TOWARDS ONE-STEP CAUSAL VIDEO GENERATION VIA ADVERSARIAL SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent hybrid video generation models combine autoregressive temporal dynamics with diffusion-based spatial denoising, but their sequential, iterative nature leads to error accumulation and long inference times. In this work, we propose a distillation-based framework for efficient causal video generation that enables high-quality synthesis with extreme limited denoising steps. Our approach builds upon Distribution Matching Distillation (DMD) framework and proposes a novel form of Adversarial Self-Distillation (ASD) strategy, which aligns the outputs of the student model's n-step denoising process with its $(n+1)$-step version in the distribution level. This design provides smoother supervision by bridging small intra-student gaps and more informative guidance by combining teacher knowledge with locally consistent student behavior, substantially improving training stability and generation quality in extremely few-step scenarios. In addition, we present a First-Frame Enhancement (FFE) strategy, which allocates more denoising steps to the initial frames to mitigate error propagation while applying larger skipping steps to later frames. Extensive experiments on VBench demonstrate that our method surpasses state-of-the-art approaches in both one-step and two-step video generation. Notably, our framework produces a single distilled model that flexibly supports multiple inference-step settings, eliminating the need for repeated re-distillation and enabling efficient, high-quality video synthesis.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021a; Nichol & Dhariwal, 2021) have achieved remarkable progress in high-quality image (Dhariwal & Nichol, 2021; Nichol et al., 2022; Rombach et al., 2022) and video generation (Ho et al., 2022; Brooks et al., 2024; Blattmann et al., 2023). However, their application to long-form video (Chen et al., 2023; Zhang et al., 2023) and interactive settings (Che et al., 2025) remains limited by the high computational cost of iterative denoising and the reliance on bidirectional attention, which requires synthesizing entire sequences jointly. Autoregressive methods (Liang et al., 2022; Ge et al., 2022; Wang et al., 2024), in contrast, allow causal frame-by-frame generation, but often suffer from error accumulation that compromises realism. Recent works (Chen et al., 2024; Jin et al., 2025; Li et al., 2025) have attempted to combine autoregressive temporal modeling with diffusion-based spatial refinement, yet these hybrid approaches inherit the efficiency bottleneck of multi-step denoising.

To alleviate this limitation, distillation has emerged as a promising direction for accelerating diffusion models (Sauer et al., 2024b; Yin et al., 2024b; Lu et al., 2025). By reducing a multi-step model to a few-step counterpart, distillation greatly improves efficiency while preserving generation quality. Nonetheless, existing distillation objective primarily focuses on aligning the prediction distributions of the few-step student with those of the multi-step teacher. When the student performs only 1- or 2-step generation, the discrepancy from the multi-step teacher becomes excessively large, making direct alignment unstable and causing severe quality degradation (Cheng et al., 2025; Yin et al., 2025; Huang et al., 2025). In other words, the fewer the denoising steps, the larger the semantic and statistical gap to be bridged, which explains why extremely few-step distillation is particularly challenging.

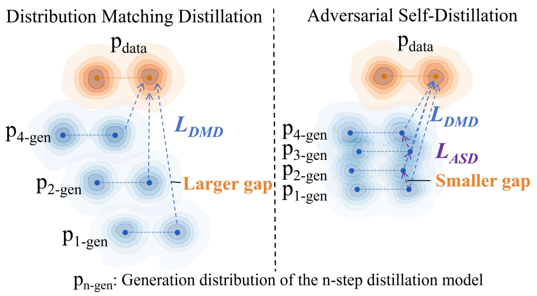


Figure 1: Conceptual illustration of the different alignment strategies of DMD and ASD during distillation. The points plotted in the figure indicate the multiple modes (or peaks) of the real data distribution, highlighting its multimodal nature.

In this paper, we introduce a novel method to address the challenge of high-quality video generation with minimal denoising steps (e.g. 1 and 2 steps). We extend the distribution matching distillation (DMD) framework (Yin et al., 2024b;a), and propose a novel Adversarial Self-Distillation (ASD) strategy. Unlike prior distillation methods (Song et al., 2023; Song & Dhariwal, 2024; Lin et al., 2024) that rely solely on supervision from the multi-step teacher, ASD provides the student with additional guidance from its own intermediate variants with slightly different denoising steps. Specifically, a discriminator is employed to adversarially match the n-step and $(n{+}1)$-step denoising distribution. As shown in Fig. 1, this step-wise self-alignment has two advantages: (1) it produces smoother supervision by bridging smaller step-to-step gaps instead of the large teacher–student gap, and (2) it yields more informative signals, since the student learns from both global teacher knowledge and its own locally consistent behavior. Together, these factors substantially enhance training stability and generation quality under extreme few-step settings. Fig. 2 also demonstrates that our method substantially reduces the performance degradation associated with a decreasing number of inference steps. Moreover, our framework introduces a **step-unified** design: instead of training a separate distilled model for each desired step size, a single student model trained with ASD can flexibly support multiple inference-step configurations at deployment. This property significantly improves practical usability, as it removes the need for repetitive re-distillation, which is especially valuable in scenarios such as dynamically balancing speed and quality trade-offs and accommodating deployment across varied resource settings.

To further improve video fidelity, we first conduct empirical analysis and observe that later frames exhibit higher generative redundancy compared to the first frame. Motivated by this, we propose a frame-wise inference strategy with varying denoising strengths. Unlike previous methods that treat all frames equally, we allocate more denoising steps to the crucial initial frames to mitigate error accumulation, while later frames are generated with larger skipping steps. This First-Frame Enhancement (FFE) strategy maintains a low overall computational cost while notably improving visual quality. Extensive experiments on VBench demonstrate that our method surpasses state-of-the-art approaches in both one-step and two-step video generation. Importantly, it achieves both efficiency and flexibility by using a single distilled model to support a wide range of inference settings.

In summary, our contributions are as follows:

- We propose a unified adversarial self-distillation strategy that aligns prediction distributions across different denoising steps of the student model, significantly improving few-step generation quality.
- We propose a frame-wise inference strategy that allocates more denoising steps to crucial initial frames, reducing error accumulation and improving video fidelity.
- Our experiments demonstrate that our method surpasses the state-of-the-art in few-step generation quality while eliminating the need for separate distillation training for each desired step size.

2 Related Work

2.1 Autoregressive/Diffusion Video Generation

Diffusion (Blattmann et al., 2023; Ho et al., 2022; Yang et al., 2025) and autoregressive models (Ge et al., 2022; Kondratyuk et al., 2024; Yu et al., 2024) are the two dominant paradigms in video generation due to their ability to produce high-quality results. Diffusion models employ bidirectional attention (Bao et al., 2022; Peebles & Xie, 2023) to denoise and synthesize all frames simultaneously, achieving strong temporal consistency but preventing frame-wise editing or interactive generation. In

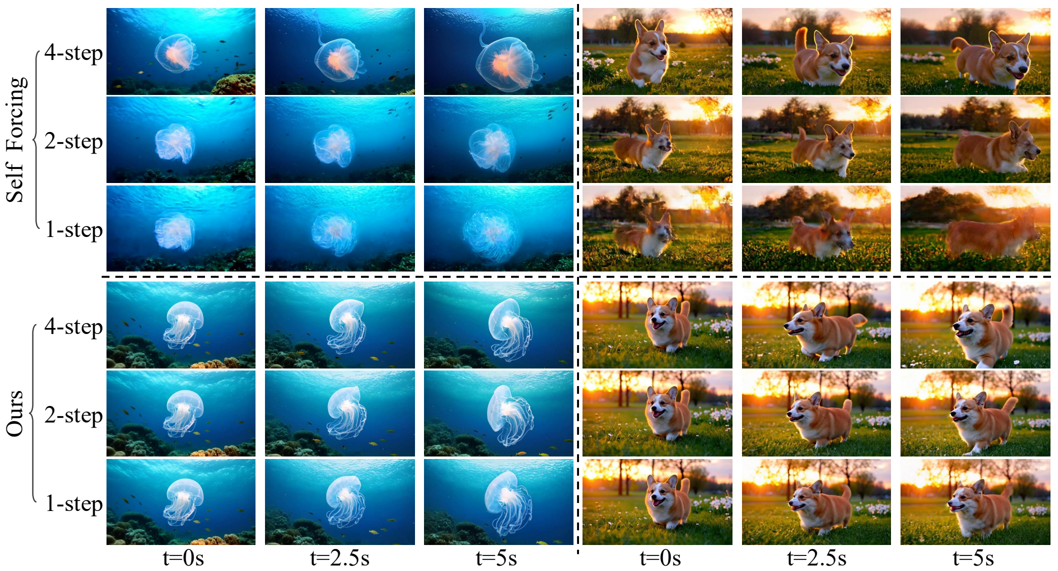


Figure 2: Qualitative results of Self Forcing (Huang et al., 2025) and ours under 4-step, 2-step and 1-step generation. Our method consistently maintains high-quality generation across 4-step, 2-step, and 1-step inference.

contrast, autoregressive models generate frames sequentially by predicting the next token conditioned on previously generated frames. However, this strong dependence on earlier outputs often leads to error accumulation, making it difficult to synthesize highly realistic long videos (Che et al., 2025).

