Zero-shot Motion Control for Multiple Objects Using Video Diffusion Models

Ryota Kaji[0009−0005−0300−0123] and Keiji Yanai[0000−0002−0431−183X]

Department of Informatics, The University of Electro-Communications, Tokyo, Japan
kaji-r@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. In recent years, numerous text-to-video models trained on large-scale video datasets have been developed, achieving remarkable generation quality. This progress has brought significant attention to video generation, building on the success of image generation models. However, many of these large-scale video generation models primarily focus on enhancing generation quality and often lack the ability to finely condition the content of the generated videos. Consequently, users benefit from reduced video production costs but frequently need to generate multiple videos to achieve their desired results. This research proposes a method for zero-shot control of multiple objects in generated videos by manipulating the Cross, Spatial, and Temporal Attention Maps of large-scale pre-trained text-to-video diffusion models. The method allows control over generated videos using text that describes the entire scene and bounding boxes that represent the trajectories of individual objects as conditioning inputs. By introducing Multi Masked Attention to manage multiple objects and Latent Alignment to ensure faithful adherence to input conditions, the proposed approach enables stable control of different movements for multiple objects in the generated videos.

Keywords: Video Generation · Diffusion Models · Zero-shot · Motion Control

1 Introduction

In recent years, numerous text-to-video (T2V) models trained on large-scale video datasets have been published [1, 2, 4, 19, 25], with models like Sora[2] achieving remarkably high generation quality, drawing attention to video generation following the success of image generation. Video is currently a trending medium, and once sufficiently capable video generation models become easily accessible, their applications could span numerous fields, including education, communication tools, and commercial applications.

However, many of the recently announced large-scale video generation models primarily focus on improving video quality, often lacking fine-grained control over the generated video content. For example, even with simple prompts like "a dog walking on grass," it is extremely difficult to express spatial conditions through language alone, such as where the dog starts and ends its walk, or at

what speed it moves. Such issues necessitate conditioning beyond language and represent a crucial challenge in the development of video generation models. The pioneering work PEEKABOO[9] proposed a method for zero-shot control of generated videos in large-scale video generation models by conditioning on both the text prompt for the desired video and the trajectory of the object to be moved within the prompt. However, its control is limited to a single object, restricting its practical applications.

Therefore, this research proposes a method for zero-shot trajectory control of multiple objects. By inputting two conditions - the text prompt for the desired video and the trajectories of each object - we enable control of multiple object trajectories by adjusting the initial latent variables and attention maps within a pre-trained T2V model. Specifically, we not only straightforwardly extend the single-object trajectory control method using masked attention in spatial, temporal, and cross attention to multiple objects but also introduce Latent Alignment for more faithful multiple objects trajectory control. Through experiments, we demonstrate the effectiveness of these methods in extending control to multiple objects.

2 Related Works

Recent text-based video generation models using Latent Diffusion Models (LDM) [17] structure have shown remarkable progress [1, 2, 4, 5, 19, 25]. VideoLDM [1] and LVDM [6] extend pretrained LDM [17] to video generation by adding temporal attention for fine-tuning in addition to the existing spatial and cross attention. Magicvideo [25] efficiently extends to video with a small number of parameters by adding an Adapter in addition to temporal attention. AnimateDiff [5] prevents quality deterioration by adding a LoRA [7] mechanism in addition to the added temporal attention-only learning pipeline. ModelScope [19] adds temporal convolution to UNet [18] in addition to temporal attention and efficiently extends it to text-to-video by jointly training images and videos. This pipeline of adding temporal layers for fine-tuning to pre-trained spatial layers has been widely adopted as an efficient method for constructing video generation models. However, these large-scale text-to-video models are often trained on vast amounts of video data to improve generation quality, with little consideration given to spatiotemporal control.

