Global Time, Local Blocks: Generating Minute-Long Videos from Action Schedules

Yang Chen[1] Chengsong Qu[1] Runting Li[1]
Xuemei Zhou[2] Teng Long[3] Vincent Tao Hu[1]

[1] School of Software Engineering, Huazhong University of Science and Technology
[2] Centrum voor Wiskunde en Informatica (CWI) [3] University of Trento

Project Page: https://vibe-hust.github.io/timeweave_page/
Code: https://github.com/vibe-hust/timeweave_code

Abstract

Generating long videos from action schedules requires each action to occur within its prescribed interval, preserve intended overlaps with other actions, and remain attached to the correct participant throughout a minute-long rollout. Rolling video generators receive only a single prompt or a sequence of prompt switches, while interval-conditioned models are limited to short bidirectional clips. We introduce TimeWeave*, which generates long videos from explicit event timelines with a few-step causal generator distilled on 10-second sequences. Two parameter-free conditioning mechanisms connect this short-sequence training to minute-long inference.* Rolling Timeline Conditioning *(RTC) supplies each generation block with the actions whose intervals intersect a window around it, keeping their original intervals on a global timeline so that ongoing and concurrent actions are neither clipped nor rescaled.* Global Context Anchoring *(GCA) keeps scene and identity descriptions untimed and lets them share one softmax with the interval-conditioned actions, so persistent context remains available throughout the rollout. On 128 one-minute event timelines, TimeWeave achieves the best scores on all automatic metrics, improving the Temporal-Following Score from 0.564 to 0.610 and temporal localization mIoU from 0.200 to 0.251 over the strongest baseline, and human raters rank it above Self-Forcing and LongLive in event fulfillment, timing, continuity, and visual quality. Code and video examples are available at https://vibe-hust.github.io/timeweave_page/.*

1. Introduction

Video diffusion models now synthesize realistic clips of a few seconds from text [1, 9, 13, 26, 34]. Extending them to minute-long narratives raises three coupled challenges: generating far beyond the training horizon, controlling *what* happens *when* at a fine temporal granularity, and preventing errors from accumulating as the rollout grows. Consider a one-minute video in which person A pours coffee from 6 to 14 seconds while person B reads a newspaper from 4 to 30 seconds: each action must occur within its interval, preserve intended overlap, and stay attached to the right participant across the minute. We study *long video generation from action schedules*.

Existing approaches to long video generation follow two main directions. The first builds on a strong short-clip generator and extends it by continuation, hierarchical synthesis, or noise rescheduling [8, 12, 19, 27, 36]. The second converts a pretrained short-clip model into a causal generator that rolls out block by block on top of its own cached history [3, 4, 11, 23, 33, 38]. Training directly on minute-long videos would be the most direct solution, but the quadratic cost of full spatiotemporal attention makes it impractical; rolling post-training on short sequences, as in Self-Forcing [11] and LongLive [33], offers a far more affordable path and is the setting we adopt.

Yet visual horizon extension alone does not yield structured narrative control. Causal rollouts achieve scalability by restricting visual self-attention to a local block and a rolling cache, whereas existing text conditioning remains coarse: a single prompt cannot schedule multiple events, and discrete prompt switching [33] partitions time into disjoint segments that reset context at switch boundaries, sever actions spanning blocks, and conflate concurrent participants. Continuous interval encodings [24, 31] bind captions to overlapping spans, yet naively distilling an interval-conditioned model into a causal student fails at minute scale: off-interval actions crowd cross-attention, and full-rollout persistent spans distort positional response.

We introduce *TimeWeave* to address both issues. A language-model agent converts a narrative prompt into a structured timeline separating time-invariant scene and participant anchors from subject-bound timed actions, rather than inferring persistence from span. Next, we distill a few-step causal student from a bidirectional interval-conditioned teacher solely on 10-second clips [11, 38], bridged to minute-long inference by two parameter-free mechanisms. *Rolling Timeline Conditioning* (RTC) decouples *which* actions enter each block from *where* they sit in time, pruning actions outside a sliding window W_k while preserving unclipped, unscaled global intervals for active and overlapping spans. *Global Context Anchoring* (GCA) routes persistent anchors through untimed logits in the same softmax as interval-modulated actions, preserving scene and identity context throughout.

On 128 one-minute event timelines, TimeWeave, trained only on 10-second sequences, achieves the best scores on all automatic metrics, improving the Temporal-Following Score from 0.564 to 0.610 and temporal localization mIoU from 0.200 to 0.251 over the strongest baseline, LongLive Interactive [33] (Table 2). In a human study, raters also rank TimeWeave above Self-Forcing and LongLive in event fulfillment, timing, continuity, and visual quality (Table 5), and removing RTC reduces the Temporal-Following Score from 0.610 to 0.476 (Table 4).

In summary, our contributions are:

- We formulate long video generation from action schedules and present TimeWeave, coupling agent-structured timelines with a causal generator distilled on short clips for minute-long rollouts.
- We identify the two-scale positional mismatch between short-clip interval encoding and long causal rollouts, and resolve it with Rolling Timeline Conditioning and Global Context Anchoring—two parameter-free mechanisms whose local-clip isomorphism we establish analytically and verify empirically.
- We construct a 60-second benchmark of single- and multi-subject timelines with temporal localization metrics and show, through automatic metrics, a human study, and ablations, that TimeWeave follows event timelines more faithfully than existing long video generators.

2. Related Work

Video Diffusion Models. Diffusion models have become the dominant approach to video synthesis. Early video diffusion models extend image U-Nets with temporal layers [9], and latent formulations scale training to large datasets [1]. Recent systems adopt diffusion transformers [17] with full spatiotemporal attention over a compressed video latent, including CogVideoX [34], Hunyuan-Video [13], and Wan [26, 28]; the latter two use flow matching [15]. Because full attention scales quadratically with frame count, these models jointly denoise 5–10 s clips and cannot be trained directly at minute scale; we distill our causal generator from an interval-conditioned Wan2.2-TI2V-5B [28] teacher.

Extending short-clip generators. One line of work extends a pretrained short-clip generator without changing its bidirectional formulation. Phenaki [27] generates variable-length videos from a sequence of prompts with a masked token model, and NUWA-XL [36] synthesizes long videos hierarchically through coarse-to-fine diffusion. StreamingT2V [8] extends videos chunk by chunk with short- and long-term memory modules, while FreeNoise [19] and FIFO-Diffusion [12] reach longer horizons in a training-free manner through noise rescheduling and diagonal denoising, respectively.

Autoregressive video diffusion. A second line converts video diffusion into causal, block-wise generation. Diffusion Forcing [3] trains sequence models with independently sampled per-token noise levels, and Rolling Diffusion [22] denoises a sliding window with noise levels that increase toward later frames; MaskFlow [7] applies related ideas with discrete flows. CausVid [38] distills a bidirectional teacher into a few-step causal student through ODE initialization and distribution matching [37], and Self-Forcing [11] closes the train–test gap by training on self-generated histories with KV caching. LongLive [33] extends streaming generation to minute-long videos through long-video tuning, short-window attention, and KV recaching at prompt switches, and LongLive-2.0 [5] improves the efficiency of this setting. Large-scale systems such as MAGI-1 [23] and SkyReels-V2 [4] adopt chunk-wise autoregressive or diffusion-forcing generation, and FramePack [39] compresses visual history to reduce drift. While these methods stabilize long visual rollouts, they rely on coarse single-prompt or segment-switched text; TimeWeave instead makes fine-grained timeline conditioning locally isomorphic to short-clip causal distillation.

Positional extrapolation. Length extension has also been studied via rotary position embeddings [25]: YaRN [18] and LongRoPE [6] interpolate frequencies, StreamingLLM [32] bounds cached positions, RIFLEx [41] adjusts temporal frequency to curb repetition, and Infinity-RoPE [35] reanchors visual frames. Whereas these works modify visual token positions, TimeWeave preserves global text–video coordinates without time rescaling or frequency modification.

Temporal and Multi-Event Control. Multi-prompt methods assign successive prompts to disjoint segments: MEVG [16] and DiTCtrl [2] smooth prompt transitions

Table 1. **TimeWeave unifies start/end interval conditioning, overlapping events, and causal KV-cached generation.** ✓ and ✗ indicate whether each method's original formulation provides the mechanism.