Recent works (Chen et al., 2024; Hu et al., 2025; Jin et al., 2025; Li et al., 2025; Yin et al., 2025; Weng et al., 2024) have proposed hybrid frameworks that integrate diffusion and autoregressive generation. A common design is to model temporal dynamics autoregressively while applying diffusion-based iterative denoising spatially. Despite their improved flexibility, their sequential and iterative nature introduces more complex generation processes and prolonged inference times. Building upon this line of research, our work adopts a distillation-based framework that significantly reduces generation latency, enhancing real-time interactivity in video synthesis.

2.2 Adversarial Distillation

A line of work (Song et al., 2023; Song & Dhariwal, 2024; Liu et al., 2022; Meng et al., 2023; Berthelot et al., 2023) focuses on distilling the multi-step generation process into a few steps to improve the sampling efficiency of diffusion models. These methods train a student model to approximate the ordinary differential equation (ODE) trajectory of the original teacher model. Adversarial distillation uses a discriminator to align the distribution of the student model with the target distribution. ADD (Sauer et al., 2024b) and UFOGen (Xu et al., 2024) aim for the student model's final output to be indistinguishable from real data.

Further advancements have sought to preserve the generation trajectory itself, not just the final output. SDXL-Lightning (Lin et al., 2024) and LADD (Sauer et al., 2024a) achieve this by aligning the intermediate denoising states of the student model with those of the teacher model. Zhang et al. (2024) and Lin et al. (2025) fine-tune a pre-trained diffusion video model for AR generation via adversarial training, but they directly align one-step outputs with real data. Our work employs a discriminator to align the generated results of an n-step denoising process with those of an $(n+1)$-step process. This approach enhances the model's performance and consistency across various limited-step scenarios.

2.3 Score Distillation

Existing methods (Zhou et al., 2024; Lu et al., 2025) also utilize score-based models to achieve distribution matching across different noise states during distillation. For example, Dreamfusion (Poole et al., 2023) and ProlificDreamer (Wang et al., 2023) leverage a T2I diffusion model's score function

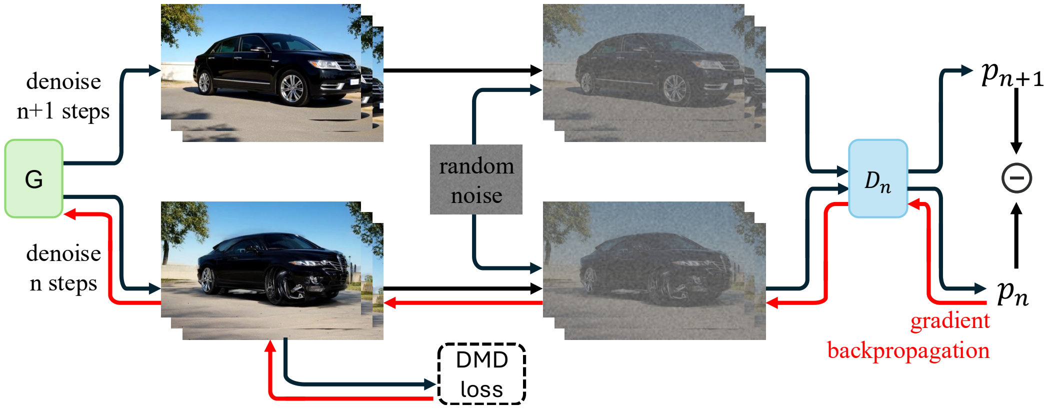


Figure 3: Pipeline of our proposed adversarial self-distillation process. We employ a discriminator D_n to align the randomly noised n-step video with the $(n+1)$-step one through calculating the ASD loss in Equation (4). The generator G is optimized using a combined objective function that includes the standard DMD loss and the ASD loss. Note that distillation is performed in the latent space, while the pixel domain is primarily used for visual analysis and display.

to guide text-conditioned 3D generation. Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) extends this approach to accelerate diffusion model generation by employing the original model as a score estimator for real data and a fine-tuned version as a score estimator for generated data.

Our approach is similar in that it aims to maintain a consistent sample probability across distributions under a particular noise level. To achieve this, we align the noisy version of the predictions from our few-step model's n-step and $(n+1)$-step denoising.

3 METHODOLOGY

To achieve high-quality and flexible causal video generation with minimal computational overhead, our distillation process employs two key strategies. We use Distribution Matching Distillation (DMD) to distill a multi-step generative model into a few-step student model through score-based alignment. Meantime, we introduce a novel form of self-supervision by aligning the student model's n-step denoising with its (n+1)-step version in the distribution level. This dual supervision not only enhances the generation quality but also enables the final model to produce consistent, hight-quality output across a variety of few-step settings. The few-step generator, the generator's score estimator and the discriminator are trained alternately to facilitate this process. The pipeline of the Adversarial Self-Distillation (ASD) process is depicted in Fig. 3. During inference, we further improve efficiency and quality by employing a First-Frame Enhancement (FFE) strategy, which uses a more intensive denoising process for the initial frame while reducing the number of steps for subsequent frames.

3.1 PRELIMINARY: DISTRIBUTION MATCHING DISTILLATION

Distribution Matching Distillation (DMD) (Yin et al., 2024a;b) presents a score-based approach to align a few-step student model with a multi-step teacher model. Unlike other distillation methods (Song et al., 2023; Song & Dhariwal, 2024), DMD not only accelerates generation but also improves output quality by aligning the output distributions. In diffusion models, the score function is represented by the gradient of the log probability of the distribution:

$$s_\theta(\boldsymbol{x}_t, t) = \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) = -\frac{\epsilon_\theta(\boldsymbol{x}_t, t)}{\sigma_t}, \quad (1)$$

where $\boldsymbol{x}_t$ is the noisy sample at timestep t with $p(\boldsymbol{x}_t)$ as its corresponding data distribution, $t \in \{0, ..., T\}$. $p(\boldsymbol{x}_T)$ is a standard Gaussian distribution. $\epsilon_\theta(\boldsymbol{x}_t, t)$ is the predicted output of the generative model parameterized by θ and σ_t is predefined by the noise schedule (Karras et al., 2022; Song et al., 2021b). DMD minimizes the reverse Kullback-Leibler (KL) divergence between the

