Motion-Zero: A Zero-Shot Trajectory Control Framework of Moving Object for Diffusion-Based Video Generation

Changgu Chen[1] Junwei Shu[1], Gaoqi He[1], Changbo Wang[2*], Yang Li[1*],

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]School of Data Science and Engineering, East China Normal University, Shanghai, China
{52215901006, 51265902091}@stu.ecnu.edu.cn, {gqhe, cbwang, yli}@cs.ecnu.edu.cn,

Abstract

Recent large-scale pre-trained diffusion models have demonstrated a powerful generative ability to produce high-quality videos from detailed text descriptions. However, exerting control over the motion of objects in videos generated by any video diffusion model remains a challenging problem. In this paper, we propose a novel zero-shot moving object trajectory control framework, Motion-Zero, to enable arbitrary single-object-trajectory control for the text-to-video diffusion model. To this end, an initial noise prior module is designed to provide a position-based prior to improve the stability of the appearance of the moving object and the accuracy of position. In addition, based on the attention map of the U-Net, spatial constraints are directly applied to the denoising process of diffusion models, which further ensures the positional consistency of moving objects during the inference. Furthermore, temporal consistency is guaranteed with a proposed shift temporal attention mechanism. Our method can be flexibly applied to various state-of-the-art video diffusion models without any training process. Extensive experiments demonstrate our proposed method can control the motion trajectories of arbitrary objects while preserving the original ability to generate high-quality videos.

Introduction

In recent years, the generative capabilities of diffusion models have been widely recognized in both text-to-image (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Song, Meng, and Ermon 2020) and text-to-video domains (Guo et al. 2023; Ho et al. 2022). Although these video models are capable of producing high-definition, high-resolution, and fluid video animations, the dynamic motion trajectories of the generated objects are relatively random in existing text-to-video models (Blattmann et al. 2023; Zhang et al. 2023). Thus, accurate control of an object's motion trajectories in a generated video remains rudimentary.

Various strategies have been proposed to address this problem. Some methods (Zhang, Rao, and Agrawala 2023; Wu et al. 2023a) try to control the trajectory of the moving object by providing a detailed conditional motion sequence, such as a dancing skeleton. However, the cost of acquiring conditional control sequences is non-negligible and inhibits the diverse generation of video outputs by the user. To obtain variety over the generated video with more control capability, i.e. trajectories of moving objects, several methods (Wang et al. 2023d,c; Yin et al. 2023) harness a substantial dataset of motion and trajectory pairs to train the baseline models. Although the results are very impressive and promising, these methods are solely applicable to the base models on which they were trained and cannot be applied to other models directly. Furthermore, all of these methods require extensive training and significant computational resources, preventing ordinary users from using them.

Nevertheless, the fundamental text-to-video diffusion models (Blattmann et al. 2023; Zhang et al. 2023) have undergone an extremely large scale training dataset (Bain et al. 2021; Xue et al. 2022). Theoretically, the pre-trained model should inherently have learned extensive knowledge about the dynamics of a variety of object movements. However, the semantics of the latent space in different modules of a video diffusion model are not explicitly defined, which makes it difficult to manually control generated video in desired motion dynamics. To some extent, the methods mentioned above try to align and define those latent spaces in the baseline model by training with additional labeled datasets, and successfully harnessing the intrinsic knowledge of pre-trained models to control the motion. In addition, we observe that the initial noise plays a significant inspirational role in the generation of the videos. The same initial noise tends to produce videos with similar content. With these two characteristics, we can manipulate the trajectory of moving objects.

In this paper, we propose a zero-shot trajectory control framework, Motion-Zero, to guide the motion of generated objects in video. Our proposed method can be easily applied to pre-trained video diffusion model while achieving universality and plug-and-play capabilities of controlling. To this end, we first design an Initial Noise Prior Module to provide noises based on the motion trajectories given by the user. Then, inspired by the fact that the values of the cross-attention map within the U-Net largely determine the generation location of the subject mentioned in the prompt (Xie et al. 2023; Epstein et al. 2023), Spatial Constraints with an attention similarity loss over these attention maps are proposed to achieve precise manipulation of object positioning within individual frames. Furthermore, we observe that merely imposing spatial constraints on the positions of objects can negatively

*Corresponding authors.

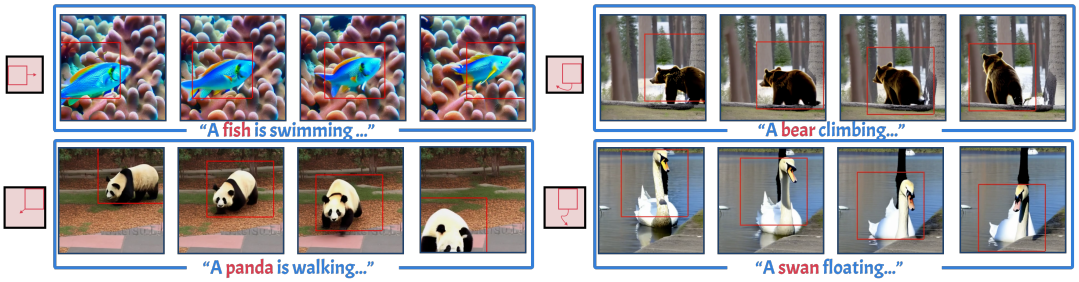

Figure 1: Our Motion-Zero framework endows different pre-trained video diffusion models with the capability to manipulate object trajectories directly, circumventing the need for supplementary training. By designating the target entity in the input prompts and a sequence of bounding boxes, users can intuitively direct the motion path of the object within the generated video.

impact the overall coherence and video quality. To preserve the continuity throughout the video sequence, a novel Shift Temporal Attention Mechanism is proposed to further maintain the baseline model focus on the same objects across the time axis. Through the implementation of these modules, our proposed framework can effectively generate high-quality video controlled by bounding-box trajectories without any training process. Samples generated by Motion-Zero can be viewed in Fig.1. The main innovations of our work can be summarized as follows,

- We propose a zero-shot framework Motion-Zero, which is capable of controlling generated object for arbitrary trajectory within a pre-trained video generation diffusion model. Our Motion-Zero is plug-and-play and without any additional training.
- Our Initial Noise Prior Module is a novel way to acquire semantic initial noise for high-quality generation. Moreover, Spatial Constraints and Shift Temporal Attention Mechanism respectively exploit the cross-attention and temporal-attention dimensions to obtain spatial control and temporal consistency.
- Extensive experiments demonstrate that our proposed work can be applied to pre-trained video diffusion models. Our method enables the original baseline to generate objects with arbitrary trajectories.

