

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

Minghong Cai^{1†} Xiaodong Cun² Xiaoyu Li^{3✉} Wenze Liu¹ Zhaoyang Zhang³
 Yong Zhang⁴ Ying Shan³ 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



Figure 1. Our method **DiTCtrl** takes multiple text prompts as input and demonstrates superior capability in generating longer videos with multiple events, long-range coherence and smooth transitions as output.

Abstract

Sora-like video generation models have achieved remarkable progress with a Multi-Modal Diffusion Transformer (MM-DiT) architecture. However, the current video generation models predominantly focus on single-prompt, struggling to generate coherent scenes with multiple sequential prompts that better reflect real-world dynamic scenarios. While some pioneering works have explored multi-prompt video generation, they face significant challenges including strict training data requirements, weak prompt following, and unnatural transitions. To address these problems, we propose DiTCtrl, a training-free multi-prompt video generation method under MM-DiT architectures for the first time. Our key idea is to take the multi-prompt video generation task as temporal video editing with smooth transitions. To achieve this goal, we first analyze MM-DiT’s attention mechanism, finding that the 3D full attention behaves similarly to that of the cross/self-attention blocks in the UNet-like diffusion models, enabling mask-guided precise semantic

control across different prompts with attention sharing for multi-prompt video generation. Based on our careful design, the video generated by DiTCtrl achieves smooth transitions and consistent object motion given multiple sequential prompts without additional training. Besides, we also present MPVBench, a new benchmark specially designed for multi-prompt video generation to evaluate the performance of multi-prompt generation. Extensive experiments demonstrate that our method achieves state-of-the-art performance without additional training. Code is available at <https://github.com/TencentARC/DiTCtrl>.

1. Introduction

Text-to-video (T2V) generation has made remarkable progress in the AIGC era [10, 44, 50], and breakthroughs such as Sora [30] have demonstrated impressive capabilities in generating longer videos through DiT [32] architecture and large-scale pre-training. However, feeding *sequential* prompts into current state-of-the-art text-to-video generation models (e.g., Kling [2], Gen3 [1], CogVideoX [48]) directly will produce isolated video sequences without natu-

[†] Work done during an internship at Tencent ARC Lab.

✉ Corresponding Author

ral transitions, as shown in Fig. 6 (First row). On the other hand, when multiple prompts are consolidated into a single-prompt describing long-term temporal changes, the generated results fail to capture semantic transitions effectively, as demonstrated in Fig. 6 (Second row). This limitation stems from their fundamental design and single-prompt training paradigm, making them inadequate for depicting real-world scenarios’ dynamic, multi-event nature.

Although pioneering works [4, 29, 42] have begun exploring multi-prompt video generation, they face significant challenges. *e.g.*, training such extended video generation models [29, 42] from scratch would require unprecedented computational resources and datasets that are practically unfeasible when the model size increases. Current zero-shot longer video generation methods [23, 36, 43] still mainly focus on the single prompt situation with longer length. Moreover, all previous works [4, 23, 29, 36, 42] are specifically designed under UNet architecture which restricts the abilities of more complex motions and increase the difficulties in multi-prompt generation. However, since Sora’s [30] groundbreaking demonstration of two-minute video generation, highlighting the scalability potential of DiT architectures [32]. Subsequent explorations have led to influential developments, notably in image generation models (Stable Diffusion 3 [13], FLUX.1 [6]) and video generations (CogVideoX [48], Mochi1 [14]). They [6, 13, 14, 48] all adopt a specific kind of DiT architecture, *i.e.*, Multi-Modal Diffusion Transformer (MM-DiT [13]) as the basic unit. This architecture effectively maps text and images (or video) into a unified sequence for attention computation, enabling deeper model scale abilities and achieving superior performances.

Thus, to keep the abilities of the pre-trained single prompt T2V model and take advantage of the performance of the diffusion transformer, we propose DiTCtrl, a *training-free* multi-prompt video generation method under the pre-trained MM-DiT video generation model. Our key observation is that the multi-prompt video generation can be considered a two-step problem: 1) *Video editing over time: The new video is generated through the previous video with a new prompt.* 2) *Video transition over time: Two generated videos need to keep a smooth transition between clips.* Thus, to perform consistent video editing, inspired by the UNet-based image editing techniques [9, 19], we explore the characteristic of the attention modules in the MM-DiT block for the first time, finding that the 3D full attention has similar behaviors to that of the cross-/self-attention blocks in the UNet-like diffusion models [10, 44]. We thus apply a KV-sharing method between the video clips of different prompts to maintain the semantic consistency of the key objects [9] with the 3D attention control. Besides, we utilize a latent blending strategy for transitions between clips to connect the video clip seamlessly. Finally, to systematically evaluate our method and facilitate

future research in multi-prompt video generation, we also introduce MPVBench, a new benchmark with diverse transition types and specialized metrics for assessing multi-prompt transitions. Extensive experiments on this benchmark demonstrate that our method achieves state-of-the-art performance while maintaining computational efficiency.

