StateMix: Complementary State Sources for Few-Step Diffusion Distillation

Anonymous authors
Paper under double-blind review



Figure 1: High-quality 4-step text-to-image generations by StateMix, distilled from Qwen-Image.

Abstract

Few-step diffusion distillation has advanced through a variety of training objectives. Beyond the objective itself, the states on which a student is trained also shape the distillation process. We introduce noise-start states, sampled directly from a fixed Gaussian prior at intermediate target times, as a new training-state source for distribution matching. Controlled toy experiments reveal complementary learning behaviors: noise-start training strengthens responses to changed conditions, whereas rollout-start training on states produced by the student's preceding denoising steps transmits more input-state variation. Building on this complementarity, we propose StateMix, a mixed-state training strategy that combines both sources under the same DMD2 supervision without changing inference. On text-conditioned SDXL and Qwen-Image, StateMix achieves leading human preference scores among the evaluated methods, with the mixture ratio allowing a trade-off between preference scores and similarity to the teacher's generated distribution. We further evaluate StateMix on Qwen-Image-Edit with both text and image conditions, where public benchmarks demonstrate its effectiveness. In human GSB evaluation, the student meets all eight application-specific quality criteria relative to the teacher, with no loss under the production acceptance standard and even surpassing the teacher on most evaluation dimensions.

1 Introduction

Diffusion models have made substantial progress in text-to-image generation and image editing, as demonstrated by recent models such as FLUX.2 and Qwen-Image (Black Forest Labs, 2025; Wu et al., 2025). However, generating an image typically requires many denoising steps, each

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.

involving a neural network evaluation. As model sizes grow, this repeated computation substantially increases inference time and resource requirements. To address this issue, diffusion distillation trains few-step generators that retain the capabilities of a pretrained teacher. Existing approaches include consistency distillation, such as PCM (Song et al., 2023; Luo et al., 2023; Wang et al., 2024), adversarial distillation (Sauer et al., 2024), and progressive distillation (Salimans & Ho, 2022). Distribution Matching Distillation (DMD) instead learns to align the student's generated distribution with the teacher's distribution (Yin et al., 2024b;a). The DMD family has produced strong few-step generators on SDXL and has recently been applied to larger models: Qwen-Image-Flash uses DMD2 to distill Qwen-Image into a 4-step generator (NVIDIA, 2026).

Beyond the supervision objective, the training states used for distribution matching also shape the distillation process. Existing few-step DMD2 trains on intermediate states produced by the student's preceding denoising steps, which we call rollout-start states. These states carry content already generated by the student, anchoring distribution matching to its own generation trajectories. We introduce a new training-state source: noise-start states sampled directly from a fixed Gaussian prior at intermediate target times, in place of rollout-start states. We reveal complementary learning behaviors induced by the two state sources under the same distribution matching objective, and propose **StateMix**, a mixed-state training strategy that combines the newly introduced noise-start states with the original rollout-start states while retaining the DMD2 supervision and few-step inference procedure.

This complementarity manifests in conditional response and the transmission of input-state variation. A controlled toy study shows that training on noise-start states yields stronger responses to changed conditions, whereas training on rollout-start states transmits more variation from the input state. StateMix improves the weaker response of each single-source training scheme, while exhibiting a trade-off in distribution distance under ordinary generation. These findings show that mixing the two state sources changes the student's learning behavior, but its impact on practical generation quality requires evaluation on real-image tasks.

To this end, we first evaluate StateMix on the text-conditioned generation models SDXL and Qwen-Image, where experiments demonstrate leading performance on human preference metrics. We further evaluate StateMix on Qwen-Image-Edit, which uses both text and image conditions and more closely reflects real-world application needs. Human GSB evaluation shows that StateMix achieves performance comparable to or better than the teacher across all eight evaluation dimensions, meeting the application's production quality requirements.

2 PRELIMINARIES

Distribution Matching Distillation (DMD) trains a student generator F_θ to reproduce a pretrained teacher's output distribution in one or a few sampling steps (Yin et al., 2024b;a). Training uses two diffusion models: a frozen teacher provides the *real score* s_T, while an auxiliary *fake-score* model s_ϕ learns the distribution of the student's generated samples. We follow the DMD2 training framework (Yin et al., 2024a). The fake model is used for training, not for inference.

