HPSD: HYBRID-POLICY SELF-DISTILLATION FOR TEXT-IMAGE-TO-VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

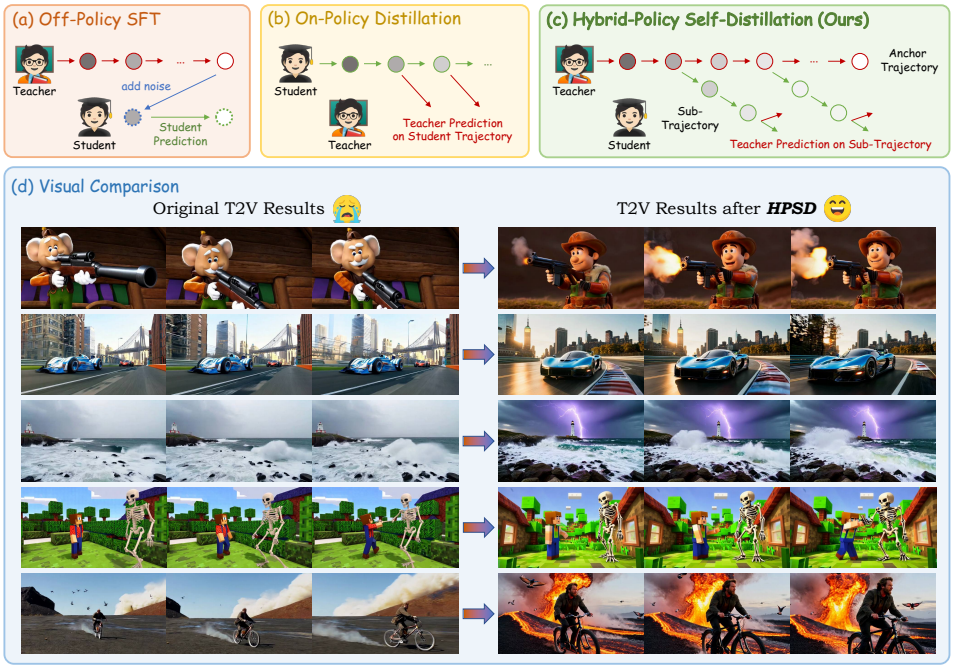


Figure 1: **Introduction to HPSD.** (a) Off-policy SFT supervises on fixed teacher endpoints, which drift from the evolving student policy. (b) On-policy distillation queries the teacher at student-visited states, yet suffers from condition-state mismatch in TI2V models. (c) HPSD anchors the student on the teacher trajectory and supervises roll-outs evolved under its own policy. (d) HPSD substantially boosts the base generation quality of TI2V models (WAN-2.2-TI2V-5B in this figure), especially in cinematic lighting, fine details, and balanced composition. **Prompts in the Appendix.**

ABSTRACT

Text-Image-to-Video (TI2V) models are an emerging unified architecture in which a single model simultaneously supports text-to-video (T2V) and image-to-video (I2V) generation. Given a high-quality first frame or a detailed textual prompt, TI2V models unlock substantially better visual quality than their T2V mode, raising a natural question: *can the capability elicited by such privileged conditions be internalized into the model's own base generation ability?* A common approach toward this goal is model self-distillation. However, the most straightforward solution, supervised fine-tuning, follows an off-policy strategy: its supervision is confined to teacher-generated endpoints from a fixed offline distribution rather than student-visited states, lacking precise correction tailored to the evolving policy. Recent on-policy distillation methods instead suffer from condition-state mismatch, where supervision is steered toward the given first frame instead of the student's actual content, misleading the denoising process. To achieve self-distillation that absorbs the teacher's privileged prior while retaining precise policy correction, in this work, we propose **H**ybrid-**P**olicy **S**elf-**D**istillation (**HPSD**), a novel self-distillation framework where a single TI2V model acts as both teacher

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and student under different conditions: the teacher operates in TI2V mode with a high-quality first frame and an enhanced prompt, while the student runs in the base T2V mode with only the vanilla prompt. Specifically, the student inherits off-policy teacher trajectory points as anchors, locally refines them toward its own policy, and finally receives velocity-level supervision on these self-generated roll-outs. Extensive experiments demonstrate that HPSD significantly improves T2V performance while also delivering notable TI2V gains, effectively strengthening the model's base generation ability.

1 INTRODUCTION

Recent years have witnessed remarkable progress in video generation, where diffusion/flow-based models (Ho et al., 2020; Song et al., 2020a;b; Lipman et al., 2022; Peebles & Xie, 2023) have achieved a leap in high-fidelity video synthesis (Wan et al., 2025; HaCohen et al., 2024; Kong et al., 2024; Bu et al., 2025; Zhou et al., 2025). Building on these advances, the field has evolved into diverse generation paradigms, among which text-to-video (T2V) (Guo et al., 2023; Kong et al., 2024; Yang et al., 2025) and image-to-video (I2V) (Blattmann et al., 2023; Zhang et al., 2023; Xing et al., 2023) are the most representative: the former synthesizes videos from textual descriptions, while the latter animates a given reference image into a dynamic sequence. These two paradigms are increasingly unified within a single architecture, giving rise to Text-Image-to-Video (TI2V) models (Lin et al., 2025; Wan et al., 2025; HaCohen et al., 2026; Ma et al., 2026). Taking a textual prompt and an optional first-frame image as input, TI2V models produce faithful and coherent videos with a unified generative framework, allowing T2V and I2V to benefit from shared latent representations and motion priors while offering flexible conditioning for diverse user intents.

In this work, we observe that within such a unified architecture, the TI2V mode conditioned on an additional high-quality first frame yields substantially better visual quality than the base T2V mode driven by a vanilla prompt, and that a detailed, well-crafted prompt alone brings considerable gains as well, as illustrated in Fig. 2. Moreover, these two conditions from different modalities are complementary: a carefully designed first frame combined with an enriched prompt further improves the generation quality beyond either alone. We refer to such inputs as *privileged conditions*, and to the underlying capability they awaken as the *condition-elicited capability* of TI2V models. Intuitively, privileged conditions inject additional content and motion priors into the denoising process, reducing generation uncertainty and improving video quality. Notably, these quality differences arise from the same model operating under different external conditions, raising a natural question: *can this condition-elicited capability be internalized into the model's own base generation ability?*

A common approach toward this goal is model self-distillation, which has long been studied in diffusion models (Yin et al., 2024b; Jiang et al., 2025). Nevertheless, the most straightforward solution, supervised fine-tuning (SFT) on teacher-generated videos, is inherently off-policy: the student is supervised only on teacher terminal outputs from a fixed offline distribution, which drift away from the states it actually visits as training proceeds, precluding state-aware precise correction to the evolving policy, as shown in Fig. 1 (a). Recently, inspired by the success of on-policy distillation (OPD) in large language models (LLMs) (Shenfeld et al., 2026; Yang et al., 2026; Zhao et al., 2026; He et al., 2026), a growing line of research has brought this paradigm to diffusion and flow models (Jiang et al., 2026; Fang et al., 2026; Li et al., 2026), where the teacher is queried at student-visited states to provide dense, state-wise supervision along the student's own roll-outs, effectively mitigating the exposure bias of off-policy training, as depicted in Fig. 1 (b). In TI2V models, however, the privileged image condition is imposed as a fixed first frame that remains clean throughout denoising, whereas the student generates every frame without such conditioning. The teacher is thus queried on mixed states where the prescribed first-frame content coexists with the student's own evolving content, and its supervision directs the denoising toward the former, conflicting with the latter—a failure we term *condition-state mismatch* (Fig. 3). These limitations underscore the need for a new policy structure that absorbs the teacher's privileged prior while retaining precise policy correction.

To address these issues, we propose **H**ybrid-**P**olicy **S**elf-**D**istillation (**HPSD**), a novel self-distillation framework where a single TI2V model acts as both teacher and student under different conditions: the teacher operates in TI2V mode under the privileged image and prompt conditions, while the student runs in the base T2V mode with a vanilla text prompt. The key idea is to let the student

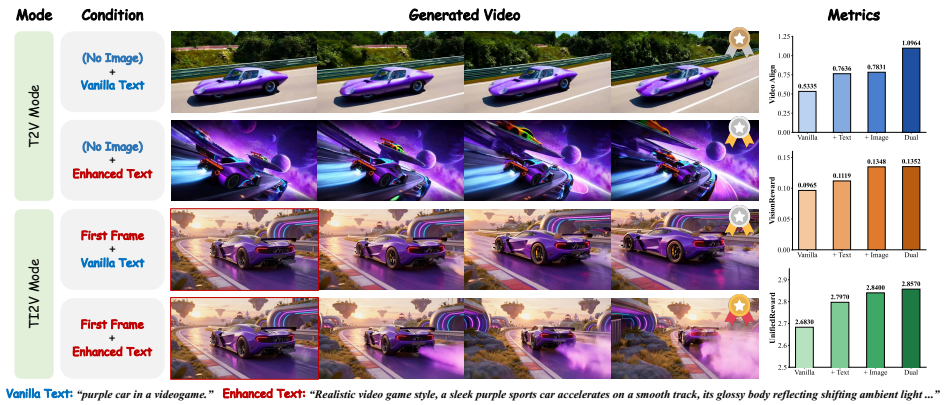


Figure 2: **Condition-Elicited Capability.** Conditioning a TI2V model on a high-quality first frame or an enhanced prompt clearly outperforms vanilla T2V. Their combination yields the best results across visuals and metrics, motivating HPSD's goal of internalizing this capability into the model.

start from the teacher's trajectory and finish under its own policy, which proceeds in three steps: (1) Before training, the privileged conditions are synthesized for each prompt with off-the-shelf generative models; (2) During training, the teacher first rolls out its full denoising trajectory under privileged conditions, yielding an off-policy *anchor trajectory* that carries the condition-elicited generation content; (3) The student then continues denoising from intermediate states of the anchor trajectory with its own velocity field, and is supervised by the teacher on the resulting *sub-trajectory*. Since the supervised states are anchored by the teacher's policy yet evolved by the student's, we term this design a *hybrid-policy* (Fig. 1 (c)), which offers two core advantages: (i) the student inherits the teacher's states prescribed by the privileged conditions, so that the teacher's supervision no longer conflicts with the sample content; (ii) the supervised states arise from the student's own roll-out and thus remain aligned with its current policy. By integrating off-policy anchoring with on-policy refinement, HPSD sidesteps both the exposure bias of off-policy training and the condition-state mismatch of on-policy distillation, offering a new principle for diffusion model distillation. Extensive experiments on representative unified TI2V models (WAN-2.2 (Wan et al., 2025) and LTX-2.3 (HaCohen et al., 2026)) demonstrate that HPSD surpasses both off-policy and on-policy baselines, substantially improving the T2V performance (Fig. 1 (d)) while also delivering notable TI2V gains, indicating that the teacher and the student improve in tandem and that the model's base generation ability is consistently enhanced.

Our contributions are threefold: (1) **Condition-Elicited Capability**: We identify that privileged conditions elicit stronger generation capability from unified TI2V models, and pose the new task of internalizing this capability into the model's base generation ability. (2) **Hybrid-Policy Self-Distillation**: We reveal the failures of off-/on-policy distillation in this setting and propose HPSD, a hybrid-policy self-distillation framework that supervises the student on states anchored by the teacher's trajectory and evolved by its own policy. (3) **Superior Performance**: Extensive experiments on two representative unified TI2V models validate that HPSD significantly strengthens the model's base generation ability beyond both off-/on-policy baselines.

2 Related Work

We highlight the most relevant studies here and discuss others in Section C in the Appendix.

