PACT: PRESERVING ACCURATE CONTINUOUS TRAJECTORIES VIA FLOW-MAP DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale Flow Matching models have set a new standard for image and video generation, but their reliance on many sequential inference steps makes them slow to deploy. While current distillation approaches like Distribution Matching Distillation (DMD) offer a way to accelerate this process, they often fail to preserve the teacher model's original data distribution, suffer from reduced visual diversity, and require highly complex training setups. To address this, we introduce PACT (Preserving Accurate Continuous Trajectories), a simple, stable, and compute-efficient framework that extends shortcut-path learning to pre-trained large-scale generators. By building an offline bank of the teacher's generation trajectories, we train a lightweight student model, using only Low-Rank Adaptation (LoRA) and new time-conditioning inputs, to predict the correct flow maps between any two visited states. This approach allows us to reuse offline data efficiently and avoids the need for auxiliary score models. We demonstrate PACT's scalability by distilling Stable Diffusion 3.5 Medium (2.5B parameters) and the LTX-2.3 video model (22B parameters) into highly efficient 4-step and 8-step generators. Our video student generates five-second audiovisual clips faster than real time, on a single GPU. Our distilled models achieve competitive visual quality while preserving the rich distributional diversity and exact seed-level behaviors of their teachers. Furthermore, PACT exhibits robust zero-shot generalization, successfully generating videos at resolutions and aspect ratios completely unseen during training.

1 INTRODUCTION

Large-scale Flow Matching models (Lipman et al., 2023) have become a powerful standard for generating high-quality images and videos, driving both creative content tools and visual world models. However, their practical use is heavily bottlenecked by slow generation speeds. Generating a single sample requires solving an ordinary differential equation over many sequential steps, resulting in a high number of function evaluations (NFEs) that can take an impractical amount of time to process; *e.g.* it can take tens of minutes to generate a short 5 second clip.

To solve this, researchers often use distillation to compress the generation process into fewer steps (Salimans & Ho, 2022; Song et al., 2023). Methods such as Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) achieve fast and high quality generation but do so by relaxing the requirement to match the teacher model's exact generation path. While this produces high-quality samples, it can reduce sample diversity relative to the teacher (Wu et al., 2026). Preserving this diversity is critical for rigorously training policies in world models, or for giving users a wide variety of accurate choices in creative applications. Furthermore, DMD training requires an auxiliary score model whose estimation accuracy can affect training stability (Yin et al., 2024a). Alternatively, Shortcut Models (Frans et al., 2025), Flow Map Matching (Boffi et al., 2025), and MeanFlow (Geng et al., 2025) learn finite-time transitions or average velocities to take larger steps along generative trajectories. Shortcut Models and MeanFlow demonstrate training from scratch, while Flow Map Matching also supports distillation from pretrained velocity fields. We build on this perspective to adapt large pretrained image and video generators.

In this work, we propose PACT (Preserving Accurate Continuous Trajectories), a simple, scalable, and compute-efficient framework that brings shortcut-path learning to large pre-trained image and video models. Our approach guarantees that the distilled student model preserves the teacher's

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.

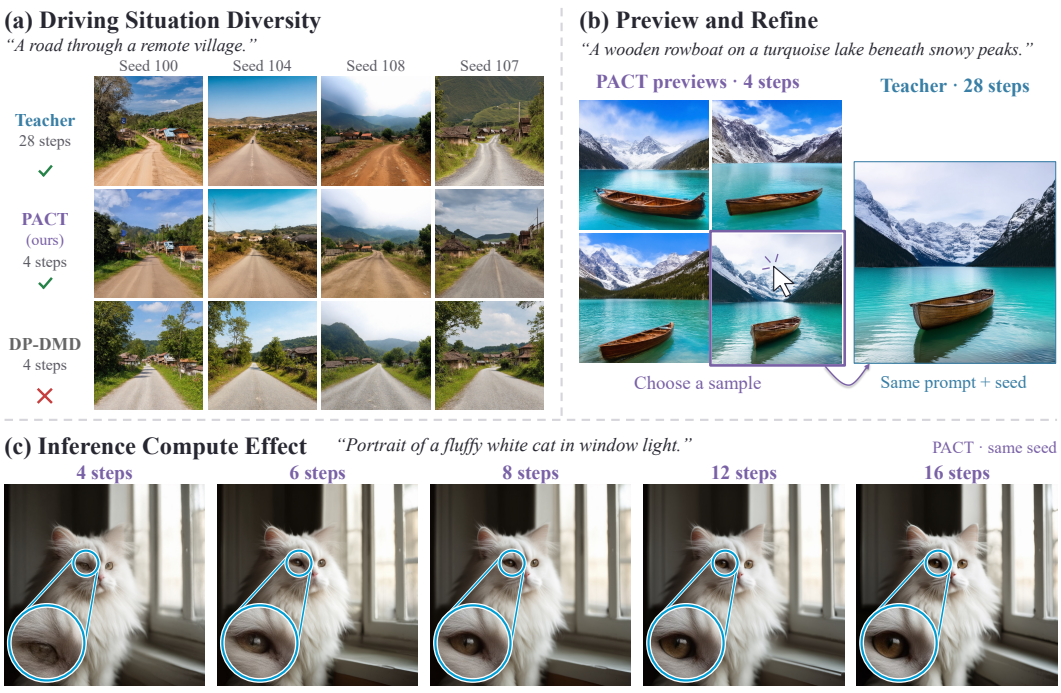


Figure 1: **Fast generation with teacher correspondence.** Selected SD3.5 examples illustrate (a) variation across matched seeds, (b) four-step PACT previews followed by teacher generation with the selected prompt and seed, and (c) changes in image detail as PACT uses more sampling steps. For more qualitative image and video samples check Appendix A.7 and the supplementary material.

generation distribution by strictly learning to follow its original trajectories. We first build an offline "trajectory bank" containing the intermediate states visited by the teacher model during standard generation. Because this bank is generated offline, a single trajectory can be reused multiple times during training by sampling different pairs of start and end states, making the process highly resource-efficient. We then slightly modify the frozen teacher model by adding lightweight Low-Rank Adaptation (LoRA) (Hu et al., 2022) parameters and new conditioning inputs for the target time and guidance scale. This allows the student to accurately predict the flow map between any two visited states.

We demonstrate the effectiveness of our method by distilling two major foundational models: Stable Diffusion 3.5 Medium (2.5 billion parameters) (Esser et al., 2024) for image generation, and LTX-2.3 (22 billion parameters) (HaCohen et al., 2025) for video generation. Across both domains, PACT achieves competitive visual quality using only a fraction of the original sampling steps. On a single NVIDIA GH200 GPU, generating a five-second audiovisual clip takes about 4.4 seconds with PACT, compared to 52 seconds for the teacher; a nearly 12× speedup that makes generation faster than playback. More importantly, our approach perfectly matches the teacher's diversity (Figure 1a), significantly outperforms DMD baselines, reproduces very similar samples to the teacher (Figure 1b), and requires no auxiliary score models to train. Finally, we showcase the robustness of our approach by demonstrating strong zero-shot generalization, proving that our distilled video student can successfully generate content at resolutions and aspect ratios it never saw during training.

2 Related Works

Progressive distillation. Progressive Distillation (Salimans & Ho, 2022) and its guided extension (Meng et al., 2023) reduce sampling budgets through successive distillation stages, halving required steps after each stage. PACT instead trains one interval-conditioned student from a fixed trajectory bank, reusing teacher states across step lengths.

Consistency models. Consistency Models (Song et al., 2023) and Latent Consistency Models (Luo et al., 2023) learn fixed-destination maps for fast sampling, rather than accepting a destination time.

Consistency Trajectory Models (Kim et al., 2024) support arbitrary trajectory segments but combine online teacher integration with consistency, denoising, and adversarial losses. PACT directly regresses interval predictions on cached teacher-state pairs, without these auxiliary objectives or online teacher evaluations.

Shortcut models, flow maps, and MeanFlow. Shortcut Models (Frans et al., 2025) learn finite jumps through flow-matching supervision and two-step self-distillation. MeanFlow (Geng et al., 2025) learns average velocity using a target computed through Jacobian–vector products. Both support different sampling budgets without a separate teacher-distillation stage. Flow Map Matching (Boffi et al., 2025) develops a framework for two-time maps, including derivative-based objectives and offline regression on composed teacher maps. DSNO (Zheng et al., 2023) also uses an offline dataset to predict teacher trajectories from initial noise. We build on this interval-map perspective to adapt pretrained image and video generators through direct regression on cached source-to-destination transitions. Our training loop avoids online teacher evaluations and Jacobian–vector products, at the cost of offline generation and storage.

Distribution matching distillation. DMD (Yin et al., 2024b) seeks to match student and teacher distributions through an approximate reverse-KL objective estimated using a frozen teacher and a learned auxiliary score model. This relaxes trajectory matching and enables high-quality fast sampling, but adds an auxiliary model to train. Original DMD also uses paired regression; DMD2 (Yin et al., 2024a) removes this term from its main distillation objective, adds real-data adversarial training, and supports multistep generation with a prescribed sampling schedule. DP-DMD (Wu et al., 2026) reports reduced diversity under DMD and mitigates it by supervising the first step with a teacher-derived target, while applying distribution matching to later steps. Our objective fits cached transitions across sampled latent states, without an auxiliary score model or adversarial loss.

3 METHOD

We begin with the notation and definitions in Section 3.1, using a simple example to explain the difference between a local velocity and a finite transition. Section 3.2 then describes how we learn these transitions from stored teacher trajectories.

3.1 PRELIMINARIES

FLOW MATCHING AND ODE SAMPLING

Figure 2a,b illustrates how flow matching turns simple training paths into a generative process. Let $x_0 \in \mathbb{R}^d$ be a clean sample drawn from the data distribution $p_{\text{data}}(\cdot \mid c)$, c is a condition, such as a text prompt. We independently draw noise $\epsilon \sim \mathcal{N}(0, I)$ and connect the two by a straight path:

$$x_t = (1-t)x_0 + t\epsilon, \qquad t \in [0,1]. \tag{1}$$

Here, t measures the noise level: $x_1 = \epsilon$ is noise and x_0 is data, so generation moves from $t = 1$ toward $t = 0$. The velocity along each such path is the constant vector $\epsilon - x_0$. Conditional flow matching (Lipman et al., 2023) trains a neural network v_ϕ to predict this velocity:

$$\mathcal{L}_{\text{CFM}}(\phi) = \mathbb{E}_{c,x_0,\epsilon,t}\left[\|v_\phi(x_t,t;c) - (\epsilon - x_0)\|_2^2\right]. \tag{2}$$

The expectation is over training examples, noise, and t at uniform between 0 and 1.

Figure 2a highlights an important feature of this supervision: several data–noise pairings can pass through the same state and ask for different velocities. Under squared error, the optimal prediction is their conditional mean, $v(x,t;c) = \mathbb{E}[\epsilon - x_0 \mid x_t = x, c]$, called the *marginal velocity*. The generative process follows this mean field. Figure 2b shows the resulting trajectories. Although the individual training paths are straight, the direction and speed of their mean can change along a trajectory, making these generative paths curved.

To generate a sample, we start from Gaussian noise and follow the learned field by solving the ordinary differential equation (ODE) $\mathrm{d}x_t/\mathrm{d}t = v_\phi(x_t,t;c)$. Euler sampling approximates this motion with a sequence of steps. For a source and destination time t, r with $0 \le r < t \le 1$, one step is

$$\widehat{x}_r = x_t + (r-t)v_\phi(x_t,t;c), \tag{3}$$

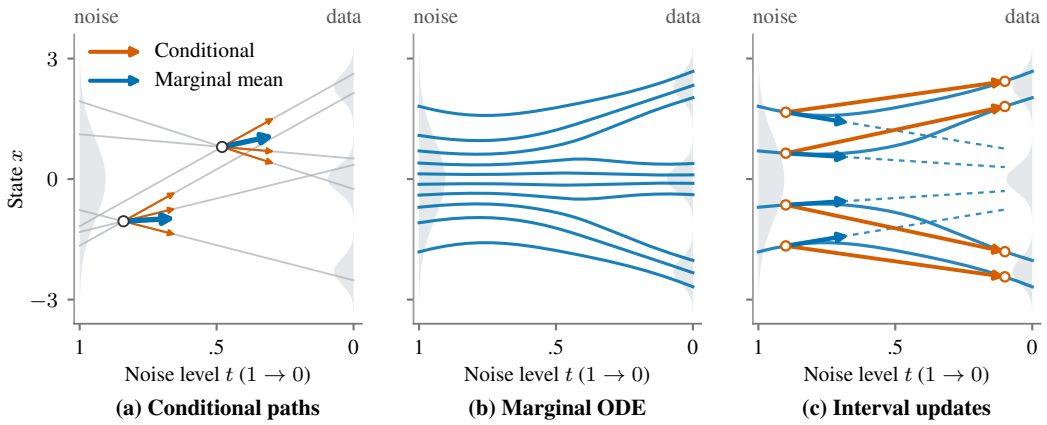


Figure 2: **From local directions to finite transitions.** An analytic example transports Gaussian noise to a mixture of three Gaussians. **(a)** Gray lines show selected straight interpolation paths through two states. Orange arrows indicate their individual denoising directions; blue arrows show the analytic conditional mean. **(b)** Ten trajectories obtained by integrating this marginal velocity field. **(c)** Four trajectories with orange arrows connecting source and destination states. Blue arrows show the instantaneous direction at each source; dashed lines extrapolate it to the same destination time. The orange chord reaches the trajectory endpoint, whereas the tangent need not.

where $\widehat{x}_r$ is the estimated destination state. This follows the local tangent, keeping the velocity fixed over the interval. A long step can therefore miss the integrated trajectory as the field changes. Conventional Euler sampling typically needs many small steps to stay close to that trajectory; our goal is to learn how to cover the same motion in larger steps.

