DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation
(Supplemental Material)

Minghong Cai[1†] Xiaodong Cun[2] Xiaoyu Li[3✉] Wenze Liu[1] Zhaoyang Zhang[3]
Yong Zhang[4] Ying Shan[3] Xiangyu Yue[1,5✉]

[1] MMLab, CUHK [2] GVC Lab, Great Bay University
[3] ARC Lab, Tencent PCG [4] Tencent AI Lab [5] SHIAE, CUHK

Overview

This supplementary material presents comprehensive experimental details, qualitative analyses, and technical implementations of our work. We provide extensive evaluations across multiple aspects, including baseline comparisons, diverse application scenarios, and ablation studies. Note that our project page shows many cases of our results, comparison and diverse application scenarios. The content is organized into six main sections:

- Section A details our experimental framework, including baseline implementations and model implementation details.
- Section B details our evaluation, including evaluation metrics, human evaluation protocols, and TSNE visualization discussion.
- Section C showcases comprehensive qualitative results across diverse domains, featuring detailed comparisons with state-of-the-art models and demonstrating the versatility of our approach.
- Section D explores various applications, including single-prompt video generation and advanced editing capabilities such as attention reweighting and word swap techniques.
- Section E presents the usage of *prompt generator*, including full descriptions used to generate individual prompts.
- Section F presents comprehensive ablation studies, including both quantitative evaluations and qualitative analyses of the masking mechanism.
- Section G discuss the inference time of alternative methods.

A. Implementation Details

Details. We implement DiTCtrl based on CogVideoX-2B [8], which is a state-of-the-art open-source text-to-video diffusion model based on MM-DiT. The hyperparameters and implementation details are shown in Tab. 1.

Table 1. Hyperparameters of DiTCtrl.

Hyperparameters	
base model	CogVideoX-2B
sampler	VPSDEDPMPP2MSampler
sample step	50
guidance scale	6
resolution	480×720
video frames	49
latent num frames	13
overlap size	6
kv-sharing steps	[2,25]
kv-sharing layers	[25,30]
threshold	0.3
λ of CSCV	10

Baselines. In experiments of our main paper, we comprehensively compare our method with previous state-of-the-art methods, including commercial and open-source techniques. We offer more details of the baselines that we use here:

- **Kling [1]**: Kling is leading closed-source commercial solutions developed by Kuaishou Technology. It can generate videos of 6s lengths, but it can only input single-prompt, so we input a single prompt describing long-term temporal changes. We use the Kling1.5 model for our visualization comparison.
- **Gen-L-Video [7]**: Gen-L-Video processes long videos as short video clips with temporal overlapping during the denoising process. We use the VideoCrafter2 [2] as the base model.
- **FreeNoise [5]**: FreeNoise reschedules the initial noise sequence and conducts temporal attention fusion based on the sliding window for temporal consistency. We use the

VideoCrafter2 [2] as the base model.

- **Video-Infinity [6]**: Video-Infinity scales up long video generation via distributed inference. We use the VideoCrafter2 [2] as the base model.
- **FreeNoise+DiT**: This is an enhanced baseline by directly incorporating FreeNoise's noise rescheduling strategy into the CogVideoX [8] framework.

For a fair comparison, all baseline methods should be aligned to use the same ratio stride. Since CogVideoX-2B has 13 latent frames, we used overlap frame 6 in our paper which is approximately 1/2 stride of the total frames ($6/13 \approx 1/2$). Other baseline methods also use this setting of same stride ratio.

Mask-guided Implementation Details. We show how mask extracted from MM-DiT attention map is utilized for mask-guided KV-sharing strategy in Fig. 1, to generate consistent video over time for multi-prompt video generation task.

Specifically, Fig. 1 illustrates our approach to generating temporally consistent videos in multi-prompt video generation tasks. When computing attention for the P_i branch latent, we utilize attention maps from both P_{i-1} and P_i branches. Specifically, we extract content from the Text-video and Video-text attention regions of their attention maps. By focusing on specified tokens (e.g., "a running horse"), we obtain and average the corresponding regional values to generate semantic mask maps. These maps are then binarized through thresholding to create foreground-background segmentation masks M_{i-1} and M_i.

Then, we leverage M_{i-1} to guide the computation of KV-sharing attention maps (calculating attention between Q_i and K_{i-1}, V_{i-1}), resulting in foreground-focused attention outputs F_{fore} and F_{back}. The final fusion is achieved through M_i as follows:

$$F_{fusion} = F_{fore} * M_i + F_{back} * (1 - M_i) \qquad (1)$$

This mask-guided approach ensures semantic consistency while maintaining smooth transitions between different prompts.

B. Evaluation details

MPVBench. We introduces a new benchmark MPVBench, which is specified designed for multi-prompt video generation task. MPVBench contains a diverse prompt dataset and a new metric customized for multi-prompt generation. Specifically, leveraging GPT-4, we produce 130 long-form prompts of 10 different transition modes (background transition, subject transition, camera transition, style transition, lighting transition, location transition, speed transition, emotion transition, clothing transition, action transition). The instruction of prompt generator is provided in Fig. 13.