Related Works

With the rapid development of deep learning (Chen et al. 2023b; Zheng et al. 2024b,a; Wang et al. 2023a; Sun et al. 2024, 2023; Li et al. 2024b,a; Wang et al. 2024c,b; Yang, Li, and Chen 2024; Cai et al. 2021, 2022, 2023; Chen et al. 2024, 2023a; Song et al. 2023b, 2024, 2023a), there are pre-trained methodologies in place that have the capability to control the motion trajectories of objects within generated videos. VideoComposer (Wang et al. 2023c) employs a two-stage training strategy to incrementally incorporate temporal information and control signals. A motion condition encoder is proposed for training, designed to enhance the model's ability to understand and integrate motion-specific information. DragNUMA (Yin et al. 2023) also employs a comparable training strategy and utilizes optical flow as a conditioning mechanism for trajectory modeling. MotionCtrl (Wang et al. 2023d) synthesizes the approaches of these two works and introduces a camera control module and a trajectory control module to further refine the management of movement within generated video content. Boximator (Wang et al. 2024a) proposes a novel self-tracking technique to simplify the learning of box-object relationships. MotionBooth (Wu et al. 2024) allows for controlling the movement of a specific object by fine-tuning a photo of that object. (Yu et al. 2024) enables the movement of a specified object in a user's input photo to generate a video. Motion-I2V (Shi et al. 2024) leverages a two-stage training process, allowing users to alter the motion in the generated video by dragging. These models typically rely on extensive training on large-scale datasets such as WebVid-10M (Bain et al. 2021) and HD-VILA-100M (Xue et al. 2022) which lead to expensive training costs. Also, these models are typically constrained to operate on the specific models they were trained on, lacking the flexibility to interchange base video diffusion models. This limitation hinders their adaptability and limits the scope of their application to only those scenarios for which they were explicitly designed. Our proposed model stands out in that it can be applied to any base diffusion model without the necessity for further training. Recently, Trailblazer (Ma, Lewis, and Kleijn 2023) enhances attention on the box area in cross-attention with a zero-shot setting, similar to our work. In contrast, our operation on cross-attention is supervised by losses, providing better interpretability and effectiveness. Peekaboo (Jain et al. 2024) uses a masked attention module to achieve control without the need for training. FreeTraj (Qiu et al. 2024) utilizes frequency fusion to generate results whose trajectory is aligned with the given box. Although these methods exploit zero-shot settings in controllable video generalization, our proposed method emphasises on the importance of the initial noise and the control of spatial and temporal consistency, and achieves superior performance.

Preliminaries

Video Diffusion Model: Video diffusion models are designed to produce high-quality and diverse videos, guided by text prompts. To save the computational costs, (Rombach et al. 2022) utilize a U-Net as a denoising model within a latent space, significantly reducing the computational load in terms of both time and space.

In detail, these models employ a Variational Autoencoder

(VAE), which comprises an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder compresses the original video from pixel space into a latent representation, and the decoder reconstructs the video from this latent space back to pixel space. The 3D U-Net typically consists of a series of down-sampling blocks, middle blocks, and up-sampling blocks. Each block is equipped with convolutional layers, spatial transformers, and temporal transformers. The optimization of the 3D U-Net (denoted as ϵ_θ) is executed through a noise prediction loss function:

$$\mathcal{L} = \mathcal{E}_{\mathbf{z}_0,c,\epsilon\sim N(0,I),t}[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})||_2^2], \tag{1}$$

where z_0 is the latent code of the training videos, c is the text prompt condition, ϵ is the Gaussian noise added to the latent code, and t is the time step. The noised latent code $\mathbf{z}_t$ is determined as:

$$\mathbf{z}_t = \sqrt{\overline{a}_t}\mathbf{z}_0 + \sqrt{1-\overline{a}_t}\epsilon, \overline{a}_t = \prod_{i=1}^{t} a_t, \tag{2}$$

where a_t is a hyper-parameter used for controlling the noise strength based on time t.

Motion Trajectory Control: Based on video diffusion, the task of motion trajectory control is to precisely control the motion trajectory of objects in generated videos. The optimization objective can be formulated as:

$$\mathcal{L} = \mathcal{E}_{\mathbf{z}_0,c,\epsilon\sim N(0,I),t}[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathcal{B})||_2^2]. \tag{3}$$

Specifically, users can input a text condition $\mathbf{c}$ along with a sequence of rectangular boxes $\mathcal{B} = \{(x_1^f, y_1^f), (x_2^f, y_2^f)\}^{N_f}$, where $(x_1^f, y_1^f), (x_2^f, y_2^f)$ are the upper left and lower right points of the box in frame f, N_f is the total frame number. The boxes $\mathcal{B}$ correspond to the position of the moving object.

Methodology

Overview

Existing trajectory control methods require large-scale video training data to optimize Eq.3, leading to a high computational cost. Differently, our work falls into the setting of zero-shot motion trajectory control. To this end, our proposed Motion-Zero framework operates entirely during the inference stage, thus eliminating the need for training and making it compatible with various pre-trained video diffusion models. The pipeline is shown in Fig.2 (a). In the following section, we provide a detailed presentation of our proposed Motion-Zero framework and its components. Firstly, we introduce a noise prior generation module to produce the initial noise for diffusion. Then, we describe the designed constraints and loss of the moving object position and spatial consistency. Finally, our shift temporal attention mechanism is presented to further improve the temporal consistency.