Algorithm 1 Adversarial Self-Distillation Process

```
Require: Few-step denoising steps $\mathcal{T} = \{t_1, \dots, t_N\}$, score function of teacher model $s_{\text{data}}$
Initialize few-step student model $G_\theta$ with the original model
Initialize generator's score function $s_{\text{gen}}$ with $s_{\text{data}}$ estimated by the original model
Initialize discriminator $D_\psi^n$ with the original model with trainable heads, $D_\psi^n$ and $s_{\text{gen}}$ share the same backbone parameters
while training do
    $n \leftarrow \text{RandomInteger}(1, N)$
    Predict few-step student model generate sample $\boldsymbol{x}_0^{n+1} = G_\theta^{n+1}(\boldsymbol{z}_1), \boldsymbol{z_1} \sim \mathcal{N}(0, I)$
    Add noise $\boldsymbol{x}_t^1 = (1-t)\boldsymbol{x}_0^{n+1} + t\epsilon,\ \epsilon \sim \mathcal{N}(0, I), t \sim \text{Uniform}(\mathcal{T})$
    Predict few-step student model generate sample $\boldsymbol{x}_0^n = G_\theta^n(\boldsymbol{z}_2), \boldsymbol{z_2} \sim \mathcal{N}(0, I)$
    Add noise $\boldsymbol{x}_t^2 = (1-t)\boldsymbol{x}_0^n + t\epsilon,\ \epsilon \sim \mathcal{N}(0, I)$
    Update few-step generator $G_\theta$ using DMD loss and ASD loss    ▷ Eq. 5
    Update discriminator $D_\psi^n$ with ASD loss    ▷ Eq. 4
    Update generator's score function $s_{\text{data}}$ using diffusion loss $\mathcal{L}_{\text{gen}}^\phi$    ▷ Eq. 3
end while
```

score of the true data distribution and the score of the student model's generated data distribution. The gradient of DMD loss can be represented by:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{DMD}} &\triangleq \mathbb{E}_t\left(\nabla_\theta \text{KL}\left(p_{\text{gen},t} \| p_{\text{data},t}\right)\right) \\ &\approx -\underset{\boldsymbol{z},t,\boldsymbol{x}_t}{\mathbb{E}}\left[\left(s_{\text{data}}\left(\boldsymbol{x}_t, t\right) - s_{\text{gen}}\left(\boldsymbol{x}_t, t\right)\right)\frac{dG_\theta(\boldsymbol{z})}{d\theta}\right],\end{aligned} \tag{2}$$

where s_{data} and s_{gen} are the score functions that trained of real data and generated data, which point towards the higher density of data for $p_{\text{gen},t}$ and $p_{\text{data},t}$. The noisy sample $\boldsymbol{x}_t$ is obtained by the diffusion forwarding process $\boldsymbol{x}_t = \alpha_t \hat{\boldsymbol{x}}_0 + \sigma_t\epsilon,\ \epsilon \sim \mathcal{N}(0, I)$, where $\hat{\boldsymbol{x}_0}$ is the generated output of the few-step student model G_θ. $\alpha_t, \sigma_T > 0$ are defined by the noise schedule. $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the Gaussian distribution.

The true data score s_{data} is estimated by the original teacher model, while the generation score s_{gen} is estimated by a "teaching assistant" (TA) model, which is a fine-tuned version of the teacher. The TA model and the few-step generator are typically trained in an alternating fashion. The TA model is fine-tuned on the standard diffusion denoising loss on the generated data from the student model:

$$\mathcal{L}_{\text{gen}}^\phi = \left\|\epsilon_{\text{gen}}^\phi(\boldsymbol{x}_t, t) - \epsilon\right\|_2^2 . \tag{3}$$

3.2 Adversarial Self-Distillation

A key drawback of DMD is its lack of flexibility. The resulting student model is optimized for a single, fixed number of steps, requiring a separate distillation for each desired configuration (e.g., 4 steps vs. 2 steps). To overcome this rigidity, we introduce a method that aligns the outputs of a single few-step model across varying step counts.

Our core idea is to align the n-step denoising distribution with the $(n{+}1)$-step version. This ensures that the model can maintain consistent, high-quality output regardless of the chosen few-step setting. Since adversarial distillation objectives (Sauer et al., 2024b; Lin et al., 2024) are known to preserve sharpness and fine details better than DMD under low-step constraints, this alignment is performed using a discriminator D_ψ^n. This discriminator's task is to make the outputs of adjacent denoising steps indistinguishable through a relativistic pairing GAN objective (Jolicoeur-Martineau, 2018). The adversarial self-distillation loss is represented as:

$$\mathcal{L}_{\text{ASD}}(\theta, \psi) = \underset{\substack{\boldsymbol{z_1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\ \boldsymbol{z_2} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}}{\mathbb{E}}\left[f\left(D_\psi^n\left(\Psi\left(G_\theta^n(\boldsymbol{z_1})\right)\right) - D_\psi^n\left(\Psi\left(G_\theta^{n+1}(\boldsymbol{z_2})\right)\right)\right)\right], \tag{4}$$

where G_θ^n tries to maxmize $\mathcal{L}_{\text{ASD}}$ and D_ψ^n is optimizied to minimize it. $f(t) = -\log\left(1 + e^{-t}\right)$ is drawn from classic GAN (Goodfellow et al., 2020; Nowozin et al., 2016) and Ψ represents the adding noise process. G_θ^n and G_θ^{n+1} are the generated output of the few-step generator with n steps and $(n{+}1)$ steps. The few-step generator is trained with both DMD loss and ASD loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DMD}} + \alpha * \mathcal{L}_{\text{ASD}}, \tag{5}$$

Algorithm 2 Inference Procedure with First-Frame Enhancement

```
Require: Intensive denoising timesteps $\{t_1, \dots, t_T\}$, reduced denosing timesteps $\{t_1, \dots, t_R\}$, video length $L$, few-step autoregressive video generator $G_\theta$,
for $i = 1$ to $L$ do
    Initialize: $x^i \sim \mathcal{N}(0, I)$
    $K = \begin{cases} T & \text{if } i = 1 \\ R & \text{otherwise} \end{cases}$
    for $j = 1$ to $K$ do
        Predict $\hat{x}_0^i = G_\theta(x^i, t_j)$
        Update $x^i = \Psi(\hat{x}_0^i, \epsilon, t_{j+1})$, where $\epsilon \sim \mathcal{N}(0, I)$
    end for
end for
Return $\{\hat{x}_0^i\}_{i=1}^L$
```

where α is the hyperparameter that weights the two loss functions. The detailed distillation process is shown in Alg. 1.

Crucially, aligning the outputs of two adjacent few-step predictions is a much more stable training objective than directly aligning a few-step model with its multi-step teacher. The smaller distributional gap between adjacent steps makes our few-step generator significantly easier to train. This approach also provides a unique benefit: the n-step generator receives a supervision signal not only from the teacher model's trajectory but also from the $(n{+}1)$-step prediction, a novel form of self-distillation that further boosts the model's generation quality.

3.3 First-Frame Enhancement

Existing causal video generation methods typically employ a uniform number of denoising steps for each frame. However, in causal generation process, the quality of subsequent frames is critically dependent on the preceding frames. The quality defects in earlier frames are highly likely to be propagated throughout the rest of the sequence. This is particularly true for the first frame, which has no prior context and must synthesize a high-quality initial state from a zero-data starting point. Consequently, the generation of the first frame requires dedicated attention to mitigate accumulated error and ensure high overall video quality. Fig. 4 shows the similarity of predicted $\hat{x}_0$ for different video frames across various denoising steps. It reveals that the first frame has low similarity across steps, indicating that each denoising step is crucial. In contrast, subsequent frames exhibit higher similarity, suggesting a greater redundancy that makes them more suitable for few-step prediction. These results are based on the Self Forcing model and are consistent across multiple prompt variants and random seeds. We include additional results across scenarios in the Appendix Section C.

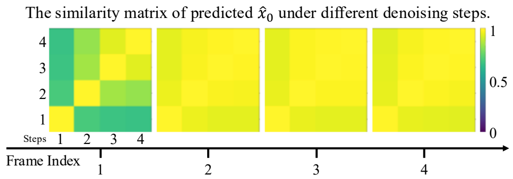


Figure 4: Cosine similarity matrices of different frames in causal diffusion video generation. Each matrix shows the similarity of the predicted $\hat{x}_0$ between different denoise steps (from 1 to 4).

Based on this observation, we propose a novel Frist-Frame Enhancement (FFE) denoising strategy, as detailed in Alg. 2. The first frame undergoes a more intensive denoising process, requiring a minimum of four steps, while subsequent frames can be generated with a significantly reduced number of steps, such as one or two. This frame-based control allows us to enhance the quality of the generated video in the few-step scenario.