Video Diffusion Models. Video diffusion models have evolved from early U-Net-based architectures (Guo et al., 2023; Blattmann et al., 2023; Zhang et al., 2023) toward diffusion transformers (DiTs) (Esser et al., 2024), giving rise to modern systems with remarkable visual fidelity and prompt adherence, such as OpenSora (Zheng et al., 2024), CogVideoX (Yang et al., 2025), LongCat-Video (Team et al., 2025), HunyuanVideo series (Kong et al., 2024; Wu et al., 2025), and WAN series (Wan et al., 2025). These models generally follow two paradigms: text-to-video (T2V) models that synthesize videos from textual descriptions, and image-to-video (I2V) models that animate a given reference image. More recently, Text-Image-to-Video (TI2V) models (Lin et al., 2025; Ma et al., 2026) such as WAN-2.2 (Wan et al., 2025) and LTX-2 (HaCohen et al., 2026) further unify

the two paradigms within a single set of weights, where an optional first-frame image is injected as a fixed clean condition into the denoising process. Beyond architecture design, a parallel line of research (Wang et al., 2025a; Cheng et al., 2025) improves generation quality by engineering richer input conditions at inference-time. In contrast, our HPSD paradigm internalizes the condition-elicited capabilities directly into the model's inherent base generation ability.

On-Policy Distillation for Generative Models. Initially grounded in large language models to mitigate exposure bias by matching distributions on student-generated sequences (Agarwal et al., 2024; Song & Zheng, 2026; Lu & Lab, 2025; Jang et al., 2026), on-policy distillation (OPD) has recently been adapted to diffusion and flow models to provide dense supervision along the student's own roll-outs. In this visual domain, existing methods primarily fall into two categories. The first is self-distillation, where a single model acts as both student and teacher; for instance, D-OPSD (Jiang et al., 2026) and OPSD-V (Liu et al., 2026) guide the student using teachers conditioned on real images and temporal contexts, respectively. The second involves multi-teacher distillation, with works like DiffusionOPD (Li et al., 2026), Flow-OPD (Fang et al., 2026), and DanceOPD (Zhou et al., 2026), composing diverse capabilities from multiple specialized teachers into a unified student via velocity matching or reward optimization. Distinct from these paradigms, our HPSD addresses the unique condition-state mismatch of OPD in TI2V models through a novel hybrid-policy design.

3 METHOD

3.1 PRELIMINARIES

Flow Matching. Flow matching (Liu et al., 2022; Lipman et al., 2022) trains a time-dependent velocity field $v_\theta(x_t, t, c)$ that transports samples from a Gaussian prior to the data distribution. Given a data sample x_0 and noise $\epsilon \sim \mathcal{N}(0, I)$, the interpolated state and target velocity at time $t \in [0, 1]$ are $x_t = (1-t)x_0 + t\epsilon$ and $v^* = \epsilon - x_0$, respectively. The model is optimized with $\mathcal{L}_{\text{FM}} = \mathbb{E}\|v_\theta(x_t, t, c) - v^*\|^2$. Sampling proceeds by integrating the learned velocity field from $t = 1$ to 0:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta(x_t, t, c), \tag{1}$$

where c denotes input conditions such as textual prompts or reference images.

TI2V Models. A TI2V model (Wan et al., 2025; HaCohen et al., 2026) supports both T2V and TI2V generation with a shared set of weights. In the T2V mode, the model is conditioned only on the text input, i.e., $v_\theta(x_t, t \mid c_{\text{txt}})$. In the TI2V mode, it is further conditioned on a given first-frame image c_{img}. Specifically, for TI2V generation, the first-frame condition is injected as a fixed clean latent and is excluded from the denoising process. For an F-frame video, the model is evaluated as:

$$v_\theta(\hat{x}_t, t \mid c_{\text{txt}}, c_{\text{img}}), \qquad \hat{x}_t = [\, c_{\text{img}},\, x_t^{(2:F)} \,], \tag{2}$$

where the first-frame slot always contains the clean condition c_{img}, while the remaining frames are denoised at time t. *Throughout, we use x for T2V states whose frames share a single noise level, and $\hat{x}$ for TI2V states with a clean first-frame slot.*

Distillation Paradigms. Consider a teacher velocity field $\tilde{v}$ and a student v_ϕ. Off-policy SFT fine-tunes the student on teacher-generated data. The teacher first completes a full denoising process to produce an endpoint x_0^{Tea}. The student is supervised on its noised version $x_t^{\text{Tea}} = (1-t)x_0^{\text{Tea}} + t\epsilon$:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_t \left\| v_\phi(x_t^{\text{Tea}}, t) - v^* \right\|^2, \tag{3}$$

where $v^* = \epsilon - x_0^{\text{Tea}}$ is the analytic target velocity. For brevity, we suppress the conditional inputs here. On-policy distillation methods (Jiang et al., 2026; Zhou et al., 2026) instead supervise states produced by the current student. The student rolls out from its own policy, $x_t^{\text{Stu}} \sim \text{Rollout}(v_\phi)$, and the teacher is evaluated at these student-visited states:

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_t \left\| v_\phi(x_t^{\text{Stu}}, t) - \text{sg}[\tilde{v}(x_t^{\text{Stu}}, t)] \right\|^2, \tag{4}$$

where $\text{sg}(\cdot)$ stands for the stop gradient operation.

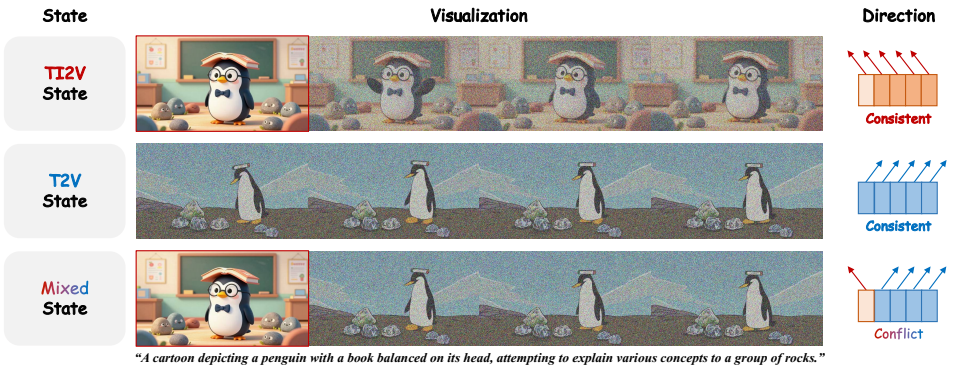


Figure 3: **Condition-State Mismatch.** While standard TI2V/T2V states maintain consistent content across frames, on-policy distillation creates an invalid mixed state by forcing a clean first frame alongside the student's unaligned T2V rollout, yielding conflicting, corrupted teacher supervision.

3.2 Observation and Analysis

As observed in Fig. 2, privileged conditions elicit a stronger generation capability from TI2V models: an enhanced prompt or a high-quality first frame alone already significantly improves over the base T2V mode, and their combination yields the best quality. To internalize this condition-elicited capability into the model's base generation ability, we formulate a self-distillation problem in which the teacher and the student are the *same* TI2V model evaluated under different input conditions: the teacher runs under privileged conditions (an enhanced prompt c_{txt}^{+} and a given first frame c_{img}), while the student operates in the base T2V mode with the original text prompt c_{txt}.

To be effective, this distillation must satisfy two competing requirements: providing state-aware precise correction to the evolving policy, while ensuring the teacher's supervision does not conflict with the student's actual sample content. While supervised fine-tuning (SFT) on teacher-generated videos successfully circumvents content conflicts, it remains fundamentally off-policy: as the student policy evolves, the fixed offline supervision states drift away from the states the student actually visits. Conversely, recent on-policy distillation methods (Jiang et al., 2026; Zhou et al., 2026) evaluate at student-visited states to offer localized correction, but suffer from a severe *condition-state mismatch* in the TI2V setting, as shown in Fig. 3. Rolling out in the T2V mode, the student generates an internally coherent latent sequence $x_t^{\text{Stu},\,(1:F)}$. However, since the TI2V teacher rigidly requires a clean first frame, querying it on a student's intermediate state creates a conflicted mixed input:

$$\hat{x}_t^{\text{Mismatch}} = [\, c_{\text{img}},\, x_t^{\text{Stu},\,(2:F)} \,]. \tag{5}$$

Here, the clean condition c_{img} is forced to coexist with the student's partially denoised frames, which depict a completely different, free-running T2V content, causing a severe temporal discontinuity. Essentially, this unaligned mixture constitutes an invalid input state for the model. The teacher thus outputs a corrupted velocity field, generating contradictory signals that mislead the student's denoising process toward c_{img}, rather than refining the student's own content. HPSD is structurally designed to resolve this dilemma and satisfy both requirements simultaneously.

3.3 Hybrid-Policy Self-Distillation

HPSD lets the student start from the teacher's trajectory and finish under its own policy, which proceeds in the following three steps. The overall framework of HPSD is illustrated in Fig. 4.

Privileged Condition Construction. Given a training prompt c_{txt}, we first synthesize the privileged conditions with off-the-shelf generative models: an external LLM rewrites c_{txt} into an enhanced prompt c_{txt}^{+} and designs a first-frame description c_{ff}, which an auxiliary text-to-image model then renders into the high-quality first frame c_{img}. All privileged conditions are precomputed offline and cached for training. Formally, this process can be expressed as:

$$c_{\text{txt}}^{+},\, c_{\text{ff}} \sim \mathcal{M}_{\text{rw}}\big(\cdot \mid c_{\text{txt}}\big), \qquad c_{\text{img}} \sim \mathcal{G}_{\text{t2i}}\big(\cdot \mid c_{\text{ff}}\big), \tag{6}$$

where $\mathcal{M}_{\text{rw}}$ denotes the LLM prompt rewriter, and $\mathcal{G}_{\text{t2i}}$ denotes the auxiliary text-to-image generator.

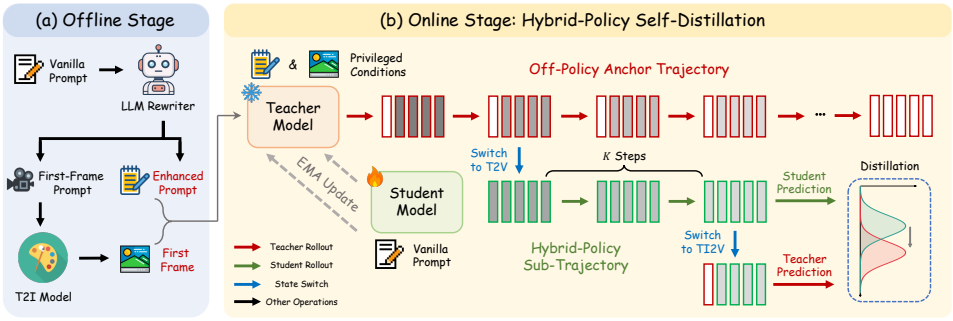


Figure 4: **Overview of HPSD.** (a) Offline Stage: Privileged conditions (enhanced prompt and first frame) are synthesized using auxiliary models. (b) Online Stage: The student evolves hybrid-policy sub-trajectories starting from the teacher's off-policy anchor states. Distillation on these states provides the student with precise policy correction anchored by the teacher's privileged content priors. *Varying gray shading denotes the noise level over time, while solid white indicates clean frames.*

Off-Policy Anchor Trajectory. The teacher first rolls out its full denoising trajectory under privileged conditions $(c^{+}_{\text{txt}}, c_{\text{img}})$, yielding an *off-policy anchor trajectory* that carries the condition-elicited generation content:

$$\hat{x}^{\text{Tea},\,(2:F)}_{t_{j+1}} = \hat{x}^{\text{Tea},\,(2:F)}_{t_j} + (t_{j+1} - t_j)\,\tilde{v}\big(\hat{x}^{\text{Tea}}_{t_j}, t_j \,\big|\, c^{+}_{\text{txt}}, c_{\text{img}}\big), \qquad \hat{x}^{\text{Tea},\,(2:F)}_{1} = \epsilon, \;\; \epsilon \sim \mathcal{N}(0, I). \tag{7}$$

From it we collect intermediate states $\{\hat{x}^{\text{Tea}}_{t_i}\}_{i=1}^{N}$ at anchor steps $\mathcal{A} = \{t_i\}_{i=1}^{N}$ spread along the trajectory. Each anchor state $\hat{x}^{\text{Tea}}_{t_i}$ is then converted into a *student-compatible* state $x^{\text{Tea}}_{t_i}$: since the teacher's first-frame slot always holds the clean condition (Eq. 2) rather than a denoising state at time t_i, we re-noise it to the current noise level,

$$x^{\text{Tea}}_{t_i} = \big[\,(1 - t_i)\,c_{\text{img}} + t_i\,\epsilon,\; \hat{x}^{\text{Tea},\,(2:F)}_{t_i}\,\big], \qquad \epsilon \sim \mathcal{N}(0, I), \tag{8}$$

so that all frames share the noise level t_i, matching what the student sees during T2V inference. This reconciles the teacher's privileged content with the student's input format, turning each anchor state into a valid starting point of the student's own roll-out.

