FROM STATIC TO DYNAMIC: MOTION-REFINED ON-POLICY DISTILLATION FROM IMAGE TO VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) specializes pretrained video diffusion models through teacher supervision along the student's own generation trajectory. Although large video models are natural teachers, developing specialized video experts can require costly video data and training, while querying them incurs substantially higher latency than querying image experts. More readily available and cheaper to query, image experts offer a cost-effective alternative, particularly for largely temporal-agnostic capabilities such as aesthetics and OCR that admit frame-level supervision. However, heterogeneous image and video latent spaces prevent direct supervision of intermediate student states, while image experts lack cross-frame motion supervision, making temporal consistency vulnerable to frame-level improvements. In this paper, we propose **MILD**, a **M**otion-Preserving **I**mage-to-Video **L**atent **D**istillation framework that transfers specialized image expertise while preserving pretrained video dynamics. MILD uses a learnable linear connector that aligns student latent states and predicted updates with those of image experts, enabling supervision transfer across heterogeneous latent spaces. We further constrain image-guided corrections around the pretrained student's predictions to preserve video dynamics and incorporate an optical-flow-based motion reward to improve motion quality and temporal consistency. Across specialized image experts and multiple video-student backbones, our method consistently outperforms video-teacher OPD baselines, with further studies demonstrating effective transfer across connector designs and heterogeneous architectures. These results establish image-to-video distillation as an effective route to improving video generation by drawing on the diverse and evolving capabilities of the image-generation ecosystem.

1 INTRODUCTION

Video diffusion models have achieved substantial progress in visual fidelity and temporal coherence (Ho et al., 2022; Blattmann et al., 2023; Wan et al., 2025). As these models become stronger, an increasingly important practical setting is to adapt a pretrained video model toward specific downstream capabilities, such as improved aesthetics and text rendering, without retraining the model from scratch. On-policy distillation (OPD) is an effective strategy for this setting which queries the teacher on states sampled from the student's own generation process, thereby steering the student's generation distribution toward that of the stronger teacher during post-training (Li et al., 2026; Fang et al., 2026). In the video domain, stronger video diffusion teachers have been shown to improve weaker video models (Yin et al., 2025; Huang et al.,

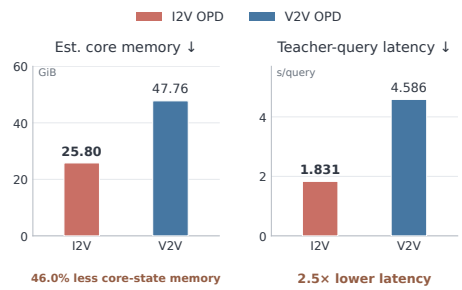

Figure 1: **Computational costs of I2V and V2V OPD.** Both train Wan2.1-1.3B using SD3.5-Medium and Wan2.1-14B teachers, respectively. Left: estimated student training-state and teacher memory, excluding activations and auxiliary modules. Right: measured teacher-query latency.

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.



Figure 2: **Qualitative comparisons of Base, V2V OPD, and I2V OPD.** Base denotes the pretrained Wan2.1-1.3B model. Top: aesthetic quality on DrawBench, text rendering on the Text Rendering benchmark, and compositional generation. Bottom: complex scene generation, count and action accuracy, and camera motion control on VBench-2.0 prompts.

2026), for example by accelerating student inference through distillation. However, OPD training requires repeated evaluations of an expensive video teacher at student-generated states, introducing substantial computational overhead. As shown in Fig. 1, video-teacher OPD has an estimated core model-state memory footprint of 47.76 GiB and a mean teacher-query latency of 4.59 seconds under the evaluated configurations. In contrast, image-teacher OPD reduces these costs to 25.80 GiB and 1.83 seconds, respectively, while image generation models can provide supervision for downstream video tasks such as visual aesthetics and text rendering (Guo et al., 2023; Kwon et al., 2024; Chen et al., 2024; Zhu et al., 2024). Qualitative comparisons in Fig. 2 further illustrate the potential of image-teacher OPD to improve visual quality and prompt adherence over the pretrained student and video-teacher OPD.

This motivates us to explore whether image generation models post-trained for specialized tasks can serve as attractive teachers for video generation, improving the student video model while maintaining training efficiency. However, applying image models to video generation faces two challenges. First, image and video models typically use different VAEs, resulting in mismatched latent spaces with different channel dimensions and spatial or temporal compression rates. Thus, the student's video latents during training cannot be directly processed by the image teacher, nor can the teacher's predictions serve as targets in the student's latent space. Second, image experts are trained on static images and do not capture temporal dynamics such as cross-frame object and scene evolution, so their frame-level supervision may disrupt temporal coherence. These challenges raise a central question: *How can we transfer specialized image expertise to a video student through on-policy distillation across heterogeneous latent spaces while preserving its learned temporal dynamics?*

To address this question, we propose **MILD**, a **M**otion-Preserving **I**mage-to-Video **L**atent **D**istillation framework that transfers specialized visual capabilities from image experts while preserving pretrained video dynamics. In particular, we first develop an adaptive connector to bridge image and video latent spaces and maintain their alignment as the student evolves throughout training. To preserve video dynamics, we anchor image-guided corrections to the student's frozen pretrained predictions and constrain their magnitude and inter-frame variation. An optical-flow-based motion reward further complements static image expertise with explicit motion feedback, improving motion quality and temporal consistency. Our main contributions are summarized as follows:

- We introduce an image-to-video OPD framework, MILD, that transfers specialized visual capabilities from image experts to video students, reducing reliance on costly specialized video teachers. Its adaptive connector supports both linear and nonlinear designs and enables transfer across different VAEs and backbone architectures.
- We combine image-expert supervision with a frozen video reference and explicit temporal constraints to improve frame quality while preserving pretrained video dynamics. An auxiliary motion reward complements image supervision with direct feedback on video motion.
- Experiments on Wan2.1-1.3B, Wan2.2-5B, and LTX-Video-2B demonstrate effective capability transfer at lower teacher-query costs on VBench-2.0 and EvalCrafter, while connector ablations and cross-architecture distillation support the framework's generality.

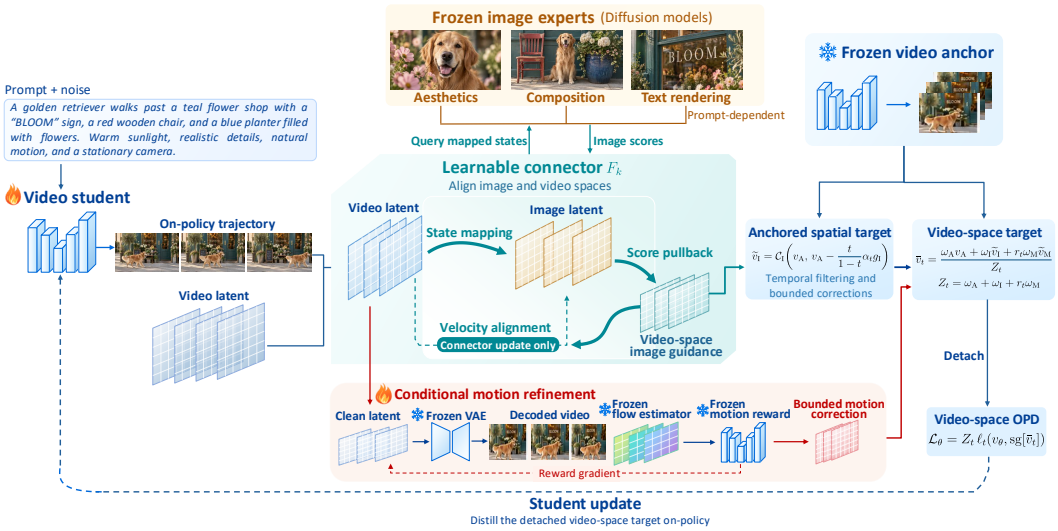

Figure 3: **Overview of MILD.** Learnable connectors transfer image-expert supervision to on-policy video states. A frozen video anchor bounds spatial corrections, while optical-flow feedback provides motion refinement. Predictions are fused into a detached distillation target; connectors use a separate alignment objective.

2 Related Work

Image priors for video generation. Prior work extends image models with temporal modules (Singer et al., 2022; Blattmann et al., 2023; Guo et al., 2023), fine-tunes spatial modules with high-quality images (Chen et al., 2024), or uses inference-time denoiser fusion (Shao et al., 2025) and image-supervised distillation (Zhai et al., 2024). We transfer specialized image predictions across heterogeneous latent spaces, using bounded corrections to a video anchor to compensate for missing temporal context.

On-policy diffusion distillation. OPD supervises students along their own generation trajectories (Lu & Lab, 2025; Yu et al., 2026). DMD2 reduces multi-step training–inference mismatch (Yin et al., 2024); CausVid distills bidirectional video teachers into autoregressive students, while Self Forcing reduces exposure bias through self-generated rollouts (Yin et al., 2025; Huang et al., 2026). DiffusionOPD and Flow-OPD consolidate specialized image teachers along student denoising trajectories (Li et al., 2026; Fang et al., 2026). We transfer image expertise to video by pulling image scores into video coordinates and integrating video denoising and motion refinement.

Motion-aware video post-training. Prior work improves motion through temporal adaptation, video guidance, or motion-specific objectives (Zhao et al., 2024; Lee et al., 2025; Li et al., 2025; Xue et al., 2025). InstructVideo uses image rewards for video fine-tuning, while VADER backpropagates reward gradients through video generation (Yuan et al., 2024; Prabhudesai et al.). Our motion feedback corrects image-induced denoising directions to construct temporally compatible targets from frame-level supervision.

3 Method

Figure 3 outlines MILD from frozen image teachers to a video student. We begin by defining the distillation objective (Section 3.1) and introducing learnable connectors to align heterogeneous representations (Section 3.2). Building on this alignment, we constrain and refine image supervision using a video anchor and motion feedback (Section 3.3). Finally, we describe target fusion and connector alignment (Section 3.4).

3.1 On-policy distillation objective

We first formulate an on-policy distillation objective that matches the video student's denoising transitions to target transitions at latent states sampled from its current generation trajectories. Let $v_\theta(x_t, t, c)$ denote the student's velocity at video latent x_t, diffusion time t, and text condition c, and let $\bar{v}(x_t, t, c)$ denote a target velocity in the same video latent space. For a reverse step from t to

$t-h$, let $\pi_u(\cdot \mid x_t, t, c)$ denote the transition induced by velocity u. The OPD objective is

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_{c,t,\, x_t \sim p_\theta(\cdot|t,c)} \left[\ell_t(v_\theta, \bar{v}) \right], \quad \ell_t(v_\theta, \bar{v}) := \frac{1}{d} D_{\text{KL}}(\pi_{v_\theta}(\cdot \mid x_t, t, c) \,\|\, \pi_{\bar{v}}(\cdot \mid x_t, t, c)), \quad (1)$$

where $p_\theta(\cdot \mid t, c)$ is the distribution of states visited by the current student at time t, and d is the number of video-latent entries. Both transitions in the KL divergence are conditioned on the same (x_t, t, c). We next obtain a closed form for this loss under the reverse-SDE Euler formulation of OPD (Li et al., 2026). We use the rectified-flow convention $x_t = (1-t)x_0 + t\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ is independent of x_0 (Liu et al., 2023). For $0 < t < 1$ and $0 < h \leq t$, the transition associated with $u \in \{v_\theta, \bar{v}\}$ is $\pi_u = \mathcal{N}(\mu_u, \sigma_t^2 h I)$, where μ_u is its reverse-SDE Euler mean and $\sigma_t > 0$ is the shared diffusion coefficient. Under this convention, the difference between the transition means is

$$\mu_{v_\theta} - \mu_{\bar{v}} = -h \left(1 + \frac{\sigma_t^2 (1-t)}{2t} \right) (v_\theta - \bar{v}).$$

Since the transitions share the same covariance, their KL divergence depends only on this mean difference. Substitution gives

$$\ell_t(v_\theta, \bar{v}) = \frac{\|\mu_{v_\theta} - \mu_{\bar{v}}\|_{\text{av}}^2}{2\sigma_t^2 h} = w(t,h) \|v_\theta - \bar{v}\|_{\text{av}}^2, \quad (2)$$

$$w(t,h) = \frac{h}{2\sigma_t^2} \left(1 + \frac{\sigma_t^2 (1-t)}{2t} \right)^2, \quad (3)$$

where $\|q\|_{\text{av}}^2 = \|q\|_2^2 / d$. Thus, matching these Gaussian denoising transitions is equivalent to minimizing a weighted velocity error in video coordinates. The complete reverse-SDE derivation is provided in Appendix A.1.