Previous studies exploring spatiotemporal control include Follow-your-click [13] and ControlVideo [24], which propose local animation and user-friendly inputs, and Free-Bloom [8], which combines large language models with text-to-image models to generate consistent videos. SG-I2V [14], DragAnything [22], MotionCtrl [20], Direct-a-Video [23] and MotionBooth [21] can control the trajectory of objects in a video using intuitive spatial conditioning as input, but they require optimization and learning of trajectories and subjects, and require significant computational resources. Additionally, there are methods that utilize structural guidance through depth maps and reference images [3], techniques that leverage LLM's prior knowledge to employ detailed object dynamics [11], and spatiotem-

poral control methods that use dense optical flow and the first frame image as conditions [15]. These methods often require additional training data, reference videos, or separate large-scale models. In contrast, PEEKABOO [9] and Trail-Blazer [12] enable zero-shot spatiotemporal control by manipulating attention maps in large-scale text-to-video models. Therefore, this research extends the masked attention proposed in PEEKABOO [9] for multiple objects.

Regarding latent variable exploration, methods such as Text2Video-Zero [10] and PYoCo [4] exist. These techniques enable zero-shot extension to video or high-quality video generation with few parameters by utilizing latent variables aligned along the time axis of the video. In this research, we propose Latent Alignment, inspired by previous studies [4], to improve temporal consistency in generated videos.

3 Preliminaries

Controlling object trajectories in videos through attention map manipulation has been attempted in many research studies. Following the method of Jain et al. [9], our approach achieves zero-shot control of generated videos by modulating attention maps in a pre-trained T2V model based on the bounding box trajectories of each conditioned object during generation. Specifically, for each of the Spatial Attention, Cross Attention, and Temporal Attention within T2V, we adjust the attention maps to guide objects along their conditioned trajectories while using attention masks to ensure foreground and background pixels attend only to their respective regions. This computation is applied only for the initial few steps of the Diffusion Models (DM) sampling process, after which generation proceeds freely. In the Masked Attention mechanism that controls regions of focus, calculations are performed as shown in equation 1 using a binary two-dimensional mask $\mathcal{M}$ for arbitrary query Q, key K, and value V.

$$\begin{aligned} \text{MaskedAttention}(Q, K, V, M) &= \text{softmax}(\frac{QK^T}{\sqrt{d}} + \mathcal{M})V \\ \text{where} \quad \mathcal{M}[i,j] &= \begin{cases} -\infty & \text{if } M[i,j] = 0 \\ 0 & \text{if } M[i,j] = 1 \end{cases} \end{aligned} \tag{1}$$

The additive mask $\mathcal{M}$ takes large negative values when the mask M is 0, enabling control over attention regions in each attention computation. The scale value $\sqrt{d}$ represents the sequence length for each query. Regarding sequence lengths, we denote the text prompt length as l_{text}, video length as l_{video}, and the spatial dimension of the latent variables as l_{latent}. Consequently, the input mask M_{input} has dimensions $l_{video} \times l_{latent}$. Additionally, we define a function $\text{fg}(\cdot)$ that takes input of pixel or text tokens and returns 1 for the foreground video elements and 0 otherwise. For the computation of the two-dimensional binary mask M, spatial and temporal attention constrain foreground and background elements to attend only within their respective regions, while cross attention controls object position, size, and movement by incorporating only relevant text embeddings

for foreground and background elements to ensure objects are generated within their conditioned bounding boxes.

3.1 Spatial and Temporal Attention Mask

For spatial attention, a two-dimensional matrix mask M^f_{SA} of size $l_{latent} \times l_{latent}$ is calculated for each frame f. For each pair of pixels, the value is 1 if both pixels are in the foreground or both are in the background. Formally, this is expressed as in Equation 2.

$$\begin{aligned} M^f_{SA}[i,j] = &\,\mathrm{fg}(M^f_{input}[i]) * \mathrm{fg}(M^f_{input}[j]) \\ &+ (1 - \mathrm{fg}(M^f_{input}[i])) * (1 - \mathrm{fg}(M^f_{input}[j])) \end{aligned} \tag{2}$$

For temporal attention, a two-dimensional matrix mask M^i_{TA} of size $l_{video} \times l_{video}$ is calculated for the pixel at coordinate i. For each pair of frames, the value is 1 if both frames are in the foreground or both are in the background. Formally, this is expressed as in Equation 3.