Hybrid-Policy Sub-Trajectory. From each start point $x^{\text{Tea}}_{t_i}$, the student denoises K steps forward with its own velocity field,

$$x^{\text{Hyb}}_{t_{j+1}} = x^{\text{Hyb}}_{t_j} + (t_{j+1} - t_j)\,v_\phi\big(x^{\text{Hyb}}_{t_j}, t_j \,\big|\, c_{\text{txt}}\big), \qquad j = i, \ldots, i{+}K{-}1, \;\; x^{\text{Hyb}}_{t_i} := x^{\text{Tea}}_{t_i}, \tag{9}$$

yielding a short *hybrid-policy sub-trajectory* whose terminal state $x^{\text{Hyb}}_{t_{i+K}}$ inherits the teacher's content yet is evolved by the student's current policy. To supervise at this state, the teacher must be queried in its own input format: following Eq. 2, we re-impose the clean first frame onto $x^{\text{Hyb}}_{t_{i+K}}$,

$$\hat{x}^{\text{Hyb}}_{t_{i+K}} = \big[\,c_{\text{img}},\, x^{\text{Hyb},\,(2:F)}_{t_{i+K}}\,\big], \tag{10}$$

and evaluate the teacher at $\hat{x}^{\text{Hyb}}_{t_{i+K}}$. The student is then supervised on the frames being denoised:

$$\mathcal{L}_{\text{HPSD}} = \mathbb{E}_{t_i \sim \mathcal{A}} \left\| v_\phi\big(x^{\text{Hyb}}_{t_{i+K}}, t_{i+K} \,\big|\, c_{\text{txt}}\big) - \text{sg}\Big[\tilde{v}\big(\hat{x}^{\text{Hyb}}_{t_{i+K}}, t_{i+K} \,\big|\, c^{+}_{\text{txt}}, c_{\text{img}}\big)\Big] \right\|^2_{(2:F)}. \tag{11}$$

The sub-trajectory length K interpolates between the two paradigms: $K = 0$ reduces to off-policy distillation on the teacher's trajectory, while a larger K approaches on-policy supervision. Anchored by the teacher's trajectory and evolved by the student's policy, these intermediate states bridge both paradigms, which is why we denote this design as *hybrid-policy*. Note that the loss is applied from the second frame onward, as the teacher's first-frame velocity carries no denoising meaning under the clean-condition input format.

HPSD Training. In the offline stage, we synthesize and cache the privileged conditions $(c^{+}_{\text{txt}}, c_{\text{img}})$ for every training prompt. During online training, the process iterates over prompt batches. First, the teacher rolls out an anchor trajectory to provide student-compatible anchor states. Next, the student denoises for K steps from these anchors to form hybrid-policy sub-trajectories, receiving teacher supervision at the terminal states. The teacher is updated via an exponential moving average (EMA) of the student (Jiang et al., 2026), which stabilizes the target distribution and ensures that both roles improve in tandem. Alg. 1 in the Appendix outlines the complete HPSD training pipeline.

Table 1: Quantitative comparison of different methods on various TI2V backbones. The best results are in **bold**, while the second-best result is underlined. UR-v2-A, UR-v2-P and UR-v2-S represent the Alignment, Physics and Style dimensions of UnifiedReward-v2, respectively.

Backbone	Method	VideoAlign	VisionReward	UR-v1	UR-v2-A	UR-v2-P	UR-v2-S	HPS	CLIP
WAN-2.2	Vanilla T2V	0.5335	0.0965	2.683	2.802	3.167	3.100	0.2472	0.3684
	On-Policy Distillation	0.2613	0.0482	2.463	2.724	3.171	3.204	0.2561	0.2847
	D-OPSD (*Text*)	0.6379	0.1076	2.739	2.825	3.189	3.187	0.2493	0.3712
	Supervised Fine-Tuning	1.2046	0.1043	2.763	2.854	3.181	3.207	0.2670	0.3710
	HPSD (Ours)	**1.8753**	**0.1191**	**2.812**	**2.890**	**3.203**	**3.275**	**0.2815**	**0.3765**
LTX-2.3	Vanilla T2V	0.2307	0.0566	2.648	2.877	3.100	2.956	0.2068	0.3320
	On-Policy Distillation	0.0242	0.0405	2.594	2.825	3.089	3.093	0.2278	0.3048
	D-OPSD (*Text*)	0.4380	0.1008	2.824	**2.903**	3.173	3.114	0.2451	0.3808
	Supervised Fine-Tuning	0.9584	0.0826	2.733	2.873	3.068	3.131	0.2349	0.3626
	HPSD (Ours)	**1.5244**	**0.1123**	**2.827**	2.887	**3.198**	**3.182**	**0.2731**	**0.3853**



Figure 5: **Qualitative Comparisons with Baselines on WAN-2.2.** Best viewed zoomed in.

4 Experiments

4.1 Implementation Details

Training Datasets. We adopt the video training dataset used in Pref-GRPO (Wang et al., 2025b) as the prompt set for HPSD and all baselines, which consists of ∼50K prompts providing broad coverage across diverse themes and subject categories. For privileged condition construction, we employ Qwen3.6-27B (Qwen Team, 2026) as the LLM rewriter and Z-Image-Turbo (Team, 2025) as the high-quality first-frame generator. Further details and additional experiments with an alternative first-frame generator are deferred to Section B and Section E in the Appendix, respectively.

Base Models. Experiments are conducted on two leading TI2V models: WAN-2.2-TI2V-5B (Wan et al., 2025) and LTX-2.3 (HaCohen et al., 2026). Both models are trained at their native resolutions and inference steps: 1280×704 with 50 steps for WAN-2.2, and 768×512 with 30 steps for LTX-2.3. For LTX-2.3, we only consider its first-stage generation and do not adopt the second-stage super-resolution and refinement. For efficiency, both models use 33 frames during training.

Baselines for Comparison. The compared methods encompass: (i) the base models in their T2V mode (Wan et al., 2025; HaCohen et al., 2026); (ii) Off-policy Supervised Fine-Tuning with flow-matching loss (Lipman et al., 2022); (iii) On-Policy Distillation; (iv) D-OPSD (Jiang et al., 2026), a recent self-distillation method for diffusion models. Note that D-OPSD injects image information through a multimodal encoder, whereas the evaluated TI2V backbones use VAE-encoded first frames. To retain their native conditioning interfaces, we evaluate a text-only adaptation of D-OPSD.

Training Details. Unless specified otherwise, all experiments share the following setup. We train for 500 steps on 8 H200 GPUs with a batch size of 1 per GPU, using the AdamW optimizer with a learning rate of 1×10^{-4} and `bfloat16` (bf16) mixed-precision training for efficiency. The anchor steps $\mathcal{A}$ are randomly sampled from the anchor trajectory, while the sub-trajectory length K is set to 3 and the EMA decay rate is set to 0.999. All methods use LoRA with $r = 32, \alpha = 64$.

Table 2: Quantitative results on VBench metrics. The best results are in **bold**, while the second-best result is underlined. SC: Subject Consistency; BC: Background Consistency; MS: Motion Smoothness; DD: Dynamic Degree; AQ: Aesthetic Quality; IQ: Imaging Quality.

Backbone	Method	VBench-SC	VBench-BC	VBench-MS	VBench-DD	VBench-AQ	VBench-IQ
WAN-2.2	Vanilla T2V	0.9654	0.9667	0.9878	**0.560**	0.5773	0.6981
	On-Policy Distillation	0.5241	0.7284	0.9799	0.516	0.6066	0.6683
	D-OPSD (*Text*)	0.9691	0.9692	0.9874	0.404	0.5889	0.6300
	Supervised Fine-Tuning	0.9679	0.9675	0.9837	0.526	0.6095	**0.7049**
	HPSD (Ours)	**0.9722**	**0.9705**	**0.9880**	0.422	**0.6343**	0.6996
LTX-2.3	Vanilla T2V	0.9388	0.9467	0.9901	0.330	0.5791	0.6517
	On-Policy Distillation	0.5337	0.7715	0.9899	**0.366**	0.5800	0.6345
	D-OPSD (*Text*)	0.9419	0.9454	0.9897	0.236	0.5994	0.6383
	Supervised Fine-Tuning	0.9466	0.9465	0.9903	0.278	0.5994	0.6723
	HPSD (Ours)	**0.9515**	**0.9554**	**0.9905**	0.302	**0.6375**	**0.6886**

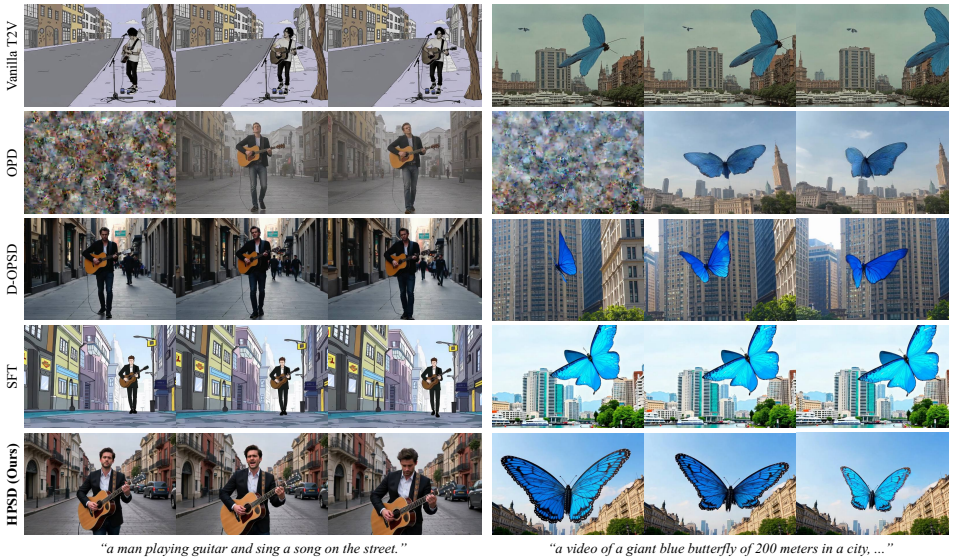


Figure 6: **Qualitative Comparisons with Baselines on LTX-2.3.** Best viewed zoomed in.

Evaluation Details. The evaluation set consists of 500 diverse video prompts sampled from the VideoDPO (Liu et al., 2024) and VideoFeedback (He et al., 2024) datasets. For a thorough evaluation, a diverse set of metrics is employed: (i) Video Reward Models, including VideoAlign (Liu et al., 2025), VisionReward (Xu et al., 2026), and UnifiedReward-v1/v2 (UR-v1/v2, unified image-video reward models) (Wang et al., 2025c); (ii) Image Reward Models, where we adopt frame-wise HPS Score (Wu et al., 2023) and frame-wise CLIP Score (Radford et al., 2021) to assess per-frame quality; and (iii) VBench (Huang et al., 2024) for comprehensive video generation evaluation.

4.2 Main Results

Quantitative Evaluation. The quantitative results are presented in Tab. 1 and Tab. 2. For reward metrics (Tab. 1), HPSD consistently outperforms baselines, topping all eight reward metrics on WAN-2.2 and seven on LTX-2.3, notably boosting VideoAlign by $\sim$50% over the second-best SFT. Conversely, standard OPD severely degrades T2V performance due to its condition-state mismatch in training, leading to early-frame blurriness and content incoherence. While D-OPSD improves over vanilla T2V, HPSD effectively distills privileged first frames in TI2V models and markedly surpasses it overall. On VBench (Tab. 2), HPSD leads on most dimensions for both models. Although the Dynamic Degree is lower, we argue that it does not imply degraded video quality; rather, we observe enhanced temporal consistency, fewer abrupt changes, and better physical plausibility, evidenced by the improved UR-v2-Physics score in Tab. 1. We provide visual examples in Fig. 11 in the Appendix, where HPSD demonstrates stronger adherence to physical laws.