3.2 Cross-architecture image guidance transfer

Video students and image experts use different VAEs and latent layouts, so their latent states and denoising velocities cannot be compared directly. We introduce a differentiable connector $F_\psi : \mathbb{R}^{d^{\text{V}}} \to \mathbb{R}^{d^{\text{I}}}$ to evaluate student states using image experts, where d^{V} and d^{I} denote the respective latent dimensions.

Transferring image guidance. We select video-latent slices at fixed temporal intervals, resize them to the experts' spatial layout, and map them into the image latent space with a learnable transformation. The K frozen experts share a VAE and input layout, allowing a common connector. To compare denoising behavior, we also map the student's velocity into image coordinates. Holding ψ fixed, the chain rule gives

$$z_t = F_\psi(x_t), \qquad \widehat{v}_\theta^{\text{I}} = J_{F_\psi}(x_t) v_\theta(x_t, t, c), \quad (4)$$

where J_{F_ψ} is the Jacobian with respect to x_t. The mapped velocity $\widehat{v}_\theta^{\text{I}}$ can therefore be compared with the expert predictions $u_k(z_t, t, c)$ at the mapped student state. Expert predictions are expressed in image coordinates and cannot directly serve as video-space updates. Since the connector need not be invertible, we convert the expert velocities into image scores s_k^{I} and pull their guidance back to the video latent space through the chain rule:

$$g_{\text{I}} = J_{F_\psi}(x_t)^\top \sum_{k=1}^{K} \beta_k(c) s_k^{\text{I}}, \qquad \beta_k(c) \geq 0, \quad \sum_k \beta_k(c) = 1. \quad (5)$$

Here, g_{I} gives a video-space direction toward greater agreement with the image experts. The fixed prompt-dependent weights $\beta_k(c)$ select relevant expertise; for example, OCR guidance is disabled when text rendering is not required. Appendix A.2 details score conversion and pullback.

Aligning the connector. Matching latent dimensions alone does not ensure that the mapped video states preserve the visual content or denoising behavior expected by the image experts. We therefore train the connector using content and velocity alignment. For content alignment, we encode the same image with both VAEs: the video VAE receives a static clip of repeated images, while the image VAE's output is replicated across the selected temporal slices. This yields paired clean latents

$(\bar{x}_0, \bar{z}_0)$. For velocity alignment, we compare the mapped student velocity with expert predictions at states sampled from the student's current trajectories:

$$\mathcal{L}_{\text{conn}} = \lambda_z \mathbb{E}\|F_\psi(\bar{x}_0) - \bar{z}_0\|^2_{\text{av}} + \frac{\lambda_v}{K}\sum_{k=1}^{K} \mathbb{E}\|\widehat{v}^{\text{I}}_\theta - u_k(z_t, t, c)\|^2_{\text{av}} + \lambda_{\text{reg}}\mathcal{R}(\psi). \tag{6}$$

The content term grounds the mapping in shared visual content, while the velocity term adapts it to the student's evolving denoising trajectories. The regularizer $\mathcal{R}(\psi)$ penalizes deviations from the initial connector parameters. During connector updates, input latents, student velocities, and expert predictions are detached, so only ψ is optimized. Although z_t depends on F_ψ, expert predictions are fixed targets; the velocity term updates the connector through $J_{F_\psi}(x_t)v_\theta$.

Our default connector is a bias-free linear map $F_\psi(x) = A_\psi x$, implemented using fixed slice selection and resizing followed by learnable spatial convolutions. Its Jacobian satisfies $J_{F_\psi}(x)v = A_\psi v$, so the same operator maps both states and velocities. Nonlinear connectors use Jacobian–vector products under the same alignment objective. Neither implementation requires an invertible mapping, a shared image–video VAE, or matching backbone architectures. Appendix A.3 provides the derivation and nonlinear training details.

3.3 Preserving video dynamics and refining motion

Transferred image guidance can improve spatial content, but it does not account for inter-frame relationships. We therefore constrain its effect using a frozen video reference and optionally add explicit motion feedback.

Preserving video dynamics. We use the velocity v_{A} predicted by a frozen video reference as an anchor and apply the transferred image guidance as a controlled correction:

$$\widetilde{v}_{\text{I}} = \mathcal{C}_{\text{I}}\left(v_{\text{A}}, v_{\text{A}} - \frac{t}{1-t}\alpha_t g_{\text{I}}\right), \qquad 0 < t < 1. \tag{7}$$

The factor $t/(1-t)$ converts a score correction into a velocity correction under the rectified-flow convention. The scale $\alpha_t \geq 0$ normalizes guidance relative to the anchor score and caps the initial correction. The operator $\mathcal{C}_{\text{I}}$ further limits its magnitude relative to the anchor and suppresses abrupt inter-frame variation. This produces an image-enhanced target while constraining departures from pretrained video behavior. Appendix A.4.1 details the guidance normalization and the temporal correction operator.

Refining motion with explicit feedback. Constraining image guidance limits temporal disruption but does not directly reward realistic motion. We hence optionally refine the image-enhanced target using a learned motion reward R. A frozen RAFT estimator (Teed & Deng, 2020) extracts optical flow from the decoded video, and a motion discriminator scores it. The discriminator is trained to distinguish real-video flows from generated and temporally corrupted flows. Specifically, we decode the clean latent prediction $\widehat{x}_0 = x_t - t\,\text{sg}[\widetilde{v}_{\text{I}}]$ and use the reward gradient to refine the target:

$$\widetilde{v}_{\text{M}} = \mathcal{C}_{\text{M}}\big(\widetilde{v}_{\text{I}}, \widetilde{v}_{\text{I}} - \gamma_t \nabla_{x_t} R\big(D^{\text{V}}(\widehat{x}_0)\big)\big)\,. \tag{8}$$

Here, D^{V} is the frozen video decoder, sg stops gradients through the image target, and $\gamma_t \geq 0$ controls the reward correction. Before temporal processing, subtracting this gradient from the velocity moves the clean prediction locally toward higher motion reward. The operator $\mathcal{C}_{\text{M}}$ controls the correction magnitude and its inter-frame variation. The decoder, RAFT, and discriminator remain frozen during refinement, while allowing gradients through their inputs. Appendix B describes discriminator training, and Appendix A.4.2 details reward differentiation and correction scaling.

3.4 Joint optimization

At each student-update state, we combine the frozen video anchor v_{A}, image-enhanced target $\widetilde{v}_{\text{I}}$, and, when applied, motion-refined target $\widetilde{v}_{\text{M}}$. Motion refinement requires differentiation through video decoding and optical-flow estimation, so we apply it only at selected states; anchor and image supervision remain active at every state. Let $r_t \in \{0, 1\}$ indicate whether motion refinement is applied at x_t. For positive weights ω_{A}, ω_{I}, and ω_{M}, define $Z_t = \omega_{\text{A}} + \omega_{\text{I}} + r_t\omega_{\text{M}}$. The fused video-space target is $\bar{v}_t = Z_t^{-1}(\omega_{\text{A}} v_{\text{A}} + \omega_{\text{I}}\widetilde{v}_{\text{I}} + r_t\omega_{\text{M}}\widetilde{v}_{\text{M}})$. Appendix B specifies the diffusion-time range eligible for motion refinement and the per-iteration activation budget.

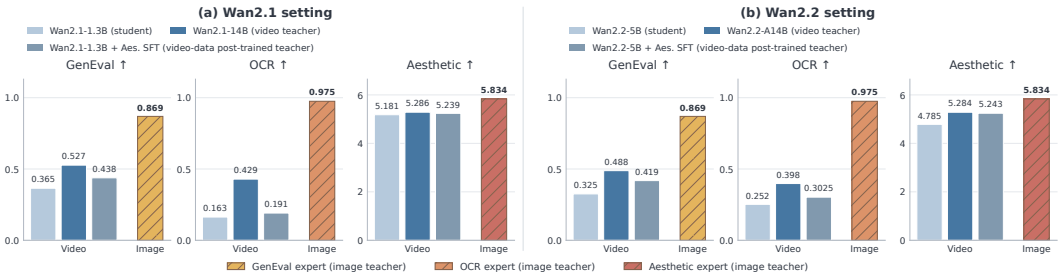

Figure 4: **Spatial capability comparisons across Wan2.1 and Wan2.2 settings.** Each setting compares the student, a larger video teacher, a video teacher post-trained on aesthetic video data, and specialized image teachers. Image experts achieve the highest GenEval, OCR, and Aesthetic scores in both settings.

For N states sampled from the current student trajectories, we optimize the student with

$$\mathcal{L}_{\text{student}} = \frac{1}{N}\sum_{j=1}^{N} Z_{t_j}\,\ell_{t_j}\left(v_\theta(x_{t_j}, t_j, c), \text{sg}[\bar{v}_{t_j}]\right). \tag{9}$$

Because ℓ_t is a weighted squared velocity error, the factor Z_t restores the component weights after target normalization: with the component targets detached, the student gradient equals that from matching each active target separately with its corresponding fusion weight.

Student updates change the on-policy states and velocities used to align the connector. After a connector-only warm-up, each iteration updates the student using Eq. (9) and the connector using Eq. (6). The fused target is detached during student updates; input states, student velocities, and image-expert predictions are detached during connector updates. Thus, each objective updates only its intended parameters. The video anchor, image experts, and auxiliary motion networks remain frozen, and inference uses only the updated student. Appendix A.5 details the two update rules.

4 Experiments

4.1 Experimental Setting

Setup. We evaluate MILD on Wan2.1-1.3B, Wan2.2-5B (Wan et al., 2025), and LTX-Video-2B (HaCohen et al., 2024), using a frozen copy of each pretrained student as its video anchor. Following DiffusionOPD (Li et al., 2026), we use three SD3.5-Medium (Esser et al., 2024) image experts specializing in compositional generation (Ghosh et al., 2023), aesthetics, and text rendering. Table 1 lists the video-teacher pairings and the image-expert backbone shared across students. Ordered as GenEval, aesthetic, and OCR experts, their image-score weights are $(0.5, 0.5, 0)$ for ordinary prompts and $(0.4, 0.4, 0.2)$ for text-related prompts. We construct a pool of 10,000 training prompts, with 5,000 from OpenVid-1M (Nan et al., 2025) and 5,000 from LVD-2M (Xiong et al., 2024). Each iteration samples one prompt and generates an on-policy trajectory for an 81-frame video at 480×832 resolution. Training comprises 50 connector-alignment iterations followed by 200 joint optimization iterations. Further implementation details and motion-reward validation are provided in Appendices B and C.1.

Table 1: **Experimental setup.** Video-teacher pairings and shared image-expert backbone. Each V2V configuration uses one video teacher.

Video student	V2V teacher	Image experts
Wan2.1-1.3B	Wan2.1-14B	SD3.5-Medium
	Wan2.2-5B	
	Wan2.2-A14B	
Wan2.2-5B	Wan2.2-A14B	
LTX-Video-2B	LTX-Video-13B	

Baselines. We compare MILD with the pretrained students and two types of V2V OPD baselines: **V2V OPD (teacher name)** uses a larger pretrained video teacher, while **V2V OPD (Aes. SFT teacher)** uses an aesthetic-SFT video teacher. To construct the latter, we fine-tune a copy of each student backbone on video–text data from AesVideo-Bench (Han et al., 2026), then freeze the resulting model as its V2V OPD teacher. Table 1 lists the video-teacher pairings, including Wan2.2-5B and Wan2.2-A14B teachers for Wan2.1-1.3B to assess transfer across video model variants. We evaluate video quality using VBench-2.0 (Zheng et al., 2025) and EvalCrafter (Liu et al., 2024).

