Context Forcing: Consistent Autoregressive Video Generation with Long Context

Shuo Chen, Cong Wei, Sun Sun, Tiancheng Shen, Ping Nie, Kai Zou, Ge Zhang, Ming-Hsuan Yang, Wenhu Chen

ICML 2026

Causal Autoregressive Video Generation

Causal autoregressive video generation

CausVid

- Distill slow bi-directional teacher model to few step autoregressive model

Self Forcing: Bridge training and inference gap

- Student self rollout, using bi-directional teacher model give whole video DMD loss.
- Trained on short clip video.

Causal autoregressive **long** video generation

LongLive & Self Forcing ++

- Student self **rollout long** sequence, using bi-directional **short teacher** model give whole video DMD loss.

InfinateRoPE

- Reduction of error accumulation by modifying positional encodings to relative.

Forgetting Drifting Dilemma in causal long video generation

Forgetting: Restricting the model to a short memory window minimizes error accumulation, but causes the model to lose track of previous subjects and scenes during long rollout.

Drifting: Maintaining a long context preserves more previous information, but exposes the model to more errors. The video distribution progressively drifts away from the real manifold.

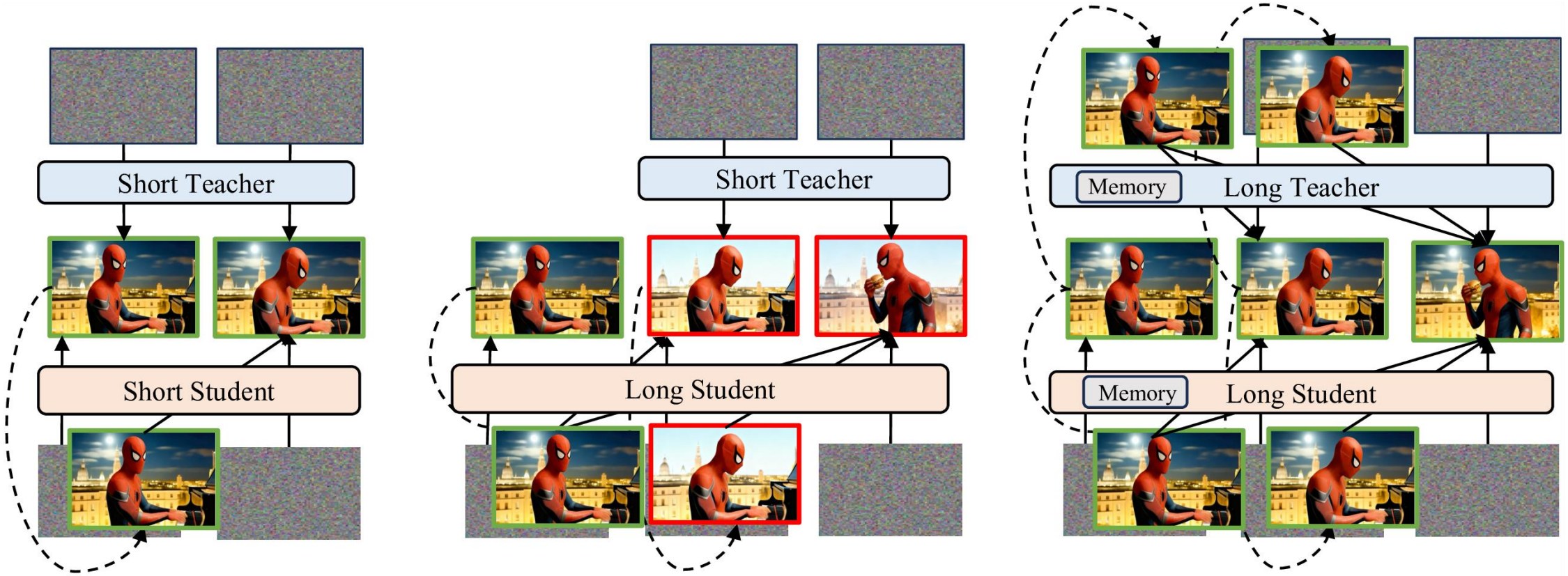
LongLive, 3.0s context (x5)

LongLive, 5.25s context (x5)

Ours, 20s+ context (x5)

[1] Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., ... & Chen, Y. (2025). Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*.

