Phased One-Step Adversarial Equilibrium for Video Diffusion Models

Jiaxiang Cheng[1], Bing Ma[1], Xuhua Ren[1*], Hongyi Henry Jin[1,2], Kai Yu[1], Peng Zhang[1], Wenyue Li[1], Yuan Zhou[1], Tianxiang Zheng[1], Qinglin Lu[1*]

[1]Tencent Hunyuan
[2]Computer Science Department, UCLA
jiaxiangcc@gmail.com

Abstract

Video diffusion generation suffers from critical sampling efficiency bottlenecks, particularly for large-scale models and long contexts. Existing video acceleration methods, adapted from image-based techniques, lack a single-step distillation ability for large-scale video models and task generalization for conditional downstream tasks. To bridge this gap, we propose the Video Phased Adversarial Equilibrium (*V-PAE*), a distillation framework that enables high-quality, single-step video generation from large-scale video models. Our approach employs a two-phase process. (i) Stability priming is a warm-up process to align the distributions of real and generated videos. It improves the stability of single-step adversarial distillation in the following process. (ii) Unified adversarial equilibrium is a flexible self-adversarial process that reuses generator parameters for the discriminator backbone. It achieves a co-evolutionary adversarial equilibrium in the Gaussian noise space. For the conditional tasks, we primarily preserve video-image subject consistency, which is caused by semantic degradation and conditional frame collapse during the distillation training in image-to-video (I2V) generation. Comprehensive experiments on VBench-I2V demonstrate that V-PAE outperforms existing acceleration methods by an average of 5.8% in the overall quality score, including semantic alignment, temporal coherence, and frame quality. In addition, our approach reduces the diffusion latency of the large-scale video model (*e.g.*, Wan2.1-I2V-14B) by 100×, while preserving competitive performance.

Project Page — https://v-pae.github.io/

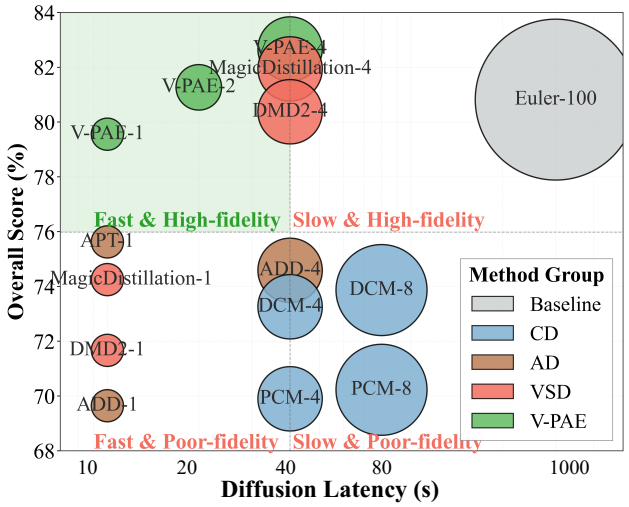

Figure 1: **Comparison between V-PAE and existing acceleration methods on VBench-I2V**. It includes three distillation paradigms: (i) Consistency Distillation (CD), (ii) Variational Score Distillation (VSD) and (iii) Adversarial Distillation (AD). For fairness, all models are distilled from Wan2.1-I2V-14B (Wan et al. 2025) using the same dataset and training cost. Diffusion latency is measured for 5-second 720 × 1280 videos on 8× H20 GPUs.

1 Introduction

Diffusion models (Song et al. 2021; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021) have witnessed a paradigm shift, extending their remarkable success from image synthesis (Rombach et al. 2022; Podell et al. 2024; Cheng et al. 2025; Chen et al. 2024; Esser et al. 2024) to video generation (Wang et al. 2023a; Guo et al. 2024; Blattmann et al. 2023; Yang et al. 2024; Kong et al. 2024; Wan et al. 2025). Contemporary video diffusion models attain impressive fidelity through two factors: architectural scaling and longer temporal context. Designs range from the small-scale U-Net (Ronneberger, Fischer, and Brox 2015) employing spatiotemporal decoupled attention to the large-scale Diffusion Transformer (DiT) (Peebles and Xie 2023) employing full attention, while temporal windows extend to hundreds of frames. However, this fidelity comes at a prohibitive computational cost. For instance, synthesizing a 5-second video with 50 diffusion steps using a large-scale video model such as Wan2.1-I2V-14B[1] (Wan et al. 2025) requires approximately 15 minutes on advanced GPUs (*e.g.*, 8×H20). This substantial latency severely limits deployment in real-time applications.

Existing video diffusion distillation methods (Wang et al. 2023b; Li et al. 2024; Zhang et al. 2025; Lin et al. 2025a; Shao et al. 2025) primarily adapt image distillation tech-

*Corresponding authors.

[1]https://huggingface.co/Wan-AI/Wan2.1-I2V-14B-720P

niques (Salimans and Ho 2022; Song et al. 2023; Luo et al. 2023; Yin et al. 2024b; Sauer et al. 2024b,a; Yin et al. 2024a), lacking critical spatiotemporal complexities for video data. We identify two fundamental limitations: (i) efficiency bottleneck, failing to distill large-scale video diffusion models (>10B) into single-step generators; and (ii) task generalization, lacking the ability to perform conditional tasks (*e.g.*, image-to-video generation (I2V)). It suffers from poor video-image subject consistency caused by semantic degradation and conditional frame collapse.

