ON-POLICY TRAJECTORY DISTILLATION FOR FEW-STEP GENERATIVE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step diffusion models promise interactive generation, but compressing a long teacher trajectory into a few student transitions often degrades fidelity. Existing approaches expose a trade-off. Trajectory and consistency distillation retain paired teacher targets but commonly query states that the few-step student does not visit, while distribution matching follows student-generated samples but replaces those targets with indirect distributional feedback that can favor modes already represented by the student. We present OPTD, which resolves this trade-off by rolling the student to an inference state and matching one coarse transition to the endpoint obtained by integrating a frozen teacher through the same interval. The base method therefore combines on-policy states with dense paired supervision using one trainable network. We add a boundary-matched GAN that compares the student's next-boundary latent with a same-condition real latent at the identical noise level in the frozen teacher's feature space. One endpoint construction supports image, bidirectional video, and causal video generation. Across image and video experiments, OPTD delivers strong and stable few-step generation quality across the evaluated settings.

1 INTRODUCTION

Few-step diffusion models promise interactive image and video generation, yet compressing a teacher trajectory into a few student transitions often degrades fidelity. A diffusion sample repeatedly evaluates a guided denoiser, and classifier-free guidance can require both conditional and unconditional evaluations at every step (Ho et al., 2020; Song et al., 2021; Esser et al., 2024; Ho & Salimans, 2022; Team Wan et al., 2025). Few-step distillation reduces this cost by asking one student transition to replace several teacher updates. Because that transition determines the state used by every later step, the objective must supervise the states and intervals encountered by the deployed student.

Trajectory distillation trains a few-step student to approximate multi-step diffusion transitions. Existing methods include teacher-based trajectory and consistency distillation, which use targets from teacher trajectories (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Lu & Song, 2025), and interval-learning methods such as MeanFlow and Shortcut Models, which learn coarse transitions without paired teacher supervision (Geng et al., 2025; Frans et al., 2025). Despite this difference, both typically rely on forward-noised data or reference trajectories rather than states visited by the current few-step sampler, leading to a train–test state mismatch analogous to exposure bias (Ranzato et al., 2016; Bengio et al., 2015).

Distribution-matching methods retain student-generated samples, so their feedback is on-policy (Sauer et al., 2024b; Yin et al., 2024b;a; Sauer et al., 2024a; Luo et al., 2025). They replace the paired teacher target with feedback from an auxiliary score model, critic, or discriminator. The resulting signal is distributional rather than paired to the teacher trajectory, and its mode-seeking tendency can favor regions the student already covers. Later methods modify the divergence or add consistency terms to reduce this limitation (Xu et al., 2025; Zheng et al., 2026; Fan et al., 2026). These two lines of work therefore separate two desirable properties. Paired methods provide direct teacher supervision but may query off-policy states, whereas distribution matching follows the student but gives up the paired target. OPTD is designed to retain both.

We retain both properties with a boundary-aligned endpoint contract. The student is rolled out with the same boundary grid and update rule used at inference. At a randomly selected boundary, the

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 1: **Overview of OPTD.** OPTD is a unified and effective on-policy trajectory distillation method for few-step generative modeling that applies to image, bidirectional video, and causal video distillation.

frozen teacher starts from the student-visited state and integrates its fine solver across the complete next student interval. The student takes one coarse update over that interval, and only the teacher's terminal state is used as the target. Detaching the rollout prefix keeps gradients on the selected transition while allowing the query to follow the student's current behavior. This couples each on-policy query with the paired endpoint that the deployed student must reach.

This endpoint contract defines the base OPTD objective. It provides dense paired supervision with one trainable student and one optimizer, while the teacher's intermediate states are used only to construct the endpoint. Two optional refinements improve complementary properties without changing this contract. A boundary-matched GAN compares the student's next-boundary latent with a same-condition real latent at the identical noise level in the frozen teacher's feature space, adding local realism feedback without replacing the paired target. A perceptual refinement compares the student and teacher clean predictions in a frozen visual feature space, adding semantic and local temporal supervision. The same contract adapts to image, bidirectional video, and causal video generation, and experiments across these settings show strong and stable few-step generation quality. Together, these design choices lead to the following contributions.

- We formulate few-step diffusion distillation as on-policy endpoint regression. A student-visited state and the frozen teacher's endpoint over the same interval form one paired training example.
- We introduce complementary boundary-matched adversarial and perceptual refinements. The former adds local realism feedback at the student's next boundary without replacing the paired endpoint target, while the latter compares clean student and teacher predictions in a frozen visual feature space.
- We instantiate the unified endpoint objective for image, bidirectional video, and causal video generation. Experiments show strong and stable few-step quality across these settings, and ablations isolate the effects of on-policy queries, interval endpoints, and the proposed refinements.

2 Related Work

Trajectory and consistency distillation. Progressive distillation shortens samplers (Salimans & Ho, 2022). Consistency models align predictions across noise levels (Song et al., 2023; Luo et al., 2023), while MeanFlow, shortcut models, and continuous-time methods learn interval transitions (Geng et al., 2025; Frans et al., 2025; Lu & Song, 2025). These methods offer dense per-sample regression but usually query data or teacher samples rather than the current student rollout. Such

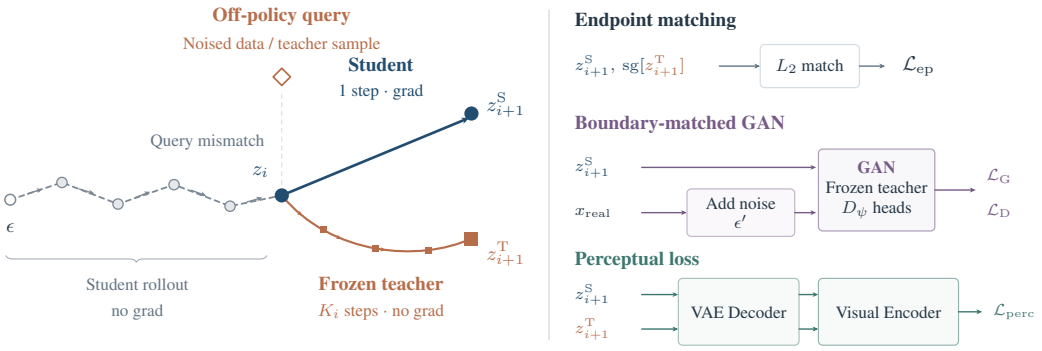

Figure 2: At a detached student query state z_i, one coarse student step is matched to the endpoint of K_i frozen-teacher steps. The endpoint is the only teacher target. The optional GAN evaluates the same student endpoint against a same-condition real latent at the identical next-boundary noise level.

queries need not lie on the deployed few-step path. OPTD keeps paired regression, queries the student trajectory, and uses teacher fine updates to construct the next endpoint.

Distribution matching and adversarial distillation. DMD and DMD2 use auxiliary models to estimate student-side score information (Yin et al., 2024b;a), while TDM and related methods match trajectory distributions (Luo et al., 2025). ADD, LADD, SDXL-Lightning, Diffusion2GAN, and ADM use adversarial feedback to improve few-step detail (Sauer et al., 2024b;a; Lin et al., 2024; Kang et al., 2024; Lu et al., 2025). These methods make feedback on student-generated samples possible, but their auxiliary optimization replaces a paired teacher target or adds a separate distributional constraint. The base OPTD objective retains the paired endpoint and therefore keeps direct per-sample pressure. Its adversarial refinement has a different contract from a generic clean-image GAN. It evaluates the student's next-boundary latent at the schedule-defined noise level, pairs the real sample with the same condition, and uses the frozen teacher as the feature backbone. This boundary match adds realism feedback without replacing endpoint regression.

On-policy distillation. On-policy distillation supervises states produced by the current student, reducing the mismatch between training queries and deployed behavior (Agarwal et al., 2024). This principle has also been explored in autoregressive video generation, where student-generated contexts are used to better align training with causal inference (Huang et al., 2025; Bandyopadhyay et al., 2026). For flow models, Flow-OPD, FlowCTS, and AnyFlow similarly leverage student-visited states for post-training, trajectory supervision, or few-step generation (Fang et al., 2026; Ye et al., 2026; Gu et al., 2026). OPTD shares this on-policy perspective but adopts a boundary-aligned endpoint objective specifically for few-step compression. From the same student-visited state, the student takes one coarse transition while the frozen teacher integrates the full corresponding interval using finer updates. Only the teacher's terminal state is used as the paired target, while its intermediate states serve only to construct that endpoint. This preserves dense per-sample supervision without matching intermediate trajectories or learning an arbitrary-time flow map.

