EventBank-G: Compact Event Memory for Controllable Multi-Shot Video Generation

Asad Khan Anoop Rehman Rohan Kanti Arpan Bhattacharya Mahbod Sabbaghi

The Intelligent Search Company

Abstract

Long-horizon video generation is often assembled from short clips, but independently prompted clips drift in state while full-story prompts blur adjacent events. We repurpose EventBank from a video-understanding representation into EventBank-G, a model-agnostic event memory layer for generation. Given an ordered script, EventBank-G builds compact event tokens that pair the current action with a persistent state ledger for subject, scene, lighting, and camera style. These tokens condition a frozen text-to-video model one shot at a time. We introduce TimeLens-Stories, a protocol that converts temporally ordered TimeLens annotations into multi-shot generation scripts, and evaluate event fidelity, event-order discriminability, state adherence, and visual continuity. Across two remote LTX-Video spot runs covering 50 stories and 150 generated shots per method, EventBank-G improves event fidelity over independent prompting (.2570 vs. .2505) and state adherence (.2458 vs. .2400), while avoiding the wrong-event leakage that appears when full-story or neighboring-event context is injected into every shot. The gains are modest and CLIP-based; we therefore frame EventBank-G as an inference-time control primitive and benchmark probe, not as a finished long-video generator.

1. Introduction

The Frames to Stories workshop focuses on long-horizon video generation: models must preserve state, follow temporal structure, support control, and be evaluated beyond single-clip image quality. This setting exposes a practical gap. Text-to-video models can generate plausible short clips, but longer videos are commonly produced as a sequence of shots. If each shot is prompted independently, identity, location, and style may reset. If every shot receives the full story, later and earlier actions contaminate the current generation.

We turn EventBank into a generation method by changing what the bank is used for. In video understanding, event tokens summarize an observed video and are searched later. For generation, the bank is created before rendering: it is a storyboard memory that says what should happen now, what should persist, and how this shot relates to neighboring events. The key hypothesis is not that event memory improves single-shot visual quality, but that a compact event state improves the long-horizon properties that matter for controllable story generation. Importantly, EventBank-G is an inference-time prompting layer: it does not train a memory module, modify the generator architecture, or retrieve from generated frames. This narrow design choice is deliberate because it isolates whether compact state tokens help a frozen text-to-video backbone.

Our contributions are: (1) EventBank-G, a frozen-generator prompting layer that converts reusable event tokens into generation-time memory; (2) TimeLens-Stories, a script-construction protocol from ordered TimeLens annotations; (3) a workshop-scale evaluation separating event fidelity from event-order confusion and persistent state; and (4) a leakage analysis showing why full-story and neighbor-conditioned prompts can over-condition the current shot.

2. EventBank for Generation

Given a story with K ordered events, EventBank-G constructs an event bank

$$\mathcal{E} = \{e_1, \ldots, e_K\}, \qquad e_k = (a_k, s),$$

where a_k is the current action and s is a persistent state ledger. The ledger records recurring subject, scene, lighting, camera, and style constraints. Unlike temporal grounding, there is no span prediction and no video encoder training: the bank is an inference-time control representation. We also test an expanded token $e_k^+ = (a_k, p_k, n_k, s)$, where p_k and n_k summarize neighboring events, to ask whether local narrative context helps or merely leaks future and past

Accepted to the ICML 2026 Workshop on From Frames to Stories (F2S). Non-archival workshop paper.

actions into the current shot.

State ledger construction. The ledger is deterministic rather than manually written. Given the ordered TimeLens event texts, we infer a coarse recurring subject from a fixed keyword list (e.g., woman, man, person, child, people, cook, player, speaker, performer), infer a setting from action words such as kitchen/home/outdoor cues, and append fixed continuity constraints: same clothing, same lighting, same camera style, and realistic handheld video. For example, a kitchen sequence yields a ledger such as "one consistent person, same indoor kitchen or home room, same clothing, same lighting, same camera style, realistic handheld video." This rule-based ledger keeps the benchmark reproducible, but it is intentionally simple; sensitivity to noisy or learned ledgers remains open.

Prompt families. We compare four ways to use the same ordered script. *Independent* uses only the current event. *Full-story* prepends the entire script to every shot. *EventBank-G + neighbors* adds the previous and next event text with an instruction not to render them. *EventBank-G* keeps only the persistent state and current action. This compact variant is designed for finite text-token budgets, where long story prompts can be truncated or can over-condition the current shot.

