ZeroTrail: Zero-Shot Trajectory Control Framework for Video Diffusion Models

Yi Lu[1,2], Minyi Lei[3], Bozheng Li[1,5], Jiawang Cao[1], Wenbo Zhu[1]

[1]Opus AI Research [2]University of Toronto [3]McMaster University [4]Brown University

Abstract

Recent large-scale text-to-video diffusion models demonstrated striking capability in synthesizing realistic clips, yet achieving effective control over objects' motion trajectories remains a challenging task. Prior attempts either required model-specific architecture modifications and costly training, or relied on zero-shot attention masks with limited effectiveness, or stacked multiple rounds of test-time latent optimization, achieving modest controllability at high computational cost and long running time. In this study, we introduce ZeroTrail, a novel zero-shot, tuning-free framework that equips video diffusion models with superior trajectory controllability without requiring alteration to the model architecture or incurring excessive inference time in multiple rounds. Our framework is composed of two key components: (i) a Trajectory Prior Injection Module (TPIM), which embeds the desired path into latent features through a single round of test-time training, and (ii) a Selective Attention Guidance Module (SAGM), which amplifies or attenuates cross-frame attention dynamically to reinforce the injected prior and preserve spatiotemporal coherence. Our framework is modular and requires no architectural modification, allowing it to be adapted to video diffusion models without requiring alterations to model architectures or additional training. Extensive experiments demonstrate that our framework consistently outperforms existing methods, accurately steering objects along complex trajectories while maintaining video diffusion models' ability to generate high-quality, consistent videos.

1 Introduction

Current text-to-video diffusion models are capable of generating high-quality videos with smooth motions following user-specified textual prompts Guo et al. [2023], Ho et al. [2022a]. Nevertheless, solely relying on text prompts suffers from limited ability to effectively control spatial layouts and motion trajectories of objects in generated videos, which remain crucial for generated videos to convey meaningful stories and maintain high expressiveness Hu and Xu [2023].

To bridge this gap, a variety of solutions have been explored. Drawing inspiration from the success of ControlNet Zhang et al. [2023] in text-to-image tasks, prior works [Wang et al., 2023a, Hu and Xu, 2023] leverage dense control signals, like skeleton tracks or edge maps, for motion guidance. However, models need expensive fine-tuning to be able to follow such dense signals. Additionally, they need to be annotated frame-wise with intensive labour cost, posing a significant barrier for general users and casual content creation, while sparse signals such as drag trajectories or bounding boxes can be annotated once and then interpolated, requiring minimal human effort. Another lines of work Wang et al. [2023b,a] fine-tune video diffusion models using large-scale datasets of video-trajectory annotation pairs. While the resulting models exhibit reasonable controllability with sparse trajectory signals, achieving this demands extensive computational resources, significant training time, vast amounts of annotated data, and often requires model-specific architectural modifications.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NextVid.

Figure 1: ZeroTrail enables diverse video diffusion foundation models to modulate moving objects' trajectories without the need for fine-tuning. By specifying the object name in the textual prompt and a sequence of bounding boxes as a control signal, the object's trajectory can be intuitively manipulated in the generated video by the end user. Zoom in for better details.

Empirical experiments in text-to-image generation Mao et al. [2023] demonstrate that the initial noisy latent plays a significant role in determining the spatial composition, while Epstein et al. [2023], Phung et al. [2023] shows that regulating the denoising process during test time could also help to customize the generated image's landscape. Considering text-to-video generation models Wang et al. [2023c], Chen et al. [2023a, 2024a] are pre-trained on large-scale datasets and demonstrated strong prior knowledge of dynamic motions and object movements, this trait could be exploited for generating controllable, smooth motions. By jointly utilizing both characteristics, effective trajectory control of moving objects can be achieved under zero-shot settings for video diffusion models.

Inspired by recent advancements in subject position control for image generation, we propose ZeroTrail, a zero-shot trajectory control framework for video diffusion models to further bridge the gap between effective trajectory enforcement and the need for a fine-tuning-free, cost-efficient method. ZeroTrail bridges this gap by introducing a lightweight yet effective recipe composed of two key components. The Trajectory Prior Injection Module (TPIM) optimizes latent representations at test-time using a cross-attention contrastive loss, encoding motion trajectory priors into noisy latents, leveraging the strong subject localization controllability of diffusion U-Nets' text-latent cross-attention layers Xie et al. [2023]. Concurrently, the Selective Attention Guidance Module (SAGM) adaptively enhances or suppresses frame-wise and frame-token cross-attention activations based on user-specified trajectories, thereby reinforcing spatial-temporal alignment and enabling fine-grained trajectory control. By orchestrating both modules during early denoising stages, ZeroTrail achieved high-quality video generation with effective trajectory control, eliminating the need for model fine-tuning. Our contribution is summarized as follows:

- We proposed ZeroTrail, an effective tuning-free trajectory control framework for pre-trained text-to-video diffusion models. It is modular and adaptable to video diffusion models out of the box without the need for fine-tuning or architecture modifications.
- Contrast to previous works, we bring the advantage of both latent optimization and attention guidance, forming a novel joint latent–trajectory alignment paradigm with superior trajectory controllability.
- Our framework demonstrates superior trajectory controllability over existing methods through extensive experiments, offering pre-trained video diffusion models with effective trajectory control capability.

2 Related Work

Video Diffusion Models Diffusion Models have demonstrated their revolutionary capability in producing high-quality images and video samples through recent groundbreaking advancements Rombach et al. [2021], Podell et al. [2023], Qin et al. [2024], Ho et al. [2022b]. The foundational DDPM Ho et al. [2020] paved the road for high-quality image generation, while Latent Diffusion Models Rombach et al. [2021] are introduced to achieve superior efficiency while maintaining satisfiable generation quality. VDM Ho et al. [2022b] expands diffusion models' capability into the field of video generation, and LVDM He et al. [2022] further proposed an efficient latent video generation approach. Blattmann et al. [2023a], Guo et al. [2023] empower pre-trained text-to-image generation models with video generation capability through inserting temporal transformer layers, while Make-a-video Singer et al. [2022] introduced the 3D U-Net architecture, which decouples attention layers into spatial, temporal, and frame-token cross-attention layers. On top of the 3D-U-Net

architecture, VideoCrafter Chen et al. [2024a] and SVD Blattmann et al. [2023b] scale up latent video diffusion models to large datasets and achieved superior video generation capability. In this work, the current state-of-the-art 3D U-Net-based video diffusion model Huang et al. [2023], Liu et al. [2023], VideoCrafter2.0 Chen et al. [2024a], referred to as "VideoCrafter" in the remaining paper, is chosen for framework implementation and comprehensive evaluations to demonstrate our framework's capability.

Trajectory Controllable Video Generation Motion control is essential for generating expressive and coherent videos. Early approaches such as Tune-A-Video Wu et al. [2022] and MotionDirector Zhao et al. [2023] achieve motion transfer by learning subject movements from existing videos. While these methods demonstrate strong generality across diverse domains, they require motion-specific training and can only learn motions from existing videos, limiting their flexibility and scalability. Chen et al. [2023b], Wang et al. [2023a] employ dense signals (e.g., depth maps, keypoints) to condition motion generation. However, these signals still correspond to non-editable motions extracted from existing footage, offering little room for customization. To tackle this problem, Wang et al. [2023b], Yin et al. [2023], Deng et al. [2023], Wang et al. [2024] leverage large-scale datasets of video-sparse motion signal (e.g., trajectories or bounding boxes) pairs as fine-tuning datasets to achieve sparse control signal interpretation through specialized training recipes or model-specific adapters. VideoComposer Wang et al. [2023a] adopts a two-stage curriculum to incorporate control and temporal consistency, and MotionCtrl Wang et al. [2023b] introduces adapters for camera and trajectory control. MotionBooth Wu et al. [2024] enables object-level animation after personalized training, while Motion-I2V Shi et al. [2024] enables drag-based trajectory control via staged training. While these methods show promising results, they rely on dedicated training pipelines and are often tightly coupled with specific architectures, which limits their generalization and demands considerable computational resources for training.