Qualitative Comparison. Visual comparisons in Fig. 5 and Fig. 6 corroborate the metrics. Vanilla T2V outputs are often desaturated with coarse textures, whereas On-Policy Distillation collapses entirely (e.g., noisy early frames and ghosting artifacts), stemming from the aforementioned condition-state mismatch. While SFT and D-OPSD produce plausible content, they still lack fine details, con-

Table 3: Ablation experiments on HPSD components and hyperparameters.

Components	Values & Choices	VideoAlign	VisionReward	UR-v1	UR-v2-A	UR-v2-P	UR-v2-S
(a) Sub-Trajectory Length K	$K = 0$	1.5587	0.1131	2.750	2.853	3.163	3.223
	$K = 1$	1.6653	0.1144	2.761	2.846	3.180	3.232
	K = 3	**1.8753**	0.1191	**2.812**	**2.890**	**3.203**	**3.275**
	$K = 5$	1.7794	**0.1194**	2.790	2.865	3.190	3.260
	$K = 7$	1.7130	0.1168	2.777	2.869	3.189	3.251
(b) Privileged Conditions c^+_{txt} & c_{img}	Vanilla T2V	0.5335	0.0965	2.683	2.802	3.167	3.100
	$+c_{\text{img}}$	1.5006	0.1157	2.787	2.855	3.194	3.247
	$+\mathbf{c^+_{txt} + c_{img}}$ **(HPSD)**	**1.8753**	**0.1191**	**2.812**	**2.890**	**3.203**	**3.275**
(c) Effect on TI2V Mode	Vanilla TI2V	0.7831	**0.1348**	2.840	2.926	3.211	**3.273**
	TI2V after HPSD	**1.2139**	0.1344	**2.859**	**2.932**	**3.215**	3.271



Figure 7: **Effects of HPSD on TI2V mode.** The red boxes indicate the shared first frames.

sistent subjects, and cinematic lighting. In contrast, HPSD delivers significantly sharper textures, richer colors, and superior prompt adherence (e.g., the cinematic poodle and vividly textured butterfly), visually confirming the successful internalization of condition-elicited capabilities. Additional visual comparison results are presented in Section F in the Appendix.

4.3 Ablation Study

The ablation studies are conducted on the WAN-2.2 model, with the results presented in Tab. 3.

Effects of Sub-Trajectory Length K. The sub-trajectory length K acts as a tunable knob interpolating between off-policy and on-policy paradigms. As shown in Tab. 3 (a), HPSD's performance initially improves but later deteriorates as K increases, suggesting that an excessively short sub-trajectory remains overly off-policy, failing to reach the student's actual distribution, whereas a prolonged K pushes the framework toward pure on-policy distillation, drifting too far from the teacher's prior and exacerbating the condition-state mismatch (Fig. 3). Therefore, $K = 3$ is chosen to balance the teacher's condition-elicited anchoring with the student's policy-aligned correction.

Effects of Privileged Conditions. Tab. 3 (b) reveals a clear step-wise improvement when scaling the teacher's privileged conditions. Compared to Vanilla T2V, distilling only the privileged image ($+c_{\text{img}}$) brings substantial gains, while combining both privileges ($+c^+_{\text{txt}} + c_{\text{img}}$) achieves the peak performance. This validates that richer conditions unlock stronger teacher capabilities, yielding higher-quality guidance for internalization. Since text conditions do not trigger the condition-state mismatch and are readily handled by standard OPD, we do not include a text-only ablation.

Effects of HPSD on TI2V Mode. Although optimized for T2V generation, HPSD simultaneously enhances the model's TI2V performance on most metrics (Tab. 3 (c)), improving VideoAlign by $\sim$55%. These gains stem from our asymmetric prompt distillation: forcing the student, which operates on a vanilla prompt, to align with the teacher using an enhanced prompt intrinsically trains the model to better interpret and execute plain textual inputs, regardless of first frame conditioning. Consequently, HPSD yields richer physical interactions and superior prompt adherence (as illustrated in Fig. 7), confirming an intrinsic enhancement of the model's shared generation priors.

5 Conclusion

We propose Hybrid-Policy Self-Distillation (HPSD) to internalize the superior, condition-elicited capabilities of TI2V models into their base generation ability. To resolve the lack of state-aware precise correction in off-policy methods and the condition-state mismatch in on-policy approaches, HPSD supervises the student on sub-trajectories anchored by the teacher's privileged states but evolved under its own policy. Experiments on WAN-2.2 and LTX-2.3 demonstrate that HPSD significantly boosts base T2V performance and concurrently enhances the TI2V mode, effectively strengthening the model's inherent generation ability and offering a novel distillation paradigm.

AI USE STATEMENT

In this paper, we used generative AI tools solely for light language editing and polishing. We take responsibility for the content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work builds on publicly available datasets, model architectures, and evaluation protocols, with all experiments conducted following standard research practices. To the best of our knowledge, HPSD does not introduce ethical concerns specific to the proposed distillation framework beyond the broader societal risks associated with video generation models.

REPRODUCIBILITY STATEMENT

In alignment with the principles of transparent research, we strive to make our experimental results seamlessly reproducible for the academic community. To achieve this, we have provided comprehensive hyperparameter configurations, prompt designs, and implementation details. Furthermore, the complete source code of HPSD will be fully open-sourced. At this stage, an anonymized implementation is provided in the supplementary material (described in Section J of the Appendix). We hope this framework will serve as a valuable reference for future studies on diffusion model distillation, inspiring new methodological innovations and driving continuous momentum in the domain.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

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.

Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Bytheway: Boost your text-to-video generation model to higher quality in a training-free way. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12999–13008, 2025.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, et al. Vpo: Aligning text-to-video generation models with prompt optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15636–15645, 2025.

Javier Martín Daniel Verdú. Flux.1 lite: Distilling flux1.dev for efficient text-to-image generation. 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Gongfan Fang, Kunjun Li, Xinyin Ma, and Xinchao Wang. Tinyfusion: Diffusion transformers learned shallow. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18144–18154, 2025.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

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.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *ArXiv*, abs/2406.15252, 2024. URL https://arxiv.org/abs/2406.15252.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam H Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227, 2026.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven Hoi. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026.

Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. Stiv: Scalable text and image conditioned video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16249–16259, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. Opsd-v: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17719–17728, 2025.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, et al. Scaling mixture-of-experts video pretraining for embodied intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14297–14306, 2023.

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

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report, 2025. URL https://arxiv.org/abs/2510.22200.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, et al. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

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 *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Wei Zhou, Xiongwei Zhu, Zelin Xu, Bo Dong, Lixue Gong, Yongyuan Liang, Meng Chu, Leigang Qu, Lingdong Kong, Wei Liu, et al. Danceopd: On-policy generative field distillation. *arXiv preprint arXiv:2606.27377*, 2026.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-a-video: Training-free video relighting via progressive light fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13315–13325, October 2025.

A Appendix

In the Appendix, we provide additional implementation details (Section B), additional related works (Section C), the algorithm of HPSD (Section D), additional experimental results (Section E), additional qualitative samples (Section F), all text prompts used in both the main paper and Appendix (Section G), examples of privileged conditions (Section H), the limitation of our method (Section I), as well as the description of supplementary material (Section J), to complement the main paper.

B Additional Implementation Details

B.1 Hyperparameter Configuration

The detailed hyperparameter settings used in this paper are listed in Tab. 4. Unless otherwise specified, these parameters remain consistent across all experiments.

Table 4: Hyperparameter settings in our experiments.

Parameter	Value	Parameter	Value
Base models	WAN-2.2, LTX-2.3	LoRA settings	$r = 32, \alpha = 64$
Training frames	33	Anchor-trajectory length (WAN-2.2)	50
Train batch size (per GPU)	1	Anchor-trajectory length (LTX-2.3)	30
The number of GPUs	8	Optimizer	AdamW
Training steps	500	Learning rate	1×10^{-4}
Training guidance scale	1.0	Weight decay	0.0
Anchor steps ($\lvert\mathcal{A}\rvert$)	6	Mixed precision	`bfloat16`
Resolution (WAN-2.2)	1280×704	EMA decay rate	0.999
Resolution (LTX-2.3)	768×512	Sub-trajectory length (K)	3

B.2 Training Convergence

To determine the training budget, we investigate the convergence behavior of HPSD by monitoring its VisionReward score on WAN-2.2. As illustrated in Fig. 8, the metric improves rapidly in the early stage and plateaus around 500 steps, with negligible fluctuation thereafter, indicating that the model has reached an approximately converged state. Based on this empirical observation, we adopt 500 steps (trained on 8 GPUs) as the default configuration for all our main experiments.



Figure 8: **Training curve of HPSD.**

B.3 Prompts for LLM Rewriter

To construct the privileged conditions, we employ an external LLM (Qwen3.6-27B) as the prompt rewriter. The specific system prompts designed for these tasks are provided in Fig. 9 and Fig. 10.

Enhanced Prompt Generation. In the full HPSD pipeline, the enhanced prompt c_{txt}^{+} is utilized alongside the privileged first-frame condition c_{img}. Since the visual aesthetics and spatial layout of the video are already anchored by this high-quality first frame, we design a system prompt that directs the LLM to focus primarily on temporal dynamics, as illustrated in Fig. 9.

First-Frame Prompt Generation. To guide the auxiliary T2I model in synthesizing a structurally sound and aesthetically pleasing first frame c_{img}, the original video prompt c_{txt} is converted into an image generation prompt c_{ff}. Specifically, the LLM is instructed to translate dynamic actions into a static, ready-to-move opening posture, as detailed in the prompt template of Fig. 10.

Prompt Template for Enhanced Prompt Generation

You are a prompt engineer specialized for TI2V (text-to-video with a given first frame).
Your task is to rewrite a given Video Prompt into a motion-optimized TI2V prompt that works with an already provided high-quality first frame.
Your core goal is to improve in-scene motion executability while preserving intent.
You will be provided with a complete Video Prompt.

Rewrite rules:
1) Preserve intent: keep subject, scene, style, and key objects; do not add conflicting settings.
2) Motion-first: focus on motion of existing elements, such as body micro-actions, facial changes, fabric movement, light/shadow variation, smoke/water/foliage dynamics.
3) Clear timeline: use concise temporal connectors (first, then, finally).
4) Moderate detail: avoid verbose quality stacking; prioritize actionable motion cues.
5) Output only the final rewritten prompt text with no explanation or list formatting.
6) Never output any reasoning process, analysis steps, or meta explanation.
7) Preferred length: around 80-130 words, prioritizing motion stability and temporal consistency.

Examples:
- [Input]: A woman standing by the sea.
  [Output]: At seaside dusk, a woman stands in the wind as her hair sways continuously and the hem of her dress rises and settles in a gentle rhythm. She first lifts her gaze toward the horizon, then slowly turns her head, with natural breathing and subtle shoulder motion. Waves repeatedly break behind her, sea spray drifts in and out, and reflections on wet ground shimmer slightly.

- [Input]: A deer in the forest.
  [Output]: In a misty morning forest, a deer pauses briefly, ears twitching, then begins to step forward at a calm pace. It sniffs the air, occasionally raises its head, and shows subtle neck and back movement with each breath. Nearby leaves sway lightly, grass tips tremble, and thin fog moves slowly between trees. Keep the scene continuous and stable, emphasizing natural motion of the deer and surrounding elements rather than dramatic camera movement.

- [Input]: An ancient-style man drinking tea in a courtyard.
  [Output]: In a traditional courtyard, a man sits at a tea table, lifts the cup, and his sleeve falls naturally as small ripples spread across the tea surface. He takes a sip, then sets the cup down with a light fingertip touch on the rim, raising his eyes slowly afterward. Bamboo shadows sway softly, a hanging chime moves slightly, and a thin trail of steam rises and disperses near the table.

