FreeAction: Training-Free Techniques for Enhanced Fidelity of Trajectory-to-Video Generation

Seungwook Kim[1] Seunghyeon Lee[2] Minsu Cho[1,3]
POSTECH[1] Ewha Womans University[2] RLWRLD[3]

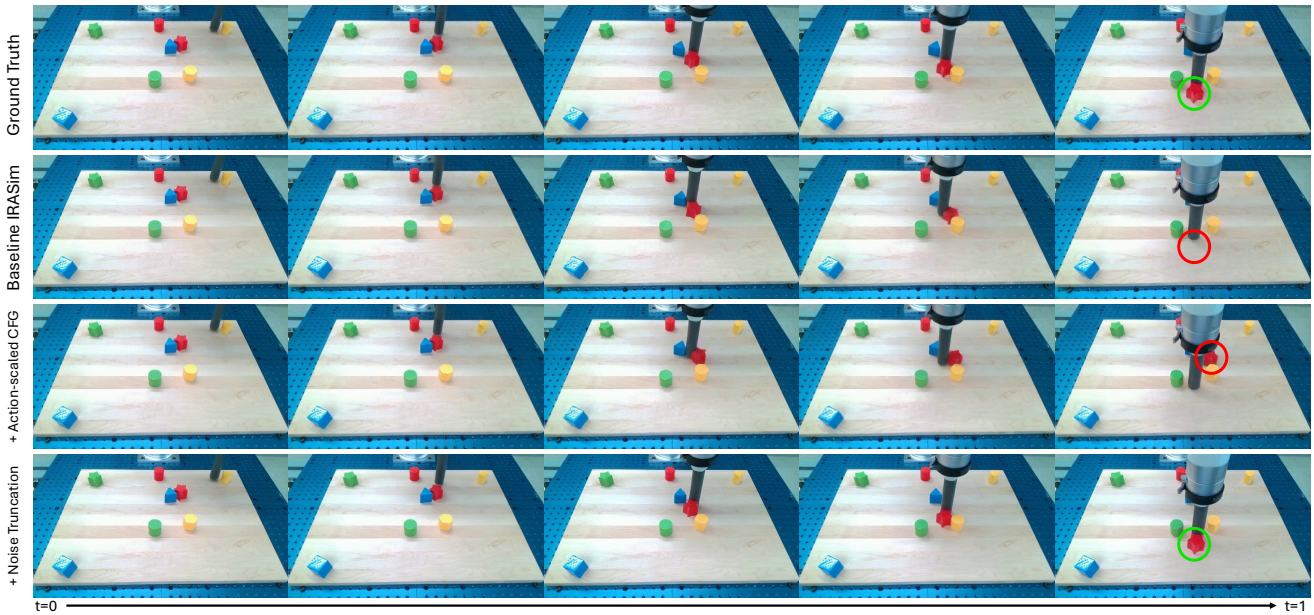

Figure 1. **Results on the LanguageTable dataset**. It can be seen that the red star block disappears in the baseline. By integrating both our action-scaled CFG and noise truncation, it can be seen that the final output is visually identical to the ground-truth video.

Abstract

Generating realistic robot videos from explicit action trajectories is a critical step toward building effective world models and robotics foundation models. We introduce two training-free, inference-time techniques that fully exploit explicit action parameters in diffusion-based robot video generation. Instead of treating action vectors as passive conditioning signals, our methods actively incorporate them to guide both the classifier-free guidance process and the initialization of Gaussian latents. First, action-scaled classifier-free guidance dynamically modulates guidance strength in proportion to action magnitude, enhancing controllability over motion intensity. Second, action-scaled noise truncation adjusts the distribution of initially sampled noise to better align with the desired motion dynamics. Experiments on real robot manipulation datasets demonstrate that these techniques significantly improve action coherence and visual quality across diverse robot environments.

1. Introduction

Generating robot videos conditioned on explicit action trajectories is a crucial component for building effective world models and robotics foundation models [4, 17, 18, 21]. By simulating robot trajectories across diverse environments and robots, such systems not only facilitate rapid prototyping and validation of control policies, but can also be used to generate rich synthetic data for downstream tasks such as policy learning and zero-shot transfer [5, 6, 10, 12, 19].