Automatic evaluation. For automatic evaluation, we generate videos using 130 prompts from our MPVBench, with three random seeds set. Then, we evaluate the generated video by three metrics: CSCV (Clip Similarity Coefficient of Variation), Motion Smoothness, Text-Image Similarity.

Human evaluation. In our user study, we combined our generated videos with those produced by four other baseline methods. We asked a total of 28 participants to evaluate the videos across four dimensions: overall preference, motion pattern, temporal consistency, and text alignment. Specifically, we asked all participants to rank the results of these methods for each of the following questions, and assigned a score from 1 (lowest quality) to 5 (highest quality) for these five methods:

- **Overall Preference**: "*Please rank the overall video preference.*" This metric evaluates participants' comprehensive assessment of the generated videos.
- **Motion Pattern**: "*How natural and realistic are the motion in the video?*" This evaluates whether the motion of objects in the generated video appears physically plausible and natural, such as whether vehicles drive realistically, animals move naturally, or human actions appear authentic.
- **Temporal Consistency**: "*How smoothly does the video content transition across different frames?*" This metric evaluates the temporal coherence of the generated video, focusing on whether the transitions between consecutive frames are natural and continuous, without abrupt changes or visual artifacts. It measures the video's ability to maintain visual continuity throughout its duration.
- **Text Alignment**: "*To what extent does the video content match the given text descriptions?*" This assesses the semantic fidelity between the generated visual content and the input text prompts, examining whether the video accurately captures and visualizes the key elements and actions described in the prompts. It measures how well the visual narrative aligns with the intended textual description.

t-SNE Visualization discussion. In the justification for the proposed CSCV metric, which evaluates the transition smoothness, We found that t-SNE visualizations of real videos from existing datasets have similar continuous trajectories due to semantic continuity. Therefore, we just present one representative case, the t-SNE of video embeddings for real videos. The selected real video in Fig.7 of main paper is the classic car video from DAVIS [4]. The car video frames are shown in Fig.2.

We also show more t-SNE visualization of our comparison cases in Fig. 8 and Fig. 9. Even when processing multi-prompt videos, our method generates continuous trajectories that are comparable to those in real videos. This showcases the exceptional transition handling capabilities and overall stability of the videos produced by DiTCtrl.

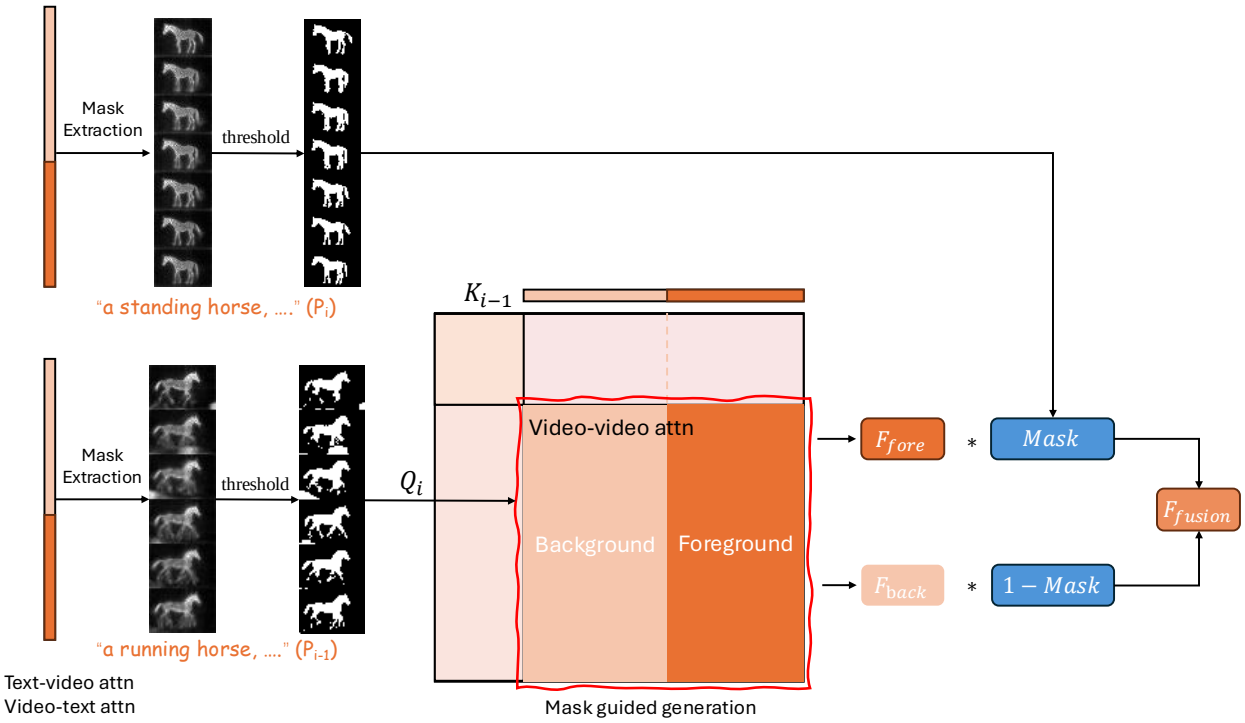


Figure 1. Mask-guided KV-sharing details.