To address the efficiency bottleneck, our objective is to distill a model capable of generating single-step videos that closely match real videos. Adversarial distillation for single-step sampling from Gaussian noise is the dominant strategy. But it faces severe challenges: the significant quality gap between real and generated videos makes discrimination trivial, yielding weak gradients and unstable training. As a result, existing adversarial methods, such as Distribution Matching Distillation (DMD2) (Yin et al. 2024a) and adversarial diffusion distillation (ADD) (Sauer et al. 2024a), begin denoising from medium-to-high signal-to-noise ratio (SNR) engines to ensure sufficient similarity between real and generated samples, which provides informative discriminator supervision. However, the mismatch between these training distributions and low-SNR sampling fundamentally limits the performance of the distilled model. To this end, we propose Video Phased Adversarial Equilibrium (**V-PAE**), which comprises two sequential processes: stability priming and unified adversarial equilibrium. (i) Stability priming is a warm-up process to align the distributions of real and generated videos, which improves the stability of single-step adversarial distillation in the subsequent process. Guided by Variational Score Distillation (VSD) (Poole et al. 2023), it relies on score-gradient differences to reduce the distributional gap. (ii) Unified adversarial equilibrium is a flexible self-adversarial process that reuses generator parameters for the discriminator backbone. It achieves a co-evolutionary adversarial equilibrium in the Gaussian noise space, improving the single-step adversarial stability.

To improve task generalization, we introduce the semantic discriminator head and the conditional Score Distillation Sampling (SDS) (Poole et al. 2023) loss. The semantic discriminator head primarily enhances the semantic perception of images and videos. It employs the multi-modal (*e.g.*, image, video, and text) attention module. By enabling cross-modal interactions with learnable query tokens, it strengthens the semantic alignment ability of the distilled video model. The conditional SDS is inspired by ADD (Sauer et al. 2024b), leveraging the continuous distribution of the pre-trained video model to prevent distribution bias. It requires only a multi-step video model as the regularization distribution, correcting conditional frame collapse based on smaller discrepancies between the conditional and generated frames.

We comprehensively evaluate V-PAE on the VBench-I2V benchmark (Huang et al. 2024) for semantic alignment, temporal coherence, and frame quality. Experiments demonstrate that our approach outperforms other acceleration methods by an average of 5.8% in the overall quality score of single-step videos. Crucially, our approach achieves single-step quality comparable to that of its 50-step video model and even exceeds it with a few steps in a zero-shot manner. As illustrated in Fig. 1, applied to the large-scale video diffusion model (*e.g.*, Wan2.1-I2V-14B (Wan et al. 2025)), our approach reduces the iterative time from nearly 15 minutes to 10 seconds, achieving a $100\times$ speedup. This overcomes the computational barrier, enabling real-time, high-fidelity video synthesis for interactive applications.

2 Related Work

Video Diffusion Models. The video generation field has seen rapid progress through diffusion models (Song et al. 2021; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021), enabling applications such as conditional generation (HaCohen et al. 2024; Kong et al. 2024; Wan et al. 2025; Yang et al. 2024) and video editing (Chai et al. 2023; Feng et al. 2024; Singer et al. 2024). Early works (Wang et al. 2023a; Guo et al. 2024; Blattmann et al. 2023) insert a lightweight temporal attention into U-Net (Ronneberger, Fischer, and Brox 2015), using spatiotemporal decoupling to model video distribution. This mechanism struggles with long video generation due to limited global interaction. Later works inherit weights from large-scale image diffusion models (Peebles and Xie 2023; Esser et al. 2024; Chen et al. 2024) for video synthesis. Recent works (Yang et al. 2024; HaCohen et al. 2024; Kong et al. 2024; Wan et al. 2025) abandon reliance on image diffusion models. Instead, they build comprehensive video synthesis systems through data expansion, model scaling, and training optimization procedures. However, they remain limited in computational efficiency and inference latency, primarily due to the large-scale model parameters and long-video temporal complexity.

Distillation Sampling. Reducing diffusion sampling steps remains a central challenge. In image distillation, Progressive Distillation (PD) (Salimans and Ho 2022; Lin, Wang, and Yang 2024) progressively compresses teacher-student pipelines but faces inherent performance limits due to error accumulation. Consistency Distillation (CD) (Song et al. 2023; Luo et al. 2023) enables single-step generation along the Probability Flow (PF) Ordinary Differential Equation (ODE); however, training stability remains an issue. Variational Score Distillation (VSD) (Yin et al. 2024b,a) minimizes gradient differences between real and fake score estimates; however, fidelity constraints persist in single-step sampling. Adversarial Distillation (AD) (Sauer et al. 2024b; Lin, Wang, and Yang 2024; Sauer et al. 2024a) uses a discriminator to reduce the distribution gap between real and generated data; however, effective distillation is limited to four uniformly spaced timesteps. In video distillation, most works simply adapt image techniques. Dual-Expert Consistency Model (DCM) (Wang et al. 2023b; Li et al. 2024; Lv et al. 2025) applies consistency distillation without video-specific designs. MagicDistillation (MD) (Shao et al. 2025) achieves few-step synthesis via weak-to-strong score matching, but the quality degrades at single-step sampling. Adversarial Post-Training (APT) (Lin et al. 2025a) deploys one-step adversarial distillation for videos, but its applicability is restricted to small-scale models and short clips.