$$\begin{aligned} M^i_{TA}[f,k] = &\,\mathrm{fg}(M^f_{input}[i]) * \mathrm{fg}(M^k_{input}[i]) \\ &+ (1 - \mathrm{fg}(M^f_{input}[i])) * (1 - \mathrm{fg}(M^k_{input}[i])) \end{aligned} \tag{3}$$

3.2 Cross Attention Mask

For cross attention, a two-dimensional matrix mask M^f_{CA} of size $l_{video} \times l_{text}$ is calculated for each frame f. For each pair of pixel and text tokens, the value is 1 if both are in the foreground or both are in the background. Formally, this is expressed as in Equation 4.

$$\begin{aligned} M^f_{CA}[i,j] = &\,\mathrm{fg}(M^f_{input}[i]) * \mathrm{fg}(T[j]) \\ &+ (1 - \mathrm{fg}(M^f_{input}[i])) * (1 - \mathrm{fg}(T[j])) \end{aligned} \tag{4}$$

4 Method

Figure 1 shows an overview of the proposed method. This approach generates videos based on input text describing the desired video and mask images converted from the bounding-box trajectories of each object. The method consists of two main components: Multi Masked Attention, which extends the masked attention from the previous research PEEKABOO [9] to handle multiple objects, and Latent Alignment, which enhances fidelity to the given conditions.

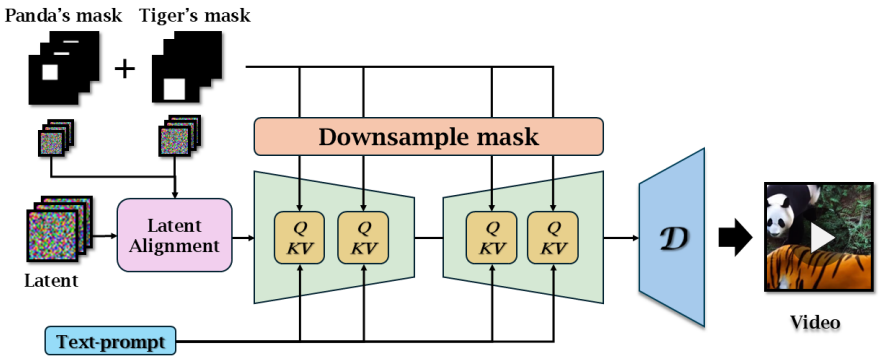

Fig. 1. Overview of the proposed method. For each input masked image, latent variables of the same size are sampled for each object and merged with the base initial latent variable (Latent Alignment). The latent variable is then input to U-Net and generated according to the trajectory conditioned through Multi Masked Attention. The mask images used for Multi Masked Attention are merged into a single mask image before each block and downsampled to the same size as the latent variable.

4.1 Multi Masked Attention

In controlling multiple objects in video generation, designing the attention mechanism to properly separate individual objects is crucial, as each object requires its own distinct trajectory conditioning. For instance, as shown in Figure 5 (w / o MA), attempting to control multiple objects in a shared space often results in undesired fusion of the target objects into a single entity. Our proposed Masked Attention method extends conventional approaches that design attention masks to prevent mutual pixel reference between foreground and background. Specifically, we introduce separate attention masks for objects within the foreground itself. As illustrated in Figure 2, we prioritize the placement of object masks similarly to segmentation, with objects closer to the foreground receiving higher priority. This design prevents objects from referencing each other with spatio-temporal attention, enabling a more independent generation of each object. Formally, for conditioned n objects $\mathbf{O} = [o_1, o_2, ..., o_n]$, we first transform each input mask $M^{o_i}_{input}$ into $\hat{M}^{o_i}_{input}$ so that it does not overlap with other input masks of objects. The attention masks are then generated using equations 2, 3, and 4. The input mask $\hat{M}^{o_i}_{input}$ for each object is defined by Equation 5.

$$\hat{M}^{o_i}_{input} = M^{o_i}_{input} \odot \prod_{j=1}^{i-1} (1 - M^{o_j}_{input}) \tag{5}$$

