Zo3T: Zero-Shot 3D-Aware Trajectory-Guided Image-to-Video Generation via Test-Time Training

Ruicheng Zhang[1]*, Jun Zhou[1]* †, Zunnan Xu[1]*, Zihao Liu[1], Jiehui Huang[2], Mingyang Zhang[3], Yu Sun[4], Xiu Li[1]†
[1] Tsinghua University
[2] Hong Kong University of Science and Technology
[3] China University of Geosciences
[4] Sun Yat-sen University

Abstract

Trajectory-Guided image-to-video (I2V) generation aims to synthesize videos that adhere to user-specified motion instructions. Existing methods typically rely on computationally expensive fine-tuning on scarce annotated datasets. Although some zero-shot methods attempt to trajectory control in the latent space, they may yield unrealistic motion by neglecting 3D perspective and creating a misalignment between the manipulated latents and the network's noise predictions. To address these challenges, we introduce Zo3T, a novel zero-shot test-time-training framework for trajectory-guided generation with three core innovations: First, we incorporate a 3D-Aware Kinematic Projection, leveraging inferring scene depth to derive perspective-correct affine transformations for target regions. Second, we introduce Trajectory-Guided Test-Time LoRA, a mechanism that dynamically injects and optimizes ephemeral LoRA adapters into the denoising network alongside the latent state. Driven by a regional feature consistency loss, this co-adaptation effectively enforces motion constraints while allowing the pre-trained model to locally adapt its internal representations to the manipulated latent, thereby ensuring generative fidelity and on-manifold adherence. Finally, we develop Guidance Field Rectification, which refines the denoising evolutionary path by optimizing the conditional guidance field through a one-step lookahead strategy, ensuring efficient generative progression towards the target trajectory. Zo3T significantly enhances 3D realism and motion accuracy in trajectory-controlled I2V generation, demonstrating superior performance over existing training-based and zero-shot approaches.

Home Page — https://richard-zhang-ai.github.io/

1 Introduction

Recent advances in text- and image-driven video diffusion models have demonstrated remarkable capabilities in generating photorealistic and semantically coherent videos (Ho et al. 2022; Blattmann et al. 2023; Zhou et al. 2025; Yuan et al. 2025; Chen et al. 2023a; Xu et al. 2025; Ma et al. 2025c). These developments have paved the way for controllable object motion animation (He et al. 2024a; Namekata et al. 2024; Wang et al. 2025; Ma et al. 2025b), which plays a vital role in practical applications such as virtual reality, gaming, advertising, and digital art (Ma et al. 2025a).

Early works (Wu et al. 2024; Wang et al. 2024b; Kong et al. 2024) motion-controllable video generation employed annotated bounding boxes or trajectory points to fine-tune video diffusion models. However, their high computational cost severely limits their practical applicability. In contrast, training-free approaches have attracted significant attention due to their efficiency. These methods modify either the latent representations of target objects through attention mechanisms (Ma, Lewis, and Kleijn 2023; Namekata et al. 2024) or the noise construction (Qiu et al. 2024) during the denoising process, thereby enabling accurate motion control and visually coherent results. However, this paradigm of direct latent space manipulation harbors a fundamental tension. The pre-trained denoising network ϵ_θ has learned a delicate mapping from a specific distribution of noisy latents to their corresponding noise predictions. Aggressively editing $\mathbf{z}_t$ to enforce motion pushes it "off-manifold", creating a misalignment between the manipulated latent and the model's learned prior. This **model-data misalignment** forces the network to denoise a state it is never trained to see, frequently causing a catastrophic degradation in visual quality, manifesting as textural collapse, loss of identity, and other temporal artifacts (He et al. 2024b; Garibi et al. 2024). Moreover, most existing approaches guide motion using 2D trajectories paired with fixed-size bounding boxes or masks. This representation suffers from inherent ambiguity, as it fails to capture the perspective scaling an object should undergo as its depth changes. As shown in Figure 5(a), this lack of 3D awareness leads to physical implausibility and visual distortions, such as unrealistic scaling and motion patterns.

To address these limitations, we propose **Zo3T**, a novel Test-Time Training (TTT) framework for zero-shot image-to-video (I2V) generation with 3D-aware trajectory control. To mitigate the model-data misalignment in the trajectory guidance process, we introduce a "soft-editing" strategy that co-adapts the data and the model through Trajectory-Guided Test-Time Training. This mechanism injects and optimizes

*Equal contribution.
†Corresponding author.

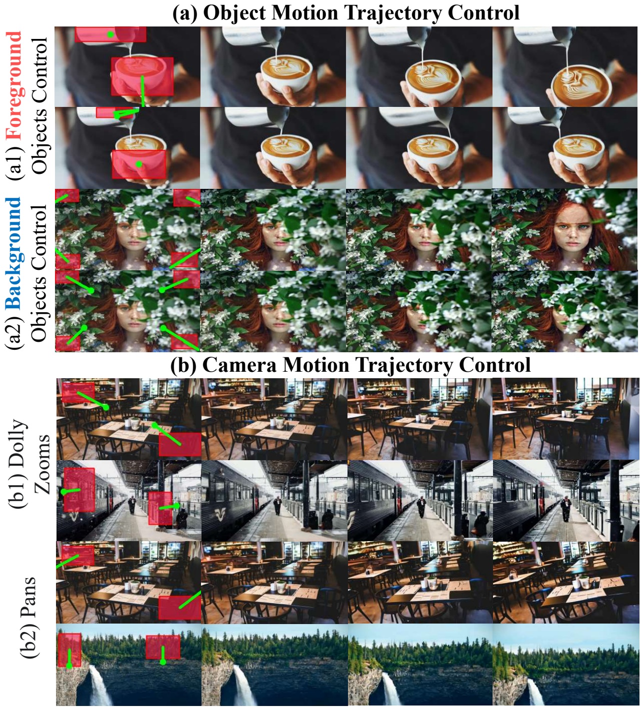

Figure 1: Versatile Trajectory Control with Our Method. Given a set of bounding boxes with corresponding trajectories, our framework enables precise control over diverse object and camera motions. By leveraging the inherent knowledge of a pre-trained video diffusion model, we achieve zero-shot trajectory guidance without any fine-tuning.