Given condition c and a prepared input z_k, the student first predicts a clean sample. We then add noise so that both score models can evaluate it:

$$y = F_\theta(z_k, t_k, c), \qquad x_\tau = \alpha_\tau y + \sigma_\tau \xi, \qquad \xi \sim \mathcal{N}(0, I). \tag{1}$$

Here t_k selects the student's generation step, whereas τ selects the noise level for score evaluation. Both score models receive the same x_τ and condition. For the basic distribution-matching component, their difference gives the student update:

$$\nabla_\theta \mathcal{L}_{\mathrm{DM}} = \mathbb{E}\left[w(\tau)\alpha_\tau \left(\frac{\partial y}{\partial \theta}\right)^\top \big(s_\phi(x_\tau, \tau, c) - s_T(x_\tau, \tau, c)\big) \right]. \tag{2}$$

This score difference matches distributions rather than requiring a paired target image; $w(\tau)$ weights noise levels. Training alternates between updating the student with both score models fixed and fitting the fake model to detached student samples. The teacher remains frozen.

For few-step training, earlier student steps prepare z_k through rollout or backward simulation (Yin et al., 2024a). This prefix is detached (stop-gradient, denoted sg); only the selected forward receives

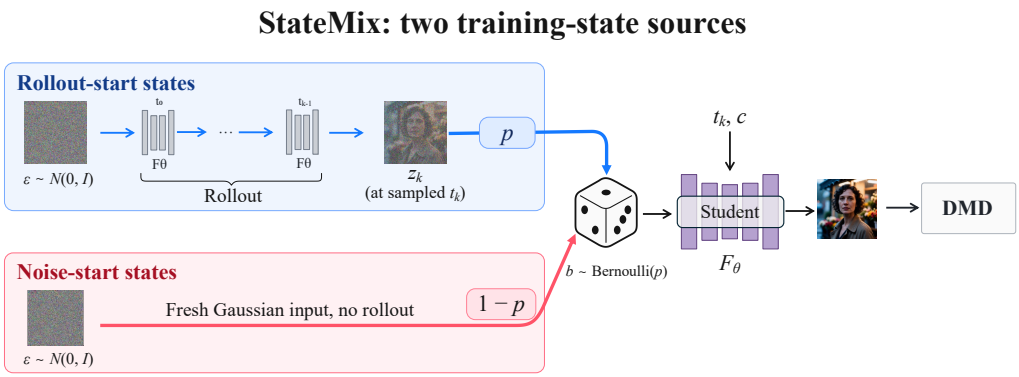

Figure 2: StateMix selects a noise-start state with probability $1 - p$ or a detached rollout-start state with probability p, using one shared predictor at target time t_k. Score supervision uses a separate time τ. Only the selected path is evaluated; inference is unchanged. Image tiles are illustrations, not experimental outputs.

gradients. StateMix changes this input preparation. Further details on DMD2 and its proxy-loss implementation are given in Appendix A.5.

3 StateMix: Complementary State Sources

StateMix exploits complementary learning behaviors through mixed-state training, combining our newly introduced noise-start states with the original rollout-start states under the same DMD2 supervision (Figure 2).

3.1 Training-State Sources and Selection

Let $q(k)$ sample one of K target times. For condition c and independent $\epsilon, \epsilon' \sim \mathcal{N}(0, I)$, we define the rollout-start state z_k^{roll} and the noise-start state z_k^{noise} as

$$z_k^{\text{roll}} = \text{sg}[R_{\theta,<k}(\epsilon, c)], \qquad z_k^{\text{noise}} = \epsilon'. \tag{3}$$

Rollout-start states inherit the student's preceding denoising steps; noise-start states are drawn directly from the fixed Gaussian prior, including at intermediate times. The sources coincide at $k = 0$. At $k > 0$, noise-start states differ from forward-noised clean inputs; sampling multiple target times also differs from 1-step DMD's fixed generator time.

A draw $b \sim \text{Bernoulli}(p)$ selects the input and prediction:

$$z_k = \begin{cases} z_k^{\text{roll}}, & b = 1, \\ z_k^{\text{noise}}, & b = 0, \end{cases} \qquad y = F_\theta(z_k, t_k, c). \tag{4}$$

Thus $p = 0$ and $p = 1$ recover the pure noise-start and rollout-start recipes. Only one source is evaluated per update, with routing synchronized across workers. Writing $\nu = \mathcal{N}(0, I)$ and holding the input sampler's parameters $\bar{\theta}$ fixed during differentiation gives

$$\mu^p_{\bar{\theta},k}(\cdot \mid c) = p\, \mu^{\text{roll}}_{\bar{\theta},k}(\cdot \mid c) + (1 - p)\nu. \tag{5}$$

Here $\bar{\theta}$ is a detached view of the current student. Distribution matching supervises the resulting predictions without requiring the inputs to follow the diffusion marginal at t_k (Appendix A.5). Algorithm 1 gives the update procedure.