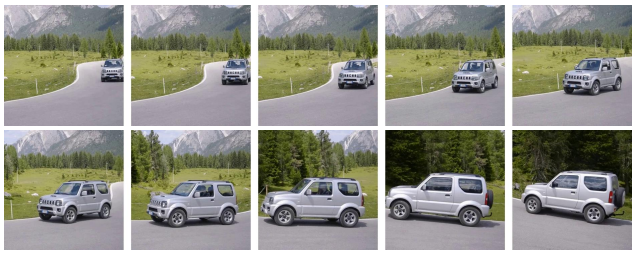


Figure 2. **real video example from DAVIS [4].**

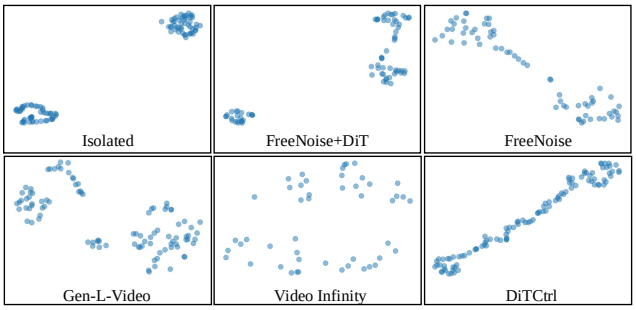


Figure 3. **t-SNE Visualization** of Fig. 8

C. More Qualitative Results

More results are provided in Fig. 6 and Fig. 7. Our method DiTCtrl can generate multi-prompt videos with good temporal consistency and strong prompt-following capabilities, demonstrating cinematographic-style transitions in depicting the boy's riding sequence. We also give more qualitative comparisons with state-of-the-art multi-prompt video generation methods [5–7], our reproduced FreeNoise+DiT, and leading commercial solutions Kling [1]. We show the *motion transition* case, and *background transition* case in Fig. 8 and Fig. 9. Our comparative analysis reveals distinct characteristics and limitations of existing approaches. Gen-L-Video [7] suffers from severe temporal jittering, compromising overall video quality. Video-Infinity [6] and FreeNoise [5] both demonstrate successful scene-level semantic changes but lack physically plausible motion. For instance, in Fig. 8, dark knight appear to be in motion while remaining spatially fixed, which is a limitation inherent to their UNet-based abilities. In contrast, FreeNoise+DiT leverages the DiT architecture's abilities to achieve more realistic object motion but struggles with semantic transitions, resulting in noticeable discontinuities between segments. Our proposed DiTCtrl method preserves the inherent capabilities of the pre-trained DiT model while addressing these limitations, en-

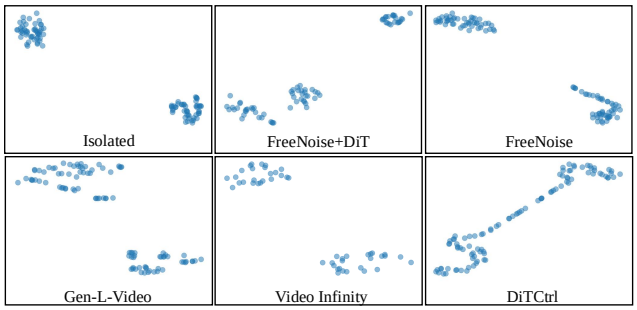


Figure 4. **t-SNE Visualization** of Fig. 9

abling smooth semantic transitions and maintaining motion coherence throughout the video sequence. More comparison of visualization case and our results are shown in our project page.

D. Applications

Based on our exhaustive analysis and exploration of attention control in MM-DiT architecture, our method could be applied to other tasks like single prompt longer video generation and video editing and achieves promising results.

D.1. Single-prompt Longer Video Generation

Although our primary objective is to address multi-prompt video generation, we discover that our method demonstrates remarkable effectiveness in single-prompt longer video generation as well. Our method can naturally work on single-prompt longer video generation. As illustrated in Fig. 10, our approach successfully generates longer videos, while maintaining consistent motion patterns and environmental coherence.

D.2. Video Editing

In this work, we conduct an in-depth analysis of MM-DiT's attention maps, which can be categorized into four components: Text-to-Video and Video-to-Text Attention, Text-to-Text and Video-to-Video Attention. Through our analysis of Text-to-Video and Video-to-Text Attention, we observe that semantic maps can be obtained by specifying token indices, suggesting potential for semantic control. We have emphasized the use of extracted foreground-background segmentation semantic maps to guide video generation, effectively preventing semantic confusion between foreground and background elements. In this section, we demonstrate video editing capabilities through two approaches: *Reweight* and *Word Swap*.