an ephemeral LoRA module within the denoising network at inference time. A trajectory-consistency loss simultaneously updates both the latent state $\mathbf{z}_t$ and the temporary LoRA weights. This co-adaptation realigns the model to the edited latent, ensuring the denoising process remains on a stable, high-fidelity manifold while adhering to the specified motion. Building on this, we further introduce Guidance Field Rectification (GFR) via re-evaluating noise scores, which adjusts the guidance field adaptively by maximizing the generation of instance features in the trajectory motion area at each denoising step to adaptively adjust the guidance field. This optimization encourages the latent features to evolve toward the desired trajectory, ensuring controllable and high-quality video synthesis. To ensure physical plausibility, Zo3T projects user-specified 2D trajectories into a 3D space, leveraging depth cues. This enables perspective-correct motion, ensuring realistic scaling and movement. While recent I2V methods like LeviTor (Wang et al. 2025) and ObjCtrl-2.5D (Wang et al. 2024a) also incorporate 3D information, LeviTor requires costly training on extensive masked datasets, and ObjCtrl-2.5D's core assumption of equating object motion with camera movement constrains its applicability in complex scenes. In contrast, our approach offers greater flexibility and directness, requiring no external motion priors or restrictive assumptions.

By jointly leveraging our 3D-aware trajectory control, test-time adaptation strategy, and guided field rectification module, ours framework achieves precise motion control and camera movement, as shown in Figure 1. Extensive experiments validate the effectiveness of Zo3T in generating visually coherent video sequences across a diverse range of trajectory control scenarios in a zero-shot setting. In summary, the main contributions of our framework can be summarized as follows:

- We propose Zo3T, a 3D-aware test-time training framework for zero-shot controllable video generation that enables precise control over both target object motion and camera movement.
- Zo3T integrates lightweight test-time LoRA modules during trajectory-guided generation test time to adaptively guide the generation process and maintain generative fidelity during latent manipulation. Furthermore, we refine the guidance field by re-evaluating noise scores to enforce trajectory fidelity.
- Extensive experiments demonstrate that our proposed method outperforms both training-based and training-free methods in terms of trajectory control and fidelity of generated videos.

2 Related Work

Controllable Generation with Diffusion Models

The success of diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Song et al. 2021) have turned focus from unconditional to fine-grained, controllable synthesis. While extensively explored in image generation (Zhang, Rao, and Agrawala 2023; Li et al. 2023; Avrahami et al. 2023; Mi et al. 2025), extending such control to the video domain introduces the formidable challenge of maintaining temporal coherence while manipulating spatial content (Guo et al. 2024). The dominant approach to this problem has been supervised fine-tuning, where a large, pre-trained video model is adapted to new control modalities on specialized datasets. This paradigm includes methods conditioned on dense signals, such as per-frame pose or depth maps (Zhao et al. 2023; Ma et al. 2023b; Chen et al. 2023b; Jin et al. 2024), which draw inspiration from image-centric architectures like ControlNet (Zhang, Rao, and Agrawala 2023).

To offer a more intuitive user interface, a significant lineage of work has focused on sparse trajectory control (Ma, Lewis, and Kleijn 2023; Guo et al. 2023; Zhang et al. 2024; Ma et al. 2023a). Early attempts often relied on intermediate representations like optical flow (Hao, Huang, and Belongie 2018), which could introduce visual artifacts. More recent methods integrate trajectory guidance directly into the diffusion backbone. For example, DragNUWA (Yin et al. 2023) demonstrated a framework for fusing multi-modal trajectory inputs through a multi-scale architecture. Recognizing the semantic ambiguity of a single moving point, DragAnything (Wu et al. 2024) advanced this concept to the entity-level by leveraging semantic features from the diffusion U-Net to guide the motion of entire objects beyond pixels. Despite their advances, these supervised methods are fundamentally constrained by the high computational cost