Flow maps and average velocity

Figure 2c shows the transition we would like to predict. From a state x_t, the local velocity points along the tangent to the trajectory. The arrow joining x_t to a destination state x_r, however, accounts for all the changes in velocity between them. Predicting that transition would let us reach x_r without evaluating the field at every intermediate step.

A *flow map* expresses precisely this idea (Boffi et al., 2025): $F(x_t, t, r) = x_r$ returns the state reached by integrating the field from time t to time r. For clarity, we leave the fixed condition c implicit here. Rather than predict x_r directly, we describe the transition through its *average velocity* (Geng et al., 2025),

$$u(x_t, t, r) = \frac{x_r - x_t}{r - t}, \qquad x_r = x_t + (r - t)u(x_t, t, r). \tag{4}$$

The numerator is the displacement shown by the arrow in Figure 2c. Dividing by the time interval gives the constant velocity that would produce the same endpoint. It summarizes the motion over the interval, even when the actual trajectory bends along the way.

This representation lets us retain a pretrained model's velocity output while changing what it learns to predict. We train a student u_θ that also receives the destination time r and takes the step

$$\widehat{x}_r = x_t + (r - t)u_\theta(x_t, t, r). \tag{5}$$

For a short interval, average velocity approaches local velocity; for a longer interval, it must account for the motion ahead. At inference, we choose a descending schedule $1 = s_0 > s_1 > \cdots > s_K = 0$ and apply this update between successive noise levels, where K is the number of steps. The destination input lets the same student serve different step counts and schedules.

Classifier-free guidance

Classifier-free guidance (CFG) (Ho & Salimans, 2022) also fits this picture. For a guidance scale w, it combines the prediction under condition c with one under a null or negative condition c_-:

$$v^w(x, t; c) = v_\phi(x, t; c_-) + w\big[v_\phi(x, t; c) - v_\phi(x, t; c_-)\big]. \tag{6}$$

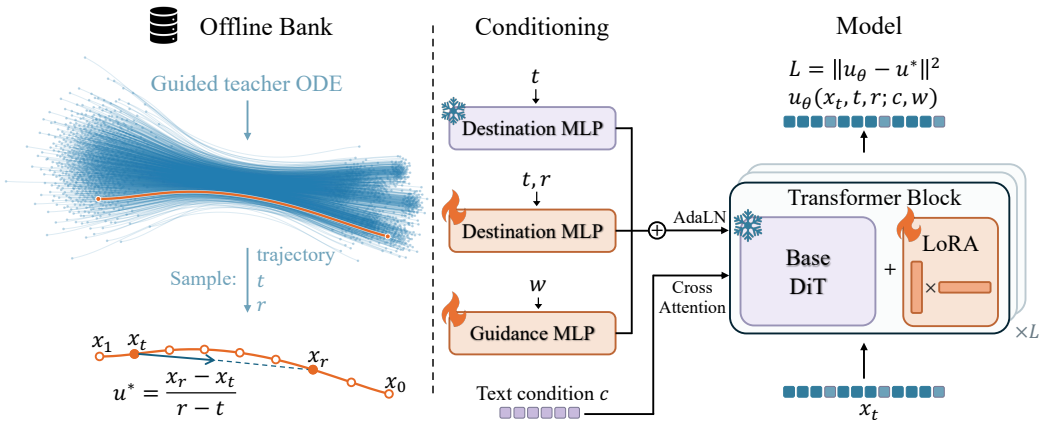


Figure 3: **Adapting a pretrained generator to predict interval-average velocity.** Two states from a stored teacher trajectory provide the source x_t and the target velocity $(x_r - x_t)/(r - t)$. The student reuses the pretrained backbone, adding trainable LoRA updates and destination and guidance branches to its time-conditioning pathway. The new branches modify the existing adaptive normalization, while the original backbone weights remain frozen.

For a fixed w, this is simply another velocity field: its guidance can be absorbed into the trajectories we distill, without changing the flow-map formulation. A student trained on those trajectories can incorporate that fixed guidance in a single prediction (Meng et al., 2023). We keep guidance strength available as a control by learning from trajectories at different scales and conditioning student on w. The next section describes how we learn this destination and guidance conditioned model.

3.2 PACT: Flow-Map Distillation

Figure 3 presents our approach. A teacher trajectory already contains examples of the transitions we want to learn: any two of its states specify where a step should start and end. Building on offline flow-map regression (Boffi et al., 2025), we collect these trajectories once and use their state pairs to adapt a pretrained generator. In practice, the states are compressed latent representations.

Offline trajectory bank

For each condition c, noise seed, and guidance scale w, we run the frozen teacher on a descending schedule $1 = t_0 > \cdots > t_N = 0$, where N is the number of teacher steps. Writing its guided velocity as $v_T(x, t; c, w)$, Euler integration gives

$$x_{t_{i+1}} = x_{t_i} + (t_{i+1} - t_i) v_T(x_{t_i}, t_i; c, w), \qquad i = 0, \ldots, N-1. \tag{7}$$

We save all $N + 1$ states with their noise levels, condition, and guidance scale in a bank $\mathcal{B}$. These states approximate the teacher's continuous trajectory through numerical integration.

Keeping the intermediate states makes each rollout useful beyond its final sample. Nearby states provide examples of short transitions, while distant states show how to cover a larger part of the trajectory in one step. We can draw new pairs from the same bank throughout training, reusing the teacher computation across updates and training runs. Collecting trajectories at several guidance scales also gives the student examples of how guidance changes these transitions. The bank construction and native teacher guidance settings are detailed in Appendix A.1.

Destination and guidance conditioning

The backbone in Figure 3 already knows how to predict local velocity from a noisy state, its time, and a condition. We build on this starting point by giving it two additional controls: the destination time r and the guidance scale w. Together, they specify how far to move and which guided trajectory to follow.

We encode these controls with two small multilayer perceptrons (MLPs) and add their outputs to the pretrained time embedding. One branch encodes the source, destination, and their interval; the other encodes the guidance scale. These residuals enter the existing adaptive layer normalization (AdaLN) pathway, through which time conditioning modulates the transformer's features. Low-rank adaptation (LoRA) (Hu et al., 2022) adds trainable low-rank updates to selected backbone projections while keeping their pretrained weights frozen.

Both additions start with zero effect: the new conditioning branches have zero-output final layers, and each LoRA update has one factor initialized to zero. The student therefore begins with the pretrained conditional prediction, then learns to account for the requested interval and guidance.

LOSS AND SUPERVISION

Training draws two states x_t, x_r from the same bank trajectory, with $r < t$, together with their condition c and guidance scale w. The transition itself supplies the target:

$$u^\star = \frac{x_r - x_t}{r - t}, \qquad \mathcal{L}(\theta) = \mathbb{E}_{\mathcal{B}} \left[\frac{1}{d} \left\| u_\theta(x_t, t, r; c, w) - u^\star \right\|_2^2 \right]. \tag{8}$$

Here, the expectation samples a trajectory and a source–destination pair from the bank, and d counts the generated coordinates. We sample pairs according to their noise-level gap, so the student learns from a range of interval lengths; the sampling policy and loss weighting are given in Appendix A.1.

Every target comes directly from two stored states, so a training update needs no teacher evaluation or differentiation through a teacher trajectory. The learned model then turns the same interval prediction into a sampling step through Equation equation 5, requiring only a single forward pass.

4 EXPERIMENTS

Can learning a teacher's trajectories preserve its generation behavior with fewer sampling steps? We study this question through three complementary measurements: generation quality, variation across seeds, and agreement with the teacher for the same seed. We first evaluate Stable Diffusion 3.5 Medium, where an open distribution-matching baseline allows a direct comparison, then scale to the 22-billion-parameter LTX-2.3 video model. Together, these experiments test how much of the teacher's quality and diversity survives distillation, how the student benefits from additional inference compute, and whether it transfers beyond the training resolution.

4.1 IMAGE GENERATION WITH STABLE DIFFUSION 3.5 MEDIUM

Setup. We distill SD3.5-M (Esser et al., 2024) and compare four-step PACT with its teacher, the released DP-DMD checkpoint (Wu et al., 2026), and published DMD variants. Dataset, training, sampling, and evaluation details are in Appendix A.3.1.

With compiled inference on a single NVIDIA GH200 GPU, PACT reduces image generation time from 1.71 to 0.27 seconds ($6.4\times$), including text encoding and VAE decoding. Appendix A.4 reports full timings, with image postprocessing and file saving measured separately.

Metrics. GenEval tests compositional prompt following (Ghosh et al., 2023); COCO-10K and Pick-a-Pic (PaP) cover natural-image captions and user-written prompts (Lin et al., 2014; Kirstain et al., 2023). CLIPScore measures text–image alignment, and ImageReward estimates human preference (Hessel et al., 2021; Xu et al., 2023). COCO FID compares generated and real-image feature distributions; lower is better (Heusel et al., 2017). CLIP and DINOv3 feature distances measure diversity across seeds, while similarity to the teacher for matched prompts and seeds measures sample correspondence. For the latter, we use global CLIP features and local DINOv3 patches.

FOUR-STEP QUALITY AND DIVERSITY

As shown in Figure 4, PACT's performance in compositional accuracy and text-image alignment is bounded by, yet very close to, the teacher model. While baselines like DP-DMD can sometimes exceed the teacher's performance on these specific quality metrics, the entire DMD family suffers from a significant drop in generation diversity. Even DP-DMD, which is specifically designed to mitigate

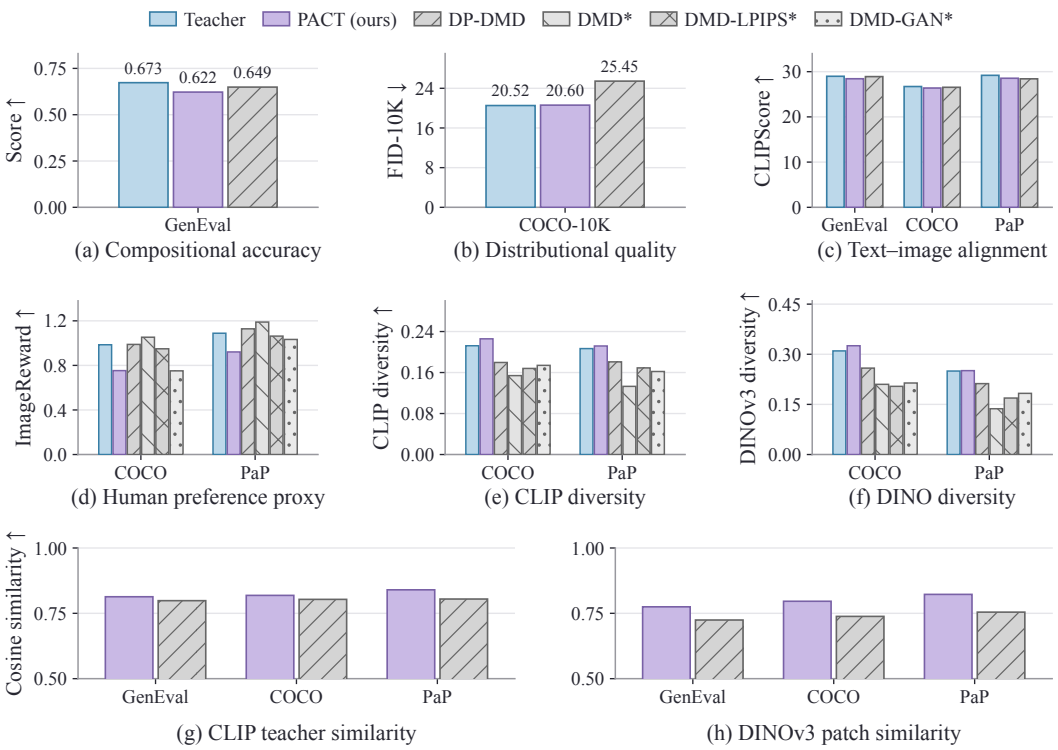


Figure 4: **SD3.5-M quality, diversity, and teacher agreement.** (a–f) Four-step PACT retains near-teacher quality on several metrics and preserves more diversity than the DMD baselines. (g–h) For matched prompts and seeds, PACT follows the teacher more closely than DP-DMD in global CLIP and local DINOv3 patch features. Starred baselines use reported results, not our reproductions.

this issue, fails to maintain the original distribution. In contrast, PACT successfully matches the teacher's diversity both globally (measured by CLIP) and locally (measured by DINO). Figure 1a and the qualitative examples in Appendix A.7 visually illustrate this preserved diversity across different random seeds.