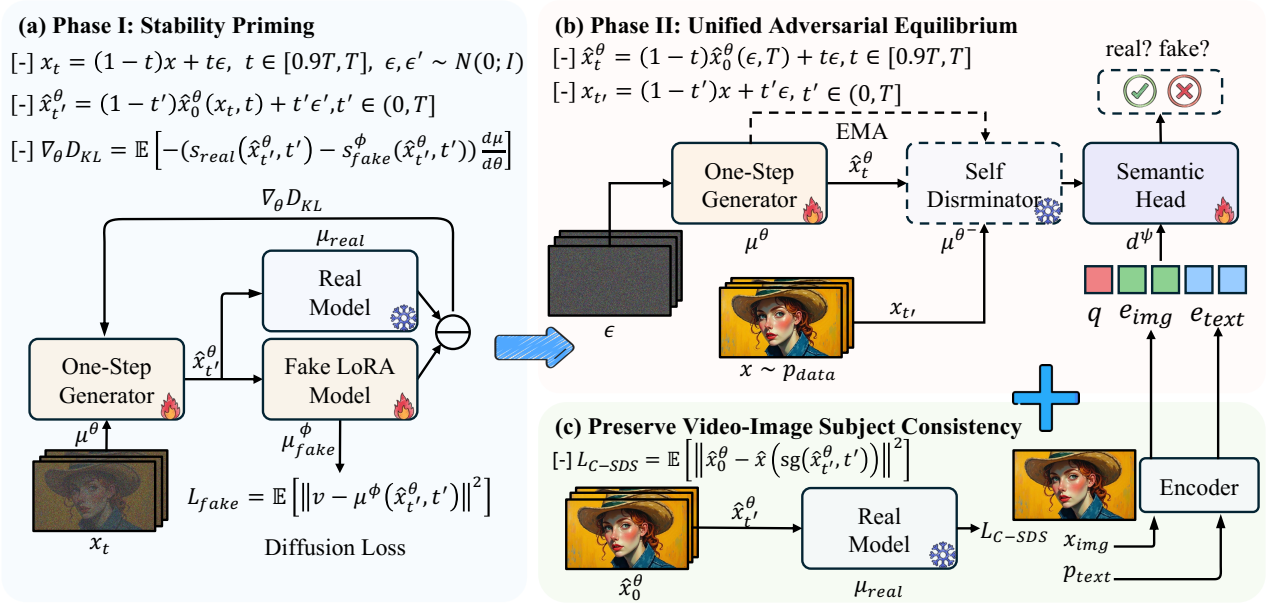

Figure 2: **Overview.** *V-PAE* first aligns the distributions of generated and real videos in the (a) *stability priming* process. Building on this process, it reuses the generator parameters for the discriminator backbone, which achieves a co-evolutionary adversarial training in the (b) *unified adversarial equilibrium* process. For the conditional generation, we also provide the conditional SDS loss and semantic discriminator to (c) *preserve video-image subject consistency*.

3 Methodology

Our objective is to distill video diffusion models that can generate single-step video $\hat{x}^\theta_0$ close to real video $x \sim p_{\text{data}}$. To this end, we propose V-PAE, a two-phase, one-step equilibrium distillation framework. First, we introduce a stability priming process in Sec. 3.1, which is a warm-up process to align the distributions between $\hat{x}^\theta_0$ and x. It improves the stability of single-step adversarial distillation in the following process. Second, we introduce the unified adversarial equilibrium process in Sec. 3.2, which is a flexible self-adversarial process that reuses generator parameters for the discriminator backbone. It achieves a co-evolutionary equilibrium. For the conditional tasks, such as image-to-video (I2V) generation, we primarily preserve the video-image subject consistency in Sec. 3.3, which avoids semantic degradation and conditional frame collapse. Finally, we provide the total training optimization objective in Sec 3.4.

3.1 Phase I: Stability Priming

Adversarial distillation faces severe training instability for the single-step video $\hat{x}^\theta_0$ generated from Gaussian noise ϵ. It is caused by the distribution mismatch between the $\hat{x}^\theta_0$ and the real video $x \sim p_{\text{data}}$. To this end, we design a stability priming process to narrow their distribution distance. We define three models: a priming generator μ^θ to align the single-step distribution with the real distribution, a real model μ_{real}, and a fake model μ^ϕ_{fake} for score-based estimates of distribution matching distillation. As illustrated in Fig. 2 (a), the μ^θ directly generates $\hat{x}^\theta_0 = f^\theta(x_t, t)$ in the low signal-to-noise ratio (SNR) regime, where $t \in [0.9T, T]$. The $\hat{x}^\theta_0$ is perturbed by varying Gaussian noise ϵ to obtain $\hat{x}^\theta_{t'} = (1-t')\hat{x}^\theta_0 + t'\epsilon$, which inputs μ_{real} and μ^ϕ_{fake} to compute the distribution matching loss with the score gradient difference, as shown in Eq. 1:

