Prompt Relay: Inference-Time Temporal Prompt Routing for Multi-Event Video Generation

Anonymous CVPR submission

Paper ID *****

Figure 1. **Prompt Relay** is an inference-time, training-free, plug-and-play method for enabling fine-grained temporal control by routing each textual prompt to its intended time segment, allowing multiple events to occur in the correct order without semantic interference.

Abstract

Video diffusion models have achieved remarkable progress in generating high-quality videos. However, these models struggle to represent the temporal succession of multiple events in real-world videos and lack explicit mechanisms to control when semantic concepts appear, how long they persist, and the order in which multiple events occur. Such control is especially important for movie-grade synthesis, where coherent storytelling depends on precise timing, duration, and transitions between events. When using a single paragraph-style prompt to describe a sequence of complex events, models often exhibit temporal entanglement, where semantics intended for different moments interfere with one another, resulting in poor text-video alignment. To address these limitations, we propose Prompt Relay, an inference-time method to enable fine-grained temporal control in multi-event video generation. We apply a penalty in the cross-attention mechanism to regulate how each query attends to keys intended for different moments in the video. This significantly improves temporal prompt alignment, reduces semantic interference and improves visual quality.

1. Introduction

Recent advances in video diffusion models have enabled the generation of high-quality videos conditioned on textual prompts, achieving impressive visual fidelity and motion coherence [2–4, 21, 26]. Despite this progress, existing models have no mechanism to allow explicit user control over the temporal structure of the video. As a result, modeling movie-grade videos composed of a succession of events, actions, or camera motions, each occurring within a specific segment of the video and in a specific order, remains challenging. Moreover, using a single paragraph-style prompt to describe a succession of complex events often leads to semantic entanglement, where concepts intended for different moments interfere with one another because the model cannot cleanly separate when each event should apply. As a result, multiple incompatible semantics may be represented simultaneously, leading to degraded text–video alignment. Recent works have begun to address temporal controllability in video generation by introducing explicit event-level conditioning [25]. However, these methods typically require training on massive amounts of temporally annotated data. In this paper, we propose Prompt Relay, an elegant