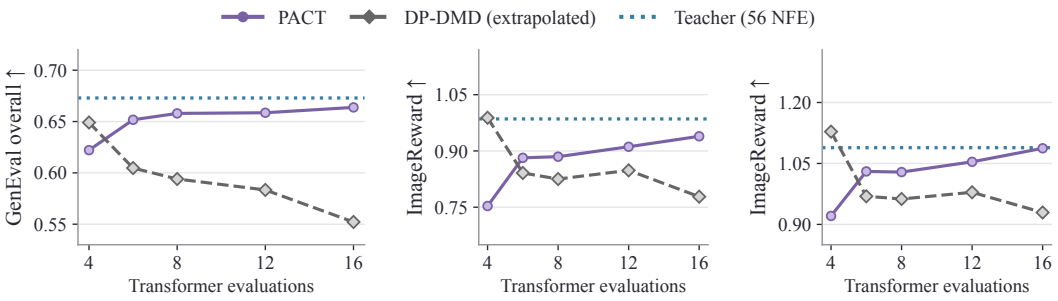
This lack of diversity in baseline methods also negatively impacts their COCO FID scores. Because FID evaluates the overall distribution of images, both quality and diversity matter; consequently, PACT successfully matches the teacher's FID score, whereas DP-DMD clearly lags behind. The primary area where PACT currently falls behind the baselines is in the human preference proxy (ImageReward). However, as we demonstrate in Figure 5, this gap substantially reduces when increasing test-time compute with additional sampling steps.

Preserving the Teacher's Seed-to-Image Behavior

Figure 4(g–h) shows that, for the same prompt and noise seed, our student produces images closer to the teacher's than DP-DMD does. This agreement holds both globally, through CLIP features, and locally, through corresponding DINO patches. It is consistent with our training objective: each supervised transition belongs to a specific teacher trajectory, rather than only matching the teacher's output distribution. Beyond retaining the pretrained model's generation behavior, this suggests a practical preview-and-refine workflow: use the fast student to explore prompts and seeds, then render promising choices with the teacher for higher-fidelity sampling. Figure 1b and Appendix A.7 provide qualitative examples of this strict teacher correspondence.

Trading Inference Compute for Quality

Figure 5 shows our student's flexibility in using additional inference compute to approach the teacher's compositional accuracy and preference scores, without retraining. Much of the gain al-



(a) GenEval (b) COCO-10K (c) Pick-a-Pic

Figure 5: **More inference compute improves the same trained student.** The largest early gain comes from four to six transformer evaluations (NFE), especially in ImageReward, while DP-DMD degrades when getting further from its optimized schedule, demonstrating higher flexibility of PACT.

ready appears when moving from four to six evaluations, especially in ImageReward, substantially narrowing the gap observed in the four-step comparison. Figure 1c and Appendix A.7 visually demonstrate how additional sampling steps improve fine image details. In contrast, extending the four-step DP-DMD model with these schedules hurts performance relative to its native setting.

4.2 Scaling to LTX-2.3 Video Generation

Setup. We distill the audiovisual LTX-2.3 model (HaCohen et al., 2025) and compare eight-step PACT with its teacher and the official full and LoRA distilled checkpoints. These checkpoints serve as open-weight industrial references with an unpublished distillation recipe; dataset, training, generation, and evaluation details are in Appendix A.3.2.

PACT generates a five-second audiovisual clip from text in 4.4 seconds end to end, on a single NVIDIA GH200 GPU, roughly $12\times$ faster than the teacher. Full measurements are in Appendix A.5. We include qualitative examples of our generations in the supplementary materials.

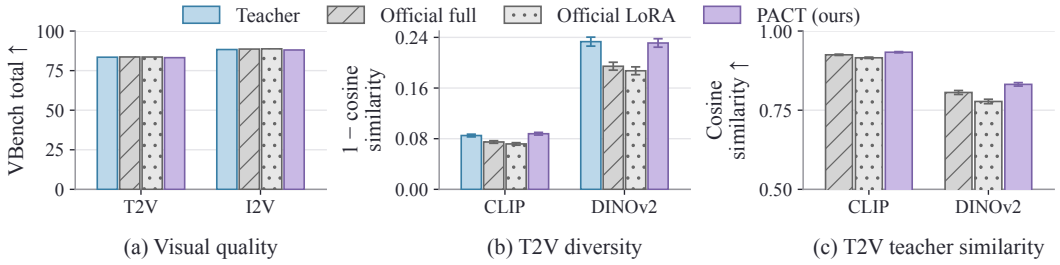
Metrics. We evaluate text-to-video (T2V) with VBench and image-to-video (I2V) with VBench++ (Huang et al., 2024; 2026), assessing visual quality, motion, and agreement with text or the conditioning image. CLIP and DINOv2 features pooled across video frames measure diversity and matched-seed teacher similarity, following the image comparisons.

Eight-Step Quality, Teacher Preservation, and Diversity

Figure 6 brings together video quality, diversity, and agreement with the teacher. Our student retains nearly the teacher's VBench performance on both T2V and I2V, with only a small decrease, while the official distilled models show small improvements over the teacher. The distinction is again in how the models preserve its generation behavior. For T2V, our student closely retains both CLIP and DINO diversity, whereas the official distilled models produce less varied samples. The same diversity trend holds for I2V, with smaller absolute differences between students in a setting where the input image constrains the range of possible outputs; we report these results in Appendix A.3.2. As in the image experiments, paired T2V samples also agree more closely with the teacher under our method. This is consistent with learning transitions along the teacher's own trajectories: reducing the sampling cost need not discard the variation and seed-specific behavior learned during pretraining. Qualitative video examples are provided in the supplementary material.

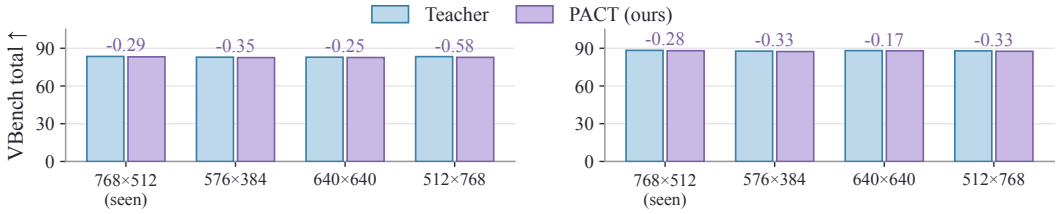
Beyond the Training Resolution

Figure 7 tests a less direct form of transfer: can a student trained on one spatial grid retain the teacher's ability to generate at other resolutions? Although the teacher supports different resolutions and aspect ratios, we use only 768×512 videos for bank generation and distillation. We then evaluate the same student on smaller landscape, square, and portrait videos, without further training or schedule changes. This is more than a change in output size. The smaller landscape grid uses



(a) Visual quality (b) T2V diversity (c) T2V teacher similarity

Figure 6: **LTX-2.3 quality and teacher preservation at eight steps.** VBench totals remain close to the 30-step teacher on both tasks. T2V diversity and paired teacher similarity favor our student over the official distilled baselines. Features pool sixteen evenly spaced frames using CLIP and DINOv2 (Oquab et al., 2024); error bars are 95% prompt-bootstrap intervals. The similarity axis starts at 0.5. Detailed components and I2V feature results are in the appendix.



(a) Text-to-video (b) Image-to-video

Figure 7: **Transfer beyond the training resolution.** Only 768×512 is used during distillation. The same student transfers to smaller landscape, square, and portrait grids. Purple labels give student-minus-teacher gaps in percentage points. Comparisons are paired within each geometry.

a little over half as many video tokens, reducing the dimension of the state on which the learned transitions act; changing aspect ratio rearranges the token grid and its spatial positional relationships. The student must therefore apply its learned jumps to state geometries absent from its trajectory bank. Despite this change, the student surprisingly stays close to the teacher across all three unseen geometries, with VBench total-score gaps below 0.6 percentage points on both tasks. The gaps remain comparable to those at the training resolution, suggesting that the student has retained the teacher's spatial flexibility rather than specializing to a single grid.

5 Conclusion

PACT provides a stable and compute-efficient framework to accelerate large-scale generative models without sacrificing their core capabilities. By learning transitions directly from an offline bank of teacher trajectories, our approach trains student models to take large sampling steps while faithfully matching the teacher's exact output distribution. This design bypasses the instability of multi-model distillation techniques and the massive compute requirements of training shortcut models from scratch. As demonstrated on Stable Diffusion 3.5 Medium and the 22-billion parameter LTX-2.3 model, PACT drastically reduces the necessary function evaluations while maintaining highly competitive image and video quality.

Beyond accelerating generation, PACT uniquely preserves the rich visual diversity and seed-specific behavior of the original pre-trained models. This preservation is essential for demanding downstream tasks, such as generating diverse choices in creative workflows or simulating highly varied environments for world models. Furthermore, the distilled LTX student retains the spatial flexibility of its teacher, demonstrating surprising zero-shot generalization to resolutions and aspect ratios entirely unseen during training. By showing that large models can learn their own flow maps efficiently through offline caching, PACT offers a practical, scalable path for deploying state-of-the-art visual generators.

AI USE STATEMENT

We used generative AI tools to assist with research methodology and experimental design, method implementation, and qualitative and thematic data analysis. We did not use generative AI tools to propose or refine hypotheses, clean or reformat datasets, or interpret results. The remaining required-disclosure categories were not applicable. Additional uses included suggesting experimental parameters; creating and editing code, research artifacts, and scientific figures; drafting and polishing manuscript text; finding, summarizing, and analyzing relevant literature; and brainstorming and suggesting titles. The authors reviewed and revised the manuscript and take responsibility for its final content, including all AI-assisted text, claims, code, and figures.

REPRODUCIBILITY STATEMENT

We have made every effort to support reproducibility by documenting implementation details, experimental settings, and training hyperparameters in Appendices A.1 and A.3. Appendices A.4 and A.5 specify the generation-latency measurement protocols. Upon acceptance, we will publicly release our code and trained checkpoints.

REFERENCES

Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=cqDH0e6ak2.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One Step Diffusion via Shortcut Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uWj4s7rMnR.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Wbr51vK331.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2025.

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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

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*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ymjI8feDTD.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. URL https://arxiv.org/abs/2310.04378.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, June 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=2NlGyqNjns. Featured Certification.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. 2024. URL https://openreview.net/forum?id=pYNl76onJL.

Wenhao Wang and Yi Yang. TIP-I2V: A Million-Scale Real Text and Image Prompt Dataset for Image-to-Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14898–14908, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_TIP-I2V_A_Million-Scale_Real_Text_and_Image_Prompt_Dataset_for_ICCV_2025_paper.html.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 893–911, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.51. URL https://aclanthology.org/2023.acl-long.51/.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-Preserved Distribution Matching Distillation for Fast Visual Synthesis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=R3JzLU2qZj.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

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 *NeurIPS*, 2024a.

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

Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42390–42402. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zheng23d.html.

A Appendix

A.1 Method details

Teacher trajectories. The teacher velocity $v_T(x, t; c, w)$ includes all guidance operations used to generate the bank. For plain CFG, it is given by Equation equation 6; any additional native guidance is included in the same field. The negative or null reference prompt is fixed within a bank and is not a separate student input. Bank entries record the actual scheduler noise levels and guidance settings, not merely the step indices. Joint audio-video entries contain aligned audio and video states from the same rollout. Teacher-generated states provide the training targets; image-conditioned video additionally requires reference images. The cost of generating and storing the bank is shared across subsequent student updates and experiments.

Pair sampling and loss normalization. A trajectory with $N+1$ saved states supplies $N(N+1)/2$ candidate pairs before filtering. These pairs reuse one rollout and are therefore correlated. We require an actual noise-level gap $t - r \geq 0.05$ to avoid amplifying storage error through very small finite differences. We group eligible pairs into gap intervals bounded by $0, 0.25, 0.5, 0.75, 1$, select a group with probabilities $(0.3, 0.3, 0.3, 0.1)$, and sample uniformly among its eligible pairs. SD3.5 training draws eight pairs per loaded trajectory; LTX training draws one. We compute finite differences in float32. We implement the mean squared error (MSE) in Equation equation 8 with a binary mask M over latent coordinates. For an error vector a, with coordinates indexed by ℓ, the normalized squared error is

$$\|a\|_M^2 = \frac{\sum_\ell M_\ell a_\ell^2}{\sum_\ell M_\ell}. \tag{9}$$

We normalize each sample before averaging over the batch. The mask is all ones for text-to-image generation and excludes fixed reference tokens for image-conditioned video. For joint audio-video training, we normalize each modality separately so that its contribution is not set by its number of coordinates.

Interval-dependent loss weighting. Equation equation 8 gives the average-velocity regression objective. For SD3.5, we weight it by $1 + (t - r)^2$. This is equivalent to adding endpoint MSE with unit weight, since

$$\|x_t + (r - t)u_\theta - x_r\|_M^2 = (t - r)^2 \|u_\theta - u^\star\|_M^2. \tag{10}$$

LTX uses unweighted velocity MSE and records endpoint error only as a diagnostic. Joint audio-video training sums the separately normalized video and audio losses with equal weight.

Exact maps and the diagonal. For a well-posed velocity field, the endpoint $x_r = F(x_t, t, r)$ satisfies

$$x_r = x_t + \int_t^r v(x_s, s)\,\mathrm{d}s, \qquad u(x_t, t, r) = \frac{1}{t - r}\int_r^t v(x_s, s)\,\mathrm{d}s, \tag{11}$$

where x_s lies on the integrated trajectory through (x_t, t) and conditioning is fixed and omitted. For a continuous field, this average approaches $v(x_t, t)$ as r approaches t. At $r = t$, the student update returns its input regardless of its predicted velocity. This identity therefore does not enforce agreement with the teacher velocity on the diagonal. Training always uses $r < t$; numerical bank states approximate the exact endpoints above.