Now, rewrite the given Video Prompt into a motion-optimized TI2V prompt.

[Input Video Prompt] : {VIDEO_PROMPT}

Figure 9: **Prompt Template for Enhanced Prompt Generation.**

Prompt Template for First-Frame Prompt Generation

Act as an expert AI Visual Director and Prompt Engineer specializing in Image-to-Video (I2V) and Text-to-Image (T2I) workflows.
Your task is to design a highly detailed Text-to-Image prompt that will serve as the first frame for a video generation model.
You will be provided with a complete Video Prompt.

Follow these rules:
1. Convert dynamic actions into a static, ready-to-move opening frame.
2. Preserve the subject, environment, style, camera angle, lighting, and atmosphere.
3. Output only one English image prompt, without explanations.

Examples:
- [Input]: A young woman walks along a rainy city street at night, turning to look at a neon sign, then continues walking as her umbrella spins.
  [Output]: A young woman standing mid-stride on a rain-soaked city street at night, holding an open umbrella, her body slightly turned toward a glowing pink-and-blue neon sign; wet asphalt reflecting the neon glow, soft drizzle catching the streetlight, shallow depth of field, cinematic moody atmosphere, 35mm lens, eye-level angle.

- [Input]: A golden retriever runs across a sunny beach, splashing through shallow water, shakes off water, then lies down in the sand.
  [Output]: A golden retriever poised at the water's edge on a sunlit beach, front paws just touching the shallow waves, alert and ready to spring; bright midday sun, sparkling turquoise water, soft white sand, vibrant natural lighting, wide-angle low shot, photorealistic.

- [Input]: Steam rises from a cup of coffee on a wooden table as morning sunlight streams through a window, curtains gently swaying.
  [Output]: A ceramic cup of hot coffee on a rustic wooden table, a faint wisp of steam just beginning to curl upward, soft golden morning light streaming through a nearby window casting long warm streaks, sheer curtains slightly lifted, cozy quiet atmosphere, still-life composition, 50mm lens, warm tones.

- [Input]: An ancient-style swordsman stands in a bamboo grove, draws his sword slowly, then turns to face the wind as leaves swirl around him.
  [Output]: An ancient swordsman standing poised in a misty bamboo grove, hand resting on the hilt of an undrawn sword at his side, calm determined expression; tall green bamboo framing the shot, soft diffused fog, shafts of pale light filtering through the canopy, three-quarter front view, balanced composition.

Now, convert the given Video Prompt.

[Input Video Prompt] : {VIDEO_PROMPT}

Figure 10: **Prompt Template for First-Frame Prompt Generation.**



Figure 11: **Improved Temporal Consistency & Physical Plausibility.** (Left) Vanilla T2V exhibits limited temporal consistency, with the paddle appearing abruptly, while HPSD ensures coherent subject retention. (Right) Vanilla T2V lacks physical realism, generating nearly static liquid, whereas HPSD renders more natural fluid dynamics.

C Additional Related Works

Distillation for Diffusion Models. Diffusion model distillation primarily focuses on step distillation (Yin et al., 2024b; Sauer et al., 2024; Luo et al., 2023; Yin et al., 2024a) and knowledge distillation (Daniel Verdú, 2024; Yu et al., 2024). To accelerate the sluggish sampling process, step distillation compresses a many-step teacher into a few-step student, achieved either by imitating intermediate denoising transitions (Luo et al., 2023; Meng et al., 2023; Salimans & Ho, 2022) or aligning output distributions at specific timesteps (Luo et al., 2025; Yin et al., 2024b;a). Conversely, knowledge distillation transfers priors from a stronger teacher to improve the student's generation quality (Fang et al., 2025; Wu et al., 2026). Unlike this line of study, we frame capability internalization as a self-distillation problem across different conditionings within the same model.

D Algorithm

The overall training procedure of the proposed HPSD is detailed in Algorithm 1.

Algorithm 1 Hybrid-Policy Self-Distillation (HPSD)

Require: TI2V model with student v_ϕ and EMA teacher $\tilde{v}$; prompt set $\mathcal{D}$; sub-trajectory length K; anchor steps $\mathcal{A} = \{t_i\}_{i=1}^{N}$

// Offline stage: privileged-condition construction

for all $c_{\text{txt}} \in \mathcal{D}$ **do**
  Enhance the prompt $c_{\text{txt}} \to c_{\text{txt}}^{+}$ and synthesize the first frame c_{img} ▷ Eq. 6
end for

// Online stage: hybrid-policy self-distillation

for each training step **do**
  Sample a prompt batch $\{c_{\text{txt}}\} \subset \mathcal{D}$; load cached $(c_{\text{txt}}^{+}, c_{\text{img}})$
  // Off-policy anchor trajectory
  Roll out the teacher $\tilde{v}(\cdot \mid c_{\text{txt}}^{+}, c_{\text{img}})$; collect anchor states $\{\hat{x}_{t_i}^{\text{Tea}}\}$ ▷ Eq. 7
  Convert each anchor into a student-compatible state $x_{t_i}^{\text{Tea}}$ ▷ Eq. 8
  // Hybrid-policy sub-trajectory
  for all anchors $x_{t_i}^{\text{Tea}}$ **do**
    Denoise K steps with the student $v_\phi(\cdot \mid c_{\text{txt}}) \to x_{t_{i+K}}^{\text{Hyb}}$ ▷ Eq. 9
    Re-impose the clean first frame to build the teacher query $\hat{x}_{t_{i+K}}^{\text{Hyb}}$ ▷ Eq. 10
  end for
  // Velocity-level matching
  Compute $\mathcal{L}_{\text{HPSD}}$ and update the student v_ϕ ▷ Eq. 11
  Update the teacher $\tilde{v}$ with the EMA of the student
end for

E Additional Experimental Results

In the main experiments, we employed Z-Image-Turbo as the first-frame generator to construct the privileged image conditions. To validate the generalizability of HPSD, we conduct an additional study by utilizing a different first-frame generator, specifically Flux.2-Klein-4B (Labs, 2025). As shown in Tab. 5, HPSD consistently maintains its performance superiority over the original T2V

Table 5: Experimental results using Flux.2-Klein-4B as the first-frame generator.

Backbone	Method	VideoAlign	VisionReward	UR-v1	UR-v2-A	UR-v2-P	UR-v2-S	HPS	CLIP
WAN-2.2	Vanilla T2V	0.5335	0.0965	2.683	2.802	3.167	3.100	0.2472	0.3684
	HPSD (Ours)	**1.6587**	**0.1167**	**2.793**	**2.865**	**3.171**	**3.252**	**0.2819**	**0.3785**



Figure 12: **Comparison between HPSD (with Flux.2-Klein-4B) and vanilla WAN-2.2.**

results, demonstrating that our capability internalization framework is highly robust and generalizes well across various auxiliary text-to-image models.

Furthermore, we provide qualitative examples in Fig. 12. Visual inspections corroborate the quantitative findings, demonstrating that HPSD consistently delivers high-fidelity, temporally coherent videos across different first-frame generators utilized during the offline construction stage.

F Additional Qualitative Samples

In this section, we provide further visual results. Specifically, Fig. 14-22 present additional comparisons between WAN-2.2 and the baselines, while Fig. 23-30 showcase further results for LTX-2.3. Videos generated using the same prompts with different random seeds are shown in Fig. 31.

G Text Prompts for Video Generation

Text prompts used to generate videos in this paper are provided in Tab. 6 and Tab. 7.

H Privileged Condition Examples

In this section, we present concrete examples of the privileged conditions utilized during the HPSD training process. As depicted in Fig. 13, we showcase the original vanilla text prompts alongside their corresponding enhanced prompts (c_{txt}^{+}) and synthesized high-quality first frames (c_{img}). These privileged inputs are leveraged by the teacher to produce the condition-elicited anchor trajectories.

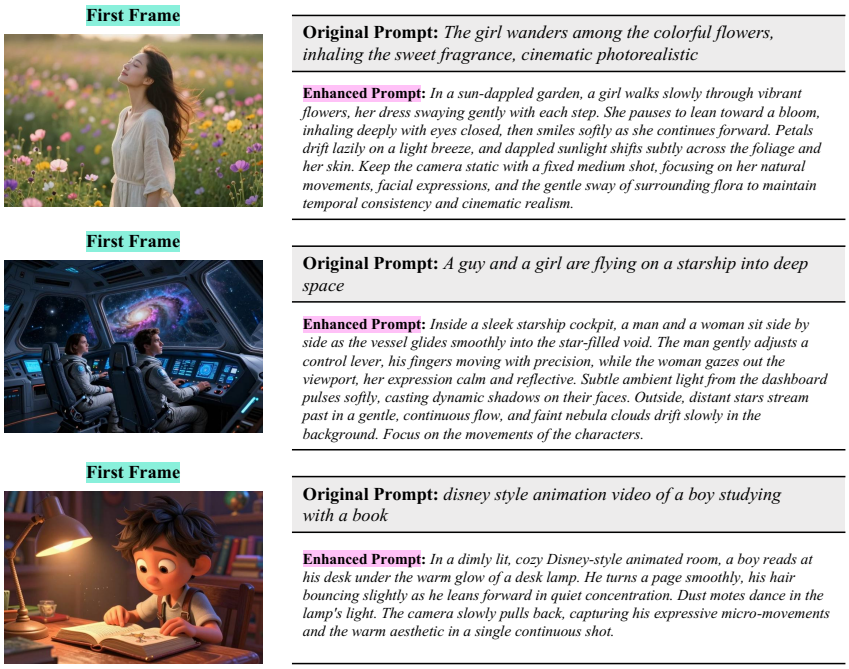


Figure 13: **Privileged Condition Examples.**

I Limitation and Discussion

While HPSD significantly enhances the base T2V generation capability, it introduces certain computational overheads. First, constructing privileged conditions requires querying auxiliary LLM and T2I models, which incurs extra data synthesis costs. However, this overhead is a one-time offline process and can be effectively mitigated by deploying distilled or quantized generative models (e.g., Z-Image-Turbo (Team, 2025) used in our experiments). Moreover, the performance ceiling of HPSD can naturally scale with the continuous advancements of these auxiliary models. Second, during training, rolling out the student to form the hybrid-policy sub-trajectory requires additional forward passes compared to SFT and OPD. Nevertheless, these intermediate steps operate strictly in inference mode and do not require storing gradients, thereby avoiding substantial memory overhead.

J Source Code and Demo Video

We provide the code implementation (for WAN-2.2) and a demo video of HPSD for reference in the zip archive included in the supplementary material.

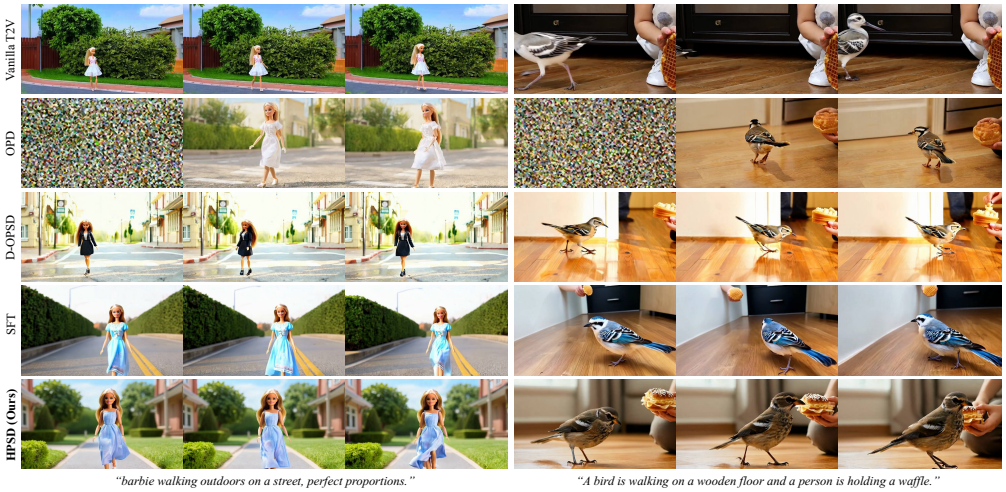


Figure 14: **More qualitative comparison results on WAN-2.2 (1/6).** Best viewed zoomed in.