3. TimeLens-Stories

TimeLens-Bench (Zhang et al., 2025) was designed for temporal grounding, but its annotations already describe ordered events in real videos. We convert it into a generation benchmark by sorting annotations within each video by timestamp and selecting the first three events from videos with at least three annotations. Each resulting item is a three-shot script grounded in everyday human activity, e.g., opening a pantry door, laughing, and checking a bottle. This produces realistic action sequences without collecting a new video-generation dataset. The three-shot setting is not a full long-film benchmark. We use it as a controlled workshop-scale test of state persistence and wrong-event leakage under fixed generation cost.

4. Experiments

Remote generation. All video generation was run remotely with SkyPilot managed spot jobs; no local GPU computation was used. We used the frozen LTX-Video text-to-video model (Lightricks, 2024), generated 50 three-shot TimeLens-Stories per method across two completed remote runs, and fixed seeds across prompt families within each story. Each clip used 25 frames at 320×576 resolution with 25 denoising steps and guidance scale 3.5. The two runs used different story samples and seeds, then were aggregated for the final table.

Table 1. TimeLens-Stories generation results with a frozen LTX-Video backbone over 50 stories and 150 shots per method. Lower Leak is better. EventBank-G has the strongest Event, Margin, and Top-1 among tested prompts and improves State over independent prompting, but independent prompting has lower Leak and higher Adjacent similarity.

Method	Event	Leak↓	Margin	Top-1	State	Adj.	Motion
Independent	.2505	**.2400**	.0105	.6133	.2400	.8068	.0083
Full-story	.2502	.2594	-.0092	.3733	.2488	.8510	.0049
EventBank-G + neighbors	.2546	.2568	-.0023	.4333	**.2522**	.8036	.0050
EventBank-G	**.2570**	.2458	**.0111**	**.6267**	.2458	.7747	.0067

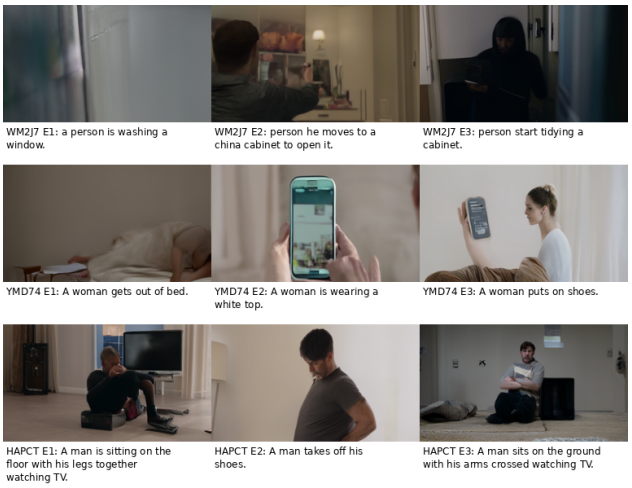

Figure 1. Generated frames from compact EventBank-G. Each row is one TimeLens-Stories script and each column is the rendered shot for the corresponding event token.

Metrics. We evaluate generated keyframes with CLIP ViT-B/32 (Radford et al., 2021). *Event* is CLIP similarity to the current event text. *Leak* is the strongest wrong-event similarity from the same story. *Margin* is Event minus Leak, measuring event-order discriminability. *Top-1* is whether the current event is the closest event text. *State* is similarity to the persistent state ledger. *Adjacent* is visual embedding similarity between neighboring shots. *Motion* is average frame difference, included as a guard against static outputs. All confidence intervals below are story-level paired bootstraps.

5. Results and Discussion

Table 1 supports three conclusions. First, compact event memory modestly improves the stateful generation objective that independent prompting misses. Compared with independent prompting, EventBank-G raises Event from .2505 to .2570 and State from .2400 to .2458. Paired bootstrap deltas are positive for both Event (+.0065, 95% CI [.0019,.0111]) and State (+.0058, 95% CI [.0027,.0090]).

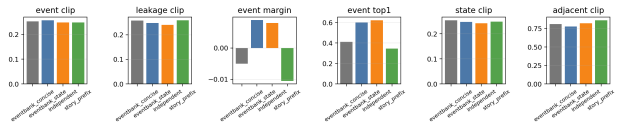

Figure 2. Metric comparison across prompt families. Leak, Margin, and Top-1 expose whether a plausible frame depicts the intended story moment or a neighboring event; final aggregate values are reported in Table 1.