Initial Noise Prior Module

According to the theory of DDIM Inversion (Ho, Jain, and Abbeel 2020), the initial noise has a significant impact on the final generated outcome. We introduce Initial Noise Prior Module (INPM) to leverage this property to provide a strong prior for the position of the moving object. Several steps are involved to integrate a moving object into a sequence of frames with a coherent prior as shown in Fig.2 (b). Firstly, given a prompt $\mathbf{c}$ and the boxes in the same location with $\{\mathcal{B}^0\}^{N_f}$, a meta video V_{meta} is sampled, $\mathbf{z}^* \sim \mathcal{N}(0, \mathbf{I})$ as latent input, using the baseline video diffusion model and our proposed spatial constraints (introduced in the next section). This generated video has the target object staying at the location with $\{\mathcal{B}^0\}^{N_f}$ due to the spatial constraints. It is noteworthy that controlling the model to generate standing-still objects is much easier than controlling the movement of objects. Then, a video latent $\mathbf{z}_{meta}$ is generated based on V_{meta} from Encoder $\mathcal{E}$. Once $\mathbf{z}_{meta}$ is prepared, we perform a DDIM Inversion to obtain the corresponding noise latent representation $\mathbf{z}_I$. We crop the latent representation within the box $\mathcal{B}^0$ for each frame, creating a sequence of latent patches containing the visual target. Subsequently, we use a local mixup operation (Zhang et al. 2017) to mix the latent patches and the initial noise $\mathbf{z}^*$ in the range of $\mathcal{B}^f$ frame by frame. Our INPM allows us to set a coherent prior in the corresponding object's position in the initial noises. It also ensures that the animated object maintains consistency in appearance and movement across the video frames, without incurring additional computational costs during the generation process. Details are shown in Alg.1.

Algorithm 1: Initial Noise Prior Module

Input: $\mathbf{c}$, $\mathcal{B}$
Parameter: N_f, λ_p
Function: $\mathbf{Pix}(\mathbf{a}, \mathbf{b})$ means the elements get from tensor $\mathbf{a}$ in the range of box $\mathbf{b}$
Output: Initial noise $\mathbf{z}_T$

$\mathbf{z}^* \sim \mathcal{N}(0, \mathbf{I})$ ▷ random sample the first latent code
$\mathbf{V}_{meta} \leftarrow \mathbf{VideoDiffusion}(\mathbf{z}^*, \mathbf{c}, \mathcal{B}_0)$ ▷ Video Diffusion using SC, the first box and $\mathbf{c}$ as condition
$\mathbf{z}_{meta} \leftarrow \mathbf{Encoder}(\mathbf{V}_{meta})$
$\mathbf{z}_I \leftarrow \mathbf{DDIM\_Inverse}(\mathbf{z}_{meta})$ ▷ inverse the $\mathbf{z}_{meta}$
$\mathbf{z}_T \leftarrow \mathbf{z}^*$
for all f=1,2,...,N_f **do**
  $\mathbf{Pix}(\mathbf{z}_T^f, \mathcal{B}^f) \leftarrow \lambda_p \cdot \mathbf{Pix}(\mathbf{z}_I^f, \mathcal{B}^\mathbf{0}) +$ $(1\text{-}\lambda_p) \cdot \mathbf{Pix}(\mathbf{z}_T^f, \mathcal{B}^f)$ ▷ local mixup operation
end for
return $\mathbf{z}_T$ ▷ initial latent with position prior

Spatial Constraints with Attention Map

The INPM alone is insufficient for precise manipulation of an object's trajectory. To further improve the capability of control, we introduce Spatial Constraints (SC) deployed at each denoising step t to optimize the intermediate latent representation $\mathbf{z}_t$. This optimization is crucial for enhancing the accuracy of the moving object position and preserving spatial consistency. Within the conditional denoising architecture (Rombach et al. 2022), cross-attention serves as the pivotal bridge that connects the text prompt with the content generated. During the denoising steps, conditioned on the prompt $\mathbf{c}$ and the intermediate features $\mathbf{z}_t$, the corresponding cross-attention map can be obtained as $\mathbf{A}$:

$$\mathbf{A} = \mathbf{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}), \tag{4}$$

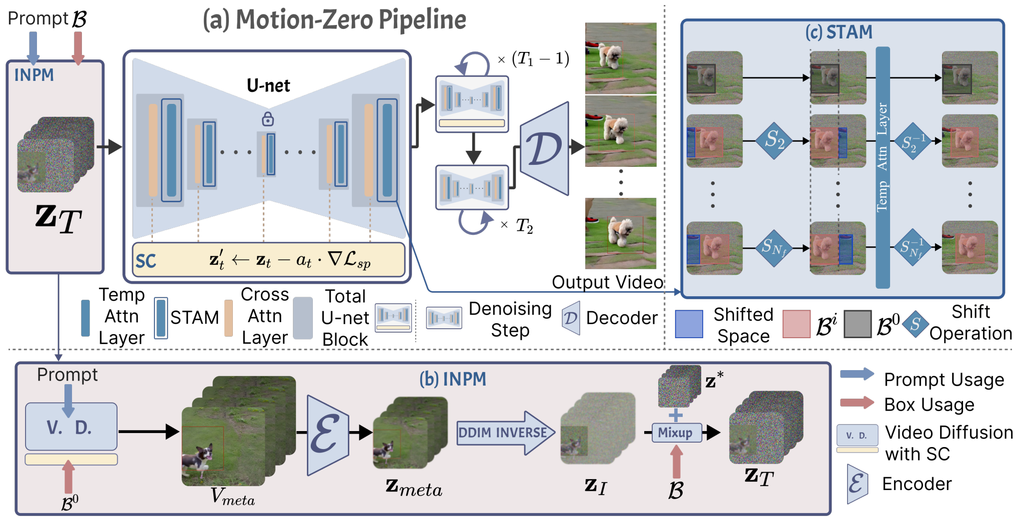

Figure 2: Overview of our Motion-Zero. The total pipeline is shown in (a). Given the box condition $\mathcal{B}$ and the prompt condition, we generate the prior latents $\mathbf{z}_T$ by our Initial Noise Prior Module (INPM) as shown on (b). At timestep t, $\mathbf{z}_t$ is firstly optimized to $\mathbf{z}'_t$ by the Spatial Constraints (SC). Subsequently, $\mathbf{z}'_t$ is passed to the UNet with Shift Temporal Attention Module (STAM) as demonstrated on (c). All the parameters of the video diffusion are frozen. T_1 represents the number of timesteps during which SC and STAM are applied, and T_2 denotes the number of timesteps where the original video diffusion process is utilized.

