LongLive

- □□□□□LONGLIVE: REAL-TIME INTERACTIVE LONG VIDEO GENERATION
- □□□□□https://arxiv.org/abs/2509.22622
- ICLR 2026

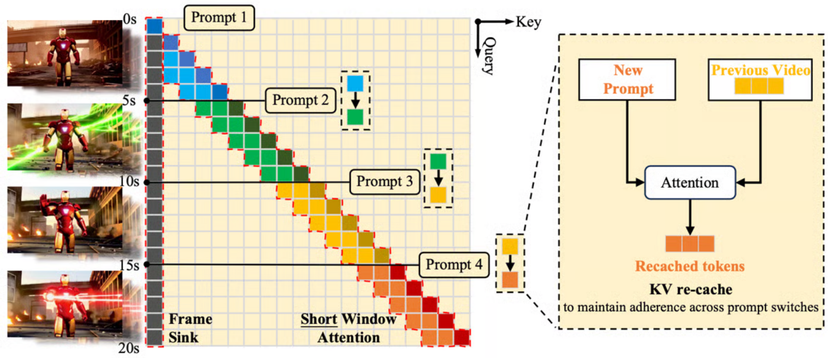

Figure 2: The framework of LONGLIVE. (Left) LONGLIVE processes sequential user prompts and generates a corresponding long video using efficient **short window attention** and **frame sink**. Compared to the normal attention window of 5s, our short window only uses half the size, with the help of frame sink, which maintains the long-range consistency. (Right) To maintain consistency when the prompt switches, LONGLIVE employs a **KV-recache** technique that updates cached key–value states by combining previous videos with new prompt embeddings through cross-attention layers.

□□□□□LongLive□□□□□□□□□□□□□□□□□□□□□□□□□□□

□□□KV cache□□□prompt□□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□LongLive□□□□□□□□□□□□□□□□□□□□□□KV-recache□□□□□prompt□□train-long-test-long□□□□□frame sink□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□attention□□□□prompt□□□□□□□□□□□□□□□□□□□□□□□□□□□

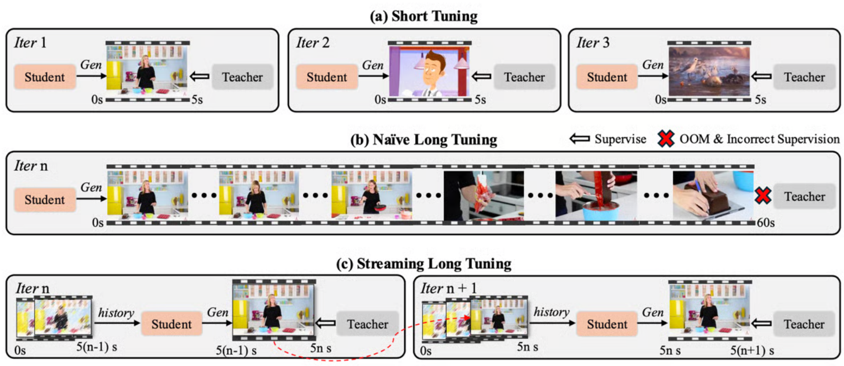

Figure 4: The streaming long tuning pipeline. (a) **Short tuning**: only 5s clips are supervised, like Self-Forcing (Huang et al., 2025), leading to quality loss on long videos. (b) **Naïve long tuning**: naively scaling to long sequences causes incorrect teacher supervision and OOM. (c) **Streaming long tuning**: our approach trains on long sequences by reusing the historical KV cache each iteration to generate the next 5s clip, then supervising it with the teacher.

□□clip□□□□□clip□□□□□□□□□□□□□clip□□□□clip□□□□□□□□□□□□□□□□□□□□□□□□frame sink□□□□□□□KV-cache□□□□□□□□

- □□□VidProM□□□prompt□□Qwen2□□□□□prompt□□□□□
- □□□12h 64 H100
- □□□https://github.com/NVlabs/LongLive

Newer Older

2026-5-6
CausVid

2026-5-6
ShotStream