Context Long Tuning: Bridge Teacher-Student Mismatch



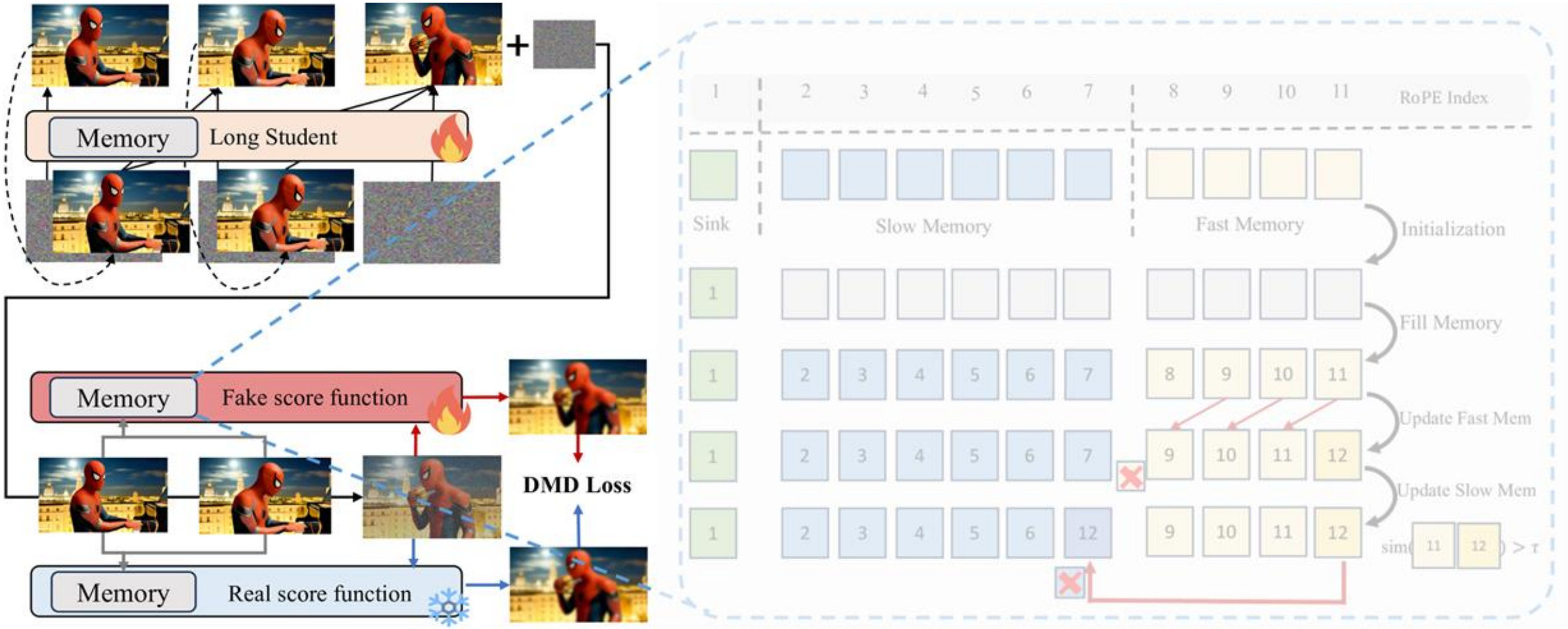
(a) Self-Forcing Training
short teacher → short student

(b) Memoryless Long Tuning
short teacher → long student

(c) **Context Long Tuning(ours)**
long teacher → long student

Context Forcing: The student is supervised by a long-context teacher aware of the full generation history, resolving the mismatch in (b).