Attention Re-weighting. As illustrated in Fig. 11, we can achieve semantic enhancement or attenuation by increasing or decreasing the values in rows or columns corresponding to token j in the Text-to-Video and Video-to-Text Attention maps. In Fig. 11 (a), we demonstrate semantic attenuation by reducing Text-Video Attention values in the row and Video-Text Attention values in the column corresponding to "pink". In Fig. 11 (b), we achieve semantic enhancement by increasing Text-Video Attention values in the row and Video-Text Attention values in the column corresponding to "snowy". These results validate the semantic control capabilities of Text-Video and Video-Text Attention in MM-DiT.

Word Swap. Building upon the concept introduced in Prompt-to-prompt [3], this approach allows users to swap tokens in the original prompt with alternatives (e.g., changing P ="a large bear" to "a large lion"). The primary challenge lies in maintaining the original composition while accurately reflecting the content of the modified prompt. Our DiTCtrl method incorporates KV-sharing, similar to the word swap mechanism in [3], where we share key-value pairs from the previous prompt to compute the corresponding video for the subsequent prompt across selected layers and steps. Specifically, DiTCtrl (without latent-blending strategy) enables token-replacement video editing while ensuring consistency in other content elements, as demonstrated in Fig. 12. This implementation validates the feasibility of prompt-to-prompt-style video editing within the MM-DiT architecture.

E. Prompt Generator

In this section, we provide additional information of the prompt generator that is described in our main paper. We use GPT4 for longer multi-prompt generation, our prompts are shown in Fig. 13. This figure shows the generation process of "background transition", and we generate 10 different transition modes (background transition, subject transition, camera transition, style transition, lighting transition, location transition, speed transition, emotion transition, clothing transition, action transition).

F. Ablation Study

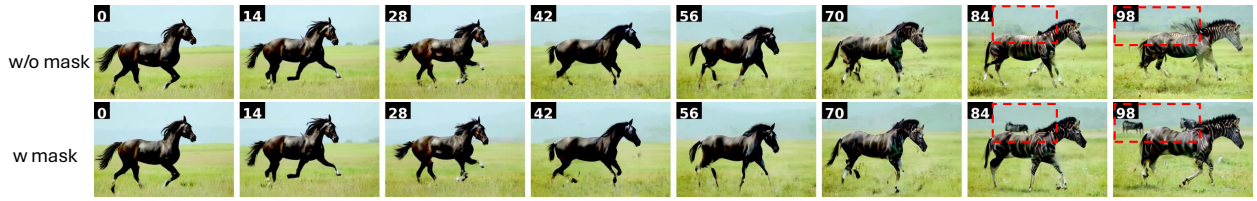
F.1. Quantitative Results of Components

As shown in Tab. 2, our latent blending strategy (second row) demonstrates superior video consistency compared to isolated clips (first row), as evidenced by higher CSCV scores - our proposed metric for evaluating multi-prompt transition smoothness. Furthermore, our KV-Sharing mechanism further improves the CSCV value, achieving enhanced stability. The mask-guided approach(fourth row) and its unmasked counterpart(third row) report comparable scores, suggesting that the contribution of masking foreground object to overall frame transition smoothness is modest. However, our qualitative analysis in Section F.2 reveals that the mask-guided method yields superior visual results.

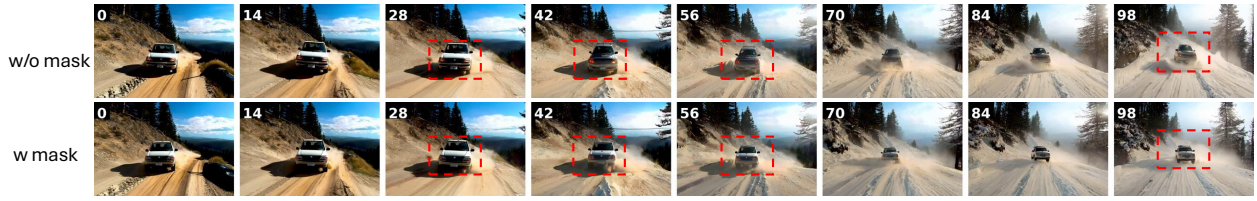
Additionally, in our evaluation of motion smoothness, our full method (DiTCtrl) achieves optimal performance. Regarding the Text-Image similarity metric, we observe a slight expected decrease with our approach. This is attributable to our methodology where the latent representation of the latter video segments incorporates keys and values from preceding segments to maintain consistency. This inherently introduces semantic information from previous segments, marginally reducing the current segment's alignment with its corresponding text prompt. However, this trade-off is justified as our method achieves stable transitions and effectively conveys both semantic elements, resulting in higher user study scores as shown in Tab. 2.