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{z}_t, \mathbf{K} = \mathbf{W}_K \mathbf{c}, \tag{5}$$

where $\mathbf{Q}, \mathbf{K}$ are the query and key with the $\mathbf{z}_t$ and $\mathbf{c}$, respectively. $\mathbf{W}_Q, \mathbf{W}_K$ are two learnable matrices, which are frozen in our settings. Assuming a maximum number N_p of prompt tokens $\{\mathbf{p}_1, ..., \mathbf{p}_{N_p}\}$ in condition $\mathbf{c}$, at time step t, it is feasible to derive N_p cross-attention maps $\{\mathbf{A}_1, ..., \mathbf{A}_{N_p}\}$.

When a user specifies the index k of prompt tokens $\mathbf{p}_k$ intended to control the trajectory of an object, along with the box $\mathcal{B}$, the box-loss can be employed to ensure that our model controls the object to appear within the input box in every frame. Specifically, the box-loss is achieved through an optimization approach that maximizes the values of the $\mathbf{A}_k$ for the corresponding prompt tokens inside the box, while minimizing the values outside the box. Frame by frame, we scale down the user-specified coordinate box to the corresponding coordinates in the latent space and construct a mask $\mathbf{M}^f$ where 1 for areas in the box and 0 otherwise. To ensure the values inside the box for attention are maximized, we propose an intuitive solution $\mathcal{L}_i$ as

$$\mathcal{L}_i{}^f = 1 - \frac{1}{P}\sum \mathbf{g}(\mathbf{A}_k^f \cdot \mathbf{M}^f, P), \tag{6}$$

where $\mathbf{g}(\cdot, P)$ means the top P highest value will be selected. If we employ all the values within the attention maps, it could disrupt the stability of the denoising process. Conversely, utilizing too few values could result in a less pronounced control effect. On the other hand, we aim for the attention values outside the box to be as minimal as possible which is formulated as

$$\mathcal{L}_o{}^f = \frac{1}{P}\sum \mathbf{g}(\mathbf{A}_k^f \cdot (1 - \mathbf{M}^f), P). \tag{7}$$

Within these two constraints, we can ensure that the object generated in each frame is contained within the box. However, there is no guarantee that the object will be positioned at the center of the box. To mitigate this, we propose another center-loss $\mathcal{L}_c$ to encourage the centroid of the $\mathbf{A}_k^f$ to closely align with the center of the box as

$$(W_{\mathbf{A}_k^f}, H_{\mathbf{A}_k^f}) = \frac{1}{\sum_{w,h}\mathbf{A}_{k,w,h}^f}(\textstyle\sum_{w,h} w \cdot \mathbf{A}_{k,w,h}^f, \sum_{w,h} h \cdot \mathbf{A}_{k,w,h}^f), \tag{8}$$

$$\mathcal{L}_c{}^f = ||(\frac{x_1^f + x_2^f}{2}, \frac{y_1^f + y_2^f}{2}) - (W_{\mathbf{A}_k^f}, H_{\mathbf{A}_k^f})||_1, \tag{9}$$

where $\mathbf{A}_{k,w,h}^f$ is the value of $\mathbf{A}_k^f$ at the position of (w, h), $(W_{\mathbf{A}_k^f}, H_{\mathbf{A}_k^f})$ is the centroid position of $\mathbf{A}_k^f$ with the k-th token concept. After establishing robust control over the position of the object, we recognize unexpected variations in the object's appearance due to the substantial extent of movement. To ensure appearance consistency across frames, we strive to maintain the uniformity of the $\mathbf{A}_k^f$ within the box. To this end, we introduce a similarity loss $\mathcal{L}_s$ as

$$\mathcal{L}_s = 1 - \frac{1}{N_f - 1}\sum_{f=1}^{N_f - 1} \mathbf{Sim}(\mathbf{Pix}(\mathbf{A}_i^f, \mathcal{B}^f), \mathbf{Pix}(\mathbf{A}_i^{f+1}, \mathcal{B}^{f+1})), \tag{10}$$

where $\mathbf{Sim}(\cdot, \cdot)$ means the similarity of two elements and $\mathbf{Pix}()$ is a function getting corresponding elements of the map $\mathbf{A}^f$ within a box range $\mathcal{B}^f$. Cosine similarity is adopted here. At each timestep, the overall spatial constraints $\mathcal{L}_{sp}$ are formulated as follows:

$$\mathcal{L}_{sp} = \sum_f (\lambda_i \mathcal{L}_i{}^f + \lambda_o \mathcal{L}_o{}^f + \lambda_c \mathcal{L}_c{}^f) + \lambda_s \mathcal{L}_s, \tag{11}$$

where $\lambda_i, \lambda_o, \lambda_c, \lambda_s$ are hyper-parameters. By minimizing and calculating the gradient of the Eq.11, we can optimize our latent $\mathbf{z}_t$ in Eq.2 as follows:

$$\mathbf{z}'_t \leftarrow \mathbf{z}_t - \beta_t \cdot \nabla\mathcal{L}_{sp}, \tag{12}$$

where β_t linearly decays at each timestep t. Specifically, before denoising with the U-Net at each timestep t, we update $\mathbf{z}_t$ to $\mathbf{z}'_t$ using Eq.12. Then, we continue the denoising process using $\mathbf{z}'_t$. Under the combined effect of the aforementioned constraints, the latent variable $\mathbf{z}_t$ at each timestep will gradually shift towards generating high-response attention in the specified position, while ensuring that the appearance attributes of the object within the box remain unchanged. Consequently, the target object is synthesized within the bounding box area provided by the user.