3.2 What Distribution is Matched?

For a fixed input sampler $\bar{\theta}$, let $Q^0_{k,r}$ be the law of $F_\theta(z_k, t_k, c)$ under source $r \in \{\text{roll}, \text{noise}\}$. These are *training clean predictions*. Their mixture over (k, r) is generally different from P^{infer}_θ, the output

law of the complete K-step sampler. Re-noising training predictions yields the distribution learned by the fake-score model.

At a fixed condition and score time, index the re-noised component laws by $i = (k, r)$ and write $Q = \sum_i a_i Q_i$. The weights a_i are the conditional probabilities of these groups; they account for any dependence of score-time sampling on the group. A fake score receiving (x_τ, τ, c) without (k, r) fits this pooled law. For a common teacher distribution P_T, the following identity holds when the KL terms are finite:

$$\sum_i a_i D_{\mathrm{KL}}(Q_i \| P_T) = D_{\mathrm{KL}}(Q \| P_T) + \sum_i a_i D_{\mathrm{KL}}(Q_i \| Q). \tag{6}$$

The last term is the weighted Jensen–Shannon divergence. Consequently, pooled matching permits distinct component distributions and does not establish separate matching at each source/time or optimal final inference. Appendix A.5 gives the proof and the score-estimation assumptions for this idealized DM interpretation.

The corresponding local gradient decomposition and detached-prefix analysis are given in Appendix A.5.

3.3 Inference and training cost

State mixing is used only during training; inference retains the original K-step sampler and model architecture. Noise-start training avoids the preceding rollout, while rollout-start training requires k additional student forwards at target index k. The expected number of student forwards needed to construct one training prediction is therefore $1 + p\sum_k q(k)k$, or $1 + p(K-1)/2$ for uniform k. This counts only these student forwards, not the total training cost. Detailed training-cost analysis is provided in Appendix C.2.

4 Toy study: condition response and state transmission

We construct a controlled conditional generation task in which the student maps Gaussian noise to the target distribution specified by a condition, to examine how training-state sources affect conditional response, input-variation transmission and distribution matching. Specifically, we use a six-mode conditional Gaussian mixture and compare noise-start training, rollout-start training and their mixture (StateMix, $p = 0.5$) under the same MLP architecture and DMD2 training recipe. The condition specifies the target mode; condition switches and within-mode perturbations probe conditional response and input-variation transmission, respectively. Full protocols are provided in Appendix A.

Following a changed condition. Figure 3(a) feeds all students the same intermediate point cloud from A, but requests B. The student trained on noise-start states produces orange outputs near B's blue contours; training on rollout-start states keeps outputs closer to A (gray dashes). StateMix shifts farther toward B. This tests whether the new condition overrides shared input content.

Retaining input variation. With the condition fixed in (b), circles and crosses track negative and positive changes along the horizontal within-mode axis. The student trained on noise-start states nearly merges paired outputs into a compact cloud, suppressing variation. The student trained on rollout-start states preserves more pair separation; StateMix retains an intermediate amount. Pair separation measures transmission of the controlled change, rather than cloud width alone.

Complementarity with a distributional cost. Panel (c) measures condition-switch success: the percentage of outputs assigned to target mode B after changing the condition from A to B with fixed input states. Higher values indicate stronger conditional response. Panel (d) measures tangential response gain: output separation under paired input perturbations, normalized by the imposed clean-space perturbation scale. Near-zero gain indicates suppressed input variation. Across five training seeds, StateMix strengthens conditional response relative to rollout-start training and variation transmission relative to noise-start training.