3 Method

OPTD is built around one contract. The student visits a state, the teacher integrates the student's next interval from that same state, and the student matches the resulting endpoint. This contract pairs the inference state with the transition that the deployed student must execute. Figure 2 summarizes the construction.

3.1 Diffusion Model Setup

Let $x \sim p_{\mathrm{data}}(x \mid c)$ be a clean latent under condition c, and let $\epsilon \sim \mathcal{N}(0, I)$. We use the linear flow path in Eq. 1,

$$z_\sigma = (1-\sigma)x + \sigma\epsilon, \qquad \sigma \in [0,1], \tag{1}$$

where sampling proceeds from noise at $\sigma = 1$ to data at $\sigma = 0$. A conditional velocity field $v(z_\sigma, \sigma, c)$ defines the ordinary differential equation $\mathrm{d}z_\sigma/\mathrm{d}\sigma = v(z_\sigma, \sigma, c)$. For two decreasing noise levels $0 \le \sigma' < \sigma$, one Euler transition is given by Eq. 2,

$$\mathrm{Euler}(z, \sigma, \sigma', v, c) = z + (\sigma' - \sigma)v(z, \sigma, c). \tag{2}$$

The frozen teacher has velocity field v_ϕ. Its classifier-free guided field is defined in Eq. 3,

$$\tilde{v}_\phi(z, \sigma, c) = v_\phi(z, \sigma, \varnothing) + w\big(v_\phi(z, \sigma, c) - v_\phi(z, \sigma, \varnothing)\big), \tag{3}$$

where w is the teacher guidance scale. The deployed student uses one conditional evaluation at each boundary and is denoted by v_θ. We write $\Phi_\phi(z, \sigma, \sigma', c)$ for one fine teacher-solver update driven by $\tilde{v}_\phi$.

3.2 On-Policy Endpoint Distillation

The teacher schedule is $1 = \sigma^{\mathrm{t}}_0 > \sigma^{\mathrm{t}}_1 > \cdots > \sigma^{\mathrm{t}}_{N_{\mathrm{t}}} = 0$. Student boundaries form a strict subsequence. Let $0 = b_0 < b_1 < \cdots < b_{N_{\mathrm{s}}} = N_{\mathrm{t}}$ and define the coupled schedule in Eq. 4,

$$\sigma^{\mathrm{s}}_i = \sigma^{\mathrm{t}}_{b_i}. \tag{4}$$

The teacher takes $K_i = b_{i+1} - b_i$ fine updates inside student interval i.

3.2.1 Student-visited query and endpoint target

We sample a condition c, initial noise ϵ, and an interval index $i \sim \mathcal{U}\{0, \ldots, N_{\mathrm{s}} - 1\}$. Starting from $\bar{z}_0 = \epsilon$, the current student is rolled out without gradients to boundary i according to Eq. 5,

$$\bar{z}_{j+1} = \mathrm{Euler}(\bar{z}_j, \sigma^{\mathrm{s}}_j, \sigma^{\mathrm{s}}_{j+1}, v_\theta, c), \qquad 0 \le j < i. \tag{5}$$

The query is $z_i = \mathrm{sg}[\bar{z}_i]$. Rollout and deployment use the same boundary grid and Euler rule, which makes z_i an inference state of the current student rather than an independently constructed noisy state.

With gradients, the student takes one transition given by Eq. 6,

$$z^{\mathrm{S}}_{i+1} = z_i + (\sigma^{\mathrm{s}}_{i+1} - \sigma^{\mathrm{s}}_i)v_\theta(z_i, \sigma^{\mathrm{s}}_i, c). \tag{6}$$

Starting from exactly the same z_i, the teacher integrates the fine schedule as in Eq. 7,

$$u^{\mathrm{T}}_{k+1} = \Phi_\phi(u^{\mathrm{T}}_k, \sigma^{\mathrm{t}}_k, \sigma^{\mathrm{t}}_{k+1}, c), \qquad k = b_i, \ldots, b_{i+1} - 1, \tag{7}$$

with $u^{\mathrm{T}}_{b_i} = z_i$. The terminal state $z^{\mathrm{T}}_{i+1} = u^{\mathrm{T}}_{b_{i+1}}$ is the only teacher target. The teacher's intermediate states are not separate supervision targets. The endpoint loss is given by Eq. 8,

$$\mathcal{L}_{\mathrm{ep}} = \mathbb{E}_{c,\epsilon,i}\left[\left\| z^{\mathrm{S}}_{i+1} - \mathrm{sg}\big[z^{\mathrm{T}}_{i+1}\big]\right\|^2_2\right]. \tag{8}$$

For $\Delta_i = \sigma^{\mathrm{s}}_{i+1} - \sigma^{\mathrm{s}}_i$, define the teacher interval secant $\bar{v}^{\mathrm{T}}_i = (z^{\mathrm{T}}_{i+1} - z_i)/\Delta_i$. Then, as shown in Eq. 9,

$$\left\| z^{\mathrm{S}}_{i+1} - \mathrm{sg}\big[z^{\mathrm{T}}_{i+1}\big]\right\|^2_2 = \Delta^2_i \left\| v_\theta(z_i, \sigma^{\mathrm{s}}_i, c) - \mathrm{sg}\big[\bar{v}^{\mathrm{T}}_i\big]\right\|^2_2. \tag{9}$$

Thus endpoint matching is a full-interval secant regression with the expected step-size weight. The innovation is the coupling of the query state to that interval target. Uniform boundary sampling gives each student interval direct paired supervision. Detaching the prefix confines gradients to the selected transition. The base objective therefore uses one trainable network and one optimizer, without a student-side score or learned critic.

The complete parameter update is specified in Algorithm 1 in Appendix A. It makes explicit that the rollout prefix and teacher endpoint are detached, while gradients pass only through the selected student transition. When either refinement is enabled, Algorithm 2 constructs the same-boundary real input, performs the separate discriminator update, and returns the student-side adversarial and perceptual terms. Thus the algorithms use the same objects as Eqs. 8–13 rather than introducing an additional training target.

3.2.2 BOUNDARY-MATCHED ADVERSARIAL REFINEMENT

Endpoint regression supplies a paired conditional target. It does not by itself judge whether the student endpoint has the local statistics of the data marginal. We add an adversarial term at precisely the same object rather than moving the discriminator to a clean estimate. Let z^{S}_{i+1} be the student endpoint and let x_{real} be the real latent paired with condition c. The real input is forward-noised to the student's next boundary according to Eq. 10,

$$r_{i+1} = (1 - \sigma^{\mathrm{s}}_{i+1})x_{\mathrm{real}} + \sigma^{\mathrm{s}}_{i+1}\epsilon', \qquad \epsilon' \sim \mathcal{N}(0, I). \tag{10}$$

The fake is z^{S}_{i+1} itself. Both inputs therefore share the condition and noise level. The discriminator never receives a clean conditional mean at an arbitrary noise level. The adversarial gradient therefore acts on the transition that the sampler actually executes, rather than on a separate auxiliary prediction. The complete discriminator and generator update is given in Algorithm 2.

A frozen teacher transformer extracts multiscale features from both inputs. We write $\mathcal{F}_\phi$ for this frozen feature map, and trainable heads D_ψ map its features to logits. With non-saturating softplus losses in Eqs. 11 and 12,

$$\mathcal{L}_{\mathrm{G}} = \mathbb{E}\left[\mathrm{softplus}\big(-D_\psi(z^{\mathrm{S}}_{i+1}, \sigma^{\mathrm{s}}_{i+1}, c)\big)\right], \tag{11}$$

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}\left[\mathrm{softplus}\big(D_\psi(\mathrm{sg}\big[z^{\mathrm{S}}_{i+1}\big], \sigma^{\mathrm{s}}_{i+1}, c)\big) + \mathrm{softplus}\big(-D_\psi(r_{i+1}, \sigma^{\mathrm{s}}_{i+1}, c)\big)\right]. \tag{12}$$

The complete student objective is given by Eq. 13,

$$\mathcal{L} = \mathcal{L}_{\mathrm{ep}} + \alpha_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}} + \alpha_{\mathrm{adv}}\mathcal{L}_{\mathrm{G}}, \qquad \alpha_{\mathrm{perc}}, \alpha_{\mathrm{adv}} \geq 0. \tag{13}$$

The discriminator heads use $\mathcal{L}_{\mathrm{D}}$ with a separate optimizer, and the teacher backbone is frozen in both turns. Setting both auxiliary weights to zero recovers base OPTD. The GAN is consequently an optional refinement with an extra trainable head and optimizer, whereas the base method remains a one-network method.