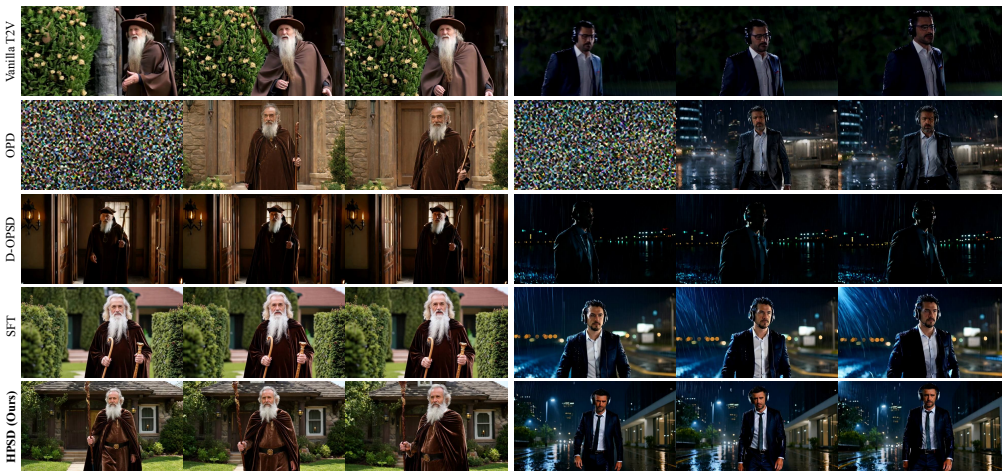


Figure 15: **More qualitative comparison results on WAN-2.2 (2/6).** Best viewed zoomed in.

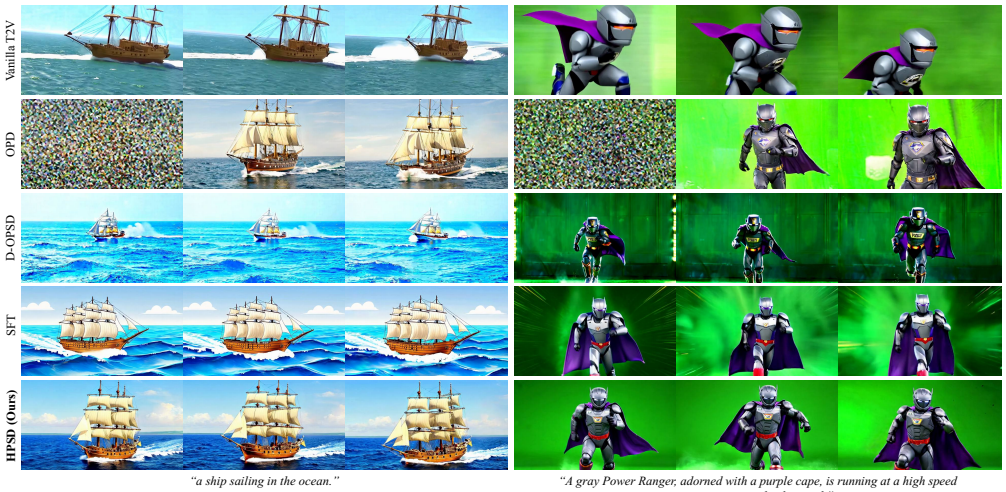


Figure 16: **More qualitative comparison results on WAN-2.2 (3/6).** Best viewed zoomed in.



Figure 17: **More qualitative comparison results on WAN-2.2 (4/6).** Best viewed zoomed in.

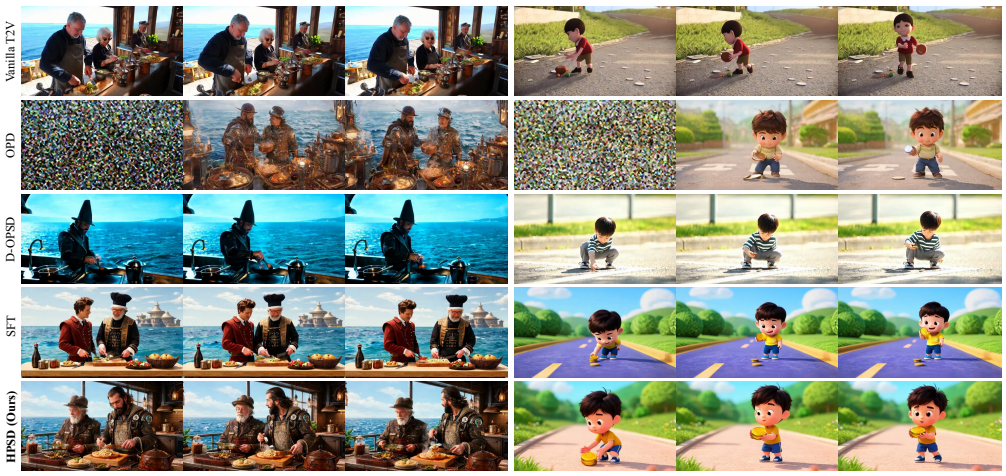


Figure 18: **More qualitative comparison results on WAN-2.2 (5/6).** Best viewed zoomed in.

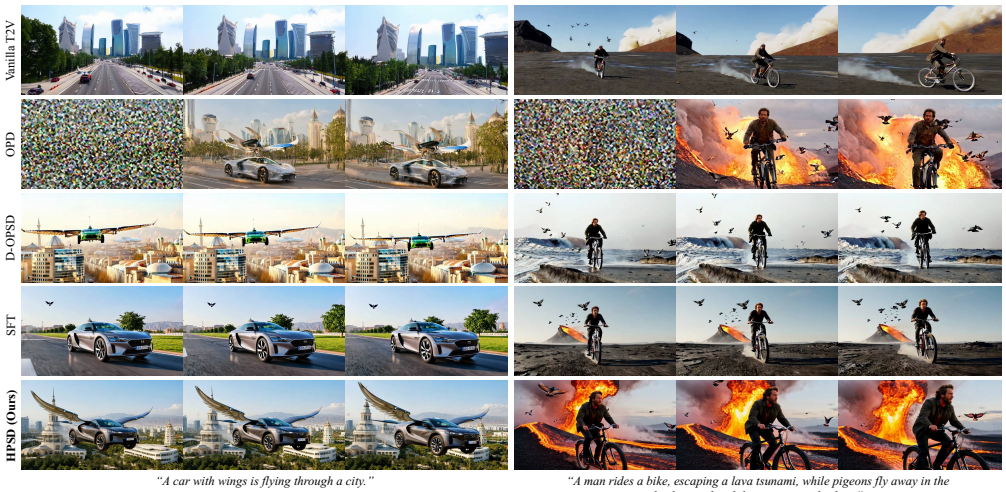


Figure 19: **More qualitative comparison results on WAN-2.2 (6/6).** Best viewed zoomed in.



Figure 20: **Comparison between HPSD and vanilla WAN-2.2 (1/3).** Best viewed zoomed in.

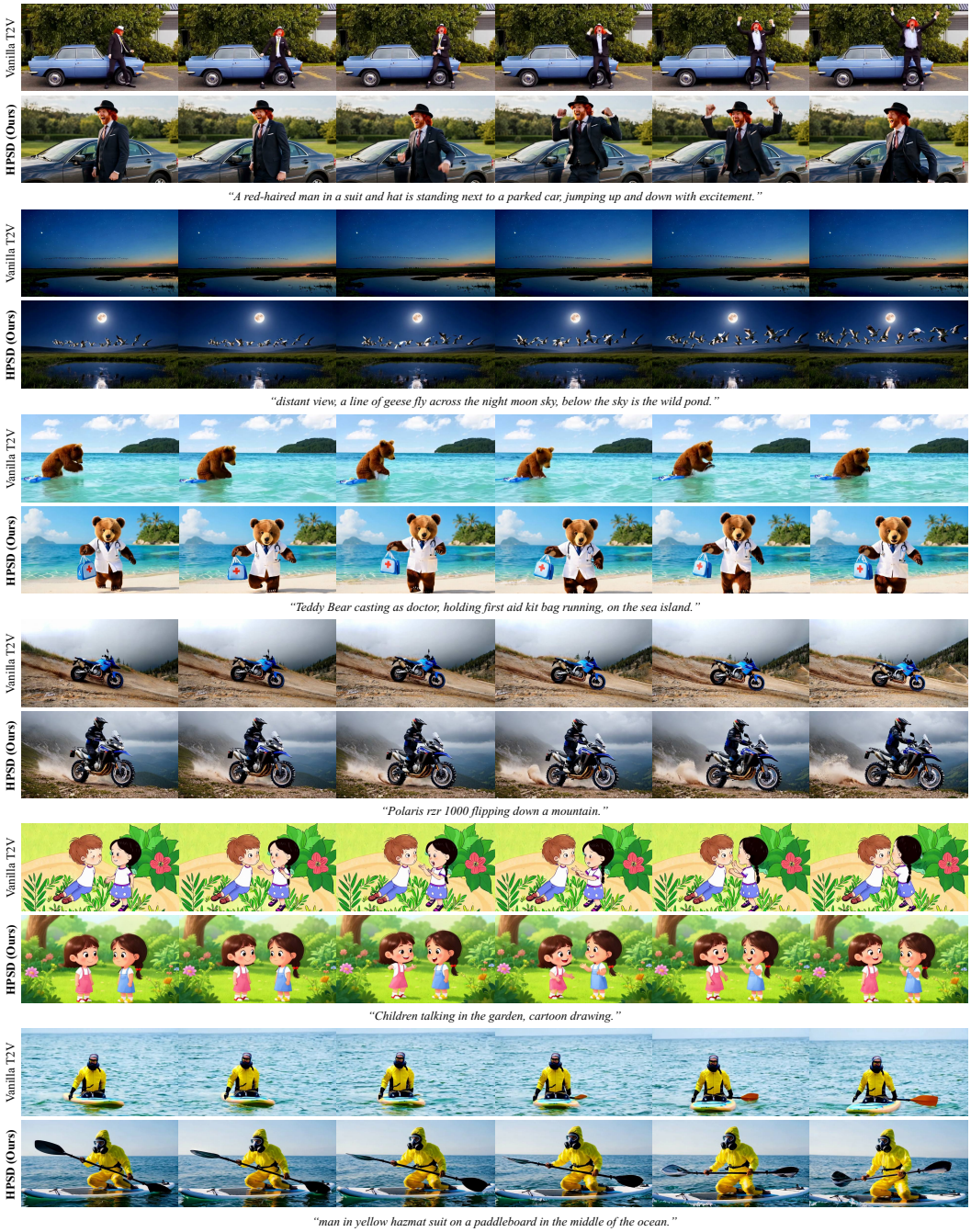


Figure 21: **Comparison between HPSD and vanilla WAN-2.2 (2/3).** Best viewed zoomed in.