4.2 Main Results

Table 2: **Evaluation results on VBench-2.0 and EvalCrafter.** MILD outperforms both V2V OPD baselines on the aggregate scores of both benchmarks across all three student backbones. Compared with the larger-teacher V2V OPD baseline, MILD reduces the sum of per-GPU peak allocated memory by 39.32%, 64.07%, and 19.53% on Wan2.1-1.3B, Wan2.2-5B, and LTX-Video-2B, respectively (Figure 5). Bold and underlined values indicate the best and second-best results within each student group.

	VBench-2.0 ↑						EvalCrafter ↑				
Method	**Creativity**	**Commonsense**	**Controllability**	**Human Fidelity**	**Physics**	**AVG.**	**Visual Quality**	**Text–Video Alignment**	**Motion Quality**	**Temporal Consistency**	**Final Sum Score**
Large video teachers (reference)											
Wan2.1-14B	48.91	59.95	27.44	90.94	44.09	54.27	66.25	56.78	53.97	63.93	240.93
Wan2.2-A14B	52.98	67.73	35.36	82.59	51.60	58.05	66.47	60.22	54.58	63.10	244.37
LTX-Video-13B	55.02	40.84	13.90	87.60	43.14	48.10	64.08	51.99	56.00	65.04	237.11
Wan2.1-1.3B											
Base	45.86	59.68	22.85	85.16	42.07	51.12	65.78	56.51	53.36	63.22	238.87
V2V OPD (Wan2.1-14B)	47.34	63.71	24.67	85.17	45.32	53.24	66.64	56.38	53.70	63.03	239.75
V2V OPD (Aes. SFT teacher)	**48.12**	61.13	24.94	**86.45**	44.94	53.12	65.47	**58.62**	54.45	63.16	241.70
MILD	47.89	**64.56**	**25.17**	85.47	**46.52**	**53.92**	**66.84**	57.89	**54.58**	**63.79**	**243.10**
Wan2.2-5B											
Base	48.20	58.50	20.02	81.89	49.31	51.58	61.48	55.63	54.00	61.93	233.04
V2V OPD (Wan2.2-A14B)	48.29	60.81	20.79	81.98	51.60	52.69	62.06	57.51	54.07	**63.10**	236.74
V2V OPD (Aes. SFT teacher)	48.64	61.69	19.07	83.75	50.62	52.75	62.23	57.79	54.11	62.73	236.86
MILD	**49.50**	**62.84**	**21.80**	**84.89**	**54.29**	**54.66**	**63.76**	**59.12**	**54.43**	62.79	**240.10**
LTX-Video-2B											
Base	47.95	43.72	12.17	67.35	37.77	41.79	56.31	46.78	54.96	62.76	220.81
V2V OPD (LTX-13B)	56.68	33.38	16.22	68.82	45.43	44.11	56.50	47.05	55.19	64.39	223.13
V2V OPD (Aes. SFT teacher)	56.75	45.36	15.97	69.15	47.14	46.87	**56.62**	47.61	55.21	64.21	223.65
MILD	**57.23**	**48.85**	**16.78**	**72.38**	**50.06**	**49.06**	56.58	**48.41**	**55.30**	**64.51**	**224.80**

Table 3: **Cross-architecture video-to-video OPD on VBench-2.0.** All OPD variants use Wan2.1-1.3B as the student, with the teacher specified in parentheses. Distillation from Wan2.2 teachers improves the student's overall performance, showing that the proposed connector transfers supervision between heterogeneous video diffusion models.

Method	**Creativity**	**Commonsense**	**Controllability**	**Human Fidelity**	**Physics**	**AVG.**
Wan2.1-1.3B	45.86	59.68	22.85	85.16	42.07	51.12
V2V OPD (Wan2.1-14B)	47.34	63.71	24.67	85.17	45.32	53.24
V2V OPD (Wan2.2-A14B)	48.74	65.72	24.59	82.37	42.91	52.87
V2V OPD (Wan2.2-5B)	48.29	58.50	23.70	82.51	43.14	51.23

Quantitative Analysis. Table 2 compares MILD with the pretrained students and V2V OPD baselines across three student backbones on VBench-2.0 and EvalCrafter. MILD achieves stronger aggregate benchmark performance while reducing GPU memory usage, as shown in Figure 5. On Wan2.1-1.3B, Wan2.2-5B, and LTX-Video-2B, MILD achieves VBench-2.0 averages of 53.92, 54.66, and 49.06, exceeding the best V2V OPD results of 53.24, 52.75, and 46.87, respectively. It also surpasses both V2V OPD baselines in EvalCrafter final score across all three backbones. Meanwhile, the sum of per-GPU peak allocated memory is reduced by 39.32%, 64.07%, and 19.53%, respectively, relative to the V2V OPD baseline. These results demonstrate that specialized image-expert supervision can outperform video-to-video distillation while requiring less aggregate GPU memory. Figure 4 compares specialized image experts with students, larger video teachers, and video teachers post-trained on aesthetic video data across the Wan2.1 and Wan2.2 settings. The GenEval and OCR experts score 0.869 and 0.975, respectively, compared with 0.527 and 0.429 for Wan2.1-14B and 0.488 and 0.398 for Wan2.2-A14B. The aesthetic expert scores 5.834, exceeding all evaluated video models, whose highest score is 5.286. These capability advantages support using specialized image experts as sources of spatial supervision for video post-training. Beyond appearance, MILD improves motion quality, temporal consistency, controllability, and physics over each pretrained student, indicating that transferring static image expertise is compatible with improved temporal behavior. Table 3 further demonstrates the connector's applicability to cross-architecture video-to-video OPD. Distilling Wan2.2-A14B or Wan2.2-5B into Wan2.1-1.3B raises its VBench-2.0 average from 51.12 to 52.87 or 51.23, respec-

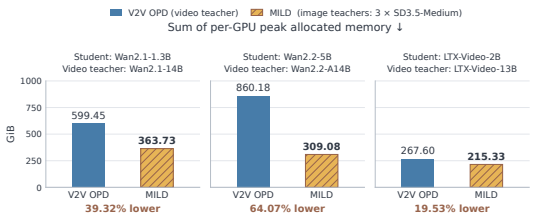

Figure 5: **GPU memory comparison across student backbones.** MILD reduces the sum of per-GPU peak allocated memory relative to V2V OPD across all three backbones. The metric sums each participating GPU's peak allocated memory.

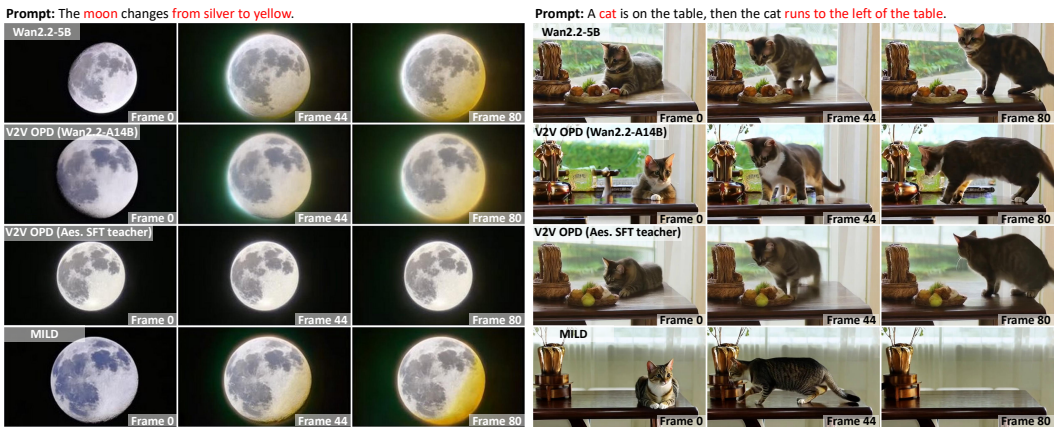

Figure 6: **Qualitative comparisons using Wan2.2-5B as the student.** Frames 0, 44, and 80 illustrate MILD's ability to accurately follow prompts involving color changes and directional motion. In contrast, baselines fail to respond correctly to the moon's color change and produces artifacts (left), while the cat's motion direction is incorrect (right). More visualizations are in Appendix C.3.

Table 4: **Frame-level spatial capability evaluation using Wan2.2-5B.** We evaluate the middle frame of each generated video, using DrawBench prompts for the five model-based metrics and FlowGRPO evaluation splits for GenEval and OCR, consistent with DiffusionOPD (Li et al., 2026). MILD achieves the highest score on every metric among the compared student variants.

	DrawBench					FlowGRPO splits	
Method	**Aesthetic**	**PickScore**	**HPSv2**	**CLIPScore**	**ImageReward**	**GenEval**	**OCR**
Base	4.7855	0.7841	0.2293	0.2372	−0.4390	0.3252	0.2522
V2V OPD (Aes. SFT teacher)	4.9169	0.7931	0.2321	0.2414	−0.3098	0.4023	0.2617
V2V OPD (Wan2.2-A14B)	4.8217	0.7819	0.2253	0.2338	−0.4610	0.4067	0.2556
MILD	**5.0305**	**0.8111**	**0.2470**	**0.2438**	**0.0320**	**0.4454**	**0.2697**

tively. This extension shows that the connector provides a supervision interface for heterogeneous diffusion models beyond the image-to-video setting. Architectural differences, particularly between the mixture-of-experts Wan2.2-A14B teacher and the dense student, may affect transfer effectiveness; understanding and mitigating these effects remains future work. Appendix C.2 provides further comparisons with video models fine-tuned on aesthetic video data, showing that MILD retains an overall advantage and supporting the effectiveness of transferring specialized image expertise to video generation.

Qualitative Analysis. Figure 6 compares frames 0, 44, and 80 from videos generated with Wan2.2-5B as the student. For the moon prompt, MILD shows a clear transition toward yellow while retaining visible lunar texture, whereas V2V OPD with the aesthetic SFT teacher shows little color change. For the cat prompt, MILD depicts the cat moving left across the table and leaving the frame by the final sampled frame, while the compared baselines keep the cat visible on the table. These cases suggest that MILD transfers spatial expertise while retaining the ability to generate prompt-aligned temporal changes. Appendix C.3 provides additional cases using Wan2.2-5B as the student, while Appendix C.4 compares the aesthetic image expert with the video teacher post-trained on aesthetic video data to illustrate differences in visual expertise available for distillation.

4.3 Spatial Capability Transfer

We evaluate spatial capability transfer using the evaluation splits and metrics adopted by DiffusionOPD (Li et al., 2026). For each evaluation prompt, we first generate a video and extract its middle frame for image-level assessment, using the same frame-selection protocol across all methods. On DrawBench (Saharia et al., 2022), we evaluate Aesthetic (Schuhmann, 2022), PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), CLIPScore (Hessel et al., 2021), and ImageReward (Xu et al., 2023) to assess visual aesthetics, predicted human preference, and text–image alignment. For compositional generation and text rendering, we use the GenEval (Ghosh et al., 2023) and OCR evaluation splits provided by FlowGRPO (Liu et al., 2026), applying their respective scoring procedures to the extracted middle frames. As shown in Table 4, MILD leads the compared student variants on all seven frame-level metrics. It achieves a PickScore of 0.8111, an ImageReward score of 0.0320, and an OCR score of 0.2697; its GenEval score also exceeds the strongest V2V OPD result (0.4454 versus 0.4067). These results support the transfer of specialized

Table 5: **Ablation study on Wan2.2-5B using VBench-2.0.** The full model uses auxiliary motion refinement, joint optimization, the default connector, and prompt-dependent image-score weights that enable OCR guidance only for text-related prompts. Single-expert variants use only the indicated image expert.