3.2.3 OPTIONAL PERCEPTUAL REFINEMENT

A second optional refinement compares clean predictions in a frozen feature space. The student clean estimate from the same forward pass is $\hat{x}^{\mathrm{S}}_{0,i} = z_i - \sigma^{\mathrm{s}}_i v_\theta(z_i, \sigma^{\mathrm{s}}_i, c)$. The teacher continues from z_i to $\sigma = 0$ to obtain $x^{\mathrm{T}}_{0,i}$. If $\mathcal{V}$ is a frozen latent decoder and F a frozen visual encoder, a token-level cosine loss is

$$\mathcal{L}_{\mathrm{perc}} = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left(1 - \frac{\langle f^{\mathrm{S}}_{i,m}, \mathrm{sg}\big[f^{\mathrm{T}}_{i,m}\big]\rangle}{\|f^{\mathrm{S}}_{i,m}\|_2 \|f^{\mathrm{T}}_{i,m}\|_2}\right)\right]. \tag{14}$$

The perceptual loss in Eq. 14 compares $f^{\mathrm{S}}_{i,m} = F_m(\mathcal{V}(\hat{x}^{\mathrm{S}}_{0,i}))$ and $f^{\mathrm{T}}_{i,m} = F_m(\mathcal{V}(x^{\mathrm{T}}_{0,i}))$. Images use DINOv3 features, whereas videos use the frozen InternVideo3 vision tower (Yan et al., 2026). The perceptual term judges clean semantic and local temporal structure. It is distinct from both the endpoint loss and the boundary GAN, and it adds teacher rollout, decoding, and feature-extractor cost. Its clean-prediction branch is included in Algorithm 2.

3.3 MODALITY-SPECIFIC DISTILLATION INSTANTIATIONS

Image and bidirectional video. Images use the complete spatial latent. Bidirectional videos based on Wan (Team Wan et al., 2025) jointly denoise all latent frames at a shared noise level with full temporal attention. Both modes use the detached student rollout, one student transition, and one teacher endpoint described above. The optional image perceptual branch uses DINOv3 patch features. The video branch uses InternVideo3 tubelet features, which mix adjacent frames and provide local temporal sensitivity without claiming global temporal reasoning.

Causal video. Causal generation applies the endpoint contract independently to autoregressive chunks, following cache-based teacher forcing for autoregressive video diffusion (Huang et al., 2025; Bandyopadhyay et al., 2026). The student and teacher fields therefore condition on the cached history. Let $h_{<q}$ denote the clean latent chunks preceding chunk q, and let $z_{q,i}$ be the noisy state of that chunk at student boundary i. During the detached rollout, the student advances chunks in order and stores the committed clean context in a KV cache,

$$h_{<q} = \mathrm{Cache}(h_{<q-1}, \hat{x}^{\mathrm{S}}_{q-1,0}), \qquad \hat{x}^{\mathrm{S}}_{q,0} = z_{q,i} - \sigma^{\mathrm{s}}_i v_\theta(z_{q,i}, \sigma^{\mathrm{s}}_i, c, h_{<q}). \tag{15}$$

The cache contains only completed chunks, so future chunks are masked. For the selected chunk and boundary, the student endpoint is

$$z^{\mathrm{S}}_{q,i+1} = z_{q,i} + (\sigma^{\mathrm{s}}_{i+1} - \sigma^{\mathrm{s}}_{i}) v_\theta(z_{q,i}, \sigma^{\mathrm{s}}_{i}, c, h_{<q}), \tag{16}$$

where the history $h_{<q}$ is detached. The teacher receives the same query and history and integrates the fine schedule across that interval,

$$u^{\mathrm{T}}_{q,k+1} = \Phi_\phi(u^{\mathrm{T}}_{q,k}, \sigma^{\mathrm{t}}_{k}, \sigma^{\mathrm{t}}_{k+1}, c, h_{<q}), \quad u^{\mathrm{T}}_{q,b_i} = z_{q,i}, \quad z^{\mathrm{T}}_{q,i+1} = u^{\mathrm{T}}_{q,b_{i+1}}. \tag{17}$$

The causal endpoint loss is therefore

$$\mathcal{L}^{\mathrm{causal}}_{\mathrm{ep}} = \mathbb{E}_{q,i,c,\epsilon}\left[\left\| z^{\mathrm{S}}_{q,i+1} - \mathrm{sg}\left[z^{\mathrm{T}}_{q,i+1}\right]\right\|_2^2\right]. \tag{18}$$

Equations 15–18 define the cache state, student endpoint in Eq. 16, teacher endpoint in Eq. 17, and causal loss. In implementation, a block-causal teacher-forcing pass recomputes the selected student transition with gradients while reusing the detached cache. Thus each query sees its noisy chunk and the allowed preceding clean chunks, and no gradient crosses the rollout or cached history.

4 Experiments

We evaluate OPTD for image distillation, bidirectional video distillation, and chunk-causal video distillation. The experiments isolate the contributions of endpoint matching, perceptual supervision, and boundary-matched adversarial supervision. Additional configurations and results appear in Sections B and C. Representative qualitative examples are shown in Figure 3, with the complete multi-prompt comparison in Figure 7 and additional OPTD samples in Figure 8. The bidirectional video comparison is shown in Figure 4 and quantified in Table 2. The causal comparison is shown in Figure 5. Limitations are discussed in Appendix D.

4.1 Setup

Training setting. For image distillation, we initialize the student and frozen teacher from FLUX.2-klein-base-4B (Black Forest Labs, 2026) at 1024×1024 resolution. The frozen teacher uses 28 Euler steps with classifier-free guidance (Ho & Salimans, 2022) of 4.0, and the student uses four steps. We train on LAION captions (Schuhmann et al., 2022) with AdamW (Loshchilov & Hutter, 2019) at 5×10^{-7} for 5,000 student updates. Perceptual supervision uses frozen DINOv3 features (Siméoni et al., 2025). For the bidirectional video GAN variant, each generator update is followed by four discriminator-head updates. Complete training configurations for image, bidirectional video, and causal video distillation are given in Section B.

Baselines. For images, we compare with the pretrained SD3.5-M and SD3.5-L models from the SD3 family (Esser et al., 2024), as well as the pretrained FLUX.2-klein-base-4B model (Black Forest Labs, 2026). We also include the open distilled models SD3.5-L-Turbo and FLUX.2-klein-4B (Black Forest Labs, 2026). Method baselines built on FLUX.2 include LCM (Luo et al., 2023), DMD2 (Yin et al., 2024a), PFM (Zhao et al., 2026), TDM (Luo et al., 2025), and rCM (Zheng et al., 2026). For video, we compare pretrained Wan2.1-T2V (Team Wan et al., 2025), FastWan2.1-T2V-1.3B (Zhang et al., 2025a;b), T2V-Turbo (Li et al., 2025), AnyFlow (Gu et al., 2026), and Wan2.1-T2V variants trained with DMD2 (Yin et al., 2024a) and rCM (Zheng et al., 2026). The table also includes our base, perceptual, and GAN-refined OPTD variants. For causal video, we compare the Wan2.2-TI2V-5B OPTD model with Self-Forcing (Huang et al., 2025) under the same causal evaluation protocol.

Metrics. We assess image distributional fidelity with Fréchet Inception Distance (FID), where lower values indicate closer agreement between generated and reference image distributions (Heusel et al., 2017). We measure prompt alignment with CLIP Score (Radford et al., 2021) and human-preference-aligned quality with ImageReward (Xu et al., 2023). GenEval (Ghosh et al., 2023) measures compositional instruction following through its Overall, Single Object, Two Object, Counting, Colors, Position, and Color Attribution scores. Mean pairwise LPIPS (Zhang et al., 2018) measures sample diversity. For videos, VBench (Huang et al., 2024) evaluates visual quality, semantic alignment, temporal consistency, motion, and compositional content across all 16 dimensions. We report its Quality, Semantic, and Total scores, together with the unweighted mean of the 16 dimension scores. Higher values are better for all metrics except FID.