Method	Separate event prompts	Start/end intervals	Event overlap	Causal KV cache
Self Forcing [11]	✗	✗	✗	✓
LongLive 1.0 [33] (interactive)	✓	✗[a]	✗	✓
MinT [31]	✓	✓	✗[b]	✗
TIE [24]	✓	✓	✓	✗
TimeWeave (ours)	✓	✓	✓	✓

[a]LongLive changes the active prompt at a switching point; it does not encode each event's complete start/end interval in cross-attention. [b]MinT's time rescaling assigns each frame to one sequential event interval, without independent schedules for concurrent events.

without retraining, VideoDirectorGPT [14] plans multi-scene scripts with an LLM, and LongLive [33] switches prompts during streaming generation, but ordered segments cannot schedule overlapping multi-subject spans independently. Interval-conditioned models make timing explicit on short bidirectional clips: MinT [31] uses rescaled rotary embeddings with separate temporal and global cross-attention layers, while TIE [24] encodes overlapping spans via duration-normalized rotary interval encoding (RoTE). Table 1 contrasts these formulations: TimeWeave unifies continuous, overlapping intervals with causal KV-cached rollouts by locally routing global-time actions (RTC) alongside untimed persistent anchors in one shared softmax (GCA).

3. Method

TimeWeave generates minute-long videos from explicit event timelines using a few-step causal generator distilled solely on short clips. We first motivate the conditioning design (Section 3.1), then define agent-constructed timelines (Section 3.2) and introduce Rolling Timeline Conditioning (Section 3.3) and Global Context Anchoring (Section 3.4), followed by short-clip causal distillation and inference (Section 3.5).

3.1. Motivation: conditioning beyond the training horizon

Extending interval cross-attention from 10-second training clips to a 60-second causal rollout creates different problems for transient actions and persistent context. Actions need their original timing and a relevant local context, while scene and identity descriptions must remain available throughout the video. Figure 1 isolates these requirements through pre-softmax positional logits under uniform text content. The following analysis motivates the two conditioning mechanisms before we define the event representation and generation procedure.

We build on an interval-conditioned video diffusion model [24], where concatenated text descriptions pass their intervals $[s, e]$ to their tokens. For a video query $\mathbf{q}$ at physical time τ and a text key–value pair $(\mathbf{k}, \mathbf{v})$, cross-attention encodes $\mathbf{q}$ with RoPE [25] and $\mathbf{k}$ with rotary interval encoding (RoTE), leaving $\mathbf{v}$ unchanged:

$$\ell_{\text{int}}(\mathbf{q}, \mathbf{k}; \tau, [s, e]) = \frac{\big(\mathbf{R}(\tau)\mathbf{q}\big)^\top \mathbf{R}(c)\, \mathbf{D}(r)\, \mathbf{k}}{\sqrt{d}}, \quad (1)$$

where $c = (s + e)/2$ and $r = (e - s)/2$ are the interval center and half-length. Across $M = d/2$ rotary frequency pairs $\{\omega_m\}_{m=0}^{M-1}$ with temporal scale γ, $\mathbf{R}(t)$ rotates pair m by $\gamma\omega_m t$ and $\mathbf{D}(r)$ scales it by the mean-normalized sinc kernel $d_m(r) = \text{sinc}(\gamma\omega_m r)/C_r$, where $C_r = \frac{1}{M}\sum_{n=0}^{M-1} \text{sinc}(\gamma\omega_n r)$ (Appendix B). Since $\mathbf{R}(\tau)^\top \mathbf{R}(c) = \mathbf{R}(c - \tau)$, the logit depends on time only through the offset $c - \tau$ and half-length r. Under unit-energy, content-matched query–key pairs, Equation (1) reduces to the pure positional response $S(\tau; [s, e]) = \frac{1}{M}\sum_{m=0}^{M-1} \cos\big(\gamma\omega_m(c - \tau)\big) d_m(r)$, plotted in Figure 1 to isolate positional geometry.

Two-scale breakdown on long rollouts. Suppose a causal student is distilled on short clips of horizon T_{train} (10 s) and rolled out to $T \gg T_{\text{train}}$ (60 s). Evaluating $S(\tau; [s, e])$ reveals opposing failure modes for transient actions and persistent scene and identity descriptions: *(i) Transient actions.* Individual actions typically satisfy $2r \lesssim T_{\text{train}}$. Because $S(\tau; [s, e])$ is shift-invariant in (τ, s, e), keeping unscaled global time gives a late action (e.g., 46–54 s) the exact onset and offset sharpness learned inside a 10 s clip, whereas rescaling $[0, T] \to [0, T_{\text{train}}]$ contracts $r \mapsto (T_{\text{train}}/T)r$ and blurs its boundaries (Figure 1a). Yet $S(\tau; [s, e])$ decays only to a positive baseline (≈ 0.75) outside $[s, e]$ rather than zero; passing all N actions in a timeline to every block therefore accumulates $\mathcal{O}(N)$ off-interval competitors in the softmax denominator (Figure 1c, top-left). *(ii) Persistent invariants.* Scene and identity descriptions hold across the entire video regardless of T. Encoding them as a bounded span $[0, T]$ sets $r = T/2$, inflating r six-fold at $T = 60$ s and driving $d_m(T/2)$ far outside its T_{train} calibration—including mid-frequency sign inversions (Figure 1b)—which depresses persistent logits toward $\tau = 0$ and $\tau = T$ (Figure 1c, top row).

Full-timeline synthesis and local-clip isomorphism. Figure 1c visualizes how both mechanisms interact on a 60 s two-participant script with three persistent anchors (G0–G2) and nine sequential or concurrent actions (E1–E9, prescribed spans in white boxes). Native RoTE (top-left) suffers simultaneously from persistent-row droop toward $\tau =$

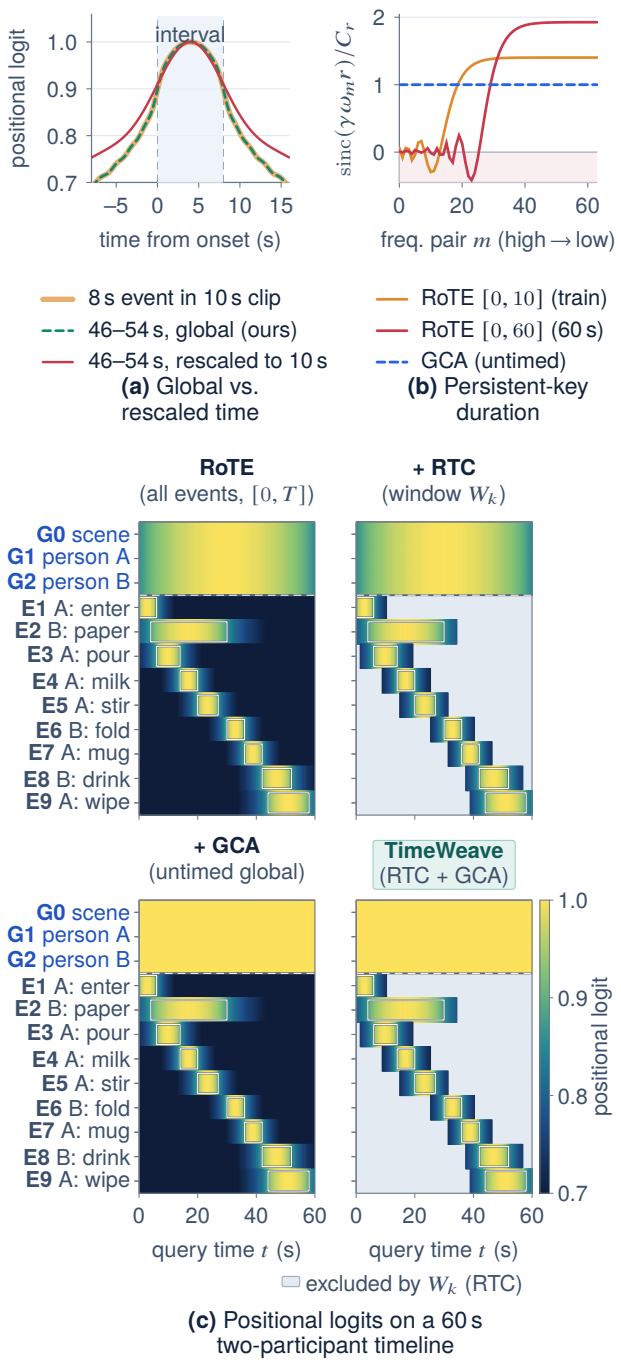

Figure 1. **TimeWeave preserves short-clip positional geometry on minute-long rollouts by locally filtering global-time actions and keeping persistent context untimed.** Pre-softmax positional logits under uniform content: (a) global vs. rescaled action timing, (b) persistent-key duration modulation across rotary frequencies, and (c) resulting responses across a 60 s two-participant timeline (white boxes: prescribed intervals; light gray: excluded by W_k).

$0, 60$ s and dense off-interval action competition across the minute. Adding RTC alone (top-right) excludes actions disjoint from a local window W_k around the current generation block prior to text encoding (light-gray cells) while keeping unclipped global profiles on overlapping spans such as E2 (4–30 s) and E3–E5, yet leaves G0–G2 attenuated at both ends; adding GCA alone (bottom-left) restores a flat unit response (1.0) on G0–G2 without removing distant-action crowding. Combining both in TimeWeave (bottom-right) achieves *local-clip isomorphism*: each causal block attends only to locally active global-time spans (Section 3.3) alongside stationary, T-independent persistent anchors in one shared softmax (Section 3.4). Figure 2 summarizes how these mechanisms fit into the causal generator; we next define its timeline inputs and each mechanism.

