Infinity-RoPE: Action-Controllable Infinite Video Generation Emerges From Autoregressive Self-Rollout

Supplementary Material

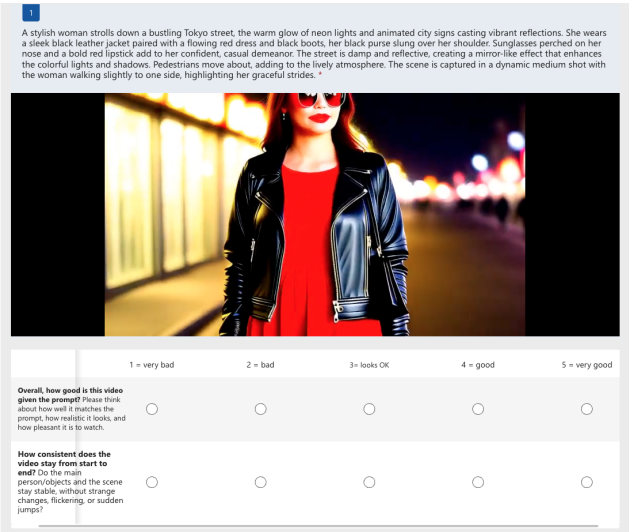


Figure 9. **User Study Interface.** User Study Interface for Long Video Generation

Table of Contents

A. Videos and Website

To facilitate comprehensive evaluation and enhance result accessibility, we provide 100+ video results including motivation examples, qualitative results, ablation studies, qualitative comparisons in infinity-rope.github.io.

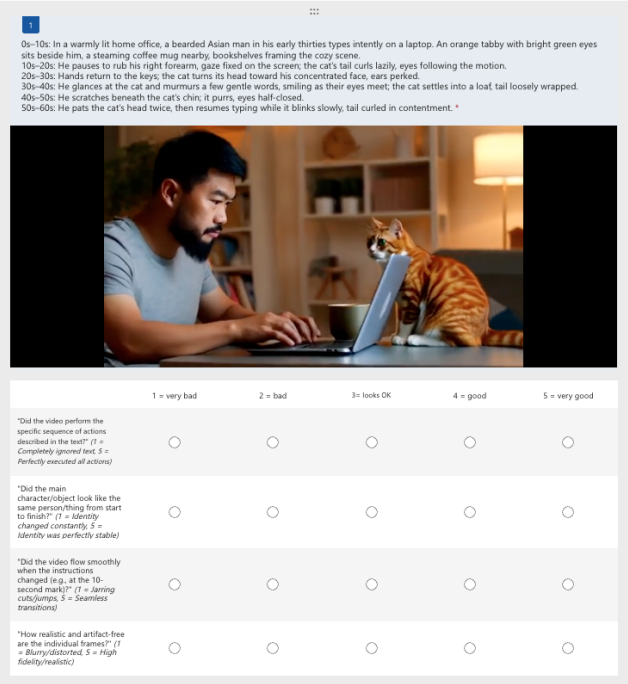


Figure 10. **User Study Interface.** User Study Interface for Action Controlled Long Video Generation

B. User Studies

In Figure 9 and Figure 10, we present the interfaces used in our two user studies. For the long-form video generation user study, participants were shown videos generated from provided prompts and were asked two questions: "Overall, how good is this video given the prompt?" and "How consistent does the video stay from start to end?" These questions were designed to evaluate prompt adherence and temporal consistency. For the action-controlled video generation user study, we compare our method against LongLive [33], SkyReels-V2 [4], and Self-Forcing [15]. In this study, participants were asked four questions: "Did the video perform the specific sequence of actions described in the text?" to measure prompt responsiveness, "Did the main character or object look like the same person or thing from start to finish?" to assess subject consistency, "Did the video flow smoothly when the instructions changed?" to examine motion smoothness at action transition points, and "How realistic and artifact-free are the individual frames?" to evaluate overall video quality.

Figure 11. **Ultra-Long Video Generation Enabled by Block-Relativistic RoPE.** Block-Relativistic RoPE reformulates temporal encoding as a moving frame of reference, allowing the model to preserve relative temporal geometry far beyond the base model's positional horizon. This enables continuous, stable, and fully coherent video generation over extremely long durations without retraining or increased cache size.

