Test-Time Domain Adaptation for Interactive Video Generation

Anonymous CVPR submission

Paper ID *****

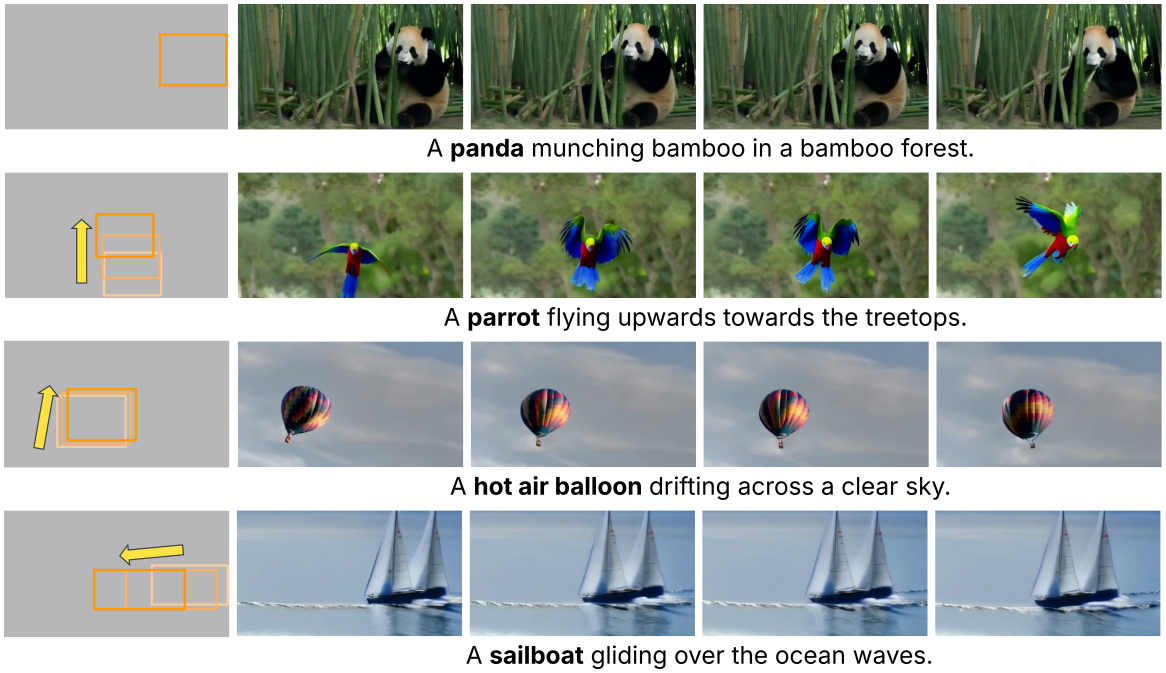


Figure 1. **Trajectory control with test-time domain adaptation:** We propose Mask Normalization with Temporal Intrinsic Denoising to achieve precise trajectory control at test-time. Provided with bounding box sequences (left panels) and subject-specific captions (bottom), our framework generates consistent, user-conditioned and high-quality video frames without any finetuning.

Abstract

Text-conditioned diffusion models have emerged as powerful tools for video synthesis, yet enabling Interactive Video Generation (IVG), where users explicitly control object trajectories, remains challenging. While recent training-free approaches utilize attention masking for guidance, they often trade off perceptual quality for control. In this work, we identify the root causes of this degradation as two distinct domain shifts: **1)** internal covariate shift *induced by applying masks to pretrained models, and* **2)** an initialization gap *where random noise lacks alignment with trajectory conditions. We propose a test time domain adaptation framework to resolve these shifts. To this end, we first introduce Mask Normalization, a pre-normalization layer that mitigates* **(1)**, *i.e., covariate shift via feature distribution alignment. Next, a Temporal Intrinsic Prior that enforces spatio-temporal consistency during denoising is introduced to bridge the initialization gap, thus addressing* **(2)**. *Extensive evaluations on popular dataset demonstrate that our approach outperforms the state-of-the-art IVG methods in both perceptual quality and trajectory adherence.*

1. Introduction

Interactive Video Generation (IVG) enables users to control subject placement across frames in text-to-video models by specifying bounding boxes [3, 4]. However, balancing precise trajectory control with temporal consistency remains a significant challenge. While training-based methods of-

fer robust guidance, they suffer from prohibitive computational costs and data dependency [6, 7, 11, 13]. In contrast, training-free solutions [3, 4, 9] leverage attention masking and noise initialization but typically sacrifice perceptual quality for control. In this work, we identify the pitfalls of current training-free approaches and propose a framework to simultaneously enhance perceptual quality and control.