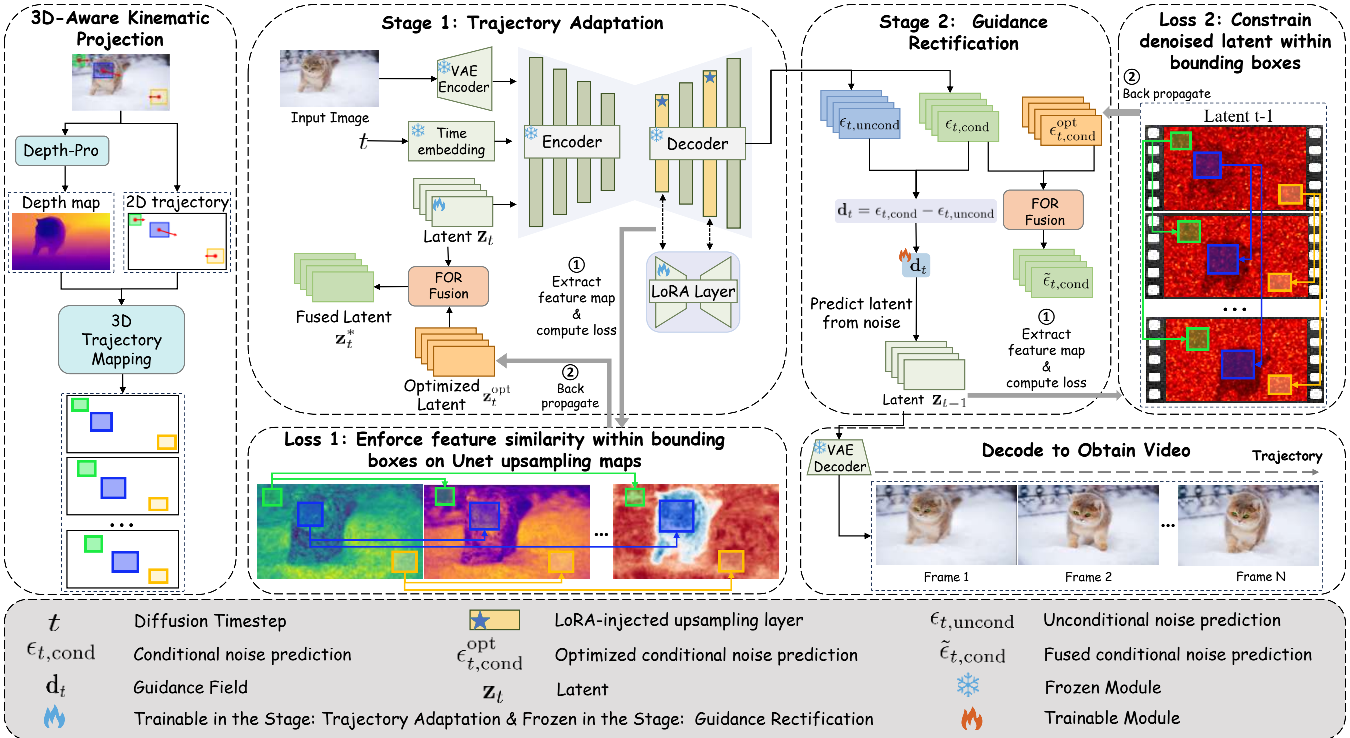

Figure 2: An overview of our zero-shot trajectory-guided video generation framework. Our method optimizes a pre-trained video diffusion model at specific denoising timesteps via two key stages. First, **Test-Time Training (TTT)** adapts the latent state and an ephemeral adapter to maintain semantic consistency along the trajectory. Second, **Guidance Field Rectification** refines the denoising direction using a one-step lookahead optimization to ensure precise path execution.

of retraining and their dependence on large-scale, annotated video datasets, which limits their flexibility and scalability.

Latent Manipulation for Zero-shot Optimization

Instead of fine-tuning, an alternative approach is to manipulate the generation process at inference, leaving the model weights frozen. This approach was first validated in the image domain (Hertz et al. 2023; Cao et al. 2023). Inspired by DragGAN (Pan et al. 2023), methods like DragDiffusion (Shi et al. 2024) established that precise, point-based spatial edits could be achieved by directly optimizing the noisy latent code z_t with respect to supervisory signals derived from the U-Net's internal feature maps.

However, translating this zero-shot optimization to video is non-trivial. A key obstacle is the weak temporal correspondence of feature maps within standard video diffusion models (Tang et al. 2023; Luo et al. 2023), which makes them unreliable for guiding motion. An intuitive solution involves a modification to the self-attention mechanism to enforce cross-frame feature alignment (Hertz et al. 2023; Cao et al. 2023; Geyer et al. 2023), thereby creating a usable guidance signal for latent optimization. However, a more fundamental challenge pervades these test-time optimization techniques. The direct, and often aggressive, manipulation of z_t can push the latent representation off the manifold learned by the pretrained model (He et al. 2024b; Garibi et al. 2024). This "manifold deviation" disrupts the delicate denoising process, frequently causing a significant degradation in visual quality, manifesting as textural collapse, loss of identity, and other temporal artifacts. Thus, a critical open question remains: how to enforce precise, user-defined control without sacrificing the generative fidelity inherent to the foundational model. Our work directly addresses this tension by introducing a new framework that co-adapts the model and the latent state, ensuring that the guided generation remains on-manifold.

3 Methodology

Zo3T is architected to function within the latent space of a pre-trained image-to-video model, such as Stable Video Diffusion (SVD), without requiring any fine-tuning. As illustrated in Figure 2, our method comprises three main components. First, we project user-defined 2D trajectories into a pseudo-3D space, leveraging monocular depth estimation to derive perspective-aware affine transformations. Second, we perform a **Test-Time Training (TTT)** to co-adapt the latent state and an ephemeral model adapter to enforce feature-space kinematic constraints during the structure-forming stages of denoising. Finally, the **Guidance Field Rectification** strategy refines the denoising direction via a one-step lookahead optimization, ensuring precise denoising path execution for trajectory control.

3D-Aware Kinematic Projection

To endow the model with a foundational awareness of 3D space, we establish a kinematic model to simulate perspective projection. Given a pre-trained monocular depth estimation network $\mathcal{D}_{\text{depth}}$, we compute the depth map $M_D = \mathcal{D}_{\text{depth}}(I_0)$ of the initial frame I_0. For an object centered at pixel coordinates $\mathbf{p}_k = (u_k, v_k)$ in frame k, we approximate its depth as $d_k = M_D(\mathbf{p}_k)$. The perspective scaling factor σ_k relative to the initial frame is then derived from a pinhole camera approximation: $\sigma_k = d_0/d_k$. This scaling allows us to define a time-varying affine transformation $\mathbf{A}_k$ that maps the initial bounding box $\mathcal{B}_0$ to its perspective-aware counterpart $\mathcal{B}_k$ for each frame k:

$$\mathbf{A}_k = \begin{pmatrix} \sigma_k & 0 & u_k - \sigma_k u_0 \\ 0 & \sigma_k & v_k - \sigma_k v_0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1}$$

From these transformations, we derive a set of time-varying binary masks $\mathcal{M}_k(\mathbf{p}) = [\mathbf{p} \in \mathbf{A}_k \mathcal{B}_0]$, which serve as the kinematic prior for subsequent optimization stages.

Zero-Shot Tajectory-Guided Adaptation via Test-Time Training

Video diffusion models learn a data distribution $p_\theta(\mathbf{x}_0)$ by reversing a noise-corruption process over T timesteps. To reduce computational cost, latent diffusion models like SVD operate in a compressed latent space, where a Variational Autoencoder (VAE) maps a raw video $\mathbf{x}_0$ to a latent representation $\mathbf{z}_0$. The reverse process is driven by a denoiser $\epsilon_\theta(\mathbf{z}_t, t, c)$ trained to predict the noise at timestep t. This network, which implicitly approximates the score function $\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t|c)$, defines the vector field that guides samples back towards the learned data manifold.