Several works have attempted to achieve trajectory control under zero-shot settings. SG-I2V Namekata et al. [2024] achieves motion control through latent optimization and high-frequency replacement while being constrained to the image-to-video generation scope. FreeTraj Qiu et al. [2024] blends the initial noise via bounding-box-driven patching to guide object motion with limited effectiveness, as its controllability is mainly contributed by attention-level guidance. MotionZero Chen et al. [2024b] applied latent optimization alone, while requiring two-pass video generation and DDIM inversion, making it computationally intensive. Other approaches like Trailblazer Ma et al. [2023] and Peekaboo Jain et al. [2023] incorporate attention control through masked attention maps alone but fall short in trajectory alignment, suffer from generating still videos or videos with deviated motions. In contrast, ZeroTrail blends the advantages of both latent optimization and attention guidance. By introducing a joint guidance strategy, our framework demonstrates superior trajectory adherence with reduced inference-time computational cost.

3 Preliminaries

Video Diffusion Model. Designed for generating high-quality and diverse videos, video diffusion models involve a tractable forward diffusion process to add noise to the Gaussian video latent $x_0 \sim p(x_0)$ and learn a denoising model to revert this process in inference. To reduce computational complexity and exploit inter-frame redundancy, the Latent Diffusion Model (LDM) Rombach et al. [2021] is widely adopted, which operates diffusion and denoising procedures in a latent space.

Coupled with a Variational Auto Encoder Kingma and Welling [2019] composed of an encoder $\mathcal{E}$ responsible for projecting original video frames x_0 from the pixel space $\mathbb{R}^{3\times F\times H\times W}$ into latent space $\mathbb{R}^{4\times F\times H'\times W'}$, and a decoder $\mathcal{D}$ reconstructing the video frames x_0 from latent representations z_0 back to the pixel space. The forward noise-addition process contains T timesteps, gradually adding noise to data sample x_0, producing a noisy latent x_t through the parameterization technique:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\Big)\,,$$
$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I\big)\,.$$

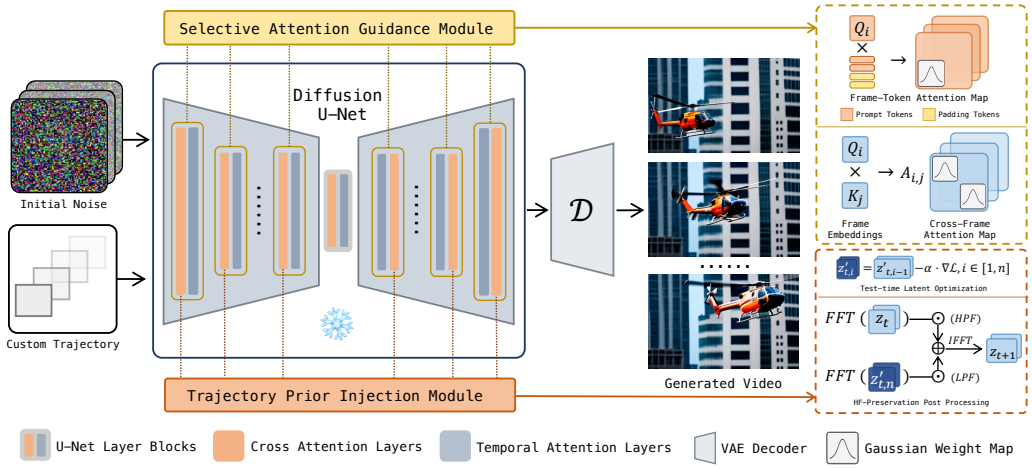

Figure 2: Overview of ZeroTrail Pipeline. Given trajectory and prompt, the latent is guided through both TPIM's test-time optimization and SAGM's spatial-temporal attention guidance. For each step, the optimized latent went through high-frequency preservation postprocessing. Both modules operate at early stages of DDIM denoising in inference time with the Diffusion U-Net being frozen, making the pipeline tuning-free and plug-and-play for Video Diffusion Models.

where q refers to the noise-addition schedule, where t is the timestep, the predefined variance schedule is denoted as β_t, $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The latent got denoised in the reverse denoising process, obtaining a less noisy latent x_{t-1} from the noisy input x_t as

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right).$$

Here $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are determined through a noise prediction network, generally a U-Net Ronneberger et al. [2015] denoted as $\epsilon_\theta(x_t, t)$. The denoising model is supervised by the objective function $\min_\theta \mathbf{E}_{t,x_0,\epsilon} \left\|\epsilon - \epsilon_\theta(x_t, t)\right\|_2^2$, , where ϵ represents the sampled ground-truth noise and θ represents the learnable network parameters. With the denoising model being trained under the above objective, during inference, the latent noise at each timestep is predicted by the denoising model and gradually generates high-quality, consistent video frames. Conventionally, mainstream LVDMs Wang et al. [2023c], Chen et al. [2024a] employ 3D-convolution modules, spatial attention layers, temporal attention layers, and conditioning attention layers.

Text Conditioning. To achieve textual control of the generated video scenes, text conditions are employed in the denoising model $z_{t-1} = f_\theta(z_t|c)$ where textual query is denoted as condition c, latents are denoted as z and the denoising model f_θ is generally a denoising 3D U-Net. During inference, textual prompts are encoded using CLIP Radford et al. [2021] into embeddings and are then cross-attended with each frame's latents at conditioning attention layers in denoising models. Across the iterative denoising process, the input noise is gradually denoised and aligned toward the desired text prompt.

Trajectory Control. The objective of video motion trajectory control is to guide the motion trajectory of objects in generated videos accurately. The optimization objective is formulated as $\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,c,\mathcal{B},\epsilon\sim\mathcal{N}(0,I),t}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c, \mathcal{B})\right\|_2^2\right]$. Specifically, users provide a text condition c along with a sequence of rectangular bounding boxes $\mathcal{B} = \{((x_{1,f}, y_{1,f}), (x_{2,f}, y_{2,f}))\}^{N_f}$, where $(x_{1,f}, y_{1,f})$ and $(x_{2,f}, y_{2,f})$ denote the top-left and bottom-right coordinates of the fth frame's bounding box respectively. Here, N_f is the total number of frames. The bounding box set $\mathcal{B}$ encodes the desired object trajectory as a sequence of bounding boxes across the video frames.

4 Method

In this section, we introduce the ZeroTrail framework, which operates completely in the early denoising stage under zero-shot settings, eliminating the need for costly fine-tuning to control object motion

trajectories in generated videos. The section is organized as follows: we first describe the Trajectory Prior Injection Module (TPIM) and its constraints and loss for effectively injecting the trajectory prior into the latents through test-time optimization. Then, we explain the Selective Attention Guidance Module (SAGM), which further enforces trajectory alignment and spatial consistency.

4.1 Trajectory Prior Injection Module

Inspired by empirical studies Ho et al. [2020] and Cao et al. [2023], Jain et al. [2023], the initial noisy latent has a significant effect in determining the generated video frames' scene construction, while frame-token cross-attention layers are mainly responsible for bridging the textual prompt and corresponding objects in generated video frames. The Trajectory Prior Injection Module (TPIM) is designed to exploit this property by injecting user-specified trajectories into the noisy latent during early denoising steps through test-time latent optimization based on a contrastive trajectory loss. Following literature Namekata et al. [2024], we leverage the high-frequency preservation technique to refine visual quality through latent-level high-frequency replacement. By injecting trajectory guidance into noisy frame latents, TPIM steers the denoising process naturally, providing strong priors for effective trajectory control.

4.1.1 Contrastive Trajectory Prior Loss

Formally, let $\mathbf{z}^{(t)} \in \mathbb{R}^{C\times H\times W}$ denote the latent at diffusion step t and let $\mathcal{B}_f = [x_0, y_0, x_1, y_1]$ be the user-drawn bounding box for frame $f \in \{1, \dots, N\}$. Foreground and background indicator functions are defined as

$$\mathbb{1}_{\mathrm{fg}_f}(x,y) = \begin{cases} 1, & (x,y) \in \mathcal{B}_f \\ 0, & \text{otherwise} \end{cases}, \mathbb{1}_{\mathrm{bg}_f} = 1 - \mathbb{1}_{\mathrm{fg}_f}.$$

Frame–token cross-attention scores are denoted by $A_f(x, y, t)$. To encourage higher attention along the trajectory while suppressing it elsewhere, a contrastive loss is adopted. Positive and negative pools are constructed by selecting the top-K_{in} and top-K_{out} scores inside and outside the box, respectively:

$$\mathcal{P}_f = \mathrm{TopK}_{\text{in}}\big\{A_f(x,y,t) \mid \mathbb{1}_{\mathrm{fg}_f}(x,y) = 1\big\},$$
$$\mathcal{N}_f = \mathrm{TopK}_{\text{out}}\big\{A_f(x,y,t) \mid \mathbb{1}_{\mathrm{fg}_f}(x,y) = 0\big\}.$$