In this paper, we hypothesize that such pitfalls stems from two specific domain shifts overlooked in masked diffusion: ***(i)* Internal Covariate Shift**: Since base models are not trained with masked attention, forcing masks at test time fundamentally alters activation statistics, creating a distributional mismatch between the masked outputs and the features the model expects. ***(ii)* Initialization Gap**: This is due to signal leakage at train time, where the model implicitly expects structurally informative latents [8]. Standard inference, however, initializes with pure Gaussian noise, resulting in training-inference domain gap. Therefore, we ask: *Can we interpret, and ultimately improve, both control and perceptual quality in masked video diffusion models via test time domain adaptation?*

To bridge these gaps, we introduce a training-free approach. First, to address covariate shift, we propose **Mask Normalization**, a novel pre-normalization layer that aligns the feature distributions of masked attention outputs with those of unmasked priors. Second, to address the initialization gap, we employ **Temporal Intrinsic Denoising**. Inspired by Deep Image Prior (DIP) [5], this method leverages the model's intrinsic diffusion prior, guided by a classifier-conditioned temporal prior, to enforce spatio-temporal coherence explicitly. Our approach maintains control while preserving the perceptual quality (see Fig. 1). Results suggests that our approach outperforms the existing training-free IVG methods, when tested on popular dataset.

2. Method

First, we introduce masking-based trajectory control and analyze how it induces a variance shift in activations, thereby degrading perceptual quality. To address this, we subsequently propose mask normalization. Next, we uncover the limitation of the initialization gap at test time in video diffusion models and introduce a lightweight temporal prior to bridge it via temporal intrinsic denoising.

2.1. Preliminaries and Masking

Problem Setup. Given a text prompt y and a bounding box trajectory mask $b \in \{0,1\}^{N\times H\times W}$ (N $H \times W$ frames), our goal is to generate a video z_0. We assume a pre-trained, frozen T2V diffusion model (in our case, Zeroscope[1]).

Trajectory Control. For masking, we adhere to Peekaboo's [3] scheme. To enforce trajectory alignment, we constrain attention such that foreground tokens only attend to foreground regions defined by b. We generate binary masks for spatial self-attention (M_{self}) and cross-attention (M_{cross}):

$$M_{\text{self}}[i] = (M_v[i] \cdot M_v[i]^\top) + ((1 - M_v[i]) \cdot (1 - M_v[i])^\top)$$
$$M_{\text{cross}}[i] = (M_v[i] \cdot M_y[i]^\top) + ((1 - M_v[i]) \cdot (1 - M_y[i])^\top)$$

These masks are applied to the U-Net decoder layers only during the initial "frozen" denoising steps. After these steps, the masks are removed, and the denoising process continues normally.

2.2. Mask Normalization

Variance Shift Analysis. To investigate the impact of masking on feature statistics, we extracted activations from the transformer layer in the second decoder block of Zeroscope during inference. We fixed the trajectory control to the first four "frozen" diffusion steps and compared naive masking against a baseline (unmasked) generation. Our analysis reveals that applying masks to models trained without them induces severe internal covariate shift. As shown in Fig. 2, masked activations exhibit significantly lower variance than the baseline, followed by unstable spikes immediately after masking ends (step 5). This discrepancy suggests that standard LayerNorm is insufficient to handle the distribution shift caused by inference-time masking.