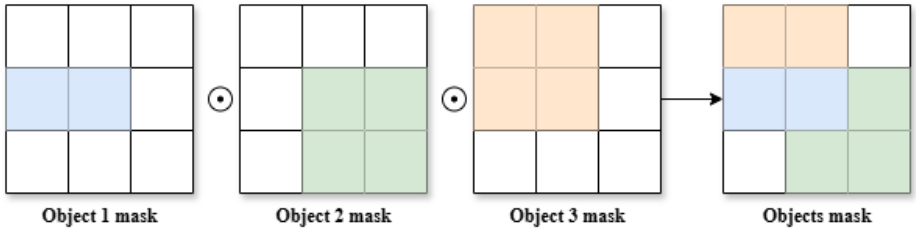

Fig. 2. Overview of Multi Masked Attention. This is an example of three object masks for a 3x3 mask image, adjusted so that the masks do not overlap each other as a way of expressing depth.

4.2 Latent Alignment

In video generation using video diffusion models, the initial latent variables during generation significantly impact the quality of the generated video. As a result, various studies have been conducted on noise design [4, 10, 16].

In this research, we introduce Latent Alignment to control multiple objects more faithfully to the input mask. This is based on the observation that content between video frames is generally similar, and latent variables between similar content are also similar. Latent Alignment uses similar latent variables across frames for the input mask areas of each object in the latent variables used for diffusion model sampling.

Specifically, in addition to the initial latent variable $\mathbf{z}$, we sample $\mathbf{z}_{o_i}$ as the base noise for each object.

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{z}_{o_i} \sim \mathcal{N}\left(\mathbf{0}, \frac{\alpha^2}{1+\alpha^2}\mathbf{I}\right) \tag{6}$$

At this point, the size of the base noise for each object is the same as the size of the input bounding box for each object. Then, after calculating $\mathbf{z}^f_{o_i}$ for frame f as represented by Equation 7, it is integrated into the initial latent variable $\mathbf{z}$ based on the input mask $\hat{M}^{o_i}_{input}$ for each object.

$$\begin{gathered} \epsilon^f_{o_i} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{1+\alpha^2}\mathbf{I}\right), \quad \mathbf{z}^f_{o_i} = \mathbf{z}_{o_i} + \epsilon^f_{o_i} \\ \mathbf{z}[\hat{M}^{o_i}_{input}] = \mathbf{z}_{o_i} \end{gathered} \tag{7}$$

Here, α is a hyperparameter that controls the proportion of base noise between each frame. When $\alpha \to \infty$, the object noise for each frame becomes equal to the base noise, and when $\alpha = 0$, the object noise for each frame becomes independent.

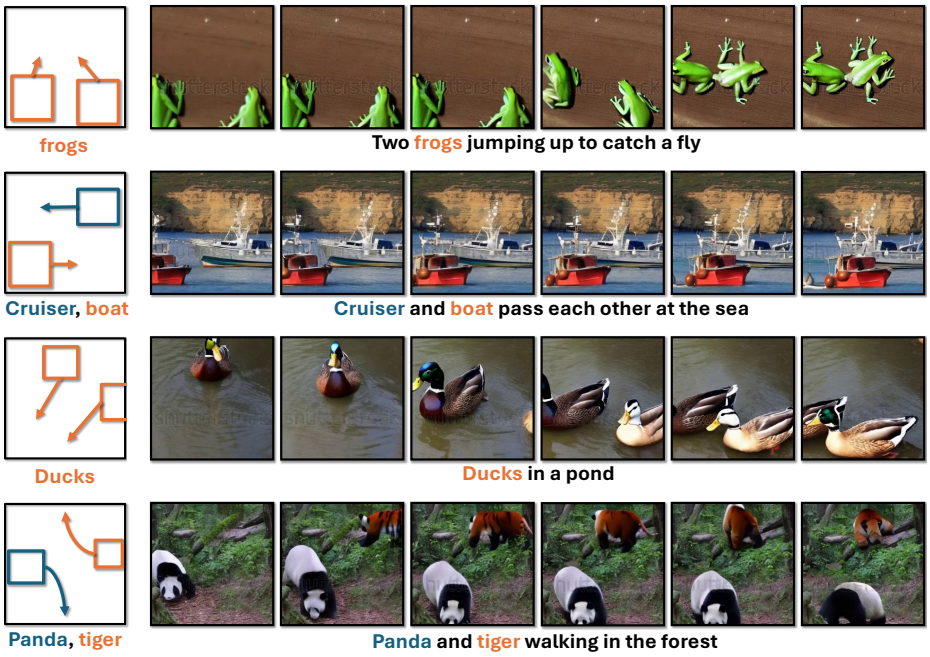

Fig. 3. Generated videos by our method.