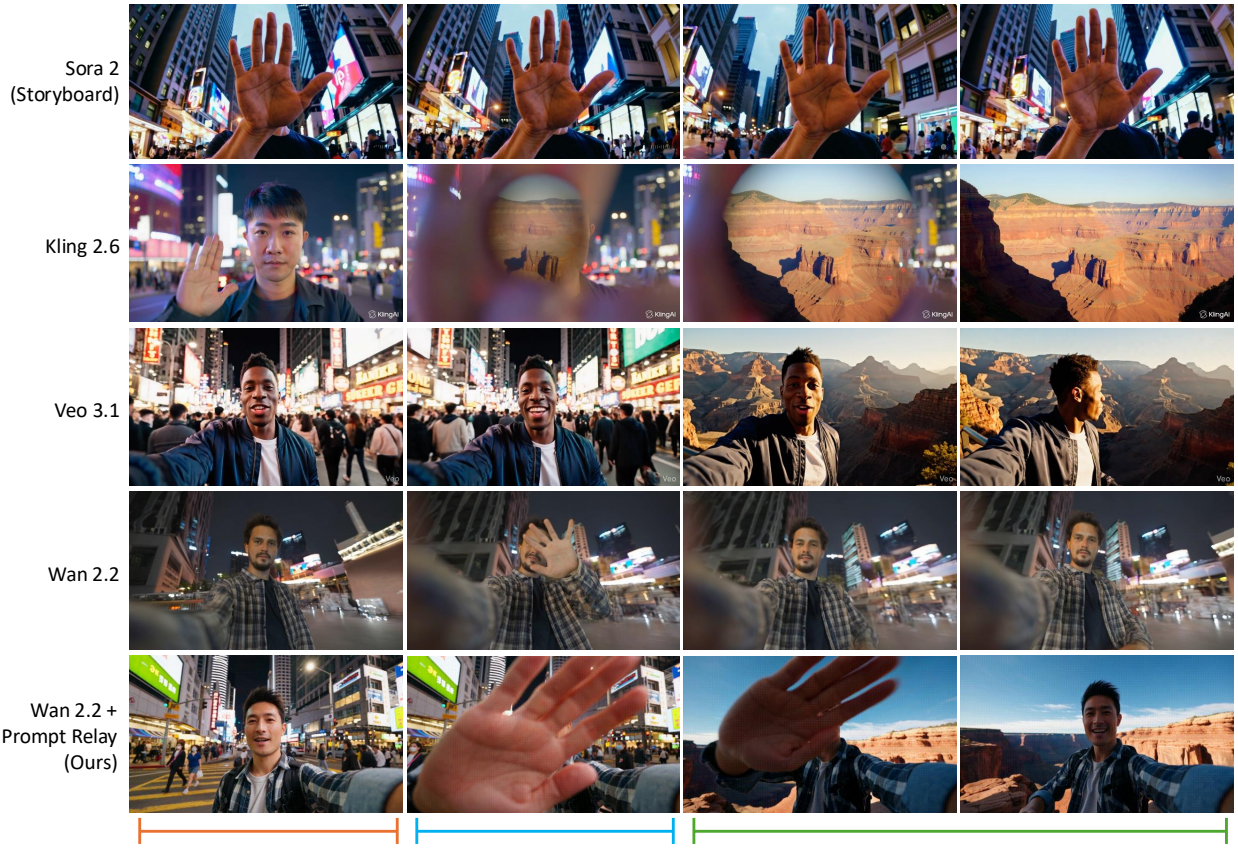

Figure 2. **Qualitative Comparison.** Given a multi-event prompt describing a deliberate scene transition, Prompt Relay preserves correct temporal structure, ensuring that each semantic instruction influences only its intended segment while maintaining global visual coherence.

attention-level routing mechanism for fine-grained temporal control and multi-event video generation. Prompt Relay operates entirely at inference time and is plug-and-play compatible with existing video diffusion backbones.

- We introduce a test-time, plug-and-play method that enforces each semantic instruction to influence only its intended temporal segment.
- We propose an adaptive temporal partitioning scheme that supports variable-length segments, enabling fine-grained control over event duration.
- We demonstrate that Prompt Relay substantially improves temporal prompt alignment, reduces semantic interference and enhances visual quality in multi-event video generation.

2. Related Works

Controllable Video Generation. Video generation has seen rapid progress in recent years, with applications spanning motion control [6, 9, 22–24], viewpoint control [7, 14, 20], identity control [16, 17, 28] and editing [8, 18]. However, most models remain limited in the ability to generate coherent multi-event videos. Because the attention mechanism allows every pixel to attend to every prompt token, models struggle to associate semantic concepts with their intended temporal intervals, leading to temporal misalignment and semantic entanglement. This challenge motivates us to provide explicit temporal control at inference time.

Attention-Based Control in Diffusion Models. Attention manipulation has emerged as a key mechanism for control-

Figure 3. **Temporal Cross-Attention Routing.** Each textual prompt is associated with a specific temporal segment of the video. The attention penalty varies smoothly across time, allowing video tokens to attend strongly to their corresponding prompt within the assigned interval while suppressing attention to temporally irrelevant prompts. This enables multiple events (e.g., pouring cereal followed by pouring milk) to occur in the correct order without semantic interference.

lable diffusion generation. Prior work has explored attention for spatial [11–13, 15, 27], identity [10, 29] and motion control [18, 19, 22]. In contrast, attention-based temporal control remains largely underexplored.

Multi-Event Video Generation. A notable approach to temporal modeling for multi-event video generation is MinT [25], which introduces a trainable temporal cross-attention module that binds event descriptions to predefined time intervals, but requires additional training, architectural modifications, and temporally annotated data.

3. Prompt Relay

Given a sequence of temporally-constrained text prompts $\{(p_s, t_s^{\text{start}}, t_s^{\text{end}})\}_{s=1}^{N}$, our goal is to generate a video such that each arbitrary prompt p_s is realized within its designated temporal interval $[t_s^{\text{start}}, t_s^{\text{end}}]$. The generated video should preserve global coherence while ensuring that each prompt influences only its assigned temporal region.