Mask Normalization. To resolve this, we propose Mask Normalization, a pre-normalization layer that aligns the statistics of masked attention outputs ($\mathcal{A}_m$) with unmasked priors ($\mathcal{A}_u$). Inspired by style transfer [12], we treat $\mathcal{A}_u$ as the target "content" distribution and $\mathcal{A}_u$ as the reference "style" distribution. We employ Exact Feature Distribution Matching (EFDM) [12], based on Exact Histogram Matching [1], to map the empirical Cumulative Distribution Function (CDF) of $\mathcal{A}_m$ to match $\mathcal{A}_u$:

$$\mathcal{A}_m = \sigma\left(\frac{QK^\top}{\sqrt{d}} \circ \mathcal{M}\right)V \quad \text{and} \quad \mathcal{A}_u = \sigma\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

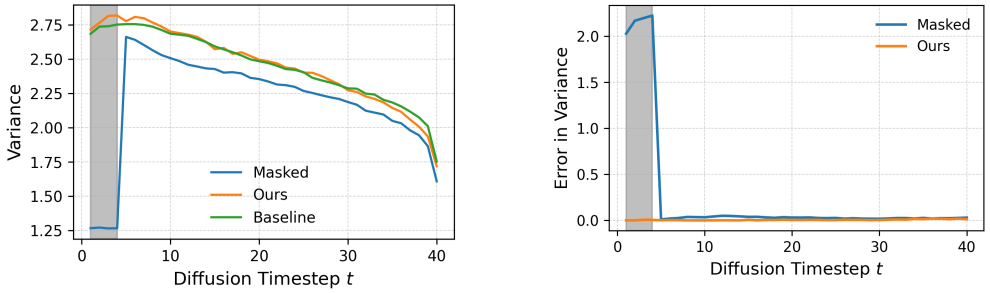
This process, illustrated in Fig. 3, effectively regularizes feature variance without retraining, ensuring the model operates within its expected activation statistics.

2.3. Temporal Intrinsic Denoising

An inherent problem of the *initialization gap* [8] persists with the Standard Gaussian initialization for trajectory-controlled generation, leading to the loss of spatial correlation information—often the critical ones.

Temporal Prior (τ). We explicitly enforce temporal consistency by maximizing the Pearson correlation (ρ) of foreground features across consecutive frames. We define the

[1] https://huggingface.co/cerspense/zeroscope_v2_576w



(a) Variance distribution curve (b) Variance error curve

Figure 2. **Effect of attention masking on activation variance.** (a) Masking (blue) causes a drop in variance compared to baseline (green), while our method (orange) aligns closely with the baseline. (b) Variance error increases progressively with masking, whereas our method maintains low error.

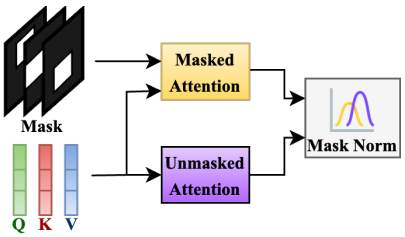


Figure 3. Mask Normalization aligns the empirical cumulative distributions of masked and unmasked attention outputs using EFDM.

consistency metric τ as:

$$\tau(\tilde{b}, z_t) = \frac{1}{N-1} \sum_{i=1}^{N-1} \rho\Big(c(u_t^i),\, c(u_t^{i+1})\Big), \qquad (1)$$

where, $c(\cdot)$ extracts foreground pixels u_t defined by bounding box b. Note that we use $\tilde{b}$ to denote that we derive explicit cross-token and cross-frame prior from b.

Temporal Intrinsic Denoising (TID). Inspired by Deep Image Prior [5], we treat the latent at step t as a "noisy image" to be refined towards the trajectory manifold. We modify the update rule of Xiao et al. [10] to integrate classifier guidance into intrinsic denoising. The update step that we call Temporal Intrinsic Denoising is repeated M times before each diffusion sampling step, and is given as:

$$z_t^m = z_t^{m-1} + \eta_l \Big[\underbrace{c_g \nabla_z \tau\big(\tilde{b},\, \hat{z}_{0,t}^{m-1}\big)}_{\text{classifier guidance}} + \underbrace{\nabla_z \log p_t\big(z_t^{m-1} \mid y, b\big)}_{\text{classifier-free guidance}} \Big] + \eta_k \epsilon_k^m. \qquad (2)$$

Here, $\hat{z}_{0,t}$ is the one-shot clean approximation computed via Tweedie's formula [2]. The pseudocode is outlined in Algorithm 1. This process steers the latent towards high-probability regions consistent with the user's trajectory control. Note that setting $c_g = 0$ corresponds to **Intrinsic Denoising (ID)**.