5 Experiments

In this method, we utilized ModelScope [19] as a pre-trained text-to-video model and applied Multi Masked Attention to the spatial attention, cross attention, and temporal attention that compose the 3D U-Net within the model. The generated videos have a resolution of 16 frames at 256×256 pixels, and we used DDIM Sampler as the generation sampler. Following [9], Multi Masked Attention was applied only for the first 2 steps out of 40 generation steps, with the remaining 38 steps generated without attention map constraints. The parameter α for Latent Alignment was empirically set to 0.2.

For evaluating the proposed method, we conducted qualitative assessments of each generated result. Section 5.1 describes the generation results based on text and input masks, while Section 5.2 discusses the results of an ablation study on Multi Masked Attention and Latent Alignment.

5.1 Qualitative Results

Figure 3 shows the input text, bounding box trajectories, and generation results. On the left are the initial position bounding boxes for each object and the direction of their subsequent trajectories, with colors corresponding to the objects in the input text prompt described below the generation results. Looking

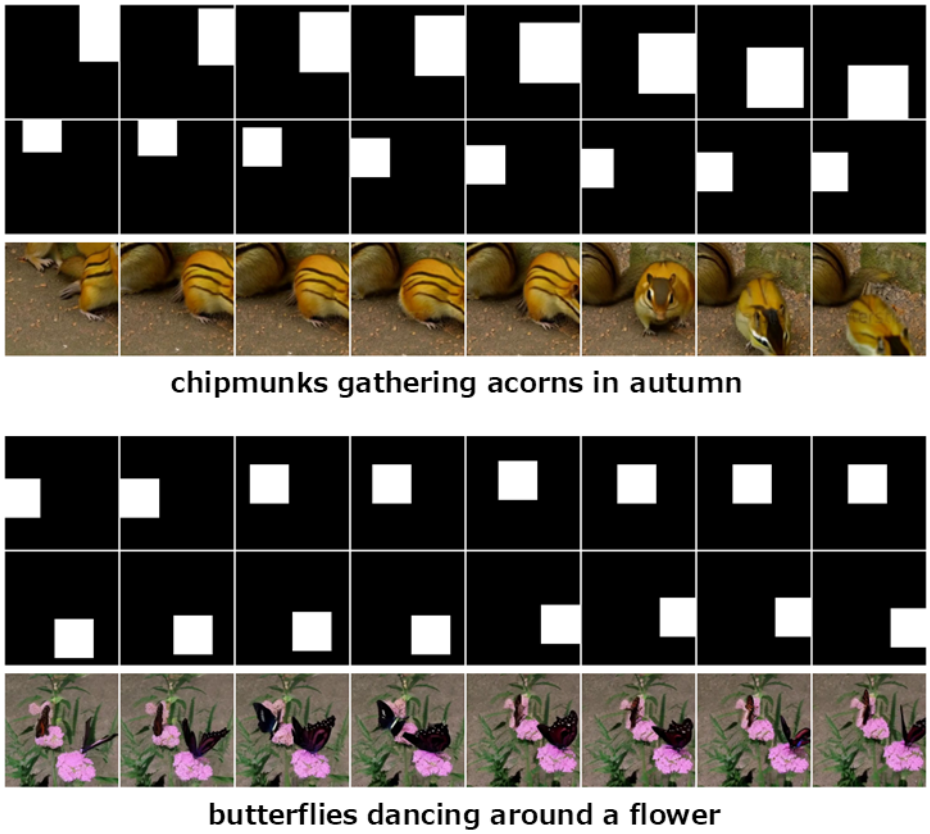

Fig. 4. Condition masks and generated videos.

at the generation results in the top two rows of Figure 3, we can see that the two frogs and two ships are generated according to the trajectories conditioned at input. For the duck generation results in the third row and the panda and tiger results in the fourth row, although one of the objects does not appear in the video in the initial frames, we can see that the generation mostly follows the given conditions.