Shift Temporal Attention Mechanism

After applying the spatial constraints on attention maps, pretrained video diffusion models still encounter difficulties in generating sequences of continuous actions. Within the temporal module of the diffusion process, the latent representation is reshaped into the following configuration:

$$z'_{(b\cdot H\cdot W+h\cdot W+w,f,c)} = z_{(b,f,c,h,w)} \tag{13}$$

where z' represents the result of applying the rearrange operation to z. Within the temporal transformer, attention is focused on the same pixel across different frames. This leads to a scenario where, if the extent of motion is too large, the same position in different frames could undergo significant semantic changes, resulting in a lack of coherence in the generated dynamics.

To overcome this inconsistency, we propose a Shift Temporal Attention Mechanism (STAM) to improve the dynamics of the moving object in different frames. Specifically, we shift the elements of $\mathbf{z}^f$ inside the $\mathcal{B}^f$ range with the elements inside the $\mathcal{B}^0$ range and use the overlapped parts to fill in the vacated spaces, as shown in Fig.2 (c). Therefore, the subsequent frames within the box range can be aligned with the box range of the first frame. The steps are shown:

$$\begin{aligned} \mathbf{z}_w^f &= \mathbf{Shift}(\mathbf{z}^f, \mathcal{B}^f, \mathcal{B}^0), \\ \mathbf{z}_w{}' &= \mathbf{TemporalAttention}(\mathbf{z}_w), \\ \mathbf{z}^{f\prime} &= \mathbf{Shift}(\mathbf{z}_w^{f\prime}, \mathcal{B}^0, \mathcal{B}^f), \end{aligned} \tag{14}$$

where $\mathbf{Shift}(\cdot, a, b)$ is the shift operation, $\mathbf{z}_w^f$ means the shifted latent $\mathbf{z}^f$ of frame f and $\mathbf{z}_w = [\mathbf{z}_w^1, .., \mathbf{z}_w^f]$.By applying STAM to a $\mathbf{TemporalAttention}$ in the baseline video diffusion models, we can achieve coherence in the motion of moving objects without additional training, and without incurring extra computational costs during inference. Note that we do not use STAM in INPM as the motion of objects in the meta video V_{meta} generated within INPM occurs at the same location. Therefore, there is no need to shift temporal attention to align the positions of moving objects.

Experiments

We evaluate the effectiveness of our method from both qualitative and quantitative perspectives. In our experiments, the default baseline is ZeroScope (Sterling 2023). ModelScope (Wang et al. 2023b) is also employed to show that our method can be applied to various video diffusion baselines. TrailBlazer (Ma, Lewis, and Kleijn 2023) and Peekaboo (Jain et al. 2024) are involved in experiments to demonstrate the control ability of our proposed method. Following TrailBlazer, 33 prompts containing different moving objects and motion patterns are employed as the evaluation dataset. For simple motions, we employed eight movement trajectories. For experiments involving complex trajectories, we utilized 17 randomly generated motion curves. Please refer to the appendix for specific prompts and trajectory parameters.

Implementation Details Our algorithm is fully implemented during the inference stage, thus it does not require any training. The hyper-parameter λ_i, λ_o are set to 1, λ_c is set to 0.05, λ_s is set to 0.5, λ_p is set to 0.8. To balance the trade-off between the size of GPU memory consumption and the semantic information retained in the attention map $\mathbf{A}$, we choose a 48×48 size for $\mathbf{A}$ when the output resolution is 384×384. We use DDIM (Song, Meng, and Ermon 2020) as our sampling method. In the experiment, T_1 in Fig.2 is set as 10, which means the SC and the STAM are employed during the first 10 timesteps; T_2 is set to 20, thus the total denoising timestep T is 30. All of our experiments are conducted on a single NVIDIA A100 GPU.

Comparisons with SOTA Methods

Qualitative Analysis. The results of the qualitative experiments with simple and complex trajectories are shown in Fig.3 and Fig.4, respectively. Our proposed Motion-Zero (+Ours) applied on other baselines (ModelScope and ZeroScope) can greatly increase the controllability of the objects' motion trajectories. In Fig.3, baselines (+Ours) refers to applying MotionZero to the baseline methods, while original baselines (+Prompt) have an extra prompt added to indicate the motion of objects: *moving from left to right.* As shown in Fig. 3, Motion-Zero correctly guides the motion of fishes by following the specific trajectories indicated by the red boxes. In contrast, both original ModelScope (+Prompt) and ZeroScope (+Prompt) fail to control the objects following the expected trajectories. Fig.3 (c) demonstrates the generated results of TrailBlazer. It shows that the fish generated by TrailBlazer does not strictly follow the red bounding boxes. Fig.3 (d) demonstrates the results of Peekaboo. It is observed that the object detaches from the box in the last few frames, and the movement direction of the fish on the left does not align with the direction of the box.

Fig.4 indicates the comparison between ours, TrailBlazer, and Peekaboo in complex trajectories setting. Our method effectively controls the motion of objects for any trajectory, e.g. a flying back rocket. On the contrary, TrailBlazer exhibits cases where moving objects are coupled with the background motion, e.g. the penguin is relatively stationary. Furthermore, the rocket and the rabbit are losing control and the motion becomes irrelevant with the bounding boxes. The penguins generated by Peekaboo lack dynamism, maintaining a single pose while only the scene moves. It fails to properly generate rockets, which might be due to the unusual nature of rotating and moving rockets. Regarding the rabbit scene, the content generated by Peekaboo exhibited frame skipping.