C. Additional Clarifications

Clarification on KV Flush description in the abstract. In the abstract of the main paper, we state that KV Flush preserves "two anchor tokens". More precisely, KV Flush preserves two anchor latent frames: the global attention sink and the last generated latent frame(s) before the prompt change. All earlier frames are removed from the cache.

D. More Discussion on Qualitative Results

D.1. Discussion on Long Video Generation Results

In the main paper, we compare ∞-RoPE against NOVA [6], MAGI-1 [30], SkyReels-V2 [4], CausVid [37], Self-Forcing [15], and Rolling-Forcing [22] across both short (5 s) and long (60 s, 120 s, 240 s) generation settings. Our quantitative evaluations consistently show that, in the long-duration regime, ∞-RoPE outperforms prior autoregressive approaches in terms of Subject Consistency, Background Consistency, and Dynamic Degree, while ranking first or second in Motion Smoothness and Temporal Flickering.

To validate that these quantitative trends align with perceptual quality, we provide **supplementary videos** with qualitative comparisons at all four durations. The qualitative results corroborate the numerical findings. As rollout length increases, Rolling-Forcing tends to repeatedly regenerate/spawn similar characters with minimal scene evolution, a limitation stemming from its training paradigm. SkyReels-V2 exhibits large, unstable camera motions that reduce subject consistency in long sequences. Pyramidal Flow frequently resets scene content every 5 seconds, resulting in low subject and background continuity. Meanwhile, both CausVid and Self-Forcing gradually accumulate exposure bias in extended rollouts.

In contrast, ∞-RoPE maintains highly dynamic scenes with stable subject and background appearance across all tested durations, despite relying solely on the pretrained Self-Forcing model, which natively supports only 5-second generation at 16 FPS. These results highlight the robustness and scalability of our method for long-form autoregressive video generation.

D.2. Discussion on Action Control Results

Beyond autoregressive baselines, we also compare ∞-RoPE with LongLive [33], SkyReels-V2 [4], and Self-Forcing [15] for action-controlled video generation. LongLive introduces *KV-Recache*, a cache management mechanism designed to enable prompt-dependent action transitions in autoregressive models. At each transition

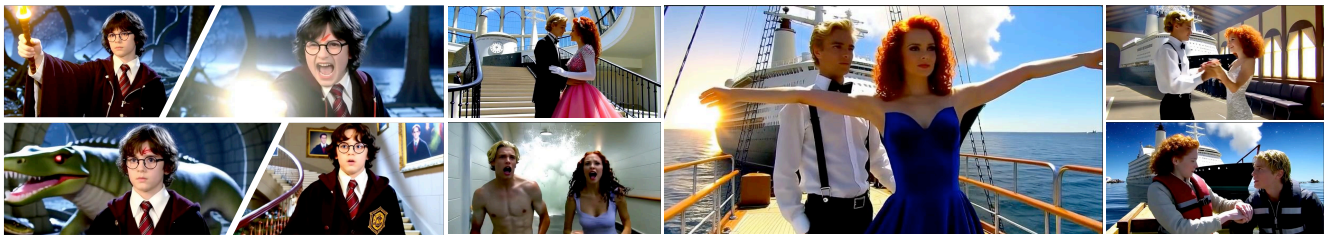


Figure 12. **Dynamic Scene Cut.** RoPE Cut enables controlled cinematic transitions by introducing discontinuities in temporal RoPE coordinates. This allows a single autoregressive rollout to produce diverse environments and background changes while preserving subject identity and temporal coherence.

point, KV-Recache extracts all cached tokens, applies cross-attention with the new prompt, rebuilds a modified cache, and conditions subsequent frames on these recached latent tokens. This procedure aims to overwrite residual semantics from the previous prompt and improve responsiveness to new user instructions. However, the approach presents two main limitations:

1. **Incomplete removal of previous prompt content.** During long rollouts, embeddings from earlier prompts accumulate in the cache and are not fully erased. This reduces prompt responsiveness over time and increases the delay before the new action appears. As demonstrated in the **Action Control Comparison** section of `website.html`, LongLive shows reduced responsiveness and significant identity and background drift, while ∞-RoPE preserves subject identity and background stability and responds to new prompts immediately.
2. **Additional latency proportional to cache size and number of transitions.** KV-Recache requires reconstructing the entire cache at every prompt change. The resulting overhead increases with the cache size and the number of action transitions. By contrast, ∞-RoPE achieves instant prompt responsiveness by simply flushing the stale cache content, which is implemented as a local update to the cache end index without proportional computation.