3.2. Problem formulation and event timelines

A video of duration T seconds is specified by an event timeline $\mathcal{C} = (\mathcal{G}, \mathcal{E})$. The persistent set $\mathcal{G} = \{g_j\}_{j=1}^{M}$ contains a scene description and one appearance description per participant; these hold for the entire video and carry no interval. The action set $\mathcal{E} = \{(u_i, d_i, s_i, e_i)\}_{i=1}^{N}$ contains, for each action, a participant identifier u_i, a self-contained description d_i of a visible action, and its prescribed interval $[s_i, e_i] \subseteq [0, T]$ in seconds. Actions may be sequential, concurrent, or staggered, expressing joint interactions via time-aligned actions across tracks. Generating videos from these timelines requires each action to occur on the right participant within its interval while scene and identities remain consistent; video time is distinct from diffusion noise level.

From narratives to timelines. A language-model agent converts a narrative prompt into $\mathcal{C}$, separating scene and appearance facts from dynamics, decomposing dynamics into observable subject-explicit actions, and placing them on the timeline with intended order and concurrency. Persistent and action descriptions receive explicit role tags rather than duration heuristics. Short training clips and minute-long test prompts share this decomposition, aligning text distributions without a separate trained rewriter (Appendix A).

3.3. Rolling Timeline Conditioning

The generator produces the video block by block. Let block k cover the interval $B_k = [a_k, b_k]$ on the global timeline, from the start of its first latent to the end of its last. RTC supplies block k with a condition $\mathcal{C}_k$ built from a window

$$W_k = [\ell_k, h_k] = \big[\max(0, a_k - \rho),\ \min(T, b_k + \rho)\big], \quad (2)$$

where the margin $\rho \geq 0$ adds nearby past and future context. The condition contains every persistent description and every action whose interval intersects the window:

$$\mathcal{C}_k = \mathcal{G} \cup \{(u_i, d_i, s_i, e_i) \in \mathcal{E} : e_i \geq \ell_k,\ s_i \leq h_k\}. \quad (3)$$

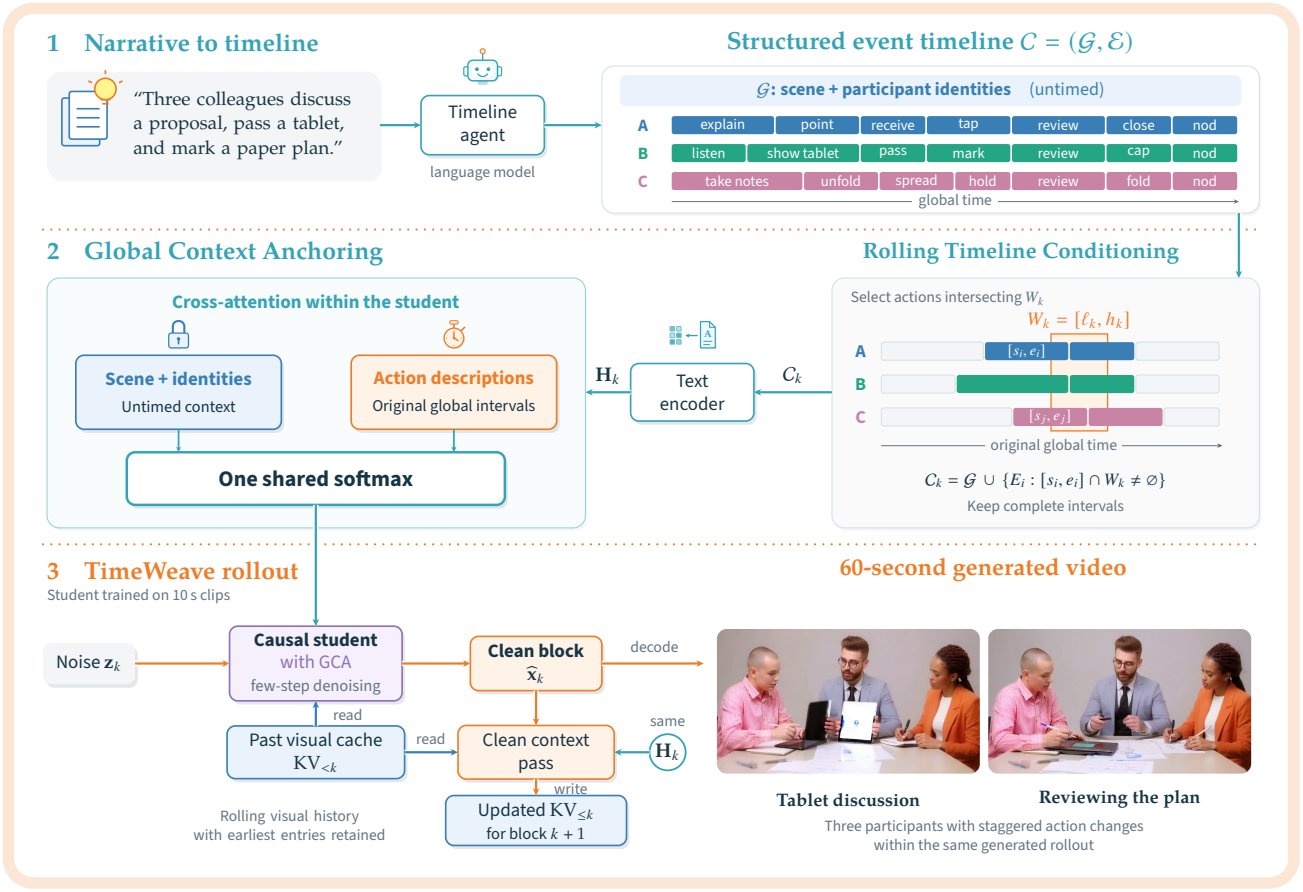

Figure 2. **TimeWeave bridges 10-second causal distillation and minute-long timeline following with two parameter-free conditioning mechanisms.** *(1) Narrative to timeline:* A language-model agent constructs persistent scene and identity descriptions and participant-specific action intervals. *(2) Conditioning:* RTC selects actions intersecting W_k before text encoding, preserving complete global intervals; GCA combines untimed context and interval-conditioned actions in one shared softmax inside the student. *(3) Causal rollout:* The student denoises each block using the current condition and retained visual cache. A clean context pass uses the same condition and writes the cache for the next block. The example shows three participants with staggered action changes.

Selection uses full intervals, keeping ongoing spans active and selecting overlapping actions together.

Global timeline. Selected actions keep their original intervals $[s_i, e_i]$, and video queries keep their global times τ_j; neither is clipped to W_k nor rescaled to the training horizon. Window edges therefore never become artificial action boundaries. By the relative-offset form of Equation (1), shifting a query and an action by a common Δ leaves the logit unchanged:

$$\ell_{\text{int}}(\mathbf{q}, \mathbf{k}; \tau + \Delta, [s + \Delta, e + \Delta]) = \ell_{\text{int}}(\mathbf{q}, \mathbf{k}; \tau, [s, e]). \quad (4)$$

For fixed content features, an action late in a long video thus receives the same positional profile as an action of equal duration inside a training clip, whereas rescaling time to the training horizon would shrink its duration and blur its boundaries (Figure 1 a and c). The global timeline preserves the frame–action offsets that the student saw during training; RTC restricts the set of attended actions so that the text condition of each block resembles a training-length timeline. Since changing selected subsets and cached history also alters content features, Equation (4) is positional and verified empirically.

Text encoding. Selection precedes text encoding, keeping text length bounded for any video duration; persistent descriptions and selected actions are repacked in source order, with $\mathcal{C}_k$ fixed across block denoising and cache update. Each block retains persistent context, complete intervals for cross-block actions, and past visual outcomes in the rolling cache.

3.4. Global Context Anchoring

RTC decides which actions are visible; GCA decides how persistent descriptions are represented. Instead of assigning them an interval, GCA removes temporal encoding from their logits while keeping them in the same attention as the actions. Let $\mathcal{K}_g$ and $\mathcal{K}_e$ index the text tokens of persistent descriptions and selected actions. For a video query at time τ,

$$\begin{aligned} \ell_{g,n} &= \frac{\mathbf{q}^\top \mathbf{k}_n}{\sqrt{d}}, && n \in \mathcal{K}_g, \\ \ell_{e,n} &= \ell_{\text{int}}\big(\mathbf{q}, \mathbf{k}_n; \tau, [s_n, e_n]\big), && n \in \mathcal{K}_e, \\ \mathbf{o} &= \sum_{n \in \mathcal{K}_g \cup \mathcal{K}_e} \text{softmax}\big([\boldsymbol{\ell}_g; \boldsymbol{\ell}_e]\big)_n \, \mathbf{v}_n. \end{aligned} \tag{5}$$

Persistent logits involve neither τ nor any interval, so their positional contribution is constant over the rollout and independent of T; no duration modulation is applied (Figure 1b,c). Actions retain native RoTE. Both groups share a single softmax, so persistent and timed descriptions compete for the same attention mass within one layer, unlike designs that process global and temporal captions in separate cross-attention layers [31].