$$\nabla_\theta D_{\text{KL}} = \mathop{\mathbb{E}}_{\epsilon\sim\mathcal{N}(0;\mathbf{I})}\left[-(s_{\text{real}}(\hat{x}^\theta_{t'}) - s^\phi_{\text{fake}}(\hat{x}^\theta_{t'}))\frac{d\mu}{d\theta}\right], \quad (1)$$

where $t' \in (0, T]$, $s_{\text{real}} = -\frac{\hat{x}^\theta_{t'} + (1-t')\mu_{\text{real}}}{t'}$, and $s_{\text{fake}} = -\frac{\hat{x}^\theta_{t'} + (1-t')\mu^\phi_{\text{fake}}}{t'}$ are under standard score-based definitions (Song et al. 2021).

Efficient Distribution Track. The real score is fixed during training as the real distribution. The fake score is dynamic during training as the single-step distribution, which is updated with the standard diffusion loss (Ho, Jain, and Abbeel 2020) as shown in Eq. 2.

$$\mathcal{L}_{\text{fake}} = \mathop{\mathbb{E}}_{\epsilon\sim\mathcal{N}(0;\mathbf{I})}[||v - \mu^\phi_{\text{fake}}(\hat{x}^\theta_{t'}, t')||^2], \quad (2)$$

To improve the stability and efficiency of tracking the distribution of the $\hat{x}^\theta_0$, we introduce the lightweight Low-Rank Adaptation (LoRA) (Hu et al. 2023) to μ^ϕ_{fake}. Compared to full-parameter training, this strategy enables a more stable and faster distribution tracking ability for the large-scale video model, even with only a small amount $\hat{x}^\theta_0$. For stable launching, we additionally enable zero-parameter initialization in the adaptation weights.

3.2 Phase II: Unified Adversarial Equilibrium

Building on the stability priming process, we propose the unified adversarial equilibrium process to guide the μ^θ in

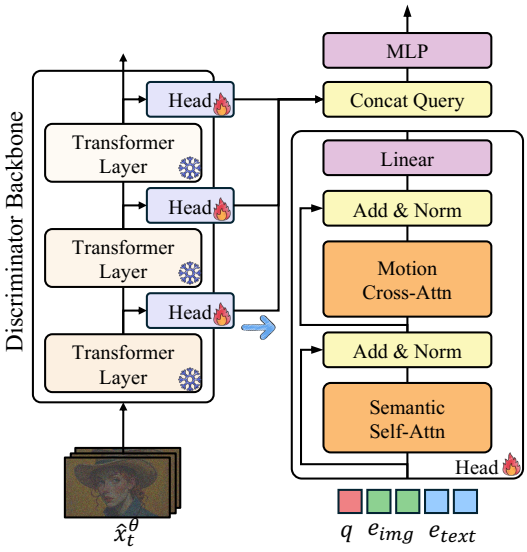

Figure 3: The semantic discriminator head architecture.

generating high-quality single-step videos $\hat{x}_0^\theta$ from Gaussian noise ϵ. In traditional adversarial distillation (Sauer et al. 2024b,a), the discriminator backbone is either frozen or fully-parameter-trained. The frozen backbone easily leads to significant asymmetry in the trained parameters between the generator and discriminator. It tends to degrade the overall quality of the single-step video. The trained backbone faces the memory challenge for the large-scale video diffusion model (*e.g.*, Wan2.1-I2V-14B), as shown in Tab. 3.

Self Discriminator. Instead, we introduce a flexible adversarial paradigm. It enables the μ^θ as a self discriminator backbone, paired with lightweight heads d^ψ for computing logits. The key property is efficient co-evolution. With the limited memory, the self discriminator backbone maintains a comparable learning capacity. It enables a stable and high-quality Nash equilibrium while generating the single-step video from Gaussian noise. As illustrated in Fig. 2 (b), the μ^θ first samples $\hat{x}_0^\theta = f^\theta(\epsilon, T)$ from the endpoint. Then $\hat{x}_0^\theta$ is perturbed with varying noise levels in low-SNR regimes. The noisy samples $\hat{x}_t^\theta$ are fed into the self discriminator backbone μ^θ to extract multi-layer features, which are used to compute logits by d^ψ. Following the Hinge loss (Miyato et al. 2018), our unified adversarial loss is as shown in Eq. 3:

$$\begin{aligned}\mathcal{L}_{\text{UAE}} = &\underbrace{\mathbb{E}_{\hat{x}_0^\theta = f^\theta(\epsilon,T)}\left[\log D_\theta^\psi\left(\hat{x}_t^\theta, t; \mu^\theta\right)\right]}_{\text{Generator objective}} \\ &+ \underbrace{\mathbb{E}_{x\sim p_{\text{data}}}\left[\log\left(1 - D_{\theta^-}^\psi\left(x_t, t; \mu^{\theta^-}\right)\right)\right]}_{\text{Discriminator objective}},\end{aligned} \tag{3}$$

where $D_{\theta^-}^\psi = \left[\mu^{\theta^-}; d^\psi\right]$ is trained using the Exponential Moving Average (EMA) weights θ^- for unified adversarial stability, and the d^ψ represents the discriminator heads.