4 Experiments

4.1 Implementations

Our approach employs a causal video distillation framework based on the Self Forcing (Huang et al., 2025) training paradigm. The causal model architecture builds upon the Wan2.1-T2V-1.3B (Wang

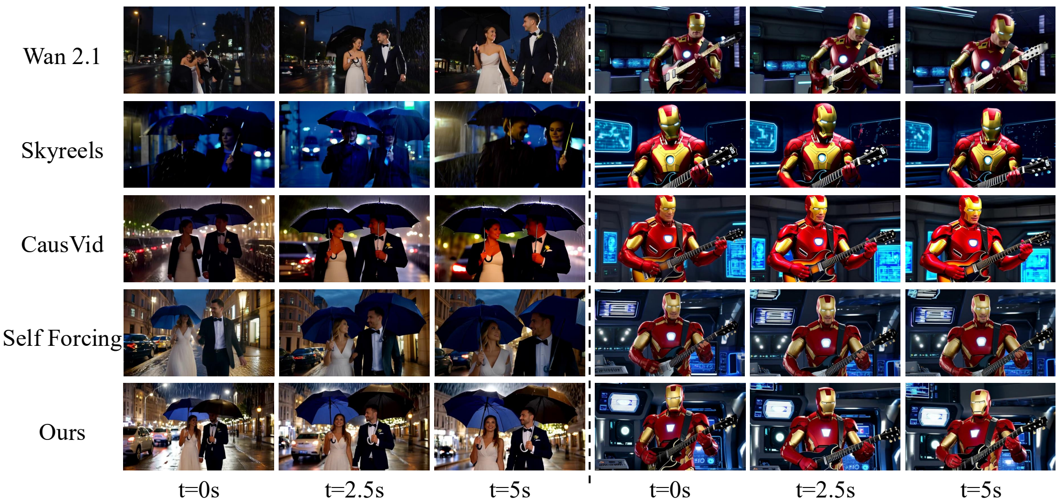


Figure 5: Qualitative comparisons. We visualize videos generated by Ours against those by Wan2.1 (Wang et al., 2025), SkyReels-V2 (Chen et al., 2025), CausVid (Yin et al., 2025) and Self Forcing (Huang et al., 2025) at 4-step generation. All models share the same architecture with 1.3B parameters.

et al., 2025) backbone, a Flow Matching based model (Lipman et al., 2023). We adopt CausVid's (Yin et al., 2025) initialization protocol to stabilize early causal training phases via asymmetric distillation from a pre-trained bidirectional teacher model. For training data, we utilize the exact text prompts from Self Forcing for fair comparison. The diffusion process is optimized with a 4-step denoising schedule during training, while using our step-skipping strategy during inference to accelerate generation. Adversarial training objective is incorporated via integration of RpGAN (Jolicoeur-Martineau, 2019) objectives with R1 (Mescheder et al., 2018) and R2 regularization terms following R3GAN (Huang et al., 2024a). And we use the frozen fake score function as the backbone of the discriminator following DMD2 Yin et al. (2024a). To reduce computational cost and the number of parameters, we model discriminators D^n's output logits as the n-th dimensional output logit of the final layer, thus different D^n share the same backbone and classifier head. The performance is rigorously assessed using VBench (Huang et al., 2024b) for multidimensional evaluation and user preference studies to quantify human-perceived visual quality and semantic alignment.

4.2 Comparison with Existing Baselines

Following the protocol of Self-Forcing, we compare with representative open-source text-to-video models under various inference steps, ranging from 64 to 4, ensuring a comprehensive assessment. For fair comparison under 2-step and 1-step generation, we additionally train 2-step and 1-step distilled versions of Self-Forcing and evaluate them against our model. As shown in Tab. 1, our model achieves slightly better performance compared to Self-Forcing under 4-step generation. Notably, in both the 2-step and 1-step settings, our approach outperforms the specifically distilled versions of Self-Forcing without requiring additional parameter optimization. For instance, under 1-step inference, our model exceeds Self-Forcing by 3.27 points in Total Score.

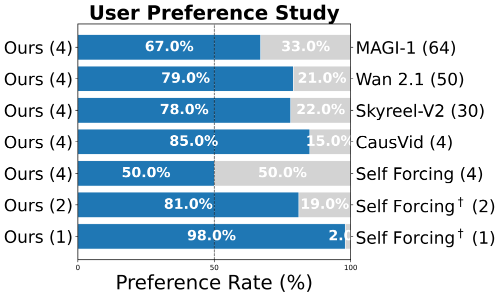


Figure 6: User preference study. "(n)" indicates the number of denoising steps for each method. Self Forcing under 1 & 2 step generation is retrained for fair comparison, denoted as Self Forcing†.

As illustrated in Fig. 5, our method is capable of generating high-fidelity videos using only approximately 8% and 13% of the denoising steps required

Table 1: Comparison with relevant baselines. We compare our method with representative open-source video generation models of similar parameter sizes and resolutions. For simplicity, n-step generation with FFE strategy is denoted with n^*. Self Forcing under 1 & 2 step generation is retrained for fair comparison, denoted as Self Forcing†.

Model	#Params	Resolution	Denoising Steps↓	Evaluation scores ↑		
				Total Score	Quality Score	Semantic Score
Many Steps						
MAGI-1 (Sand-AI, 2025)	4.5B	832×480	64	79.18	82.04	67.74
Wan2.1 (Wang et al., 2025)	1.3B	832×480	50	**84.26**	**85.30**	**80.09**
SkyReels-V2 (Chen et al., 2025)	1.3B	960×540	30	82.67	84.70	74.53
NOVA (Deng et al., 2025)	0.6B	768×480	25	80.12	80.39	79.05
LTX-Video (HaCohen et al., 2024)	1.9B	768×512	20	80.00	82.30	70.79
Pyramid Flow (Jin et al., 2025)	2B	640×384	20	81.72	84.74	69.62
4 Steps						
CausVid (Yin et al., 2025)	1.3B	832×480	4	81.20	84.05	69.80
Self Forcing (Huang et al., 2025)	1.3B	832×480	4	84.31	85.07	**81.28**
Ours	1.3B	832×480	4	**84.38**	**85.16**	81.25
2 Steps						
Self Forcing† (Huang et al., 2025)	1.3B	832×480	2	83.49	84.20	80.62
Ours	1.3B	832×480	2^*	**84.32**	**85.15**	**81.02**
1 Steps						
Self Forcing† (Huang et al., 2025)	1.3B	832×480	1	80.62	81.19	78.35
Ours	1.3B	832×480	1^*	**83.89**	**84.55**	**81.24**

by Wan 2.1(Wang et al., 2025) and SkyReels(Chen et al., 2025), respectively, while achieving visually superior results. This leads to a substantial reduction in computational cost. Moreover, under the same number of denoising steps, our approach yields improved detail and visual quality compared to CausVid(Yin et al., 2025) and better video details compared to Self-forcing(Huang et al., 2025). Fig. 6 shows the user study results comparing our model against several important baselines. Our approach is consistently preferred over many-step baselines, including the many-step Wan2.1 that our model is initialized from. Meanwhile, our results are 70% better than CausVid (Yin et al., 2025) and on par with Self Forcing, since both our model and Self Forcing use identical DMD-based supervision in the 4-step setting. For the extreme 1-step and 2-step generation setting, our method yields a significantly 96% and 62% preference over Self-Forcing, respectively. This demonstrates the effectiveness of our proposed ASD and FFE strategy in improving extremely few-step video generation.

4.3 Ablation Study

To better understand the contribution of each component to skip-step generation, we conducted a comprehensive ablation study, as shown in Tab. 2. As shown in the first two rows, using identical training data, our ASD training method consistently improves the quality of both 1-step and 2-step generation during inference. For example, it raises the Total Score by 2.52 and the Semantic Score by 6.84 under one-step generation. Comparison between the first and third rows reveals that the FFE inference strategy leads to substantial gains in overall video quality—notably, a 4.19 increase in Total Score and a 10.64 improvement in Semantic Score for one-step generation, even exceeding the quality of the original two-step generation results (row 1). Moreover, combining both methodology further enhances performance, as demonstrated in the last row.