A recent line of work builds on diffusion-based video generative models to produce robot videos with high visual fidelity [7, 21]. However, such existing trajectory-conditioned models [21] treat action trajectories as passive conditioning signals *e.g.,* injecting the action trajectories as conditions into the diffusion process via the Ada-LN layers [11] in a Diffusion Transformer (DiT)[15] architecture. In this paper, we posit that action trajectories are currently underutilized and can be leveraged more effectively to enhance trajectory coherence in generated robot videos.

To address this gap, we present two training-free,

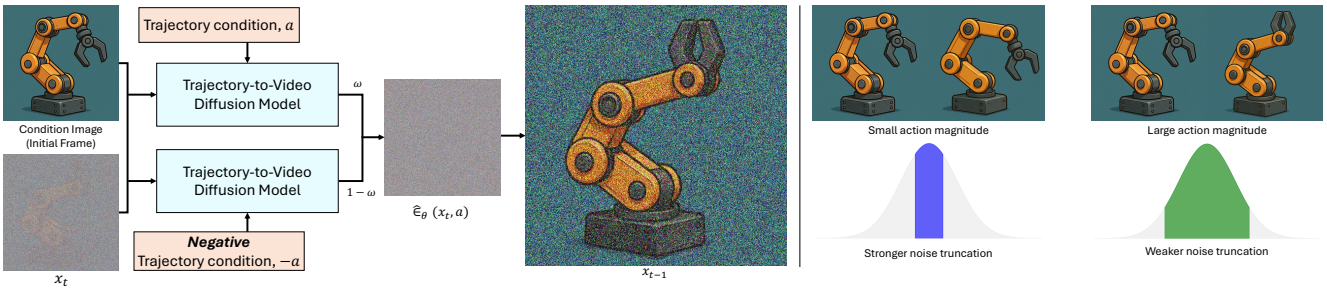

Figure 2. **Illustration of our training-free inference-time techniques**. (Left) Our action-scaled classifier-free guidance to steer the generation away from the negative trajectory, and towards our desired trajectory. We set $\omega = \lambda\|a\|_2 \mathbf{1}_{\{t>T/2\}}$ so that the guidance strength varies in proportion to the action magnitude, and the guidance is applied only in the earlier steps. (Right) Our action-scaled noise truncation scheme, where we manipulate the strength of truncation based on the magnitude of action.

inference-time techniques that actively exploit action trajectory information to steer the diffusion process toward improved action coherence in generated videos. The first modulates classifier-free guidance strength according to action magnitude, while the second manipulates initial noise sampling to enhance the fidelity and determinism of the trajectory-to-video generation process, reflecting the intended motion dynamics.

Our contributions are as follows:

- We propose *action-scaled classifier-free guidance*, which dynamically adjusts the guidance weight in proportion to the action magnitude, improving trajectory coherence.
- We introduce *action-scaled noise truncation* that modifies the initially sampled Gaussian noise to better align generated motion with the specified action parameters.
- We demonstrate on three real-robot manipulation datasets that our methods yield substantial improvements in both action fidelity and visual quality.

2. Action-scaled CFG and Noise Truncation

Preliminary: IRASim. IRASim [21] is a DiT [14]-based action trajectory-to-video generative diffusion model, which functions as an interactive real-robot simulator. In essence, given an initial frame I^1 and an action trajectory $a^{1:N-1}$ as conditions, IRASim generates $N-1$ future frames $I^{1:N}$ which reflect the results of the action trajectory on the initial frame. IRASim injects action trajectories as conditions via the AdaLN layers [11]. In the following, we introduce how we analyze and manipulate the conditioning and noise initialization mechanisms in the inference pipeline of IRASim to improve the visual quality and action coherency of generated videos.