The contributions of this paper can be summarized as:

- We propose DiTCtrl, the first tuning-free approach based on MM-DiT architecture for coherent multi-prompt video generation. Our method incorporates a novel KV-sharing mechanism and latent blending strategy, enabling seamless transitions between different prompts without additional training.
- We pioneer the analysis of MM-DiT’s attention mechanism, finding that 3D full attention has similar behaviors to that of the cross/self-attention blocks in the UNet-like diffusion models, enabling mask-guided precise semantic control across different prompts for enhanced generation consistency.
- We introduce MPVBench, a new benchmark specially designed for multi-prompt video generation, featuring diverse transition types and specialized metrics for multi-prompt video evaluation.
- Extensive experiments demonstrate that our method achieves state-of-the-art performance on multi-prompt video generation while maintaining computational efficiency.

2. Related Work

Video Diffusion Model. Diffusion models have achieved significant success in the field of text-to-image generation [28, 37–39], and these advancements have also propelled progress in video generation from text or images [3, 7, 8, 10, 11, 15, 16, 21, 40]. Among these methods, AnimateDiff [16] attempts to turn existing text-to-image diffusion models with a motion module. Other models such as Imagen Video [21] and Make-a-Video [40] train a cascade model of spatial and temporal layers directly in pixel space. To improve efficiency, many other works [3, 7, 8, 10, 11, 15] generate the videos in latent space, leveraging an auto-encoder to compress the video into a compact latent. Notably, most of these text-to-video models utilize a U-Net architecture. Subsequently, the introduction of Sora [30] demonstrates the scalability and advantages of diffusion transformer architecture [32]. Recent works such as CogVideoX [48], Mochi1 [14], and Movie Gen [33] have adopted the DiT architecture and achieved impressive results. In this work, we build upon the open-source model CogVideoX [48], a DiT-based architecture, to explore attention control mechanisms for multi-prompt long video generation.

Long Video Generation. Training diffusion models on long videos often requires significant computational resources. Consequently, current video diffusion models are typically trained on videos with a limited number of frames. As a result, the quality of generated videos often degrades significantly during inference when generating longer videos. To address this problem, some works [17, 18, 42, 46] employ an autoregressive mechanism for long video generation. However, due to error accumulation, these methods often suffer from quality degradation after a few iterations. Alternatively, tuning-free methods [4, 23, 36, 41, 43] have been developed to extend off-the-shelf short-video diffusion models for generating long videos without additional training. For instance, Gen-L-Video [43] processes long videos as short video clips with temporal overlapping during the denoising process. FreeNoise [36] explores the influence of initial noises and conducts temporal attention fusion based on the sliding window for temporal consistency. MultiDiffusion [5] and Mimicmotion [49] introduces the latents blending strategy to achieve smooth transitions. Inspired by these works, we propose a novel KV-sharing mechanism and latent blending strategy for seamless transitions between different segments without additional training.

Image/Video Editing with Attention Control. Attention control is gaining popularity due to its ability to perform zero-shot image or video editing without the need for additional data. In the realm of image editing, MasaCtrl [9] enhances the existing self-attention mechanism in diffusion models by introducing mutual self-attention. This allows for querying correlated content and textures from source images, ensuring consistent and coherent edits. Prompt-to-Prompt [19] utilizes cross-attention layers to control the relation between text prompts and images, which has also been adopted in many image editing works [12, 31, 47]. When it comes to video editing [24, 25, 35], temporal consistency needs to be considered during attention control. Video-P2P [25] extend the cross-attention control from Prompt-to-Prompt to video editing. FateZero [35] fuses self-attention with a blending mask obtained by cross-attention features from the source prompt. However, all these works are designed for video-to-video translation editing with structure preservation. Differently, we aim for appearance-consistent video editing over time. Besides, none of these works explore attention control in diffusion transformers. In this paper, we are the first to analyze how the full attention in diffusion transformers could be utilized for video editing over time in multi-prompt video generation.

3. Method

We tackle the challenge of zero-shot, multi-prompt longer video generation without the need for model training or optimization. This allows us to generate high-quality videos with

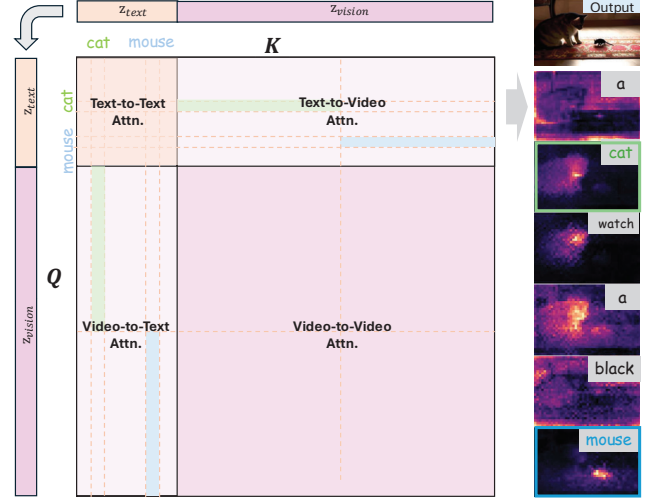


Figure 2. **MM-DiT Attention Analysis.** We find the attention matrix in MM-DiT attention can be divided into four different regions. As for the prompt of “a cat watch a black mouse”, each text token shows a high-light response using the average of the text-to-video and video-to-text attention.