Conditioning branches. Each conditioning branch is a two-layer MLP (Linear–SiLU–Linear), with zero-initialized final weights and biases. SD3.5 uses Fourier features of $(t, r, r - t)$ and multiplies the interval branch by $r - t$. LTX uses its native sinusoidal time features for $(t, r, t - r)$ and a gate $t - r$. The two sign conventions parameterize the interval differently, but both branches vanish at $r = t$. Both guidance branches use features of $w - 1$ and a gate $w - 1$. SD3.5 adds the residuals to its shared time/text embedding before the existing modulation layers. LTX uses separate video and audio conditioners with their native hidden widths, adding residuals before the latent and, where present, prompt AdaLN projections. Fixed reference tokens retain their native conditioning. In the joint model, both conditioners receive the variable video CFG scale because the streams interact; audio guidance settings remain fixed in the teacher policy. The LTX bank also uses spatiotemporal guidance, modality guidance, and rescaling, all held fixed when varying video CFG.

Backbone adaptation. For a pretrained projection W, LoRA uses $W + \alpha BA/q$, with adapter rank q, scaling α, and trainable factors A, B. We initialize $B = 0$ and keep W frozen. SD3.5 adapts image and context attention query, key, value, and output projections. Joint LTX adapts video and audio self- and text-attention, both directions of cross-modal attention, and the video and audio feed-forward projections. Both backbones use $q = \alpha = 384$ and keep the native normalization layers and output heads frozen. The existing output head predicts velocity and is reused without modification. During image-conditioned generation, reference coordinates are held fixed and the interval update acts only on generated coordinates.

A.2 Analytic illustration

Figure 2 uses $x_0 \sim \sum_{k=1}^{3} \pi_k \mathcal{N}(\mu_k, a^2)$, with means $(-2.3, 0, 2.3)$, weights $(0.3, 0.4, 0.3)$, and $a = 0.32$, independently of $\epsilon \sim \mathcal{N}(0, 1)$. The interpolation marginal is a Gaussian mixture with means $(1-t)\mu_k$ and common variance $D_t = (1-t)^2 a^2 + t^2$. Given state x at time t, the posterior weight of mixture component k and the conditional mean of x_0 within that component are, respectively,

$$\rho_k(x,t) = \frac{\pi_k \exp\left[-\frac{(x-(1-t)\mu_k)^2}{2D_t}\right]}{\sum_j \pi_j \exp\left[-\frac{(x-(1-t)\mu_j)^2}{2D_t}\right]}, \quad m_k(x,t) = \mu_k + \frac{(1-t)a^2}{D_t}\left[x - (1-t)\mu_k\right]. \tag{12}$$

The analytic marginal velocity is

$$v^*(x,t) = \sum_k \rho_k(x,t)\left[\frac{t}{D_t}\left[x - (1-t)\mu_k\right] - m_k(x,t)\right]. \tag{13}$$

For panel (a), we choose $(t, x) = (0.84, -1.05)$ and $(0.48, 0.80)$. At each state, the conditional distribution of x_0 is the mixture with weights ρ_k, means m_k, and component variance $a^2 t^2 / D_t$. We select quantiles $(0.12, 0.80, 0.90)$ at the first state and $(0.03, 0.30, 0.94)$ at the second, then set $\epsilon = [x - (1-t)x_0]/t$. These choices give six separated interpolation paths, each passing through its selected state. The blue arrows show the analytic conditional mean over all compatible paths, including those not drawn. All arrows follow decreasing t. For panels (b,c), we integrate the analytic field with DOP853, using relative tolerance 10^{-11} and absolute tolerance 10^{-12}. The ten initial states are evenly spaced Gaussian quantiles between 0.035 and 0.965. Across 1,001 times, the maximum deviation from the conserved analytic marginal quantile is below 2.7×10^{-12}. Panel (c) retains trajectories 0, 2, 7, and 9 (ordered by initial state), with source time $t = 0.9$ and destination $r = 0.1$. Dashed lines extrapolate the source tangent to $x_t + (r-t)v(x_t, t)$, the one-step Euler estimate at the same destination time. Endpoint density silhouettes are independently rescaled for display and do not share a horizontal density scale.

A.3 Experimental details

A.3.1 Image generation

Trajectory bank and training. SD3.5-M is a 2.5-billion-parameter text-to-image model (Esser et al., 2024). We use 247,260 SD3.5-M trajectories: 244,760 for training and 2,500 for validation. The prompts comprise 217,309 DiffusionDB entries and 29,951 Pick-a-Pic entries (Wang et al., 2023; Kirstain et al., 2023). We merge the 200,000-prompt DiffusionDB selection used by DP-DMD with our earlier DiffusionDB/Pick-a-Pic bank, remove overlapping training prompts, and retain the earlier validation split. Normalized training and validation prompts are disjoint. The teacher generates 1024×1024 images with 28 Euler steps, storing the initial noise and all 28 updated states. Each prompt receives one CFG value from $\{1, 2.5, 4.5, 7\}$ with probabilities $(0.10, 0.20, 0.55, 0.15)$. Text conditioning is cached and the text encoders and VAE remain frozen.

The student uses rank-384 LoRA with scaling 384 and the conditioning branches described in Appendix A.1. We train for 190,000 AdamW updates with a global batch of 32, learning rate 1.5×10^{-4}, 500 warmup updates, and cosine decay to zero. Weight decay is 0.01 and the gradient norm is clipped at 1. Training uses bfloat16 model computation and float32 losses. Eight transition pairs are drawn per trajectory load; they are streamed into optimizer batches, rather than multiplying the stated global batch by eight. The selected checkpoint uses the final raw weights.

Evaluation. Teacher, PACT, and our DP-DMD reproduction share prompts and a deterministic seed mapping at 1024×1024. GenEval (Ghosh et al., 2023) contains 553 prompts with four seeds each; the COCO-10K quality panel (Lin et al., 2014) contains 10,000 captions with one seed each; the safe-for-work Pick-a-Pic (PaP) panel contains 1,024 user-written prompts with nine seeds each. The teacher uses CFG 4.5 and 28 denoising steps, requiring 56 conditional/unconditional transformer evaluations. The released four-step DP-DMD checkpoint (Wu et al., 2026) uses its native CFG 1.0. Our student uses learned CFG 4.5, without external CFG, so each denoising step requires one transformer evaluation. The main comparison uses four such evaluations. Additional DMD, DMD-LPIPS, and DMD-GAN values in Figure 4 come from Wu et al. (2026), rather than our reproductions; Appendix A.6 details their protocol differences. We report the number of transformer evaluations as NFE for SD3.5; it is distinct from wall-clock latency.

Quality metrics. GenEval checks whether requested objects, counts, colors, and spatial relations appear in the image. Its overall score is the unweighted mean of six category accuracies; image accuracy and the fraction of prompts with at least one correct image are reported separately. CLIP-Score measures text–image agreement (Hessel et al., 2021), while PickScore and ImageReward use learned human-preference predictors (Kirstain et al., 2023; Xu et al., 2023). We retain both the raw PickScore logit and the raw score divided by 26, with the latter labeled normalized PickScore in the tables. Edge strength sums the mean absolute horizontal and vertical differences in grayscale pixel values. It is a supplementary image statistic, not a quality score with an assumed optimum. FID measures the difference between generated and real-image feature distributions (Heusel et al., 2017); we compare 10,000 generated images with the 10,000 unique COCO-2014 validation images specified by DMD2's COCO-10K subset. Both sets are center-cropped and resized to 512×512 using Lanczos interpolation, then evaluated with DMD2's clean-fid Inception-v3 implementation (Yin et al., 2024a). This is FID-10K under this preprocessing protocol, rather than a directly interchangeable FID-50K estimate.

Diversity and teacher similarity. For diversity, we average pairwise cosine distance among images generated from the same prompt. A separate COCO-10K diversity evaluation uses nine matched seeds for each of the 10,000 captions, giving 90,000 images per model and 36 unordered pairs per prompt; the final score weights prompts equally. Paired teacher similarity compares images with the same prompt and initial-noise seed. We use the unprojected CLIP ViT-L/14 image CLS features (Radford et al., 2021) and DINOv3 ViT-B/16 (Siméoni et al., 2026), shared across all models. DINO patch similarity averages cosine similarity at corresponding 14×14 patch positions. Image-level standard errors in the detailed similarity tables describe the recorded sample averages; they do not account for dependence among seeds sharing a prompt.

Inference schedules. The four-step student uses noise levels $(1, 0.977903, 0.851349, 0.546666, 0)$. For the inference-compute sweep, schedules are selected using PACT only, on a four-prompt calibration panel, using a composite of perceptual loss and sharpness deficit. The selected schedules have high-noise powers 2.5, 2, and 1.5 at 6, 8, and 12 NFE, respectively; 16 NFE uses the native scheduler. The full arrays appear in Table 10. Applying these schedules to DP-DMD is an unoptimized extrapolation beyond its released four-step setting.

A.3.2 VIDEO GENERATION

Trajectory bank and training. The LTX-2.3 22B dev teacher (HaCohen et al., 2025) supplies 50,000 joint audio-video trajectories: 25,000 text-to-video prompts from VidProM (Wang & Yang, 2024) and 25,000 image-conditioned cases from TIP-I2V (Wang & Yang, 2025). Each task reserves 10% for validation, leaving 45,000 training trajectories in total. Candidate selection uses native LTX text embeddings, a 256-dimensional PCA projection, and 2,500 clusters per task; ten primary examples are selected per cluster from a 40,000-example candidate pool. Gemma-3-12B enhances the prompts, using the reference image for image-conditioned cases. Generation uses 768×512, 121 frames, 24 frames per second, and 30 Euler steps. Half of the bank uses the native schedule and half uses its jittered variant. Video CFG values $\{2, 2.5, 3, 3.5, 4\}$ have probabilities $(0.15, 0.20, 0.30, 0.20, 0.15)$. The teacher additionally uses spatiotemporal guidance 1, rescaling 0.7, and modality guidance 3, with perturbed block 28. Audio uses CFG 7 and the same fixed additional guidance settings. We store 31 aligned states for each modality in bfloat16.

The student uses rank-384 LoRA with scaling 384 and jointly learns audio and video transitions with equal modality loss weights. We train with AdamW, global batch 64, bfloat16 computation, activation checkpointing, and gradient norm clipping at 1. The first 20,000 updates use a peak learning rate of 2.5×10^{-5}, 250 warmup updates, and cosine decay to 2.5×10^{-6}. A further 10,000 updates preserve optimizer moments, warm up from 2.5×10^{-6} to 1.25×10^{-5} over 1,000 updates, then decay to 2.5×10^{-6}. Exponential moving averaging (EMA), with decay 0.999, starts at update 20,000. We use the 30,000-update raw checkpoint throughout the main comparisons and include EMA as an ablation. The 10,000- and 20,000-update runs have different learning-rate histories, so their comparison tests the longer training recipe rather than update count alone.

Models and generation. All comparisons generate joint audio and video at the bank resolution and clip length. They share the same enhanced prompt text and seeds, while VBench scores use the original benchmark captions. The teacher uses 30 steps with video CFG 3 and audio CFG 7, together with the additional guidance above. Our student uses eight steps and learned video CFG 3 without additional external guidance. Its noise levels are $(1, 0.981628, 0.957281, 0.923478, 0.873381, 0.791468, 0.633332, 0.2, 0)$. Official distilled full and LoRA checkpoints use eight steps and native unity guidance. The release provides weights and inference settings but does not specify the full LTX-2.3 distillation training recipe; we therefore treat these models as open-weight industrial references. The LoRA comparison uses strength 0.8 in a controlled single-stage pipeline; it is not the complete upstream two-stage production recipe. Because native teacher guidance evaluates multiple branches, we report denoising steps for LTX rather than infer transformer evaluations or measured speedups from the step ratio.

VBench evaluation. We evaluate text-to-video (T2V) with VBench (Huang et al., 2024) and image-to-video (I2V) with the image-conditioned benchmark in VBench++ (Huang et al., 2026). These benchmarks assess visual consistency, motion, appearance, and agreement with text or the input image. Text-to-video evaluation has 944 unique prompts, with five seeds per prompt except for 75 temporal-flickering prompts that use 25 seeds, giving 6,220 videos per model. Image-to-video evaluation has 1,118 prompt/image cases over 355 reference images, each with five seeds, giving 5,590 videos per model. T2V totals combine normalized quality and semantic groups with weights 0.8 and 0.2; I2V totals combine quality and image conditioning with equal weight. Dynamic degree has weight 0.5 within each quality group; camera motion has weight 0.1 within I2V conditioning. Individual dimensions below are multiplied by 100 before display, whereas aggregates use the official per-dimension normalization. I2V temporal flickering is supplementary and excluded from its official total. We do not average T2V and I2V into an additional benchmark score. These metrics assess visual content; they do not evaluate audio quality or audio-video synchronization. No confidence intervals are available for the VBench aggregate comparisons.

Visual diversity, similarity, and uncertainty. We reuse the five matched seeds for all 944 T2V and 1,118 I2V cases, excluding the extra temporal-flickering seeds so each case has equal weight. From each clip, we extract frames $0, 8, \ldots, 120$ and compute projected CLIP ViT-L/14 features (768 dimensions) and DINOv2-L CLS features (1,024 dimensions). We average the 16 raw frame features in float32, then normalize the pooled embedding; frames are not individually normalized first. Diversity averages cosine distance over the ten unordered seed pairs; teacher similarity averages the five matched-seed cosine similarities. Final scores weight cases equally. A sensitivity analysis omits frame zero before pooling; this removes the direct contribution of the conditioning frame but does not remove image conditioning from later frames.