D.3. Discussion on Dynamic Scene Cut Results

Autoregressive video diffusion models naturally produce temporally smooth sequences, which often results in limited scene dynamism. To introduce cinematic variation without compromising coherence, we propose RoPE Cut, a mechanism that applies controlled discontinuities to the temporal RoPE coordinates. This technique enables intentional scene shifts while preserving overall generative stability.

We present our Dynamic Scene Cut results in the **Supplementary Videos** and in Fig. 12. Using RoPE Cut, we generate trailer-style sequences for several films, including Harry Potter, Titanic, Game of Thrones, The Shawshank Redemption, Barbie, and Interstellar. As demonstrated by these results, RoPE Cut produces dynamic, diverse scenes with varying backgrounds and environments within a single continuous generation stream, while consistently maintaining subject identity and visual fidelity.

E. Additional Ablation Studies

E.1. Ablation on f_0

Block-Relativistic RoPE can operate in both settings where the KV cache size is less than or equal to the pretrained model's natural temporal horizon f_{limit}, as well as in settings where the KV cache size exceeds f_{limit}. In our **supplementary videos**, we demonstrate examples from both regimes. The motivation behind ∞-RoPE comes from a key observation: regardless of the KV cache size used during generation, the model can always exploit the full f_{limit} range without being constrained by the cache length. This perspective provides several advantages. First, when the KV cache exceeds f_{limit}, the model can still fully utilize the available f_{limit} range through the semanticization process without ever stepping beyond the pretrained model's RoPE horizon. Second, when performing RoPE Cut to obtain dynamic scene transitions in the regime where the KV cache is much smaller than f_{limit}, the model can again leverage the entire f_{limit} horizon to generate large, controlled changes in the environment. In our qualitative experiments, we use a KV cache size of 6 and evaluate generation quality for initial frame indices $f_0 \in \{6, 9, 12, 15, 18, 21\}$. Across all settings, the model produces smooth motion and high-quality subject preservation, further demonstrating the relativistic behavior of ∞-RoPE and its ability to maintain consistency even as the temporal reference frame shifts.

E.2. Ablation on Temporal Jump Index Δ

We conduct both qualitative and quantitative experiments on the temporal jump index Δ, which controls the magnitude of temporal discontinuity and enables abrupt environment and action changes within a single generation. Qualitative results are provided in `website.html` under the **Supplementary Material** and illustrated in Figure 14. Our observations show that larger values of Δ produce more pronounced background changes, resulting in stronger cinematic transitions.

Temporal Jump Index Δ	Subject Consistency ↑	Background Consistency ↑	Temporal Smoothness ↑
$\Delta = 6$	90.74	88.98	0.98
$\Delta = 21$	90.04	87.57	0.98
$\Delta = 45$	88.47	84.48	0.96
$\Delta = 90$	87.69	82.27	0.95

Table 4. **Ablation Study on Temporal Jump Index Δ.** We evaluate Subject Consistency, Background Consistency, and Temporal Smoothness for different values of Δ on 40-second video generation. A total of 20 videos are generated, and each video undergoes a scene cut every 10 seconds. Smaller jump values ($\Delta = 6$ and 21) fall within the training horizon and yield smoother transitions, while larger jump values ($\Delta = 45$ and 90) produce more pronounced scene changes accompanied by stronger transition-edge artifacts.

We evaluate $\Delta \in \{6, 21, 45, 90\}$. Notably, $\Delta = 6$ and $\Delta = 21$ lie within the training horizon of the pretrained model, whereas $\Delta = 45$ and $\Delta = 90$ fall outside that range. For in-horizon values, transitions remain smooth with minimal artifacts. For out-of-horizon values, the model produces more dramatic scene transitions at the cost of a visible transition edge. This edge effect arises because the block $\mathbf{B}_{f \to f+\Delta}$ is rotated by a RoPE angle that does not appear in the training data and must therefore be extrapolated. The resulting artifact reflects the inherent extrapolation behavior of the base model's RoPE formulation.