Gradient Regularization. To mitigate the gradient explosion of the discriminator heads, we further introduce a spatiotemporal differential R1 regularization. It treats temporal dimensions as intrinsic properties of the video, which enforce structured perturbations that jointly constrain spatial and temporal quality. Specifically, we apply composite noise perturbations to the single-step video as $\tilde{x}_t^\theta = \hat{x}_t^\theta + \sigma_s\epsilon_s + \sigma_t\epsilon_t$. The gradient regularization for the video model is as shown in Eq 4:

$$\mathcal{L}_{\text{STR1}} = \mathbb{E}_{\hat{x}_0^\theta = f(\epsilon,T)}\left[\frac{\|D_{\theta^-}^\psi(\tilde{x}_t^\theta, t) - D_{\theta^-}^\psi(\hat{x}_t^\theta, t)\|^2}{\|\sigma_s\epsilon_s + \sigma_t\epsilon_t\|^2}\right], \tag{4}$$

where $\sigma_s = 0.01$ is for perturbing the pixel distribution per frame, and $\sigma_t = 0.1$ is for disturbing the temporal distribution across the frames.

3.3 Preserve Video-Image Subject Consistency

Compared to text-to-video (T2V) generation, image-to-video (I2V) generation is the primary video application. However, its primary challenge is that the score of video-image subject consistency significantly degrades during the adversarial distillation process. This phenomenon is caused by semantic degradation and conditional frame collapse. To this end, we introduce the semantic discriminator head and conditional Score Distillation Sampling (SDS) loss.

Semantic Discriminator Head. To improve semantic alignment, our discriminator heads integrate multi-modal information (*e.g.*, video, image, and text). Inspired by Adversarial Post-Training (APT) (Lin et al. 2025a), we introduce a learnable logit query $q \in \mathbb{R}^{n\times d}$. As illustrated in Fig. 3, the q concatenates with the conditional image embeds $e_{\text{img}} \in \mathbb{R}^{s\times d}$ and the text embeds $e_{\text{text}} \in \mathbb{R}^{s\times d}$, which fuse the semantic information. To distinguish between different modalities, we incorporate modality position indices. After the semantic self-attention module, the multi-layer features from the self discriminator backbone μ^θ are cross-attended with the q, which enhances sensitivity to temporal quality. Finally, all processed queries are concatenated along the channel dimension and projected to a scalar logit by a Feed-Forward Network (FFN) layer.

Conditional SDS Loss. To prevent the conditional frame collapse, the conditional SDS loss leverages the distributional stability of the μ_{real} to reduce the discrepancy between continuous frames. As illustrated in Fig. 2 (c), the μ_{real} can maintain the consistency between the conditional and generated frames. Inspired by SDS (Poole et al. 2023), this loss can be defined as shown in Eq. 5:

$$\mathcal{L}_{\text{C-SDS}} = \mathbb{E}_{\hat{x}_0^\theta}\left[\|\hat{x}_0^\theta - f_{\text{real}}(\text{sg}(\hat{x}_{t'}^\theta), t')\|^2\right], \tag{5}$$

Unlike the poor effect in image adversarial distillation (Sauer et al. 2024b,a), this mechanism significantly avoids disturbances in the conditional frame.

3.4 Total Training Objective

Based on the processes mentioned above, our training objective for the single-step generator can be defined as:

$$\mathcal{L}_{\text{G}} = \mathcal{L}_{\text{UAE-G}} + \lambda\mathcal{L}_{\text{C-SDS}}, \tag{6}$$

A dignified woman dressed in elegant attire looks straight ahead, her face showing a gentle smile

Figure 4: **Qualitative comparison with Wan2.1-I2V-14B.** We compare our method against the baseline using both 100-NFE and 1-NFE sampling. For 100-NFE, videos are generated with 50 denoising steps and a guidance scale of 5.0.

Methods	Type	I2V Score (↑)	Semantic Alignment (SA)		Temporal Coherence (TC)		Frame Quality (FQ)		Quality Score (↑)	Latency Time (s)
			Subject Consistency (↑)	Background Consistency (↑)	Motion Smoothness (↑)	Dynamic Degree (↑)	Aesthetic Quality (↑)	Image Quality (↑)		
100-NFE										
Baseline	Euler	92.90	94.86	97.07	97.90	51.38	64.75	70.44	80.82	890.15
4-NFE										
Baseline	Euler	75.97	79.59	81.24	80.92	42.7	54.34	57.55	64.47	37.48
DMD2	VSD	91.9	93.19	95.37	96.05	50.61	64.9	70.65	80.38	
MD	VSD	94.24	**96.83**	97.90	**99.76**	50.60	65.4	69.82	81.94	
LCM	CD	79.79	81.21	83.87	81.39	43.86	54.18	58.26	68.94	
PCM	CD	80.54	80.82	84.12	83.01	44.89	55.67	60.27	69.9	
DCM	CD	83.01	85.49	87.27	87.74	46.25	59.33	63.75	73.26	
ADD	AD	84.48	87.53	90.07	90.77	46.71	59.06	63.5	74.59	
V-PAE	AD	**94.93**	94.21	**98.43**	98.64	**52.14**	**65.73**	**70.76**	**82.24**	
1-NFE										
Baseline	Euler	69.59	70.73	74.95	72.94	38.51	49.72	54.06	61.52	9.37
DMD2	VSD	83.15	81.91	85.47	85.87	45.78	57.04	62.47	71.67	
MD	VSD	84.02	87.87	87.9	90.04	46.32	58.85	64.76	74.25	
LCM	CD	76.92	79.1	81.0	82.55	43.06	53.94	59.15	67.96	
ADD	AD	79.38	82.4	83.83	83.01	44.15	54.63	60.12	69.65	
APT	AD	84.87	87.11	88.77	90.94	48.29	60.33	64.69	75.21	
V-PAE	AD	**91.54**	**94.14**	**95.8**	**94.99**	**49.54**	**62.28**	**68.66**	**79.56**	