We report pointwise 95% percentile intervals from 10,000 bootstrap resamples of prompt/image cases. Paired differences resample the same cases across models. The intervals condition on the generated seeds and fixed encoders, do not measure retraining variation, and are not adjusted for multiple comparisons. I2V cases sharing a reference image are not clustered together during resampling. These video features differ from the image-evaluation features, so their absolute scores should not be compared across backbones.

Unseen resolutions and aspect ratios. We keep the 30,000-update raw student and its eight-step schedule fixed at 576×384, 640×640, and 512×768, with a matched teacher at each shape. Clip length and frame rate are unchanged. The 768×512 reference reuses the earlier evaluation. Portrait I2V is a custom extension with centered exact 2:3 crops of the official square reference

images. Reference crops change across aspect ratios; therefore, teacher–student gaps within each shape are the meaningful comparisons, while absolute cross-shape changes also reflect different image conditioning.

A.3.3 COMPLETE IMAGE RESULTS

The tables report the four-step comparison and the inference-compute sweep from 4 to 16 NFE. Scores retain the metric scales defined above. PACT denotes the 190,000-update raw checkpoint; a dash marks an unmeasured value. Fractional scores use three decimals, while FID and raw PickScore use two. Inference schedules retain six decimals for reproducibility.

Table 1: GenEval image metrics at the reported budgets. †: DP-DMD extrapolation with schedules selected for our student, without baseline retuning. Pick raw is the PickScore logit; Pick norm. divides it by 26. Edge denotes grayscale edge strength. – indicates an unmeasured value.

Model	NFE	CLIPScore	Pick raw	Pick norm.	ImageReward	Edge	CLIP div.	DINO div.
Teacher	56	28.980	22.76	0.875	0.921	0.046	0.237	0.361
DP-DMD	4	28.909	22.60	0.869	0.968	0.038	0.206	0.313
PACT	4	28.434	22.21	0.854	0.666	0.038	0.245	0.376
PACT	6	28.573	22.47	0.864	0.803	0.039	0.239	0.365
PACT	8	28.701	22.53	0.867	0.824	0.040	0.238	0.363
PACT	12	28.639	22.57	0.868	0.865	0.041	0.243	0.367
PACT	16	28.610	22.63	0.870	0.911	0.039	0.240	0.361
DP-DMD†	6	28.352	22.40	0.862	0.693	0.020	0.253	0.339
DP-DMD†	8	28.193	22.36	0.860	0.667	0.022	0.259	0.346
DP-DMD†	12	28.241	22.43	0.863	0.737	0.026	0.254	0.343
DP-DMD†	16	27.989	22.28	0.857	0.651	0.024	0.263	0.351

Table 2: COCO-10K (FID: lower is better) image metrics at the reported budgets. †: DP-DMD extrapolation with schedules selected for our student, without baseline retuning. Pick raw is the PickScore logit; Pick norm. divides it by 26. Edge denotes grayscale edge strength. – indicates an unmeasured value.

Model	NFE	CLIPScore	Pick raw	Pick norm.	ImageReward	Edge	FID
Teacher	56	26.714	22.41	0.862	0.985	0.049	20.52
DP-DMD	4	26.544	22.31	0.858	0.988	0.045	25.45
PACT	4	26.383	21.93	0.844	0.753	0.044	20.60
PACT	6	26.440	22.17	0.853	0.882	0.049	20.21
PACT	8	26.576	22.22	0.855	0.885	0.048	21.73
PACT	12	26.526	22.26	0.856	0.911	0.046	21.24
PACT	16	26.437	22.29	0.857	0.939	0.045	22.43
DP-DMD†	6	26.337	22.27	0.856	0.841	0.033	27.83
DP-DMD†	8	26.326	22.26	0.856	0.825	0.034	28.12
DP-DMD†	12	26.331	22.27	0.856	0.848	0.038	27.72
DP-DMD†	16	26.257	22.15	0.852	0.778	0.037	29.68

Table 3: Pick-a-Pic image metrics at the reported budgets. †: DP-DMD extrapolation with schedules selected for our student, without baseline retuning. Pick raw is the PickScore logit; Pick norm. divides it by 26. Edge denotes grayscale edge strength. – indicates an unmeasured value.

Model	NFE	CLIPScore	Pick raw	Pick norm.	ImageReward	Edge	CLIP div.	DINO div.
Teacher	56	29.190	21.89	0.842	1.089	0.043	0.207	0.250
DP-DMD	4	28.419	21.77	0.837	1.128	0.040	0.181	0.212
PACT	4	28.545	21.63	0.832	0.921	0.037	0.212	0.251
PACT	6	28.670	21.80	0.838	1.030	0.040	0.208	0.246
PACT	8	28.752	21.74	0.836	1.029	0.039	0.213	0.245
PACT	12	28.791	21.79	0.838	1.054	0.039	0.216	0.247
PACT	16	28.715	21.75	0.837	1.087	0.038	0.214	0.248
DP-DMD†	6	27.764	21.54	0.828	0.969	0.029	0.206	0.238
DP-DMD†	8	27.705	21.51	0.827	0.963	0.031	0.209	0.244
DP-DMD†	12	27.712	21.51	0.827	0.979	0.035	0.211	0.247
DP-DMD†	16	27.550	21.39	0.823	0.929	0.034	0.216	0.251

Table 4: COCO-10K diversity from nine seeds per prompt (90,000 images per model). Values are mean ± prompt-level standard error over 10,000 prompts. Greater distance indicates more variation across seeds.

Model	CLIP diversity	DINOv3 diversity
Teacher	0.212 ± 0.001	0.310 ± 0.001
DP-DMD	0.180 ± 0.001	0.259 ± 0.001
PACT	0.226 ± 0.001	0.325 ± 0.001

Table 5: Detailed GenEval summary. Values are fractions; all are higher-is-better. NFE counts transformer evaluations, including both teacher CFG branches. † denotes unoptimized DP-DMD extrapolation.

Model	NFE	Overall	Image acc.	Prompt-any
Teacher	56	0.673	0.662	0.837
DP-DMD	4	0.649	0.638	0.823
PACT	4	0.622	0.609	0.808
PACT	6	0.652	0.640	0.826
PACT	8	0.658	0.646	0.837
PACT	12	0.659	0.646	0.825
PACT	16	0.664	0.652	0.839
DP-DMD†	6	0.605	0.592	0.799
DP-DMD†	8	0.594	0.581	0.792
DP-DMD†	12	0.583	0.571	0.781
DP-DMD†	16	0.552	0.539	0.769

Table 6: Detailed GenEval categories. Values are fractions; all are higher-is-better. NFE counts transformer evaluations, including both teacher CFG branches. † denotes unoptimized DP-DMD extrapolation.

Model	NFE	Single	Two	Count	Color	Position	Color attr.
Teacher	56	0.994	0.843	0.594	0.819	0.217	0.570
DP-DMD	4	0.997	0.823	0.544	0.798	0.258	0.475
PACT	4	0.991	0.712	0.556	0.811	0.188	0.475
PACT	6	0.988	0.785	0.594	0.811	0.182	0.550
PACT	8	0.991	0.823	0.591	0.801	0.198	0.545
PACT	12	0.988	0.828	0.609	0.811	0.190	0.525
PACT	16	0.991	0.859	0.613	0.806	0.198	0.517
DP-DMD†	6	0.959	0.760	0.544	0.750	0.237	0.378
DP-DMD†	8	0.944	0.735	0.550	0.718	0.240	0.378
DP-DMD†	12	0.944	0.742	0.537	0.691	0.225	0.360
DP-DMD†	16	0.912	0.674	0.519	0.665	0.242	0.300

Table 7: GenEval paired teacher similarity (higher is closer to the teacher). Values are mean ± image-level standard error; repeated seeds within a prompt are not clustered. Teacher self-similarity is one. † denotes unoptimized extrapolation.

Model	NFE	CLIP global	DINO global	DINO patch
DP-DMD	4	0.798 ± 0.002	0.695 ± 0.003	0.724 ± 0.002
PACT	4	0.813 ± 0.002	0.724 ± 0.003	0.775 ± 0.002
PACT	6	0.830 ± 0.002	0.746 ± 0.003	0.782 ± 0.002
PACT	8	0.837 ± 0.002	0.756 ± 0.003	0.786 ± 0.002
PACT	12	0.837 ± 0.002	0.757 ± 0.003	0.786 ± 0.002
PACT	16	0.835 ± 0.002	0.757 ± 0.003	0.784 ± 0.002
DP-DMD†	6	0.723 ± 0.003	0.623 ± 0.004	0.656 ± 0.002
DP-DMD†	8	0.714 ± 0.003	0.612 ± 0.004	0.654 ± 0.002
DP-DMD†	12	0.717 ± 0.003	0.615 ± 0.004	0.660 ± 0.002
DP-DMD†	16	0.699 ± 0.003	0.594 ± 0.004	0.646 ± 0.002

Table 8: COCO-10K paired teacher similarity (higher is closer to the teacher). Values are mean ± image-level standard error; repeated seeds within a prompt are not clustered. Teacher self-similarity is one. † denotes unoptimized extrapolation.

Model	NFE	CLIP global	DINO global	DINO patch
DP-DMD	4	0.803 ± 0.001	0.719 ± 0.001	0.738 ± 0.001
PACT	4	0.819 ± 0.001	0.751 ± 0.001	0.796 ± 0.001
PACT	6	0.838 ± 0.001	0.773 ± 0.001	0.804 ± 0.001
PACT	8	0.843 ± 0.001	0.783 ± 0.001	0.805 ± 0.001
PACT	12	0.844 ± 0.001	0.783 ± 0.001	0.805 ± 0.001
PACT	16	0.837 ± 0.001	0.776 ± 0.001	0.798 ± 0.001
DP-DMD†	6	0.760 ± 0.001	0.677 ± 0.002	0.694 ± 0.001
DP-DMD†	8	0.748 ± 0.001	0.664 ± 0.002	0.688 ± 0.001
DP-DMD†	12	0.743 ± 0.001	0.657 ± 0.002	0.688 ± 0.001
DP-DMD†	16	0.727 ± 0.001	0.642 ± 0.002	0.676 ± 0.001

Table 9: Pick-a-Pic paired teacher similarity (higher is closer to the teacher). Values are mean ± image-level standard error; repeated seeds within a prompt are not clustered. Teacher self-similarity is one. † denotes unoptimized extrapolation.

Model	NFE	CLIP global	DINO global	DINO patch
DP-DMD	4	0.805 ± 0.001	0.772 ± 0.001	0.755 ± 0.001
PACT	4	0.840 ± 0.001	0.819 ± 0.001	0.822 ± 0.001
PACT	6	0.853 ± 0.001	0.835 ± 0.001	0.827 ± 0.001
PACT	8	0.857 ± 0.001	0.842 ± 0.001	0.829 ± 0.001
PACT	12	0.859 ± 0.001	0.843 ± 0.001	0.829 ± 0.001
PACT	16	0.851 ± 0.001	0.835 ± 0.001	0.820 ± 0.001
DP-DMD†	6	0.758 ± 0.001	0.721 ± 0.002	0.708 ± 0.001
DP-DMD†	8	0.749 ± 0.001	0.709 ± 0.002	0.704 ± 0.001
DP-DMD†	12	0.745 ± 0.001	0.702 ± 0.002	0.704 ± 0.001
DP-DMD†	16	0.733 ± 0.001	0.690 ± 0.002	0.693 ± 0.001

Table 10: Selected SD3.5 student schedules, in descending noise level (rounded to six decimals). Higher-budget DP-DMD uses the same arrays only as an unoptimized extrapolation.

NFE	Schedule	Noise levels
4	Power 2.75	1.000000, 0.977903, 0.851349, 0.546666, 0.000000
6	high noise power 2.5	1.000000, 0.988660, 0.935850, 0.823223, 0.637113, 0.366062 0.000000
8	high noise power 2	1.000000, 0.984375, 0.937500, 0.859375, 0.750000, 0.609375 0.437500, 0.234375, 0.000000
12	high noise power 1.5	1.000000, 0.975944, 0.931959, 0.875000, 0.807550, 0.731043 0.646447, 0.554472, 0.455669, 0.350481, 0.239274, 0.122358 0.000000
16	native	1.000000, 0.976817, 0.951380, 0.923343, 0.892285, 0.857692 0.818923, 0.775174, 0.725417, 0.668327, 0.602151, 0.524533 0.432225, 0.320621, 0.182965, 0.008929, 0.000000

A.3.4 Complete video results

The main comparison uses PACT at 30,000 updates without EMA (30k raw). We also report earlier checkpoints, EMA, and an alternative 20k inference schedule. VBench tables use a 0–100 scale; feature scores retain their native cosine scale. In paired-difference tables, both differences and confidence bounds are multiplied by 1,000 to keep small changes readable.

Table 11: LTX VBench aggregates on a 0–100 scale, including training, EMA, and schedule comparisons. Official full/LoRA denote the released distilled baselines. The 20k old row changes only the inference schedule; the 10k/20k comparison also changes learning-rate history.