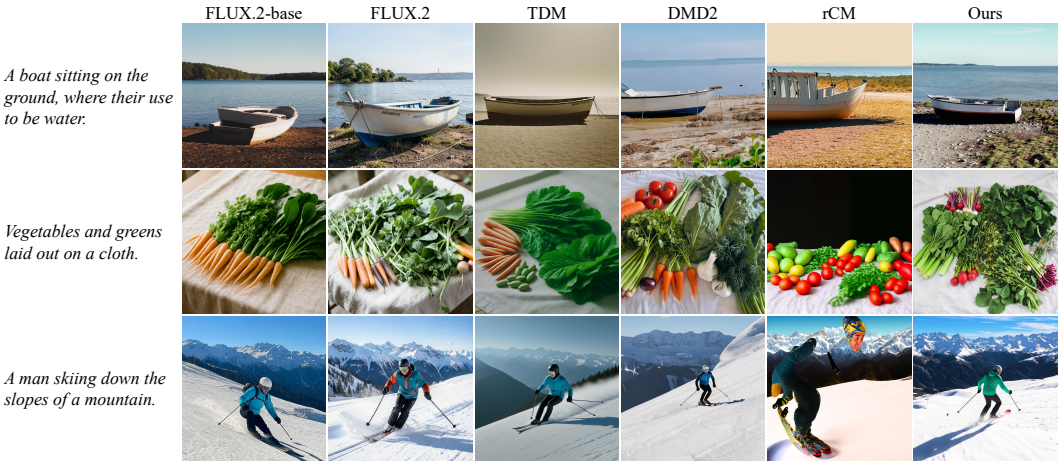

Figure 3: **Qualitative image comparison.** FLUX.2-klein-base uses 50 sampling steps.

Table 1: **Image generation results.** NFE denotes the number of function evaluations. GenEval reports the Overall score; component scores are given in Table 3. Red and orange backgrounds mark the best and second-best values among the Distillation Methods, respectively. After excluding colored cells, bold marks the best remaining value in each full column. A red cell is bold when it is also the full-column best.

Method	NFE	FID ↓	CLIP ↑	IR ↑	LPIPS ↑	GenEval ↑
Pretrained Models						
SD3.5-M	50	20.83	0.3423	1.0137	0.7292	0.700
SD3.5-L	50	19.99	0.3388	1.0722	0.7233	0.710
FLUX.2-klein-base-4B	50	21.03	0.3369	0.9492	**0.7436**	0.760
Released Distilled Models						
SD3.5-L-Turbo	4	26.26	0.3399	0.9524	0.6745	0.700
FLUX.2-klein-4B	4	22.44	**0.3427**	**1.1096**	0.6860	**0.820**
Distillation Methods						
LCM	4	32.51	0.3261	0.4669	0.6709	0.680
DMD2	4	18.91	0.3322	0.7767	0.6940	0.690
PFM	4	49.09	0.2771	0.4676	0.4322	0.370
TDM	4	31.31	0.3292	0.7706	0.6754	0.740
rCM	4	32.52	0.3160	0.6005	0.7261	0.589
OPTD	4	21.95	0.3363	0.8349	0.7152	0.745
OPTD+perceptual	4	23.62	0.3420	1.0093	0.6913	0.750
OPTD+GAN	4	**15.06**	0.3360	0.7221	0.7053	0.780
OPTD+GAN+perceptual	4	19.23	0.3369	0.8651	0.6982	0.791
OPTD ($4 \rightarrow 2$)	2	**16.86**	0.3361	0.6257	0.6952	0.756
OPTD ($2 \rightarrow 1$)	1	16.82	0.3356	0.5922	0.6984	0.749

4.2 Main Results

Image results. Figure 3 presents representative qualitative comparisons, with the complete grid provided in Figure 7. Across the boat, vegetable, skiing, and bedroom prompts, OPTD preserves the requested content, coherent composition, and local detail. Compared with the other distilled models, its outputs show fewer severe crops, missing objects, and texture artifacts, while maintaining the scene structure across diverse prompts. Table 1 summarizes the quantitative comparison. Base OPTD improves CLIP, ImageReward, and GenEval over the non-OPTD distillation baselines while retaining competitive LPIPS diversity. The refinement variants separate the roles of the auxiliary objectives. OPTD+perceptual gives the strongest alignment and preference scores within the method

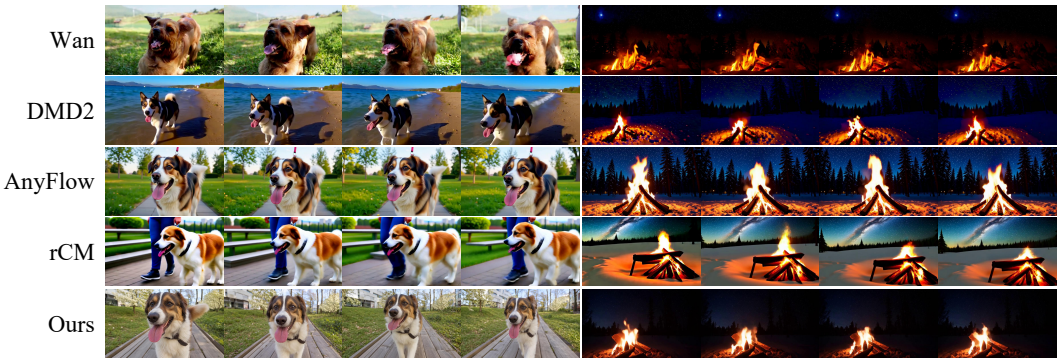

Figure 4: **Qualitative bidirectional video comparison.** The clips compare four-step Wan2.1-T2V-1.3B distillation methods at 832×480 resolution over 81 frames. Distribution-matching methods such as DMD2 tend to produce a narrow, mode-seeking appearance with oversaturated content, whereas Ours preserves more natural colors, textures, and motion in the shown examples.

Table 2: **Bidirectional video generation results.** Scores are from VBench. Total is VBench's weighted score, while Mean is the unweighted mean of its 16 dimensions. Red backgrounds mark the best value among the Distillation Methods. Bold marks the best value in each full column. A red cell is bold when it is also the full-column best.

Method	Steps	Resolution	Frames	Total ↑	Quality ↑	Semantic ↑	Mean ↑
Pretrained Models							
Wan2.1-T2V-1.3B	50	832×480	81	75.31	79.57	58.27	0.5844
Released Distilled Models							
FastWan2.1-T2V-1.3B	3	832×480	81	80.76	84.34	66.47	0.6501
T2V-Turbo-v2	16	512×320	16	**85.54**	**85.70**	**84.92**	**0.7382**
AnyFlow	4	832×480	81	83.82	85.21	78.24	0.7036
Distillation Methods							
DMD2	4	832×480	81	81.15	84.15	69.15	0.6804
rCM	4	832×480	81	79.09	82.38	65.95	0.6386
OPTD	4	832×480	81	72.82	77.46	54.23	0.5574
OPTD+perceptual	4	832×480	81	74.76	79.09	57.43	0.5803
OPTD+GAN	4	832×480	81	82.64	85.44	71.45	0.6837

group, whereas OPTD+GAN gives the lowest FID in the full table. Combining both refinements produces the strongest GenEval score. The two-step and one-step continuations retain lower FID than four-step DMD2, showing that endpoint supervision remains useful as the sampling budget is reduced.

Bidirectional video results. Figure 4 provides a qualitative comparison under the common 81-frame, 832×480 setting. DMD2 and similar distribution-matching methods often exhibit an oversaturated, mode-seeking appearance, resulting in less natural scenes. In contrast, our method preserves more plausible colors and textures with smoother temporal evolution. Consistent quantitative improvements are reported in Table 2. Among the distillation methods, OPTD+GAN achieves the best results across all four summary metrics, improving over base OPTD by 13.5% in Total and 31.8% in Semantic score. It also slightly surpasses the released AnyFlow model in Quality, while AnyFlow remains stronger in semantic alignment. The detailed metrics further show broad gains across motion, appearance, and action dimensions, supporting the boundary-matched GAN as an effective realism refinement complementary to the paired endpoint objective.