Their averages are $\mu_f^+ = \frac{1}{K_{\text{in}}}\sum_{a\in\mathcal{P}_f} a$ and $\mu_f^- = \frac{1}{K_{\text{out}}}\sum_{a\in\mathcal{N}_f} a$. With μ_f^+ and μ_f^- denoting the average values of these two sets and τ a temperature hyper-parameter, the InfoNCE loss for the frame f is

$$\mathcal{L}_f = -\log \frac{\exp\big(\mu_f^+/\tau\big)}{\exp\big(\mu_f^+/\tau\big) + \exp\big(\mu_f^-/\tau\big)}$$

And the overall objective for all frames is

$$\mathcal{L}_{\text{ctr}} = \frac{1}{N}\sum_{f=1}^{N} \mathcal{L}_f.$$

Starting from the first denoising step, the latent is updated for five iterations by gradient descent according to $\mathbf{z}^{(t,s+1)} = \mathbf{z}^{(t,s)} - \eta\,\nabla_{\mathbf{z}}\mathcal{L}_{\text{ctr}}$, where $s \in [0, 5]$, η is the learning rate and $\nabla_{\mathbf{z}}\mathcal{L}_{\text{ctr}}$ is the gradient of the loss function described above.

4.1.2 FFT-based Latent Visual Artifact Patching

To mitigate visual artifacts introduced by latent updates, the amplitude of the high-frequency Fourier components of $\mathbf{z}^{(t,i)}$ is optionally replaced by that of the original $\mathbf{z}^{(t,i-1)}$ following the equation below:

$$\mathcal{F}^{\text{low}}_{\mathbf{z}^{(t,i)}} = \mathrm{FFT}_{2\mathrm{D}}\big(\mathbf{z}^{(t,i)}\big) \odot \mathbf{H}_\gamma,$$
$$\mathcal{F}^{\text{high}}_{\mathbf{z}^{(t,i-1)}} = \mathrm{FFT}_{2\mathrm{D}}\big(\mathbf{z}^{(t,i-1)}\big) \odot \big(\mathbf{1} - \mathbf{H}_\gamma\big),$$
$$\tilde{\mathbf{z}}^{(t)} = \mathrm{IFFT}_{2\mathrm{D}}\big(\mathcal{F}^{\text{low}}_{\mathbf{z}^{(t,i)}} + \mathcal{F}^{\text{high}}_{\mathbf{z}^{(t,i-1)}}\big)$$

Where $\text{FFT}_{2\text{D}}$ / $\text{IFFT}_{2\text{D}}$ denote the 2-D (inverse) Fast Fourier Transform applied frame-wise, and $\mathbf{H}_\gamma$ be the frequency response of a 2-D low-pass Butterworth filter with cut-off frequency γ. The latent postprocessing is done after all optimization epochs and is involved in the first $i \in [1, 5]$ denoising steps, preserving the target motion encoded in the low-frequency part of $\mathbf{z}^{(t,i)}$ while reinstating the high-frequency details from $\mathbf{z}^{(t,i-1)}$.

4.2 Selective Attention Guidance Module

SAGM operates on the five early iterations and modifies both token–frame and frame–frame cross-attention layers in the denoising U-Net Rombach et al. [2021] module.

4.2.1 Spatial Frame-Token cross-attention

Given queries $Q \in \mathbb{R}^{N_F \times d_h \times d}$ from spatial tokens and keys $K \in \mathbb{R}^{N_F \times |W| \times d}$ from prompt tokens, the cross attention map is computed as $A_s = \text{Softmax}(\frac{Q_s K_s^\top}{\sqrt{d}}) \in \mathbb{R}^{N_F \times d_h \times |W|}$, where $d_h = w \times h$ is defined by the spatial resolution at specific attention layers. d refers to the feature dimension of both keys and queries, $|W|$ is defined as 77 for CLIP Radford et al. [2021] text embeddings. For simplicity, the batch size and attention head dimensions are omitted.

The denoising path is then guided through the adaptive editing of the frame-token cross attention layers for the attention maps $A_s^i \in \mathbb{R}^{N_F \times d_h}$ related to specific prompt word tokens, whose indices are represented as $i \in [0, |W|]$. Given the bounding box $\mathcal{B}$, the frame-token cross attention guidance is defined as:

$$W_s(x, y) = \begin{cases} 1 - c_w, & (x, y) \in \mathcal{B}, \\ c_w, & \text{otherwise} \end{cases}$$

$$S_s(x, y) = \begin{cases} C \cdot g(x, y), & (x, y) \in \mathcal{B} \\ 0, & \text{otherwise}, \end{cases}$$

where (x, y) are indices of the attention map and $g(\cdot, \cdot)$ refers to Gaussian map with size ($\sigma_x = b_w/2, \sigma_y = b_h/2$). C is dynamically scaled according to the current attention map size $a_w \times a_h$ and number of prompt word tokens $|W_p|$: $C = \frac{c_s}{a_w \times a_h \times |W_p|}$. Given subject prompt token indices, each cross-attention score at location $(x, y) \in A_s$ is adjusted as $A_s^i(x, y) = A_s(x, y) \odot W_s(x, y) + S_s(x, y)$, where $\odot$ refers to element-wise product (Hadamard Product), scaling the cross attention map with weight matrix W.

Temporal Frame-wise cross-attention The temporal frame-wise attention guidance is introduced to better regulate temporal correlation and consistency. Different from frame-token cross-attention, the temporal attention map is obtained through $A_t = \text{Softmax}(\frac{Q_t K_t^\top}{\sqrt{d}}) \in \mathbb{R}^{d_h \times N_F \times N_F}$.

The spatial dimension of the attention map is denoted as d_h, $Q_t \in \mathbb{R}^{d_h \times N_F \times d}$, $K_t \in \mathbb{R}^{d_h \times N_F \times d}$. The cross-attention score at location $(x, y) \in A_t$ is then adjusted as $A_t^i(x, y) = A_t(x, y) \odot W_t(x, y)$, where $W_t(\cdot, \cdot)$ is defined as the same as $W_s(\cdot, \cdot)$.

5 Experiments

Overview. The evaluation is conducted both quantitatively and qualitatively to assess the performance and effectiveness of our framework thoroughly. VideoCrafter Chen et al. [2024a] is selected as the base video diffusion model in our experiments. To evaluate ZeroTrail's trajectory controllability, recent works including FreeTraj Qiu et al. [2024], TrailBlazer Ma et al. [2023], and Peekaboo Jain et al. [2023] are compared. MotionZero Chen et al. [2024b] is excluded from comparison due to the absence of the official implementation at the time of this study.

Implementation Details. All experiments are conducted on a single NVIDIA A100 GPU. We adopt DDIM sampling Song et al. [2020] with 50 steps for video generation. During the first 5 denoising

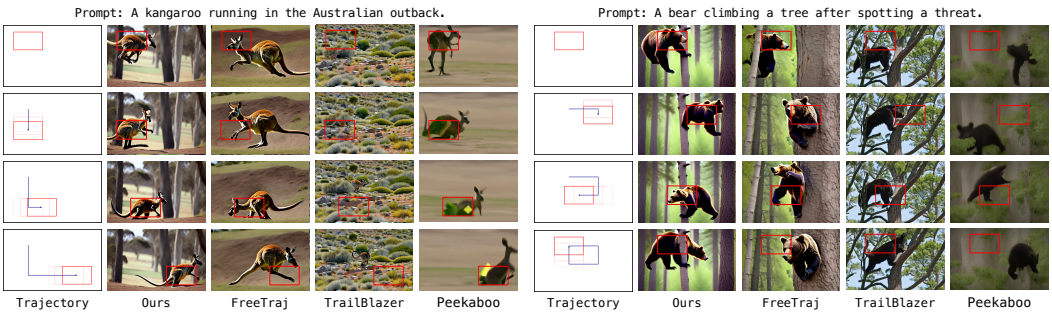

Figure 3: Qualitative comparison result on complex trajectory control cases. ZeroTrail demonstrates better trajectory guidance capability while retaining high-quality scenes compared to other SOTAs.

steps, both TPIM and SAGM modules are jointly applied. For **TPIM**, the latent is optimized for 5 epochs during denoising with an initial learning rate of 3.5×10^{-2} and linearly decaying to zero. In the contrastive loss, the temperature is fixed as $\tau = 0.07$ and the sampling ratios are selected as $p_{\text{in}} = 0.25$ (inside the guidance region) and $p_{\text{out}} = 0.1$ (outside). High-frequency Fourier amplitudes of the optimized latent are partially replaced by those of the original latent to suppress artifacts. Specifically, we resample 25% of the spatial spectrum ($d_s = 0.25$) and 10% of the temporal spectrum ($d_t = 0.10$). For **SAGM**, it is similarly activated during the first 5 denoising steps. c_w is set to 0.01 for both token-frame and temporal cross attention layers while c_s is set to 0.25. The remaining steps proceed without intervention.