smooth and precise inter-prompt transitions, covering various transition types (*e.g.*, style, camera movement, and location changes). Formally, given a pre-trained single prompt text-to-video diffusion model $\mathcal{F}$ and a sequence of n prompts $\{P_1, P_2, \dots, P_n\}$, the proposed DiTCtrl can generate a coherent longer video $\mathcal{V}_{\{1, \dots, n\}}$ that faithfully follows these prompts over time, which can be formulated as:

$$\mathcal{V}_{\{1, \dots, n\}} = \text{DiTCtrl}\{\mathcal{F}(P_1), \dots, \mathcal{F}(P_n)\}. \quad (1)$$

Below, we first give a careful analysis of MM-DiT’s attention mechanisms (Sec. 3.1). This analysis enables us to design a mask-guided full-attention KV-sharing mechanism for video editing over time (Sec. 3.2) in multi-prompt video generation. Finally, to ensure temporal coherence across different semantic segments, we further incorporate a latent blending strategy that enables smooth transitions in longer videos with multiple prompts (Sec. 3.3).

3.1. MM-DiT Attention Mechanism Analysis

The MM-DiT is the fundamental architecture of the current SOTA method of Text-to-image/Video models [6, 13, 14, 48], which is fundamentally distinct from prior UNet architectures since it maps text and videos into a unified sequence for attention computation. Although it has been widely utilized, the properties of its inner attention mechanism remain insufficiently explored, which restricts its applications in our multi-prompt longer video generation task. Therefore, for the first time, we conducted a comprehensive analysis of the regional attention patterns in the 3D full attention map based on the open-source video model, *i.e.* CogVideoX [48].

As shown in Fig. 2, due to the concatenation of the vision and text prompt, each attention matrix can be decomposed into four distinct regions, corresponding to different attention operations: video-to-video attention, text-to-text attention, text-to-video attention, and video-to-text attention. Below, we give the details of each region-inspired previous UNet-like structure with individual attentions [19].

Text-to-Video and Video-to-Text Attention. Previous UNet-like architectures incorporate cross-attention for video-text alignment. In MM-DiT, the text-to-video and video-to-text attention play a similar role. To validate its efficiency, we conduct a detailed analysis of the attention patterns, as illustrated in Fig. 2. Specifically, we compute the averaged attention values across all layers and attention heads, then extract attention values by selecting specific columns or rows corresponding to token indices in both text-to-video and video-to-text regions. These attention values are subsequently reshaped into an $F \times H \times W$ format, allowing us to visualize the semantic activation maps for individual frames. As demonstrated in Fig. 2, these visualizations show remarkable precision in token-level semantic localization, effectively capturing fine-grained relationships between textual descriptions and visual elements. This discovered capability for precise semantic control and localization provides a strong foundation for adapting established image/video editing techniques to enhance the consistency and quality of multi-prompt video generation.

Text-to-Text and Video-to-Video Attention. Text-to-text and video-to-video regional attention are somehow new from the respective UNet structure. As illustrated in Fig. 3, our analysis reveals similar patterns in both components. In the text-to-text attention component (Fig. 3(a)(b), where (a) represents the attention pattern for shorter prompts and (b) illustrates the pattern for longer prompts), we observe a prominent diagonal pattern, indicating that each text token primarily attends to its neighboring tokens. Notably, there are distinct vertical lines that shift backward as the text sequence length increases, suggesting that all tokens maintain significant attention to the special tokens at the end of the text sequence. For the video-to-video attention component, since MMDiT flat the spatial and temporal token for 3D attention calculation, our analysis at the single-frame level reveals a distinctive diagonal pattern in spatial attention (Fig. 3(c)). More significantly, when examining attention maps constructed from tokens at identical spatial positions across different frames, we also observe a pronounced diagonal pattern (Fig. 3(d)). This characteristic mirrors those found in recent UNet-based video models of the spatial-attention and temporal attention, such as VideoCrafter [26] and Lavie [45], aligning with the findings reported in [27]. Since previous works only train the specific part of the diffusion model for more advanced control and generation, our