GCA reuses the backbone's query, key, and value projections and adds no parameters. It is implemented exactly as one standard attention call by concatenating features: the query becomes $[\mathbf{q}, \mathbf{R}(\tau)\mathbf{q}]$, a persistent key becomes $[\mathbf{k}, \mathbf{0}]$, an action key becomes $[\mathbf{0}, \mathbf{R}(c)\mathbf{D}(r)\mathbf{k}]$, values are zero-padded to the same width, and the softmax scale stays $1/\sqrt{d}$ of the original head width. This avoids materializing dense logits. Without GCA, persistent descriptions are encoded by native RoTE with support $[0, T]$, which is the configuration removed in our ablation (Table 4).

3.5. Causal distillation and long-rollout inference

Short-sequence causal distillation. We distill a causal student from the frozen bidirectional interval-conditioned teacher solely on short sequences. The student shares the teacher's weights and architecture, replacing bidirectional video attention with block-causal self-attention and native persistent-text RoTE with GCA, while the teacher and auxiliary score model keep native interval encoding for all descriptions (Table S1). Following CausVid [38] and Self-Forcing [11], training proceeds in two stages (Appendix C): we first update all student parameters by regressing clean endpoints from independently noised blocks along guided teacher ODE trajectories, and then fine-tune the student with LoRA adapters [10] on KV-cached rollouts via distribution matching distillation [37] with the frozen teacher and a LoRA flow-matching [15] score network.

Long-rollout inference. At inference, the student generates the video block by block with a few denoising steps per block, far beyond the training horizon. Each block receives its RTC condition $\mathcal{C}_k$. After a block is generated, a context pass with the same condition writes its keys and values to a rolling visual cache, whose earliest entries are retained as an attention sink [32]. Text–video cross-attention keeps the global timeline throughout.

4. Experiments

4.1. Experimental setup

Implementation. We build on Wan2.2-TI2V-5B [28] and use its bidirectional interval-conditioned variant [24] as the frozen teacher. The student is trained only on 10-second clips (16 fps, 1280×704): ODE initialization updates all parameters for 3,000 iterations, and Self-Forcing trains rank-128 LoRA adapters for 600 iterations on 64 H100 GPUs. At inference, TimeWeave generates 60 s videos with four steps per block, an RTC margin $\rho = 4$ s matching the 10 s training horizon, and a 41-latent rolling cache.

Benchmark. We use 128 one-minute event timelines with one (32), two (64), or three (32) participants and 1,532 actions in total (12.0 per timeline, mean duration 7.9 s). The benchmark stresses concurrency: 62% of all actions overlap another action, and two or more actions run in parallel during 57% and 71% of the two- and three-participant timelines. A typical action is shorter than a training clip, while the timeline is six times longer.

Baselines. Self-Forcing [11] and LongLive [33] take the full timeline as one timestamped prompt; LongLive Interactive [33], the strongest temporal-control baseline, switches prompts to the active actions plus scene and identity text. All methods use official weights and the same structured scene, identity, and action descriptions.

Metrics. The Temporal-Following Score (TFS) [29] uses Qwen3-VL [20] to rate sampled frames against the timed reference. CLIP [21] and InternVideo2 [30] measure semantic similarity inside each requested interval. Motion Rationality (MR) adapts VBench-2.0 [42] to action intervals. For 1,276 adjacent action pairs, a vision-language judge checks presence (AP), pairwise order (Joint Success, JS), and temporal span; Time mIoU scores missing actions as zero and Loc. mIoU averages over localized ones, both penalizing early onsets and lingering actions (Appendix D).

4.2. Comparison on one-minute event timelines

TimeWeave achieves the best score in every column of Table 2. Three observations explain the gains.

Table 2. **TimeWeave outperforms all baselines on every metric across 128 one-minute event timelines, with the largest gains in action presence and temporal localization.** Time mIoU scores missing actions as zero, whereas Loc. mIoU averages over localized actions (1,276 adjacent pairs). Bold: best; underlined: second best.

Method	TFS	CLIP	InternVideo2	MR (%)	AP (%)	JS (%)	Time mIoU	Loc. mIoU
Self-Forcing [11]	0.3824	0.2421	0.0266	51.95	57.04	6.64	0.0916	0.1793
LongLive [33]	0.4595	0.2437	0.0263	53.17	59.00	6.24	0.0976	0.1840
LongLive Interactive [33]	<u>0.5644</u>	<u>0.2462</u>	<u>0.0307</u>	<u>70.35</u>	<u>70.06</u>	<u>19.28</u>	<u>0.2003</u>	<u>0.3291</u>
TimeWeave (Ours)	**0.6098**	**0.2473**	**0.0332**	**74.22**	**78.73**	**23.39**	**0.2508**	**0.3622**



Figure 3. **TimeWeave sustains accurate action timing and participant count through the end of a 60-second rollout, whereas baselines miss late actions or hallucinate extra participants.** On two-participant sample 067, TimeWeave executes both synchronized knee leans at 44.5 s (A) and camera waves at 57.5 s (B), while baselines remain upright, lag on an earlier pose, or introduce a third person. Green solid: matched action; red dashed: mismatch or artifact; six more examples appear in Appendix F.

Continuous intervals outperform discrete prompt switches. Packing textual timestamps into a single prompt (Self-Forcing, LongLive) leaves cross-attention largely blind to event timing (Time mIoU < 0.10, JS $< 7\%$). LongLive Interactive recovers coarse ordering by switching prompts at discrete boundaries (JS 19.3%), yet each segment receives an untimed prompt that drops continuous onsets, durations, and actions spanning a switch. Trained only on 10-second clips without minute-long tuning, TimeWeave surpasses LongLive Interactive on every metric (TFS 0.564 to 0.610, JS 19.3% to 23.4%, MR 70.4% to 74.2%) by preserving each selected action's continuous interval on the global timeline.

The gains are in *when*, not only *what*. CLIP and InternVideo2 vary little across methods (CLIP 0.242–0.247), as they score semantic agreement within an interval and barely penalize an action that appears at the wrong time. Timing-sensitive metrics separate the methods: over LongLive Interactive, TimeWeave raises AP from 70.1% to 78.7% and Time mIoU from 0.200 to 0.251, while Loc. mIoU improves only from 0.329 to 0.362. Much of the gain therefore comes from actions that are now executed at all, not merely from tighter boundaries.

The advantage grows with concurrency. On single-participant timelines (Table 3), actions are strictly sequen-

Table 3. **TimeWeave's advantage over prompt switching and the gain from GCA both grow with multi-participant concurrency.** Results on 1, 2, and 3 participants (32, 64, and 32 timelines; header percentages indicate time with ≥ 2 overlapping actions; MR and AP in %, T-mIoU: Time mIoU).

	1 participant (0%)				**2 participants** (57%)				**3 participants** (71%)			
Method	TFS	MR	AP	T-mIoU	TFS	MR	AP	T-mIoU	TFS	MR	AP	T-mIoU
LongLive Interactive	0.638	75.8	76.8	0.287	0.534	71.9	69.3	0.194	0.552	61.7	67.0	0.168
TimeWeave w/o RTC	0.449	67.2	71.7	0.230	0.483	55.7	59.3	0.129	0.489	49.6	59.9	0.110
TimeWeave w/o GCA	**0.661**	76.3	78.7	0.337	**0.562**	66.3	69.0	0.193	0.621	56.1	69.1	0.189
TimeWeave	0.614	**82.4**	**83.8**	**0.340**	0.541	**72.8**	**77.6**	**0.253**	**0.744**	**68.9**	**77.4**	**0.205**

Table 4. **RTC drives temporal alignment, while GCA substantially improves motion rationality and action presence.** Ablation on the 128 one-minute timelines under the protocol of Table 2, removing one component at a time.

Variant	**TFS**	**CLIP**	**InternVideo2**	**MR (%)**	**AP (%)**	**JS (%)**	**Time mIoU**	**Loc. mIoU**
w/o RTC	0.4759	0.2452	**0.0348**	57.04	61.89	12.79	0.1402	0.2685
w/o GCA	0.6011	**0.2477**	0.0318	66.26	70.87	22.23	0.2157	0.3571
Ours (full)	**0.6098**	0.2473	0.0332	**74.22**	**78.73**	**23.39**	**0.2508**	**0.3622**

Table 5. **Human raters prefer TimeWeave across every criterion, with the largest margins in event fulfillment and timing.** Mean ranking scores in $[0, 100]$ (100/50/0 for 1st/2nd/3rd place; protocol in Appendix E).