Following prior works, a standard set of 33 prompts covering diverse subjects and motions is adapted for evaluation, with 8 simple motion trajectories and 15 complex ones featuring diverse and compound movement patterns. All visualizations throughout the paper are uniformly sampled from the 16 generated frames. The complete list of prompts and trajectories are provided in the appendix.

5.1 Main Results.

5.1.1 Qualitative Analysis.

We demonstrate qualitative experiment results with simple and complex trajectories in Fig.3 and Fig.5. The attention guidance effect is visualized in Fig. 4. As shown in the graph, our framework enables effective trajectory control compared to the base model, reflected by the evidently-guided activations of cross-attention layers.

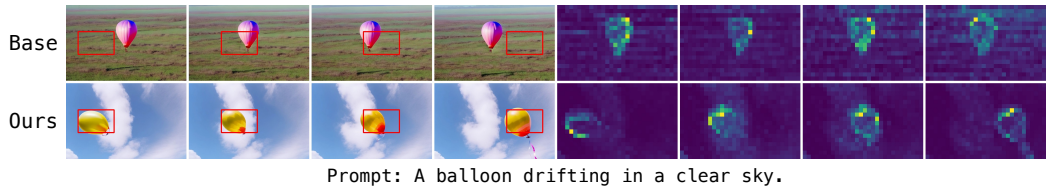

Figure 4: Attention map visualization of the base model and our framework's result. The specified trajectory is visualized as the red bounding box. Zoom in for a better view.

Fig. 5 demonstrates the result of cross comparison on a simple trajectory: moving from bottom-right to top-left. Our framework precisely guides the parrot flying towards the specified trajectory, while in FreeTraj's and Trailblazer's results, the parrot is anchored in the central position throughout the video. For Peekaboo, the motion roughly aligns with the bounding box's direction, while its visuals are noticeably inferior. In terms of complex trajectory setting, our framework significantly outperforms all existing methods as well. From Fig. 3, our method effectively controls the objects' motion for any trajectory. On the contrary, FreeTraj failed to align the object with the bounding box, while Trailblazer's results are mostly stationary or random. Peekaboo demonstrated some controllability in the left-hand side L-shape trajectory case, while its visuals demonstrated strong artifacts.

Figure 5: Qualitative comparison result on simple trajectories. ZeroTrail shows better trajectory guidance capability while retaining high-quality scenes compared to others.

Trajectory Control Metrics				**Visual Quality Metrics**			
Method	**mIoU ↑**	**AP@50 ↑**	**Central Dist. ↓**	**Method**	**Align ↑**	**Consistency ↑**	**PickScore ↑**
ZeroTrail	**33.25**	**0.1447**	**7.61%**	ZeroTrail	**0.3102**	0.9620	**20.792**
FreeTraj	24.38	0.0469	11.52%	FreeTraj	0.3072	0.9617	20.648
TrailBlazer	18.65	0.0717	17.10%	TrailBlazer	0.3042	**0.9698**	20.539
Peekaboo	13.31	0.0628	18.33%	Peekaboo	0.2939	0.9338	19.322

Table 1: Quantitative comparison between ZeroTrail and baselines. The best results are in bold. All metrics except Central Distance expect higher values for better performance.

5.1.2 Quantitative Analysis.

To evaluate trajectory controllability, mIoU, AP50, Center Distance are chosen as main evaluation metrics. mIoU evaluates the intersection-over-union between detected bounding boxes from video frames and user-specified ones. AP50 evaluates the percentage of detected bounding boxes that overlap with the user-specified ones over 50%. Center Distance evaluates the distance between the detected bounding box's central point to that of the ground truth as percentage values, evaluating the trajectory deviations. OWLViT-Large Minderer et al. [2022] is employed as the object detector to generate bounding boxes, supporting the computation of the above metrics. For visual quality, we applied the CLIP score Hessel et al. [2021] to verify the text-video alignment (Alignment) and inter-frame consistency (Consistency). PickScore Kirstain et al. [2023] is applied to evaluate user preferences over videos following previous works Wu et al. [2023a,b], Chen et al. [2024b].

As shown in the Tab. 1, our framework outperforms existing methods in terms of trajectory control effectiveness in every aspect. Through mIoU and AP50 metrics, it is clear that the size and absolute position of the generated moving objects are significantly closer-matched with the user-specified bounding box sequences, i.e., sparse trajectory signals, demonstrating the fact that our joint TPIM-SAGM attention guidance optimization design ensures the generated objects of interest are kept inside the specified bounding boxes. In addition to demonstrating leading trajectory control capabilities, our framework maintains high visual quality, as shown in Tab. 1. Despite Trailblazer achieving slightly better performance in the Consistency metric, this may be attributed to its limited ability to generate realistic motion when handling complex trajectories, preserving internal frame similarity.

We further compared our framework's average running time with baselines in Tab. 2. As Trailblazer and Peekaboo utilize ZeroScope Khachatryan et al. [2023] and FreeTraj applies VideoCrafter Chen et al. [2024a] as the base model separately, we report the increased runtime percentage for fair comparison. ZeroTrail introduces a modest inference time increase, which is a reasonable trade-off for improved motion controllability. Since MotionZero lacks a public implementation, we conservatively estimate its runtime as roughly twice that of the base model, based on its multi-stage pipeline involving meta-video generation, DDIM inversion, and final video synthesis.

5.2 Ablation Study.

Below, we report major ablation experiment results. More ablations on hyperparameters are attached in the appendix.

Impact of Module incorporation. Tab. 3 demonstrates the result of ablation experiments on different modules of ZeroTrail. According to the table, both TPIM and SAGM play crucial roles in ensuring superior trajectory guidance capability. The SAGM has the most significant impact on aligning the moving object towards the bounding box center, as removing SAGM would cause the

Method	ZT	FT	TB	PB	MZ†
Extra Time	32.1%	28.2%	24.3%	20.5%	100%

Table 2: Running time comparison between ZeroTrail (ZT) and baseline methods. "FT", "TB", "PB", and "MZ" represent FreeTraj, TrailBlazer, Peekaboo, and MotionZero. Results are reported as the percentage of additional inference time. MotionZero's runtime is approximated based on its design.

Central Distance metric to increase by 7.82%. Additionally, both TPIM and SAGM have a significant impact on trajectory control, as disabling TPIM, the mIoU scores drop from 0.3325 to 0.2719 and from 0.1447 to 0.0687 in AP@50, while removing SAGM causes the mIoU to decrease to 0.2139 and AP@50 drops to 0.0580. Overall, both modules are crucial for achieving effective trajectory control in our framework.

Module Setting	mIoU ↑	AP@50 ↑	Center Dist. (%) ↓
TPIM + SAGM	**0.3325**	**0.1447**	**7.61%**
w/o TPIM	0.2719	0.0687	8.98%
w/o SAGM	0.2139	0.0580	15.43%

Table 3: Ablation study on the impact of TPIM and SAGM modules in our framework. The best-performing metrics are stressed in bold. Apart from Center Distance, all metrics expect higher values for better performance.

Impact of FFT-based Latent Postprocessing. We further conducted an ablation experiment on the effectiveness of the FFT-based latent visual postprocessing module. As shown in Tab. 4, incorporating the FFT-based latent postprocessing improved all three visual metrics, demonstrating that postprocessing optimized latent plays a crucial role in improving the generated video's visual quality.

Module Setting	Align ↑	Consistency ↑	PickScore ↑
with FFT Fix	**0.3102**	**0.9620**	**20.792**
w/o FFT Fix	0.3094	0.9606	20.768

Table 4: Ablation study on the visual quality impact introduced by the FFT-postprocessing Module (FFT Fix). The best-performing metrics are stressed in bold. All metrics expect higher values for better performance.