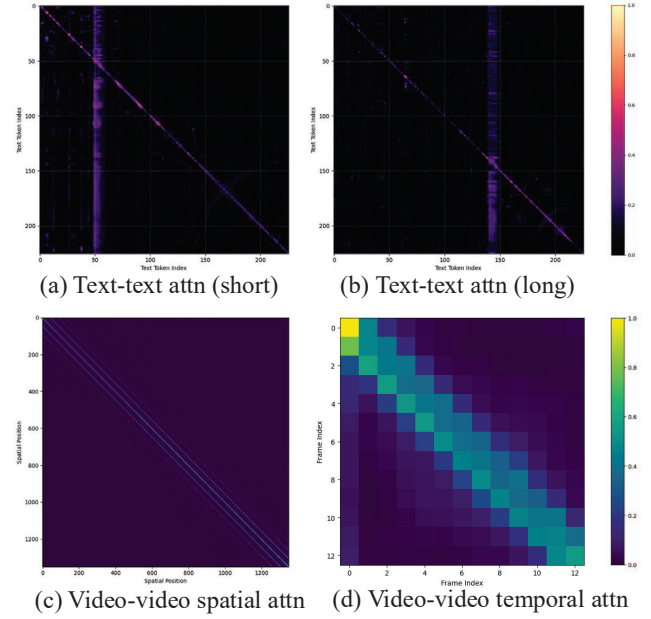


Figure 3. **MM-DiT Text-to-Text and Video-to-Video Attention Visualization.** We find that the current MM-DiT has a stronger potential to construct the individual attention in the previous UNet-like structure [10, 11, 44].

finding provides strong evidence for these methods from MM-DiT perspectives.

Overall, the presence of these consistent diagonal patterns in the MM-DiT architecture demonstrates robust frame-to-frame correlations, which proves essential for maintaining spatial-temporal coherence and preserving motion fidelity throughout the video sequence.

3.2. Consistent Video Generation Over Time

Based on the previous analysis, we propose the masked-guided KV-sharing strategy for consistent video generation over time for our multi-prompt video generation task. As shown in Fig. 4, to generate the consistent video between prompt P_{i-1} and prompt P_i , we utilize the intermediate attentions from the $i-1$ -th and i -th prompt in MM-DiT to generate the attention masks of the specified foreground object (“knight” in Fig. 4). This is achieved by averaging the Text-Video/Video-Text parts of the 3D full attention across all heads and layers with the given object tokens, then thresholded to obtain binary masks M . Inspired by MasaCtrl [9], we leverage the masks to conduct mask-guided attention fusion based on the key and values from the prompt P_{i-1} to generate the new attention features of the prompt P_i .

We denote M_{i-1} and M_i as masks extracted for the foreground objects in videos $\mathcal{V}_{i-1}$ and $\mathcal{V}_i$, respectively. With these masks, we can restrict the object in $\mathcal{V}_i$ to query con-

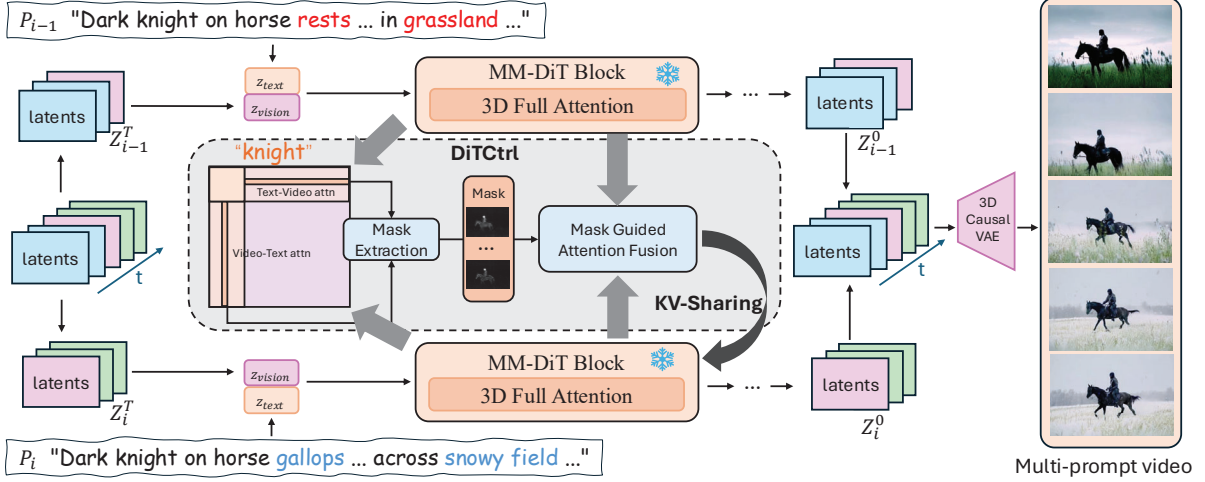


Figure 4. **Pipeline of the proposed DiTCtrl.** Note that initial latents are assumed to be 5 frames here. The first three frames are used to generate the contents of P_{i-1} , and the last three frames are used to generate contents of P_i . The pink latent represents the overlapping frame, while the blue and green latents are used to distinguish different prompt segments. Our method tries to synthesize content-consistent videos based on multi-prompts. The first video is synthesized with source text prompt P_{i-1} . During the denoising process for video synthesis, we convert the full-attention into masked-guided KV-sharing strategy to query video contents from source video $\mathcal{V}_{i-1}$, so that we can synthesize content-consistent video under the modified target prompt P_i .

tents information only from the object region in $\mathcal{V}_{i-1}$:

$$f_o^l = \text{Attention}(Q_i^l, K_{i-1}^l, V_{i-1}^l; M_{i-1}), \quad (2)$$

$$f_b^l = \text{Attention}(Q_i^l, K_{i-1}^l, V_{i-1}^l; 1 - M_{i-1}), \quad (3)$$

$$\bar{f}^l = f_o^l * M_i + f_b^l * (1 - M_i), \quad (4)$$

where $\bar{f}^l$ is the final attention output. The object regions and the background regions in the current video query the content information from corresponding restricted areas rather than all the last video features.