Method	**Event**	**Timing**	**Continuity**	**Quality**	**Overall**
Self-Forcing	2.08	2.17	1.85	1.85	0.00
LongLive	52.08	54.35	61.11	70.37	64.81
TimeWeave	**95.83**	**93.48**	**87.04**	**77.78**	**85.19**

tial and prompt switching is competitive: LongLive Interactive even obtains a higher TFS (0.638 vs. 0.614), although TimeWeave localizes actions better. As overlapping actions of several participants appear, a single active prompt can no longer express them, and the gap widens to 0.552 vs. 0.744 in TFS and 67.0% vs. 77.4% in AP for three participants—the exact regime targeted by global-timeline intervals.

4.3. Human evaluation and qualitative results

Table 5 ranks TimeWeave first in every criterion. Mirroring automatic metrics, margins are largest for Event and Timing, smaller for Continuity, and smallest for Quality (77.8 vs. 70.4): raters mainly favor action and timing fidelity without reduced visual quality. Figure 3 illustrates two failures that accumulate in long rollouts: late in the minute, the baselines miss the requested actions or hallucinate a third participant, whereas TimeWeave preserves both (further examples in Appendix F).

4.4. Ablation study

Table 4 tests *w/o RTC* by conditioning every block of the same student on the complete timeline; *w/o GCA* uses an identically trained student with native interval encoding and full-video support for persistent descriptions.

RTC routes actions in time. Removing RTC drops TFS from 0.610 to 0.476 and Time mIoU from 0.251 to 0.140, placing the full-timeline student below LongLive Interactive on both metrics. AP also falls from 78.7% to 61.9%, despite similar CLIP and InternVideo2 scores. These results show that local action selection improves both action presence and temporal alignment. Without RTC, JS remains above the single-prompt baselines (12.8% vs. 6.2–6.6%), but rises to 23.4% with local selection.

GCA keeps multi-participant actions executable. Removing GCA changes overall TFS only slightly (0.610 to 0.601), but lowers MR by 8.0 and AP by 7.9 percentage points. In Table 3, GCA raises AP by 5–9 points in every group; MR gains grow from 6.1 (one participant) to 12.8 (three). TFS falls for one and two participants but rises from 0.621 to 0.744 for three. Thus, GCA improves action presence and motion rationality across groups, with its clearest TFS benefit under the highest concurrency.

5. Conclusion

TimeWeave generates minute-long videos from explicit event timelines using a causal generator distilled solely on 10-second clips. Our analysis and experiments indicate that

faithful narrative control far beyond training requires block conditioning locally isomorphic to short training clips. RTC selects locally relevant transient events without clipping or rescaling global intervals; GCA decouples persistent scene and identity anchors from rollout duration in one shared softmax. Both are parameter-free, with the largest gains over prompt switching under multi-participant concurrency.

Limitations. Time-invariant global scene and participant descriptions suit continuous single-scene videos but cannot directly express scene transitions, camera cuts, or persistent appearance changes. RTC extrapolates timelines far beyond 10-second training with actions near training duration; substantially longer atomic actions remain outside the trained range unless decomposed into shorter phases. Evolving global states, interactive mid-rollout replanning, and synchronized audio remain promising future directions.

LLM Usage Statement

We used OpenAI Codex to edit and refine author-produced text, experimental designs and analyses, figures and visual materials, data-related materials, and research ideas, claims, and conclusions. For related work and citations, and for code and implementation, we used tool-generated work through one-shot or interactive assistance; the authors reviewed, edited, or approved that work. We did not provide the full manuscript or major parts of it to an LLM provider to obtain reviewer-style feedback. The authors take full responsibility for all submitted content, regardless of any LLM or AI-tool use.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2

[2] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, et al. DiTCtrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. *arXiv preprint arXiv:2412.18597*, 2024. 2

[3] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024. 1, 2

[4] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 1, 2

[5] Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, Yuyang Zhao, Huizi Mao, Ying-Cong Chen, Enze Xie, Xiaojuan Qi, and Song Han. LongLive-2.0: An NVFP4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026. 2

[6] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *ICML*, 2024. 2

[7] Michael Fuest, Vincent Tao Hu, and Björn Ommer. MaskFlow: Discrete flows for flexible and efficient long video generation. *arXiv preprint arXiv:2502.11234*, 2025. 2

[8] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, et al. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 1, 2

[9] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022. 1, 2

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 6, 5

[11] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 1, 2, 3, 6, 7, 5

[12] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. FIFO-Diffusion: Generating infinite videos from text without training. In *Advances in Neural Information Processing Systems*, 2024. 1, 2

[13] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2

[14] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. In *COLM*, 2024. 3

[15] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 2, 6, 5

[16] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. MEVG: Multi-event video generation with text-to-video models. In *European Conference on Computer Vision*, 2024. 2

[17] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[18] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *ICLR*, 2024. 2

[19] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *ICLR*, 2024. 1, 2

[20] Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. 6

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[22] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 42818–42835. PMLR, 2024. 2

[23] Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, et al. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 1, 2

[24] Zhilei Shu, Shangwen Zhu, Zihang Liang, Xiaofan Li, Qianyu Peng, Xinyu Cui, Bo Ye, Yiming Li, Fan Cheng, Jian Zhao, Yang Cao, Zheng-Jun Zha, and Ruili Feng. TIE: Time interval encoding for video generation over events. *arXiv preprint arXiv:2605.10543*, 2026. 1, 3, 6, 4, 5

[25] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 2, 3

[26] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2

[27] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, et al. Phenaki: Variable length video generation from open domain textual description. In *International Conference on Learning Representations*, 2023. 1, 2

[28] Wan-AI. Wan2.2-TI2V-5B. Official model card, 2025. 2, 6, 5

[29] Jing Wang, Xiangxin Zhou, Jiajun Liang, Kaiqi Liu, Wanyuan Pang, Zhenyu Xie, Tianyu Pang, and Xiaodan Liang. TempAct: Advancing temporal plausibility in autoregressive video generation via planner-executor RL. *arXiv preprint arXiv:2606.28016*, 2026. 6

[30] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024. 6

[31] Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 3, 6

[32] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. 2, 6, 5

[33] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025. 1, 2, 3, 6, 7

[34] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. 1, 2

[35] Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-RoPE: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *CVPR*, 2026. 2

[36] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, et al. NUWA-XL: Diffusion over diffusion for extremely long video generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. 1, 2

[37] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 2, 6, 5

[38] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. 1, 2, 6

[39] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *NeurIPS*, 2025. 2

[40] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data, 2024. 6

[41] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. RIFLEx: A free lunch for length extrapolation in video diffusion transformers. In *ICML*, 2025. 2

[42] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. 6

Global Time, Local Blocks: Generating Minute-Long Videos from Action Schedules

Supplementary Material

Contents

Main Paper

1. Introduction 1

2. Related Work 2
- Video Diffusion Models. 2
- Extending short-clip generators. 2
- Autoregressive video diffusion. 2
- Positional extrapolation. 2
- Temporal and Multi-Event Control. 2

3. Method 3
- 3.1. Motivation: conditioning beyond the training horizon 3
  - Two-scale breakdown on long rollouts. 3
  - Full-timeline synthesis and local-clip isomorphism. 3
- 3.2. Problem formulation and event timelines 4
  - From narratives to timelines. 4
- 3.3. Rolling Timeline Conditioning 4
  - Global timeline. 5
  - Text encoding. 5
- 3.4. Global Context Anchoring 6
- 3.5. Causal distillation and long-rollout inference 6
  - Short-sequence causal distillation. 6
  - Long-rollout inference. 6

4. Experiments 6
- 4.1. Experimental setup 6
  - Implementation. 6
  - Benchmark. 6
  - Baselines. 6
  - Metrics. 6
- 4.2. Comparison on one-minute event timelines 6
  - Continuous intervals outperform discrete prompt switches. 7
  - The gains are in *when*, not only *what*. 7
  - The advantage grows with concurrency. 7
- 4.3. Human evaluation and qualitative results 8
- 4.4. Ablation study 8
  - RTC routes actions in time. 8
  - GCA keeps multi-participant actions executable. 8

5. Conclusion 8
- Limitations. 9

Appendix

A Agent-based recaptioning and timeline construction **2**
A.1. Semantic decomposition and multi-track consistency . . . 2
A.2 One-minute multi-participant benchmark construction . . . 2

B Interval-conditioned backbone and temporal translation **3**
B.1. Training roles and cross-attention routing . . . 3
B.2. Interval encoding and global temporal translation . . . 4

C Training and long-rollout inference **5**
C.1. Stage 1: ODE initialization . . . 5
C.2. Stage 2: Self-Forcing distribution matching . . . 5
C.3. Minute-long causal inference . . . 5

D Datasets and evaluation details **5**
D.1. Corpus inventory and baseline conditioning . . . 6
D.2. Automatic evaluation metrics . . . 6

E Human evaluation protocol **7**
E.1. Stimuli and blind viewing procedure . . . 7
E.2. Evaluation criteria and ranking aggregation . . . 7

F. Additional qualitative comparisons **8**

G Additional statements **11**

A. Agent-based recaptioning and timeline construction

Following Section 3.2, a language-model agent converts a narrative prompt and target duration T into a structured timeline $\mathcal{C} = (\mathcal{G}, \mathcal{E})$ that explicitly separates time-invariant context $\mathcal{G}$ from timed, subject-bound actions $\mathcal{E} = \{(u_i, d_i, s_i, e_i)\}_{i=1}^{N}$. We apply the same structural representation to both short training clips and minute-long evaluation prompts.