Figure 4 shows the input masks, text prompts, and the corresponding generation results. The leftmost frame is the starting frame, and the rightmost is the final frame. Each row displays 8 frames out of the 16 generated frames, with a frame skip of 2. Each row of input masks corresponds to a single object. Examining the generation results, it's evident that both examples generally adhere to the input masks. The upper generation example demonstrates results for different mask sizes, showing that the proportion of the screen occupied by each object increases in proportion to its mask size. The lower example shows the result when guidance is applied in almost the same direction for each object. In this case, it's clear that the objects don't have their own individual dynamics, but rather the entire scene moves as if the camera is panning.

5.2 Ablation Study

Figure 5 shows the results of an ablation study on the two components: Masked Attention and Latent Alignment. In Figure 5, MA stands for Multi Masked Attention, and LA stands for Latent Alignment. Looking at the results without Multi Masked Attention, we can see that only the panda appears in the video, and its movement is completely different from the mask trajectory. This indicates that Multi Masked Attention is an essential component. Next, examining the results without Latent Alignment, we can see that while there is slight guidance from Multi Masked Attention, it is insufficient for conditioning on its own. This is evident as the tiger disappears in the latter half of the video and the number of pandas increases. These results demonstrate that Multi Masked Attention alone is not enough for proper conditioning.

6 Limitation

There are several challenges with the current approach. First, there is an issue with maintaining consistency between frames. Upon observing multiple generated examples, it is often noticeable that parts of objects change between the first and last frames. It is necessary to explore whether improving the base text-to-video model is essential to address this problem. Second, there is an observed issue where parts of multiple generated objects become mixed together. For instance, in generations such as a panda and a tiger walking, results show the tiger's head being replaced with the panda's head. This suggests a need for a structure that generates objects more independently from each other during the generation process.