Quantitative Analysis. Following LOVEU-TGVE competition (Wu et al. 2023b), we use the CLIP score (Hessel

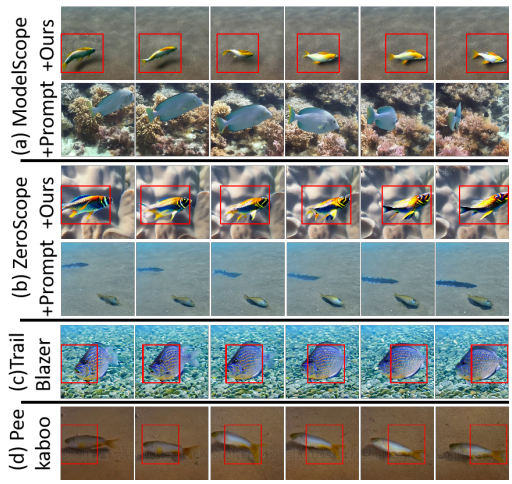

Figure 3: Quality comparison results on different methods. We take one frame from every three frames. The input prompt: *A fish is swimming in the sea.* We employed ModelScope (a) and ZeroScope (b) as our baseline models and compared the effect of incorporating additional prompts with the integration of our Motion-Zero. In addition, we conducted a comparative analysis with TrailBlazer and Peekaboo.

	Align	Cons.	Pick.	mIoU	AP50	Cov.	CD(↓)
ZeroScope	20.31	0.88	18.98	-	-	-	-
+Ours	21.96	**0.94**	**19.89**	**0.54**	**0.64**	**0.96**	**0.07**
ModelScope	20.55	0.91	18.70	-	-	-	-
+Ours	**23.54**	0.90	19.56	**0.54**	**0.64**	0.88	0.08
TrailBlazer	21.53	0.92	19.64	0.52	**0.64**	0.91	0.14
Peekaboo	20.04	0.83	18.70	0.43	0.45	0.89	0.11
TrailBlazer (c.)	**23.34**	0.92	18.81	0.50	0.61	0.91	0.13
Peekaboo (c.)	20.89	0.84	18.53	0.41	0.42	0.88	0.11
Ours (c.)	22.40	**0.95**	**19.19**	**0.55**	**0.67**	**0.97**	**0.07**

Table 1: Automatic metric on baseline methods and SOTA methods. The left half indicates the quality of the generation, while the right half demonstrates the control capability of the model. All metrics expect CD to be such that higher values(↑) indicate better performance. (c.) means the method is tested in complex trajectories. Align means Text Align, Cons. means Consistency and Pick. means PickScore.

	Align	Cons.	Pick.	mIoU	AP50	Cov.	CD(↓)
w/o INPM	20.30	0.93	19.29	0.40	0.41	0.86	0.09
w/o SC	21.27	0.92	19.10	0.18	0.17	0.71	0.22
w/o STAM	20.64	0.93	19.27	0.43	0.47	0.83	0.13
Ours	**21.96**	**0.94**	**19.89**	**0.54**	**0.64**	**0.96**	**0.07**

Table 2: The table of ablation study on different modules.

et al. 2021) to verify text-video consistency (Text Align) and inter-frame consistency (Consistency). The PickScore (Kirstain et al. 2023) is employed to predict user preferences of our model. To further evaluate the control capability of our model, we employ metrics including mIoU, AP50, Cov., and CD. mIoU stands for mean Intersection-over-Union of the detected bounding boxes and the input box on the generated video. This metric is primarily used to assess whether the model can effectively control the position and size of moving objects. We use the OWLViT-large detector (Minderer et al. 2022) to detect object boxes in the generated video. AP50 refers to average precision@50%, which is used to determine if the overlap between the detected boxes and the user-provided boxes is greater than 50%. Cov. is used to assess the probability of the detector detecting moving objects. CD represents the distance between the center of the moving object in the generated video and the center of the user-provided box, reflecting whether the generated center of the moving object is consistent with the user input.

As shown in Tab.1, the left part indicates the generation quality, and it shows that our proposed Motion-Zero framework does not compromise the baseline model's performance but increases video generation quality instead. This demonstrates that our method exploits and preserves the model's generation capability even with the zero-shot setting. In addition, the motion subject constrained by the box results in improved semantic accuracy and temporal consistency. This indicates that our proposed Motion-Zero can get a better generation quality score due to the effectiveness of consistency between the prompt and motion.

For the left in Tab.1, our method outperforms Trailblazer in terms of control performance for both simple and complex trajectories and surpasses Peekaboo in every aspect. Trailblazer performs better on the Align metric because it manually modifies cross-attention rather than deriving it through loss. However, this diminishes its performance on other metrics. Through mIoU and AP50 metrics, we can see that the size and position of the moving objects generated by our method have a closer match for the boxes provided by the users. This confirms that L_o and L_i indeed ensure that the generated objects remain within the boxes given by the users. Higher Cov. score demonstrates that the objects generated by our method are clearer. This proves that our INPM and STAM modules enable smoother motion of objects, while L_s maintains the consistency of the objects. At the same time, CD indicates that the centers of the objects generated by our method are closer to the centers of the boxes provided. This reflects that L_c can position the generated objects at the center of the box.