3.3. Latent Blending Strategy for Transition

While our previous methods enable semantic consistency between adjacent video segments, achieving smooth transitions between different semantic segments still needs to be carefully designed. Thus, we propose a latent blending strategy to ensure temporal coherence across different semantic segments, inspired by recent works [5, 36, 49].

As illustrated in Fig. 5, our approach introduces overlapped regions between adjacent semantic video segments. For each frame position in the overlapped region, we apply a position-dependent weight function that follows a symmetric distribution, in which frames closer to their respective segments receive higher weights while those at the boundaries receive lower weights. This weighting scheme ensures smooth transitions between different semantic contexts.

The blended latent feature $\mathbf{z}_t$ for frame t is calculated as:

$$\mathbf{z}_t = \frac{\sum_{i=1}^n w(t_i) \cdot \mathbf{z}_{t_i}}{\sum_{i=1}^n w(t_i)}, \quad (5)$$

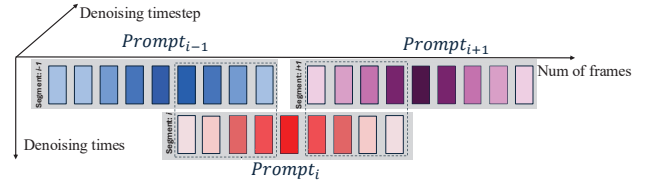


Figure 5. **Latent blending strategy** for video transition between video clips.

$$w(t_i) = \min\left(\frac{2(t_i + 0.5)}{T}, 2 - \frac{2(t_i + 0.5)}{T}\right), \quad (6)$$

where $\mathbf{z}_{t_i}$ is latent feature corresponding to t -th frame in i -th latent segment, n is number of overlapped segments, T denotes the frames number of one latent segment and $w(t_i)$ is a position-dependent weight function.

To conclude, our approach employs the latent blending strategy and kv-sharing mechanism simultaneously during each denoising step. We process segment pairs sequentially, feeding them as one batch into the MM-DiT block for mask-guided kv-sharing (Fig. 4), then blend their denoised latents progressively (Fig. 5).

4. Experiments

Setup. We implement DiTCtrl based on CogVideoX-2B [48], which is an open-source text-to-video diffusion model based on MM-DiT. In our experimental setup, we generate multi-prompt conditioned videos, where each video clip is composed of 49 frames with a resolution of 480×720 . The sample step is configured to 50. The kv-sharing steps are set

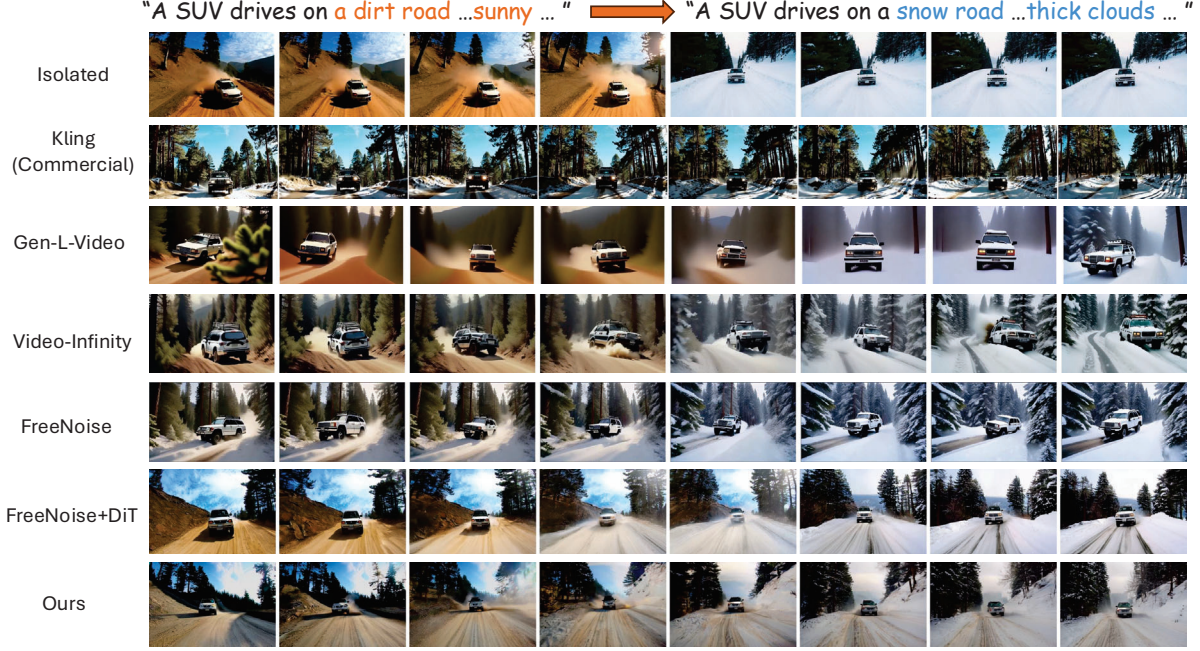


Figure 6. Comparison of generation results across methods. FreeNoise+DiT is our implementation of FreeNoise on CogVideoX.