Variant	Creativity	Commonsense	Controllability	Human Fidelity	Physics	AVG.
Pretrained model (reference)						
Wan2.2-5B (Base)	48.20	58.50	20.02	81.89	49.31	51.58
Supervision and optimization strategy						
w/o image experts	45.97	60.79	20.48	82.11	50.95	52.06
w/o motion refinement	48.64	61.75	19.73	83.45	53.68	53.45
V2V OPD (Wan2.2-A14B) + motion refinement	46.86	61.68	20.40	83.61	52.74	53.06
Alternating optimization	48.42	**62.84**	20.34	83.83	53.93	53.87
MILD (joint optimization)	**49.50**	**62.84**	**21.80**	**84.89**	**54.29**	**54.66**
Connector architecture						
1×1 linear connector	48.20	60.80	19.62	83.75	52.37	52.95
Nonlinear connector	48.74	61.39	19.68	84.33	52.78	53.38
MILD (default connector)	**49.50**	**62.84**	**21.80**	**84.89**	**54.29**	**54.66**
Expert selection, weighting, and routing						
GenEval expert only	47.99	61.38	19.53	83.12	49.78	52.36
Aesthetic expert only	49.26	61.09	19.70	84.52	51.58	53.23
Uniform weights $(1/3,\ 1/3,\ 1/3)$	48.68	59.66	20.01	83.16	53.21	52.94
MILD (prompt-dependent weights)	**49.50**	**62.84**	**21.80**	**84.89**	**54.29**	**54.66**

image expertise to video generation and complement the video-level evaluations of motion quality and temporal consistency. Qualitative comparisons are provided in Appendix C.5.

4.4 Ablation Study

We ablate image guidance, motion refinement, optimization strategy, connector architecture, and expert configuration on Wan2.2-5B using VBench-2.0, keeping all other settings fixed unless otherwise stated. Results are reported in Table 5.

Supervision and optimization strategy. We remove image-expert guidance while retaining motion feedback, or remove motion refinement while retaining image guidance; both variants retain the video anchor. The full model achieves an average score of 54.66, compared with 52.06 without image experts and 53.45 without motion refinement, and outperforms both variants across all five categories. It also exceeds V2V OPD with Wan2.2-A14B under the same motion-refinement procedure (53.06), supporting the effectiveness of image-expert transfer when both approaches receive motion feedback. Joint optimization outperforms an alternating schedule of one connector update followed by 16 student updates (54.66 versus 53.87), consistent with the benefit of adapting the connector as the student policy evolves. Both schedules retain the separate objectives and gradient-detachment rules in Section 3.4.

Connector architecture. We replace the default spatial linear connector with a 1×1 linear or nonlinear connector under the same alignment objective. Linear connectors apply the same operator to states and velocities, while the nonlinear variant uses Jacobian–vector products. The 1×1 linear and nonlinear variants score 52.95 and 53.38, respectively, both above the pretrained student's 51.58 but below the default connector's 54.66. These results show that the framework supports different connector architectures, while the default design performs best among those tested.

Expert selection, weighting, and activation. We compare the default multi-expert policy with GenEval-only, aesthetic-only, and uniformly weighted variants, keeping the image-guidance scales and target-fusion weights fixed. With experts ordered as GenEval, aesthetic, and OCR, these variants use weights $(1, 0, 0)$, $(0, 1, 0)$, and $(1/3, 1/3, 1/3)$, respectively, for every prompt. The default policy scores 54.66, compared with 53.23 for the stronger single-expert variant and 52.94 for uniform weighting. This comparison supports the combined use of specialized experts with prompt-dependent weighting and activation.

5 Conclusion

We presented MILD for transferring specialized image expertise to video diffusion models through on-policy distillation. Learnable connectors align heterogeneous latent representations, while video-anchor constraints and motion feedback support spatial improvements alongside temporal consistency. Experiments across three video backbones demonstrate effective capability transfer and consistent aggregate gains over video-teacher OPD baselines. Our results establish that image experts can be a practical source of diverse supervision for video post-training.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing and LaTeX formatting, as well as feedback on methodological descriptions and interpretation of experimental results. AI-assisted suggestions were revised under author direction. The authors take full responsibility for the final content, including all scientific claims, experimental results, and references.

REFERENCES

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575. IEEE, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320. IEEE, 2024.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4DNex: Feed-forward 4d generative modeling made easy. In *1st Workshop on Reliable and Interactive World Model in Computer Vision Non Archival*, 2025.

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 *International conference on machine learning*, 2024.

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.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

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.

Yujin Han, Yujie Wei, Yefei He, Xinyu Liu, Tianle Li, Zichao Yu, Andi Han, Shiwei Zhang, Tingyu Weng, and Difan Zou. Aesrm: Improving video aesthetics with expert-level feedback. *arXiv preprint arXiv:2604.28078*, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Mingi Kwon, Seoung Wug Oh, Yang Zhou, Difan Liu, Joon-Young Lee, Haoran Cai, Baqiao Liu, Feng Liu, and Youngjung Uh. Harivo: Harnessing text-to-image models for video generation. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2024.

Dohun Lee, Bryan Sangwoo Kim, Geon Yeong Park, and Jong Chul Ye. Videoguide: Improving video diffusion models without training through a teacher's guide. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2599–2608. IEEE, 2025.

Jiachen Li, Qian Long, Jian Skyler Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Wang. T2v-turbo-v2: Enhancing video model post-training through data, reward, and conditional guidance design. In *International Conference on Learning Representations*, volume 2025, pp. 92279–92305, 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.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149. IEEE, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, volume 2025, pp. 1045–1064, 2025.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients (2024). *arXiv preprint arXiv:2407.08737*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Remi Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Christoph Schuhmann. LAION-Aesthetics, 2022. URL https://laion.ai/blog/laion-aesthetics/.

Shitong Shao, Lichen Bai, Haoyi Xiong, Zeke Xie, et al. Iv-mixed sampler: Leveraging image diffusion models for enhanced video synthesis. In *International Conference on Learning Representations*, volume 2025, pp. 12396–12424, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

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.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2096–2105. IEEE, 2023.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. *Advances in Neural Information Processing Systems*, 37:16623–16644, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Haotian Xue, Qi Chen, Zhonghao Wang, Xun Huang, Eli Shechtman, Jinrong Xie, and Yongxin Chen. Mogan: Improving motion quality in video diffusion via few-step motion adversarial post-training. *arXiv preprint arXiv:2511.21592*, 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. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement. *arXiv preprint arXiv:2608.09836*, 2026.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6463–6474. IEEE, 2024.

Yuanhao Zhai, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, David Doermann, Junsong Yuan, and Lijuan Wang. Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation. *Advances in Neural Information Processing Systems*, 37:111000–111021, 2024.

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

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Yuanzhi Zhu, Hanshu Yan, Huan Yang, Kai Zhang, and Junnan Li. Accelerating video diffusion models via distribution matching. *arXiv preprint arXiv:2412.05899*, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Method **3**

4 Experiments **6**

5 Conclusion **9**

A Detailed derivations **14**

B Implementation Details **19**

C Additional Analyses **21**

Roadmap. The appendix is organized as follows:

- Appendix A provides the detailed derivations behind the proposed method: the closed-form OPD loss from rectified flow, the connector velocity mapping, the anchored image and motion targets together with their correction bounds, and the gradients used for joint optimization.
- Appendix B gives implementation details: model configurations, training hyperparameters, and evaluation protocols.
- Appendix C presents additional analyses, including the motion-reward validation.

A Detailed derivations

We use the notation of Section 3 and suppress the conditioning variable c where unambiguous. Local transition calculations use $0 < t < 1$, $0 < h \leq t$, and a diffusion coefficient $\sigma_t > 0$ that depends only on time. Density calculations assume the regularity and boundary conditions needed to interchange derivatives and integrals. The SDE argument further assumes well-posedness and a matched initial marginal. Connector parameters are held fixed when differentiating a path with respect to diffusion time.

A.1 From rectified flow to the closed-form OPD loss

We derive the Gaussian transition kernels used in Eq. (1) and evaluate their KL divergence. The rectified-flow interpolation $x_t = (1-t)x_0 + t\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ is independent of x_0, defines the reference conditional density $p_t(x \mid x_0, c) = \mathcal{N}(x; (1-t)x_0, t^2 I)$. Differentiating its marginal gives

$$\begin{aligned}\nabla_x \log p_t(x \mid c) &= \frac{\int p_{\text{data}}(x_0 \mid c) \nabla_x p_t(x \mid x_0, c)\, \mathrm{d}x_0}{p_t(x \mid c)} \\ &= \mathbb{E}\left[-\frac{x - (1-t)x_0}{t^2} \,\middle|\, x_t = x, c \right] \\ &= -\frac{1}{t}\mathbb{E}[\epsilon \mid x_t = x, c].\end{aligned}$$

The regression-optimal rectified-flow velocity is $v_t(x, c) = \mathbb{E}[\epsilon - x_0 \mid x_t = x, c]$. Since $\epsilon = x_t + (1-t)(\epsilon - x_0)$, we have

$$\mathbb{E}[\epsilon \mid x_t = x, c] = x + (1-t)v_t(x, c).$$

Consequently, the reference score and its regression-optimal velocity satisfy

$$\nabla_x \log p_t(x \mid c) = -\frac{x + (1-t)v_t(x, c)}{t}.$$

For a learned or constructed velocity u, we use the same algebraic form to define the score estimate $\mathcal{S}_t(x, u) := -[x + (1-t)u]/t$. This estimate need not equal the score of the reference marginal.

For the exact velocity field v_t and reference score $s_t = \nabla \log p_t$, the flow satisfies $\partial_t p_t = -\nabla \cdot (p_t v_t)$. Introducing the increasing reverse clock $\tau = 1 - t$, consider

$$\mathrm{d}Y_\tau = \left[-v_t(Y_\tau, c) + \frac{\sigma_t^2}{2} s_t(Y_\tau, c) \right] \mathrm{d}\tau + \sigma_t \, \mathrm{d}B_\tau, \qquad t = 1 - \tau.$$

For $q_\tau = p_{1-\tau}$, the corresponding Fokker–Planck operator gives

$$\begin{aligned}&-\nabla \cdot \left[q_\tau \left(-v_t + \frac{\sigma_t^2}{2} s_t \right) \right] + \frac{\sigma_t^2}{2} \Delta q_\tau \\ &\quad = \nabla \cdot (q_\tau v_t) - \frac{\sigma_t^2}{2} \nabla \cdot (q_\tau s_t) + \frac{\sigma_t^2}{2} \Delta q_\tau \\ &\quad = \nabla \cdot (q_\tau v_t) = \partial_\tau q_\tau,\end{aligned}$$

where the second equality uses $q_\tau s_t = \nabla q_\tau$. Under the stated assumptions, this continuous-time SDE has the reference flow marginals. This result applies to the exact fields v_t and s_t.

For distillation, we use the corresponding drift expression to construct local transitions from predicted velocities. In decreasing diffusion time, substituting $\mathcal{S}_t(x,u)$ gives

$$\begin{aligned} b_t(x,u) &= u - \frac{\sigma_t^2}{2}\mathcal{S}_t(x,u) \\ &= \frac{\sigma_t^2}{2t}x + \chi_t u, \qquad \chi_t = 1 + \frac{\sigma_t^2(1-t)}{2t}. \end{aligned} \tag{10}$$

Euler–Maruyama discretization over a step $t \to t-h$ yields

$$x_{t-h} = x_t - h b_t(x_t,u) + \sigma_t\sqrt{h}\,\xi, \qquad \xi \sim \mathcal{N}(0,I).$$

At fixed step size h, the conditional transition is therefore

$$\pi_u(\cdot \mid x_t,t,c) = \mathcal{N}(\mu_u, \sigma_t^2 hI), \qquad \mu_u = \left(1 - \frac{h\sigma_t^2}{2t}\right)x_t - h\chi_t u, \tag{11}$$

where u is evaluated at (x_t,t,c). Evaluating the student and target velocities at the same state cancels the state-dependent term in their means:

$$\mu_{v_\theta} - \mu_{\bar{v}} = -h\chi_t(v_\theta - \bar{v}).$$

Because the two Gaussian kernels share the covariance $\sigma_t^2 hI$, their KL divergence is

$$\begin{aligned} D_{\mathrm{KL}}(\pi_{v_\theta}\|\pi_{\bar{v}}) &= \frac{\|\mu_{v_\theta} - \mu_{\bar{v}}\|_2^2}{2\sigma_t^2 h} \\ &= \frac{h\chi_t^2}{2\sigma_t^2}\|v_\theta - \bar{v}\|_2^2. \end{aligned}$$

Dividing by the number of video-latent entries d and substituting χ_t gives

$$\ell_t(v_\theta,\bar{v}) = \frac{1}{d}D_{\mathrm{KL}}(\pi_{v_\theta}\|\pi_{\bar{v}}) = \frac{h}{2\sigma_t^2}\left(1 + \frac{\sigma_t^2(1-t)}{2t}\right)^2 \|v_\theta - \bar{v}\|_{\mathrm{av}}^2,$$

which establishes Eqs. (2)–(3). The KL equality is exact for the constructed Gaussian kernels, while the Euler step approximates the continuous-time dynamics. The OPD objective in Eq. (1) evaluates this local loss at states visited by the current student; it does not assume that $p_\theta(\cdot \mid t,c)$ equals the reference marginal p_t.