F.2. Mask-guided Generation Analysis

We present comparative results in Fig. 5 to demonstrate the effectiveness of our mask-guided KV-sharing strategy. In



(a) ID transition: horse to zebra



(b) Background transition: dirt-road to snow-road

Figure 5. Ablation study of mask-guided KV-sharing results. First row shows our model without mask-guided KV-sharing, while the second row demonstrates our full model with mask-guided KV-sharing. The prompt for (a) transitions from "A powerful horse gallops across a field..." to "A striking zebra leads its herd across the field...". The prompt for (b) evolves from "A white SUV drives a dirt road..." to "A white SUV powers through snow..."

Method	CSCV	Motion smoothness	Text-Image similarity
Isolated	72.37%	97.78%	**32.05%**
DiTCtrl(w/o kv-sharing)	81.79%	97.35%	31.37%
DiTCtrl(w/o mask-guided)	**84.92%**	97.76%	30.66%
DiTCtrl(full)	84.90%	**97.80%**	30.68%

Table 2. Comparison of metrics for ablation.

Fig. 5 (a), while the first prompt describes a single horse, the second prompt emphasizes a zebra leading its herd. Without mask-guided KV-sharing (first row), we observe that the model fails to properly generate the zebra herd and exhibits background inconsistencies. In contrast, our full model with mask-guided KV-sharing (second row) successfully maintains scene coherence while incorporating the herd elements.

Similarly, in Fig. 5 (b), the transition sequence in the first row (without mask-guided KV-sharing) shows notable deformations in the vehicle's appearance, including undesired color variations. The second row, implementing our mask-guided approach, better preserves the vehicle's original appearance, color, and shape throughout the transition. These results validate both the effectiveness of our mask-guided approach and the feasibility of leveraging semantic maps extracted from MM-DiT's Text-Video and Video-Text Attention for application in Video-Video Attention.

G. Inference Time

We present a comparison of the inference times on a single A100 GPU, with the variation based on the number of prompts (N). For a fair assessment, when 2 prompts are input, each method is tasked to generate approximately 100 frames. When the number of prompts increases to 3, the generation target is set at approximately 150 frames. As depicted in Table 3, our method (without mask) demonstrates competitive efficiency in terms of elapsed time, and also achieves satisfactory video transition effects. When the mask-guided approach is further employed, it yields even more superior visual outcomes. Despite the sixfold increase in runtime, the method remains Pareto optimal.

References

[1] Kling. https://kling.kuaishou.com/en, 2024. 1, 3

[2] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 1, 2

[3] 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. 4

[4] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *ArXiv*, abs/1704.00675, 2017. 2, 3

[5] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 1, 3

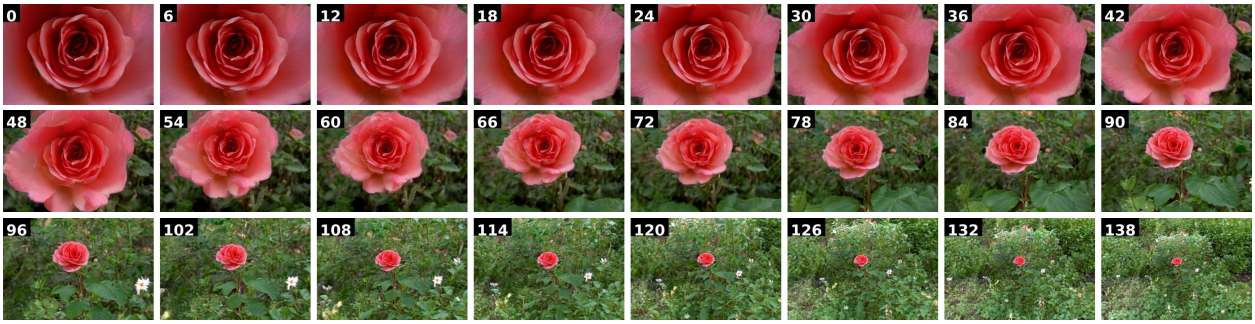
[6] Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. Video-infinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260*, 2024. 2, 3