Model	Steps	T2V quality	T2V semantic	T2V total	I2V quality	I2V cond.	I2V total
Teacher	30	84.86	77.95	83.48	80.21	96.43	88.32
Official full	8	85.23	77.15	83.62	80.53	96.70	88.62
Official LoRA	8	85.06	77.75	83.60	80.98	96.53	88.76
PACT 20k	8	84.50	77.32	83.06	79.29	96.45	87.87
PACT 30k raw	8	84.59	77.56	83.19	79.50	96.59	88.04
PACT 30k EMA	8	84.69	77.60	83.28	79.45	96.48	87.97
PACT 10k	8	84.28	77.15	82.85	79.24	96.52	87.88
PACT 20k old	8	84.47	77.17	83.01	79.22	96.45	87.83

Table 12: T2V individual VBench dimensions, displayed as 100 times the native score. These columns are not the normalized inputs to the aggregate formula. Full/LoRA denote the official distilled baselines; numbered columns denote PACT training updates in thousands. All students use eight steps; the teacher uses 30.

Dimension	Teacher	Full	LoRA	20k	30k raw	30k EMA	10k	20k old
Aesthetic quality	64.83	65.41	65.37	64.34	64.29	64.37	64.30	64.43
Appearance style	20.85	21.05	21.07	21.39	21.45	21.50	21.34	21.38
Background consistency	96.28	95.97	95.88	96.08	96.18	96.16	95.99	96.10
Color	73.35	72.54	73.84	73.20	73.34	75.28	75.04	72.59
Dynamic degree	76.94	84.72	87.50	76.11	76.94	77.78	75.83	75.83
Human action	97.80	97.60	98.00	97.00	98.20	98.60	97.60	97.40
Imaging quality	67.39	66.30	65.06	67.47	67.47	67.61	67.28	67.35
Motion smoothness	98.71	98.81	98.73	98.48	98.47	98.47	98.40	98.48
Multiple objects	79.04	76.62	76.51	74.30	75.47	74.80	74.30	74.15
Object class	95.00	93.39	93.24	94.27	94.70	93.97	93.84	93.96
Overall consistency	26.65	26.33	26.51	26.76	26.78	26.77	26.61	26.78
Scene	53.92	51.47	52.17	51.56	52.15	51.51	51.74	51.57
Spatial relationship	76.27	77.44	79.47	78.00	76.43	76.42	74.94	76.99
Subject consistency	93.10	92.68	92.16	92.72	92.79	92.84	92.18	92.67
Temporal flickering	99.10	98.95	98.89	99.10	99.12	99.12	99.08	99.11
Temporal style	24.89	24.82	25.02	24.77	24.69	24.74	24.71	24.88

Table 13: I2V individual VBench dimensions, displayed as 100 times the native score. These columns are not the normalized inputs to the aggregate formula. Full/LoRA denote the official distilled baselines; numbered columns denote PACT training updates in thousands. All students use eight steps; the teacher uses 30. Temporal flickering is supplementary and excluded from the official I2V total.

Dimension	Teacher	Full	LoRA	20k	30k raw	30k EMA	10k	20k old
Aesthetic quality	63.38	63.05	62.62	62.51	62.85	62.81	62.65	62.44
Background consistency	96.97	96.97	96.52	96.55	96.84	96.79	96.61	96.56
Camera motion	57.06	59.21	60.81	61.91	61.13	60.84	62.62	61.65
Dynamic degree	36.26	39.92	50.41	32.76	32.03	32.03	31.87	32.36
I2V background	99.01	99.11	98.98	98.91	98.99	98.94	98.95	98.91
I2V subject	98.41	98.60	98.30	98.14	98.36	98.26	98.17	98.16
Imaging quality	71.68	71.73	71.20	71.40	71.44	71.49	71.39	71.37
Motion smoothness	98.81	98.93	98.93	98.58	98.66	98.64	98.55	98.58
Subject consistency	95.98	95.83	94.81	95.26	95.69	95.58	95.35	95.23
Temporal flickering	98.02	97.97	97.76	97.85	97.97	97.94	97.85	97.86

Table 14: CLIP visual feature scores using all 16 sampled frames. Brackets give pointwise 95% case-bootstrap intervals; see Appendix A.3.2 for the uncertainty protocol.

Task	Model	Diversity	95% CI	Teacher similarity	95% CI
T2V	Teacher	0.085	[0.083, 0.087]	1.000	[1.000, 1.000]
T2V	Official full	0.075	[0.073, 0.077]	0.925	[0.923, 0.927]
T2V	Official LoRA	0.072	[0.070, 0.074]	0.916	[0.914, 0.918]
T2V	PACT 20k	0.088	[0.086, 0.091]	0.933	[0.931, 0.935]
T2V	PACT 30k raw	0.088	[0.086, 0.090]	0.933	[0.931, 0.935]
T2V	PACT 30k EMA	0.087	[0.085, 0.089]	0.934	[0.932, 0.936]
I2V	Teacher	0.022	[0.021, 0.022]	1.000	[1.000, 1.000]
I2V	Official full	0.015	[0.015, 0.016]	0.983	[0.982, 0.983]
I2V	Official LoRA	0.016	[0.015, 0.016]	0.978	[0.977, 0.979]
I2V	PACT 20k	0.028	[0.027, 0.029]	0.977	[0.976, 0.978]
I2V	PACT 30k raw	0.026	[0.025, 0.027]	0.979	[0.978, 0.980]
I2V	PACT 30k EMA	0.026	[0.025, 0.027]	0.979	[0.978, 0.980]

Table 15: DINO visual feature scores using all 16 sampled frames. Brackets give pointwise 95% case-bootstrap intervals; see Appendix A.3.2 for the uncertainty protocol.

Task	Model	Diversity	95% CI	Teacher similarity	95% CI
T2V	Teacher	0.233	[0.226, 0.240]	1.000	[1.000, 1.000]
T2V	Official full	0.194	[0.188, 0.201]	0.806	[0.800, 0.812]
T2V	Official LoRA	0.187	[0.181, 0.193]	0.778	[0.771, 0.785]
T2V	PACT 20k	0.236	[0.229, 0.243]	0.831	[0.826, 0.836]
T2V	PACT 30k raw	0.231	[0.225, 0.238]	0.832	[0.826, 0.837]
T2V	PACT 30k EMA	0.232	[0.225, 0.238]	0.832	[0.827, 0.838]
I2V	Teacher	0.063	[0.059, 0.066]	1.000	[1.000, 1.000]
I2V	Official full	0.044	[0.042, 0.047]	0.952	[0.949, 0.955]
I2V	Official LoRA	0.049	[0.046, 0.052]	0.940	[0.937, 0.943]
I2V	PACT 20k	0.083	[0.079, 0.087]	0.938	[0.935, 0.942]
I2V	PACT 30k raw	0.075	[0.071, 0.080]	0.945	[0.942, 0.948]
I2V	PACT 30k EMA	0.077	[0.073, 0.081]	0.944	[0.941, 0.947]

Table 16: CLIP visual feature scores after excluding frame zero. Brackets give pointwise 95% case-bootstrap intervals; see Appendix A.3.2 for the uncertainty protocol.

Task	Model	Diversity	95% CI	Teacher similarity	95% CI
T2V	Teacher	0.086	[0.084, 0.088]	1.000	[1.000, 1.000]
T2V	Official full	0.076	[0.074, 0.078]	0.924	[0.922, 0.926]
T2V	Official LoRA	0.072	[0.070, 0.074]	0.915	[0.913, 0.917]
T2V	PACT 20k	0.089	[0.087, 0.092]	0.933	[0.931, 0.934]
T2V	PACT 30k raw	0.089	[0.087, 0.091]	0.932	[0.931, 0.934]
T2V	PACT 30k EMA	0.088	[0.086, 0.090]	0.933	[0.932, 0.935]
I2V	Teacher	0.024	[0.023, 0.025]	1.000	[1.000, 1.000]
I2V	Official full	0.017	[0.016, 0.018]	0.981	[0.980, 0.981]
I2V	Official LoRA	0.018	[0.017, 0.018]	0.975	[0.974, 0.976]
I2V	PACT 20k	0.031	[0.030, 0.032]	0.974	[0.973, 0.975]
I2V	PACT 30k raw	0.029	[0.028, 0.030]	0.977	[0.976, 0.978]
I2V	PACT 30k EMA	0.029	[0.028, 0.030]	0.977	[0.976, 0.978]

Table 17: DINO visual feature scores after excluding frame zero. Brackets give pointwise 95% case-bootstrap intervals; see Appendix A.3.2 for the uncertainty protocol.

Task	Model	Diversity	95% CI	Teacher similarity	95% CI
T2V	Teacher	0.235	[0.228, 0.242]	1.000	[1.000, 1.000]
T2V	Official full	0.196	[0.190, 0.202]	0.805	[0.798, 0.811]
T2V	Official LoRA	0.189	[0.183, 0.195]	0.776	[0.769, 0.783]
T2V	PACT 20k	0.238	[0.231, 0.245]	0.830	[0.824, 0.835]
T2V	PACT 30k raw	0.233	[0.226, 0.240]	0.830	[0.825, 0.836]
T2V	PACT 30k EMA	0.233	[0.227, 0.240]	0.831	[0.826, 0.836]
I2V	Teacher	0.071	[0.067, 0.075]	1.000	[1.000, 1.000]
I2V	Official full	0.050	[0.047, 0.053]	0.946	[0.943, 0.949]
I2V	Official LoRA	0.055	[0.052, 0.058]	0.932	[0.928, 0.936]
I2V	PACT 20k	0.094	[0.089, 0.099]	0.930	[0.927, 0.934]
I2V	PACT 30k raw	0.085	[0.081, 0.090]	0.938	[0.934, 0.941]
I2V	PACT 30k EMA	0.087	[0.082, 0.092]	0.937	[0.933, 0.940]

Table 18: CLIP paired differences. All differences and interval bounds are multiplied by 1,000. Differences subtract the second model from the first; positive values indicate greater diversity or teacher similarity. Brackets give pointwise 95% paired case-bootstrap intervals (Appendix A.3.2).

Task	Comparison	Δ div.	95% CI	Δ sim.	95% CI
T2V	PACT 30k raw − PACT 20k	-0.58	[-1.44, 0.28]	-0.15	[-0.73, 0.43]
I2V	PACT 30k raw − PACT 20k	-1.72	[-2.14, -1.30]	2.63	[2.27, 3.00]
T2V	PACT 30k EMA − PACT 20k	-1.11	[-1.91, -0.31]	0.74	[0.17, 1.30]
I2V	PACT 30k EMA − PACT 20k	-1.75	[-2.15, -1.34]	2.61	[2.27, 2.96]
T2V	PACT 30k raw − Teacher	2.82	[1.43, 4.23]	-66.89	[-68.74, -65.03]
I2V	PACT 30k raw − Teacher	4.33	[3.67, 4.97]	-20.68	[-21.59, -19.81]
T2V	PACT 30k EMA − Teacher	2.30	[0.93, 3.68]	-66.01	[-67.77, -64.23]
I2V	PACT 30k EMA − Teacher	4.29	[3.62, 4.95]	-20.70	[-21.60, -19.83]
T2V	Official full − Teacher	-10.23	[-11.57, -8.87]	-75.04	[-76.92, -73.08]
I2V	Official full − Teacher	-6.18	[-6.83, -5.53]	-17.34	[-18.09, -16.62]
T2V	Official LoRA − Teacher	-13.30	[-14.79, -11.85]	-84.46	[-86.46, -82.39]
I2V	Official LoRA − Teacher	-5.93	[-6.63, -5.25]	-21.84	[-22.72, -21.00]
T2V	PACT 30k EMA − PACT 30k raw	-0.52	[-1.11, 0.05]	0.89	[0.46, 1.31]
I2V	PACT 30k EMA − PACT 30k raw	-0.03	[-0.30, 0.23]	-0.02	[-0.20, 0.16]

Table 19: DINO paired differences. All differences and interval bounds are multiplied by 1,000. Differences subtract the second model from the first; positive values indicate greater diversity or teacher similarity. Brackets give pointwise 95% paired case-bootstrap intervals (Appendix A.3.2).

Task	Comparison	Δ div.	95% CI	Δ sim.	95% CI
T2V	PACT 30k raw − PACT 20k	-5.05	[-7.51, -2.67]	0.51	[-1.11, 2.17]
I2V	PACT 30k raw − PACT 20k	-7.50	[-8.94, -6.07]	6.27	[5.23, 7.30]
T2V	PACT 30k EMA − PACT 20k	-4.59	[-6.92, -2.22]	1.13	[-0.43, 2.73]
I2V	PACT 30k EMA − PACT 20k	-6.22	[-7.58, -4.84]	5.47	[4.52, 6.44]
T2V	PACT 30k raw − Teacher	-2.13	[-6.00, 1.78]	-168.37	[-173.77, -162.95]
I2V	PACT 30k raw − Teacher	12.69	[10.65, 14.69]	-55.24	[-58.48, -52.18]
T2V	PACT 30k EMA − Teacher	-1.67	[-5.56, 2.23]	-167.75	[-173.12, -162.40]
I2V	PACT 30k EMA − Teacher	13.97	[11.85, 16.09]	-56.04	[-59.34, -52.94]
T2V	Official full − Teacher	-38.84	[-42.96, -34.71]	-193.97	[-200.27, -187.81]
I2V	Official full − Teacher	-18.50	[-20.72, -16.36]	-47.71	[-50.58, -45.08]
T2V	Official LoRA − Teacher	-46.00	[-50.56, -41.54]	-222.42	[-229.38, -215.48]
I2V	Official LoRA − Teacher	-14.15	[-16.51, -11.86]	-60.08	[-63.37, -57.02]
T2V	PACT 30k EMA − PACT 30k raw	0.46	[-1.10, 2.02]	0.62	[-0.48, 1.71]
I2V	PACT 30k EMA − PACT 30k raw	1.28	[0.45, 2.14]	-0.80	[-1.39, -0.23]

Table 20: CLIP paired differences after excluding frame zero. All differences and interval bounds are multiplied by 1,000. Differences subtract the second model from the first; positive values indicate greater diversity or teacher similarity. Brackets give pointwise 95% paired case-bootstrap intervals (Appendix A.3.2).