Ablation Study

Quantitative Analysis. We conduct ablation experiments on different components of Motion-Zero, as shown in Tab.2. It indicates that removing individual modules leads to a significant decrease in the model's performance. We observe that when SC is removed, the scores of Align do not decrease significantly. However, when STAM or INPM is removed, the scores of Align decrease noticeably, demonstrating the significant role of the STAM and INPM modules in improving the consistency between text and generated content. The SC has the greatest impact on location control as the scores drop from 0.54 to 0.18 in mIoU and from 0.64 to 0.17 in AP50. Without the SC module, there is no guarantee for objects' position and spatial consistency between frames. Additionally, the absence of STAM and INPM also results in

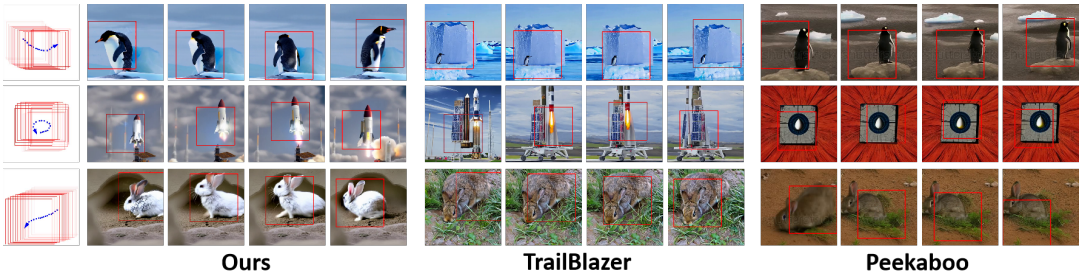

Figure 4: Quality comparison results with complex trajectories. We take one frame from every three frames. The input prompt of the first row: *A penguin standing on an iceberg.* The second row: *A rocket launching into space from a launchpad.* The third row: *A rabbit burrowing downwards into its warren.* **Zoom in for the best view.**

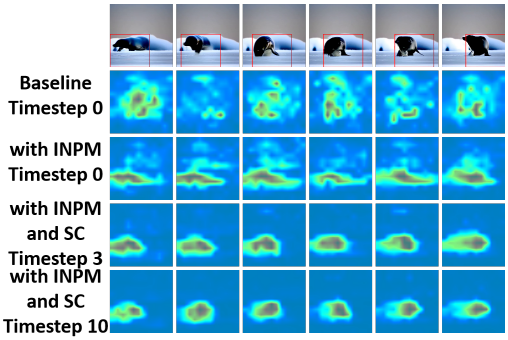

Figure 5: Attention maps with different components. Prompt: *A seal walking on the ice.*

inferior performance in temporal consistency for moving objects as the Cov. drops and it becomes difficult for the model to generate expected positioned objects as mIoU and AP50 decrease. Overall, the three modules are all important for video generation and the STAM and INPM have more impact on generation quality.

Visualization of Attention Maps. We visualize the attention maps in the cross-attention block, as shown in Fig.5. The second row displays the cross-attention map at step 0 under random initial noise conditions. It is observed that the attention values are random. As shown in the third row, when we introduce INPM, we can see that the areas with high attention response at step 0 are already approximately located at the positions of the user's box. This reflects that INPM can provide strong prior information for the position of moving objects. After optimization by SC, in the fourth and fifth rows, we can see that the areas with high response in the attention map are accurately located at the box position and are close to the position of the generated moving object. This demonstrates that SC can enforce high attention values inside the box while keeping the low response values outside the box.

User Study

We randomly pick 3 videos from 10 videos generated by each model, 15 in total to be used in our user study. Every video has 3 dimensions of evaluation rating: Appearance, Consistency, and Control, and the scores 1 to 5 indicate appearance, consistency, and control capabilities from low to high. We randomly choose 60 users from our university through an incentivized questionnaire to evaluate each dimension of the 15 videos. At the beginning of our questionnaire, we provide users with a detailed explanation of the task and the scoring guidelines. Only after confirming that the user fully understands these instructions do they proceed with the questionnaire. The results are shown in Tab.3. For baseline models, additional prompts are used to control the movements. The voters prefer our methods from all aspects compared with baselines, Trailblazer, and Peekaboo. To further validate the effectiveness of our user study, we use the Cronbach's α coefficient to assess the internal consistency (reliability) of the questionnaire. The Cronbach's α coefficient for our questionnaire is 0.901, indicating a high level of reliability. We use the Friedman Test to confirm that our results are statistically significant. The results indicate significant differences among the five methods (Friedman Test, $X^2 = 1534.29$, $p < 0.001$).

	Appearance	Consistency	Control
ZeroScope	2.97	2.80	2.27
ModelScope	2.57	2.55	2.23
Trailblazer	3.32	3.32	3.40
Peekaboo	3.96	3.84	3.55
Ours	**4.75**	**4.57**	**4.67**

Table 3: The table of user study.

Conclusion

In this paper, we proposed a novel zero-shot framework, Motion-Zero, for arbitrary object motion trajectory control that can be applied to various video diffusion models. Unlike previous methods require extensive training, our method enabled motion-control video generation without any fine-tuning of the baseline model. Our Initial Noise Prior Module was proposed to acquire prior initial noise for a high-quality generation. Moreover, Spatial Constraints and Shift Temporal Attention Mechanism respectively exploited the cross-attention dimension and temporal-attention dimension to obtain spatial control and temporal consistency. Extensive experiments demonstrated the efficacy and generalization of our proposed approach.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62472178, 62376244), Fundamental Research Funds for the Central Universities, and Shanghai Urban Digital Transformation Special Fund Project (202301027). This work is also sponsored by Natural Science Foundation of Chongqing, China (CSTB2022NSCQ-MSX0552) and the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (No.MAIS2024111).

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Cai, Y.; Chen, L.; Guan, H.; Lin, S.; Lu, C.; Wang, C.; and He, G. 2023. Explicit invariant feature induced cross-domain crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 259–267.

Cai, Y.; Chen, L.; Ma, Z.; Lu, C.; Wang, C.; and He, G. 2021. Leveraging intra-domain knowledge to strengthen cross-domain crowd counting. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Cai, Y.; Ma, Z.; Lu, C.; Wang, C.; and He, G. 2022. Global Representation Guided Adaptive Fusion Network for Stable Video Crowd Counting. *IEEE Transactions on Multimedia*, 25: 5222–5233.

Chen, L.; Cai, Y.; Lu, C.; Wang, C.; and He, G. 2023a. Video-based spatio-temporal scene graph generation with efficient self-supervision tasks. *Multimedia Tools and Applications*, 82(25): 38947–38966.