[7] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 1, 3

[8] 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

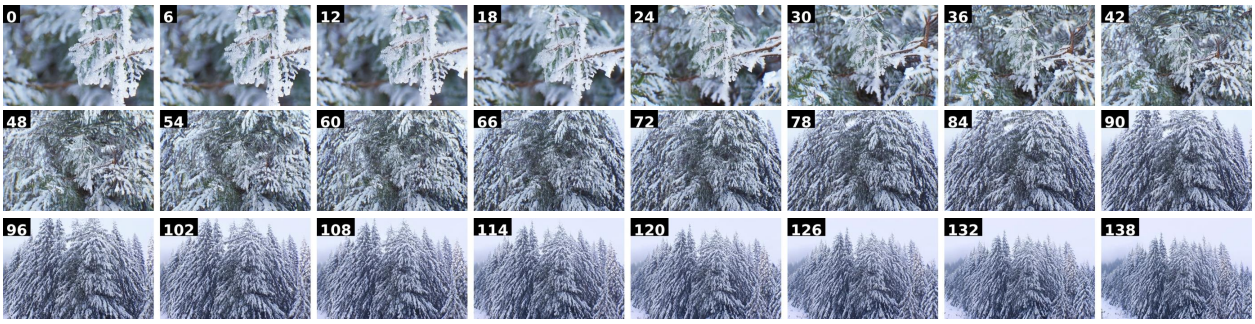
	Gen-L-Video	FreeNoise	FreeNoise+DiT	Video-Infinity	Ours(w/o mask)	Ours(w/ mask)
N=2	9.1min	6.1min	5.3min	1.2min (2 gpu)	5.3min	∼39min
N=3	13.6min	9.2min	10.6min	1.2min (3 gpu)	10.6min	∼78min

Table 3. Inference time comparison with the number of prompts N



(a) "**Close-up shot → medium shot → wide shot** of a blooming rose, cinematic"

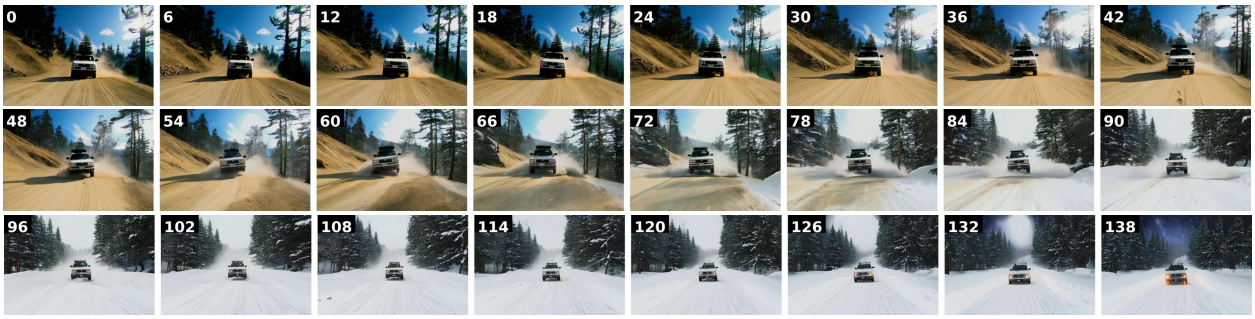
(b) "Boy cycling **through corridor → to doors → into garden**, cinematic, 4K"



(c) "Frosty pine: **close-up shot → medium shot → forest vista**, cinematic"

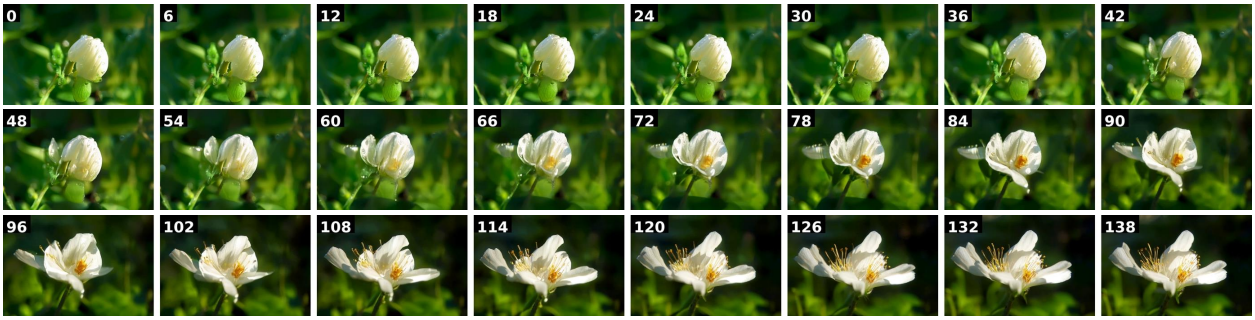
Figure 6. More multi-prompt results

diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2



(a) "A white SUV driving on **dirt road → snowy path → starry night**"

(b) "Dark knight **rests in grassland → gallops across snowy fields → desert**"



(c) "A flower **bud emerges → unfolds gracefully → stands in full bloom**"

Figure 7. More multi-prompt results

Figure 8. Motion and background transition.