3.1. Preliminaries

Cross-attention is a mechanism that enables a diffusion model to incorporate external conditioning information, such as text prompts, into the generation process. Given a latent representation at diffusion step t, denoted as $\phi(z_t)$, and a set of conditioning embeddings $\psi(P)$ derived from an input prompt P, cross-attention computes interactions between the two through learned projections.

$$\text{Attn}(\phi(z_t), \psi(P)) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \qquad (1)$$

where $Q = \ell_Q\phi(z_t)$ are query vectors derived from latent features, $K = \ell_K\psi(P)$ and $V = \ell_V\psi(P)$ are key and value vectors projected from the conditioning embeddings, and d denotes the projection dimensionality. Each attention weight reflects how strongly a latent query attends to a particular conditioning token. Through this operation, semantic information from the conditioning input is selectively injected into the latent representation, allowing different queries to respond to different aspects of the prompt. However, because attention is computed globally over all conditioning tokens, multiple semantic concepts may compete for influence over the same latent queries. When these concepts correspond to different temporal regions, unrestricted attention can lead to interference between instructions.

3.2. Temporal Cross-Attention Routing

In order to enforce the association between each prompt p_s and its assigned temporal interval $[t_s^{\text{start}}, t_s^{\text{end}}]$, we introduce a penalty term $C(Q, K)$ into the cross-attention logits:

$$\text{Attn}(\phi(z_t), \psi(P)) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} - C(Q, K)\right)V. \qquad (2)$$

The role of $C(Q, K)$ is to suppress the attention between key and query tokens whenever they do not belong to the same interval $[t_s^{\text{start}}, t_s^{\text{end}}]$. This allows each prompt to guide generation only within its intended segment, without leaking semantic concepts into other parts of the video. For any arbitrary query token indexed by i and any key token j belonging to p_s, the penalty is defined as:

$$C(i, j) = \frac{\text{ReLU}(|f(i) - m_s| - w)^2}{2\sigma^2}, m_s = \frac{t_s^{\text{start}} + t_s^{\text{end}}}{2}. \qquad (3)$$

Here, $f(i)$ denotes the latent frame index associated with query token i, and m_s denotes the midpoint of the corresponding temporal segment. The parameter $w (w < m_s)$ defines a local window around the segment midpoint within which no penalty is applied, while σ controls the rate at which attention decays outside this window. Query tokens

Metric	Sora 2 (Storyboard)	Kling 2.6	Veo 3.1	Wan 2.2	Wan 2.2 + Prompt Relay
Temporal Alignment (↓)	4.67	1.30	3.93	4.00	**1.10**
Transition Naturalness (↓)	4.60	4.43	1.30	3.50	**1.17**
Visual Quality (↓)	3.67	2.50	**2.0**	4.00	2.83

Table 1. Human preference scores for multi-event video generation. (lower values indicate better rankings)

within the window incur zero penalty and can attend freely to their associated prompt tokens. Beyond this region, attention is smoothly attenuated as a function of the temporal distance between the query and the segment midpoint.

3.3. Boundary-Aware Decay

To suppress semantic interference across temporal segments, attention between queries near segment boundaries and prompt tokens from neighboring segments should be negligible. We therefore choose the decay parameter σ so that the attention prior sufficiently decreases near segment endpoints. Since our penalty subtracts $C(i,j)$ from the logits, it applies a multiplicative factor $\exp\big(-C(i,j)\big)$ to the unnormalized attention scores before softmax. This prior is 1 inside the "free-attention" window and decays toward the segment boundaries. Let the endpoint distance from the segment midpoint be $L = |f(i) - m_s|$. We choose σ such that the prior reaches a small user-defined value $\epsilon \in (0,1)$ at the endpoints:

$$\exp\left(-\frac{(L-w)^2}{2\sigma^2}\right) = \epsilon \;\Rightarrow\; \sigma = \frac{L-w}{\sqrt{2\ln(1/\epsilon)}}. \quad (4)$$

This formulation ensures smooth transitions between neighboring prompts while preventing destructive interference across segments. As a result, each textual instruction primarily influences its intended temporal region, allowing the model to focus on one semantic concept at a time while maintaining global temporal coherence.