We qualitatively compare the quality of generated outputs with and without our proposed ASD training and FFE inference strategies, as illustrated in Fig. 7. For simplicity, n-step generation with the FFE inference strategy is denoted as n^*-step. When adopting the ASD training strategy, the generation quality at 2^*-step is comparable to that achieved with 4 steps, while even under the challenging 1^*-step setting, video quality is well maintained. Moreover, at both 2-step and 1-step inference, the frame quality is noticeably superior to variants trained without the ASD strategy. This

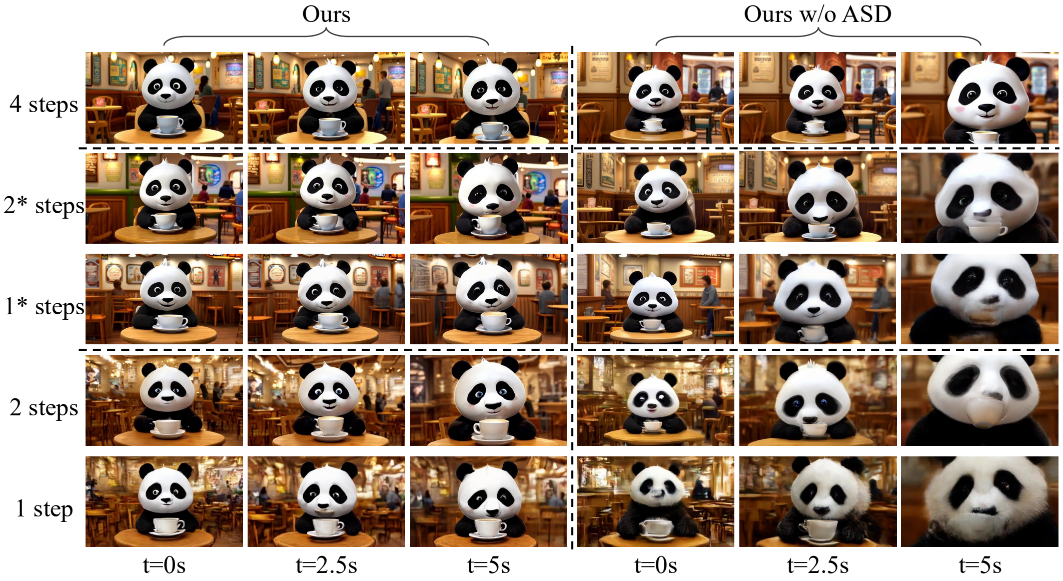


Figure 7: Qualitative comparison illustrating the effects of training and inference strategies on video generation. Left: results using the ASD training method; right: results without it. For simplicity, inference with the FFE strategy is denoted with *.

Table 2: Ablation Study. The symbols ✓ and × denote the inclusion and exclusion of the corresponding strategy, respectively. ASD refers to the Adversarial Self-Distillation strategy in Section 3.2, and FFE denotes the First-Frame Enhancement strategy introduced in Section 3.3.

Methods		2-step Generation ↑			1-step Generation ↑		
ASD	FFE	Total Score	Quality Score	Semantic Score	Total Score	Quality Score	Semantic Score
×	×	82.61	83.84	77.68	78.13	80.33	69.31
✓	×	83.28	84.17	79.69	80.65	81.77	76.15
×	✓	83.80	84.65	80.40	83.04	83.81	79.95
✓	✓	**84.32**	**85.15**	**81.02**	**83.89**	**84.55**	**81.24**

shows the effectiveness of our ASD training strategy under skip-step generation. More specifically, the variant trained without the ASD strategy exhibits obvious error accumulation under the 2^* setting, resulting in noticeable background shifts and character blurring at t = 5s. And under the 1^*-step condition, the severe blurring occurs as early as t = 2.5s. As demonstrated by the comparison between rows 2 & 4 and rows 3 & 5, the FFE inference strategy that increases the denoising steps for the initial frame substantially enhances overall video quality by effectively mitigating artifacts such as background and subject blurring in the first frame.

5 Conclusion

In this work, we introduce an Adversarial Self-Distillation training objective for causal video diffusion models' distillation, along with a First-Frame Enhancement inference strategy for efficient sampling. The adversarial objective encourages the n-step generated video distribution to approximate that of the $(n+1)$-step generation, thereby progressively improving few-step generation quality. During inference, we explicitly distinguish between the first frame and subsequent frames in causal video generation, allocating different numbers of denoising steps to enhance overall video quality with low computational overhead. Experiments demonstrate that our distilled model outperforms other baselines under 1- and 2-step generation configurations.

6 Reproducibility statement

A comprehensive description of the implementation is provided in Section 4.1 and Section A. The source code and additional comparison videos are included in the supplementary material and will be released in the future.

7 Ethics Statement

This work focuses on improving the efficiency of diffusion-based video generation through distillation. Our contributions are methodological and technical in nature, and do not involve human subjects, personal or sensitive data, or medical applications. The experiments are conducted on publicly available benchmark datasets (e.g., VBench), which are widely used in prior research and do not contain personally identifiable information.

We confirm that this research adheres to the ICLR Code of Ethics and complies with all relevant institutional and legal standards. No conflicts of interest or external sponsorship influencing the work are present.

References

Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *ICLR*, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NeurIPS*, 37:24081–24125, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2023.

Jiaxiang Cheng, Bing Ma, Xuhua Ren, Hongyi Jin, Kai Yu, Peng Zhang, Wenyue Li, Yuan Zhou, Tianxiang Zheng, and Qinglin Lu. Pose: Phased one-step adversarial equilibrium for video diffusion models. *arXiv preprint arXiv:2508.21019*, 2025.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *ICLR*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, pp. 102–118. Springer, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 35:8633–8646, 2022.

Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. In *ICLR Workshop on World Models: Understanding, Modelling and Scaling*, 2025.

Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The gan is dead; long live the gan! a modern gan baseline. *NeurIPS*, 37:44177–44215, 2024a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

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. In *CVPR*, 2024b.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025.

A Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *ICLR*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 35:26565–26577, 2022.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: a large language model for zero-shot video generation. In *ICML*, pp. 25105–25124, 2024.

Zongyi Li, Shujie Hu, Shujie Liu, Long Zhou, Jeongsoo Choi, Lingwei Meng, Xun Guo, Jinyu Li, Hefei Ling, and Furu Wei. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. In *ICLR*, 2025.

Jian Liang, Chenfei Wu, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *NeurIPS*, 35:15420–15432, 2022.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. *arXiv preprint arXiv:2506.09350*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2022.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. *arXiv preprint arXiv:2507.18569*, 2025.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, pp. 14297–14306, 2023.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, volume 162, pp. 16784–16804. PMLR, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pp. 8162–8171. PMLR, 2021.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *NeurIPS*, 29, 2016.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Sand-AI. Magi-1: Autoregressive video generation at scale, 2025. URL `https://static.magi.world/static/files/MAGI_1.pdf`.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia*, pp. 1–11, 2024a.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, pp. 87–103. Springer, 2024b.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021a.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *ICLR*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, pp. 32211–32252. PMLR, 2023.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *NeurIPS*, 36:8406–8441, 2023.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *CVPR*, pp. 7395–7405, 2024.

Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *CVPR*, pp. 8196–8206, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *NeurIPS*, 37: 47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *ICLR*, 2024.

Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, pp. 10188–10198. IEEE, 2023.

Zhixing Zhang, Yanyu Li, Yushu Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, Sergey Tulyakov, et al. Sf-v: Single forward video generation model. *NeurIPS*, 2024.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *ICML*, 2024.

A Details of Implementations

Our implementation is largely based on the code from Self-Forcing(Huang et al., 2025). Specifically, we adopt the chunk-wise DMD (Yin et al., 2024b) training variant of Self-Forcing, with all training conducted in the VAE latent space using a chunk size of 3. Following Self-Forcing, we utilize the Wan 2.1 14B model as the real score estimator. The model is trained with a batch size of 8 for 3,000 steps.