as [2,25], and the kv-sharing layers are specified as [25,30]. For latent sampling, the number of frames is set to 13, and the overlap size is set to 6 in our experiments. All these experiments are conducted on a single NVIDIA A100 GPU. **Baselines.** We mainly compare the proposed tuning-free method to current state-of-the-art multi-prompt video generation methods [36, 41, 43] and leading commercial solutions Kling [2]. Gen-L-Video, FreeNoise, and Video-Infinity are built upon the VideoCrafter2 [11] framework. To ensure a fair comparison of base models, we implement FreeNoise [36] as an enhanced baseline by directly incorporating their noise rescheduling strategy into the CogVideoX framework.

4.1. MPVBench

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. Then, for multi-prompt video generation, we observe that the distribution of the CLIP features differs between single-prompt and multi-prompt scenarios. As shown in Fig. 7, the feature points of real video (from DAVIS [34]) follow a continuous curve, while those of two concatenated isolated videos follow two continuous curves with a breakpoint in the middle. Since the common CLIP similarity calculates the average of neighborhood similarities, the difference between real video and isolated video only occurs at the breakpoint, which becomes very small when divided by the number of frames. To address this limitation, we propose CSCV (Clip Similarity Coefficient of Variation), a metric specifically designed to evaluate the

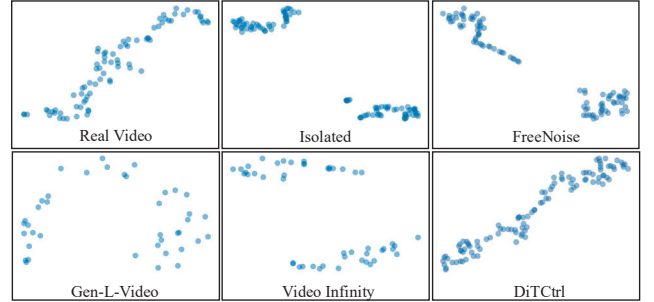


Figure 7. **t-SNE visualization of CLIP embeddings.** Each point represents the CLIP embedding of a single video frame after dimensionality reduction. The visualization demonstrates that conventional multi-prompt videos form distinct clusters, while our method produces a more continuous distribution, indicating smoother semantic transitions. More details and discussions will be given in the supplementary.

transition smoothness of multi-prompt videos, defined as:

$$s_i = \mathbf{x}_i^\top \mathbf{x}_{i+1}, \quad i = 1, \dots, n-1 \quad (7)$$

$$\text{score} = \frac{1}{1 + \lambda \cdot \frac{\sigma(s)}{\mu(s)}}, \quad (8)$$

where $\mathbf{x}_i$ denotes frame features, σ and μ are standard deviation and average respectively. The Coefficient of Variation $CV = \sigma(s)/\mu(s)$ describes the degree of uniformity, which can largely punish the isolated situation. The function $\frac{1}{1+\lambda(\cdot)}$ projects the score to $[0, 1]$, the larger the better. We also re-

Method	CSCV	Motion smoothness	Text-Image similarity
Gen-L-Video	67.28%	97.66%	30.60%
FreeNoise	84.37%	97.22%	32.69%
FreeNoise+DiT	78.74%	97.76%	30.90%
Video-Infinity	74.97%	97.31%	32.35%
DiTCtrl(w/o kv-sharing)	81.79%	97.35%	31.37%
DiTCtrl(Ours)	84.90%	97.80%	30.68%

Table 1. **Evaluation metrics.** Comparison of performance metrics for various video generation methods as benchmarked by MPVBench. Bold values represent the best performance within each group.

port the Text-Image similarity by using CLIP Similarity [20] to assess the alignment between given prompts and output video clips, and Motion smoothness from VBench [22] to evaluate whether the motion in the generated video is smooth, and follows the physical law of the real world.

4.2. Qualitative Results

The qualitative comparison with previous multi-prompt video generation methods is shown in Fig. 6. Notably, when multiple prompts are consolidated into a single-prompt describing long-term temporal changes, the generated result by Kling [2] fails to capture semantic transitions effectively, where the sun remains present while no clouds appear, which does not align with the text. Gen-L-Video [43] suffers from severe temporal jittering, compromising overall video quality. Video-Infinity [41] and FreeNoise [36] both demonstrate successful scene-level semantic changes but lack physically plausible motion. For instance, in Fig. 6, vehicles appear to be in motion while remaining spatially fixed, which is a limitation inherent to their UNet base-model 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 DiTCtrl preserves the inherent capabilities of the pre-trained DiT model while addressing these limitations, enabling smooth semantic transitions and maintaining motion coherence throughout the video sequence. For a more comprehensive evaluation, we provide additional comparisons with extensive qualitative examples in the supplementary.