2.1. Action-scaled Classifier-free Guidance

Classifier-Free Guidance (CFG) [8] is a standard method for improving the fidelity of diffusion-generated outputs. In text-to-image models, CFG is implemented by randomly dropping text prompts during training, enabling the diffusion model to act as both a conditional and an unconditional denoiser. Given a text prompt c, and an unconditional token $\varnothing$, the standard CFG update at step t is as follows:

$$\hat{\epsilon}\theta(x_t, c) = \epsilon\theta(x_t, \varnothing) + \omega\left[\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing)\right], \quad (1)$$

where $\omega = 0$ recovers unconditional prediction, $\omega = 1$ recovers purely conditional prediction, and $\omega > 1$ amplifies conditioning. However, this "prompt-dropping" strategy substantially increases the training complexity [1, 9], and directly extending prompt-dropping to action-to-video models introduces semantic inconsistencies, since a zero action should represent no motion rather than a neutral prior.

To overcome these challenges, we propose **action-scaled classifier-free guidance**. Instead of learning an unconditional denoiser, we synthesize a *negative* action condition by negating the input action, *i.e.,* $-a$. Concretely, if $\epsilon_\theta(x_t, a)$ denotes the predicted noise given action a, and $\epsilon_\theta(x_t, -a)$ the prediction given the negated action, our guided update at timestep t takes the form:

$$\hat{\epsilon}_\theta(x_t, a) = \epsilon_\theta(x_t, a) + \omega\left[\epsilon_\theta(x_t, a) - \epsilon_\theta(x_t, -a)\right]. \quad (2)$$

Intuitively, this encourages samples to move closer to the trajectory specified by a and farther from its negation.

However, two practical issues remain in the above formulation: (1) small-magnitude actions should result in proportionally weaker repulsion from the negative condition, and (2) the high-frequency texture details, which are largely independent of motion, should be preserved. We address both by scaling the guidance weights as $\omega = \lambda\|a\|_2 \mathbf{1}_{\{t>T/2\}}$, where $|a|_2$ is the ℓ_2-norm of the action vector, t is the current step, and T is the total number of sampling steps. λ is a hyperparameter[1]. This formulation increases guidance strength for larger actions, minimizes perturbations for small ones, and applies the effect only in early steps (low-frequency generation) to maintain texture fidelity in later refinements. We illustrate our action-scaled classifier-free guidance in the left of Fig. 2.

[1] We set $\lambda = 1$ in our experiments.

Table 1. **Quantitative results on RT-1, Bridge, and Language-Table datasets.** Across both short- and long-trajectory evaluation settings, using our test-time techniques improves results compared to the baseline IRASim [21].

Dataset	Method	PSNR↑	SSIM↑	Latent L2↓
Short-trajectory				
RT-1 [3]	IRASim-Frame-Ada	26.024	0.833	0.2100
	+ Action-scaled CFG	<u>26.198</u>	<u>0.837</u>	<u>0.2068</u>
	+ Action-scaled Noise Truncation	**26.435**	**0.840**	**0.1629**
Bridge [16]	IRASim-Frame-Ada	25.340	0.834	0.1939
	+ Action-scaled CFG	<u>25.398</u>	<u>0.835</u>	<u>0.1938</u>
	+ Action-scaled Noise Truncation	**25.770**	**0.843**	**0.1662**
Language-Table [13]	IRASim-Frame-Ada	28.794	0.888	0.1663
	+ Action-scaled CFG	<u>29.021</u>	<u>0.890</u>	<u>0.1653</u>
	+ Action-scaled Noise Truncation	**29.514**	**0.902**	**0.1326**
Long-trajectory				
RT-1 [3]	IRASim-Frame-Ada	21.729	0.760	0.2408
	+ Action-scaled CFG	<u>21.984</u>	<u>0.763</u>	<u>0.2355</u>
	+ Action-scaled Noise Truncation	**22.299**	**0.776**	**0.1891**
Bridge [16]	IRASim-Frame-Ada	21.536	0.769	0.2306
	+ Action-scaled CFG	<u>21.601</u>	<u>0.771</u>	<u>0.2302</u>
	+ Action-scaled Noise Truncation	**22.035**	**0.785**	**0.1961**
Language-Table [13]	IRASim-Frame-Ada	24.861	0.852	0.1730
	+ Action-scaled CFG	<u>24.920</u>	<u>0.853</u>	<u>0.1719</u>
	+ Action-scaled Noise Truncation	**25.120**	**0.867**	**0.1394**

2.2. Action-scaled Noise Truncation

In conventional text-to-video generation, models are expected to balance fidelity and diversity. In contrast, in the current task of trajectory-to-video, the goal is a single faithful realization of the given specific action trajectory a. We therefore propose to regulate the diversity of generated videos by manipulating the initial Gaussian latent $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, via action-scaled noise truncation, in the spirit of BigGAN's truncation trick [2]. This truncation constrains sampling to high-density regions, yielding higher fidelity while sacrificing diversity - a trade-off that is desirable for trajectory-to-video.