Algorithm 1 Diffusion Sampling with TID

for $t = T$ **downto** 1 **do**
  $z_t^0 \leftarrow z_t$
  for $m = 1$ **to** M **do**
    $\mathcal{G}_g \leftarrow \nabla_z \tau\big(\tilde{b},\, \hat{z}_{0,t}^{m-1}\big)$
    $\mathcal{G}_p \leftarrow \nabla_z \log p_t\big(z_t^{m-1} | y, b\big)$
    $z_t^m \leftarrow z_t^{m-1} + \eta_l \big[c_g \mathcal{G}_g + \mathcal{G}_p\big] + \eta_k \epsilon_k^m$
  end for
  $z_{t-1} \leftarrow \text{DDIMStep}(z_t^M, y, b, t)$
end for
Output: z_0

3. Experiments and Results

Following prior work [3, 4], our experiments are conducted on Zeroscope. All experiments were performed on a single NVIDIA A100 GPU. We compare Mask Normalization and Temporal Intrinsic Denoising, against Peekaboo [3] and Trailblazer [4], evaluated on randomly generated bounding boxes (static and dynamic), as proposed by Jain et al. [3].

Quantitative Results. As shown in Tab. 1, Peekaboo achieves high coverage but poor spatial precision (low mIoU), indicating the subject often drifts outside the designated path. Trailblazer offers strong dynamic control but degrades significantly on static trajectories and semantic fidelity (lowest CLIP-SIM). Our method (Mask Norm + TID) balances these trade-offs, achieving the highest static mIoU (33.82%) and competitive dynamic performance without sacrificing semantics. This is reinforced by Tab. 2, where we achieve the best perceptual quality (lowest FID/KID). We note that the deceptively low JeDi scores for baselines reflect their tendency to ignore "unnatural" trajectory constraints in favor of natural motion priors, whereas our method enforces the user's specific intent.

Table 1. **Conditional evaluation.** Best results in green, second-best in blue. While baselines trade off semantic quality (CLIP-SIM) for control (mIoU), our full method (Mask Norm + TID) achieves consistent performance across both static and dynamic trajectories.

Model	Static			Dynamic		
	CLIP-SIM↑	CoV↑	mIoU↑	CLIP-SIM↑	CoV↑	mIoU↑
Peekaboo [3]	31.47	40	29.92%	31.55	45	29.76%
Trailblazer [4]	31.50	35	21.18%	30.03	42	36.01%
Mask Norm (Ours)	31.66	38	22.09%	30.64	32	16.27%
Mask Norm + ID (Ours)	32.06	38	25.97%	32.17	49	22.32%
Mask Norm + TID (Full)	31.48	41	33.82%	31.80	46	34.73%