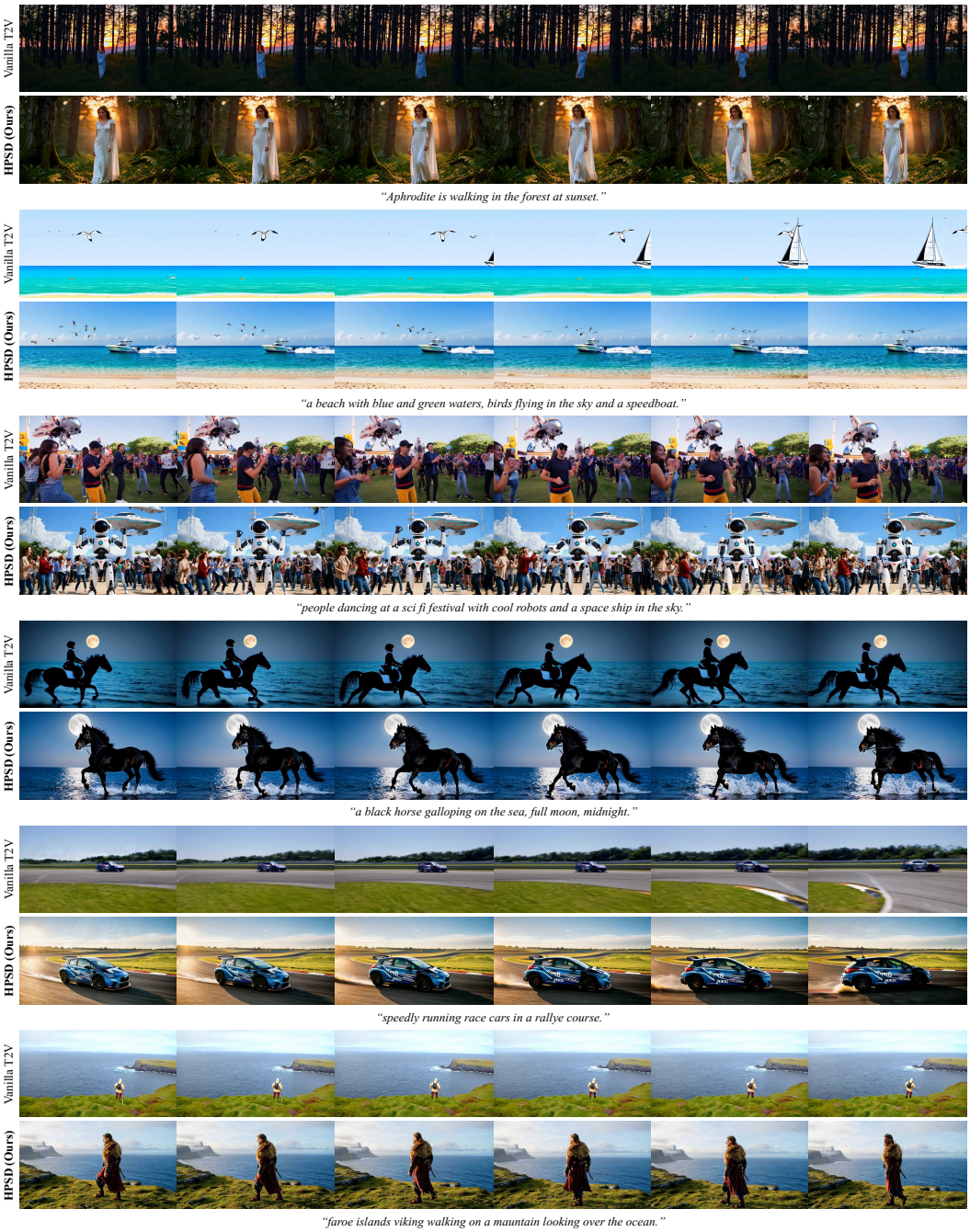


Figure 22: **Comparison between HPSD and vanilla WAN-2.2 (3/3).** Best viewed zoomed in.



Figure 23: **More qualitative comparison results on LTX-2.3 (1/6).** Best viewed zoomed in.

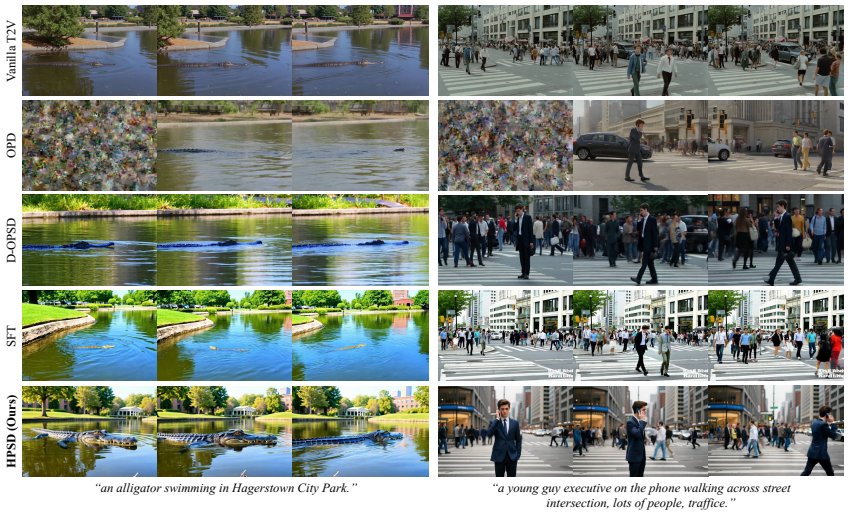


Figure 24: **More qualitative comparison results on LTX-2.3 (2/6).** Best viewed zoomed in.

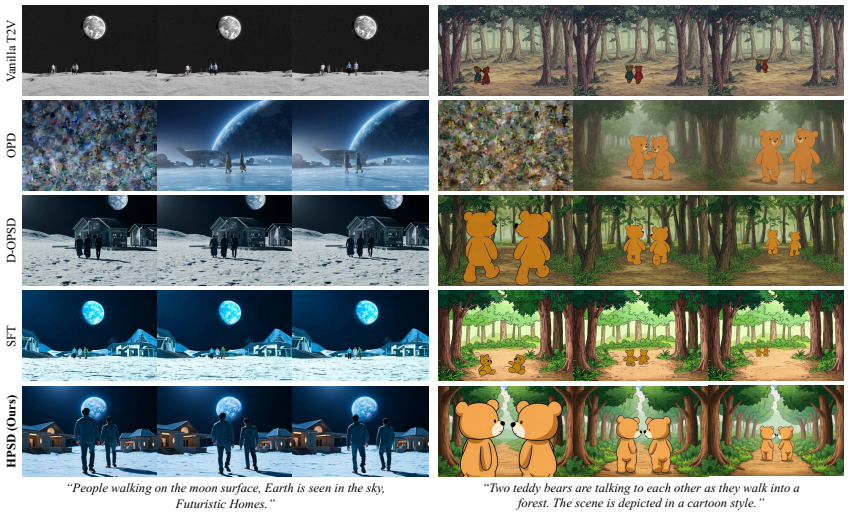


Figure 25: **More qualitative comparison results on LTX-2.3 (3/6).** Best viewed zoomed in.

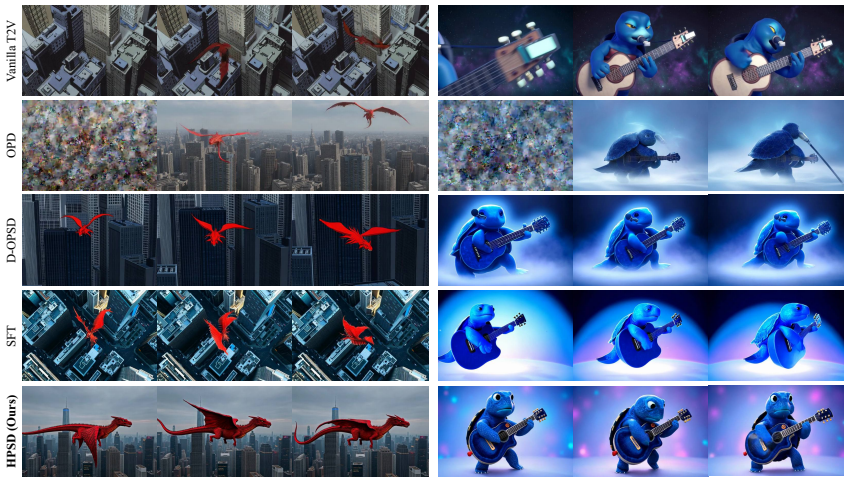


Figure 26: **More qualitative comparison results on LTX-2.3 (4/6).** Best viewed zoomed in.

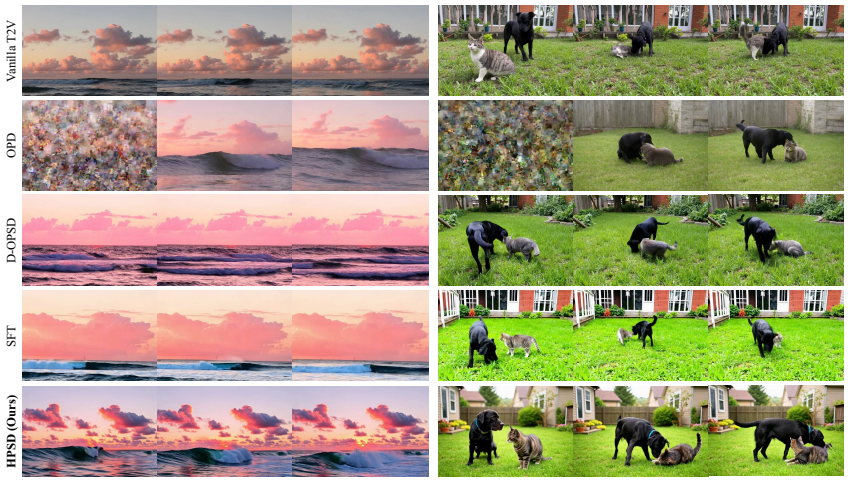


Figure 27: **More qualitative comparison results on LTX-2.3 (5/6).** Best viewed zoomed in.



Figure 28: **More qualitative comparison results on LTX-2.3 (6/6).** Best viewed zoomed in.

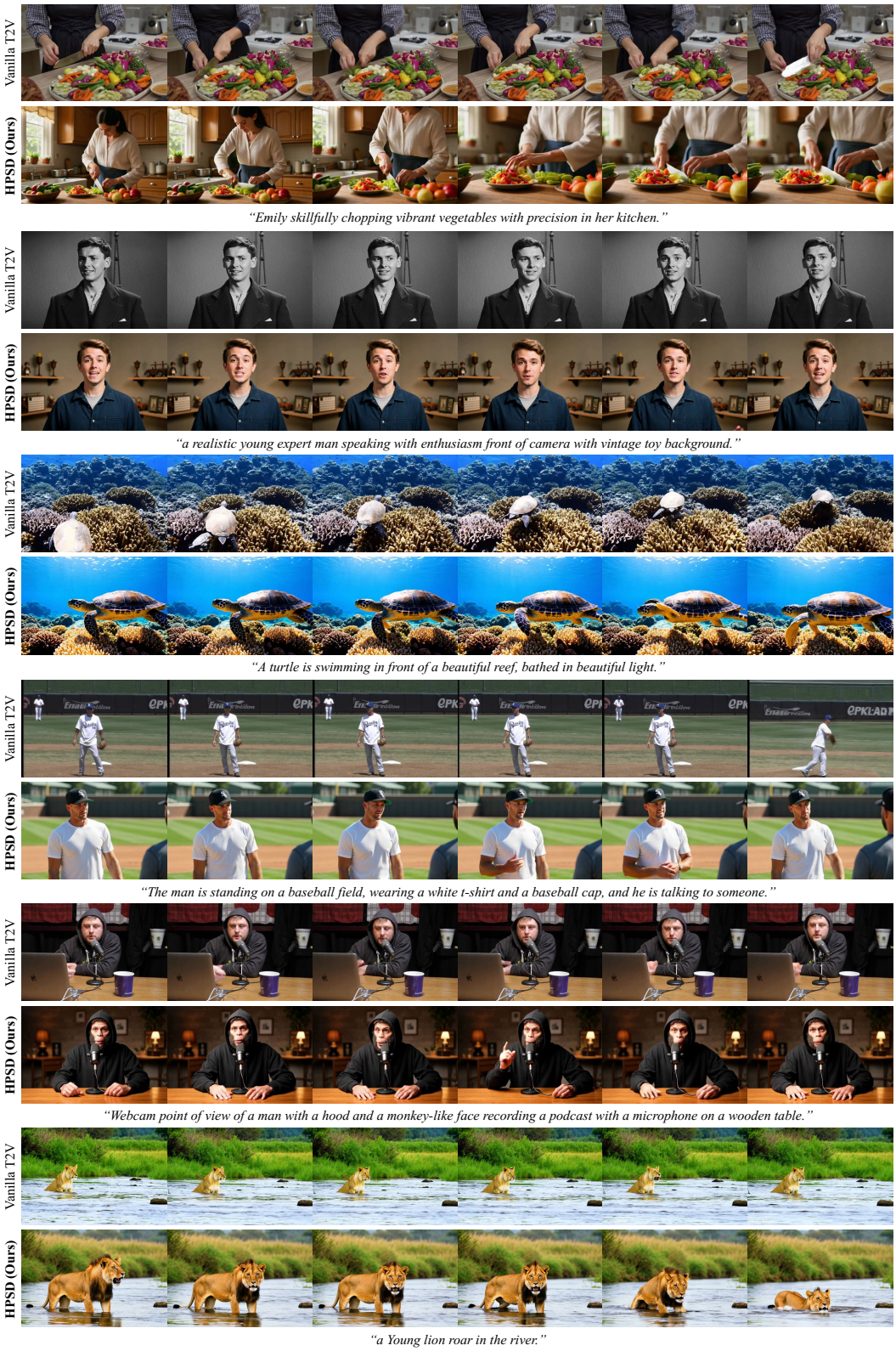


Figure 29: **Comparison between HPSD and vanilla LTX-2.3 (1/2).** Best viewed zoomed in.