For Adversarial Self-Distillation, we incorporate the R1 and R2 regularizations (Mescheder et al., 2018) from the R3GAN (Huang et al., 2024a) objective.

$$\mathcal{L}_{\text{reg}} = \frac{1}{2}\mathbb{E}_{t,x_t^{n+1},x_t^n,\epsilon,\hat{\epsilon}}\left[\|f_\psi(x_t^{n+1}) - f_\psi(x_t^{n+1} + \sigma \cdot \epsilon)\|_2^2 + \|f_\psi(x_t^n) - f_\psi(x_t^n + \sigma \cdot \hat{\epsilon})\|_2^2\right] \tag{6}$$

$$\mathcal{L}_D(\psi) = -\mathbb{E}_{t,x_t^{n+1},x_t^n}\left[\log\left(\text{sigmoid}\left(f_\psi(x_t^{n+1}) - f_\psi(x_t^n)\right)\right)\right] + \lambda\mathcal{L}_{\text{reg}} \tag{7}$$

$$\mathcal{L}_G(\theta) = -\mathbb{E}_{t,x_t^{n+1},x_t^n}\left[\log\left(\text{sigmoid}\left(f_\psi(x_t^n) - f_\psi(x_t^{n+1})\right)\right)\right] \tag{8}$$

where $x_t^{n+1} \sim p_{\theta,\text{n+1},t}$, $x_t^n \sim p_{\theta,\text{n},t}$ are the noisy (n+1)-step generation data and n-step generation data, respectively, ϵ and $\hat{\epsilon}$ are Gaussian noise sampled from $\mathcal{N}(0,1)$, and f_ψ is the critic network

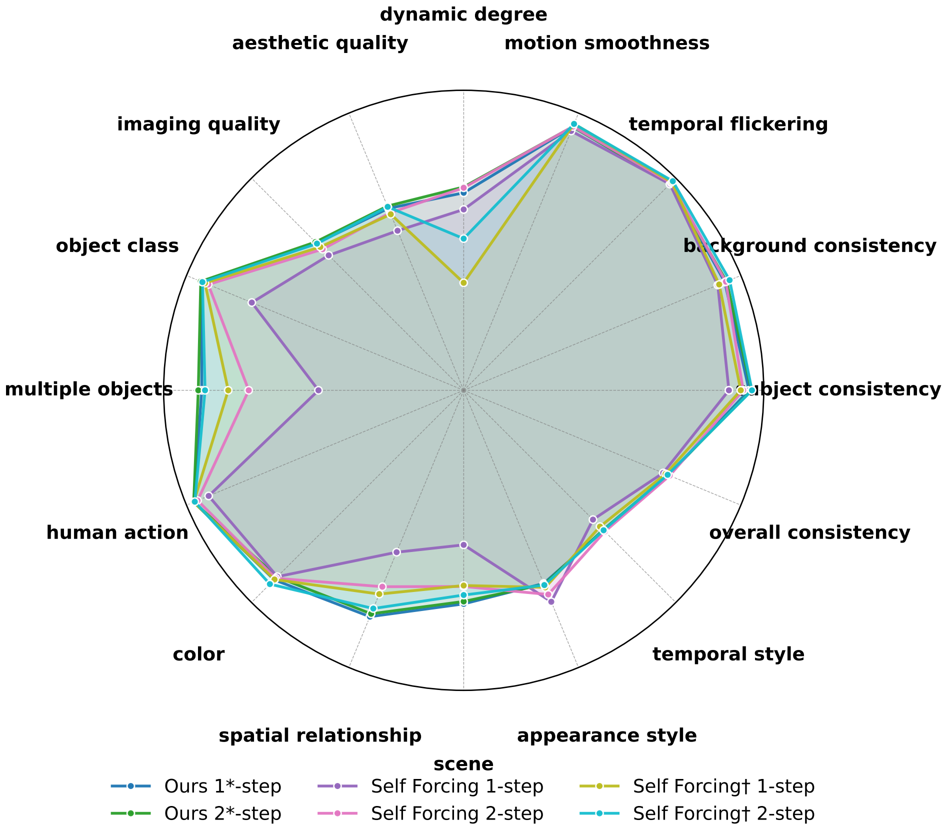


Figure 8: VBench scores visualization. We compare our results with Self Forcing Huang et al. (2025) variants using all 16 VBench metrics under 1-/2-step setting. The retrained 1-/2-step version denotes as Self Forcing†.

(discriminator) of GAN. We use $\lambda = 600$, $\sigma = 0.05$ for all experiments. The generator, fake score estimator, and discriminator are optimized alternately with a ratio of 1:4:1, following(Yin et al., 2024a; Huang et al., 2025). Following Self-Forcing, we inserted additional cross-attention layers and classification heads (serving as discriminator heads) at layers 12, 21, and 29 of the fake score model. These discriminator heads operate solely on backbone features and query tokens, without access to the timestep t. This implementation detail will be explicitly included in the revised manuscript. Note that we discard ASD Loss for the last step denoising. For simplicity, we reserve the index n of the n-th discriminator D_n that discriminates the n-step generation distribution from that of $(n+1)$-step here.

For the retrained Self-Forcing baselines, we use exactly the same hyperparameters as the official implementation, except for changing the student model's number of denoising step list from [1000, 750, 500, 250] to [1000, 500] or [1000]. Specifically, we used: real score CFG weight: 3.0, optimizers: AdamW for both generator and discriminator with β_1 = 0, β_2 = 0.999, ϵ = 1e-8, weight decay = 0.01, learning rate (generator): $2e-6$, learning rate (discriminator): $4e-7$, generator/discriminator update ratio: 5:1. Training is monitored until convergence.

B VBench Scores Across All Dimensions

In Figure 8, we evaluate our method and Self-Forcing across all 16 VBench metrics under both 1-step and 2-step generation settings. Our method consistently outperforms Self Forcing, achieving notably higher scores in semantic alignment-particularly in object class, multiple objects, spatial relationships, and scene. Moreover, it demonstrates superior performance in dynamic modeling, as evidenced by a significantly higher dynamic degree score.

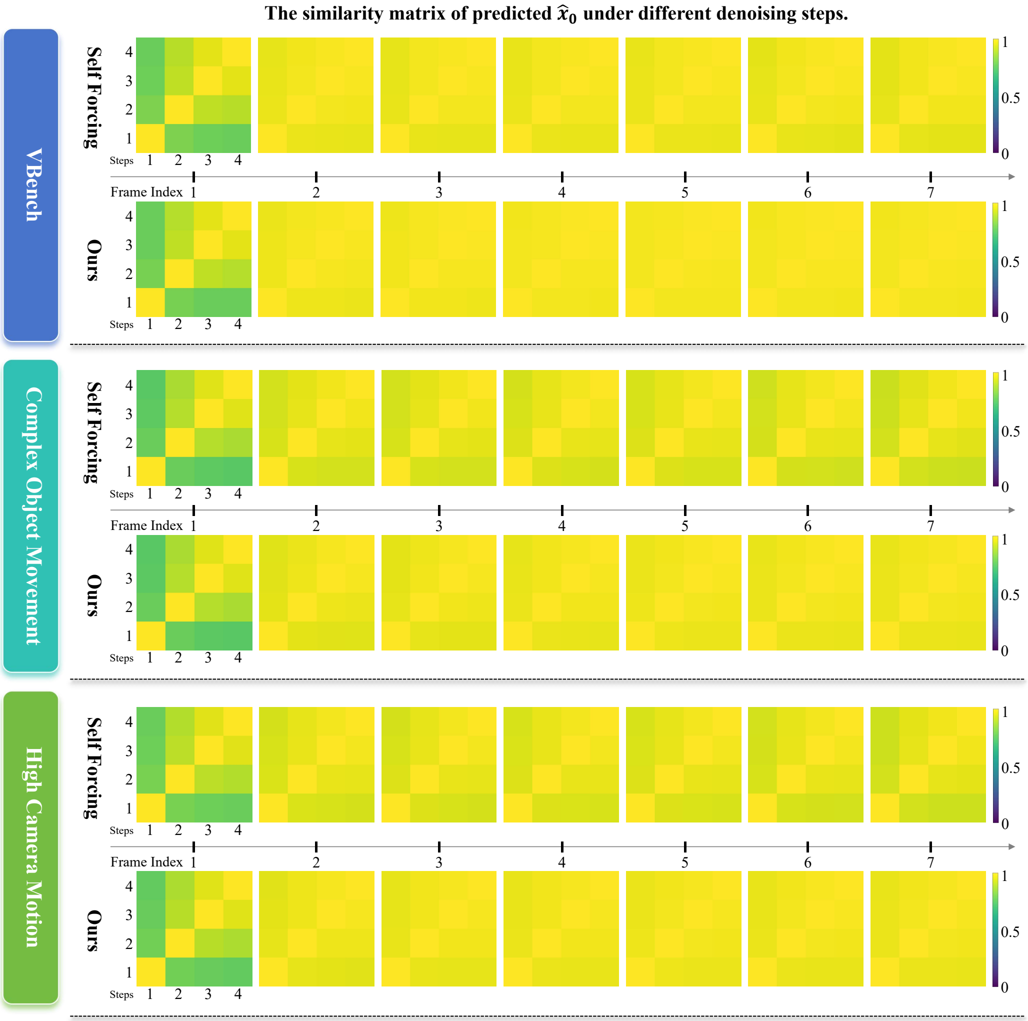


Figure 9: Cosine similarity matrices of predicted $\hat{x_0}$ across denoising steps (1 to 4) in causal diffusion video generation, comparing our method and Self-Forcing (Huang et al., 2025). We report results averaged separately for VBench, complex object motion, and high camera motion.