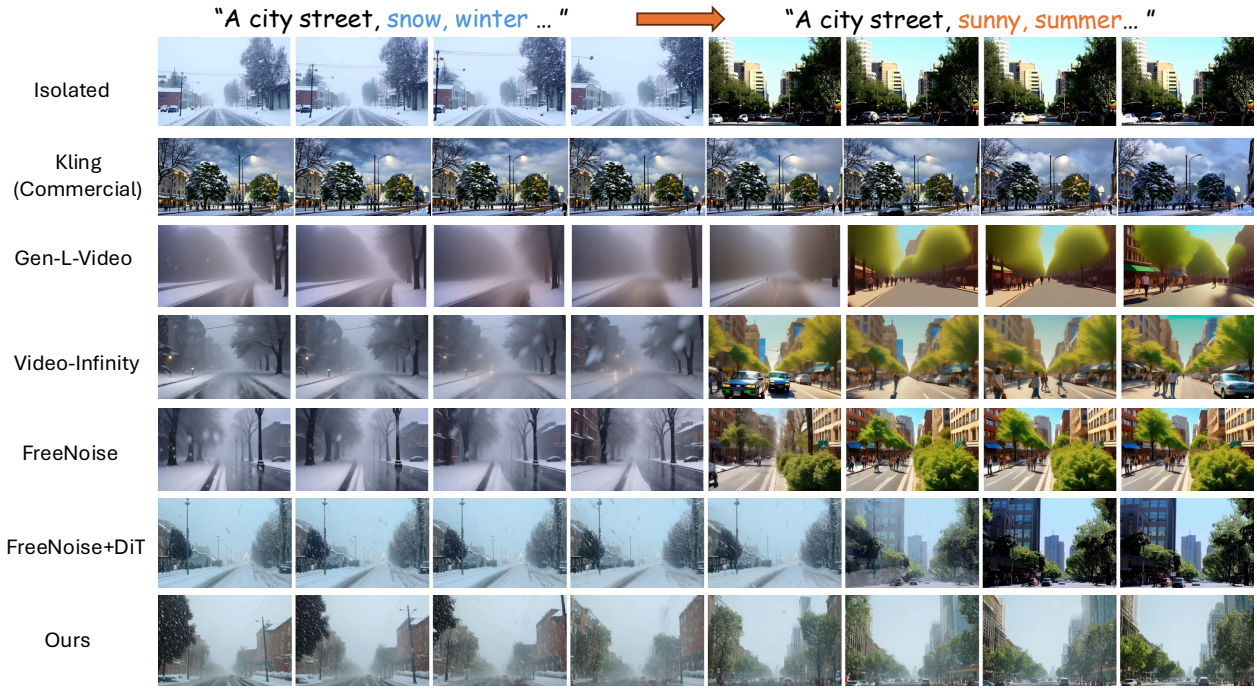
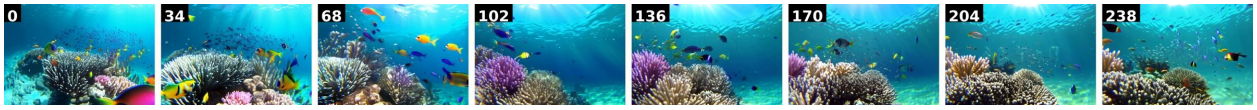


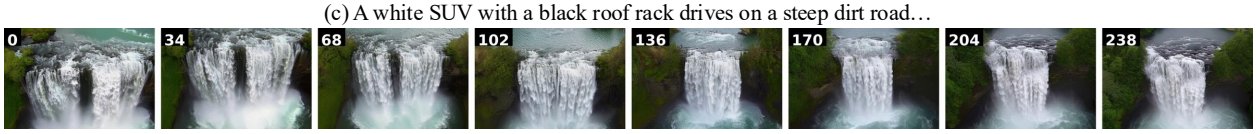
Figure 9. Background transition.



(a) A vibrant school of tropical fish weaves through an intricate coral reef system…

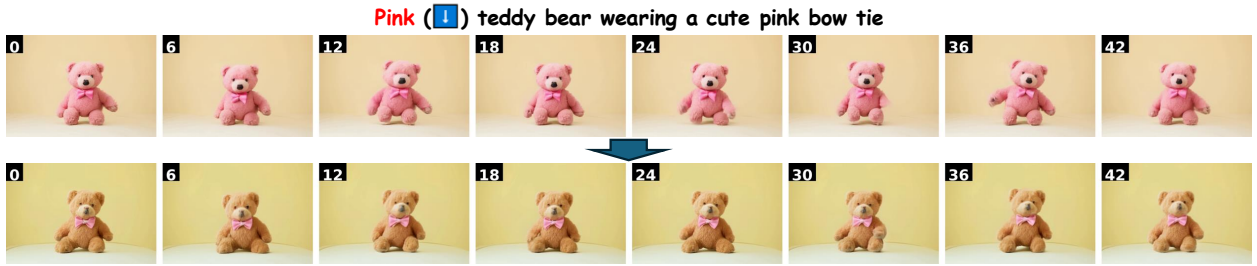
(b) A magnificent hot air balloon ascends into the dawn sky…