6 Conclusion

In this work, we introduce ZeroTrail, a novel zero-shot trajectory control framework that achieves arbitrary object trajectory control that is applicable to diverse video diffusion models. Different from existing works, which either require computationally intensive fine-tuning or suffer from suboptimal motion controllability, our trajectory control framework is effective, plug-and-play, and does not require extra fine-tuning of the foundational model. By guiding the denoising process through the joint application of both the Trajectory Prior Injection Module and the Selective Attention Guidance Module, our framework achieved superior trajectory guidance, excels previous works in terms of efficacy and generality, which is demonstrated by extensive experiments.

References

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Y. Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ArXiv*, abs/2307.04725, 2023. URL `https://api.semanticscholar.org/CorpusID:259501509`.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen

video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022a. URL https://api.semanticscholar.org/CorpusID:252715883.

Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *ArXiv*, abs/2307.14073, 2023. URL https://api.semanticscholar.org/CorpusID:260164708.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. URL https://api.semanticscholar.org/CorpusID:256827727.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *ArXiv*, abs/2306.02018, 2023a. URL https://api.semanticscholar.org/CorpusID:259075720.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *ACM SIGGRAPH 2024 Conference Papers*, 2023b. URL https://api.semanticscholar.org/CorpusID:265696111.

Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. In *European Conference on Computer Vision*, 2023. URL https://api.semanticscholar.org/CorpusID:266210446.

Dave Epstein, A. Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *ArXiv*, abs/2306.00986, 2023. URL https://api.semanticscholar.org/CorpusID:258999106.

Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7932–7942, 2023. URL https://api.semanticscholar.org/CorpusID:259108247.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *ArXiv*, abs/2308.06571, 2023c. URL https://api.semanticscholar.org/CorpusID:260887737.

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, 2023a.

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, 2024a.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7418–7427, 2023. URL https://api.semanticscholar.org/CorpusID:259991581.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. URL https://api.semanticscholar.org/CorpusID:245335280.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhen fei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators. *ArXiv*, abs/2410.18072, 2024. URL https://api.semanticscholar.org/CorpusID:273532769.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *ArXiv*, abs/2204.03458, 2022b. URL https://api.semanticscholar.org/CorpusID:248006185.

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL https://api.semanticscholar.org/CorpusID:219955663.

Yin-Yin He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *ArXiv*, abs/2211.13221, 2022. URL https://api.semanticscholar.org/CorpusID:253802030.

A. 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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023a. URL https://api.semanticscholar.org/CorpusID:258187553.

Uriel Singer, Adam Polyak, Thomas Hayes, Xiaoyue Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *ArXiv*, abs/2209.14792, 2022. URL https://api.semanticscholar.org/CorpusID:252595919.

A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023b. URL https://api.semanticscholar.org/CorpusID:265312551.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2023. URL https://api.semanticscholar.org/CorpusID:265506207.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149, 2023. URL https://api.semanticscholar.org/CorpusID:264172222.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7589–7599, 2022. URL https://api.semanticscholar.org/CorpusID:254974187.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023.

Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang-Jin Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *ArXiv*, abs/2305.13840, 2023b. URL https://api.semanticscholar.org/CorpusID:258841645.

Sheng-Siang Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *ArXiv*, abs/2308.08089, 2023. URL https://api.semanticscholar.org/CorpusID:260925229.

Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. Dragvideo: Interactive drag-style video editing. *ArXiv*, abs/2312.02216, 2023. URL https://api.semanticscholar.org/CorpusID:265659457.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *ArXiv*, abs/2402.01566, 2024. URL https://api.semanticscholar.org/CorpusID:267406297.

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *ArXiv*, abs/2406.17758, 2024. URL https://api.semanticscholar.org/CorpusID:270710803.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Y. Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Da, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *ACM SIGGRAPH 2024 Conference Papers*, 2024. URL https://api.semanticscholar.org/CorpusID:267311454.

Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *ArXiv*, abs/2411.04989, 2024. URL https://api.semanticscholar.org/CorpusID:273878003.

Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *ArXiv*, abs/2406.16863, 2024. URL https://api.semanticscholar.org/CorpusID:270702952.

Changgu Chen, Junwei Shu, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. Motion-zero: Zero-shot moving object control framework for diffusion-based video generation. *ArXiv*, abs/2401.10150, 2024b. URL https://api.semanticscholar.org/CorpusID:267035012.

Wan-Duo Kurt Ma, J. P. Lewis, and W. Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. *SIGGRAPH Asia 2024 Conference Papers*, 2023. URL https://api.semanticscholar.org/CorpusID:266725649.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Singh Behl. Peekaboo: Interactive video generation via masked-diffusion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8079–8088, 2023. URL https://api.semanticscholar.org/CorpusID:266174002.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12:307–392, 2019. URL https://api.semanticscholar.org/CorpusID:174802445.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. URL https://api.semanticscholar.org/CorpusID:3719281.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Ming 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. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22503–22513, 2023. URL https://api.semanticscholar.org/CorpusID:258179432.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. URL https://api.semanticscholar.org/CorpusID:222140788.

Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *ArXiv*, abs/2205.06230, 2022. URL https://api.semanticscholar.org/CorpusID:248721818.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718, 2021. URL https://api.semanticscholar.org/CorpusID:233296711.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *ArXiv*, abs/2305.01569, 2023. URL https://api.semanticscholar.org/CorpusID:258437096.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023a.

Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, Rui He, Feng Hu, Junhua Hu, Hai Huang, Hanyu Zhu, Xu Cheng, Jie Tang, Mike Zheng Shou, Kurt Keutzer, and Forrest Iandola. Cvpr 2023 text guided video editing competition, 2023b.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15908–15918, 2023. URL https://api.semanticscholar.org/CorpusID:257687280.

A Appendix

A.1 Overview

In this section, we supply additional ablation experiments on the implementation of the ZeroTrail pipeline. The code will be publicly available after the paper's acceptance. We organize this section as follows. Firstly, we provide ablation studies regarding the choice of hyperparameters and their impact on our framework. Secondly, we provide the details of the evaluation dataset used in this work. Finally, we discuss limitations and potential future works, along with this work's broader impacts.

A.2 General Ablation Studies

User Study We randomly picked 5 prompts that are used for quantitative evaluation. For each selected prompt, 4 videos generated with simple trajectories and 4 videos produced under the guidance of complex trajectories are randomly chosen, resulting in a total of 40 videos per model. Participants were requested to evaluate the videos generated by each method and rate the videos across three criteria: Visual Quality, Trajectory Control, and Consistency, scaling from 1 to 5. As shown in Tab. 5, our framework consistently outperforms all baseline methods by a significant margin, highlighting the effectiveness of our approach.

Method	**Appearance** ↑	**Consistency** ↑	**Control** ↑
Peekaboo	3.15	2.61	3.45
Trailblazer	3.96	2.85	3.61
FreeTraj	4.15	3.21	3.95
Ours	**4.35**	**3.97**	**4.52**

Table 5: User-study averaged ratings on Appearance Quality, Temporal Consistency, and Trajectory Controllability. Best-performing metrics are stressed in bold text, and our method is highlighted in gray. All metrics expect higher values.

Ablation on Intervened Steps. In this ablation, we study the impact of the number of intervened initial denoising steps on our framework's performance. As shown in Fig. 6, the rooster remains still across the video when ZeroTrail is applied on a small number of denoising steps, while as the incurred denoising step rises to 5, our framework demonstrated noticeable trajectory controllability as the rooster moves naturally according to the trajectory bounding box while maintaining its appearance consistent. As the number of interventions continues to increase, despite the trajectory control remaining valid, visual artifacts begin to appear. When the intervention step is 10, the chicken only had its head visible in the video. For the extreme case of 20 denoising steps, a texture-like visual artifact is clearly visible across the video frames. To conclude, a smaller step of denoising intervention could lead to inadequate trajectory controllability, while excessively high intervention steps could cause the degradation of visual quality and subject consistency. Therefore, we set our framework's denoising intervention steps to 5.

A.3 Ablation Studies on TPIM

Ablation on Learning Rate. The impact of TPIM's test-time latent optimization learning rate value on overall generation quality is evaluated in this ablation experiment. As illustrated in Fig. 7, using suboptimal learning rates could impair trajectory guidance or visual quality. In the first row, an insufficiently small learning rate fails to enforce the moving trajectory, as evidenced by the lion initially moving towards the opposite direction as specified by the intended right-to-left trajectory. In contrast, applying an adequate learning rate of 0.035 yields natural motion with more precise trajectory control, as the lion's location and scale are better aligned with the specified bounding boxes. However, further increasing the learning rate introduces visual artifacts, as the vertical bar one shown in the third row. The ablation demonstrates that an optimal learning rate value is required to achieve the balance between trajectory adherence and visual fidelity. In our framework, the learning rate is set to 0.035.