Given a well-defined kinematic prior $\mathcal{M}_k(\mathbf{p})$ that provides per-frame target regions for guidance, the primary challenge in zero-shot trajectory-guided generation lies in enforcing motion constraints within these regions without disrupting the delicate balance between the latent state $\mathbf{z}_t$ and the denoising network ϵ_θ. Prior zero-shot methods (Namekata et al. 2024; Zhang, Alcazar, and Ghanem 2025) often employ "hard edits" to the latents $\mathbf{z}_t$ to realize object movement. Such approaches can push the latent state off the learned data manifold into configurations the pre-trained model cannot coherently interpret, resulting in significant generative artifacts.

To resolve this challenge, we introduce a **Test-Time Training (TTT)** paradigm that performs a "soft adaptation" of both the data and the model. Inspired by previous drag-based control methods (Namekata et al. 2024; Yang et al. 2024; Pan et al. 2023), we leverage the cross-frame similarity of deep features within target regions as trajectory guiding signals. At a specific denoising step $t \in T$, we introduce an ephemeral low-rank perturbation $\Delta\theta_{\text{LoRA}}$ to the model's parameters and co-adapt it with the latent state $\mathbf{z}_t$ by enforcing feature consistency within the target regions. This TTT process seeks a new state-operator equilibrium $(\mathbf{z}_t^*, \theta' = \theta \oplus \Delta\theta_{\text{LoRA}})$ while minimizing a test-time objective functional, $\mathcal{J}_{\text{TTT}}$:

$$(\mathbf{z}_t^*, \theta_{\text{LoRA}}^*) = \text{argmin}_{\mathbf{z}_t, \theta_{\text{LoRA}}} \mathcal{J}_{\text{TTT}}(\mathbf{z}_t, \theta'). \tag{2}$$

The test-time objective $\mathcal{J}_{\text{TTT}}(\mathbf{z}_t, \theta')$ is a feature-space consistency loss that enforces alignment between the features of a tracked object in subsequent frames and its features in the first frame:

$$\mathcal{J}_{\text{TTT}} = \sum_{b=1}^{M} \sum_{l \in \mathcal{L}} w_l \sum_{k=2}^{N_f} \left\| \mathcal{G}_b \odot \left(F_{l,k}[\mathcal{M}_{b,k}] - (F_{l,1})_{\text{frozen}} [\mathcal{M}_{b,1}] \right) \right\|_F^2. \tag{3}$$

Here, $F_{l,k} = \phi_l(\mathbf{z}_t; \theta')_k$ represents the full feature map from layer l at frame k, produced by the adapted model θ'. The notation $[\mathcal{M}_{b,k}]$ denotes cropping the feature map to to the region defined by the mask for object b at frame k. Since the 3D-aware kinematic projection can yields regions of varying sizes across frames, we interpolate all cropped feature maps to a uniform size before computing the pixel-wise loss. The reference features from the first frame, $(F_{l,1})_{\text{frozen}}$, are detached from the computation graph (stop-gradient) to provide a stable optimization target. To mitigate boundary effects, a Gaussian heatmap $\mathcal{G}_b$ focuses the loss on the object's center via the Hadamard product ($\odot$) (Namekata et al. 2024). For the selection of feature layers $\mathcal{L}$, we follow the methodology of (Namekata et al. 2024; Wu et al. 2023), which has been proven to yield superior cross-frame spatial-semantic alignment. The TTT process, implemented as a short iterative loop at specific denoising steps, ensures the resulting trajectory is both geometrically accurate and semantically coherent. Trajectory-Guided TTT steers the model, via a semantic loss, to adjust the content layout in subsequent frames, thereby realizing controllable object motion.

Guidance Field Rectification via One-Step Lookahead

While the TTT process successfully constrains the solution space to ensure semantic consistency along the trajectory, it does not prescribe an optimal path within that space. The resulting denoising direction is therefore not guaranteed to be the most efficient or precise for trajectory adherence. Furthermore, the strict spatial consistency enforced by TTT can sometimes lead to unnatural or rigid visual results. To resolve this, we introduce a lookahead-based guidance rectification strategy inspired by the principles of Classifier-Free Guidance (CFG) (Ho and Salimans 2022).

Within the CFG framework, the denoising direction is synthesized from a conditional prediction, which drives the generation towards the control target, and an unconditional prediction, which maintains generative naturality. Their difference can be interpreted as the **guidance field**, $\mathbf{d}_t = \boldsymbol{\epsilon}_t^{\text{cond}} - \boldsymbol{\epsilon}_t^{\text{uncond}}$, encapsulating the directional force for control (Li et al. 2025). However, this standard guidance field provides a semantically plausible path of least resistance, which may not align with the globally optimal path for precise trajectory control.

To overcome this challenge, we propose Guidance Field Rectification (GFR) by optimizing it with foresight. This is achieved via a **one-step lookahead** optimization, where our objective is to find an optimal field, $\mathbf{d}_t^*$, that steers the probability flow ODE in a direction that is demonstrably optimal for the subsequent state's alignment with the kinematic prior. This lookahead rectification stage contrasts fundamentally with TTT: whereas TTT adapts the state-operator pair $(\mathbf{z}_t, \theta')$ in *feature space* to ensure representational capacity, this stage rectifies the guidance field $\mathbf{d}_t$ by minimizing a kinematic loss in *latent space*. To this end, we define a cost functional, $\mathcal{J}_{\text{guide}}$, that measures the kinematic inconsistency of the *prospective* latent state at timestep $t-1$:

$$\mathcal{J}_{\text{guide}}(\mathbf{d}) = \frac{1}{2}\sum_{b=1}^{M}\sum_{k=2}^{N_f}\int_{\mathcal{B}_{b,k}} \mathcal{G}_b(\mathbf{p}) \Big\| \text{Pooling}\big(\mathcal{P}(\mathbf{z}_t^*, \mathbf{d})_k[\mathcal{B}_{b,1}]\big) - \text{Pooling}\big(\mathcal{P}(\mathbf{z}_t^*, \mathbf{d})_1[\mathcal{B}_{b,1}]\big)_{\text{frozen}} \Big\|_2^2 \, d\mathbf{p}. \tag{4}$$