(c) A white SUV with a black roof rack drives on a steep dirt road…

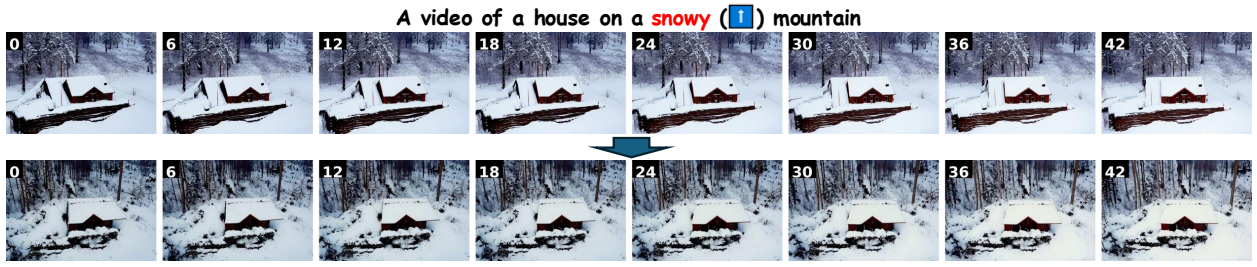


(d) The video captures the majestic beauty of a waterfall cascading down a cliff into a serene lake…

Figure 10. Visualization of single prompt longer video generation.

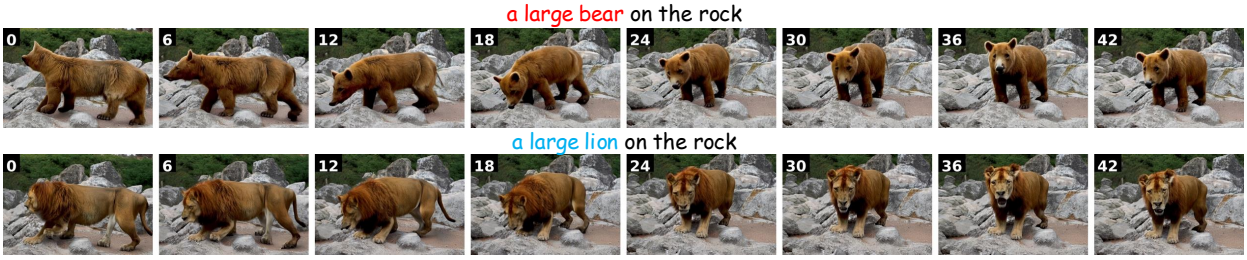


(a) Weaken token - *pink*



(b) Strengthen token - *snowy*

Figure 11. Reweighting example of Video Editing.



(a) ID editing

(b) Color editing

Figure 12. Word Swap example of Video Editing.

You are part of a team of bots that creates multi-prompt videos. You work with an assistant bot that will draw anything you say in square brackets.

For example , outputting “a beautiful morning in the woods with the sun peaking through the trees” will trigger your partner bot to output an video of a forest morning , as described. You will assist people to create detailed, amazing videos by generating prompt groups. These grouped prompts are used to generate a single scenario, controlling the video content progression over time to create multi-prompt videos. Therefore, these prompts should not differ too much. The way to accomplish this is to first generate short prompts according to a given category, and then, extend them. When you extend the prompts, you should always keep them similar.

1. Taking two prompts in a group for example. There are some instances for generating short prompts:
Given the category “Background transition”:
" A jeep car is running on the beach, sunny.;\
A jeep car is running on the beach, night. "
You can see the generated short prompts only differ a little. And the sentences have no logic relation. Therefore, words like “the same” in the prompts are prohibited.

2. There are some rules for extending the prompts:

- Please give me prompts that are exactly same but can highlight the core differences in description.
- When modifications are requested , you should not simply make the description longer . You should refactor the entire description to integrate the suggestions.
- Video descriptions should have similar number of words as examples below. Maximum words of one prompt are 226.

Here are some examples. You should generate prompts with similar number of words as below:
"A dark knight rests motionless atop a majestic black horse in the middle of a vast grassland. The rider's armor gleams dully in the diffused light, while tall grass sways gently in the breeze. The overcast sky creates a moody atmosphere as the horse and rider remain still, surveying the expansive landscape that stretches to the horizon.;\
A dark knight guides the majestic black horse at a steady gallop across a snow-covered field. The rider's armor contrasts sharply against the white landscape, while snowflakes swirl in their wake. The overcast sky and blanket of snow create a stark winter atmosphere as the horse and rider move purposefully through the pristine terrain.;\
A dark knight guides the majestic black horse at a steady gallop across the vast desert expanse. The rider's armor shimmers brilliantly in the harsh sunlight, while sand particles dance in their wake. The blazing sky and endless dunes create a scorching atmosphere as the horse and rider move purposefully through the sun-baked terrain.;"

Let us start! The first category is “Background transition”. For 2-prompt group, 3-prompt group, 4-prompt group and 5-prompt group, first generate 13 groups of short prompts and then extend them. Give me BOTH the short prompt groups, and the extended ones.

Figure 13. Our instruction to create multiple individual long prompts based on short prompts group of specified types