Causal video results. Figure 5 compares the causal Wan2.2-TI2V-5B setting with Self-Forcing (Huang et al., 2025). Self-Forcing exhibits visible residual exposure bias. The generated subject changes shape across frames, and viewpoint transitions introduce abrupt distortions and in-

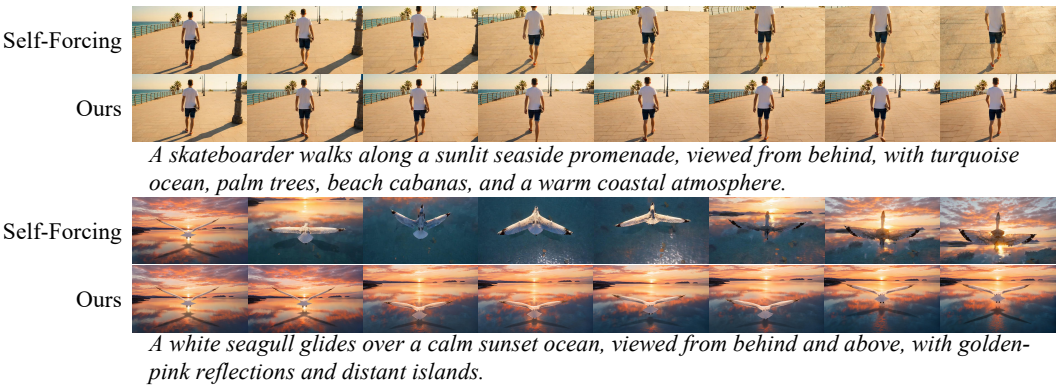

Figure 5: **Qualitative causal video comparison.** The rows show sampled frames for a skateboarder walking along a seaside promenade and a seagull gliding over the ocean.

consistent geometry. OPTD keeps the subject identity and scene structure stable while the camera viewpoint changes, with smoother appearance transitions and more coherent motion in both examples. These results are consistent with the endpoint contract for causal generation. Each selected chunk is trained from the student's own cached history, while the teacher endpoint supplies a paired target for the same chunk and interval.

4.3 Ablations and analysis

We analyze the effects of the loss components and sampling budget using the quantitative results in Tables 1 and 3.

Loss components. We compare OPTD with GAN, perceptual, and combined refinement. The qualitative ablation in Figure 6 and the quantitative results in Tables 1 and 3 show complementary effects. GAN refinement produces more realistic textures and backgrounds, improving distributional fidelity and compositional scores as shown by its lower FID and higher GenEval. Perceptual refinement produces cleaner and more visually pleasing details, leading to stronger CLIP and ImageReward scores. The two objectives can be combined with tunable relative weights to trade off realism and perceptual quality. In our setting, the combined model obtains the highest GenEval score and is especially strong on counting and positioning. The video ablation shows a complementary pattern. When the base model is weak, perceptual supervision yields only marginal gains over base OPTD, whereas the boundary-matched GAN leads to a substantial recovery in video quality. We attribute this improvement to the GAN's ability to encourage the student to match local statistics of real data, while the endpoint loss constrains the generated trajectory to remain consistent with the teacher's conditional dynamics.

Fewer sampling steps. We continue distilling OPTD from four to two and then one sampling step. Both compressed models outperformed four-step DMD2 in distributional fidelity and composition, with comparable diversity but lower preference alignment. Four-step OPTD remained stronger in preference alignment. One-step and two-step OPTD had similar FID and diversity, while two-step OPTD was better in composition and preference alignment.

5 Conclusion

We presented OPTD, a unified framework for few-step diffusion distillation that matches each student transition to a frozen teacher endpoint computed from the same student-visited state and over the same sampling interval. This on-policy pairing provides direct supervision for the transitions used at inference time. Boundary-matched adversarial and perceptual refinements further improve distributional fidelity and visual quality while preserving the paired endpoint objective. Overall, OPTD generates realistic and semantically faithful images and videos in few steps while maintaining sample diversity. We hope this framework provides a useful foundation for future work on efficient diffusion distillation across different generation settings.

AI USE STATEMENT

We used generative AI tools solely to assist with language editing, organization, terminology consistency, and proofreading of the manuscript and its LaTeX source. These tools were used only to improve the presentation of ideas and technical content already developed by the authors. They were not used to generate or analyze research data, design the methodology or experiments, implement or run experiments, interpret experimental results, or generate scientific claims, mathematical content, or citations. The generative models developed and evaluated as part of this work constitute the subject of the research and are described in the corresponding methodology and experimental sections; their use is distinct from the AI-assisted tools disclosed here. The authors reviewed all AI-assisted edits and take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies the efficiency and fidelity of generative-model distillation. We use standard research data and pretrained models. The released code and evaluation protocol will follow the licenses and usage requirements of the underlying models and datasets.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide detailed descriptions of the proposed method, model architecture, training and evaluation procedures, hyperparameter settings, and experimental protocols. To further facilitate independent reproduction and verification of our results, we will publicly release the complete source code, including training and evaluation scripts and configuration files, together with all trained model checkpoints used in our experiments. We will also provide instructions and dependencies required to reproduce the reported results. All released materials will be made publicly available upon publication.

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.

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-matched distillation: Teacher causality for autoregressive video distillation. *arXiv preprint arXiv:2608.13391*, 2026.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. `https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence`, 2026. Base model weights at `https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B`.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41667–41676, 2026.

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

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. Anyflow: Any-step video diffusion model with on-policy flow map distillation. *arXiv preprint arXiv:2605.13724*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, 2024.

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.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 2025.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829. IEEE, 2025.

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 *IEEE/CVF International Conference on Computer Vision*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

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

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8428–8437. IEEE, 2025.

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. In *Advances in Neural Information Processing Systems*, 2023.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. InternVideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 2026. URL https://arxiv.org/abs/2606.12195.

Kaiyang Ye, Yuan Ge, Junxiang Zhang, Bei Li, Ziming Zhu, Haishu Zhao, Xiaoqian Liu, Chenglong Wang, Jingbo Zhu, Zhengtao Yu, and Tong Xiao. FlowCTS: On-policy continuous trajectory supervision of flow models. *arXiv preprint arXiv:2607.24522*, 2026.

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 *Advances in 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. *arXiv preprint arXiv:2505.13389*, 2025a.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Chuyang Zhao, Yifei Song, Hongfa Wang, Jianlong Yuan, Yuan Zhang, Siming Fu, Zhineng Chen, Huilin Deng, Haoyang Huang, and Nan Duan. Perceptual flow matching for few-step generative modeling. *arXiv preprint arXiv:2607.03524*, 2026. URL https://arxiv.org/abs/2607.03524.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *International Conference on Learning Representations*, 2026.

A Detailed training algorithms

Algorithm 1 gives one endpoint-distillation update. It constructs the query by replaying the deployed student sampler, integrates the frozen teacher only over the selected interval, and keeps the prefix and teacher target out of the gradient graph. Algorithm 2 expands the optional boundary-matched GAN and perceptual branches. The discriminator and student optimizers are separate, and the teacher feature backbone remains frozen in both updates.

B Implementation details

The image implementation uses the FLUX OPD systems and the video implementation uses the trajectory OPD recipe. Both detach the rollout prefix and compute the teacher endpoint with frozen guided updates. When enabled, the perceptual decoder and feature encoder are frozen. The GAN uses softplus generator and discriminator losses, a frozen teacher feature backbone, same-condition real data, and separate optimizer state for the trainable heads.