Ablation on optimization steps The impact of the number of test-time latent refinement epochs on generation quality is examined in this ablation study. As shown in Fig. 8, using a few optimization steps (e.g., 3) results in insufficient guidance of the object's trajectory. Visualized in the first row, the fish fails to follow the specified path, remains largely misaligned with the bounding boxes throughout the sequence, and has its head turned backwards in the middle of the generated video, contradicting the specified trajectory. On the contrary, increasing the number of optimization steps could result in better trajectory adherence but come at the cost of suboptimal subject preservation capability. In this case, the fish exhibits noticeable appearance changes across frames, suggesting overfitting to the trajectory signal. In contrast, using 5 optimization steps offers a more favorable balance, achieving satisfactory trajectory control while maintaining the subject's visual consistency and quality. Accordingly, our framework adopts 5 optimization steps by default.

Ablation on Parameter p_{in} The impact of the within-bounding box sampling ratio p_{in}'s value is studied in this ablation experiment. Setting p_{in} too high could lead to visual artifacts such as washed-out or overexposed colors, possibly due to excessive alternation during the test-time optimization process. Conversely, setting it too low may weaken the trajectory prior injection, causing suboptimal control over the object's size and position, reducing the trajectory alignment's effectiveness. In Fig. 9, when p_{in} is set to 0.1, the rhino's head is clearly outside the bounding box at the first frame. In contrast, with $p_{\text{in}} = 0.25$, the rhino's position is much better constrained within the target region throughout the video. Although a high value, such as 0.4, might lead to stricter spatial adherence, it may also introduce visual artifacts such as the background desaturation shown in the third row. This highlights the importance of selecting appropriate p_{in} to achieve a balance between visual quality and effective trajectory control. In our framework, the default value of p_{in} is set to 0.25.

Ablation on Parameter p_{out} The impact of the outside-bounding box sampling ratio p_{out} is analyzed in this ablation experiment. Setting p_{out} too high can destabilize the visual generation process, which might be due to the introduction of excessive attention correlation from unrelated regions. As shown in Fig. 10, when p_{out} is increased to 0.15 and 0.20, the visual quality deteriorates noticeably—blurry artifacts emerge and roosters begin to appear along the entire motion trace, indicating chaotic generation.

Conversely, setting p_{out} too low may weaken trajectory guidance. In the first row, where $p_{\text{out}} = 0.05$, the rooster's movement and size slightly deviate from the specified bounding box comparing to the case where p_{out} is set to 0.1, and the subject consistency is compromised, which is evidenced by visual artifacts such as a third foot appearing in the second frame. These observations highlight the importance of selecting an appropriate p_{out} to achieve both visual coherence and accurate trajectory alignment. In our framework, we use $p_{\text{out}} = 0.1$ as the default setting.

A.4 Ablation Studies on SAGM

Ablation on parameter c_w The coefficient c_w controls the strength of the foreground attention guidance. A larger value reduces the differentiation between foreground and background attention, which could negatively impact localization. As illustrated in Fig. 11, when $c_w = 0.01$, the rooster's position in the first frame shows a slight deviation compared to other settings, indicating less precise localization. In contrast, setting c_w too high (e.g., 0.1) destabilizes the generation process, resulting in blurry artifacts and incomplete object shapes, such as a malformed chicken body. These findings suggest that excessively amplifying the attention contrast can compromise visual fidelity and frame stability. Based on experimental observations, we set $c_w = 0.01$ in our pipeline to maintain a balance between spatial precision and generation stability.

Ablation on parameter c_s We investigate the impact of the weighting coefficient c_s, which scales the Gaussian mask. As shown in Fig. 12, setting c_s too low compromises trajectory adherence. In the 0.1 case, the rooster struggles to follow the trajectory in the initial frames and progressively detaches from the bounding box in later frames, indicating insufficient motion guidance. On the other hand, increasing c_s excessively can lead to visual artifacts. In the 0.4 case, blurry artifacts appear in the foreground, partially obscuring the rooster and degrading visual fidelity. These results suggest that c_s must be carefully tuned to balance trajectory precision and appearance quality. Empirically, we find that setting $c_s = 0.25$ offers the best trade-off, enabling accurate control while preserving clean and coherent visual outputs.

A.5 Evaluation Dataset Details

In this section, we list all prompts used for evaluating the model. Similar to MotionZeroChen et al. [2024b], FreeTrajQiu et al. [2024], and TrailBlazerMa et al. [2023], we applied 33 diverse prompts featuring various subjects and motions in the evaluation dataset:

- A woodpecker climbing a tree trunk.
- A squirrel maneuvering on a tree after gathering nuts.
- A bird snatching fish from water.
- A frog leaping to catch a fly.
- A parrot flying among treetops.
- A squirrel jumping between trees.
- A rabbit digging in its warren.
- A satellite orbiting Earth in outer space.
- A skateboarder performing tricks at a skate park.
- A leaf drifting gently.
- A paper plane gliding in the air.
- A bear climbing a tree after spotting a threat.
- A duck diving to search for food.
- A kangaroo hopping on a gentle slope.
- An owl swooping to catch prey at night.
- A balloon drifting in a clear sky.
- A bus traversing London streets.
- A plane flying high in the sky.
- A helicopter hovering near city buildings.
- A streetcar trundling along tracks in a historic district.
- A rocket launching from a launchpad.
- A deer bounding in a snowy field.
- A horse galloping in a meadow.
- A fox prowling in a forest clearing.
- A swan floating gracefully on a lake.
- A panda munching bamboo in a bamboo forest.
- A penguin waddling on an iceberg.
- A lion prowling in savanna grass.
- An owl gliding silently at night.
- A dolphin just breaking the ocean surface.
- A camel trudging in a desert landscape.
- A kangaroo running in the Australian outback.
- A colorful hot air balloon floating in the sky.

Following previous works, we applied 8 simple base trajectories and 15 complex, diverse trajectories to evaluate the effectiveness and robustness of our framework as shown in Fig. 13, Fig. 14, and Fig. 15.

A.6 Limitations and Future Works

The plug-and-play, model-agnostic design of our framework made it applicable to diverse video diffusion models. As current video diffusion models still occasionally suffer from generating visual artifacts and misunderstanding the textual prompt, the framework could still generate undesirable videos due to the constraints of the base model's capability.

Despite demonstrating superior trajectory controllability and comparable visual quality, the hyperparameters applied in our implementation are empirical values that are suitable for most cases. Some of them might need to be adjusted when porting the framework to a new base model to achieve optimal performance.

Currently, we achieve effective trajectory controllability through user-specified trajectories, which may not be aligned with the description and underlying semantical of the textual prompt. For instance, the user may instruct the model to generate objects described as "standing still" while also expecting the object to follow along the trajectory. In certain cases, such a contradiction may prevent the model from producing desirable results, and achieving semantically-aware trajectory control will be the direction of our future work.

A.7 Broader Impacts

As our framework applies to video diffusion models, it may inherit the societal and public impacts, both positive and negative, of those with video diffusion models and generative video editing technologies. Additionally, owing to its plug-and-play and model-agnostic design, the introduction of ZeroTrail could facilitate the application and adaptation of Generative-AI-based content creation, potentially accelerating innovation and productivity within the creative industries.

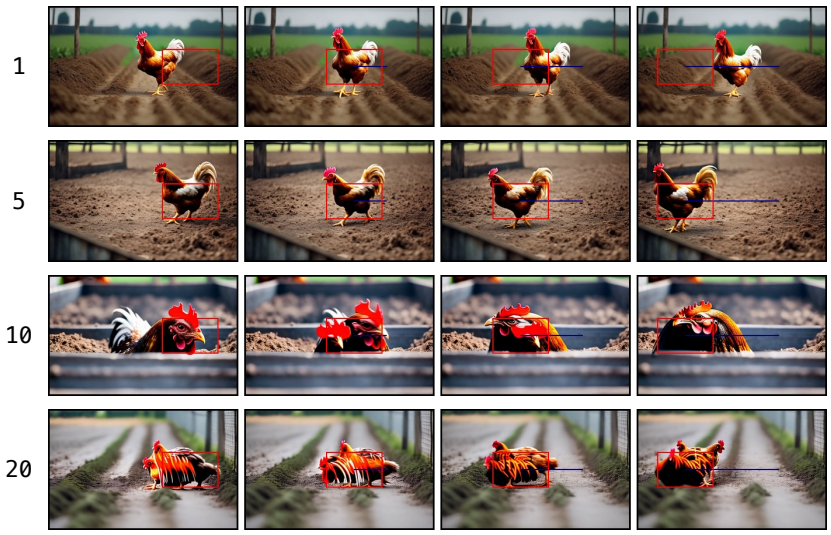

Figure 6: Ablation study on total intervened denoising steps from 0 to 20. Zoom in for a better view.