C Causal Denoising Trajectory Similarity Across Scenarios

To further validate the generality of the phenomenon in Section 3.3, we apply the analysis to both our method and Self-Forcing across three diverse scenarios: VBench(Huang et al., 2024b), complex object motion, and high camera motion. For each, we record denoising trajectories from 200 video samples, yielding Figure 9. The figure shows that, across all scenarios, the first frame exhibits low similarity across denoising steps, indicating each step is critical. In contrast, subsequent frames show high inter-step similarity, suggesting greater redundancy and thus greater suitability for few-step prediction. Moreover, our method achieves even higher similarity in later frames, demonstrating that ASD enables the student to match multi-step quality with fewer steps, effectively supporting step-skipping inference.

D Reduced Distributional Gaps Between Adjacent Denoising Steps

To empirically validate the smoother transitions between adjacent denoising steps, we conduct a quantitative analysis using the Fréchet Video Distance (FVD). Table 3 presents FVD values computed between outputs at step n and step $n+1$ (i.e., adjacent steps), as well as between step n and the 50-step 14B teacher model.

Table 3: Fréchet Video Distance (FVD) between video distributions at different denoising steps. Lower FVD values indicate smaller distributional gaps.

Comparison	**Step 1**	**Step 2**	**Step 3**	**Step 4**
$n \to n+1$	732	1136	441	N/A
$n \to$ teacher	1836	1646	1454	1448

As illustrated in Table 3, the FVD between adjacent denoising steps is consistently lower than that between each step and the final teacher output. For instance, at step 1, the adjacent-step FVD is 732, much smaller than 1836 for the teacher-step comparison. These results provide quantitative support that the distributional discrepancy between consecutive steps is substantially smaller than the divergence from the teacher distribution. The Adjacent-Step Distribution (ASD) objective capitalizes on this property by aligning each step with its immediate successor, which exhibits lower distributional divergence. Overall, these findings corroborate the core assumption of our approach: transitions between adjacent denoising steps are indeed smoother.

E Few-step Long-video Generation Evaluation

To demonstrate that our method enables high-quality, long-video generation with few (1 or 2) denoising steps—and effectively mitigates error accumulation—we generate 20-second continuations using both 1-step and 2-step inference for our approach and Self-Forcing, and evaluate them on VBench(Huang et al., 2024b). As shown in Table 4, our method overall outperforms Self-Forcing both quantitatively and qualitatively in long-video generation. For visual comparison, Figure 10 presents side-by-side 2*-step (ours) vs. 2-step (Self-Forcing) 20-second samples. Our results exhibit clearly superior performance in motion dynamics (rows 3, 4, 6), visual detail and fidelity (rows 1, 2, 3, 6), and camera motion realism (rows 1, 2, 4).

These results confirm that ASD not only reduces computational cost but also enhances temporal coherence and visual quality in extended video generation.

Table 4: VBench evaluation scores for our method vs. Self-Forcing under few-step long-video (20-second) inference settings

Method	Total Score	Quality Score	Semantic Score
Self Forcing(Huang et al., 2025) (2-step)	0.8250	0.8293	0.8076
Ours (2*-step)	0.8263	0.8329	0.7998
Self Forcing(Huang et al., 2025) (1-step)	0.8066	0.8101	0.7923
Ours (1*-step)	0.8200	0.8248	0.8011

F Hyperparameter Analysis

To investigate the impact of the ASD loss on model performance, we evaluated different variants of the hyperparameter α in the Equation (5) and computed the corresponding Total Score on VBench under three inference settings, as illustrated in Figure 11. The results indicate that at 4-step generation, the Total Scores across different model variants are comparable. When using 2^*-step generation, the variant without ASD Loss ($\alpha = 0$) exhibits a noticeable performance drop, while other variants maintain performance levels similar to those observed in 4-step generation. Under 1^*-step generation, all methods experience a decline in performance, with the most pronounced degradation occurring when α is set to zero. These findings demonstrate that the proposed ASD training objective enhances video quality under skip generation settings and remains robust across a wide range of hyperparameter values. While absolute values vary slightly across α, the trend is consistent: skip-step performance improves markedly whenever $\alpha \neq 0$.

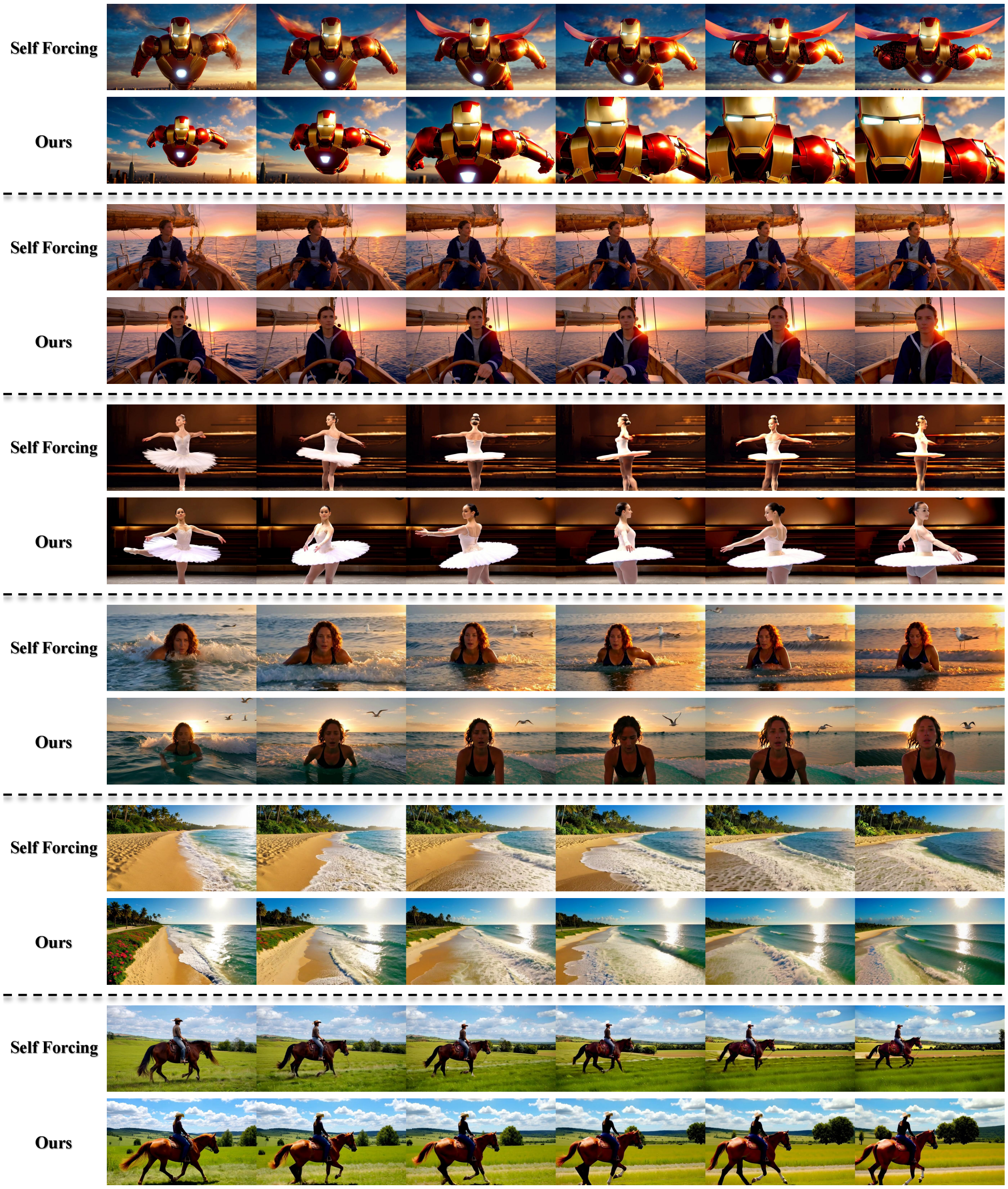


Figure 10: Our method versus Self-Forcing on uniformly sampled frames (including first and last) from 20-second long videos.