Here, $\mathcal{P}(\mathbf{z}_t^*, \mathbf{d})$ denotes the one-step DDIM solver that evolves the current state $\mathbf{z}_t^*$ to a prospective state $\mathbf{z}_{t-1}$ using a candidate field $\mathbf{d}$. Unlike the pixel-wise MSE loss used in our TTT stage, this guidance loss is computed on spatially pooled features, capturing a holistic representation of the target region. This form of supervision serves a dual purpose: It further guides the object towards the target trajectory while relaxing the strict spatial constraints of TTT, thereby mitigating potential unnatural or rigid visual generation. The optimal field $\mathbf{d}_t^*$ is found by evolving an initial field, $\mathbf{d}_0 = \mathbf{d}_t$, along the negative gradient flow of this functional, as described by $\partial \mathbf{d}/\partial \tau = -\nabla_{\mathbf{d}} \mathcal{J}_{\text{guide}}(\mathbf{d})$. The final, rectified conditional noise estimate is then recomposed:

$$\boldsymbol{\epsilon}_t^{\text{cond, opt}} = \boldsymbol{\epsilon}_t^{\text{uncond}} + \mathbf{d}_t^*. \tag{5}$$

This lookahead rectification strategy enables precise steering of the generative trajectory by maximizing its alignment with motion constraints, achieved with minimal overhead via a short, localized optimization loop.

Inference Strategies

Fidelity Preservation via Fourier Orthogonal Recomposition. While our iterative optimizations achieve effective trajectory control, they risk causing a distributional shift from the diffusion model's learned prior, leading perceptual quality degradation. Recognizing that motion is predominantly encoded in low-frequency components (Wu et al. 2025), we introduce the Fourier Orthogonal Recomposition (FOR) strategy. This method preserves the high-frequency textural details from the original signal $\mathbf{x}$, which are crucial for maintaining distributional fidelity, while integrating the low-frequency structural modifications from our optimization $\mathbf{x}^*$. We define a generalized fusion operator $\mathcal{M}_{fuse}$ as:

$$\mathcal{M}_{fuse}(\mathbf{x}^*, \mathbf{x}) \triangleq \mathcal{F}^{-1}\big(\mathcal{F}(\mathbf{x}^*) \odot \mathbf{H}_\gamma + \mathcal{F}(\mathbf{x}) \odot (1 - \mathbf{H}_\gamma)\big), \tag{6}$$

where $\mathcal{F}$ is the Fourier Transform and $\mathbf{H}_\gamma$ is an ideal low-pass filter mask. This operator is applied twice at the denoising step t: first to the latent states ($\tilde{\mathbf{z}}_t = \mathcal{M}_{fuse}(\mathbf{z}_t^*, \mathbf{z}_t)$) following TTT, and second to conditional noise predictions ($\tilde{\epsilon}_{t,\text{cond}} = \mathcal{M}_{fuse}(\boldsymbol{\epsilon}_{t,\text{cond}}^*, \boldsymbol{\epsilon}_{t,\text{cond}})$) after guidance rectification. By drawing low-frequency structure from optimization and high-frequency details from the original signal, the module preserves control accuracy and visual fidelity.

Selective Timestep Optimization. The coarse structure and motion dynamics of the generated video are largely determined during the early stages of the denoising process (Wang and Vastola 2023). To maximize computational efficiency and ensure structural stability, we apply our trajectory guidance optimizations selectively. Specifically, both the TTT adaptation and GFR are performed only during the early-to-mid denoising phase, from timestep $t = 45$ down to $t = 30$ in a 50-step schedule. This selective application aligns with established practices in the controllable generation literature (Namekata et al. 2024), striking an effective balance between precise control and generative quality.

4 Experiments

Experiment Settings

Implementation Details. Our framework is built upon the Stable Video Diffusion (SVD) (Blattmann et al. 2023), initialized with its official pre-trained weights. All experiments are configured to generate 14-frame videos at a resolution of 576×1024. The optimization process employs the Adam optimizer with learning rates of 0.25, 0.01, and 0.05 for the latent, LoRA weights, and the guidance field, respectively. We employ DepthPro (Bochkovskii et al. 2025) for monocular depth estimation, the output of which is used to construct 3D-aware trajectories and the corresponding affine transformations of target regions. For the Fourier Orthogonal Recomposition (FOR) module, the low-pass filter's cutoff frequency is set to 0.6. A comprehensive ablation analysis of these experimental parameters is detailed in the ***Appendix***.

Evaluation Metrics. We conduct a comprehensive evaluation using both quantitative metrics and a user study. For quantitative assessment, we measure ***Video Quality*** via FID, FVD, and four key VBench (Huang et al. 2024) metrics (subject consistency, background consistency, aesthetic quality, and imaging quality), and ***Motion Control Accuracy*** using Object Motion Conformity (ObjMC) (Namekata et al. 2024)—the Euclidean distance between target and actual trajectories. In parallel, our user study involved six expert evaluators who assessed 50 videos generated from a custom dataset of images, each with a predefined trajectory.

Evaluation Datasets. For a fair comparison, we adopt the open-source dataset used in (Wu et al. 2024). Recognizing that many baseline I2V models excel on common, in-distribution trajectories (Wang et al. 2024a), we augment this set by applying a "mirroring" transformation to most paths, resulting in a more challenging set of 143 samples featuring more diverse and less common motion paths. Following the protocols in (Wu et al. 2024), FID and FVD metrics are computed on the VIPSeg (Miao et al. 2022) dataset.