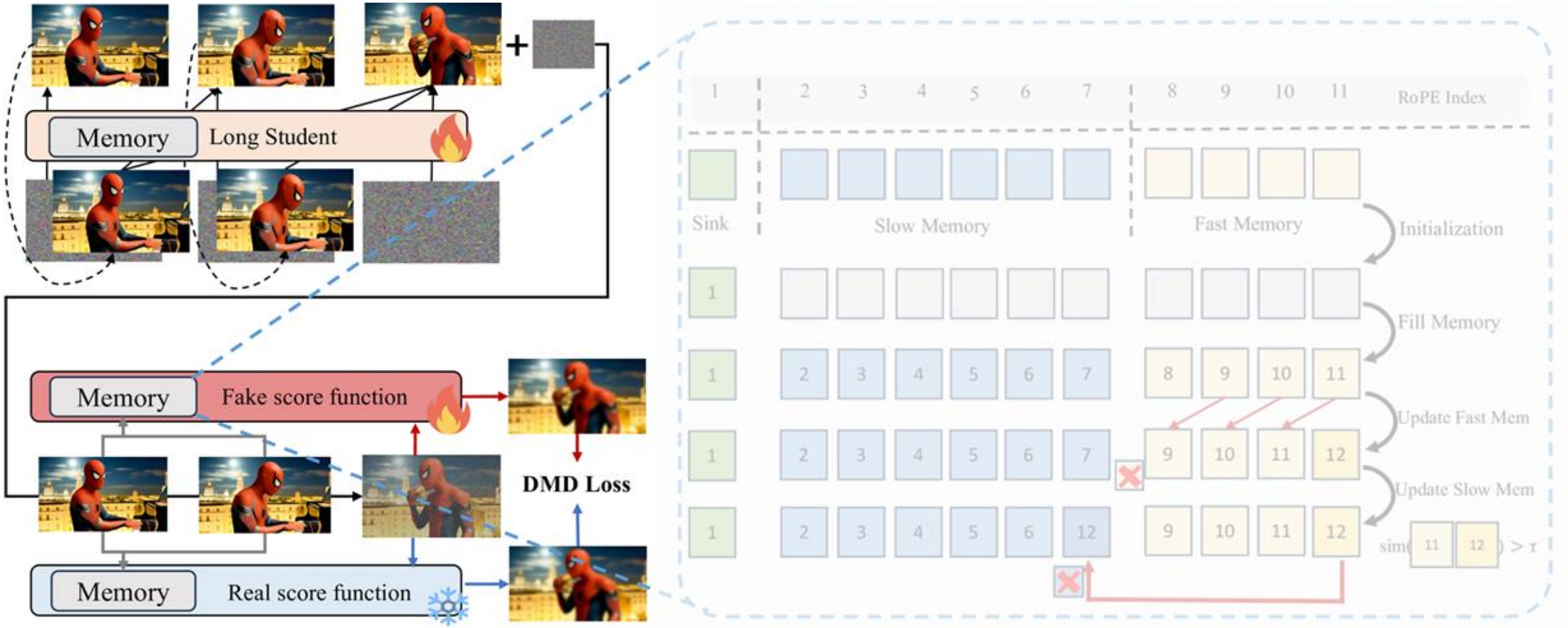
Context Distribution Matching Distillation



Context Forcing: The student is supervised by a long-context teacher aware of the full generation history, resolving the teacher-student mismatch.