We implement min–max truncation with a sigmoid mapping from action magnitude to an element-wise truncation limit, centered at the dataset mean:

$$\tau(a) = \tau_{\min} + (\tau_{\max} - \tau_{\min})\,\sigma\big(\|a\|_2 - \mu_{\text{act}}\big), \tag{3}$$

where $\sigma()$ is the sigmoid function, μ_{act} is the dataset mean of $\|a\|_2$[2] and $0 < \tau_{\min} \leq \tau_{\max}$[3]. We then draw the initial latent by sampling each entry independently from a zero-mean truncated normal with limit $\tau(a)$:

$$z_{0,i} \sim \mathcal{N}(0,1) \text{ conditioned on } |z_{0,i}| \leq \tau(a) \quad \forall i$$

[2]We pre-calculate the mean each dataset from their train set.
[3]We set $\tau_{\min} = 0.5$ and $\tau_{\max} = 1.5$ in our experiments.

Small actions ($\|a\|_2 \ll \mu_{\text{act}}$) induce strong truncation ($\tau \approx \tau_{\min}$) to aim for nearly deterministic appearance; large actions relax truncation ($\tau \approx \tau_{\max}$) to retain the variation needed to express substantial motion and appearance change from the previous frame. This plug-and-play modification requires no additional training. We illustrate our action-scaled noise truncation in the right of Fig. 2.

3. Experimental Results

Evaluation setup. Following IRASim [21], we evaluate our method on the real-robot manipulation datasets of (1) RT-1 [3] with 7 DoF action space, (2) Bridge [16] with 7 DoF action space, and (3) LanguageTable [13] with 2 DoF action space. In short-trajectory evaluation (single generation pass), we use a single frame as reference and 15 actions to predict the next 15 frames. For the long-trajectory evaluation, we use a single frame as reference, and generate the video autoregressively where the final frame from the previous pass serves as the reference frame for the next pass. Noting that Latent L2 loss and PSNR best align with human preferences [21], we report Latent L2, PSNR and additionally SSIM to evaluate the performance of our method.

Results and analyses. We show the results of our two training-free techniques in Table 1, on both the short- and long-trajectory evaluation settings.