Versatile Trajectory Control Modes

Zo3T provides flexible trajectory control for any designated entity via a bounding box. This includes enforcing regional stillness by setting the trajectory vector to zero. Figure 1 showcases two primary control modalities offered by our framework. (1) **Object Motion Trajectory Control:** By

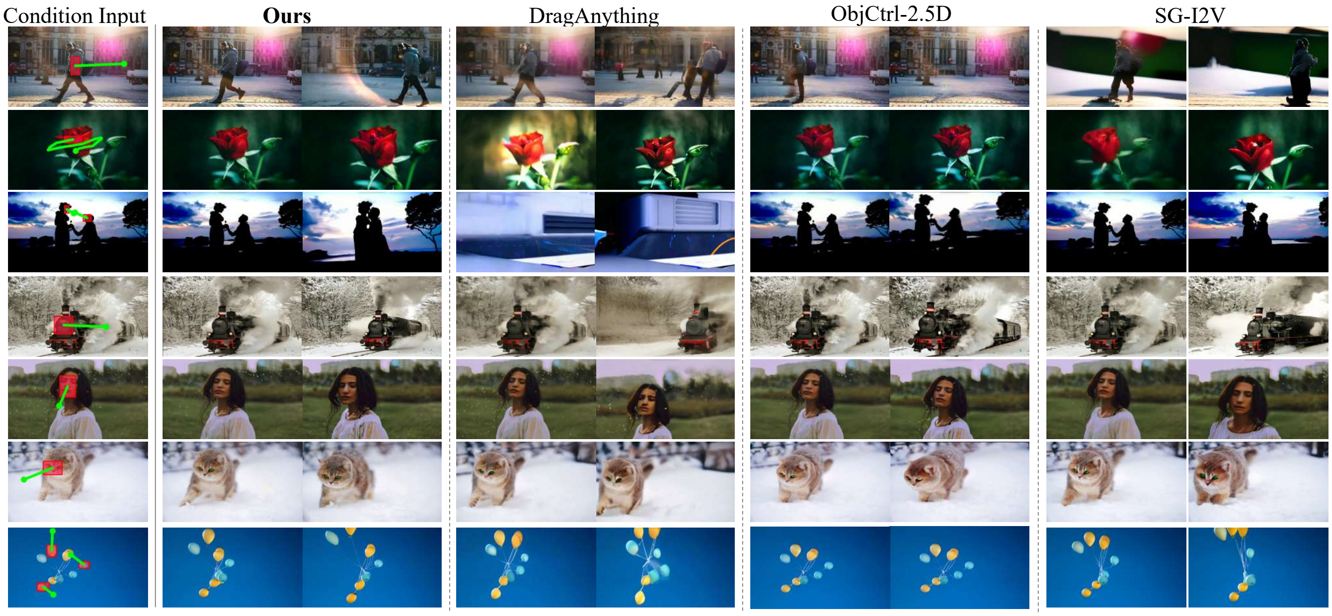

Figure 3: Qualitative Comparison with SOTA Methods.

Method	Zero-shot	FID (↓)	FVD (↓)	ObjMC (↓)	Subject Consist.(↑)	Bkg. Consist.(↑)	Aesthetic Quality(↑)	Imaging Quality(↑)	Resolution	Backbone
DragNUWA (Yin et al. 2023)	✗	126.31	251.04	**10.84**	0.9177	0.9272	0.5412	0.5433	320 × 576	SVD
DragAnything (Wu et al. 2024)	✗	119.07	266.42	11.64	0.9204	0.9262	0.5469	0.5711	320 × 576	SVD
LeviTor (Wang et al. 2025)	✗	79.41	207.23	12.06	0.9482	0.9511	0.6417	0.6314	288 × 512	SVD
FreeTraj (Qiu et al. 2024)	✓	92.12	230.21	31.63	0.9236	0.9281	0.5841	0.5903	320 × 512	VideoCrafter2
ObjCtrl-2.5D (Wang et al. 2024a)	✓	81.06	212.17	18.72	0.9512	0.9544	0.6344	0.6429	320 × 576	SVD
SG-I2V (Namekata et al. 2024)	✓	79.36	209.53	14.43	0.9448	0.9517	0.6370	0.6317	576 × 1024	SVD
Ours	✓	**74.83**	**197.63**	12.74	**0.9760**	**0.9682**	**0.6779**	**0.6820**	576 × 1024	SVD

Table 1: Quantitative comparison on the VIPSeg dataset. Despite being a zero-shot approach, our method achieves a small gap in motion fidelity (ObjMC) compared to supervised baselines, without degrading video quality (FID, FVD). Furthermore, our approach outperforms other zero-shot baselines across all metrics. (↓) indicates lower is better, (↑) indicates higher is better.

defining a bounding box around a target object and specifying a desired trajectory, Zo3T can direct the object's movement along the prescribed path. As shown in Figure 1(a), Zo3T can effectively control both foreground (a1) and background (a2) objects, ensuring they accurately track their respective trajectories. (2) **Camera Motion Trajectory Control:** Camera motion is simulated by defining a bounding box over a background region and assigning it a trajectory inverse to the intended camera movement. As shown in Figure 1(b), Zo3T can can effectively perform complex camera operations, including dolly zooms (b1) and pans (b2).

Comparisons with State-of-the-Art Methods

We perform a comprehensive evaluation by comparing Zo3T with SOTA supervised and zero-shot baselines. For supervised baselines, we select DragNUWA (Yin et al. 2023), DragAnything (Wu et al. 2024) and LeviTor (Wang et al. 2025). The zero-shot methods include SG-I2V (Namekata et al. 2024), ObjCtrl-2.5D (Wang et al. 2024a), and an adapted version of FreeTraj (Qiu et al. 2024) for our I2V task, following the procedure in (Namekata et al. 2024). For a fair evaluation, all video outputs are resized to a 320×576 resolution to align with the supervised models.

Comparison with Supervised Methods. Supervised methods, trained on tracker-derived trajectories, exhibit slightly higher motion precision (ObjMC) but suffer from significant visual artifacts like object distortion, flickering, and style collapse due to their prioritization of trajectory adherence over generative fidelity (Table 1, Figure 3). In contrast, our method integrates a Test-Time Training (TTT) generative guidance paradigm with the Fourier Orthogonal Recomposition (FOR) module to restore high-frequency details, yielding superior generation quality (FID, FVD) and more natural motion. Crucially, supervised methods are often limited to lower resolutions (320×576) by the high cost of fine-tuning, whereas our zero-shot framework operates directly at the native high resolution of SVD (576×1024) without requiring any external training or external knowledge.