Task	Comparison	Δ div.	95% CI	Δ sim.	95% CI
T2V	PACT 30k raw − PACT 20k	-0.60	[-1.46, 0.27]	-0.14	[-0.73, 0.44]
I2V	PACT 30k raw − PACT 20k	-1.96	[-2.44, -1.49]	2.95	[2.55, 3.38]
T2V	PACT 30k EMA − PACT 20k	-1.10	[-1.91, -0.28]	0.74	[0.18, 1.31]
I2V	PACT 30k EMA − PACT 20k	-1.98	[-2.43, -1.52]	2.94	[2.55, 3.33]
T2V	PACT 30k raw − Teacher	3.03	[1.64, 4.45]	-67.60	[-69.46, -65.73]
I2V	PACT 30k raw − Teacher	4.91	[4.18, 5.64]	-23.27	[-24.29, -22.29]
T2V	PACT 30k EMA − Teacher	2.53	[1.16, 3.92]	-66.71	[-68.48, -64.93]
I2V	PACT 30k EMA − Teacher	4.90	[4.14, 5.65]	-23.29	[-24.31, -22.31]
T2V	Official full − Teacher	-10.24	[-11.59, -8.88]	-75.76	[-77.65, -73.79]
I2V	Official full − Teacher	-6.95	[-7.68, -6.21]	-19.49	[-20.34, -18.68]
T2V	Official LoRA − Teacher	-13.37	[-14.87, -11.90]	-85.23	[-87.23, -83.14]
I2V	Official LoRA − Teacher	-6.66	[-7.46, -5.89]	-24.55	[-25.54, -23.61]
T2V	PACT 30k EMA − PACT 30k raw	-0.50	[-1.09, 0.09]	0.89	[0.46, 1.32]
I2V	PACT 30k EMA − PACT 30k raw	-0.02	[-0.32, 0.28]	-0.02	[-0.22, 0.19]

Table 21: DINO paired differences after excluding frame zero. All differences and interval bounds are multiplied by 1,000. Differences subtract the second model from the first; positive values indicate greater diversity or teacher similarity. Brackets give pointwise 95% paired case-bootstrap intervals (Appendix A.3.2).

Task	Comparison	Δ div.	95% CI	Δ sim.	95% CI
T2V	PACT 30k raw − PACT 20k	-5.04	[-7.51, -2.64]	0.55	[-1.08, 2.23]
I2V	PACT 30k raw − PACT 20k	-8.39	[-9.99, -6.78]	7.11	[5.93, 8.27]
T2V	PACT 30k EMA − PACT 20k	-4.60	[-6.96, -2.22]	1.22	[-0.35, 2.83]
I2V	PACT 30k EMA − PACT 20k	-6.98	[-8.50, -5.42]	6.21	[5.14, 7.31]
T2V	PACT 30k raw − Teacher	-1.76	[-5.67, 2.18]	-169.64	[-175.05, -164.18]
I2V	PACT 30k raw − Teacher	14.31	[12.02, 16.56]	-62.40	[-66.05, -58.97]
T2V	PACT 30k EMA − Teacher	-1.32	[-5.22, 2.62]	-168.97	[-174.37, -163.62]
I2V	PACT 30k EMA − Teacher	15.73	[13.34, 18.11]	-63.30	[-67.00, -59.83]
T2V	Official full − Teacher	-38.72	[-42.89, -34.54]	-195.23	[-201.55, -189.05]
I2V	Official full − Teacher	-20.89	[-23.40, -18.49]	-53.88	[-57.11, -50.91]
T2V	Official LoRA − Teacher	-46.01	[-50.58, -41.54]	-223.76	[-230.72, -216.85]
I2V	Official LoRA − Teacher	-16.02	[-18.68, -13.44]	-67.86	[-71.55, -64.38]
T2V	PACT 30k EMA − PACT 30k raw	0.44	[-1.13, 2.00]	0.67	[-0.43, 1.76]
I2V	PACT 30k EMA − PACT 30k raw	1.41	[0.47, 2.38]	-0.90	[-1.56, -0.25]

Table 22: LTX results at each spatial shape. Baseline 768×512 reuses the earlier evaluation; all other shapes are unseen in training. Compare models within each shape because I2V reference crops change across aspect ratios. Portrait I2V uses a custom centered crop extension.

Shape	Model	T2V quality	Semantic	T2V total	I2V quality	Conditioning	I2V total
768×512	Teacher	84.86	77.95	83.48	80.21	96.43	88.32
768×512	PACT	84.59	77.56	83.19	79.50	96.59	88.04
512×768	Teacher	84.99	76.76	83.34	79.98	95.87	87.93
512×768	PACT	84.17	77.12	82.76	79.13	96.06	87.59
576×384	Teacher	84.68	75.80	82.90	79.67	95.76	87.71
576×384	PACT	84.38	75.26	82.56	79.00	95.76	87.38
640×640	Teacher	84.13	77.87	82.87	79.83	96.41	88.12
640×640	PACT	83.98	77.21	82.63	79.21	96.70	87.96

Table 23: T2V dimensions by spatial shape. Each entry is 100 times the native score.

Dimension	768×512 Teacher	768×512 PACT	512×768 Teacher	512×768 PACT	576×384 Teacher	576×384 PACT	640×640 Teacher	640×640 PACT
Aesthetic quality	64.83	64.29	63.52	62.84	63.83	63.28	63.46	63.40
Appearance style	20.85	21.45	20.75	21.33	20.70	21.27	20.91	21.52
Background consistency	96.28	96.18	96.16	95.95	96.61	96.37	96.47	96.39
Color	73.35	73.34	73.34	75.68	75.66	72.73	73.86	71.31
Dynamic degree	76.94	76.94	80.83	78.89	78.89	78.89	73.33	74.44
Human action	97.80	98.20	98.00	98.60	97.40	98.40	98.20	98.00
Imaging quality	67.39	67.47	67.87	66.28	66.26	66.47	66.35	66.47
Motion smoothness	98.71	98.47	98.67	98.41	98.63	98.40	98.68	98.42
Multiple objects	79.04	75.47	78.99	74.95	75.53	71.22	81.08	76.43
Object class	95.00	94.70	95.09	94.02	92.82	94.13	93.05	92.82
Overall consistency	26.65	26.78	26.74	26.75	26.55	26.55	26.69	26.84
Scene	53.92	52.15	50.58	53.04	52.14	50.83	51.10	50.57
Spatial relationship	76.27	76.43	69.27	70.31	63.49	63.72	76.91	77.56
Subject consistency	93.10	92.79	93.19	92.23	93.04	92.55	92.42	91.90
Temporal flickering	99.10	99.12	99.08	99.15	99.07	99.10	99.14	99.15
Temporal style	24.89	24.69	24.94	24.73	24.95	24.74	25.13	24.85

Table 24: I2V dimensions by spatial shape. Each entry is 100 times the native score. Temporal flickering is supplementary and excluded from the official I2V total.

Dimension	768×512 Teacher	768×512 PACT	512×768 Teacher	512×768 PACT	576×384 Teacher	576×384 PACT	640×640 Teacher	640×640 PACT
Aesthetic quality	63.38	62.85	61.94	61.64	61.10	60.28	63.52	63.32
Background consistency	96.97	96.84	96.30	96.39	97.00	96.73	97.06	97.16
Camera motion	57.06	61.13	55.49	60.24	56.85	60.60	56.12	61.70
Dynamic degree	36.26	32.03	42.28	33.90	37.40	34.72	33.58	26.91
I2V background	99.01	98.99	98.77	98.85	98.46	98.27	98.90	98.97
I2V subject	98.41	98.36	97.82	97.66	97.86	97.76	98.59	98.55
Imaging quality	71.68	71.44	71.02	70.68	71.24	71.25	71.14	71.19
Motion smoothness	98.81	98.66	98.75	98.68	98.66	98.47	98.82	98.75
Subject consistency	95.98	95.69	95.08	95.28	95.67	95.24	95.55	95.71
Temporal flickering	98.02	97.97	97.87	98.03	97.84	97.71	98.03	98.11

A.4 SD3.5-M GENERATION LATENCY

We time 1024×1024 image generation at batch size one on a single NVIDIA GH200 GPU, using bfloat16 and resident model weights. The teacher uses 28 steps with CFG 4.5, batching conditional and unconditional predictions into 28 transformer calls of batch size two (56 single-example evaluations). The 190,000-update PACT checkpoint uses four batch-one calls with learned CFG 4.5; its LoRA weights are merged and its additional conditioner remains in float32. After five warmup generations, each configuration processes ten matched prompts, seeds, and initial latents. We synchronize CUDA at timing boundaries. Four independent single-GPU replicas share a node's CPU and storage resources.

Timing boundaries. Model time includes fresh tokenization and encoding with both CLIP encoders and T5, denoising, and VAE decoding. Text inputs use a maximum sequence length of 256, and the fixed empty negative embedding is cached once. PIL image postprocessing and PNG saving are measured separately and excluded from model time, as are loading, LoRA merging, initial noise creation, compilation, and warmup. Thus, model time is a steady-state measurement through decoded image generation, not latency through file saving.

We compare eager execution with `torch.compile` in `reduce-overhead` mode for the denoiser; text encoders and the VAE remain eager. No new compilation occurs during measured samples. The environment uses PyTorch 2.9.0a0, Diffusers 0.40.0, and PEFT 0.20.0. Table 25 reports timings and speedups for both execution modes. Encoding and decoding remain a fixed cost, so the reduction in total model time is smaller than the denoising speedup.

Table 25: **Single-GPU SD3.5-M generation latency.** Times are in seconds over ten matched cases per configuration. Model time includes text encoding, denoising, and VAE decoding; postprocessing and PNG saving are separate. SD is the sample standard deviation; P90 uses linear interpolation. Speedups are ratios of mean times, comparing the same execution mode unless stated otherwise.

(a) Mean time per stage

Model	Text	Denoise	VAE decode	Model total	Postproc.	PNG save
Teacher (eager)	0.050	2.460	0.091	2.601	0.006	0.282
PACT (eager)	0.049	0.215	0.088	0.353	0.006	0.295
Teacher (compiled)	0.055	1.561	0.094	1.710	0.006	0.283
PACT (compiled)	0.054	0.121	0.091	0.266	0.006	0.295

(b) Model-time variability

Model	Mean $\pm$ SD	Median	P90	Min	Max
Teacher (eager)	2.601 ± 0.003	2.600	2.604	2.598	2.609
PACT (eager)	0.353 ± 0.005	0.353	0.357	0.342	0.359
Teacher (compiled)	1.710 ± 0.004	1.709	1.715	1.703	1.716
PACT (compiled)	0.266 ± 0.003	0.265	0.271	0.263	0.271

(c) Speedups

Comparison	Denoising	Model time
Teacher / PACT (both eager)	$11.42\times$	$7.37\times$
Teacher / PACT (both compiled)	$12.90\times$	$6.43\times$
Teacher: eager / compiled	$1.58\times$	$1.52\times$
PACT: eager / compiled	$1.78\times$	$1.33\times$

A.5 LTX-2.3 GENERATION LATENCY

We measure how the reduction in sampling steps translates into generation time, including the work before and after denoising. Each sample is a 768×512 audiovisual clip with 121 frames at 24 frames per second (approximately five seconds) and 48 kHz audio. We use the same raw 30,000-update student as in the main quality comparison, with its LoRA weights merged for inference. The teacher takes 30 denoising steps with native guidance, requiring 120 transformer forward calls; the student takes eight steps and eight calls.

Timing protocol. Each model runs at batch size one on a single NVIDIA GH200 GPU in bfloat16, with its weights resident in memory. We evaluate both eager execution and `torch.compile` with the Inductor backend, using the same mode for teacher–student comparisons. Compilation applies to transformer blocks; text encoders, video/audio decoders, and the vocoder remain eager. The software environment uses PyTorch 2.9.0a0 and CUDA 13.0. For each task and configuration, five warmup generations precede ten measured generations with matched prompts, conditioning images where applicable, and seeds. We synchronize the GPU at timing boundaries and verify that no compilation occurs during measured generations. The four configurations run as independent single-GPU replicas on one node, sharing its CPU and storage resources.

Positive text encoding and I2V image encoding run afresh for each sample. The teacher's fixed negative prompt is encoded once; prompt enhancement is performed offline. *Model time* includes text/image encoding, denoising, and video/audio decoding, including the vocoder. *End-to-end time* additionally includes MP4 encoding, audio/video muxing, and remaining generation overhead. Neither includes model loading, warmup, compilation, or offline prompt enhancement. These are therefore steady-state generation times, rather than cold-start latency.

Table 26 reports the full timing breakdown and variation across samples. With compilation, both T2V and I2V complete in less than the clip's playback duration, including audio generation and file encoding. Table 27 separates the speedup in denoising from that of the complete pipeline: the fixed cost of encoding and decoding reduces the overall gain. Speedups are ratios of mean times; the reported standard deviations describe variation across the ten cases, not confidence intervals.