4.3. Quantitative Results

We conduct the automatic evaluation with our MPVBench. From Table 1 one can see that our method achieves the highest CSCV score, demonstrating superior transition handling and overall stability in generation patterns. While FreeNoise ranks second with relatively strong stability, other methods significantly lag behind in this aspect, which is consistent with the t-SNE visualization of CLIP embedding as shown in Fig. 7. In terms of motion smoothness, our approach

Method	Overall preference	Motion Pattern	Temporal Consistency	Text Alignment
Gen-L-Video	1.15	1.14	1.08	1.25
FreeNoise	3.02	2.90	2.99	3.08
FreeNoise+DiT	3.81	3.93	3.75	3.78
Video-Infinity	2.90	2.85	2.91	2.98
DiTCtrl(Ours)	4.11	4.17	4.26	3.91

Table 2. **User study.** Human evaluation of different video generation methods across multiple aspects. Scores range from 1 to 5, with higher scores indicating better performance. Bold values represent the best performance within each metric.

exhibits superior performance in motion quality and consistency. Regarding Text-Image Similarity metrics, although FreeNoise and Video-Infinity achieve higher scores, this can be attributed to our method’s kv-sharing mechanism, where subsequent video segments inherently learn from preceding semantic content.

As shown in Fig. 6, our design choice allows the road surface to gradually transition to snowy conditions while retaining features from the previous scene. Despite potentially lower text-image alignment scores, it ensures superior semantic continuity in the sequences. In practice, this trade-off doesn’t negatively impact the visual quality in multi-prompt scenarios, as demonstrated by our user study in Table 2.

4.4. Human Evaluation

We invited 28 users to evaluate five models: Gen-L-Video [43], Video-Infinity [41], FreeNoise [36], FreeNoise+DiT and our method. We employ a Forced Ranking Scale, where items are ranked from 1 to 5, with the highest rank receiving a score of 5 and the lowest rank receiving a score of 1. Participants score each method considering overall preference, motion pattern, temporal consistency and text alignment over 16 videos generated by different scenarios. As clearly indicated in Table 2, generated videos from our method significantly outperform other state-of-the-art approaches in all four criteria, demonstrating superior capability in producing videos with natural semantic transitions that better align with human preferences for visual coherence and continuity.

4.5. More Applications

Single-prompt Longer Video Generation. Our method can naturally work on single-prompt longer video generation. As illustrated in Fig. 8, using the prompt “A white SUV drives on a steep dirt road”, our approach successfully generates videos that are more than 12 times longer than the original length, while maintaining consistent motion patterns and environmental coherence.

Video Editing. We show how we use our methods to achieve video editing performance (“word swap” and “reweight”) in Fig. 9.



Figure 8. Single prompt longer video generation example.

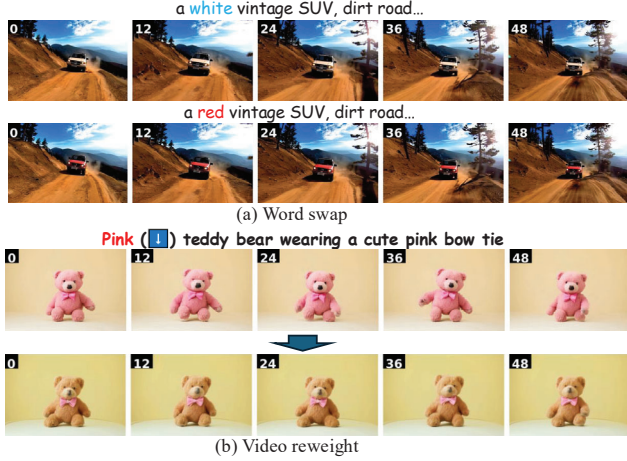


Figure 9. Video editing example.

- **Word Swap:** removing our latent blending strategy of our approach DiTCtrl, we can achieve the video editing performance of Word Swap. Specifically, we just use masked-guided KV-sharing strategy to share keys and values from source prompt P_{source} branch, so that we can synthesize a new video to preserve the original composition while also addressing the content of the new prompt P_{target} .
- **Reweight:** Similar to prompt-to-prompt [19], through reweighting the specific columns and rows corresponding to specified token (e.g. “pink”) in the MM-DiT’s Text-Video attention and Video-Text attention, we can also achieve the video editing performance of reweight.

4.6. Ablation Study

We conducted ablation studies to validate the effectiveness of DiTCtrl’s key components: latent blending strategy, KV-sharing mechanism, and mask-guided generation as shown in Fig. 10. The first row shows results that directly using text-to-video models results in abrupt scene changes and disconnected motion patterns, failing to maintain continuity in the athlete’s movements from surfing to skiing. The second row demonstrates that DiTCtrl without the latent blending strategy achieves basic video editing capabilities but lacks smooth transitions between scenes. Without KV-sharing (third row), DiTCtrl exhibits unstable environmental transitions and significant motion artifacts, with inconsistent character scaling and deformed movements. Moreover, DiTCtrl without mask guidance (fourth row) improves motion coherence and transitions but struggles with object attribute confusion across different prompts and environments.