A.2 Mapping image guidance to video latents

A.2.1 State and velocity mapping

Consider a differentiable rectified-flow path $x(t)$ and a connector F_ψ whose parameters remain fixed along that path. For $z(t) = F_\psi(x(t))$, differentiability gives

$$F_\psi(x(t+\Delta t)) - F_\psi(x(t)) = J_{F_\psi}(x(t))\dot{x}(t)\Delta t + o(|\Delta t|).$$

Dividing by Δt and taking the limit yields

$$\dot{z}(t) = J_{F_\psi}(x(t))\dot{x}(t).$$

Setting $\dot{x}(t) = v_\theta(x(t),t,c)$ gives the mapped velocity in Eq. (4). The connector may change between training iterations but is fixed within this local diffusion-time derivative. This chain-rule mapping requires no inverse of F_ψ and does not imply exact transport of stochastic transition distributions.

For a bias-free linear connector $F_\psi(x) = A_\psi x$, the Jacobian is constant and $J_{F_\psi}(x)v = A_\psi v$. Fixed linear slice selection and spatial resizing can be included in A_ψ, so the complete operator maps both states and velocities.

A.2.2 VELOCITY-TO-SCORE CONVERSION AND SCORE PULLBACK

The image-expert predictions $u_k(z_t, t, c)$ are velocities in image latent coordinates. Under the rectified-flow convention, we convert each prediction into the image-space score estimate

$$s_k^{\mathrm{I}} = -\frac{z_t + (1-t)u_k(z_t, t, c)}{t}.$$

As in the video-space calculation, this expression equals an image reference marginal's score when u_k is its regression-optimal velocity. For learned experts, it is the score estimate used to construct guidance.

To derive the pullback, first consider exact image scores $s_k^{\mathrm{I}}(z, t, c) = \nabla_z \log p_{k,t}^{\mathrm{I}}(z \mid c)$. With the prompt-dependent weights $\beta_k(c)$ fixed, define $\Phi_t(x) = \sum_{k=1}^K \beta_k(c) \log p_{k,t}^{\mathrm{I}}(F_\psi(x) \mid c)$. For any direction δx,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a}\Phi_t(x + a\delta x)\Big|_{a=0} &= \sum_{k=1}^K \beta_k(c)(s_k^{\mathrm{I}})^\top J_{F_\psi}(x)\delta x \\ &= \left[J_{F_\psi}(x)^\top \sum_{k=1}^K \beta_k(c) s_k^{\mathrm{I}}\right]^\top \delta x.\end{aligned}$$

Because this holds for every δx, it gives $\nabla_x \Phi_t(x) = J_{F_\psi}(x)^\top \sum_k \beta_k(c) s_k^{\mathrm{I}}$, as in Eq. (5). The vector–Jacobian product uses the scores evaluated at the queried state and does not require derivatives through the image experts. It also requires neither an inverse connector nor a change-of-variables determinant. The potential Φ_t need not define a normalized density over video latents. For learned score estimates, the same pullback defines the local image-expert direction used by the method.

A.3 CONNECTOR ALIGNMENT AND NONLINEAR TRAINING

The connector objective in Eq. (6) combines paired-latent content alignment with velocity alignment on student-visited states. In the content term, $\bar{x}_0$ and $\bar{z}_0$ are fixed encodings of the same visual content under the video and image VAEs. In the velocity term, the sampled video state x_t and student velocity $v_\theta(x_t, t, c)$ are fixed inputs, while each expert prediction is a detached target. Thus, the expert's input $z_t = F_\psi(x_t)$ may change as ψ changes, but gradients are not propagated through the expert prediction $u_k(z_t, t, c)$.

For a nonlinear connector, velocity alignment differentiates the Jacobian–vector product $J_{F_\psi}(x_t)v^{\mathrm{det}}$, where $v^{\mathrm{det}} = \mathrm{sg}[v_\theta(x_t, t, c)]$. Assuming the required mixed derivatives exist, its derivative for output coordinate j and parameter coordinate p is

$$\frac{\partial [J_{F_\psi}(x_t)v^{\mathrm{det}}]_j}{\partial \psi_p} = \sum_i \frac{\partial^2 F_{\psi,j}(x_t)}{\partial \psi_p \, \partial x_i} v_i^{\mathrm{det}}.$$

These mixed input–parameter derivatives allow the velocity term to train a nonlinear connector. For the default linear connector $F_\psi(x) = A_\psi x$, the corresponding term reduces to $A_\psi v^{\mathrm{det}}$. In either case, the connector's velocity term is an image-space mean-squared error; it does not use the reverse-SDE weight $w(t, h)$. The full parameter gradient is given in Appendix A.5.

A.4 ANCHORED IMAGE AND MOTION TARGETS

A.4.1 IMAGE-TARGET CONVERSION AND CORRECTION BOUND

Let $s_{\mathrm{A}} = \mathcal{S}_t(x_t, v_{\mathrm{A}})$. For a provisional nonnegative scale a_t, adding image guidance g_{I} to the anchor score and converting back to velocity gives

$$\begin{aligned}\frac{-t(s_{\mathrm{A}} + a_t g_{\mathrm{I}}) - x_t}{1-t} &= \frac{-t s_{\mathrm{A}} - x_t}{1-t} - \frac{t}{1-t} a_t g_{\mathrm{I}} \\ &= v_{\mathrm{A}} - \frac{t}{1-t} a_t g_{\mathrm{I}}.\end{aligned}$$

This establishes the sign and time factor of the image-guidance correction in Eq. (7) before applying $\mathcal{C}_{\mathrm{I}}$.

To set its strength, let M be a binary mask selecting video-latent coordinates covered by image supervision. For one sample, define the stabilized RMS

$$r_M(q) = \sqrt{\frac{\|M \odot q\|_2^2}{\|M\|_1 + \varepsilon}}, \qquad \varepsilon > 0. \tag{12}$$

Let $\kappa_{\mathrm{I}} \geq 0$ be the image-score RMS multiplier and $\rho_{\mathrm{I}}^{(0)} \geq 0$ the initial velocity-correction cap relative to the anchor. We normalize the image contribution and rescale its induced velocity correction if necessary:

$$\begin{aligned} a_t &= \kappa_{\mathrm{I}} \frac{r_M(s_{\mathrm{A}})}{r_M(g_{\mathrm{I}}) + \varepsilon}, \\ \delta v_{\mathrm{I}} &= -\frac{t}{1-t} a_t g_{\mathrm{I}}, \\ c_{\mathrm{I}} &= \min\left(1, \frac{\rho_{\mathrm{I}}^{(0)} r_M(v_{\mathrm{A}})}{r_M(\delta v_{\mathrm{I}}) + \varepsilon}\right). \end{aligned}$$

By homogeneity, $r_M(a_t g_{\mathrm{I}}) \leq \kappa_{\mathrm{I}} r_M(s_{\mathrm{A}})$. Setting $\alpha_t = c_{\mathrm{I}} a_t$ gives the initial image-enhanced proposal $v_{\mathrm{I}}^{(0)} = v_{\mathrm{A}} + c_{\mathrm{I}} \delta v_{\mathrm{I}} = v_{\mathrm{A}} - [t/(1-t)]\alpha_t g_{\mathrm{I}}$. Its correction obeys

$$\begin{aligned} r_M(v_{\mathrm{I}}^{(0)} - v_{\mathrm{A}}) &= c_{\mathrm{I}} r_M(\delta v_{\mathrm{I}}) \\ &\leq \rho_{\mathrm{I}}^{(0)} r_M(v_{\mathrm{A}}). \end{aligned} \tag{13}$$

This bound applies to the initial proposal. The processed target $\widetilde{v}_{\mathrm{I}} = \mathcal{C}_{\mathrm{I}}(v_{\mathrm{A}}, v_{\mathrm{I}}^{(0)})$ additionally applies temporal processing and magnitude control. A bound on $\widetilde{v}_{\mathrm{I}}$ requires the corresponding properties of $\mathcal{C}_{\mathrm{I}}$; it does not follow from Eq. (13) alone.

Temporal correction constraints. The image and motion branches use the same correction operator with different settings. For a base velocity b and proposed velocity p, define $\Delta = p - b$ and its temporal mean $\mu_\Delta = T^{-1} \sum_{\tau=1}^{T} \Delta_\tau$, broadcast across temporal slices. We separate the shared component from the temporal residual and filter the latter:

$$e = \Delta - \mu_\Delta, \qquad \widehat{\Delta} = \mu_\Delta + \eta \mathcal{H} e, \tag{14}$$

where η controls the residual contribution. The three-point temporal filter is

$$[\mathcal{H}e]_\tau = \frac{1}{4} e_{\tau-1} + \frac{1}{2} e_\tau + \frac{1}{4} e_{\tau+1}, \qquad e_0 = e_1, \quad e_{T+1} = e_T.$$

The filtered correction is scaled and added to the base:

$$\mathcal{C}_{\eta,\boldsymbol{\rho}}(b, p) = b + s_b(\widehat{\Delta})\widehat{\Delta}, \qquad 0 \leq s_b(\widehat{\Delta}) \leq 1, \tag{15}$$

where $\boldsymbol{\rho} = (\rho_g, \rho_f, \rho_d)$ specifies global, per-slice, and temporal-difference bounds. For one sample, let r denote RMS over all latent entries and r_τ RMS over channels and spatial entries at temporal slice τ. Define the adjacent-slice difference $[\mathrm{D}b]_\tau = b_{\tau+1} - b_\tau$. The per-slice and temporal-difference reference levels are

$$\begin{aligned} q_{b,\tau} &= \max(r_\tau(b), \xi r(b)), \\ q_{d,\tau} &= \max(r_\tau(\mathrm{D}b), \xi r(b)), \end{aligned}$$

where ξ supplies a floor when a local reference is small. The scaling factors are

$$\begin{aligned} s_g &= \frac{\rho_g r(b)}{r(\widehat{\Delta}) + \varepsilon}, \\ s_f &= \min_\tau \frac{\rho_f q_{b,\tau}}{r_\tau(\widehat{\Delta}) + \varepsilon}, \\ s_d &= \min_{\tau < T} \frac{\rho_d q_{d,\tau}}{r_\tau(\mathrm{D}\widehat{\Delta}) + \varepsilon}, \\ s_b(\widehat{\Delta}) &= \min(1, s_g, s_f, s_d). \end{aligned}$$

One scalar is applied to every entry of each sample. Writing $\Delta_{\text{safe}} = s_b(\widehat{\Delta})\widehat{\Delta}$, linearity of D and homogeneity of RMS give

$$\begin{aligned} r(\Delta_{\text{safe}}) &\le \rho_g r(b), \\ r_\tau(\Delta_{\text{safe}}) &\le \rho_f q_{b,\tau}, \\ r_\tau(\mathrm{D}\Delta_{\text{safe}}) &\le \rho_d q_{d,\tau}. \end{aligned}$$

The operators in the main text are $\mathcal{C}_{\mathrm{I}} = \mathcal{C}_{\eta_{\mathrm{I}},\boldsymbol{\rho}_{\mathrm{I}}}$ and $\mathcal{C}_{\mathrm{M}} = \mathcal{C}_{\eta_{\mathrm{M}},\boldsymbol{\rho}_{\mathrm{M}}}$. Their parameter values are reported in Appendix B.