In Table 4, we present the quantitative results for temporal jump indices $\Delta \in \{6, 21, 45, 90\}$. The results show that increasing the temporal jump index leads to lower background consistency due to the more abrupt scene transitions. However, this reduction occurs while subject consistency and temporal smoothness remain high, indicating that the model preserves identity and motion stability even under large scene changes.

F. Action Control Quantitative Comparison

To evaluate the action-controlled video generation capability of ∞-RoPE, we compare it against Self-Forcing, SkyReels-V2, and LongLive both quantitatively in Table 5 and qualitatively in the **Supplementary Videos** under the Action Control Comparison section. The user study demonstrates that ∞-RoPE outperforms existing autoregressive action-controlled generators, while the qualitative results highlight the advantage of **KV Flush** over **KV Recache**. With KV Flush, prompt transitions take effect immediately, whereas KV Recache exhibits a noticeable delay in executing the new action. Moreover, KV Flush requires only a local update of the KV cache end index, while KV Recache performs cross-attention with the new prompt over all cached tokens across all layers, resulting in significantly higher computational cost.

Method	Text Align. ↑	Subject Consist. ↑	Motion Smoothness ↑	Video Quality ↑
Self Forcing	1.88	2.00	1.81	1.64
SkyReels-V2	2.21	2.12	2.14	1.81
LongLive	3.19	3.29	3.10	2.98
Ours	**3.86**	**3.95**	**3.74**	**3.38**

Table 5. **User Study on Action Controlled Video Generation.** ∞-RoPE consistently obtain higher Text Alignment, Subject Consistency, Motion Smoothness and Video Quality scores.

G. Interpretability via Attention Maps

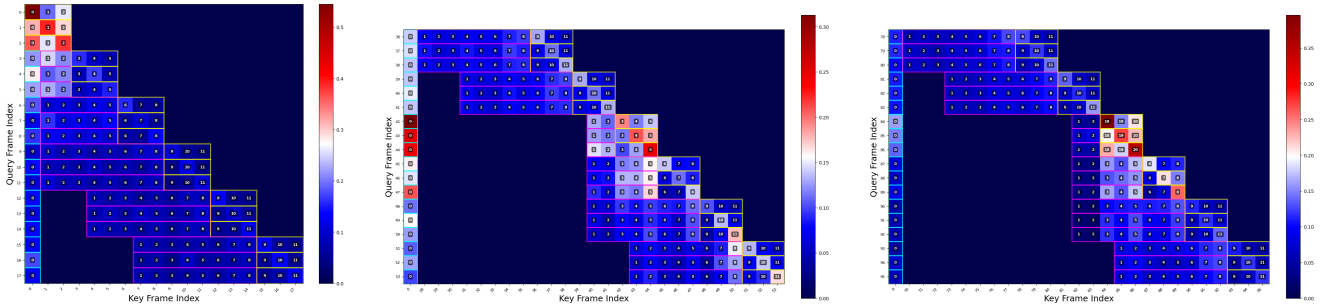
Construction of frame-level attention maps. For each video, we extract self-attention weights from the middle transformer block of the denoiser at a fixed denoising step and aggregate them at the frame level. Let the video consist of T frames, and let each frame t be represented by a set of latent tokens. Denote by $a_{(t,i)\to(s,j)}$ the self-attention weight from query token (t, i) (frame t, spatial index i) to key token (s, j) (frame s, spatial index j), averaged over attention heads. We then construct a $T \times T$ frame–frame attention matrix M by summing over all token pairs between frames:

$$M_{t,s} = \sum_{i \in \text{frame } t} \sum_{j \in \text{frame } s} a_{(t,i)\to(s,j)}.$$

Each cell (t, s) in the attention map therefore corresponds to the total attention mass from all tokens of frame t (query frame) to all tokens of frame s (key frame). Since the underlying self-attention weights are row-normalized by the softmax, each row of M is naturally normalized as well and can be interpreted as a frame-level attention distribution over the video history. A sharp diagonal structure means that each frame mainly attends to itself and its immediate temporal neighbors, while vertical stripes or off-diagonal blocks indicate longer-range dependencies or special tokens (e.g., sink tokens).

