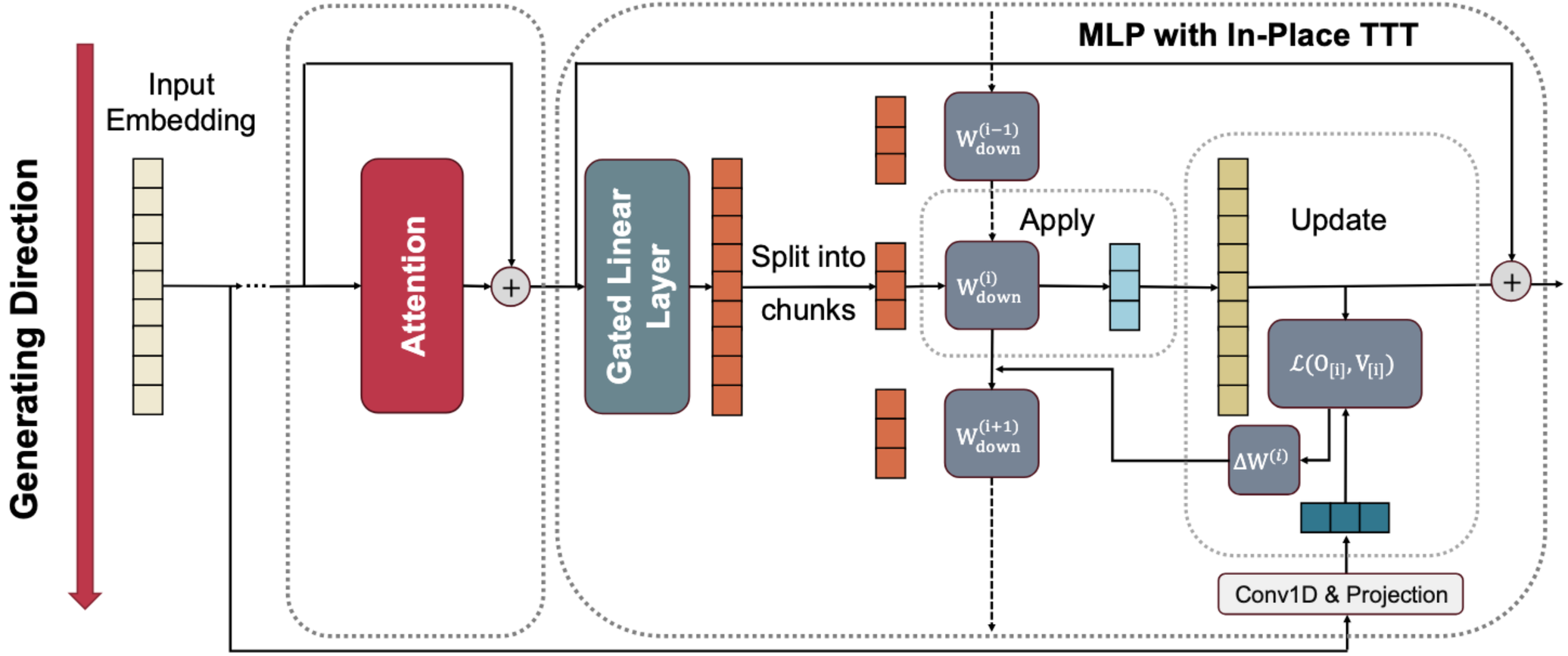
Parametric Memory

Pathwise Test-Time Correction for Autoregressive Long Video Generation



Parametric Memory

In-Place Test-Time Training



Parametric Memory

LoRA as Knowledge Memory

- ***Hybrid Memory is better***
- ***LoRA&ICL in video generation***
- ***Test Time Training***

Embrace Hybrid Memory for Robustness. Our results indicate that LoRA-based parametric memory is best viewed not as a replacement for context-based methods, but as a complementary component. While LoRA can reduce inference costs for repeated access to a stable knowledge base, peak performance on long and multi-hop settings is achieved when LoRA is combined with external context (ICL/RAG), which mitigates fragmentation and restores global coherence. Together, these findings can position LoRA as a practical complementary axis of memory alongside RAG and ICL, and motivate hybrid systems as a promising direction for robust and efficient knowledge access.

Understanding LoRA as Knowledge Memory: An Empirical Analysis **Arxiv 2026.03**

核心问题

误差累积与长程锚点之间的trade off

- Anchor frame可抑制rollout shift，但是生成过程过分绑定在初始语义上、无法适应新语义
- attention sink过度关注首帧：sink collapse，后面记忆读不出来
- 如何管理context中的memory很重要，memory的重点是选择
- TTT：不在context中分配注意力预算，避免sink collapse，将部分长期信息写进参数中

LoRA

- 和历史帧的人物/背景一致性： 与anchor frame的CLIP similarity

- diffusion loss重构：前期用anchor frame，后期转用lora：加噪训练重构前期用anchor frame生成的chunk，不过这次不用anchor frame；
- 生成两路：teacher（有anchor）+student（无anchor）

KV点积loss

需要预训练

$$W \leftarrow W - \eta \nabla_W \mathcal{L}\big(f_W(k), v\big)$$

$$\mathcal{L}\big(f_W(k_i), v_i\big) = -f_W(k_i)^\top v_i$$

$$f_W(x) = W_2\,[\mathrm{SiLU}(W_1 x) \circ (W_3 x)]$$