A.4.2 Motion-refinement sign and correction bound

Motion refinement evaluates the clean latent predicted by $\widetilde{v}_{\mathrm{I}}$. Let $b_{\mathrm{I}} = \mathrm{sg}[\widetilde{v}_{\mathrm{I}}]$ and define $f(x_0) = R(D^{\mathrm{V}}(x_0))$. Holding b_{I} fixed gives

$$\begin{aligned} \widehat{x}_0 &= x_t - t b_{\mathrm{I}}, \\ g_{\mathrm{M}} &= \nabla_x f(x - t b_{\mathrm{I}})|_{x=x_t} = \nabla f(\widehat{x}_0). \end{aligned}$$

If $y = D^{\mathrm{V}}(\widehat{x}_0)$, $m = \mathcal{O}(y)$ is the optical flow computed by the frozen estimator, and $R(y) = D_{\mathrm{M}}(\mathcal{O}(y))$, the chain rule gives

$$g_{\mathrm{M}} = J_{D^{\mathrm{V}}}(\widehat{x}_0)^\top J_{\mathcal{O}}(y)^\top \nabla_m D_{\mathrm{M}}(m).$$

The decoder, flow estimator, and discriminator remain frozen but retain these input derivatives.

Define the per-sample RMS $r(q) = \sqrt{\|q\|_2^2/d^{\mathrm{V}}}$ and let $s_{\widetilde{\mathrm{I}}} = \mathcal{S}_t(x_t, \widetilde{v}_{\mathrm{I}})$. Let $\chi_{\mathrm{M}} \ge 0$ be the gradient RMS multiplier, $\kappa_{\mathrm{M}} \ge 0$ the motion-guidance scale, $\kappa_{\max} \ge 0$ the time-factor cap, and $\rho_{\mathrm{M}}^{(0)} \ge 0$ the initial velocity-correction cap. We compute

$$\begin{aligned} \bar{g}_{\mathrm{M}} &= \chi_{\mathrm{M}} \frac{r(s_{\widetilde{\mathrm{I}}})}{r(g_{\mathrm{M}}) + \varepsilon} g_{\mathrm{M}}, \\ \delta v_{\mathrm{M}} &= -\kappa_{\mathrm{M}} \min\left(\frac{t}{1-t}, \kappa_{\max}\right) \bar{g}_{\mathrm{M}}, \\ c_{\mathrm{M}} &= \min\left(1, \frac{\rho_{\mathrm{M}}^{(0)} r(\widetilde{v}_{\mathrm{I}})}{r(\delta v_{\mathrm{M}}) + \varepsilon}\right). \end{aligned}$$

Collecting the scalar factors, define

$$\gamma_t = c_{\mathrm{M}} \kappa_{\mathrm{M}} \min\left(\frac{t}{1-t}, \kappa_{\max}\right) \chi_{\mathrm{M}} \frac{r(s_{\widetilde{\mathrm{I}}})}{r(g_{\mathrm{M}}) + \varepsilon}.$$

The initial motion-refined proposal is $v_{\mathrm{M}}^{(0)} = \widetilde{v}_{\mathrm{I}} + c_{\mathrm{M}} \delta v_{\mathrm{M}} = \widetilde{v}_{\mathrm{I}} - \gamma_t g_{\mathrm{M}}$. By homogeneity,

$$\begin{aligned} r(v_{\mathrm{M}}^{(0)} - \widetilde{v}_{\mathrm{I}}) &= c_{\mathrm{M}} r(\delta v_{\mathrm{M}}) \\ &\le \rho_{\mathrm{M}}^{(0)} r(\widetilde{v}_{\mathrm{I}}). \end{aligned} \tag{16}$$

The sign follows from its effect on the predicted clean latent:

$$x_t - t v_{\mathrm{M}}^{(0)} = \widehat{x}_0 + t \gamma_t g_{\mathrm{M}}.$$

For a nonzero g_{M} and a positive scalar $a \to 0$, differentiability gives

$$f(\widehat{x}_0 + a g_{\mathrm{M}}) = f(\widehat{x}_0) + a \|g_{\mathrm{M}}\|_2^2 + o(a).$$

Thus, a sufficiently small positive correction follows a local reward-ascent direction for the initial clean latent. This does not guarantee reward improvement after a finite correction, temporal processing, target fusion, or a full rollout. The processed target $\widetilde{v}_{\mathrm{M}} = \mathcal{C}_{\mathrm{M}}(\widetilde{v}_{\mathrm{I}}, v_{\mathrm{M}}^{(0)})$ additionally applies temporal and magnitude control. The fused target is detached before student optimization, so the student loss does not differentiate through the reward gradient.

A.5 Gradient derivation for joint optimization

We differentiate Eq. (9) with sampled states, motion-activation indicators, and constructed targets held fixed. The sampling distribution of the current student is not differentiated.

At each state, let $\mathcal{T}_t$ contain $(\omega_{\mathrm{A}}, v_{\mathrm{A}})$ and $(\omega_{\mathrm{I}}, \widetilde{v}_{\mathrm{I}})$, together with $(\omega_{\mathrm{M}}, \widetilde{v}_{\mathrm{M}})$ when $r_t = 1$. Then $Z_t = \sum_{(\omega,q)\in\mathcal{T}_t} \omega$ and $\bar{v}_t = Z_t^{-1} \sum_{(\omega,q)\in\mathcal{T}_t} \omega q$. Expanding the squared residuals gives

$$\sum_{(\omega,q)\in\mathcal{T}_t} \omega \ell_t(v,q) = Z_t \ell_t(v,\bar{v}_t) + w(t,h) \sum_{(\omega,q)\in\mathcal{T}_t} \omega \|q - \bar{v}_t\|_{\mathrm{av}}^2.$$

With the targets fixed, the second term is independent of student parameters. Hence, the fused-target loss has the same student gradient as the weighted sum of the active video-space losses, although their scalar values differ.

Let $V_{\theta,j} = J_\theta v_\theta(x_{t_j}, t_j, c)$ and $e_j = v_\theta(x_{t_j}, t_j, c) - \mathrm{sg}[\bar{v}_{t_j}]$. The student gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{student}} = \frac{2}{N} \sum_{j=1}^{N} \frac{Z_{t_j} w(t_j, h_j)}{d^{\mathrm{V}}} V_{\theta,j}^\top e_j.$$

Image expertise enters this gradient through $\bar{v}_{t_j}$; the student loss has no separate image-space prediction-matching term. Since the fused target is detached, $\nabla_\psi \mathcal{L}_{\mathrm{student}} = 0$.

For connector optimization, write $B = J_{F_\psi}(x_t)$, $v^{\mathrm{det}} = \mathrm{sg}[v_\theta(x_t, t, c)]$, and $u_k^{\mathrm{det}} = \mathrm{sg}[u_k(F_\psi(x_t), t, c)]$. Hold the latent inputs and these student and expert predictions fixed when differentiating with respect to ψ. Define $q = F_\psi(\bar{x}_0) - \bar{z}_0$ and $e_k^{\mathrm{conn}} = B v^{\mathrm{det}} - u_k^{\mathrm{det}}$. Differentiating Eq. (6) gives

$$\begin{aligned}\nabla_\psi \mathcal{L}_{\mathrm{conn}} = {} & \frac{2\lambda_z}{d^{\mathrm{I}}} \mathbb{E}_{\mathrm{pair}}\left[(J_\psi F_\psi(\bar{x}_0))^\top q\right] \\ & + \frac{2\lambda_v}{K d^{\mathrm{I}}} \sum_{k=1}^{K} \mathbb{E}_{\mathrm{rollout}}\left[(J_\psi(B v^{\mathrm{det}}))^\top e_k^{\mathrm{conn}}\right] + \lambda_{\mathrm{reg}} \nabla_\psi \mathcal{R}(\psi).\end{aligned}$$

The velocity-alignment term uses mean-squared error without the reverse-SDE weight $w(t,h)$. Its parameter derivative for a nonlinear connector is given in Appendix A.3. Detached student predictions and latent inputs imply $\nabla_\theta \mathcal{L}_{\mathrm{conn}} = 0$.

For the initialization regularizer, we use

$$\mathcal{R}(\psi) = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} \frac{\|p - p^{(0)}\|_2^2}{d_p}, \qquad \nabla_p \mathcal{R} = \frac{2(p - p^{(0)})}{|\mathcal{P}| d_p},$$

where $\mathcal{P}$ is the set of connector parameter tensors, d_p counts entries in tensor p, and $p^{(0)}$ is its initial value. Each tensor is normalized by its size before averaging.

B Implementation Details

Checkpoints and optimization. The student checkpoints are Wan2.1-T2V-1.3B, Wan2.2-TI2V-5B, and LTXV-2B-0.9.6-dev, referred to in the main text as Wan2.1-1.3B, Wan2.2-5B, and LTX-Video-2B. The image experts are SD3.5-Medium models with separate LoRA adapters (Hu et al., 2021). The GenEval, aesthetic, and OCR experts are trained for 200, 1,800, and 1,800 training steps, respectively. Expert predictions use a classifier-free guidance scale of 4.5.

For MILD post-training, we sample one prompt per iteration from the 10,000-prompt pool drawn from OpenVid-1M (Nan et al., 2025) and LVD-2M (Xiong et al., 2024). The current student generates one on-policy trajectory for an 81-frame video at 480×832 resolution, giving a global batch size of one video. We sample four states from this trajectory using the trajectory-index ranges in Table 6. The first 50 iterations update only the connector shared by the three image experts for that backbone; the following 200 iterations jointly optimize the student and connector.

Table 6: **Backbone-specific implementation settings.** Trajectory-index ranges specify candidate states for OPD supervision; temporal stride is measured in RGB-frame indices.

Setting	Wan2.1-1.3B	Wan2.2-5B	LTX-Video-2B
Student denoising steps	50	50	40
OPD trajectory-index range	2–47	2–47	2–37
Student learning rate	5×10^{-7}	3×10^{-7}	2×10^{-6}
Connector learning rate	5×10^{-6}	1×10^{-5}	2×10^{-6}
Connector hidden channels	16	48	128
Image-supervision temporal stride	4	4	8
Supervised latent slices	21	21	11
Effective image-score multiplier κ_{I}	0.10	0.10	0.20
Initial image velocity cap $\rho_{\mathrm{I}}^{(0)}$	0.10	0.10	0.25
Motion guidance scale κ_{M}	0.02	0.02	0.05
Initial motion velocity cap $\rho_{\mathrm{M}}^{(0)}$	0.05	0.05	0.10
Anchor fusion weight ω_{A}	0.15	0.15	0.15
Image-target fusion weight ω_{I}	0.10	0.10	0.10
Motion-target fusion weight ω_{M}	0.05	0.05	0.05

During joint training, losses and gradients for all sampled states are evaluated using parameters at the start of the iteration. Gradients are accumulated before either optimizer is stepped; the student optimizer is stepped first, followed by the connector optimizer. Student optimization uses AdamW (Loshchilov & Hutter, 2017) with zero weight decay, a cosine learning-rate schedule, a 5% learning-rate warm-up ratio, and a minimum learning rate of 10% of the initial value. Student and connector gradients are clipped at 1.0 and 0.1, respectively. Computation uses BF16, while master parameters and optimizer states are maintained in FP32. We maintain FP32 exponential moving-average weights with decay 0.99. The Wan samplers use a sigma shift of 5.0; LTX-Video uses a guidance rescale of 0.7 and a maximum text length of 256 tokens. Backbone-specific settings are listed in Table 6.