Table 1: **Quantitative results on VBench-I2V.** The baseline model is Wan2.1-I2V-14B (Wan et al. 2025). For each metric, the best result is highlighted in **bold**, and the second best is underlined. Various distillation methods are compared under different numbers of sampling steps. Latency time is measured on $8\times$ H20 for 5-second videos with a resolution of 720×1080.

which consists of the adversarial generative loss and the conditional SDS loss. Where the $\lambda = 10$ represents the injection strength. And our training objective for the discriminator heads can be defined as:

$$\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{UAE-D}} + \mathcal{L}_{\text{STR1}}(\sigma_s, \sigma_t), \tag{7}$$

which consists of the adversarial discriminative loss and the gradient regularization loss. Where the $\sigma_s = 0.01$ and $\sigma_t = 0.1$ represent the disturbance strengths.

4 Experiment

We first present the detailed implementation for datasets, training, evaluation metrics, and baseline methods in Sec. 4.1 . And we demonstrate the effectiveness of our approach through quantitative results and qualitative visualization comparisons in Sec. 4.2. Then we analyze the necessity of our modules in Sec. 4.3. Finally, we perform the ablation experiments in Sec. 4.4.

4.1 Implementation

Datasets Preparation. Our training dataset consists of synthetic sources and open sources. The synthetic dataset is generated by Wan2.1-T2V-14B (Wan et al. 2025), where the captions are from OpenVID (Nan et al. 2025). It is used for the distributional alignment during the distillation process. The open datasets consist of the Koala-36M (Wang et al. 2025) and the Intern4K (Lin et al. 2025b). It is used to enhance the upper bound of distillation performance.

Training Configuration. In the stability priming process, we train the model with a learning rate of 1×10^{-6} and a batch size of 32. We use the Adam optimizer (Kim and hwan Oh 2025) with $\beta = (0.9, 0.999)$. This process is trained for only 500 steps. In the adversarial process, we train the model

Figure 5: **Qualitative results of 1-NFE between V-PAE and existing acceleration distillation methods.** We evaluate against representative methods from three paradigms, including DMD2 (Yin et al. 2024a) from VSD, PCM (Wang et al. 2024) from CD, and APT (Lin et al. 2025a) from AD.

with a larger learning rate of 2×10^{-6} and a batch size of 32. We adopt the Adam optimizer with the smaller beta $\beta = (0.5, 0.999)$ for lower memory consumption. And we apply the EMA with a decay rate of 0.995. For the generator, we set $\lambda = 10$ to prevent the conditional frame collapse in I2V training. For the discriminator heads, we set $\sigma_s = 0.01$ and $\sigma_t = 0.1$ to avoid the gradient explosion of the discriminator heads . This process is trained for 1000 steps.

Evaluation Metrics. We primarily conduct the quantitative evaluation on the comprehensive benchmark VBench-I2V (Huang et al. 2024). However, we also provide the quantitative results for the traditional benchmark VFHQ (Xie et al. 2022). The former includes 7 metrics and 3 dimensions: semantic alignment, temporal coherence, and frame quality. The latter is used to measure the distributional distance between generated and real videos. Additionally, we evaluate the diffusion latency of the inference process for all experiments on $8\times$H20 GPUs.

Baseline Methods. To validate the efficiency and quality of our approach, we compare it with other primary methods across different distillation paradigms. It includes the Variational Score Distillation (VSD) (Yin et al. 2024b,a; Shao et al. 2025), the Consistency Distillation (CD) (Song et al. 2023; Kim et al. 2024; Lv et al. 2025) and the Adversarial Distillation (AD) (Lin, Wang, and Yang 2024; Sauer et al. 2024a,b; Lin et al. 2025a).

4.2 Main Results

Quantitative Result. We present the quantitative results on VBench-I2V in Tab. 1. The results demonstrate that V-PAE outperforms existing acceleration methods under an identical Number of Function Evaluations (NFE). For 1-NFE, our approach outperforms the suboptimal method by an average of 5.8% in the overall quality score, which includes semantic alignment, temporal coherence, and frame quality. Compared to the 100-NFE baseline model, our approach achieves competitive performance in 1-NFE, which is only marginally behind by 1.5% in the overall quality score. However, our approach in 4-NFE surpasses it by 3.3%. For the diffusion latency, our approach accelerates the diffusion process by $100\times$. The more quantitative results on VFHQ can be found in the Appendix.