The Event-order margin is effectively tied with independent prompting (+.0006, CI crosses zero), so we do not claim that state memory alone solves ordering against a single-event prompt. Adjacent visual similarity is also lower for EventBank-G than for independent and full-story prompts; we therefore do not present visual continuity as a solved outcome.

Second, full-story prompting is a misleading baseline. Its negative Margin and 37.3% Top-1 show that CLIP often matches the generated frame to the wrong event in the same script. This is exactly the failure mode a long-horizon video system should avoid: the shot contains plausible story content, but not necessarily the right story moment. Relative to full-story prompting, EventBank-G reduces Leak by .0135 and improves Margin by .0203 with positive paired intervals.

Third, more memory is not automatically better. EventBank-G + neighbors has the highest State score, but weak Margin and Top-1. Simply naming previous and next events, even with a negative instruction, makes the current shot more similar to those wrong events. This makes compact memory a methodological point rather than a convenience: long-horizon control has to preserve persistent state without crowding out the current event.

6. Limitations

This is a workshop-scale study. The evaluation uses CLIP keyframe metrics rather than human preference, VLM judging, or full-video temporal metrics, so the numerical gains should be read as proxy evidence rather than perceptual proof. The absolute Event and State improvements over independent prompting are small, although paired intervals are positive. The benchmark also tests short three-shot stories with one frozen LTX-Video backbone rather than long generated films or multiple generators, so longer-script scalability remains untested. Finally, the method is prompting-only: it does not feed generated keyframes back into the next shot, use explicit identity preservation, retrieve from a memory of rendered frames, compare against hierarchical or recurrent prompting baselines, ablate ledger components, or train a generator with event tokens. These are the main extensions needed for a fuller long-horizon system.

7. Conclusion

EventBank-G reframes EventBank as compact generation memory. Instead of localizing events in an observed video, it stores the event state needed to render a coherent sequence of generated shots. The strongest evidence is not a large visual-quality gain, but a more precise control result: compact persistent state improves event and state adherence over independent prompting while avoiding the leakage of full-story and neighbor-conditioned prompts. The resulting direction fits the F2S agenda: persistent state, controllable multi-shot generation, and evaluations that distinguish single-shot plausibility from story correctness.

References

Lightricks. LTX-Video. https://huggingface.co/Lightricks/LTX-Video, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Zhang, J., Wang, T., Ge, Y., Ge, Y., Li, X., Shan, Y., and Wang, L. Timelens: Rethinking video temporal grounding with multimodal LLMs. *arXiv preprint arXiv:2512.14698*, 2025.

A. Supplementary Prompt Details

Independent prompt. The independent baseline prompts each shot with only the current TimeLens event: "Photorealistic live-action smartphone video with a real human actor in a real-world environment. Current action: [event]. Natural motion, natural lighting, no captions, no text overlays."

Full-story prompt. The full-story baseline prepends the complete three-shot script: "Continuous photorealistic live-action smartphone video story with real people. Full story plan: Shot 1: [event 1] Shot 2: [event 2] Shot 3: [event 3]. Current shot: [event]. Preserve visual realism, coherent motion, no captions, no text overlays."

EventBank-G prompt. The compact event-memory prompt uses the current event and persistent state only: "Photorealistic live-action smartphone shot k/K. Keep persistent state: [ledger]. Only depict this action: [event]. Natural motion, natural lighting, no captions, no text overlays."

EventBank-G + neighbors prompt. The neighbor ablation adds local story context: "Photorealistic live-action smartphone shot k/K. Keep persistent state: [ledger]. Continuity context: this shot follows [previous event]; this shot precedes [next event]. Only show the current action: [event]. Do not show the previous or next action. No captions, no text overlays."

B. Supplementary Reproducibility Details

For each selected TimeLens video, events are sorted by timestamp and the first three annotations are retained. Stories are shuffled with a fixed seed, and every prompt family uses the same story-level seed plus the same event-index offset. Evaluation averages three keyframes per generated clip before computing CLIP similarities. Event, Leak, Margin, Top-1, State, Adjacent, and Motion are first aggregated per story; confidence intervals resample stories rather than individual frames. The reported table combines the 20-story and 30-story state-focused LTX-Video runs, giving 50 stories and 150 shots per method.