Connector implementation. The default connector applies bilinear resizing followed by 1×1, 3×3, and 1×1 convolutions. All convolutions omit biases and nonlinear activations. The input and hidden channel counts match the student's latent channel count, while the output has 16 channels. We initialize the connector as a near-identity mapping. The latent-alignment, velocity-alignment, and parameter-regularization coefficients in Eq. (6) are 0.1, 0.01, and 0.001, respectively. Temporal sampling uses the strides in Table 6, covering all temporal latent slices of the evaluated videos.

Temporal correction settings. The image and motion branches use the correction operator defined in Appendix A.4.1, with different parameters. The image branch uses $\eta_{\mathrm{I}} = 0.25$ and $\boldsymbol{\rho}_{\mathrm{I}} = (0.05, 0.05, 0.10)$, taking v_{A} as its base velocity. The motion branch uses $\eta_{\mathrm{M}} = 0.50$ and $\boldsymbol{\rho}_{\mathrm{M}} = (0.025, 0.025, 0.05)$, taking $\widetilde{v}_{\mathrm{I}}$ as its base. Both branches use the reference floor $\xi = 0.05$. The operator acts across all latent temporal slices without a coverage mask. Its final scalar scales both the shared component and filtered temporal residual; no further filtering or clipping is applied after target fusion.

Motion representation and discriminator. For each adjacent-frame flow field (u_i, v_i), we construct $M_i = (u_i, v_i, \sqrt{u_i^2 + v_i^2})$ and stack these fields along the temporal dimension. We use the torchvision implementation of RAFT-Small (Teed & Deng, 2020) with 12 flow-estimation updates. Motion volumes used for discriminator training are computed offline and cached in FP32. During OPD, flow is recomputed within the differentiable reward computation. The discriminator consists of a 3D convolutional stem and three factorized spatiotemporal residual stages, with spatial convolution followed by temporal convolution in each block. Spatial average pooling produces temporal tokens, which are augmented with temporal positional embeddings and processed by a two-layer Transformer before a scalar-logit output head.

Table 7: **Motion-discriminator training data.** Generated negatives use the corresponding video backbone; counts precede balanced sampling.

Source	Train	Validation
Real videos	3,412	379
Generated videos	3,412	379
Temporal hard negatives	3,412	379
Total	10,236	1,137

Motion-discriminator data and objective. Motion-discriminator training is separate from the 10,000-prompt MILD post-training setup. All three motion discriminators use the same real-video sources and data-construction protocol. The positive collection contains 1,791 videos from OpenVid-1M (Nan et al., 2025) and 2,000 videos from 4DNeX-10M (Chen et al., 2025). For each student backbone, we generate one negative video per positive example using its associated caption and corresponding pretrained video model: Wan2.1-1.3B, Wan2.2-5B (Wan et al., 2025), or LTX-Video-2B (HaCohen et al., 2024). Generation uses text-only conditioning.

Each positive video also contributes one temporal hard negative, constructed by uniformly sampling frame shuffling, repetition and dropping, temporal jitter, or discontinuous temporal splicing. Reversal and static-video transformations are excluded because they need not imply implausible motion. Each corrupted example remains paired with its source video within the same split. This produces 11,373 entries per discriminator, equally divided among positive videos, model-generated negatives, and temporal hard negatives. Table 7 reports the training and validation counts.

Let m^+ denote real motion volumes and m^- denote generated or temporally corrupted volumes. We use a logistic discrimination objective with R1 regularization (Mescheder et al., 2018):

$$\mathcal{L}_{D_{\mathrm{M}}} = \mathbb{E}_{m^+}[\mathrm{softplus}(-D_{\mathrm{M}}(m^+))] + \mathbb{E}_{m^-}[\mathrm{softplus}(D_{\mathrm{M}}(m^-))] + \frac{\lambda_{\mathrm{R1}}}{2}\mathbb{E}_{m^+}[\|\nabla_{m^+} D_{\mathrm{M}}(m^+)\|_2^2].$$

We use AdamW with learning rate 2×10^{-4}, batch size four, balanced sampling of positive and negative examples, and BF16 computation for 20 epochs. The R1 coefficient is $\lambda_{\mathrm{R1}} = 1$, with the penalty evaluated every 16 optimizer steps.

Prompt routing and motion-refinement settings. The prompt router detects visible-text requests using text-related keywords and quoted strings. For image-score aggregation, the normalized weights of the GenEval, aesthetic, and OCR experts are $(0.5, 0.5, 0)$ for ordinary prompts and $(0.4, 0.4, 0.2)$ for text-related prompts. These weights apply only to score aggregation. Connector training uses paired clean-latent alignment and equally weighted velocity-alignment targets from all three experts for every prompt, including when the OCR score-pullback weight is zero.

The effective image-score multiplier κ_{I} is the minimum of the configured image-score scale and score-ratio ceiling. It is 0.10 for the Wan backbones and 0.20 for LTX-Video-2B. Motion refinement uses gradient RMS multiplier $\chi_{\mathrm{M}} = 1$ and time-factor cap $\kappa_{\max} = 4$. Backbone-specific motion-guidance scales κ_{M} and initial velocity caps $\rho_{\mathrm{M}}^{(0)}$ are listed in Table 6. These coefficients are distinct from the temporal residual scales η_{I} and η_{M}.

During student updates, at most one of the four sampled states receives motion refinement, provided its diffusion time satisfies $t \in [0.2, 0.6]$. If no state is eligible, motion refinement is skipped for that iteration. The fusion weights $(\omega_{\mathrm{A}}, \omega_{\mathrm{I}}, \omega_{\mathrm{M}})$ are shared across all three backbones, whereas image and motion guidance scales are backbone-specific, as shown in Table 6. The active fusion weights sum to $Z_t = 0.25$ without motion refinement and $Z_t = 0.30$ with it. This sum determines both target normalization and the outer student-loss coefficient.

C Additional Analyses

C.1 Motion-Reward Validation

We evaluate the learned motion discriminators through three complementary metrics. AUROC measures their ability to distinguish motion volumes from real videos from those of generated or tem-

Table 8: **Validation of the motion discriminators.** Higher AUROC and pairwise accuracy are better. Reward–magnitude correlation is reported as a signed value.

Student backbone	AUROC	Pairwise accuracy	Reward–magnitude correlation
Wan2.1-1.3B	0.9552	0.9472	−0.0817
Wan2.2-5B	0.9288	0.9736	−0.1078
LTX-Video-2B	0.8925	0.9538	−0.0521

Table 9: **Comparison with direct aesthetic video fine-tuning.** Aesthetic SFT denotes video models fine-tuned on aesthetic video data and evaluated directly, without subsequent OPD. Bold indicates the better result within each backbone. Higher is better for all metrics.

	VBench-2.0						EvalCrafter				
Method	Creativity	Commonsense	Controllability	Human Fidelity	Physics	AVG.	Visual Quality	Text–Video Alignment	Motion Quality	Temporal Consistency	Final Sum Score
Wan2.1-1.3B											
Aesthetic SFT	**49.84**	61.13	24.94	**86.45**	44.94	53.46	66.38	**58.90**	54.18	63.34	242.80
MILD	47.89	**64.56**	**25.17**	85.47	**46.52**	**53.92**	**66.84**	57.89	**54.58**	**63.79**	**243.10**
Wan2.2-5B											
Aesthetic SFT	**50.65**	**63.13**	**22.94**	83.44	46.76	53.38	63.60	57.89	**54.45**	**62.88**	238.82
MILD	49.50	62.84	21.80	**84.89**	**54.29**	**54.66**	**63.76**	**59.12**	54.43	62.79	**240.10**
LTX-Video-2B											
Aesthetic SFT	57.01	47.99	15.80	**73.35**	48.72	48.57	**57.57**	47.87	55.16	64.39	**224.99**
MILD	**57.23**	**48.85**	**16.78**	72.38	**50.06**	**49.06**	56.58	**48.41**	**55.30**	**64.51**	224.80

porally corrupted videos. Pairwise accuracy measures how often a real motion volume receives a higher reward than its temporally corrupted counterpart. Reward–magnitude correlation measures the association between the discriminator logit and mean optical-flow magnitude. Together, these metrics assess discrimination performance, sensitivity to temporal corruption, and association with motion magnitude.

As shown in Table 8, all three discriminators distinguish real motion from the evaluated negative samples and consistently rank original motion above its temporally corrupted counterpart. The reward–magnitude correlations are small in absolute value, indicating limited linear association with motion magnitude on these validation samples. These findings support sensitivity to the evaluated temporal corruptions, without establishing general motion understanding or independence from motion magnitude.

Under this shared data protocol, positive and generated videos are partitioned independently, while each temporal corruption remains in the same split as its source video. The resulting split is disjoint at the video-file level, but caption-matched positive and generated videos can appear in different splits. Frame rates are also not standardized before flow extraction. The reported results therefore characterize performance under this validation protocol and do not establish generalization to unseen caption groups or robustness to frame-rate differences.

C.2 Comparison with Aesthetic Video Fine-Tuning

Quantitative comparison. We compare MILD with Aesthetic SFT video teacher, which directly fine-tunes each pretrained video backbone on aesthetic video data. Table 9 shows that MILD achieves higher VBench-2.0 average scores across all three backbones and higher EvalCrafter aggregate scores on both Wan backbones, while scoring slightly lower on LTX-Video-2B (224.80 vs. 224.99). MILD also achieves higher physics scores across all three backbones, while improvements in other dimensions vary by backbone. These results support the effectiveness of transferring specialized image expertise compared with direct aesthetic video fine-tuning.

Qualitative comparison. Figure 7 compares MILD with direct aesthetic video fine-tuning using Wan2.2-5B on VBench-2.0. For the moon prompt, MILD shows a clear transition toward yellow while retaining visible lunar texture. For the cat prompt, MILD depicts a single cat moving left and leaving the frame, whereas Aesthetic SFT introduces an additional cat and does not clearly depict

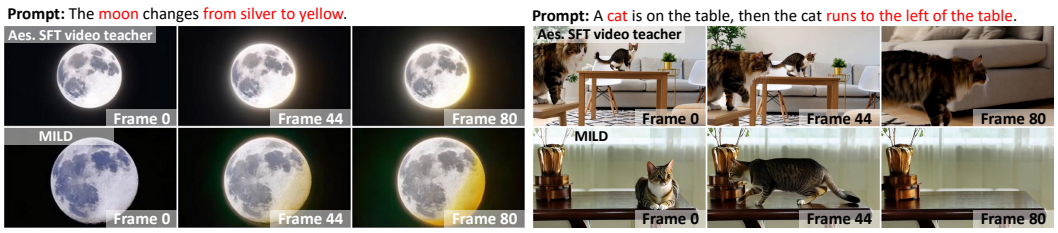

Figure 7: **Qualitative comparison with direct aesthetic video fine-tuning on VBench-2.0 using Wan2.2-5B.** Frames 0, 44, and 80 illustrate appearance changes and directional motion under matched prompts.