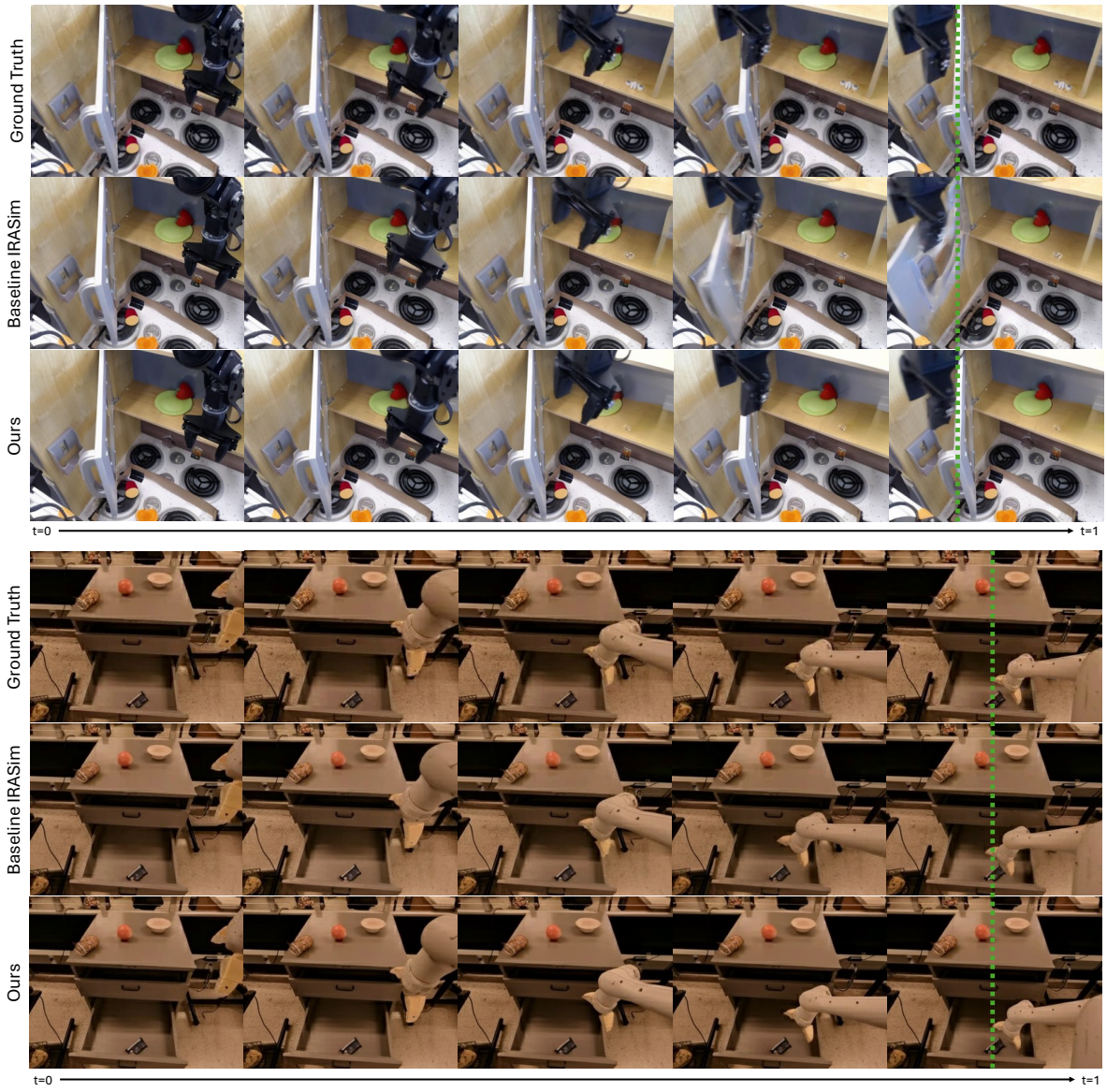

Figure 3. **Qualitative results on the Bridge (top) and RT-1 (bottom) datasets**. Applying our techniques improves coherency to the action trajectory. Note the green line, where it can be seen that ours generates results which better resemble the ground truth video.

It can be seen that our proposed techniques, action-scaled CFG and action-scaled noise truncation shows consistent improvements over the baseline IRASim-Frame-Ada on all metrics of PSNR, SSIM and Latent L2, *i.e.*, effectively achieving higher perceptual quality and fidelity on both settings of short- and long-trajectory. We also provide qualitative results on the Language-Table dataset in Figure 1, and on the Bridge and RT-1 datasets in Figure 3. In both figures, we visually show that while the baseline method already shows photorealistic outputs, applying our techniques leads to improved action coherence, showing better similarity with the ground-truth video. We provide additional experiments and analyses in the appendix.

4. Discussion and Future Direction

The proposed action-scaled classifier-free guidance and action-scaled noise truncation are training-free, inference-time controls that improve visual fidelity and action coherence in diffusion-based trajectory-to-video generation. We demonstrated the efficacy of our techniques over real-robot RT-1, Bridge and Language-Table datasets. We expect that these techniques will enable trajectory-conditioned generators to generate more coherent rollouts without having to modify or re-train the base model. Looking ahead, we will investigate effective techniques in expanded conditioning regimes - *e.g.*, models that accept *natural language* instructions or additional modalities such as depth.

Acknowledgement. This work was supported by the IITP grants (RS-2022-II220959: Few-Shot Learning of Causal Inference in Vision and Language for Decision Making (50%), RS-2025-25443730: Research and Evaluation Framework for Robotic Artificial General Intelligence with Broad Generalization Capabilities (45%), RS-2019-II191906: AI Graduate School Program at POSTECH (5%)) funded by the Korea government (MSIT).