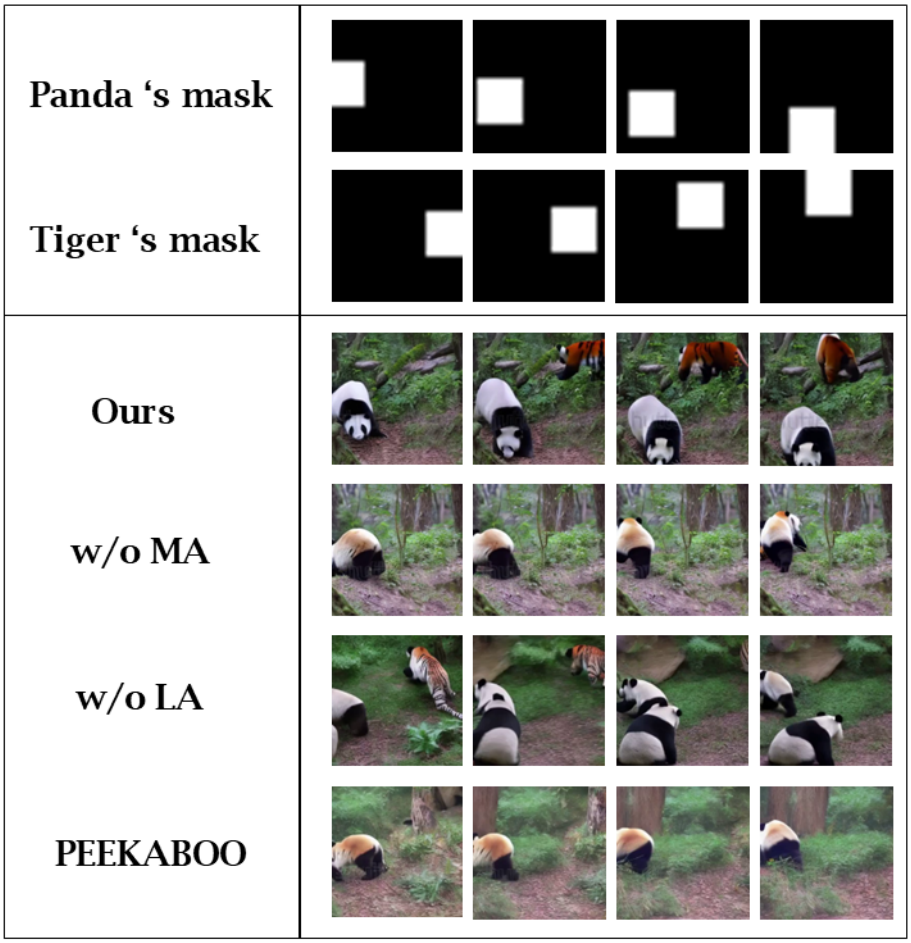

Fig. 5. Results of the ablation studies.

7 Conclusion

This research proposes a method for zero-shot control of multiple objects by manipulating the cross, spatial, and temporal attention maps of large-scale pre-trained text-to-video diffusion models. By introducing Multi Masked Attention to manage multiple objects and Latent Alignment to ensure faithful adherence to input conditions, the method enables stable control of distinct movements for various objects.

However, the proposed approach still faces several challenges. For instance, the randomness in generating base noise for objects in Latent Alignment can significantly impact the generation quality, while object mixing may occur during free generation following guidance in Multi Masked Attention. To address these challenges associated with multi-object handling, we aim to explore further improvements in future work.

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22563–22575 (2023)
2. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators
3. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Int. Conf. Comput. Vis. pp. 7346–7356 (2023)
4. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Int. Conf. Comput. Vis. pp. 22930–22941 (2023)
5. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: Int. Conf. Learn. Represent. (2023)
6. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
7. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (2022)
8. Huang, H., Feng, Y., Shi, C., Xu, L., Yu, J., Yang, S.: Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. Adv. Neural Inform. Process. Syst. **36** (2024)
9. Jain, Y., Nasery, A., Vineet, V., Behl, H.: Peekaboo: Interactive video generation via masked-diffusion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8079–8088 (2024)
10. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Int. Conf. Comput. Vis. pp. 15954–15964 (2023)

11. Lian, L., Shi, B., Yala, A., Darrell, T., Li, B.: Llm-grounded video diffusion models. In: Int. Conf. Learn. Represent. (2023)
12. Ma, W.D.K., Lewis, J.P., Kleijn, W.B.: Trailblazer: Trajectory control for diffusion-based video generation. In: ACM SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
13. Ma, Y., He, Y., Wang, H., Wang, A., Qi, C., Cai, C., Li, X., Li, Z., Shum, H.Y., Liu, W., Chen, Q.: Follow-your-click: Open-domain regional image animation via short prompts. In: arXiv preprint arXiv:2403.08268 (2024)
14. Namekata, K., Bahmani, S., Wu, Z., Kant, Y., Gilitschenski, I., Lindell, D.B.: Sg-i2v: Self-guided trajectory control in image-to-video generation. arXiv preprint arXiv:2411.04989 (2024)
15. Niu, M., Cun, X., Wang, X., Zhang, Y., Shan, Y., Zheng, Y.: Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In: Eur. Conf. Comput. Vis. (2024)
16. Qiu, H., Chen, Z., Wang, Z., He, Y., Xia, M., Liu, Z.: Freetraj: Tuning-free trajectory control in video diffusion models. arXiv preprint arXiv:2406.16863 (2024)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
19. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
20. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
21. Wu, J., Li, X., Zeng, Y., Zhang, J., Zhou, Q., Li, Y., Tong, Y., Chen, K.: Motionbooth: Motion-aware customized text-to-video generation. arXiv preprint arXiv:2406.17758 (2024)
22. Wu, W., Li, Z., Gu, Y., Zhao, R., He, Y., Zhang, D.J., Shou, M.Z., Li, Y., Gao, T., Zhang, D.: Draganything: Motion control for anything using entity representation. In: Eur. Conf. Comput. Vis. pp. 331–348. Springer (2025)
23. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
24. Zhang, Y., Wei, Y., Jiang, D., ZHANG, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. In: Int. Conf. Learn. Represent. (2023)
25. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)