A.1. Semantic decomposition and multi-track consistency

The agent extracts invariant scene setting and participant appearance traits into $\mathcal{G}$ without timestamps, while assigning postures, state changes, and observable movements to $\mathcal{E}$ with explicit participant identifiers u_i. Persistent descriptions never assert transient states that a subsequent action alters, and each action clause d_i names its acting participant and visible object motion directly so it remains self-contained under rolling selection. Rather than partitioning the rollout into equal-length segments or inferring persistence from duration heuristics, the agent assigns distinct role tags to persistent and timed entries and chooses $[s_i, e_i] \subseteq [0, T]$ to reflect plausible preparation, execution, and recovery phases.

Across concurrent tracks, the agent enforces physical and causal coherence before accepting a timeline. Actions involving shared objects or bodily contact must agree on object possession and hand availability, preventing a single participant from performing mutually incompatible concurrent motions. Joint multi-participant interactions—such as passing an object, holding a shared prop, or synchronizing a gesture—are represented by time-aligned action records on the participating tracks whose intervals overlap during contact or coordination.

A.2. One-minute multi-participant benchmark construction

The 128 one-minute evaluation timelines (32 one-, 64 two-, and 32 three-participant scenes; 1,532 actions in total, averaging 12.0 actions per timeline with a mean duration of 7.9 s) follow this exact decomposition. Each 60-second prompt starts from a 10-second source scene and its anchored participants, then introduces compatible sequential, staggered, and concurrent actions to build a coherent minute-long narrative—such as extending a brief beverage-pouring scene with stirring, taking a bite, wiping hands, and conversing—rather than unnaturally stretching single actions to fill 60 seconds. All timelines pass pre-generation checks on interval bounds, participant continuity, and causal ordering; no generated-video scores are used to filter or select prompts.

B. Interval-conditioned backbone and temporal translation

This section expands the interval-conditioned cross-attention and global time-translation property used in Sections 3.1, 3.3, and 3.4.

B.1. Training roles and cross-attention routing

Figure S1 illustrates the interval-conditioned backbone [24] shared across the frozen teacher and the auxiliary score network. In bidirectional native interval encoding, every video query at frame time τ is rotated by point-time rotary embeddings (RoPE), every text key with interval $[s, e]$ is rotated and modulated by interval center $c = (s + e)/2$ and half-duration $r = (e - s)/2$ (RoTE), and text values remain untimed. During teacher classifier-free guidance, only the conditional branch carries interval metadata, while the unconditional branch receives an empty negative prompt.

Table S1 summarizes how the causal student differs from the bidirectional score models. Reusing the exact backbone projections, the student replaces bidirectional video self-attention with block-causal attention over a rolling visual cache and routes persistent anchors $\mathcal{G}$ (along with padding tokens) through the untimed branch of Equation (5). Concatenating queries as $[\mathbf{q}, \mathbf{R}(\tau)\mathbf{q}]$, persistent keys as $[\mathbf{k}, \mathbf{0}]$, and action keys as $[\mathbf{0}, \mathbf{R}(c)\mathbf{D}(r)\mathbf{k}]$ with zero-padded values computes Equation (5) in a single standard attention call at softmax scale $1/\sqrt{d}$ without adding parameters.

Table S1. **GCA modifies only the causal student's text cross-attention, while the teacher and auxiliary score model retain native interval encoding.** Attention masks, conditioning modes, and trainable parameters across training roles.

Role	Video attention	Text conditioning	Updated parameters
ODE student	Block-causal	GCA + action interval encoding	Full model
SF generator	Block-causal	GCA + action interval encoding	LoRA
ODE / SF teacher	Bidirectional	Native interval encoding	Frozen
SF auxiliary score	Bidirectional	Native interval encoding	LoRA

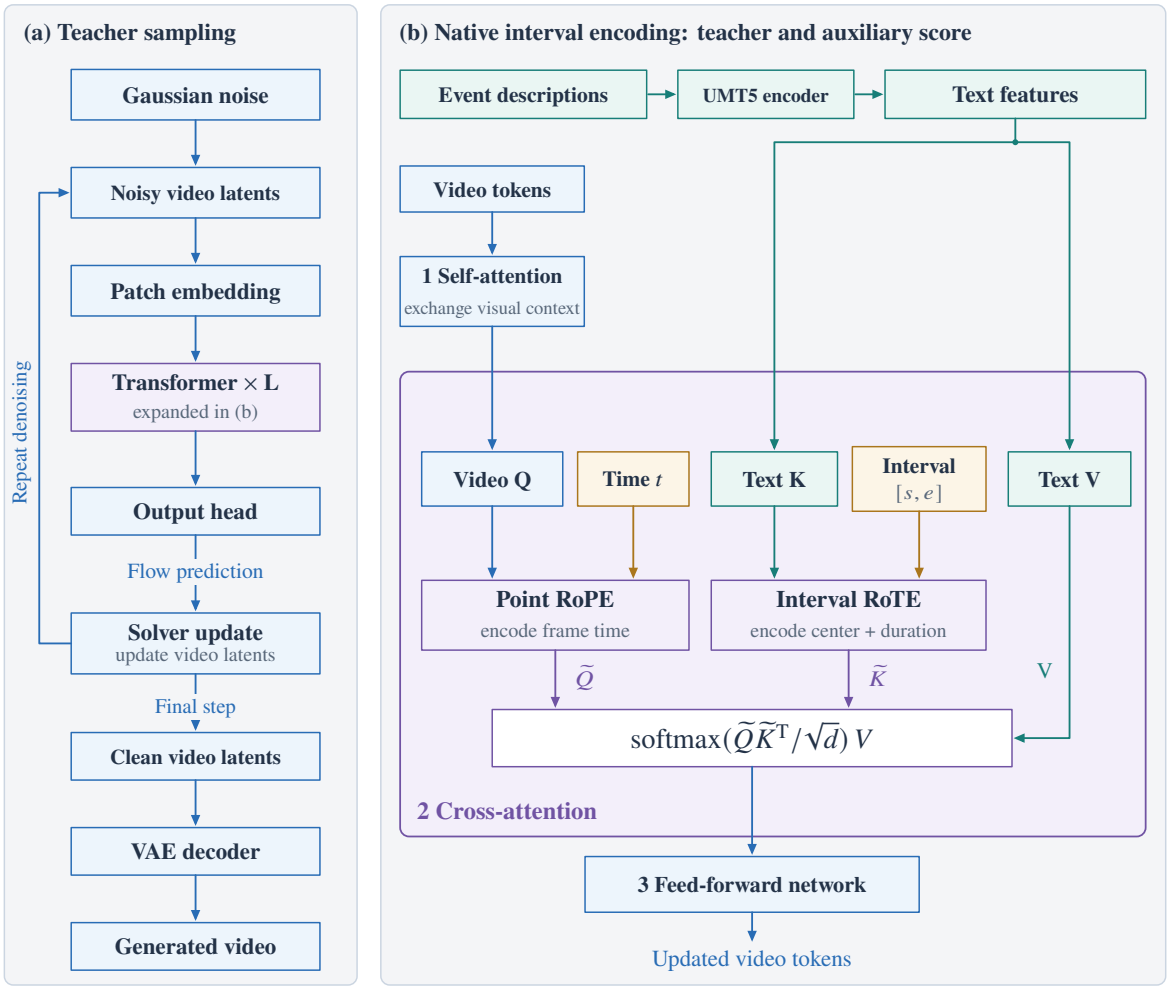

Figure S1. **The teacher and auxiliary score network encode point time on video queries and interval center and duration on text keys.** (a) Bidirectional flow-prediction stack and VAE decoding. (b) Transformer block with video self-attention and interval-conditioned cross-attention ($\widetilde{Q}$: RoPE-encoded queries; $\widetilde{K}$: RoTE-encoded keys; V: unmodified text values; gold boxes indicate physical time). The student's untimed global branch is defined in Equation (5).