4. Experiments

4.1. Experimental Setup

We construct diverse multi-event test scenarios, covering a wide range of settings including explicit scene transitions, multi-character interactions, and complex camera trajectories, randomly generated with ChatGPT [1]. These scenarios each contain 3-6 temporal events. All experiments are conducted using the state-of-the-art pretrained video generation model Wan2.2-T2V-A14B. To demonstrate the limitations of existing video generators in handling multi-event prompts, we test several state-of-the-art models, including Sora 2 [3], Veo 3.1[4], Wan 2.2[5], and Kling 2.6[2]. For Prompt Relay, we set $\epsilon = 10^{-3}$ across all experiments. In addition to selectively routing local prompts to their assigned temporal segments, we include a global prompt that conditions the entire video and provides persistent context.

4.2. Evaluation Metrics

Existing quantitative metrics test visual fidelity or global text-video alignment, but fail to capture temporal semantics or transition quality, properties that are inherently perceptual. Hence, we conduct a human preference study to evaluate multi-event video generation along three dimensions:

- **Temporal Prompt Alignment:** Whether each semantic instruction appears at its intended temporal interval.
- **Transition Naturalness:** The smoothness and perceptual continuity between consecutive events, including the absence of abrupt or unnatural scene changes.
- **Visual Quality (Supplementary):** Overall perceptual fidelity, clarity and realism

We use 20 randomly generated multi-event test scenarios. For each dimension, participants (30) are shown paired videos generated by different methods and asked to rank them according to each criterion.

4.3. Qualitative Results

As shown in Table. 1, Prompt Relay consistently outperforms baseline approaches in temporal alignment and transition naturalness. Notably Wan 2.2 with Prompt Relay consistently exhibits stronger visual quality compared to the baseline Wan 2.2. This is likely because Prompt Relay's attention routing mechanism suppresses attention between queries in a particular temporal segment and prompts belonging to other segments. By reducing unnecessary competition in the cross-attention space, the model can allocate attention more effectively to the active semantic concepts, resulting in clearer visual structure, improved temporal alignment, and more stable generation.

5. Conclusion

We present Prompt Relay, an inference-time method for multi-event video generation with fine-grained temporal control. We also show that our method also improves visual quality by reducing competition in the cross-attention space. We view our work as a pivotal step towards movie-grade, controllable video synthesis tools.

References

[1] Chatgpt 5.2. Accessed January 15, 2026 [Online], 2025. 4

[2] Kling 2.6. Accessed January 15, 2026 [Online], 2025. 1, 4

[3] Sora 2. Accessed January 15, 2026 [Online] https://sora.chatgpt.com/explore, 2025. 4

[4] Veo 3.1. Accessed January 15, 2026 [Online], 2025. 1, 4

[5] Wan 2.2. Accessed January 15, 2026 [Online], 2025. 4

[6] Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025. 2

[7] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 2

[8] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025. 2

[9] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[10] Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025. 3

[11] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 3

[12] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 2023.

[13] Gordon Chen, Ziqi Huang, Cheston Tan, and Ziwei Liu. Stencil: Subject-driven generation with context guidance. In *2025 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2025. 3

[14] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[16] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025. 2

[17] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 2

[18] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[19] Tuna Han Salih Meral, Hidir Yesiltepe, Connor Dunlop, and Pinar Yanardag. Motionflow: Attention-driven motion transfer in video diffusion models. *arXiv preprint arXiv:2412.05275*, 2024. 3

[20] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[21] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

[22] Luozhou Wang, Ziyang Mai, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Ying-Cong Chen. Motion inversion for video customization. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025. 2, 3

[23] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 2023.

[24] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2

[25] 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 Computer Vision and Pattern Recognition Conference*, 2025. 1, 3

[26] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 3

[28] Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-id: Towards universal identity-preserving video synthesis. *arXiv preprint arXiv:2503.14151*, 2025. 2

[29] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 2024. 3