Qualitative Result. We present the qualitative comparison with existing acceleration methods in Fig. 5. The results demonstrate that V-PAE produces clearer and more coherent videos in 1-NFE. The comparison with the baseline model is shown in Fig. 4. The baseline model produces videos with blurry frames, abrupt transitions between conditional and generated frames, and motion collapse due to unstable dynamic in 1-NFE. In contrast, our approach generates consistently sharp and coherent frames. The more qualitative results with other acceleration methods and the baseline model are also presented in the Appendix.

4.3 Validating the Design

We primarily validate the necessity of our design based on the following analysis. *First, the role of the stability priming process is confirmed through ablations:* (i)Omitting this process entirely, (ii) implementing consistency distillation (Song et al. 2023) of APT (Lin et al. 2025a), (iii) implementing our stability priming process. The quantitative results in Tab. 2 demonstrate that our design is well-motivated

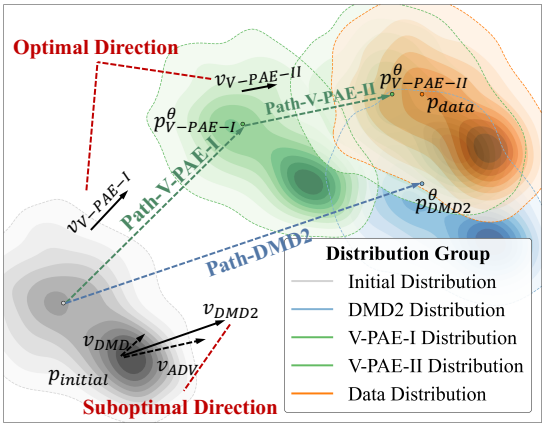

Figure 6: **Fundamental analysis of V-PAE and DMD2**. V-PAE can optimizes the distribution along an optimal trajectory through phased optimization. DMD2 simultaneously moves along a suboptimal direction without error correction.

and necessary in the first process. And its positive advantages can be extended to the following process. *Second, empirical observations reveal fundamental differences between V-PAE and DMD2.* Our approach and DMD2 are both optimized through adversarial distillation and score sampling distillation. But there are fundamental differences between them. The detailed analysis is illustrated in Fig. 6. DMD2 simultaneously optimizes the model using both the variational score gradient difference and the adversarial loss, which is a simple combination of different losses. In contrast, our core motivation is to align the distributions of real and generated videos in the first process. It can improve the stability of single-step adversarial distillation in the following process. Thus, we can achieve the high-quality adversarial equilibrium under the closer distributions.

4.4 Ablation Study

Discriminator Backbone Paradigm. We compare the quantitative performance of the frozen, fully parameterized, and self discriminator in Tab. 3. The self discriminator outperforms the other paradigms in the overall quality scores, demonstrating that the paradigm can identify more robust single-step video differences. In addition, the fully parameterized training causes the Out-of-Memory (OOM) error, rendering it incompatible with the large-scale video model.

Discriminator Head Architecture. We compare the quantitative results for the different discriminator head architectures in Tab. 3. The results show that the design of the head architecture has a significant impact on semantic alignment. For the large-scale video model, the convolution head is unable to stabilize adversarial training,which leads to a degradation in the overall score quality.

Factor of Conditional SDS. We ablate the factor of the conditional SDS loss in Tab. 3. When $\lambda = 0$, the conditional and generated frames suffer from severe collapse and discontinuities. As the $\lambda \to 10$, the overall quality score has significantly improved. And the conditional frame collapse is gradually restored. As the $\lambda \to 100$, the video sharpness is slightly impacted, and the frame quality score is reduced.

Methods	Type	SA (↑)	TC (↑)	FQ (↑)	Quality Score (↑)
First Process, 1-NFE					
Baseline	Euler	72.84	55.73	51.89	61.50
LCM	CD	80.05	62.81	56.55	67.96
V-PAE-I	VSD	**84.92**	**65.83**	**59.75**	**72.34**
Second Process, 1-NFE					
UAE + Baseline	AD	83.12	63.58	57.37	69.65
UAE + LCM	AD	87.94	69.62	62.51	75.21
UAE + V-PAE-I	AD	**94.97**	**72.26**	**65.47**	**79.56**

Table 2: **Necessity of stability priming.** The comparison of different first-process initialization methods and their effects on subsequent adversarial training.

Setting	SA (↑)	TC (↑)	FQ (↑)	Quality Score (↑)	Memory
Discriminator Head Architecture					
Conv	75.55	57.52	52.11	63.29	IM
CA	91.36	72.23	65.29	78.31	IM
CA†	**94.97**	**72.26**	**65.47**	**79.56**	IM
Discriminator Backbone Paradigm					
Frozen	92.21	70.12	63.92	77.36	IM
Trained	-	-	-	-	OOM
Unified	**94.97**	**72.26**	**65.47**	**79.56**	IM
Factor of Conditional SDS λ					
$\lambda = 0$	80.36	61.15	55.41	67.33	IM
$\lambda = 2.5$	86.83	66.18	59.96	72.87	IM
$\lambda = 10$	**94.97**	**72.26**	**65.47**	**79.56**	IM
$\lambda = 100$	92.28	70.24	63.62	77.31	IM

Table 3: **Component-wise ablation study.** Videos are generated in 1-NFE. Memory is measured on 32×H20 GPUs. Abbreviations – SA: Semantic Alignment; TC: Temporal Coherence; FQ: Frame Quality; CA†: Multi-modal Cross-Attention; OOM: Out of Memory; IM: In Memory.