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 2

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 3, 7

[4] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023. 1

[5] Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Generalist bimanual manipulation via foundation video diffusion models. *arXiv preprint arXiv:2507.12898*, 2025. 1

[6] Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control. *arXiv preprint arXiv:2506.01943*, 2025. 1

[7] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024. 1, 6

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[9] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023. 2

[10] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024. 1

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 1, 2

[12] Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025. 1

[13] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023. 3, 7

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2

[15] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[16] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023. 3, 7

[17] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024. 1

[18] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 1

[19] Xiuyu Yang, Bohan Li, Shaocong Xu, Nan Wang, Chongjie Ye, Zhaoxi Chen, Minghan Qin, Yikang Ding, Xin Jin, Hang Zhao, et al. Orv: 4d occupancy-centric robot video generation. *arXiv preprint arXiv:2506.03079*, 2025. 1

[20] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020. 6

[21] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2403.13206*, 2024. 1, 2, 3

A. Comparative analysis

We compare *action-scaled* CFG and noise truncation against their *fixed-weight* counterparts. As summarized in Table 2, the action-scaled variants achieve the best overall performance across datasets and metrics. Specifically, using a fixed CFG weight ($\omega = \text{const}$) can degrade results, likely due to miscalibrated steering, *i.e.*, the guidance becomes too strong for small actions and too weak for large ones. A fixed truncation threshold ($\tau = \text{const}$) consistently improves upon the IRASim baseline by reducing spurious variability in the initial latent; however, scaling the truncation $\tau(\mathbf{a})$ with the action norm yields the strongest gains by enforcing determinism for small motions while preserving necessary diversity for large motions.

B. Application to Visual Policy Learning

We test whether our insights transfer beyond trajectory-to-video generation. Specifically, our approach regulates sample diversity to improve fidelity; here, we evaluate the *noise truncation* component on a visual policy learning task using Prediction with Action (PAD) [7] as the baseline. Because we could not find an open-source visual policy framework that conditions on action trajectories, and PAD conditions on natural language instructions instead, we disable CFG and evaluate a fixed-threshold truncation ($\tau = 1.0$).

Experimental setup (PAD). PAD [7] is a DiT-based diffusion model that jointly predicts future visual observations and robot actions. We evaluate on the MetaWorld benchmark [20], reporting results both *including* and *excluding* `handle-pull-v2`, which is known to be unusually difficult[4]. PAD supports three generative modalities—image, depth, and action—each initialized element-wise from $\mathcal{N}(0, 1)$. We apply fixed-threshold noise truncation to different modality subsets: (1) Image; (2) Action+Depth+Image; (3) Depth+Image.

Results and analysis. Table 3 shows that truncating the initial noise for the *image* or *depth* modalities consistently improves success rates over the PAD baseline. In contrast, truncating the *action* modality harms performance: unlike image/depth, which are strongly correlated across frames, action vectors are not reliably autocorrelated[5], so reducing action-sampling variability can over-constrain the next-action prediction. Overall, trading off diversity for fidelity in modalities with strong temporal correlation (image, depth) yields gains beyond trajectory-to-video generation, and—even without modifying the action channel—the increased determinism in image/depth positively influences action prediction, improving task success. Qualitative examples in Fig. 4 illustrate more precise placements (e.g., moving the cup to the green target) when truncation is applied.

[4] Baseline PAD yields 0% success on `handle-pull-v2`.

[5] States (e.g., positions) are temporally correlated, but the *actions* that transition between them need not be.