Comparison with Zero-shot Methods. Our method demonstrates clear superiority over all zero-shot baselines (Figure 3, Table 1). FreeTraj's handcrafted noise prior compromises fine details and offers only coarse control. ObjCtrl-

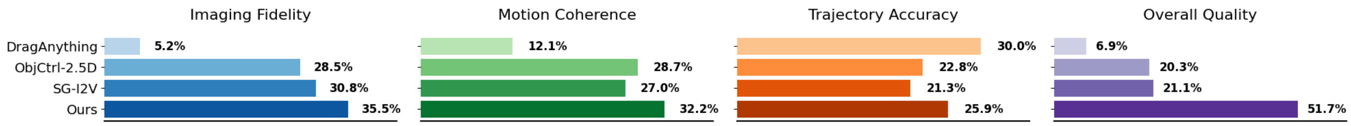

Figure 4: User Study. We evaluate subjective perceptual quality via a user study where six trained participants performe preference voting on our method against three top competitors based on imaging fidelity, motion coherence, trajectory accuracy, and overall quality. The majority of participants preferred the results obtained by our method over both training-free and training-based methods, attributing this preference to its better trajectory alignment and more natural motion generation

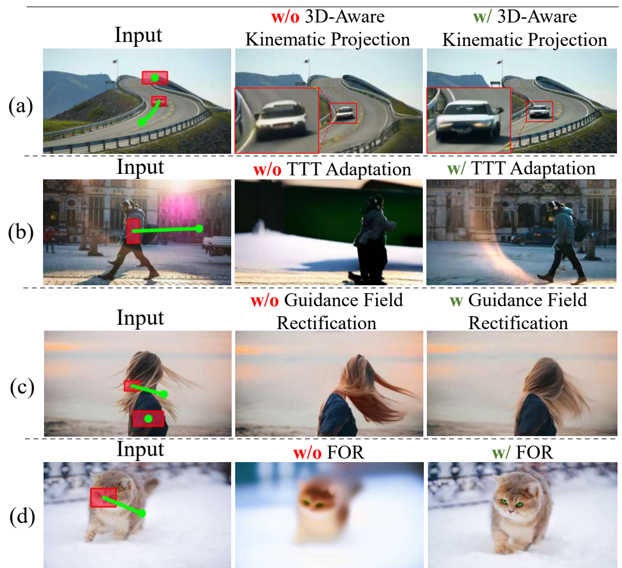

Figure 5: Qualitative Ablation Study

2.5D, by equating object motion with camera movement, is restricted to single-object control and proves ineffective in complex scenarios requiring coordinated object-background motion. Our method, conversely, offers flexible multi-region trajectory guidance, achieving more accurate and natural motion without additional dependencies. While SG-I2V strikes a reasonable balance between quality and accuracy, its "hard editing" in latent space risks manifold deviation, which can lead to artifacts and unnatural results. In contrast, our TTT paradigm ensures manifold integrity through a co-adaptive ephemeral LoRA adapter and a GFR module, enabling precise, high-fidelity trajectory guidance.

Ablation Studies

We validate each core component via quantitative (Table 2) and qualitative (Figure 5) ablations. ***Effectiveness of 3D-Aware Kinematic Projection.*** In scenes with depth changes, enforcing consistency with an unscaled 2D box causes severe object distortions (Figure 5(a)). This demonstrates the necessity of our 3D-aware projection for ensuring physically plausible motion and realistic perspective shifts. ***Effectiveness of TTT Trajectory-Guided Adaptation.*** Removing the LoRA adapter degenerates our method into "hard editing" of the latent z_t. While this strategy can roughly follow the trajectory, it suffers from a clear quality drop and causes scene collapse in difficult cases (Figure 5(b)). This confirms that TTT is critical for resolving latent-denoiser misalignment and preserving the generative manifold. ***Effectiveness of Guidance Field Rectification.*** Removing GFR leads to degradation across all metrics (Table 2). By using a one-step lookahead with holistic feature loss, GFR relaxes the strict spatial constraints imposed by TTT's pixel-wise loss. This mitigates unnatural rigidity and promotes organic motion, effectively harmonizing trajectory accuracy with natural dynamics. ***Effectiveness of Fourier Orthogonal Recomposition.*** Removing FOR has a negligible effect on motion accuracy but significantly worsens FID and FVD scores. Qualitatively, videos generated without FOR appear overly smooth and lack fine texture details (Figure 5(d)). This validates FOR's essential role in eliminating optimization-induced artifacts and restoring visual fidelity.

Variant	FID (↓)	FVD (↓)	ObjMC (↓)
Ours (Full Model)	**74.83**	**197.63**	**12.74**
(a) w/o 3D Projection	76.12	201.55	12.98
(b) w/o TTT Adaptation	89.36	219.53	14.24
(c) w/o GFR	78.91	205.18	13.92
(d) w/o FOR	95.45	221.09	12.81

Table 2: Ablation study on the core components of our framework. We start with our full model and progressively remove each key component. The results demonstrate that every module contributes positively to either video quality (FID/FVD) or motion accuracy (ObjMC).

5 Conclusion

In this paper, we introduce a novel method for trajectory-guided video generation that addresses the critical challenges of spatial ambiguity and generative quality degradation. Our core contribution is a two-stage test-time optimization framework. First, we perform test-time training to co-adapt the latent state and a lightweight model adapter, which steers the object along a perspective-aware 3D path while keeping the generation on the learned data manifold. Second, we introduce Guidance Field Rectification, a one-step lookahead optimization that refines the denoising direction for precise path execution. By integrating these optimizations with a Fourier-based fusion strategy to preserve high-frequency details, our method achieves SOTA performance, generating videos with superior motion accuracy and visual fidelity without the need for supervised fine-tuning.

Acknowledgements

This work was partly supported by Shenzhen Key Laboratory of next generation interactive media innovative technology (No:ZDSYS20210623092001004) .

References

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. SpaText: Spatio-textual representation for controllable image generation. In *CVPR*.

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*.

Bochkovskii, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S. R.; and Koltun, V. 2025. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. In *International Conference on Learning Representations*.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023a. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*.

Chen, W.; Wu, J.; Xie, P.; Wu, H.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023b. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2305.13840*.

Garibi, D.; Patashnik, O.; Voynov, A.; Averbuch-Elor, H.; and Cohen-Or, D. 2024. ReNoise: Real Image Inversion Through Iterative Noising. In *ECCV*.

Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2023. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. *arXiv preprint arxiv:2307.10373*.

Guo, J.; Zhang, D.; Liu, X.; Zhong, Z.; Zhang, Y.; Wan, P.; and Zhang, D. 2024. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*.

Guo, Y.; Yang, C.; Rao, A.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*.

Hao, Z.; Huang, X.; and Belongie, S. 2018. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7854–7863.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2024a. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*.

He, Y.; Murata, N.; Lai, C.-H.; Takida, Y.; Uesaka, T.; Kim, D.; Liao, W.-H.; Mitsufuji, Y.; Kolter, J. Z.; Salakhutdinov, R.; et al. 2024b. Manifold preserving guided diffusion. In *ICLR*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-prompt image editing with cross attention control. In *ICLR*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; 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. In *Advances in Neural Information Processing Systems*, volume 33, 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jin, X.; Xu, Z.; Ou, M.; and Yang, W. 2024. Alignment is All You Need: A Training-free Augmentation Strategy for Pose-guided Video Generation. *arXiv preprint arXiv:2408.16506*.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Li, G.; Yang, Y.; Song, C.; and Zhang, C. 2025. FlowDirector: Training-Free Flow Steering for Precise Text-to-Video Editing. *arXiv preprint arXiv: 2506.05046*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. GLIGEN: Open-set grounded text-to-image generation. In *CVPR*.

Luo, G.; Dunlap, L.; Park, D. H.; Holynski, A.; and Darrell, T. 2023. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. In *NeurIPS*.

Ma, W.-D. K.; Lewis, J.; and Kleijn, W. B. 2023. Trail-Blazer: Trajectory Control for Diffusion-Based Video Generation. *arXiv preprint arXiv:2401.00896*.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023a. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; Feng, K.; Hu, Z.; Wang, X.; Wang, Y.; Zheng, M.; He, X.; Zhu, C.; Liu, H.; He, Y.; et al. 2025a. Controllable Video Generation: A Survey. *arXiv preprint arXiv:2507.16869*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023b. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186*.

Ma, Y.; Liu, Y.; Zhu, Q.; Yang, A.; Feng, K.; Zhang, X.; Li, Z.; Han, S.; Qi, C.; and Chen, Q. 2025b. Follow-Your-Motion: Video Motion Transfer via Efficient Spatial-Temporal Decoupled Finetuning. *arXiv preprint arXiv:2506.05207*.

Ma, Y.; Yan, Z.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; et al. 2025c. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*.

Mi, Y.; Zhong, Z.; Huang, Y.; Yuan, Q.; Zhao, X.; Xu, J.; Ding, S.; Wang, S.; Guo, R.; and Zhou, S. 2025. Data Synthesis with Diverse Styles for Face Recognition via 3DMM-Guided Diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21203–21214.

Miao, J.; Wang, X.; Wu, Y.; Li, W.; Zhang, X.; Wei, Y.; and Yang, Y. 2022. Large-scale Video Panoptic Segmentation in the Wild: A Benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Namekata, K.; Bahmani, S.; Wu, Z.; Kant, Y.; Gilitschenski, I.; and Lindell, D. B. 2024. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*.

Pan, X.; Tewari, A.; Leimkühler, T.; Liu, L.; Meka, A.; and Theobalt, C. 2023. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 conference proceedings*, 1–11.

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, Y.; Xue, C.; Liew, J. H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V. Y.; and Bai, S. 2024. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *CVPR*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-based generative modeling through stochastic differential equations. In *ICLR*.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent correspondence from image diffusion. *NeurIPS*.

Wang, B.; and Vastola, J. J. 2023. Diffusion Models Generate Images Like Painters: an Analytical Theory of Outline First, Details Later. *CoRR*, abs/2303.02490.

Wang, H.; Ouyang, H.; Wang, Q.; Wang, W.; Cheng, K. L.; Chen, Q.; Shen, Y.; and Wang, L. 2025. LeviTor: 3D trajectory oriented image-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, Z.; Lan, Y.; Zhou, S.; and Loy, C. C. 2024a. ObjCtrl-2.5D: Training-free Object Control with Camera Poses. In *arXiv preprint arXiv:2412.07721*.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024b. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. 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, T.; Si, C.; Jiang, Y.; Huang, Z.; and Liu, Z. 2025. FreeInit: Bridging Initialization Gap in Video Diffusion Models. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 378–394. Cham: Springer Nature Switzerland.

Wu, W.; Li, Z.; Gu, Y.; Zhao, R.; He, Y.; Zhang, D. J.; Shou, M. Z.; Li, Y.; Gao, T.; and Zhang, D. 2024. DragAnything: Motion Control for Anything using Entity Representation. In *ECCV*.

Xu, Z.; Yu, Z.; Zhou, Z.; Zhou, J.; Jin, X.; Hong, F.-T.; Ji, X.; Zhu, J.; Cai, C.; Tang, S.; et al. 2025. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15909–15919.

Yang, S.; Hou, L.; Huang, H.; Ma, C.; Wan, P.; Zhang, D.; Chen, X.; and Liao, J. 2024. Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24)*, 12. New York, NY, USA: ACM.

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*.

Yuan, S.; Huang, J.; He, X.; Ge, Y.; Shi, Y.; Chen, L.; Luo, J.; and Yuan, L. 2025. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12978–12988.

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

Zhang, T.; Alcazar, J. C. L.; and Ghanem, B. 2025. Motion-Aware Concept Alignment for Consistent Video Editing. arXiv:2506.01004.

Zhang, Z.; Liao, J.; Li, M.; Qin, L.; and Wang, W. 2024. Tora: Trajectory-oriented Diffusion Transformer for Video Generation. *arXiv preprint arXiv:2407.21705*.

Zhao, R.; Gu, Y.; Wu, J. Z.; Zhang, D. J.; Liu, J.; Wu, W.; Keppo, J.; and Shou, M. Z. 2023. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*.

Zhou, J.; Li, J.; Xu, Z.; Li, H.; Cheng, Y.; Hong, F.-T.; Lin, Q.; Lu, Q.; and Liang, X. 2025. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13093–13103.