Figure 7: Ablation study on TPIM module's learning rate from 0.02 to 0.05. Zoom in for a better view.

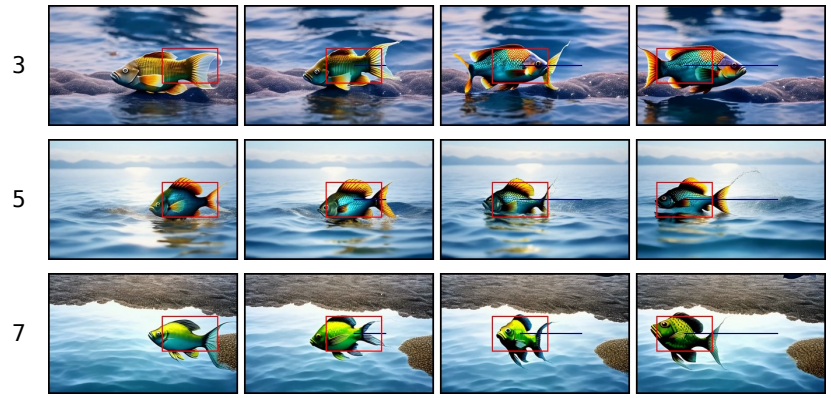

Figure 8: Ablation study on TPIM module's optimization step. Zoom in for a better view.

Figure 9: Ablation study on loss function's p_{in} parameter. Zoom in for a better view.

Figure 10: Ablation study on loss function's p_{out} parameter. Zoom in for a better view.

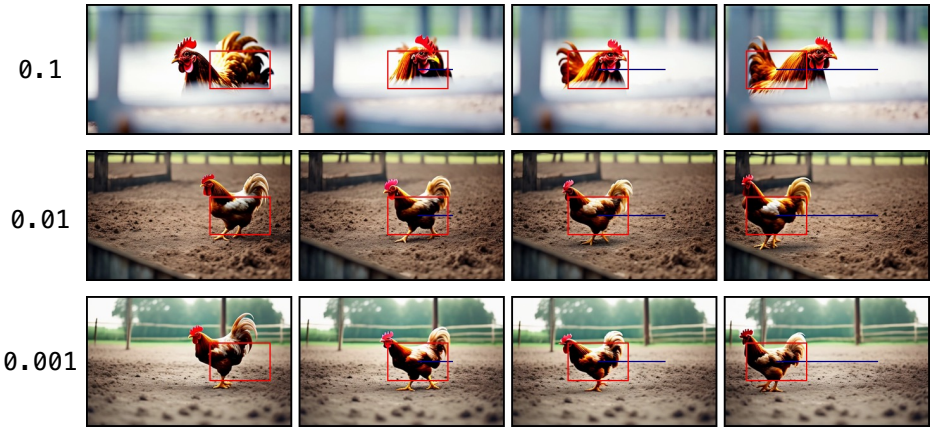

Figure 11: Ablation study on SAGM module's c_w parameter. Zoom in for a better view.

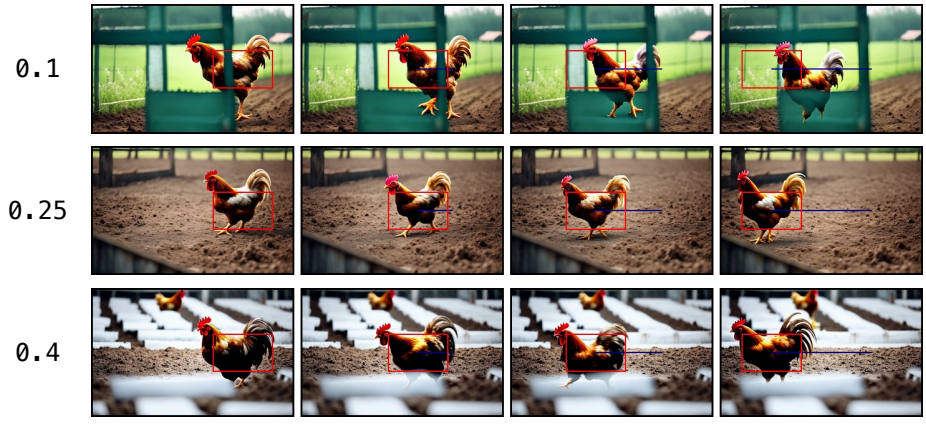

Figure 12: Ablation study on SAGM module's c_s parameter. Zoom in for a better view.

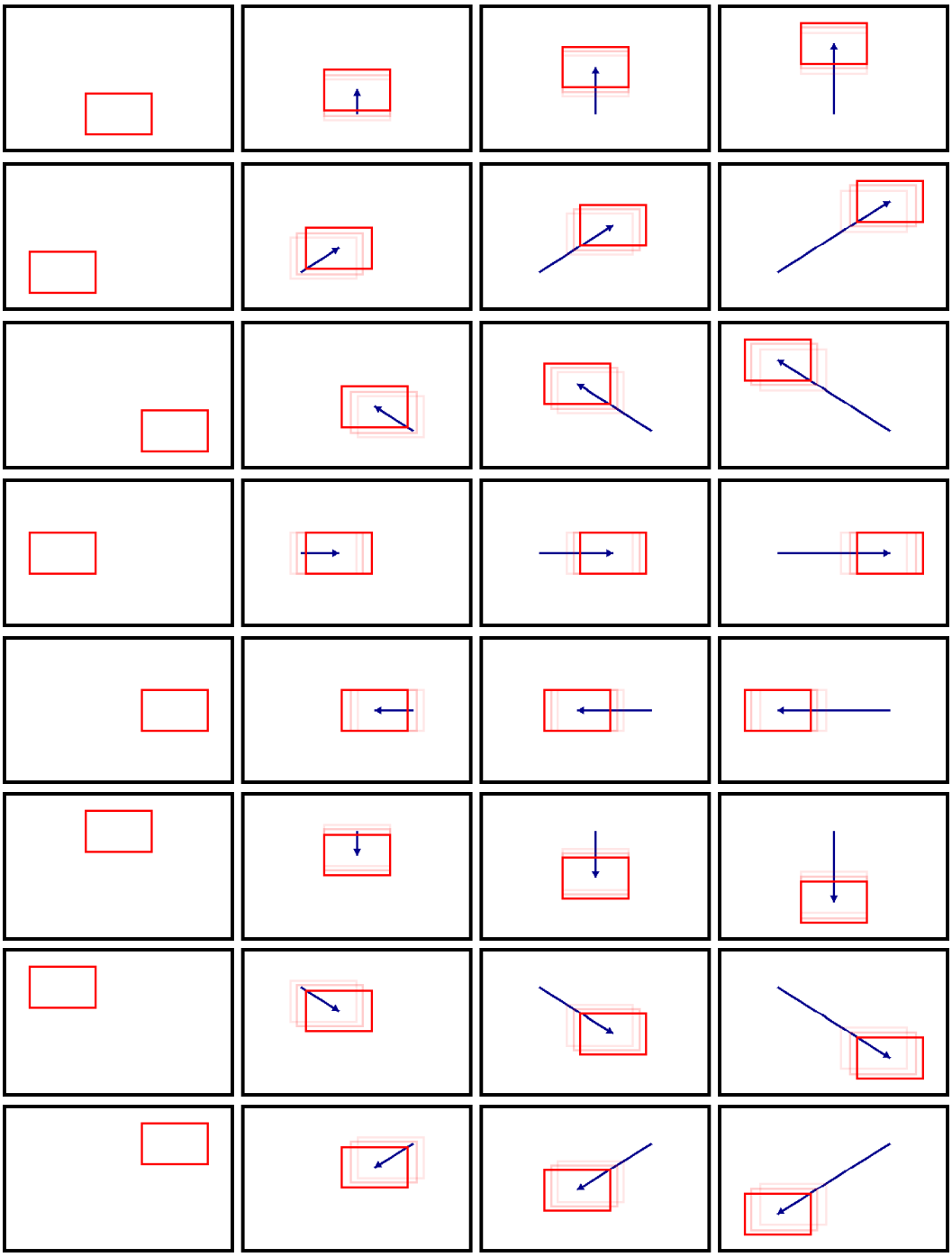

Figure 13: Visualization of simple trajectories. Each row corresponds to a trajectory and is read from left to right. Bounding boxes are in red while the moving trajectories are shown as blue arrowed lines.