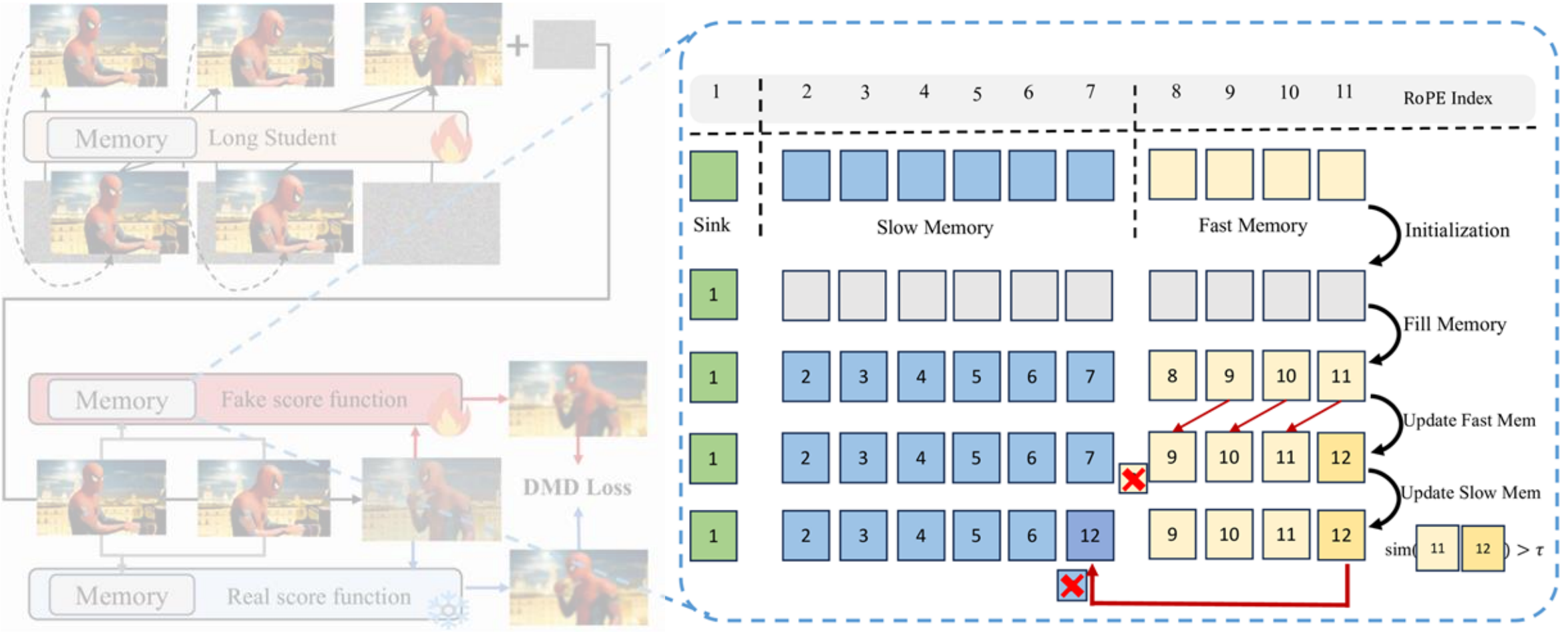
During contextual DMD training, the long teacher provides supervision to the long student by utilizing the same context memory mechanism.