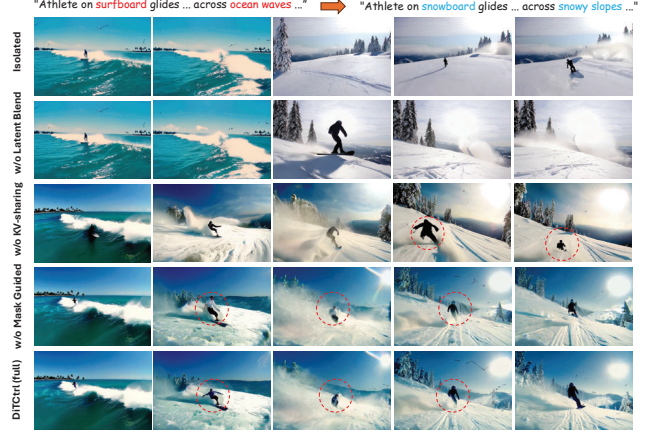


Figure 10. Visualization of ablation component in DiTCtrl.

On the other hand, The full DiTCtrl implementation provides the most precise control over generated content, demonstrating superior object consistency and smoother transitions between prompts while maintaining desired motion patterns. These results validate our analysis of MM-DiT’s attention mechanism and its role in enabling accurate semantic control.

5. Conclusion

In this paper, we introduce DiTCtrl, a novel, tuning-free method for multi-prompt video generation using the MM-DiT architecture. Our pioneering analysis of MM-DiT’s attention mechanism reveals similarities with the cross/self-attention blocks in UNet-like diffusion models, enabling mask-guided semantic control across prompts. With mask-guided kv-sharing mechanism and latent blending strategy, DiTCtrl ensures smooth transitions and consistent object motion between semantic segments, without extra training. We also present MPVBench, a new benchmark with diverse transition types and specialized metrics for assessing multi-prompt transitions.

Limitation & Future Work. While our method demonstrates state-of-the-art performance, there remain two primary limitations. First, compared to image generation models, current open-source video generation models exhibit relatively weaker conceptual composition capabilities, occasionally resulting in attribute binding errors across different semantic segments. Second, the computational overhead of DiT-based architectures presents challenges for inference speed. These limitations suggest promising directions for future research in enhancing semantic understanding and architectural efficiency.

Acknowledgements. This work is partially supported by the National Natural Science Foundation of China (Grant No. 62306261), and The Shun Hing Institute of Advanced Engineering (SHIAE) Grant (No. 8115074).

References

- [1] Gen-3. <https://runwayml.com/research/introducing-gen-3-alpha/>, 2024. 1
- [2] Kling. <https://kling.kuaishou.com/en>, 2024. 1, 6, 7
- [3] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 2
- [4] Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szepktor, Aditya Grover, and Kai-Wei Chang. Talc: Time-aligned captions for multi-scene text-to-video generation. *arXiv preprint arXiv:2405.04682*, 2024. 2, 3
- [5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 3, 5
- [6] Black Forest Labs. Announcing black forest labs. <https://blackforestlabs.ai/announcing-black-forest-labs/>, 2023. Accessed: 2024-4. 2, 3
- [7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2
- [8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2
- [9] 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*, pages 22560–22570, 2023. 2, 3, 4
- [10] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 1, 2, 4
- [11] 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. 2, 4, 6
- [12] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 3
- [13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633, 21–27 Jul 2024. 2, 3
- [14] Genmo. Mochi 1: A new sota in open-source video generation models. <https://www.genmo.ai/blog/>, 2023. Accessed: 2024-10. 2, 3
- [15] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2
- [16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yao-hui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 2
- [17] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 3
- [18] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 3
- [19] 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. 2, 3, 4, 8
- [20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7
- [21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- [22] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7
- [23] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *arXiv preprint arXiv:2405.11473*, 2024. 2, 3
- [24] Zhenyi Liao and Zhijie Deng. Lovecon: Text-driven training-free long video editing with controlnet. *arXiv preprint arXiv:2310.09711*, 2023. 3
- [25] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control, 2023. 3
- [26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 4
- [27] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Free-

- long: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024. 4
- [28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2
- [29] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, Hyeokmin Kwon, and Sangpil Kim. Mtv: Multi-text video generation with text-to-video models. *arXiv preprint arXiv:2312.04086*, 2023. 2
- [30] OpenAI. Video generation models as world simulators. <https://openai.com/index/video-generation-models-as-world-simulators/>, 2023. Accessed: 2024-2. 1, 2
- [31] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3
- [32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, October 2023. 1, 2
- [33] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2
- [34] 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. 6
- [35] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 3
- [36] 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. 2, 3, 5, 6, 7
- [37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2
- [40] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2
- [41] Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. Video-infinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260*, 2024. 3, 6, 7
- [42] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2, 3
- [43] 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. 2, 3, 6, 7
- [44] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2, 4
- [45] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 4
- [46] Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *arXiv preprint arXiv:2207.09814*, 2022. 3
- [47] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 36:26291–26303, 2023. 3
- [48] 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, 2, 3, 5
- [49] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 3, 5
- [50] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1