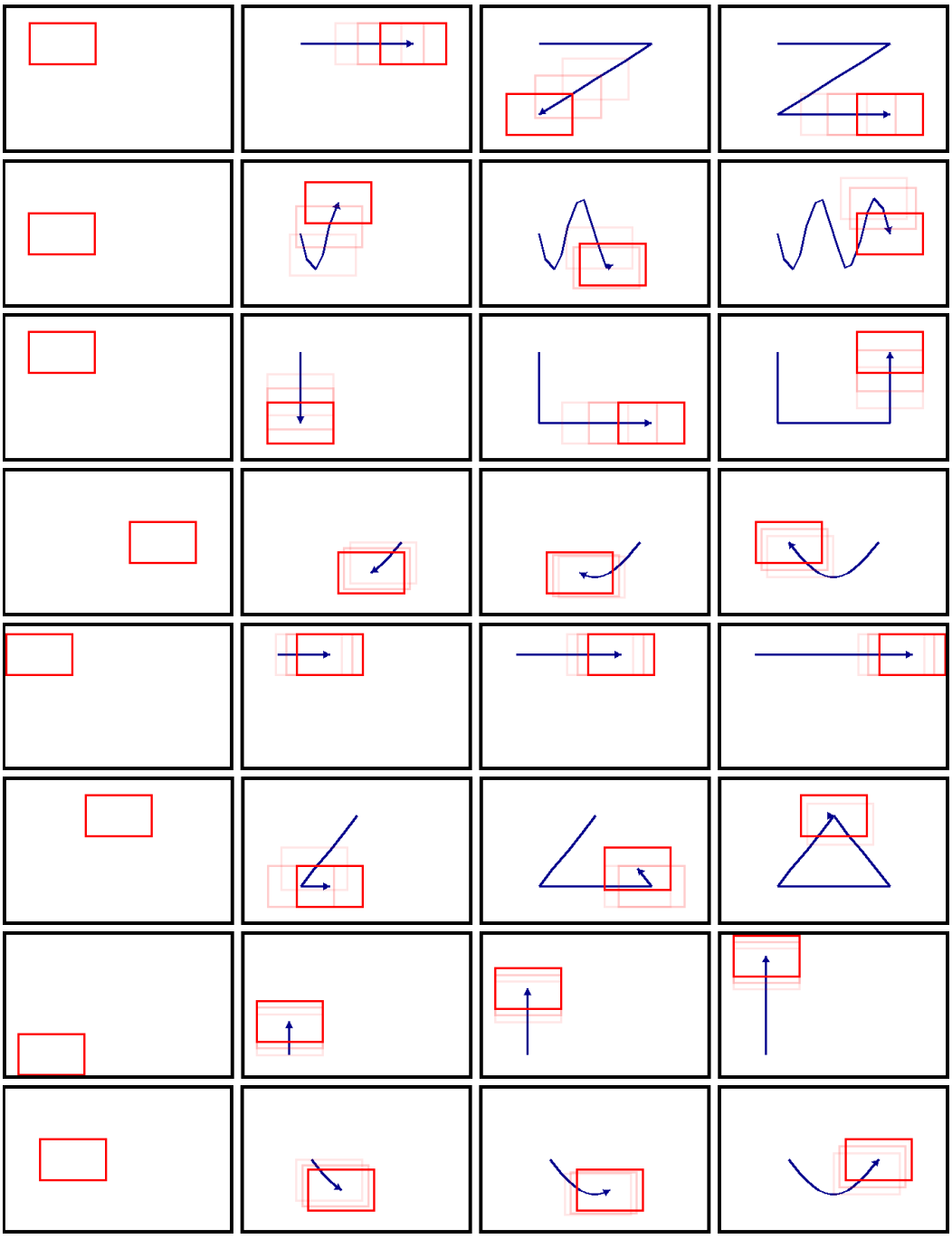

Figure 14: Visualization of the first 8 complex trajectories. Each row corresponds to a trajectory and is read from left to right. Bounding boxes are in red while the moving trajectories are shown as blue arrowed lines.

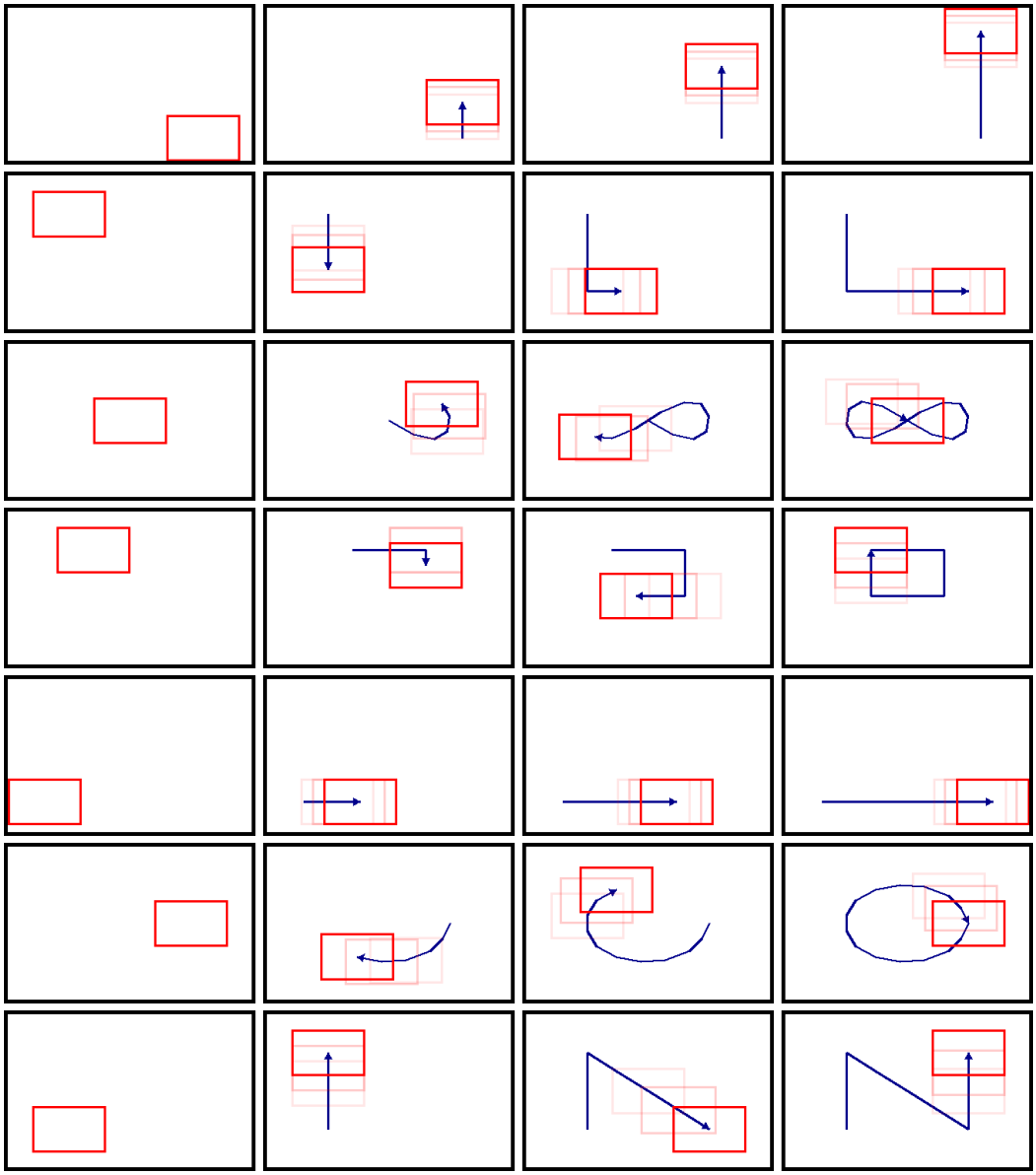

Figure 15: Visualization of the last 7 complex trajectories. Each row corresponds to a trajectory and is read from left to right. Bounding boxes are in red while the moving trajectories are shown as blue arrowed lines.