Image distillation. We initialize the student and frozen teacher from FLUX.2-klein-base-4B (Black Forest Labs, 2026) and train at 1024×1024 resolution. Training prompts come from the LAION caption subset (Schuhmann et al., 2022) distributed with DMD2 (Yin et al., 2024a). Adversarial training additionally uses precomputed latents of the corresponding real images. The teacher uses 28 Euler steps with classifier-free guidance (Ho & Salimans, 2022) of 4.0. The four-step student takes every seventh boundary of the teacher schedule and uses a single conditional model evaluation per step. We optimize the student with AdamW (Loshchilov & Hutter, 2019) at a learning rate of 5×10^{-7}, with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 0.01. The learning rate increases linearly over the first 500 training iterations and remains constant thereafter. We use a batch size of four images per GPU, one microbatch per update, BF16 mixed precision, gradient checkpointing, and a student gradient-norm limit of 1.0. Each distillation stage comprises 5,000 student updates. The endpoint loss has unit weight. Perceptual supervision uses cosine distance between normalized final-layer features of frozen DINOv3 ViT-H/16+ (Siméoni et al., 2025), with 224×224 encoder inputs and $\alpha_{\text{perc}} = 1.0$. For adversarial training, we set $\alpha_{\text{adv}} = 0.03$ and optimize the discriminator heads with AdamW at 10^{-5}, using the same weight decay and warmup as the student and a gradient-norm limit of 10.0. Both optimizers use $(\beta_1, \beta_2) = (0, 0.999)$ in GAN variants. For image training, generator and discriminator updates alternate at a 1:1 ratio, giving 10,000 training iterations with 5,000 updates to each network. We further distill the four-step OPTD

Algorithm 1. OPTD endpoint-distillation update

Input: Student parameters θ, frozen teacher parameters ϕ, teacher schedule $\{\sigma_k^{\mathrm{t}}\}_{k=0}^{N_{\mathrm{t}}}$, student boundaries $0 = b_0 < \cdots < b_{N_{\mathrm{s}}} = N_{\mathrm{t}}$, condition distribution, and optional weights $\alpha_{\mathrm{adv}}, \alpha_{\mathrm{perc}}$
Output: Updated student parameters θ

Sample condition c, initial noise $\epsilon \sim \mathcal{N}(0, I)$, and $i \sim \mathcal{U}\{0, \ldots, N_{\mathrm{s}} - 1\}$
if $\alpha_{\mathrm{adv}} > 0$ **then**
  Sample the paired clean latent x_{real} for condition c and $\epsilon' \sim \mathcal{N}(0, I)$
Set $\bar{z}_0 \leftarrow \epsilon$
for $j = 0, \ldots, i-1$ **do**
  $\bar{z}_{j+1} \leftarrow \mathrm{Euler}(\bar{z}_j, \sigma_j^{\mathrm{s}}, \sigma_{j+1}^{\mathrm{s}}, v_\theta, c)$ `// Replay the inference-time student prefix`
$z_i \leftarrow \mathrm{sg}[\bar{z}_i]$ `// No gradient through the query state`
$z_{i+1}^{\mathrm{S}} \leftarrow z_i + (\sigma_{i+1}^{\mathrm{s}} - \sigma_i^{\mathrm{s}}) v_\theta(z_i, \sigma_i^{\mathrm{s}}, c)$ `// One gradient-carrying student transition`
Set $u_{b_i}^{\mathrm{T}} \leftarrow z_i$
for $k = b_i, \ldots, b_{i+1} - 1$ **do**
  $u_{k+1}^{\mathrm{T}} \leftarrow \Phi_\phi(u_k^{\mathrm{T}}, \sigma_k^{\mathrm{t}}, \sigma_{k+1}^{\mathrm{t}}, c)$
`// Integrate the complete teacher interval`
$z_{i+1}^{\mathrm{T}} \leftarrow u_{b_{i+1}}^{\mathrm{T}}$ `// Use only the terminal teacher state as target`
$\mathcal{L}_{\mathrm{ep}} \leftarrow \| z_{i+1}^{\mathrm{S}} - \mathrm{sg}[z_{i+1}^{\mathrm{T}}] \|_2^2$
if $\alpha_{\mathrm{adv}} > 0$ ***or*** $\alpha_{\mathrm{perc}} > 0$ **then**
  $(\mathcal{L}_{\mathrm{G}}, \mathcal{L}_{\mathrm{perc}}) \leftarrow \textsc{AuxiliaryUpdate}(z_{i+1}^{\mathrm{S}}, z_i, c, i, x_{\mathrm{real}}, \epsilon'; \theta, \psi, \phi)$ `// Algorithm 2`
else
  $\mathcal{L}_{\mathrm{G}} \leftarrow 0, \mathcal{L}_{\mathrm{perc}} \leftarrow 0$
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{ep}} + \alpha_{\mathrm{adv}} \mathcal{L}_{\mathrm{G}} + \alpha_{\mathrm{perc}} \mathcal{L}_{\mathrm{perc}}$
$\theta \leftarrow \mathrm{AdamWUpdate}(\theta, \nabla_\theta \mathcal{L})$ `// Only the selected student transition receives gradient`
return θ

model into a two-step model and then a one-step model. Each stage initializes the student and teacher from the preceding stage, freezes the teacher, and uses teacher guidance 1.0 because guidance has already been distilled. The "Ours" examples in Figure 3 use the OPTD+GAN+perceptual setting.

Bidirectional video distillation. We use Wan2.1-T2V-1.3B (Team Wan et al., 2025) and train on a quality-balanced subset of Koala-36M (Wang et al., 2025). We randomly sample 81-frame clips at 16 frames per second and 480×832 resolution, retaining clips with an optical-flow magnitude of at least 1.0 and no detected watermarks. The teacher uses 48 Euler steps with guidance 5.0, while the student uses four steps with guidance 1.0. Both use a shifted linear noise schedule with shift 5.0, and the student takes every twelfth teacher boundary. The student learning rate is 10^{-6} throughout training without warmup, and the batch size is one video per GPU. AdamW parameters, weight decay, gradient accumulation, numerical precision, gradient clipping, and the budget of 5,000 student updates follow the image setting. Perceptual supervision uses the frozen InternVideo3 vision tower (Yan et al., 2026), normalized final-layer features, cosine distance, and the same perceptual weight. Its clean target is obtained with a 12-step teacher rollout. The GAN uses the image adversarial weight, optimizer moments, discriminator learning rate, with constant learning rates for both optimizers. We update the generator once for every four discriminator-head updates. Discriminator heads extract features from teacher layers 15, 22, and 29. The visual comparison in Figure 4 uses the OPTD+GAN setting.

Causal video distillation. We initialize the student and frozen teacher from a Wan2.2-TI2V-5B checkpoint (Team Wan et al., 2025). We train on 121-frame clips at 24 frames per second and 480×832 resolution. Each clip contains one conditioning latent frame followed by ten generated chunks of three latent frames each. Rollout proceeds causally and stores preceding chunks in a KV cache. We use endpoint matching with the same 48-step teacher, four-step student, guidance scales, and noise schedule as in bidirectional video distillation. Training also uses a constant learning rate of 10^{-5}, one video per GPU, and 5,000 student updates, with the remaining optimizer and numerical settings unchanged. The causal comparison in Figure 5 uses the OPTD+GAN setting as well.

Algorithm 2. Boundary-matched auxiliary-loss update

Input: Student endpoint z^{S}_{i+1}, detached query z_i, condition c, selected interval i with $\sigma^{\mathrm{s}}_i = \sigma^{\mathrm{t}}_{b_i}$ and $\sigma^{\mathrm{s}}_{i+1} = \sigma^{\mathrm{t}}_{b_{i+1}}$, optional paired clean latent x_{real} and $\epsilon' \sim \mathcal{N}(0, I)$ when $\alpha_{\mathrm{adv}} > 0$, student parameters θ, discriminator-head parameters ψ, and frozen teacher/feature modules $\phi, \mathcal{V}, F$

Output: Student-side losses $\mathcal{L}_{\mathrm{G}}, \mathcal{L}_{\mathrm{perc}}$ and updated discriminator heads ψ