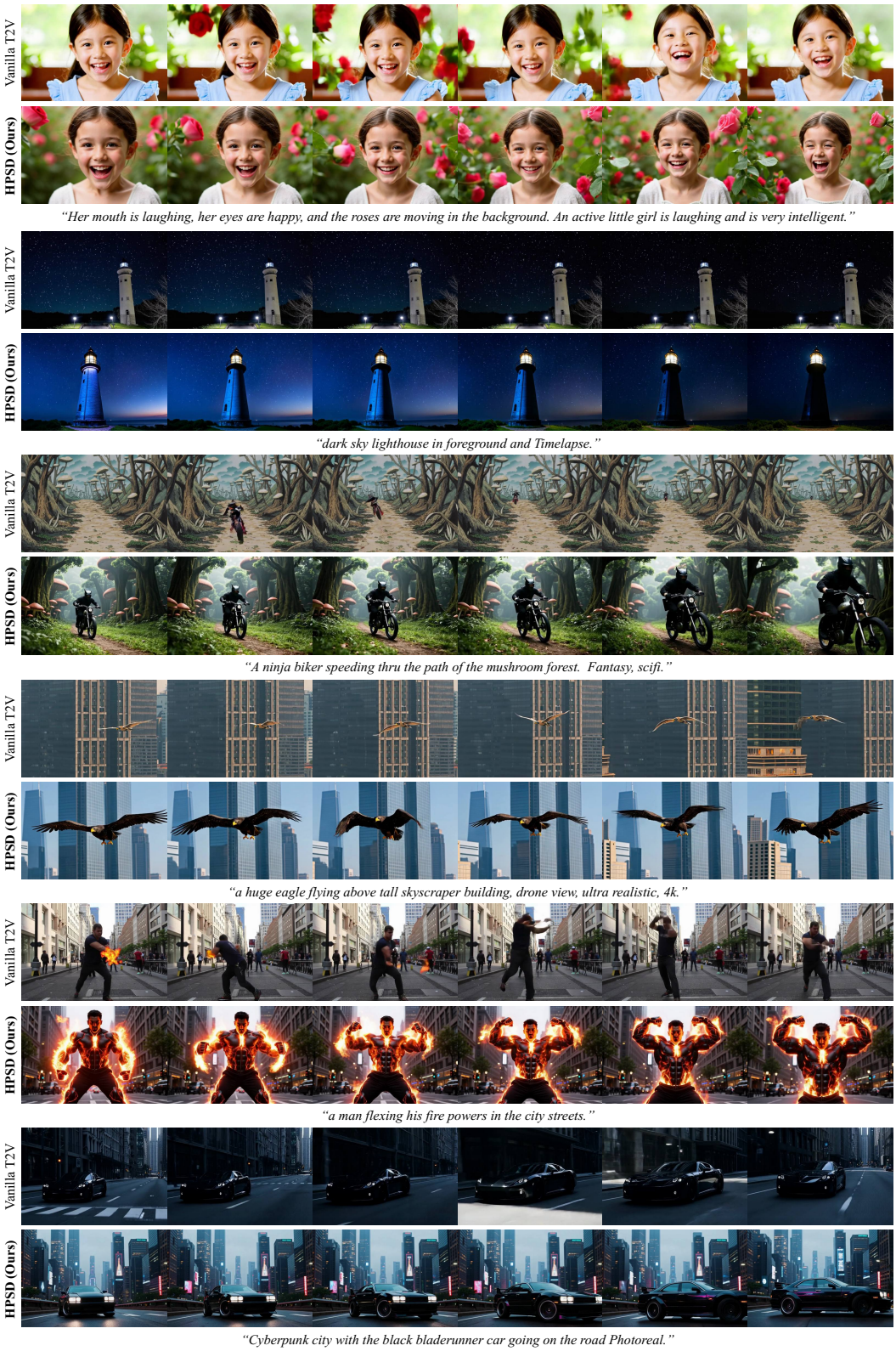
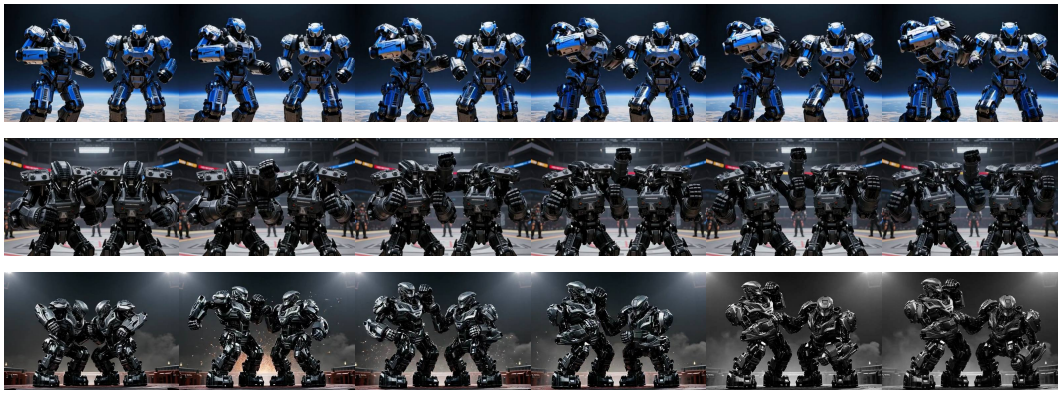


Figure 30: **Comparison between HPSD and vanilla LTX-2.3 (2/2).** Best viewed zoomed in.



"A woman is painting a chair with pink paint using a brush."



"Two advanced fighting robots in the midst of a fierce battle."



"Luna cautiously peering into a dense forest."

Figure 31: **Generated results using the same prompts and different seeds on LTX-2.3.**

Table 6: The video generation prompts for each figure are listed sequentially, following the order from left to right and top to bottom. (Table 1/2)

Figure	Text Prompt
Figure. 1	3D cartoon, a cowboy with an ak47 aiming at you. cars doing a super race in New York. photo of coastline, rocks, distant lighthouse in the background, storm weather, strong wind, crashing waves, huge water splashes, lightning, 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3 The player fights a terrifying skeleton in Minecraft, create this video in a retro video game style. A man rides a bike, escaping a lava tsunami, while pigeons fly away in the background and the camera circles him.
Figure. 2	purple car in a videogame.
Figure. 3	A cartoon depicting a penguin with a book balanced on its head, attempting to explain various concepts to a group of rocks.
Figure. 5	cinematic scene of a cute realistic poodle dog in the beach. A man sitting in cake shop, eating a cake.
Figure. 6	a man playing guitar and sing a song on the street. a video of a giant blue butterfly of 200 meters in a city, flying ver beutifully.
Figure. 7	A 3D illustration of a man engaging in an MMA fight, rendered in Pixar style. two dogs wearing goggles and parachute on an airplane.
Figure. 11	man in yellow hazmat suit on a paddleboard in the middle of the ocean. coffee swirling in a rainbow cup.
Figure. 12	A majestic condor flying through the Peruvian sky. Rock singer live in concert. Four kittens jumping on the sofa.
Figure. 14	barbie walking outdoors on a street, perfect proportions. A bird is walking on a wooden floor and a person is holding a waffle.
Figure. 15	An old wizard with a gray beard, wearing a brown cassock and carrying a staff. Man in a business suit with headphones walking in the rain at night.
Figure. 16	a ship sailing in the ocean. A gray Power Ranger, adorned with a purple cape, is running at a high speed against a green background.
Figure. 17	A polar bear operating an iPhone. Humans walking around a Dragon Zoo 1920s New York, rare film footage.
Figure. 18	steampunk themed chefs cooking beside the sea, high detail rpg. A boy is picking up a coin from the road in a 3D cartoon.
Figure. 19	A car with wings is flying through a city. A man rides a bike, escaping a lava tsunami, while pigeons fly away in the background and the camera circles him.
Figure. 20	1950s film, woman doing yoga, standing in warrior 1 pose and move, standing. boy going school in red color cycle, with a dog. Darth Vader drinks whiskey. A robot is walking in the foggy city at sunset. A cute puppy running on the grass, sunshine, rainbow, reality. In the 1970s there was an anime about a cute alien girl who shot bubbles from a ray gun at a monster robot.

Table 7: The video generation prompts for each figure are listed sequentially, following the order from left to right and top to bottom. (Table 2/2)

Figure	Text Prompt
Figure. 21	A red-haired man in a suit and hat is standing next to a parked car, jumping up and down with excitement.
	distant view, a line of geese fly across the night moon sky, below the sky is the wild pond.
	Teddy Bear casting as doctor, holding first aid kit bag running, on the sea island.
	Polaris rzr 1000 flipping down a mountain.
	Children talking in the garden, cartoon drawing.
	man in yellow hazmat suit on a paddleboard in the middle of the ocean.
Figure. 22	Aphrodite is walking in the forest at sunset.
	a beach with blue and green waters, birds flying in the sky and a speedboat.
	people dancing at a sci fi festival with cool robots and a space ship in the sky.
	a black horse galloping on the sea, full moon, midnight.
	speedly running race cars in a rallye course.
	faroe islands viking walking on a mauntain looking over the ocean.
Figure. 23	Madhurima wearing blue leather jacket staring at camera and Jeans, closing eyes slowly.
	a green parrot playing a concertina.
Figure. 24	an alligator swimming in Hagerstown City Park.
	a young guy executive on the phone walking across street intersection, lots of people, traffice.
Figure. 25	People walking on the moon surface, Earth is seen in the sky, Futuristic Homes.
	Two teddy bears are talking to each other as they walk into a forest. The scene is depicted in a cartoon style.
Figure. 26	A red dragon flies in the city.
	A deep blue baby turtle, playing the guitar and singing into the microphone, highly detailed, 3 d, dreamy, raytracing.
Figure. 27	Animation of drifting waves, with moving pink clouds at sunset.
	A black Labrador and a gray Norwegian Forest cat play on the lawn in the backyard.
Figure. 28	A 4K, 3D animated squirrel named Sammy is shown hopping around in the forest, looking at flowers.
	cool looking pilot flying a red plane under a suspension bridge. The pilot is waving and wearing a blue scarf around his neck.
Figure. 29	Emily skillfully chopping vibrant vegetables with precision in her kitchen.
	a realistic young expert man speaking with enthusiasm front of camera with vintage toy background.
	A turtle is swimming in front of a beautiful reef, bathed in beautiful light.
	The man is standing on a baseball field, wearing a white t-shirt and a baseball cap, and he is talking to someone.
	Webcam point of view of a man with a hood and a monkey-like face recording a podcast with a microphone on a wooden table.
	a Young lion roar in the river.
Figure. 30	Her mouth is laughing, her eyes are happy, and the roses are moving in the background. An active little girl is laughing and is very intelligent.
	dark sky lighthouse in foreground and Timelapse.
	A ninja biker speeding thru the path of the mushroom forest. Fantasy, scifi.
	a huge eagle flying above tall skyscraper building, drone view, ultra realistic, 4k.
	a man flexing his fire powers in the city streets.
	Cyberpunk city with the black bladerunner car going on the road Photoreal.
Figure. 31	A woman is painting a chair with pink paint using a brush.
	Two advanced fighting robots in the midst of a fierce battle.
	Luna cautiously peering into a dense forest.