G User Study Details

We randomly selected 20 prompts from the VBench(Huang et al., 2024b) prompts expanded by Self-Forcing. Five baseline methods Sand-AI (2025); Huang et al. (2025); Yin et al. (2025); Chen et al. (2025); Wang et al. (2025) were used to generate corresponding videos, which were then paired with corresponding videos generated by our method, resulting in 5 × 20 video pairs. During the user study, each participant (total 12 participants) evaluated 20 video pairs, including 4 randomly selected pairs from each baseline. The order and pairing of videos were independently randomized for every participant. Participants could not see each other's choices. The final results were averaged across all participants. The user study interface, as shown in Figure 12, displayed a video pair alongside the corresponding prompt and selection buttons. Participants could choose whether the left video

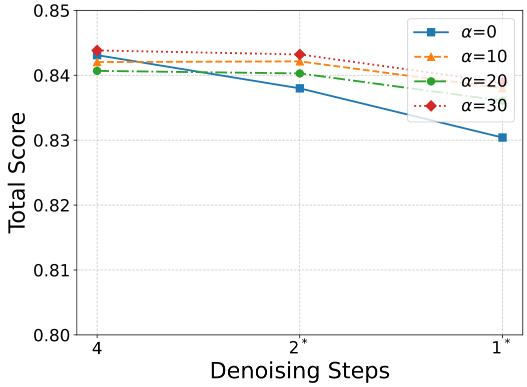
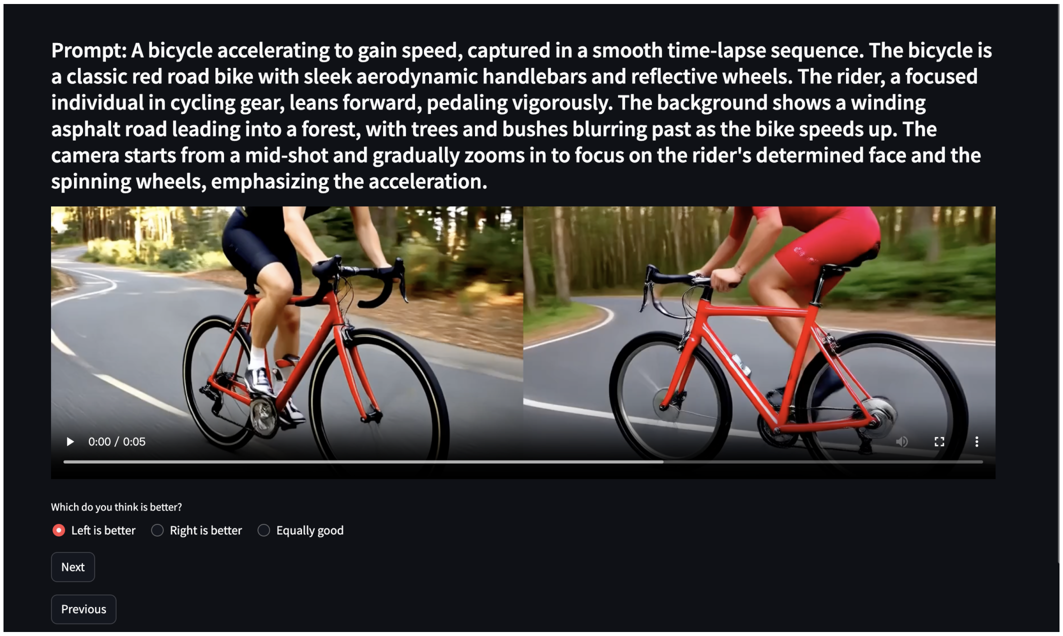


Figure 11: Sensitivity analysis. Impact of parameter α in Equation (5) on the VBench(Huang et al., 2024b) Total Score across three inference scenarios. For simplicity, inference with first-frame enhancement is denoted with *.



Prompt: A bicycle accelerating to gain speed, captured in a smooth time-lapse sequence. The bicycle is a classic red road bike with sleek aerodynamic handlebars and reflective wheels. The rider, a focused individual in cycling gear, leans forward, pedaling vigorously. The background shows a winding asphalt road leading into a forest, with trees and bushes blurring past as the bike speeds up. The camera starts from a mid-shot and gradually zooms in to focus on the rider's determined face and the spinning wheels, emphasizing the acceleration.

Which do you think is better?

● Left is better ○ Right is better ○ Equally good

Figure 12: User study interface. Participants were asked to select the higher-quality video in each pair or to indicate if the two were of similar quality.

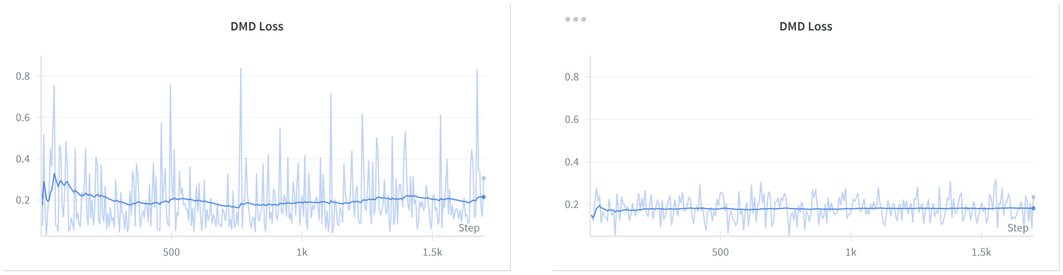
was better, the right video was better, or both were of similar quality. To ensure a fair comparison with both one-step and two-step configurations, we retrained the one-step and two-step versions of Self-Forcing.

H USAGE OF LLMS

In this work, LLMs were primarily used to assist in grammar checking. All outputs from LLMs were manually reviewed.

I TRAINING DYNAMICS ANALYSIS

Figure 13 presents the DMD loss curves of our method and Self-Forcing during training. Notably, our approach exhibits significantly reduced fluctuations in DMD loss compared to Self-Forcing. With



(a) Self Forcing

(b) Ours

Figure 13: The DMD loss curves of Self Forcing and our method during training. The DMD loss values remain around 0.2 for both methods, but our method exhibits substantially lower fluctuation.

ASD, the DMD loss shows lower mean (0.180 vs. 0.198) and dramatically reduced variance (0.00258 vs. 0.01963) compared to pure DMD training, demonstrating that the proposed ASD loss effectively stabilizes distribution matching with the teacher model during training.

Figure 14 shows that the discriminator's logit difference between n-step and (n+1)-step samples oscillates calmly around zero, indicating that the generator and discriminator co-evolve stably rather than collapsing or diverging. This behavior supports that ASD training is stable in practice, despite using an adversarial term.

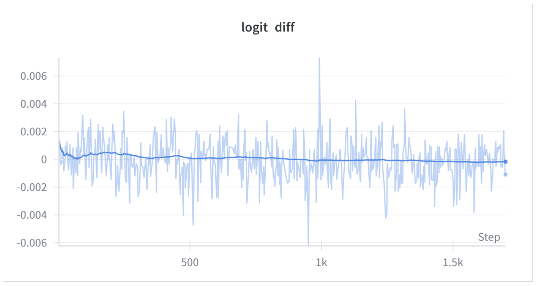


Figure 14: The discriminator's predicted logit difference between n-step and (n+1)-step samples during training.

Training Overhead. The ASD framework introduces only a lightweight discriminator on top of the existing TA score model's backbone, using shared parameters for all n-step discriminators. This results in minimal extra memory usage compared to traditional DMD. The additional GPU memory required by ASD is +1.3% compared to the Self-Forcing baseline, which is a modest increase. The additional training time is approximately +20% relative to Self-Forcing. Given that ASD allows a single model to handle multiple inference steps (1, 2, 3, 4), this overhead is relatively small compared to the computational cost of training separate models for each step.