Context Distribution Matching Distillation



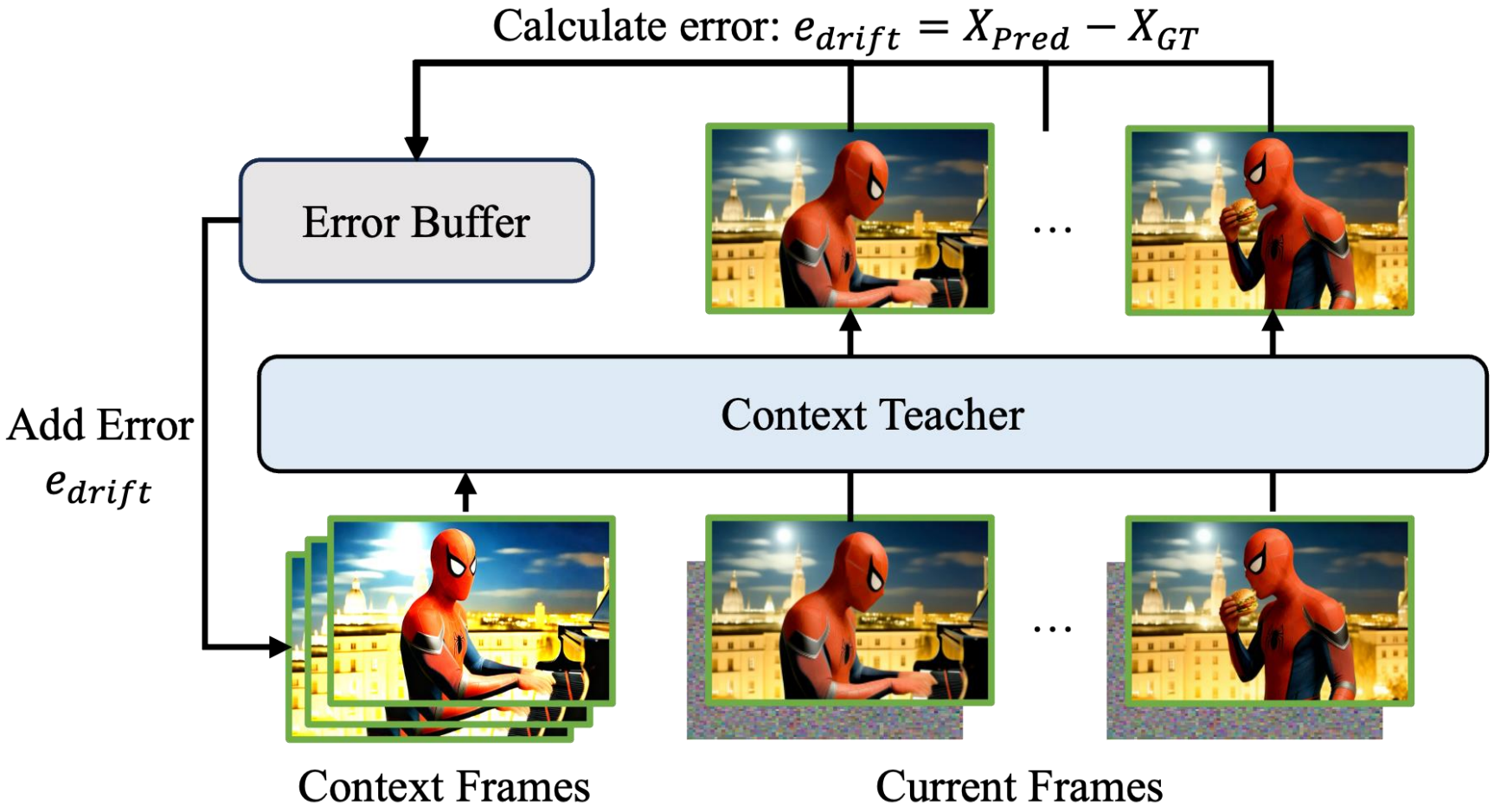
Contextual DMD: Real Score(teacher model) is context teacher which is frozen and can take clean context images and noise images as input and do denoise.
Fake score is same arch with real score, learning the student distribution.
Student: autoregressive student with context management system

Context Management System



We use KV Cache as the context memory, and we organize it into three parts: sink, slow memory and fast memory.
Attention Sink retains initial tokens to stabilize attention, following StreamingLLM
Slow Memory storing high-entropy keyframes and updating only with significant new information.
Fast Memory capturing immediate local context with short-term persistence.

Context Teacher Training



To combat exposure bias, past predictive errors are dynamically stored in an Error Buffer and probabilistically **injected into clean context and noise latents**.