5 Conclusion

In this paper, we propose the Video Phased Adversarial Equilibrium (V-PAE), a distillation framework that enables high-quality, single-step video generation from large-scale video models. Our approach employs a two-phase process. (i) Stability priming is a warm-up process designed to improve the stability of single-step adversarial distillation in the following process. (ii) Unified adversarial equilibrium is a flexible self-adversarial process that achieves a co-evolutionary adversarial equilibrium in the Gaussian noise space. We also primarily preserve video-image subject consistency in image-to-video (I2V) generation. Comprehensive experiments on VBench-I2V demonstrate that V-PAE outperforms existing acceleration methods by an average of 5.8% in the overall quality score. And it reduces the diffusion latency of the large-scale video model by 100×.

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23040–23050.

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2024. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, 74–91. Springer.

Cheng, J.; Xie, P.; Xia, X.; Li, J.; Wu, J.; Ren, Y.; Li, H.; Xiao, X.; Wen, S.; and Fu, L. 2025. Resadapter: Domain consistent resolution adapter for diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2438–2446.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Feng, R.; Weng, W.; Wang, Y.; Yuan, Y.; Bao, J.; Luo, C.; Chen, Z.; and Guo, B. 2024. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6712–6722.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *The Twelfth International Conference on Learning Representations*.

HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; et al. 2024. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2023. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Kim, D.; Lai, C.-H.; Liao, W.-H.; Murata, N.; Takida, Y.; Uesaka, T.; He, Y.; Mitsufuji, Y.; and Ermon, S. 2024. Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion. In *The Twelfth International Conference on Learning Representations*.

Kim, G. Y.; and hwan Oh, M. 2025. ADAM Optimization with Adaptive Batch Selection. In *The Thirteenth International Conference on Learning Representations*.

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

Li, J.; Feng, W.; Fu, T.-J.; Wang, X.; Basu, S.; Chen, W.; and Wang, W. Y. 2024. T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lin, S.; Wang, A.; and Yang, X. 2024. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*.

Lin, S.; Xia, X.; Ren, Y.; Yang, C.; Xiao, X.; and Jiang, L. 2025a. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*.

Lin, W.; Wei, J.; Liu, B.; Zhang, Y.; Yan, S.; and Guo, M. 2025b. CascadeV: An Implementation of Wurstchen Architecture for Video Generation. *arXiv preprint arXiv:2501.16612*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Lv, Z.; Si, C.; Pan, T.; Chen, Z.; Wong, K.-Y. K.; Qiao, Y.; and Liu, Z. 2025. DCM: Dual-Expert Consistency Model for Efficient and High-Quality Video Generation. *arXiv preprint arXiv:2506.03123*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*.

Nan, K.; Xie, R.; Zhou, P.; Fan, T.; Yang, Z.; Chen, Z.; Li, X.; Yang, J.; and Tai, Y. 2025. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation. In *The Thirteenth International Conference on Learning Representations*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Salimans, T.; and Ho, J. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*.

Sauer, A.; Boesel, F.; Dockhorn, T.; Blattmann, A.; Esser, P.; and Rombach, R. 2024a. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2024b. Adversarial Diffusion Distillation. In *European Conference on Computer Vision*, 87–103.

Shao, S.; Yi, H.; Guo, H.; Ye, T.; Zhou, D.; Lingelbach, M.; Xu, Z.; and Xie, Z. 2025. MagicDistillation: Weak-to-Strong Video Distillation for Large-Scale Few-Step Synthesis. *arXiv preprint arXiv:2503.13319*.

Singer, U.; Zohar, A.; Kirstain, Y.; Sheynin, S.; Polyak, A.; Parikh, D.; and Taigman, Y. 2024. Video Editing via Factorized Diffusion Distillation. In *European Conference on Computer Vision*, 450–466.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency models.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, F.-Y.; Huang, Z.; Bergman, A.; Shen, D.; Gao, P.; Lingelbach, M.; Sun, K.; Bian, W.; Song, G.; Liu, Y.; et al. 2024. Phased consistency models. *Advances in neural information processing systems*, 37: 83951–84009.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, Q.; Shi, Y.; Ou, J.; Chen, R.; Lin, K.; Wang, J.; Jiang, B.; Yang, H.; Zheng, M.; Tao, X.; et al. 2025. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8428–8437.

Wang, X.; Zhang, S.; Zhang, H.; Liu, Y.; Zhang, Y.; Gao, C.; and Sang, N. 2023b. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 657–666.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Yin, T.; Gharbi, M.; Park, T.; Zhang, R.; Shechtman, E.; Durand, F.; and Freeman, B. 2024a. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37: 47455–47487.

Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W. T.; and Park, T. 2024b. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6613–6623.

Zhang, H.; Chen, X.; Wang, Y.; Liu, X.; Wang, Y.; and Qiao, Y. 2025. Accvideo: Accelerating video diffusion model with synthetic dataset. *arXiv preprint arXiv:2503.19462*.