if $\alpha_{\mathrm{adv}} > 0$ **then**
  $r_{i+1} \leftarrow (1 - \sigma^{\mathrm{s}}_{i+1}) x_{\mathrm{real}} + \sigma^{\mathrm{s}}_{i+1} \epsilon'$ `// Same condition and noise level as the fake endpoint`
  $(f^{\mathrm{S}}, f^{\mathrm{R}}) \leftarrow \big(\mathcal{F}_\phi(z^{\mathrm{S}}_{i+1}, \sigma^{\mathrm{s}}_{i+1}, c), \mathcal{F}_\phi(r_{i+1}, \sigma^{\mathrm{s}}_{i+1}, c)\big)$ `// Frozen parameters; retain input gradients for` f^{S}
  $d^{\mathrm{S}} \leftarrow D_\psi(\mathrm{sg}[f^{\mathrm{S}}]), d^{\mathrm{R}} \leftarrow D_\psi(f^{\mathrm{R}})$
  $\mathcal{L}_{\mathrm{D}} \leftarrow \mathrm{softplus}(d^{\mathrm{S}}) + \mathrm{softplus}(-d^{\mathrm{R}})$
  $\psi^+ \leftarrow \mathrm{AdamWUpdate}(\psi, \nabla_\psi \mathcal{L}_{\mathrm{D}})$ `// Separate discriminator-head update`
  $\mathcal{L}_{\mathrm{G}} \leftarrow \mathrm{softplus}\big(-D_{\mathrm{sg}[\psi^+]}(f^{\mathrm{S}})\big)$ `// Gradient flows to` z^{S}_{i+1}`, not to` ϕ `or` ψ^+
  $\psi \leftarrow \psi^+$
else
  $\mathcal{L}_{\mathrm{G}} \leftarrow 0$
if $\alpha_{\mathrm{perc}} > 0$ **then**
  $\hat{x}^{\mathrm{S}}_{0,i} \leftarrow z_i - \sigma^{\mathrm{s}}_i v_\theta(z_i, \sigma^{\mathrm{s}}_i, c)$
  $x^{\mathrm{T}}_{0,i} \leftarrow \mathrm{TeacherRollout}_\phi(z_i, \sigma^{\mathrm{t}}_{b_i}, 0, c)$ `// Teacher continuation from the same query to the clean boundary`
  $f^{\mathrm{S}}_{i,m} \leftarrow F_m(\mathcal{V}(\hat{x}^{\mathrm{S}}_{0,i}))$
  $f^{\mathrm{T}}_{i,m} \leftarrow F_m(\mathcal{V}(x^{\mathrm{T}}_{0,i}))$ `// Frozen parameters; retain gradients through the student branch`
  $\mathcal{L}_{\mathrm{perc}} \leftarrow \frac{1}{M} \sum_{m=1}^{M} \left(1 - \frac{\langle f^{\mathrm{S}}_{i,m}, \mathrm{sg}[f^{\mathrm{T}}_{i,m}] \rangle}{\|f^{\mathrm{S}}_{i,m}\|_2 \|f^{\mathrm{T}}_{i,m}\|_2}\right)$
else
  $\mathcal{L}_{\mathrm{perc}} \leftarrow 0$
return $\mathcal{L}_{\mathrm{G}}, \mathcal{L}_{\mathrm{perc}}, \psi$

C ADDITIONAL RESULTS

Table 3 reports the six compositional dimensions of GenEval for the models in Table 1.

Table 4: **Detailed VBench dimension scores for bidirectional video.** The table reports the 16 VBench dimensions and their unweighted mean. Red backgrounds mark the best value among the Distillation Methods. Bold marks the best value in each full column. A red cell is bold when it is also the full-column best. The unweighted mean for rCM is computed from its reported 16 dimension scores.

Method	Subject cons.	Background cons.	Temporal flick.	Motion smooth.	Dynamic degree	Aesthetic quality	Imaging quality	Object class	Multiple objects	Human action	Color	Spatial relationship	Scene	Appearance style	Temporal style	Overall cons.	Mean (16-dim.)
Pretrained Models																	
Wan2.1-T2V-1.3B	0.924	0.955	**0.993**	0.969	0.625	0.546	0.579	0.608	0.393	0.660	0.658	0.520	0.221	0.283	0.204	0.212	0.584
Released Distilled Models																	
FastWan2.1-T2V-1.3B	0.972	0.937	0.987	**0.987**	0.736	0.602	0.702	0.756	0.538	0.680	0.820	0.682	0.302	0.278	0.210	0.214	0.650
T2V-Turbo-v2	0.952	**0.975**	0.969	0.966	0.944	**0.675**	**0.707**	**0.964**	0.597	**0.970**	**0.889**	0.468	**0.837**	**0.348**	**0.269**	**0.283**	**0.738**
AnyFlow	**0.984**	0.962	0.991	0.984	0.681	0.651	0.690	0.851	**0.742**	0.770	0.837	**0.801**	0.523	0.303	0.244	0.245	0.704
Distillation Methods																	
DMD2	0.958	0.929	0.978	0.985	**0.958**	0.580	0.660	0.835	0.664	0.770	0.861	0.747	0.285	0.190	0.246	0.240	0.680
rCM	0.954	0.941	0.975	0.983	0.708	0.590	0.660	0.756	0.563	0.690	0.741	0.640	0.302	0.286	0.216	0.213	0.639
OPTD	0.924	0.938	0.989	0.982	0.417	0.534	0.551	0.588	0.278	0.590	0.847	0.499	0.151	0.211	0.203	0.219	0.557
OPTD+perceptual	0.933	0.944	0.989	0.979	0.472	0.553	0.600	0.593	0.351	0.600	0.833	0.560	0.225	0.213	0.213	0.228	0.580
OPTD+GAN	0.938	0.954	0.990	0.986	0.847	0.654	0.677	0.863	0.554	0.790	0.866	0.733	0.345	0.220	0.257	0.266	0.684

Table 3: **GenEval compositional evaluation.** NFE denotes the number of function evaluations. SO, TO, and CA denote Single Object, Two Objects, and Color Attribution, respectively. Red highlights the best results per column among LCM, DMD2, PFM, TDM, rCM, and OPTD variants across 1–4 NFEs, including ties in the reported source values. Scores are displayed to three decimal places, while rankings use the available unrounded values. OPTD ($4 \to 2$) and OPTD ($2 \to 1$) are obtained by continuing the distillation process from the preceding student.

Method	NFE	SO ↑	TO ↑	Counting ↑	Colors ↑	Position ↑	CA ↑
Pretrained Models							
SD3.5-M	50	0.990	0.860	0.640	0.830	0.250	0.600
SD3.5-L	50	0.990	0.900	0.700	0.840	0.270	0.580
FLUX.2-klein-base-4B	50	0.990	0.860	0.750	0.860	0.510	0.590
Released Distilled Models							
SD3.5-L-Turbo	4	0.990	0.870	0.700	0.810	0.270	0.560
FLUX.2-klein-4B	4	0.996	0.940	0.840	0.870	0.660	0.630
Distillation Methods							
LCM	4	0.980	0.760	0.610	0.790	0.480	0.450
DMD2	4	0.990	0.850	0.520	0.830	0.450	0.480
PFM	4	0.870	0.260	0.410	0.470	0.140	0.093
TDM	4	0.990	0.790	0.790	0.810	0.510	0.550
rCM	4	0.940	0.680	0.540	0.750	0.280	0.350
OPTD	4	0.990	0.860	0.730	0.850	0.510	0.530
OPTD+perceptual	4	**1.000**	0.870	0.690	0.820	0.570	0.540
OPTD+GAN	4	0.981	**0.889**	0.781	0.859	0.578	0.593
OPTD+GAN+perceptual	4	0.981	0.849	**0.831**	0.859	**0.648**	0.583
OPTD ($4 \to 2$)	2	0.985	0.842	0.637	**0.879**	0.588	**0.613**
OPTD ($2 \to 1$)	1	0.985	0.835	0.628	0.859	0.578	**0.613**

Table 4 gives the complete VBench dimension scores for the bidirectional video comparison. AnyFlow is listed with the released distilled models and has the strongest semantic alignment and unweighted mean among the released 81-frame four-step comparisons. Within the Distillation Methods, OPTD+GAN leads the summary scores and improves several motion, appearance, and action dimensions. DMD2 obtains the highest Dynamic Degree score in that group, but this advantage does not yield the strongest summary scores.

Figure 6 provides a qualitative comparison of the image refinement variants discussed in Section 4.3. GAN refinement emphasizes realistic textures and backgrounds, whereas perceptual refinement em-



Figure 6: **Qualitative ablation of image refinements.** We compare four-step OPTD, OPTD with perceptual refinement, OPTD with GAN refinement, and the combination of both refinements under the same prompts.

phasizes cleaner visual details. Their combination exposes the trade-off between these two properties.

Figure 7 gives the complete qualitative comparison on four prompts. For each prompt, the upper row places OPTD beside the FLUX.2, TDM, DMD2, and rCM baselines, while the lower row adds the SD3.5 and LCM/PFM references. OPTD retains the requested scene category and coherent object layouts across the boat, vegetables, skiing, and bedroom prompts. The grid also shows the variation in composition across baselines that is not visible in the representative examples in the main text.

Additional qualitative grids, per-boundary curves, diversity samples, and ablations are organized by modality and by student step count in the accompanying code release.