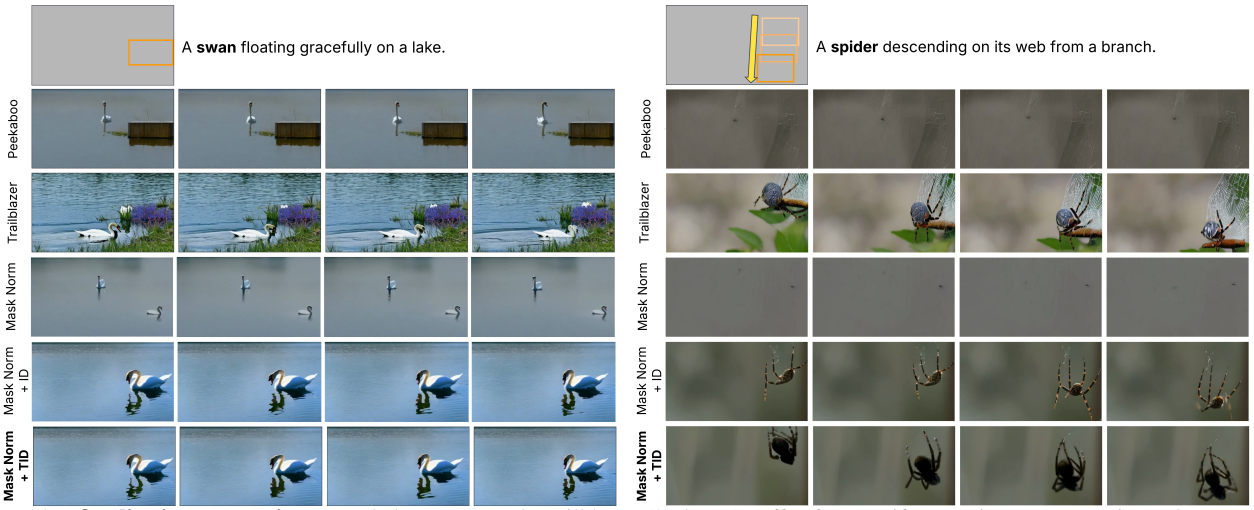


Figure 4. **Qualitative comparison.** Peekaboo (top) and Trailblazer (2nd row) suffer from artifacts or ignore constraints. Our method (bottom) harmonizes photorealism with precise trajectory adherence.

Table 2. **Unconditional evaluation.** Our method maintains high perceptual quality (low FID/KID). †As noted in text, low JeDi scores for baselines often indicate a lack of adherence to unnatural trajectory constraints rather than superior quality.

Model	FID↓	KID↓	JeDi† ↓
Peekaboo [3]	134.30	2.03% ± 0.07	1.40
Trailblazer [4]	140.07	2.18% ± 0.07	1.65
Mask Norm	151.99	2.77% ± 0.07	1.46
Mask Norm + ID	133.74	1.81% ± 0.06	1.34
Mask Norm + TID	131.19	1.92% ± 0.06	1.86

Qualitative Results. Visual inspection of Fig. 4 corroborates our quantitative findings. Peekaboo (top row) often generates the subject with significant artifacts outside the specified region, particularly in dynamic settings. Trailblazer (second row), while adhering to dynamic paths, suffers from mode collapse, producing oversaturated and unrealistic textures due to high guidance scales. In contrast, our method (bottom row) demonstrates robust spatio-temporal coherence, effectively grounding the subject to the bounding box sequence without compromising the visual integrity or natural statistics of the pre-trained diffusion priors. Additional qualitative results are shown in Fig. 1.

Ablation Study. The necessity of our components is evident in Tab. 1 and Fig. 4. Mask Normalization alone stabilizes variance but fails to anchor the object spatially. Adding Intrinsic Denoising (ID) improves texture quality but lacks temporal coherence. The full integration of the Temporal Prior (TID) is critical, yielding a ∼8-12% gain in mIoU by enforcing cross-frame consistency.

4. Conclusion

This work offers a principled approach to trajectory-controlled IVG via the lens of test time domain adaptation. By identifying internal covariate shift and the initialization gap as key barriers, we introduced Mask Normalization and Temporal Intrinsic Denoising to enable precise, training-free trajectory control. Our approach successfully bridges the gap between user constraints and diffusion priors. Future work aims to extend our approach to multi-subject scenarios and experiment with DiT.

References

[1] Dinu Coltuc, Philippe Bolon, and J-M Chassery. Exact histogram specification. *IEEE TIP*, 15(5):1143–1152, 2006. 2

[2] Bradley Efron. Tweedie's formula and selection bias. *J. Am. Stat. Assoc.*, 106(496):1602–1614, 2011. 3

[3] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *CVPR*, pages 8079–8088, 2024. 1, 2, 3, 4

[4] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia*, pages 1–11, 2024. 1, 2, 3, 4

[5] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, pages 9446–9454, 2018. 2, 3

[6] 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. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[7] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 2

[8] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *ECCV*, pages 378–394. Springer, 2024. 2

[9] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, pages 378–394. Springer, 2024. 2

[10] Jie Xiao, Ruili Feng, Han Zhang, Zhiheng Liu, Zhantao Yang, Yurui Zhu, Xueyang Fu, Kai Zhu, Yu Liu, and Zheng-Jun Zha. Dreamclean: Restoring clean image using deep diffusion prior. In *ICLR*, 2024. 3

[11] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[12] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *CVPR*, pages 8035–8045, 2022. 2

[13] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024. 2