Block-Relativistic RoPE for Infinite-length Video Generation. For standard infinite-length generation, the Block-Relativistic RoPE map (Fig. 13a) exhibits two main structures: (i) a sharp diagonal band around the main diagonal, and (ii) a persistent bright column corresponding to the global attention sink token. The diagonal band shows that each query frame primarily attends to a small window of its recent predecessors and itself, which captures local temporal continuity and smooth motion. The bright sink column indicates that the model also consistently attends to a global sink token that provides a stable global context over time.

Crucially, all tokens that lie within the active KV window share a consistent local RoPE coordinate frame: their relative temporal indices stay within the range seen during pretraining (the teacher horizon), even though the absolute



(a) Block-Relativistic RoPE (b) KV Flush (c) RoPE Cut

Figure 13. **Attention behavior of ∞-RoPE across interventions.** Frame-to-frame attention maps from the 13th DiT layer, shown as head-averaged, pixel-summed attention with query frame index q on the y-axis and key frame index k on the x-axis, for (a) a baseline ∞-RoPE rollout, (b) KV Flush at an action change, and (c) RoPE Cut. These visualizations summarize how our method structures long-horizon temporal dependencies; see Sec. 4.4 for a detailed mechanistic interpretation.

Figure 14. **Temporal Jump Index Δ Ablation.**

video length keeps growing. The attention maps do not show any drift of attention towards extremely early frames or degenerate patterns as the sequence becomes long. This supports our claim that Block-Relativistic RoPE re-anchors temporal indices within the teacher horizon instead of letting them drift to unseen absolute positions, effectively bypassing the 1024-index limit and enabling continuous, stable infinite-horizon rollouts.

KV Flush for Action-controllable Long Video Generation. For action-controllable generation, the KV Flush map (Fig. 13b) visualizes the effect of our selective cache renewal strategy when the prompt is changed mid-generation. At the moment of a prompt change, we flush the KV cache and retain only two anchors: the global attention sink token and the last few generated frames (e.g., the last one or two frames before the change). All earlier frames are removed from the cache.

In the attention map, this appears as: (i) a strong vertical column corresponding to the sink token, and (ii) a narrow band of attention centered around the last pre-flush frame(s), while attention to older frames is strongly suppressed and appears almost dark. After the flush, new frames attend primarily to the sink and these recent anchors, rather than to the distant history. This pattern confirms that the model re-orients its temporal context around a very short window of frames while keeping a stable global reference via the sink token.

Mechanistically, this behavior matches the intended design: the model preserves immediate temporal continuity (short-term motion and appearance consistency) through the last cached frames, but rapidly adapts to the new semantic guidance specified by the updated text prompt. The attention maps therefore make explicit how KV Flush balances short-term temporal coherence with fast semantic re-steering.

RoPE Cut for Dynamic Scene Transitions. For dynamic scene transitions, the RoPE Cut map (Fig. 13c) shows what happens when we perform a scene cut. At the cut point, we apply two operations simultaneously: (i) we flush the KV cache, and (ii) we discontinuously advance the RoPE indices for the new frames, assigning them to a new temporal region that does not overlap with the pre-cut segment.

In the frame-level attention maps, this dual action produces two almost disjoint diagonal blocks. The first block corresponds to the pre-cut scene: frames in this block attend strongly to each other but receive very little attention from the post-cut frames. The second block corresponds to the post-cut scene: new frames attend mainly to themselves, their nearby neighbors, and the sink token, while giving only weak attention to frames from the first block. The weak residual attention to the pre-cut frames appears as a faint background, indicating that the old scene is still technically part of the extended context but is functionally de-emphasized.

This pattern shows that RoPE Cut forces the model to reset its effective temporal context at the cut: the post-cut segment behaves like a new scene with its own local tem-

poral structure, rather than a continuation of the old one. At the same time, the sink-mediated pathway still provides a stable identity signal, which helps preserve subject identity across the scene boundary. In other words, the attention maps directly visualize how RoPE Cut implements a hard scene transition in the temporal representation, while still maintaining the global subject and style consistency learned by the model.