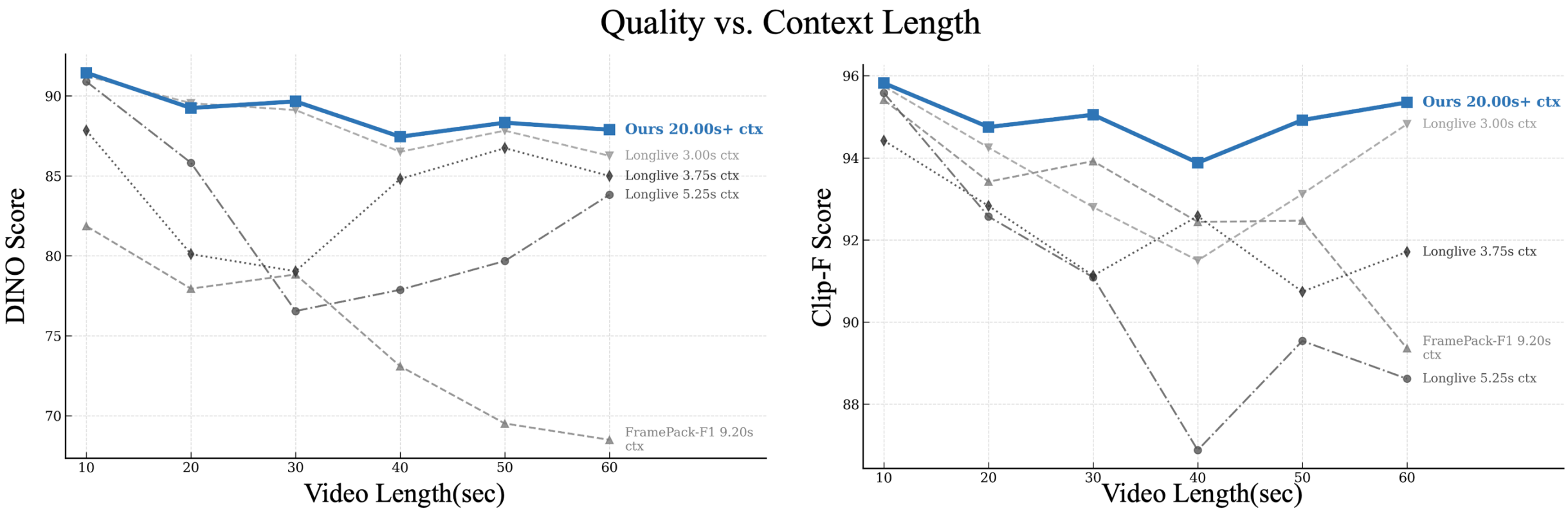
The input sequence aligns directly with the student model's **context management system** (incorporating 21 latent frames of sink, slow, and fast memory).

Long Video Generation

Qualitative Results of Context Forcing. Our method enables minute-level video generation with minimal drifting and high consistency across diverse scenarios.

Consistency Evaluation

For streaming long-context tuning baselines (e.g., LongLive),enlarging the context window during inference (3.0 → 5.25 s) causes error accumulation and distribution shift (Drifting).
Context Forcing supports 20s+ context while maintaining strong long-term consistency.



Consistency Evaluation

Our method keeps both the background and subject consistent across 1-min video, while other baselines have different levels drifting or identity shift.

LongLive[1]
(KV Size = 12)

Infinite Rope[2]
(KV Size = 6)

Ours(KV Size = 21)

[1] Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., ... & Chen, Y. (2025). Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*.
[2] Yesiltepe, H., Meral, T. H. S., Akan, A. K., Oktay, K., & Yanardag, P. (2025). Infinity-RoPE: Action-Controllable Infinite Video Generation Emerges From Autoregressive Self-Rollout. arXiv preprint arXiv:2511.20649.

Comparison with LongLive

Our method enables stable **minute-level video generation** with high **subject/background consistency** across diverse scenarios. Conversely, LongLive exhibits **flashback artifacts**(e.g., at 50s) and lacks the capacity to preserve consistent subject and background details throughout long sequences.