Table 26: **Single-GPU LTX-2.3 generation latency.** All times are seconds, measured over ten cases per task and configuration after warmup. E: eager; C: compiled transformer blocks. Model time includes input encoding and both decoders; end-to-end time also includes MP4 encoding/muxing. SD is the sample standard deviation, and P90 is the linearly interpolated 90th percentile.

(a) Mean time per stage

Task	Model	Text	Image	Denoise	Video dec.	Audio dec.	MP4
T2V	Teacher (E)	0.151	0.000	63.356	0.399	0.109	0.227
T2V	PACT (E)	0.143	0.000	4.447	0.395	0.107	0.242
T2V	Teacher (C)	0.134	0.000	51.446	0.399	0.103	0.226
T2V	PACT (C)	0.143	0.000	3.513	0.404	0.107	0.240
I2V	Teacher (E)	0.184	0.065	63.460	0.399	0.118	0.266
I2V	PACT (E)	0.145	0.064	4.450	0.395	0.108	0.275
I2V	Teacher (C)	0.136	0.062	51.386	0.399	0.104	0.265
I2V	PACT (C)	0.138	0.065	3.516	0.404	0.107	0.277

(b) Total latency and variability

		Model time			End-to-end		
Task	Model	Mean ± SD	Median	P90	Mean ± SD	Median	P90
T2V	Teacher (E)	64.014 ± 0.037	64.003	64.060	64.242 ± 0.036	64.241	64.285
T2V	PACT (E)	5.093 ± 0.006	5.094	5.098	5.337 ± 0.036	5.336	5.369
T2V	Teacher (C)	52.082 ± 0.053	52.083	52.133	52.309 ± 0.069	52.294	52.412
T2V	PACT (C)	4.167 ± 0.006	4.166	4.175	4.408 ± 0.033	4.411	4.439
I2V	Teacher (E)	64.226 ± 0.200	64.161	64.434	64.494 ± 0.218	64.426	64.646
I2V	PACT (E)	5.162 ± 0.027	5.164	5.183	5.438 ± 0.057	5.449	5.510
I2V	Teacher (C)	52.086 ± 0.078	52.103	52.174	52.353 ± 0.081	52.360	52.458
I2V	PACT (C)	4.230 ± 0.027	4.231	4.255	4.507 ± 0.048	4.510	4.568

T2V has no image conditioning; its image column rounds a negligible dispatch overhead to zero. Denoising includes latent setup and scheduling.

Table 27: **LTX-2.3 speedups from distillation and compilation.** Each entry is the ratio of mean generation times for the stated comparison. Teacher–student ratios compare the same execution mode. Compilation benefits both models; the real-time result uses compiled inference.

Task	Comparison	Denoising	Model time	End-to-end
T2V	Teacher / PACT (both eager)	14.25×	12.57×	12.04×
T2V	Teacher / PACT (both compiled)	14.64×	12.50×	11.87×
T2V	Teacher: eager / compiled	1.23×	1.23×	1.23×
T2V	PACT: eager / compiled	1.27×	1.22×	1.21×
I2V	Teacher / PACT (both eager)	14.26×	12.44×	11.86×
I2V	Teacher / PACT (both compiled)	14.62×	12.31×	11.61×
I2V	Teacher: eager / compiled	1.23×	1.23×	1.23×
I2V	PACT: eager / compiled	1.27×	1.22×	1.21×

A.6 Additional published SD3.5-M baselines

Table 28 records four-step SD3.5-M results reported by Wu et al. (2026), including DMD, DMD-LPIPS, and DMD-GAN. Figure 4 includes these reported values as starred baselines alongside our local evaluations. The published teacher uses 30 steps and CFG 3.5, whereas ours uses 28 steps and CFG 4.5. Published DINO diversity uses ViT-Large, rather than our ViT-Base encoder. The released Pick-a-Pic and COCO prompt lists match our lists, but the exact table seeds and mapping to released checkpoints are not specified. The released preference-evaluation example uses one sample per prompt, while our Pick-a-Pic preference scores average nine. Table 29 compares the reported and reproduced DP-DMD results where metric definitions align; the comparison checks agreement under these protocol differences.

Table 28: Published SD3.5-M results from DP-DMD Table 1. All student rows use four evaluations; the published teacher uses 60 evaluations (30 steps with CFG). These are the source paper's measurements, not our evaluations. DINO uses ViT-Large. PickScore is the raw logit; VQ-R1 denotes VisualQuality-R1, a learned visual-quality score, and MANIQA assesses image quality without a reference image. Reported and local measurements use different protocols, as described above.

Benchmark	Model	CLIP div.	DINO div.	ImageReward	PickScore	VQ-R1	MANIQA
Pick-a-Pic	SD3.5-M teacher	0.221	0.240	1.007	21.80	4.657	1.020
Pick-a-Pic	DMD	0.133	0.137	1.189	21.75	4.649	1.016
Pick-a-Pic	DMD-LPIPS	0.169	0.169	1.063	21.62	4.598	1.005
Pick-a-Pic	DMD-GAN	0.162	0.183	1.033	21.63	4.525	0.984
Pick-a-Pic	DP-DMD	0.182	0.179	1.142	21.76	4.646	1.017
COCO-10K	SD3.5-M teacher	0.204	0.288	0.910	22.31	4.636	1.043
COCO-10K	DMD	0.154	0.210	1.053	22.40	4.690	1.060
COCO-10K	DMD-LPIPS	0.168	0.204	0.949	22.29	4.599	1.012
COCO-10K	DMD-GAN	0.174	0.214	0.751	22.02	4.584	0.983
COCO-10K	DP-DMD	0.182	0.250	0.988	22.41	4.689	1.032

Table 29: DP-DMD: reported scores and our reproduction with the released natural checkpoint. The protocols differ as described above. DINO diversity is omitted here because the encoder differs.

Benchmark	Metric	Reported	Reproduced
Pick-a-Pic	ImageReward	1.142	1.128
Pick-a-Pic	PickScore	21.76	21.77
Pick-a-Pic	CLIP diversity	0.182	0.181
COCO-10K	ImageReward	0.988	0.988
COCO-10K	PickScore	22.41	22.31
GenEval	Overall	0.650	0.649

Table 30: Published GenEval results from DP-DMD Table 5. This table uses the original paper's rounded values and teacher protocol; the other DMD variants have no reported GenEval scores in that table.

Model	Overall	Single	Two	Counting	Colors	Position	Attribution
Teacher	0.66	0.98	0.78	0.64	0.83	0.18	0.53
DP-DMD	0.65	0.99	0.81	0.60	0.81	0.21	0.48

A.7 SD3.5 QUALITATIVE EXAMPLES

Figures 8–10 extend the selected examples in Figure 1, illustrating variation across seeds, teacher correspondence, and the effect of additional inference compute.

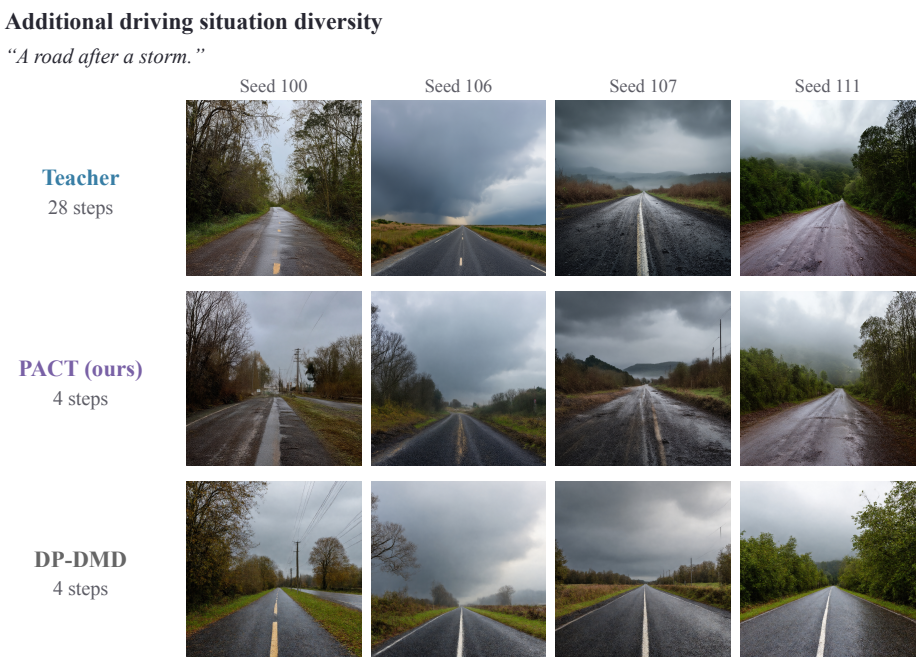


Figure 8: **Diversity across seeds.** Teacher, PACT, and DP-DMD generations for the same prompt. Each column shares a seed across models; the teacher uses 28 steps and both students use four.

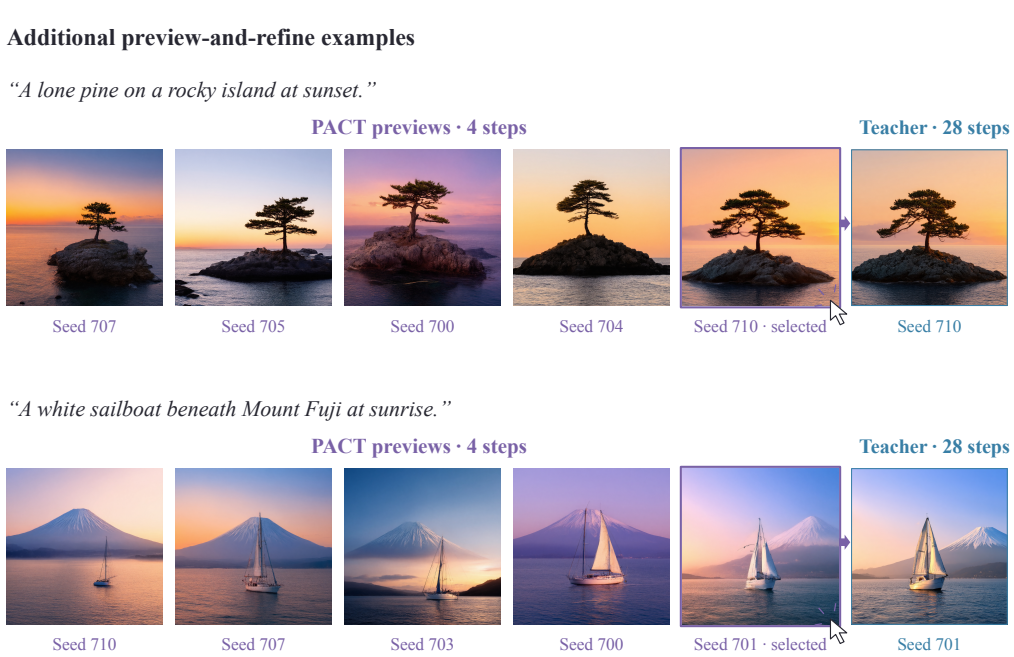


Figure 9: **Preview and refine.** Each row shows five four-step PACT previews and a 28-step teacher generation using the selected preview's prompt and seed. Purple frames mark the selections; the teacher generates from matched initial noise, without taking the preview image as input.

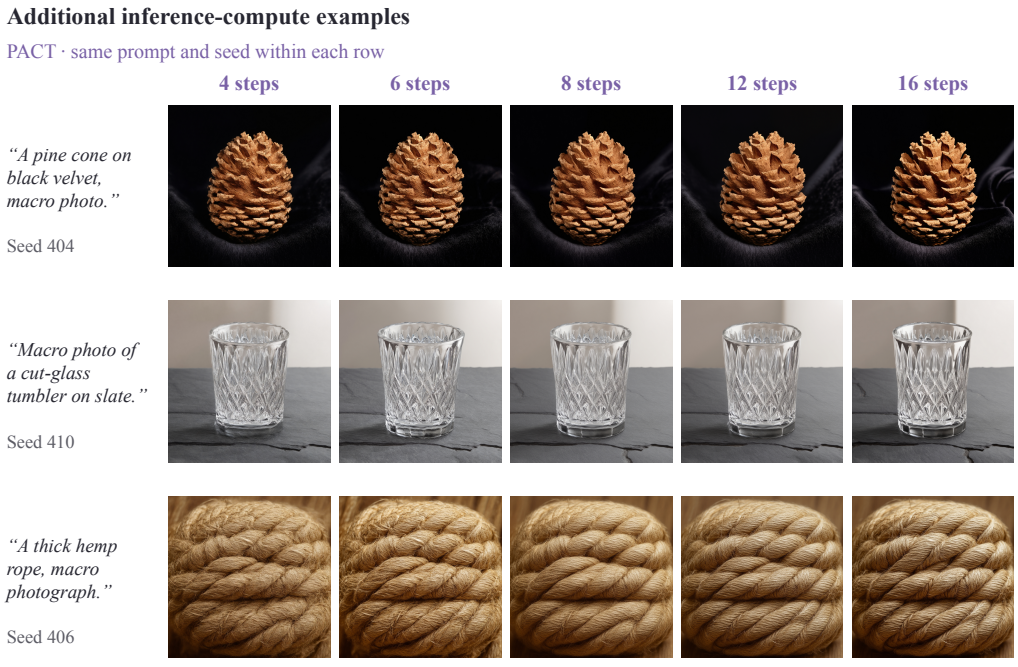


Figure 10: **More inference compute.** PACT generations with 4, 6, 8, 12, and 16 steps using the optimized schedules. Each row fixes the prompt and seed, illustrating how local detail changes with the sampling budget.