B.2. Interval encoding and global temporal translation

For an attention head of even width d with pre-rotary content vectors $\mathbf{q}, \mathbf{k} \in \mathbb{R}^d$ and action interval $[s, e]$ of center $c = (s+e)/2$ and half-duration $r = (e-s)/2$, let $\mathbf{R}(t)$ denote the block-diagonal 2×2 rotary matrix with frequency-pair angles $\omega_m t$ (absorbing the backbone's fixed time scale into ω_m). The normalized duration modulation [24] is

$$\mathbf{D}(r) = \operatorname{diag}\left(\frac{\operatorname{sinc}(\omega_1 r)}{C_r}\mathbf{I}_2, \ldots, \frac{\operatorname{sinc}(\omega_{d/2} r)}{C_r}\mathbf{I}_2\right), \qquad C_r = \frac{2}{d}\sum_{m=1}^{d/2} \operatorname{sinc}(\omega_m r), \tag{6}$$

where $\operatorname{sinc}(x) = \sin(x)/x$ is continuously extended to 1 at $x = 0$ and $C_r \neq 0$. Because rotary blocks satisfy $\mathbf{R}(\tau)^\top \mathbf{R}(c) = \mathbf{R}(c - \tau)$, the interval-conditioned cross-attention logit in Equation (1) evaluates to

$$\ell_{\text{int}}(\mathbf{q}, \mathbf{k}; \tau, [s, e]) = \frac{1}{\sqrt{d}}\, \mathbf{q}^\top \mathbf{R}(c - \tau)\mathbf{D}(r)\mathbf{k}. \tag{7}$$

Shifting both the query time and action interval by a common offset Δ maps $(\tau, c, r) \mapsto (\tau + \Delta, c + \Delta, r)$, leaving both $c - \tau$ and $\mathbf{D}(r)$ invariant and proving Equation (4). Meanwhile, persistent logits $\ell_g = d^{-1/2}\mathbf{q}^\top \mathbf{k}$ in Equation (5) depend on neither τ nor total rollout length T. Both statements characterize the explicit positional transforms at fixed content features; changing the selected text subset or visual history still alters the full attention output.

C. Training and long-rollout inference

This section provides the exact two-stage 10-second distillation recipe and the block-causal minute-long inference schedule summarized in Section 3.5 and Section 4.1.

C.1. Stage 1: ODE initialization

We build on Wan2.2-TI2V-5B [28] and use its bidirectional interval-conditioned variant [24] as a frozen teacher. At 16 fps and 1280×704 resolution with temporal stride four, each 10-second clip spans 41 video latents (161 decoded frames), partitioned into causal blocks of $[5, 6, 6, 6, 6, 6, 6]$ latents. Latent query times τ_j equal cell centers $[0, 1/f]$ for $j = 0$ and $[(4j-3)/f, (4j+1)/f]$ for $j > 0$ at frame rate f.

From 4,000 ten-second source timelines (Table S2), the teacher integrates 48 Euler steps with flow shift 5 and classifier-free guidance scale 5, storing intermediate states $\{\mathbf{z}_\sigma^{\mathrm{T}}\}$ and clean endpoint $\mathbf{z}_0^{\mathrm{T}}$ at model timesteps $\{1000, 937.5, 833.3333, 625, 0\}$. Initializing from teacher weights, the student converts video self-attention to block-causal masking and enables GCA on persistent descriptions while keeping all 10-second clip events visible. For each causal block b, we independently sample a trajectory state at noise level σ_b into $\widetilde{\mathbf{z}}$ and train all student parameters to regress the clean endpoint:

$$\mathcal{L}_{\mathrm{ODE}} = \mathbb{E}\left[\frac{\sum_b m_b \left\|\mathbf{D}_{\theta,b}(\widetilde{\mathbf{z}}, \boldsymbol{\sigma}, \mathcal{C}) - \mathbf{z}_{0,b}^{\mathrm{T}}\right\|_2^2}{\sum_b m_b n_b}\right], \qquad m_b = \mathbf{1}[\sigma_b > 0], \tag{8}$$

where n_b counts scalar entries in block b. Initialization runs for 3,000 AdamW steps (global batch size 128 across 64 H100 GPUs, constant learning rate 2×10^{-6}, weight decay 0.01, gradient clipping at 10, no warmup or EMA), with BF16 network compute and FP64 physical-time metadata.

C.2. Stage 2: Self-Forcing distribution matching

Starting from the 3,000-step ODE checkpoint, we perform rollout-based distribution matching distillation [11, 37] on 172,478 eligible 10-second timelines. Fresh rank-128 LoRA adapters [10] ($\alpha = 128$, zero dropout) are attached to both the causal student generator and a bidirectional flow-matching [15] auxiliary score model initialized from the teacher, while base weights and the real-score teacher remain frozen.

At each rollout block, the generator executes a 4-step clean-prediction and renoising trajectory at timesteps $[1000, 937.5, 833.3333, 625]$ without classifier-free guidance: intermediate prediction $\widehat{\mathbf{z}}_{0,i}$ is renoised as $\mathbf{z}_{\sigma_{i+1}} = (1 - \sigma_{i+1})\widehat{\mathbf{z}}_{0,i} + \sigma_{i+1}\boldsymbol{\epsilon}_i$, followed by a clean timestep-zero pass that writes block keys and values into a 41-latent KV cache with a 5-latent attention sink [32]. One denoising step per rollout receives supervision, with prior steps and cached historical blocks detached. Let $\widehat{\mathbf{z}}$ be the generated clip and $\mathbf{z}_\sigma = (1 - \sigma)\widehat{\mathbf{z}} + \sigma\boldsymbol{\epsilon}$ its renoised input at a shift-5 timestep clamped to $[20, 980]$. Comparing the guided teacher $\mathbf{D}_{\mathrm{T}}$ (CFG scale 4) against the conditional auxiliary score prediction $\mathbf{D}_\phi$, we optimize

$$\mathbf{g}_n = \frac{\mathbf{D}_{\phi,n} - \mathbf{D}_{\mathrm{T},n}}{\max(\mathrm{mean}_{T,C,H,W}\,|\widehat{\mathbf{z}}_n - \mathbf{D}_{\mathrm{T},n}|,\ 10^{-6})}, \qquad \mathcal{L}_{\mathrm{DMD}} = \frac{1}{2}\,\mathrm{MSE}(\widehat{\mathbf{z}}, \mathrm{sg}[\widehat{\mathbf{z}} - \mathbf{g}])\,. \tag{9}$$

Training runs for 600 outer steps on 64 H100 GPUs at global batch size 64: the auxiliary score updates every step (learning rate 2×10^{-6}) and the generator updates once every five steps (120 generator updates total, learning rate 10^{-5}), both using AdamW with $(\beta_1, \beta_2) = (0, 0.999)$ and weight decay 0.01.

C.3. Minute-long causal inference

At inference, TimeWeave generates 60-second videos (241 video latents, 961 frames at 16 fps and 1280×704) block by block—an initial 5-latent block followed by 6-latent (1.5 s) blocks—using four denoising steps per block and a timestep-zero context write into the 41-latent rolling visual KV cache, whose earliest entries are retained as an attention sink [32]. With RTC margin $\rho = 4$ s, each interior 6-latent block spans an active conditioning window W_k of 9.5 s, matching the 10-second training horizon. Selection precedes UMT5 text encoding: persistent anchors $\mathcal{G}$ and selected actions $\mathcal{E}_k$ are repacked in source order while retaining their unclipped, unscaled global intervals $[s_i, e_i] \subseteq [0, 60]$, and encoded text representations are cached whenever consecutive blocks share an identical subset $\mathcal{C}_k$.

D. Datasets and evaluation details

This section documents the prompt cohorts, baseline inference configurations, and automatic metric definitions underlying Tables 2, 3, and 4.

D.1. Corpus inventory and baseline conditioning

Table S2 summarizes the 10-second distillation sets and the 128 one-minute evaluation timelines (32 one-, 64 two-, and 32 three-participant scenes containing 1,532 action intervals with boundaries on a 0.25 s grid). Every method is evaluated across all 128 timelines at 16 fps over a 60-second rollout (961 frames, 241 latents) using fixed seed 4042, four denoising steps per block, and an empty negative prompt. All three baselines run with their official Wan2.1-1.3B weights at their native 832×480 resolution with 3-latent blocks [11, 33], using LongLive's 12-latent rolling window and 3-latent attention sink for both LongLive variants.

All methods consume the same structured scene, participant, and action descriptions. Self-Forcing [11] and standard LongLive [33] receive the complete one-minute schedule as one prompt with decimal timestamps (concisely formatted where needed so all text fits inside the 512-token encoder budget without truncation). LongLive Interactive [33] updates its prompt at each block midpoint to the currently active actions alongside persistent scene and participant text. TimeWeave retains the complete persistent context via GCA and filters actions per block via RTC ($\rho = 4$ s).