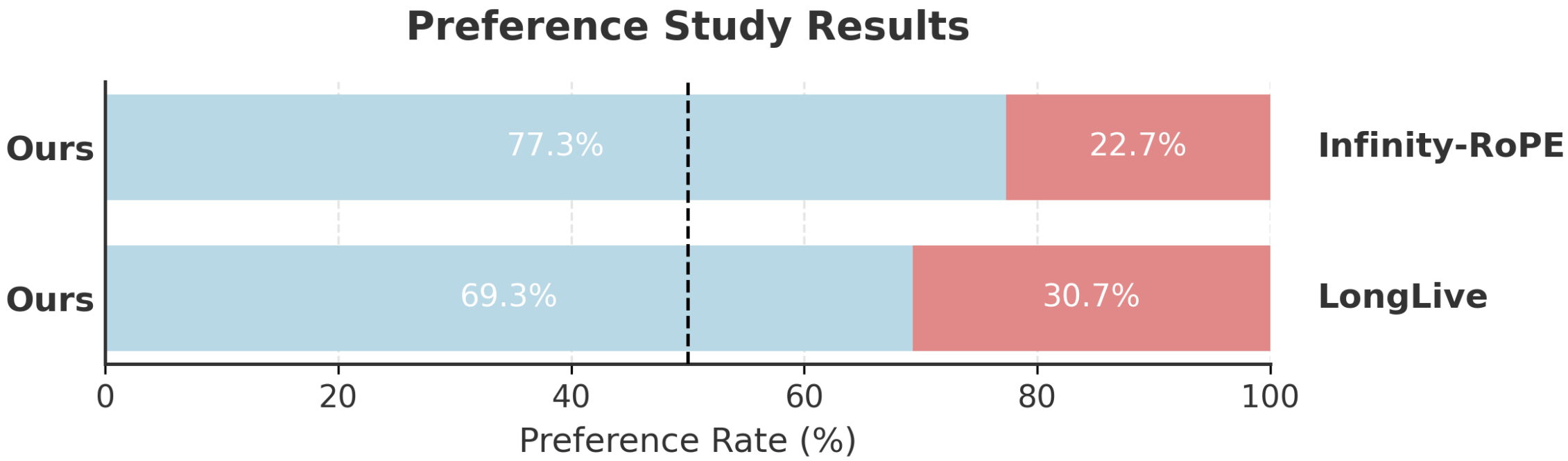
Longlive

Ours

[1] Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., ... & Chen, Y. (2025). Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*.

User Study

Human raters compare two anonymized videos and indicate which is more aesthetic, better matches the prompt, and exhibits stronger temporal consistency (subject and background coherence).



Blind preference rates (higher is better for Ours): vs. Infinity-RoPE (77.3% / 22.7%) and vs. LongLive (69.3% / 30.7%). The dashed line marks 50% (chance).

Ablation Study

Ablation study on Slow Memory Sampling Strategy, Context DMD, and Bounded Positional Encoding (evaluated on 60s).

Our Context DMD and Bounded RoPE increase the long video consistency and mitigating temporal drift during the generation process.

Model	Total Score ↑	Quality Score ↑	Semantic Score ↑	Background Consistency ↑	Subject Consistency ↑	Dynamic Degree ↑
Slow Memory Sampling Strategy						
Uniform sample, interval 1	80.82	82.20	<u>75.32</u>	92.45	92.10	52.15
Uniform sample, interval 2	<u>81.11</u>	<u>82.61</u>	75.12	93.12	92.85	<u>55.30</u>
Contextual Distillation						
w/o. Contextual Distillation	80.36	82.28	72.70	<u>93.55</u>	<u>93.20</u>	48.12
Bounded Positional Encoding						
w/o. Bounded Positional Encoding	73.52	75.44	65.82	84.68	79.24	27.45
Ours	**82.45**	**83.55**	**76.10**	**95.34**	**94.88**	**58.26**

Concluding Remarks

- We propose **Context Forcing**, a framework that enables **consistent long-video generation** by aligning student and teacher context lengths.
- Paired with a **Slow-Fast Memory system**, it achieves 2--10x **longer context information** than current state-of-the-art methods.

Future Works

- Context compression/Learnable memory module
- Apply Context Forcing into more complex setting(world model/multi shot video gen)
- improve robustness against error drifting.