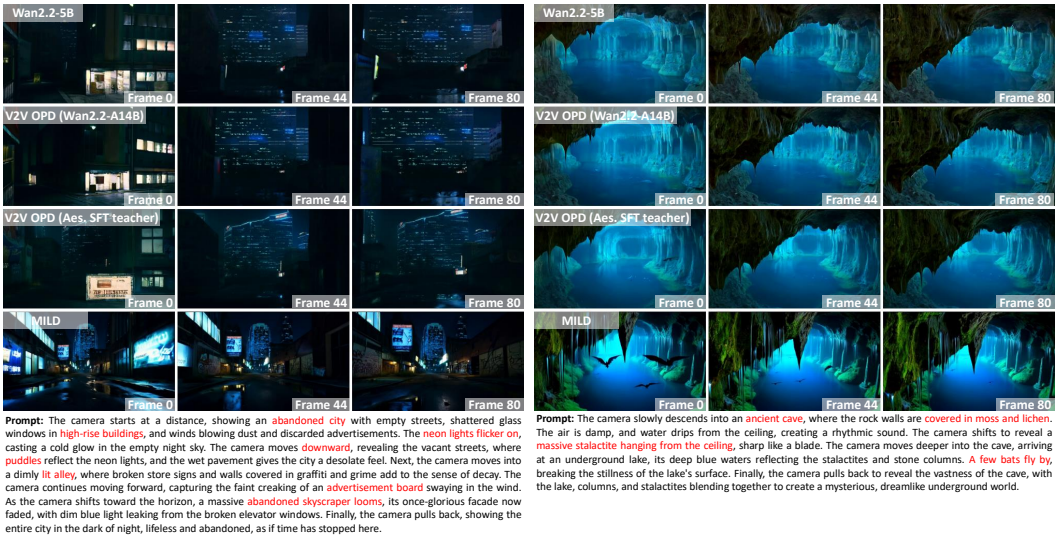

Figure 8: **Abandoned-city and underground-cave comparisons.** Left: MILD depicts prominent illuminated signs, wall markings, and wet-pavement reflections. Right: MILD renders distinct foreground stalactites, textured rock surfaces, and visible bat silhouettes. Each example shows frames 0, 44, and 80.

the requested leftward motion. These examples complement the quantitative results, illustrating MILD's ability to follow temporal instructions while preserving visual details.

C.3 Additional Video Case Studies

We present five additional qualitative comparisons using Wan2.2-5B as the student, complementing the cases in Section 4.2. Each example compares the pretrained student, V2V OPD with Wan2.2-A14B, V2V OPD with an aesthetic-SFT teacher, and MILD under the same prompt. Frames 0, 44, and 80 are shown for each generated video. These examples examine scene detail, text legibility, and subject composition.

Scene detail and spatial organization. Figure 8 compares the abandoned-city and underground-cave examples. In the abandoned city, MILD depicts prominent illuminated signs, visible wall markings, and reflections on the wet pavement. These details make the foreground street and surrounding buildings easier to distinguish than in the darker baseline outputs. In the cave, MILD renders pronounced foreground stalactites, textured green rock surfaces, and recognizable bat silhouettes above the blue underground lake. The bats are particularly visible in the first sampled frame and appear farther into the cave in the later frames. The foreground rock formations and illuminated background also provide a clear sense of depth.

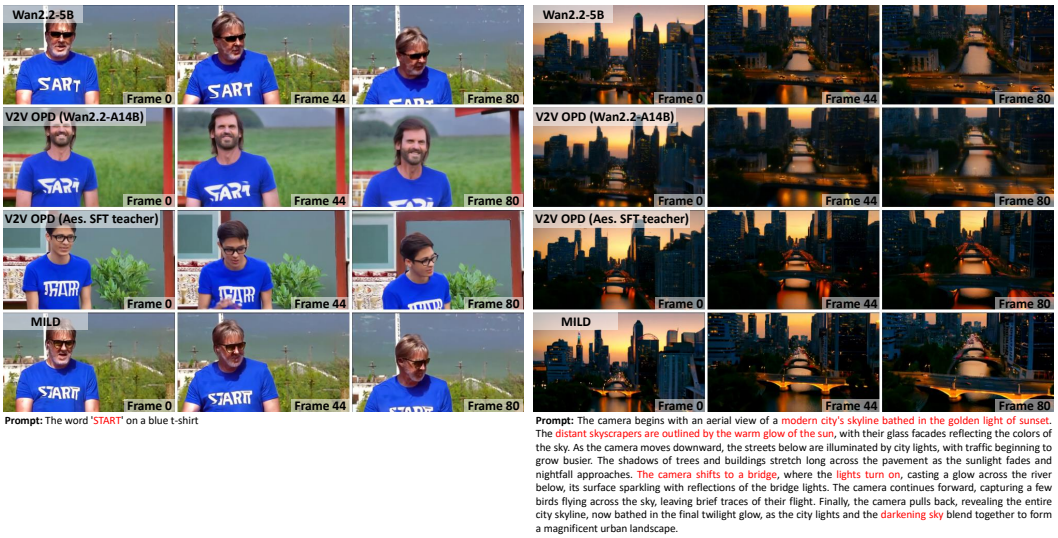

Figure 9: **Text-rendering and twilight-city comparisons.** Left: MILD renders "START" more clearly in the frames where the complete word is visible. Right: MILD depicts distinct architecture, illuminated bridge contours, and water reflections. Each example shows frames 0, 44, and 80.

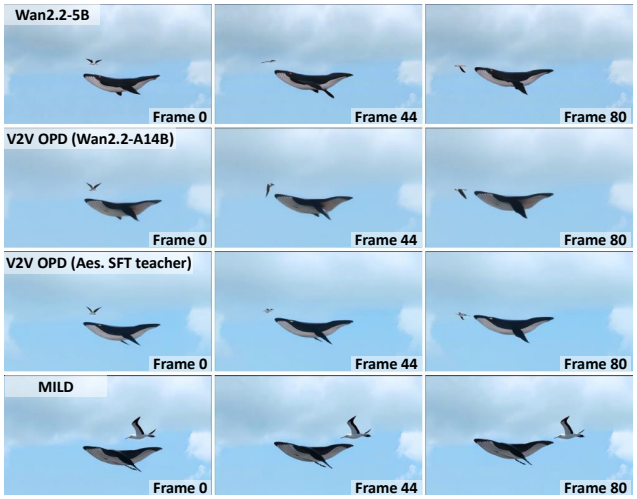

Figure 10: **Whale-and-seagull composition.** MILD depicts a more prominent seagull alongside the whale, with both subjects clearly identifiable across frames 0, 44, and 80.

Text rendering and urban scene detail. Figure 9 presents the blue-T-shirt and twilight-city examples. For the T-shirt prompt, the pretrained student renders "SART," while the V2V OPD outputs contain distorted or ambiguous lettering. MILD renders "START" more clearly in the first two sampled frames, where the word is fully visible; the lettering is partially cropped in the final frame. In the twilight-city example, MILD depicts distinct building outlines and illuminated bridges along the central waterway. The bridge contours and their reflections remain readily identifiable across the displayed viewpoints, illustrating detailed architectural content within the generated sequence.

Subject visibility and composition. Figure 10 examines the unusual composition of a whale and a seagull flying together. Although both subjects appear in the baseline outputs, their seagulls occupy relatively small image regions. MILD depicts a larger, more prominent seagull with a clearly

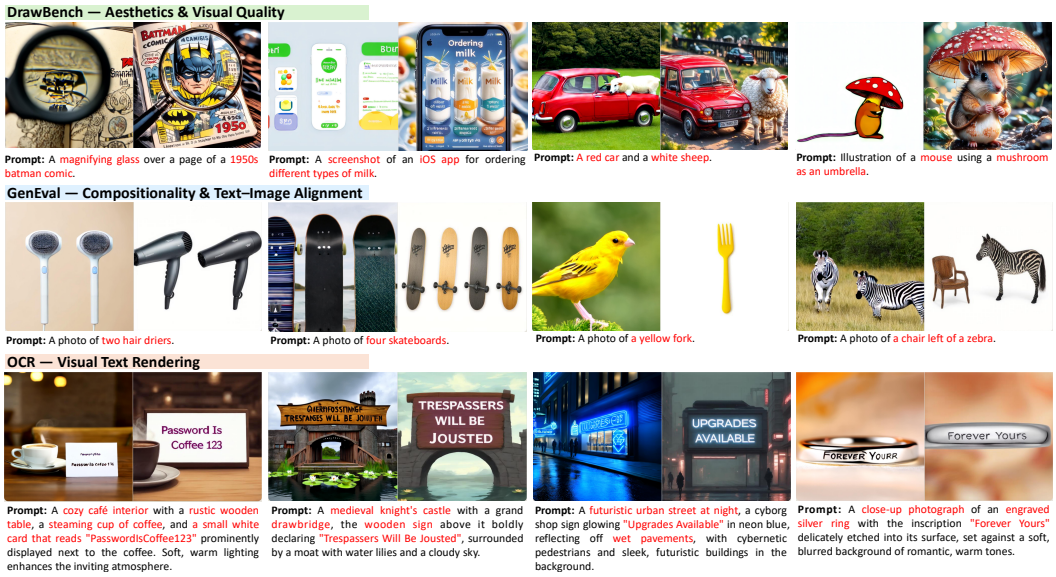

Figure 11: **Qualitative comparison of aesthetic image and video teachers.** Each pair shows the Wan2.2-5B video teacher post-trained on aesthetic video data (left) and the aesthetic image expert (right) under the same prompt. Rows show DrawBench, GenEval, and OCR examples, illustrating differences in visual detail, compositional accuracy, and text rendering available for distillation.

distinguishable body and wing silhouette alongside the whale. Both subjects remain identifiable in all three sampled frames, making the requested two-subject composition visually explicit.

C.4 Qualitative Comparison of Aesthetic Teachers

Figure 11 compares the aesthetic image expert with the Wan2.2-5B video teacher post-trained on aesthetic video data under matched prompts. These examples examine the visual expertise available for distillation in the Wan2.2-5B student setting, complementing the quantitative teacher comparisons in Figure 4. On DrawBench prompts, the image expert produces more clearly defined details in the magnifying-glass and comic example and richer texture and lighting in the mouse-and-mushroom scene. The GenEval cases further reveal differences in compositional accuracy: the image expert depicts four skateboards, a yellow fork, and a chair to the left of a zebra, whereas the video teacher misses the requested count, object identity, or relation. On OCR prompts, the image expert renders phrases such as "Trespassers Will Be Jousted," "Upgrades Available," and "Forever Yours" more legibly and accurately. However, improved legibility does not always imply exact text reproduction, as illustrated by the altered spacing in the coffee example. Together, these selected cases illustrate spatial capabilities that image experts can offer for video distillation; they do not assess temporal quality or establish student gains.

C.5 Qualitative Comparisons of Spatial Capabilities

Figure 12 compares MILD with the pretrained Wan2.2-5B student and two V2V OPD baselines under matched prompts, complementing the quantitative evaluation in Section 4.3.

Visual appearance and attribute alignment. For the black-dog prompt, MILD depicts a predominantly black coat with visible fur shading and facial details. The pretrained student and aesthetic-SFT-teacher OPD produce dogs with substantial tan or white markings, whereas the larger-video-teacher baseline also captures the requested black coat. This example illustrates MILD's ability to combine subject detail with color-attribute alignment.

Text rendering. MILD renders "FREE HUGS" clearly on the arm and "LAST CALL" legibly on the neon sign. The pretrained student and aesthetic-SFT-teacher OPD do not clearly reproduce



Figure 12: **Spatial capability comparisons using Wan2.2-5B.** Columns show the pretrained student, OPD with an aesthetic-SFT video teacher, OPD with a larger video teacher, and MILD. Matched-prompt examples illustrate color-attribute alignment, text rendering, and object composition in sampled video frames.



Figure 13: **Spatial capability comparison with direct aesthetic video fine-tuning using Wan2.2-5B.** Matched-prompt examples illustrate color-attribute alignment, text rendering, and object composition in sampled video frames.

the requested phrases, while the larger-video-teacher baseline produces incomplete or altered text, such as “Fre Hugs”. These examples illustrate improved text fidelity on both skin and illuminated signage.

Compositional generation. MILD depicts the requested two cats and one dog, three oranges, and four bowls. The pretrained student and aesthetic-SFT-teacher OPD omit one cat in the animal example. In the orange example, these two baselines show only two visible oranges, while the larger-video-teacher baseline depicts four. The bowl examples likewise show count mismatches among the baselines. Together, these cases illustrate more accurate object counts and category composition with MILD.

Comparison with direct aesthetic video fine-tuning. Figure 13 further compares MILD with Aesthetic SFT video teacher, evaluated directly without subsequent OPD. Using Wan2.2-5B, MILD more closely matches the requested black coat, clearly renders "FREE HUGS" and "LAST CALL", and satisfies the specified object counts. Aesthetic SFT exhibits color-attribute, text, and counting errors in these examples. These comparisons illustrate MILD's advantages in spatial prompt fidelity over direct aesthetic video fine-tuning.