Table 2. **Ablation experiments on RT-1, Bridge, and Language-Table datasets.**

Dataset	Method	PSNR↑	SSIM↑	Latent L2↓
vs. CFG using fixed guidance weight ω				
RT-1 [3]	$\omega = 1.0$ (baseline)	26.024	0.833	0.2100
	$\omega = 3.0$	26.052	0.832	0.2112
	Ours ($\omega = \lambda \|a\|_2 \mathbf{1}_{\{t>T/2\}}$)	26.198	0.837	0.2068
Bridge [16]	$\omega = 1.0$ (baseline)	25.340	0.834	0.1939
	$\omega = 3.0$	25.175	0.830	0.1950
	Ours ($\omega = \lambda \|a\|_2 \mathbf{1}_{\{t>T/2\}}$)	25.398	0.835	0.1938
Language-Table [13]	$\omega = 1.0$ (baseline)	28.794	0.888	0.1663
	$\omega = 3.0$	28.922	0.890	0.1654
	Ours ($\omega = \lambda \|a\|_2 \mathbf{1}_{\{t>T/2\}}$)	29.021	0.890	0.1653
vs. Fixed truncation level $\tau(a)$				
RT-1 [3]	$\tau(a) = 1.0$	26.370	0.840	0.1669
	$\tau(a) = 1.5$	26.236	0.839	0.1892
	Ours (Equation 3)	26.435	0.840	0.1629
Bridge [16]	$\tau(a) = 1.0$	25.782	0.841	0.1690
	$\tau(a) = 1.5$	25.487	0.838	0.1766
	Ours (Equation 3)	25.770	0.843	0.1662
Language-Table [13]	$\tau(a) = 1.0$	29.488	0.899	0.1369
	$\tau(a) = 1.5$	29.122	0.892	0.1459
	Ours (Equation 3)	29.514	0.902	0.1326

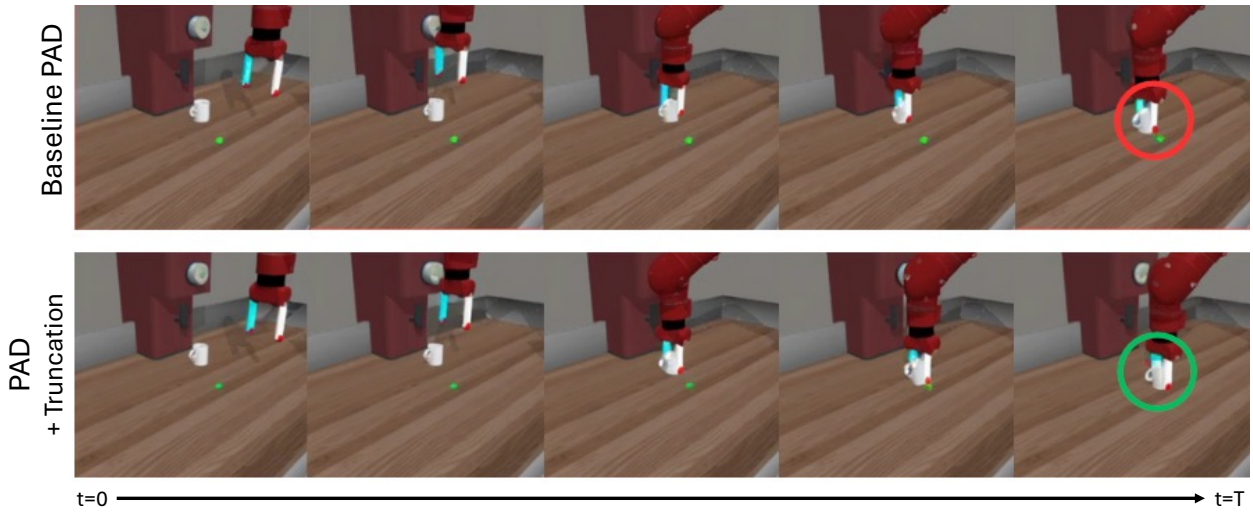

Figure 4. **Qualitative example of PAD with truncation on the Metaworld benchmark.**. The figure compares baseline PAD and PAD with truncation in the `coffee-pull-v2` task, where the cup is expected to be placed at the green dot to be counted as a success. While the baseline often fails to place the coffee cup at the target green dot, PAD with truncation successfully reaches the target, demonstrating more consistent and accurate manipulation over time.

Table 3. **Success rate comparison under different input modality configurations on the MetaWorld benchmark**, with and without the handle-pull task (50-task benchmark).

Configuration	**w/o handle-pull** (%)	**w/ handle-pull** (%)
Baseline	70.00	67.20
Truncation at Image	70.42	67.60
Truncation at Action, Depth, Image	69.38	66.60
Truncation at Depth, Image	**71.46**	**68.60**