Table S2. **TimeWeave is distilled exclusively on 10-second timeline conditions and evaluated on 60-second rollouts.** Condition counts and durations across training, validation, and evaluation sets.

Use	Conditions	Duration (s)
Eligible Self-Forcing training pool	172,478	10
ODE initialization	4,000	10
ODE validation	128	10
Main automatic comparison	128	60

D.2. Automatic evaluation metrics

Temporal-Following Score (TFS). We adapt TempAct's framewise evaluation rubric [29] with Qwen3-VL-8B-Instruct [20] under deterministic decoding. Forty frames are sampled uniformly across the 60-second rollout (1.5 s spacing) on a shared 1280×704 canvas and paired with explicit frame timestamps and the full timed reference. The judge outputs six $[0, 10]$ sub-scores normalized to $[0, 1]$ and combined with TempAct's published weights:

$$\mathrm{TFS} = 0.20\overline{F} + 0.30G + 0.20O + 0.10P + 0.10C + 0.10A, \tag{10}$$

where $\overline{F}$ is the frame-level mean, G global assessment, O temporal order, P physical plausibility, C visual consistency, and A prompt alignment. Table S3 reports each unweighted component across all 128 timelines.

Table S3. **TimeWeave leads on every sub-score of the Temporal-Following Score (TFS).** Unweighted component scores in $[0, 1]$ and their TFS weights on the 128 one-minute timelines.

Method	TFS	$\overline{F}$ 20%	G 30%	O 20%	P 10%	C 10%	A 10%
Self-Forcing [11]	0.3824	0.6304	0.2984	0.3289	0.3578	0.3477	0.3047
LongLive [33]	0.4595	0.6724	0.3750	0.4273	0.4516	0.4492	0.3695
LongLive Interactive [33]	0.5644	0.7760	0.4828	0.5188	0.5695	0.5633	0.4734
TimeWeave (Ours)	**0.6098**	**0.8511**	**0.5164**	**0.5680**	**0.6008**	**0.6000**	**0.5094**

Interval semantic similarity and Motion Rationality. Within each requested action interval $[s_e, e_e)$, CLIP ViT-L/14 [21] (nearest frame at each 0.25 s bin) and InternVideo2 Stage2-6B [30] (4-frame 1 s support centered on each bin) measure raw cosine similarity against the action text and referenced participant identity, averaged equally over actions per video and then across all 128 videos. Motion Rationality (MR) adapts VBench-2.0's binary conjunction [42] to each action interval: LLaVA-Video-7B-Qwen2 [40] inspects 64 uniformly sampled frames inside $[s_e, e_e)$ and answers three independent yes/no questions verifying (1) execution of the requested action, (2) plausible human body poses and continuous motion, and (3) plausible object contact without interpenetration. An action passes only when all three answers are yes, and MR averages the per-video pass rate in percent across all 128 videos.

Adjacent-action presence, ordering, and temporal localization. Sorting each participant's timeline yields 1,276 consecutive non-identical action pairs (A, B) with $e_A \leq s_B$ (2,552 pair-context action observations). For each pair window $[s_A, e_B]$, Qwen3-VL-8B-Instruct inspects 4 fps video frames (576×320 canvas) given only randomized `X`/`Y` labels for the source-derived core actions—withholding target timestamps and expected order. Stage 1 determines whether each dynamic action is `present`, `absent`, or `uncertain` (static holding alone does not count as execution), yielding Action Presence Rate (AP) over decidable observations. Stage 2 localizes all observed execution intervals $[s^*, e^*]$ for detected actions; Loc. mIoU averages the temporal IoU against $[s, e]$ over localized actions, while Time mIoU assigns zero to confirmed absent actions, so both metrics directly penalize premature starts and lingering actions. A pair achieves Joint Success when both actions are localized with $s_A^* < s_B^*$ and $e_A^* \leq s_B^* + 0.25$ s. Table 2 reports JS among decidable pairs, whereas Table S4 reports full-pair JS bounds across all 1,276 pairs by counting unresolved outcomes as either all failures or all successes.

Table S4. **TimeWeave localizes the most adjacent action pairs and achieves the highest joint ordering success.** Decidable presence counts (out of 2,552) and localization coverage across 1,276 adjacent pairs; JS bounds are in %.

Method	Decidable presence	Both localized	Successes	Unresolved	JS bounds
Self-Forcing [11]	2544/2552	321	71	206	[5.56, 21.71]
LongLive [33]	2544/2552	343	66	219	[5.17, 22.34]
LongLive Interactive [33]	2538/2552	<u>371</u>	<u>178</u>	353	[<u>13.95</u>, <u>41.61</u>]
TimeWeave (Ours)	**2548/2552**	**438**	**196**	438	[**15.36**, **49.69**]

E. Human evaluation protocol

This section documents the stimuli, blind interface, and ranking aggregation underlying Table 5 (TimeWeave vs. Self-Forcing and LongLive across ten benchmark scenes).

E.1. Stimuli and blind viewing procedure

Each trial compares synchronized 40-second (16 fps) excerpts of the same scene at native aspect ratio, anonymized as A/B/C and cyclically permuted across sessions so each method occupies every slot equally often. Raters see the scene and participant descriptions, interval-timed action subtitles, and the excerpt schedule (actions starting at or after 40 s are omitted; actions crossing the cutoff are marked as continuing) and must complete ≥ 39.5 s of unique playback per clip before submitting.

E.2. Evaluation criteria and ranking aggregation

Table S5 lists the five criteria rated independently per trial. Raters rank A, B, and C with ties permitted, or select *cannot judge*; playback errors are logged separately and excluded rather than penalized as ties.

Table S5. **The human study evaluates overall preference alongside event fulfillment, timing, continuity, and visual quality.** Exact rater instructions in questionnaire order.

Criterion	Instruction
Overall	Rank the videos by overall preference, considering both the requested scene/actions and visual quality.
Event	Ignoring exact timing, compare whether the correct subjects perform the requested actions. Consider missing actions, incorrect actions, and actions performed by the wrong subject.
Timing	Compare approximate action onsets, offsets, durations, order, and overlap against the requested intervals. Frame-by-frame measurement is not required.
Continuity	Compare the stability of people, objects, and the scene, and the naturalness of transitions. Consider extra or missing people, sudden appearances/disappearances, identity changes, and abrupt object or background changes.
Quality	Ignoring prompt agreement, compare visual fidelity and naturalness of motion, including deformation, flicker, blur, and abnormal movement.

For criterion c and method m, let $\mathcal{V}_c$ be valid rankings and $r_{imc} \in [1, 3]$ the occupied rank (averaged for ties). Table 5 reports the mean normalized score $\overline{q}_{mc} = \frac{1}{|\mathcal{V}_c|} \sum_{i \in \mathcal{V}_c} 50\,(3 - r_{imc}) \in [0, 100]$, giving 100/50/0 points for 1st/2nd/3rd place (75/25/50 for ties). Excluding one technical failure and cannot-judge entries, the snapshot aggregates 27 completed rankings across all ten scenes ($|\mathcal{V}_c| = 24, 23, 27, 27, 27$ for *Event*, *Timing*, *Continuity*, *Quality*, and *Overall*).

F. Additional qualitative comparisons

Six additional examples complement Figure 3, using the same shared timestamps and red/green annotation convention.



Figure S2. **TimeWeave follows sequential desk actions through 58 s, whereas baselines trigger late actions prematurely or omit them.** Single-subject sample 013 at 5.5 s (A, writing with a pencil) and 58 s (B, opening the laptop).



Figure S3. **TimeWeave distinguishes individual and shared object manipulation at their prescribed intervals.** Two-subject sample 041 at 21 s (A, individual chart lift) and 49.5 s (B, shared lift); Self-Forcing and LongLive leave the chart on the table.



Figure S4. **TimeWeave coordinates three participants across sequential shared actions without identity drift.** Three-subject sample 106 at 48 s (A, shared album lift) and 56 s (B, laughing together); baselines miss the lift, add a fourth person, or keep the album open.



Figure S5. **TimeWeave cleanly terminates completed actions instead of letting earlier poses persist.** Two-subject sample 047 (call ends at 20–26 s): TimeWeave lowers the phone at 23 s (A) and shifts to shared notes at 47 s (B), whereas Self-Forcing and LongLive keep the phone at the ear.



Figure S6. **TimeWeave executes fine-grained multi-step assembly on schedule.** Two-subject sample 058 at 30.5 s (A, adding a pinecone) and 56 s (B, holding up the finished bundle); baselines handle ribbon early or leave loose branches at 56 s.



Figure S7. **TimeWeave preserves three distinct participants and synchronizes their joint late-stage action.** Three-subject sample 105 at 28 s (A) and 49.5 s (B, three-way toast); LongLive introduces a fourth person, and baselines miss or partially execute the toast.

G. Additional statements