D LIMITATIONS

Video distillation is time-consuming because constructing each endpoint target requires teacher integration over a full student interval with classifier-free guidance. Performance also depends on the base model. When its generative capabilities are limited, endpoint matching alone lacks independent supervision to correct inherited deficiencies. Our GAN refinement therefore relies on additional real training data to provide this corrective signal.

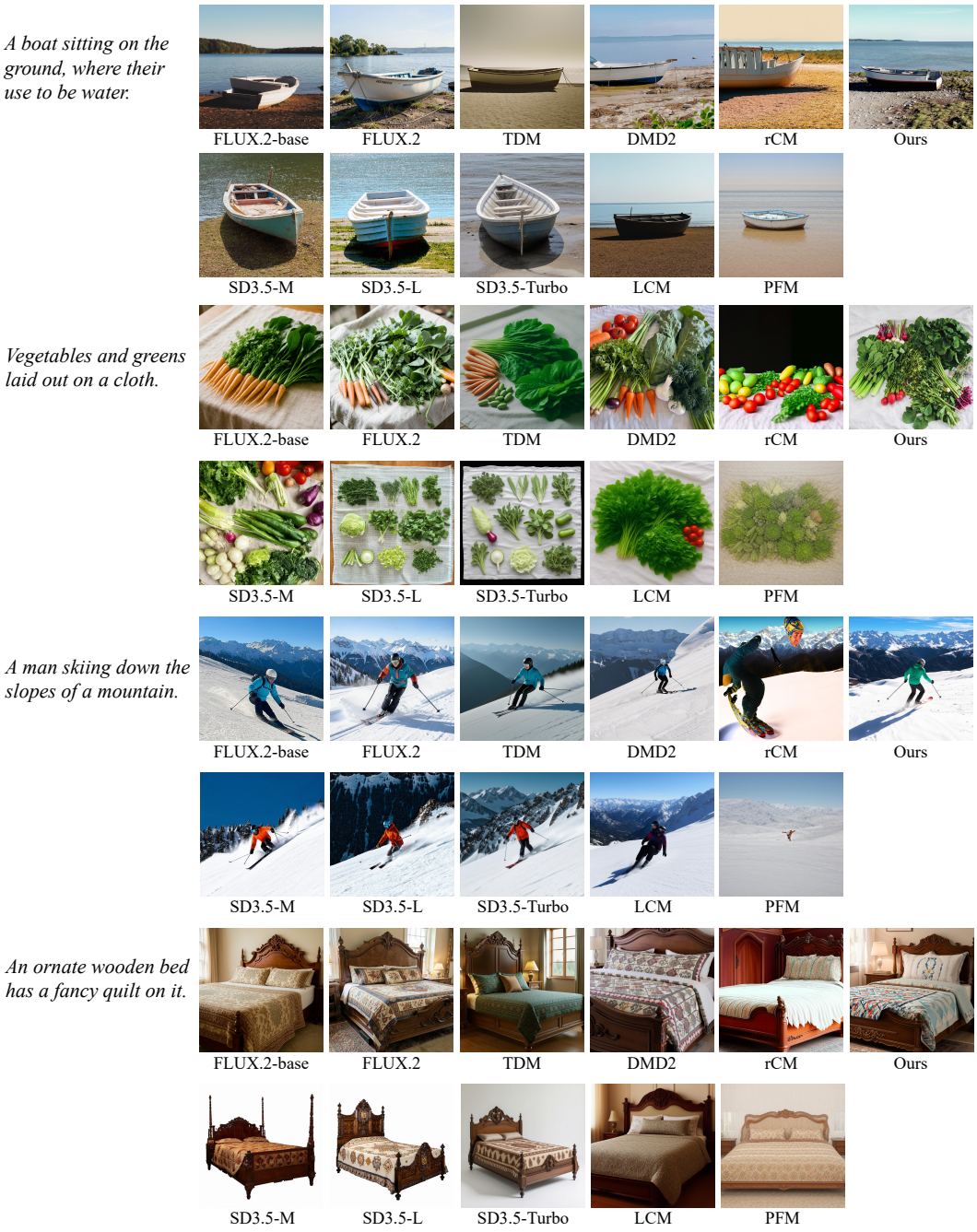

Figure 7: **Complete qualitative comparison across four prompts.** The grid compares OPTD with FLUX.2-base, FLUX.2, TDM, DMD2, rCM, SD3.5-M, SD3.5-L, SD3.5-Turbo, LCM, and PFM. The prompts cover a boat, vegetables on a cloth, a skier on a snowy slope, and an ornate bed.

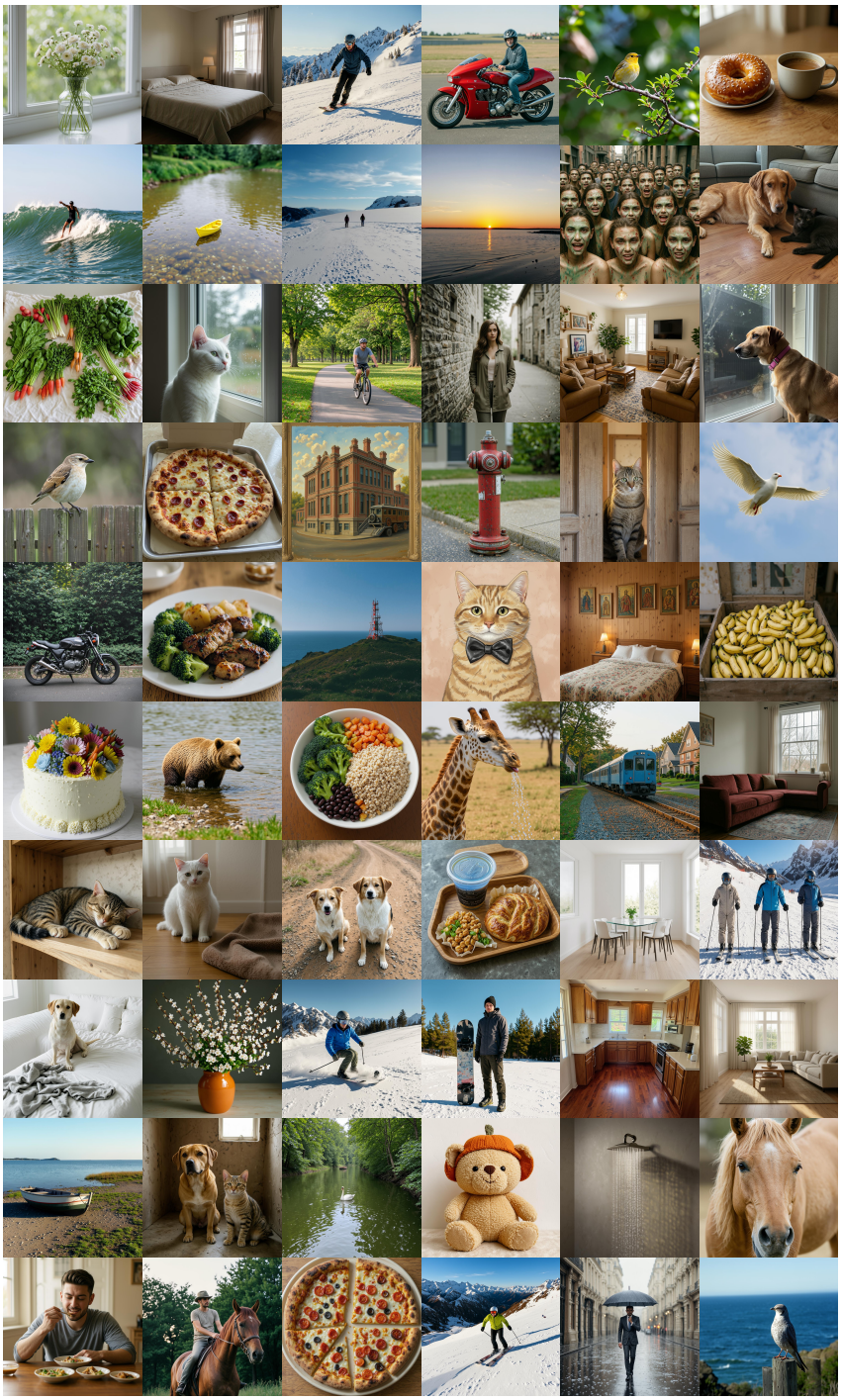

Figure 8: **Additional qualitative results.**