Chen, L.; Song, Y.; Cai, Y.; Lu, J.; Li, Y.; Xie, Y.; Wang, C.; and He, G. 2024. Multi-Prototype Space Learning for Commonsense-Based Scene Graph Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1129–1137.

Chen, Z.; Li, B.; Wu, S.; Jiang, K.; Ding, S.; and Zhang, W. 2023b. Content-based Unrestricted Adversarial Attack. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 51719–51733. Curran Associates, Inc.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A. A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jain, Y.; Nasery, A.; Vineet, V.; and Behl, H. 2024. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8079–8088.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*.

Li, Z.; Wang, W.; Cai, Y.; Qi, X.; Wang, P.; Zhang, D.; Song, H.; Jiang, B.; Huang, Z.; and Wang, T. 2024a. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model. *arXiv preprint arXiv:2408.02503*.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Tu, V.; et al. 2024b. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6657–6678.

Ma, W.-D. K.; Lewis, J. P.; and Kleijn, W. 2023. TrailBlazer: Trajectory Control for Diffusion-Based Video Generation.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; Wang, X.; Zhai, X.; Kipf, T.; and Houlsby, N. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. *arXiv preprint arXiv:2205.06230*.

Qiu, H.; Chen, Z.; Wang, Z.; He, Y.; Xia, M.; and Liu, Z. 2024. FreeTraj: Tuning-Free Trajectory Control in Video Diffusion Models. arXiv:2406.16863.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shi, X.; Huang, Z.; Wang, F.-Y.; Bian, W.; Li, D.; Zhang, Y.; Zhang, M.; Cheung, K. C.; See, S.; Qin, H.; et al. 2024. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, S.; Chen, J.; Li, C.; and Wang, C. 2023a. GVQA: Learning to Answer Questions about Graphs with Visualizations via Knowledge Base. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Song, S.; Li, C.; Li, D.; Chen, J.; and Wang, C. 2024. GraphDecoder: Recovering Diverse Network Graphs From Visualization Images via Attention-Aware Learning. *IEEE Transactions on Visualization and Computer Graphics*, 30(7): 3074–3088.

Song, S.; Li, C.; Sun, Y.; and Wang, C. 2023b. VividGraph: Learning to Extract and Redesign Network Graphs From Visualization Images. *IEEE Transactions on Visualization and Computer Graphics*, 29(7): 3169–3181.

Sterling, S. 2023. ZeroScope.

Sun, Z.; Chen, S.; Yao, T.; Yi, R.; Ding, S.; and Ma, L. 2024. Rethinking Open-World DeepFake Attribution with Multi-perspective Sensory Learning. *International Journal of Computer Vision*.

Sun, Z.; Chen, S.; Yao, T.; Yin, B.; Yi, R.; Ding, S.; and Ma, L. 2023. Contrastive Pseudo Learning for Open-World DeepFake Attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 20882–20892.

Wang, H.; Li, B.; Wu, S.; Shen, S.; Liu, F.; Ding, S.; and Zhou, A. 2023a. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17958–17968.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023b. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, J.; Zhang, Y.; Zou, J.; Zeng, Y.; Wei, G.; Yuan, L.; and Li, H. 2024a. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*.

Wang, W.; Li, Z.; Xu, Q.; Cai, Y.; Song, H.; Qi, Q.; Zhou, R.; Huang, Z.; Wang, T.; and Xiao, L. 2024b. QCRD: Quality-guided Contrastive Rationale Distillation for Large Language Models. *arXiv preprint arXiv:2405.13014*.

Wang, W.; Li, Z.; Xu, Q.; Li, L.; Cai, Y.; Jiang, B.; Song, H.; Hu, X.; Wang, P.; and Xiao, L. 2024c. Advancing Fine-Grained Visual Understanding with Multi-Scale Alignment in Multi-Modal Models. *arXiv preprint arXiv:2411.09691*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023c. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018*.

Wang, Z.; Yuan, Z.; Wang, X.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2023d. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. *arXiv preprint arXiv:2312.03641*.

Wu, J.; Li, X.; Zeng, Y.; Zhang, J.; Zhou, Q.; Li, Y.; Tong, Y.; and Chen, K. 2024. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023a. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Wu, J. Z.; Li, X.; Gao, D.; Dong, Z.; Bai, J.; Singh, A.; Xiang, X.; Li, Y.; Huang, Z.; Sun, Y.; He, R.; Hu, F.; Hu, J.; Huang, H.; Zhu, H.; Cheng, X.; Tang, J.; Shou, M. Z.; Keutzer, K.; and Iandola, F. 2023b. CVPR 2023 Text Guided Video Editing Competition. arXiv:2310.16003.

Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5036–5045.

Yang, L.; Li, Y.; and Chen, L. 2024. ClothPPO: A Proximal Policy Optimization Enhancing Framework for Robotic Cloth Manipulation with Observation-Aligned Action Spaces. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6895–6903. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yin, S.; Wu, C.; Liang, J.; Shi, J.; Li, H.; Ming, G.; and Duan, N. 2023. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*.

Yu, S.; Fang, J. Z.; Zheng, J.; Sigurdsson, G.; Ordonez, V.; Piramuthu, R.; and Bansal, M. 2024. Zero-shot controllable image-to-video animation via motion decomposition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3332–3341.

Zhang, D. J.; Wu, J. Z.; Liu, J.-W.; Zhao, R.; Ran, L.; Gu, Y.; Gao, D.; and Shou, M. Z. 2023. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *ArXiv*, abs/2302.05543.

Zheng, T.; Geng, C.; Jiang, P.; Wan, B.; Zhang, H.; Chen, J.; Wang, J.; and Li, B. 2024a. Non-uniform Timestep Sampling: Towards Faster Diffusion Model Training. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 7036–7045. ACM.

Zheng, T.; Jiang, P.; Wan, B.; Zhang, H.; Chen, J.; Wang, J.; and Li, B. 2024b. Beta-Tuned Timestep Diffusion Model. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part III*, volume 15061 of *Lecture Notes in Computer Science*, 114–130. Springer.