Panel (e) measures conditional sliced W_2^2 distance to the true target distribution under ordinary 4-step generation (lower is better). Rollout-start training achieves the lowest distance, with StateMix

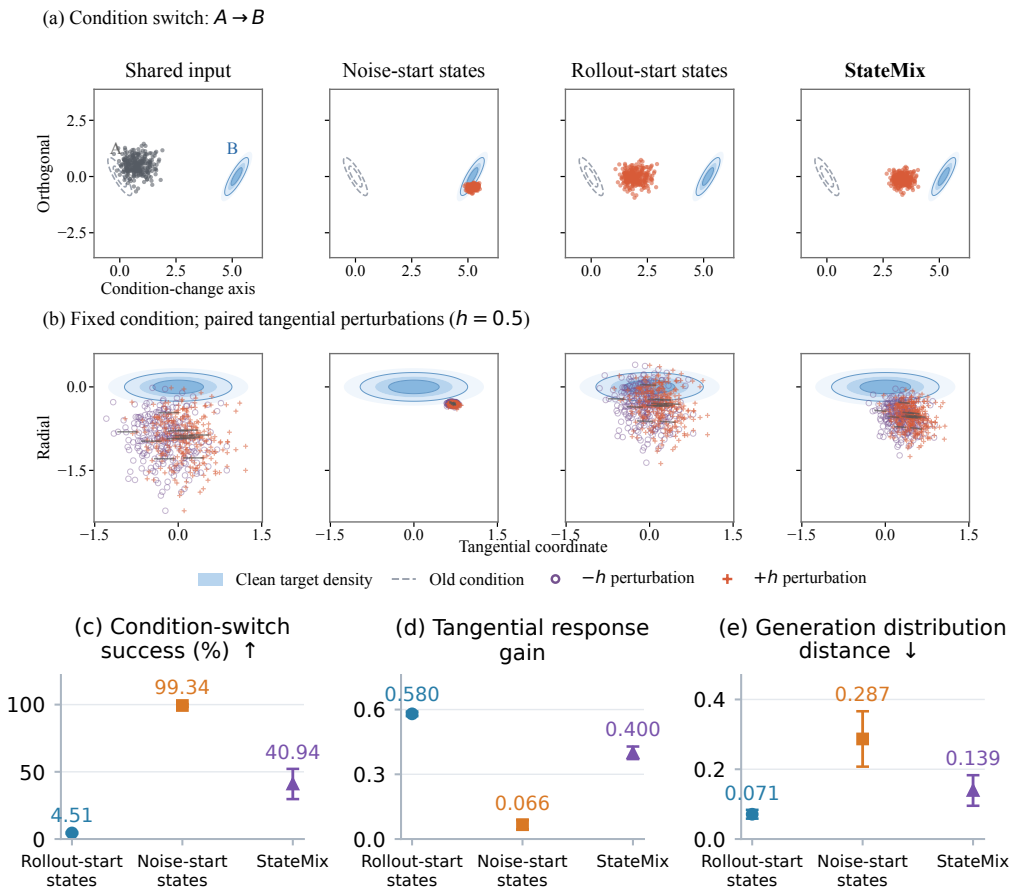

Figure 3: Complementary condition response (a,c) and variation transmission (b,d), with a distributional cost (e). Gray points are shared inputs; blue contours are clean target densities. Model labels identify their training sources. Interventions use $k = 3$ (1 step remaining), with $h = 0.5$ for variation. Scatter plots show one case; panels (c,d) report condition-switch success and tangential response gain, respectively. Panel (e) reports conditional sliced W_2^2 distance for ordinary 4-step generation (lower is better). Error bars show one standard deviation across five training seeds.

between the two pure-source schemes. Thus, improved response balance comes with a trade-off in distribution matching. Additional intervention results and ordinary-generation analyses are provided in Appendices A.2 and A.3.

5 Experiments

The toy study in Section 4 reveals complementary conditional response and input-variation transmission from noise-start and rollout-start training, alongside a trade-off in distribution matching. To further examine mixed-state training on real-world generative models, we first evaluate text-conditioned generation with SDXL and Qwen-Image, then extend the evaluation to Qwen-Image-Edit with both text and image conditions, which more closely reflects practical application needs. Through public benchmarks, qualitative comparisons and human evaluation, we assess StateMix in terms of generation quality, condition adherence and content preservation.

5.1 Text-to-image generation (T2I)

Experimental setup. We distill SDXL (Podell et al., 2024) and Qwen-Image (Wu et al., 2025) for 4-step generation. SDXL comparisons include Flash Diffusion (Chadebec et al., 2025), PCM (Wang et al., 2024), LCM (Luo et al., 2023), SenseFlow (Ge et al., 2025) and DMD2 (Yin et al., 2024a). Qwen-Image comparisons use FastGen's DMD2 implementation and PiFlow (Chen et al., 2025).

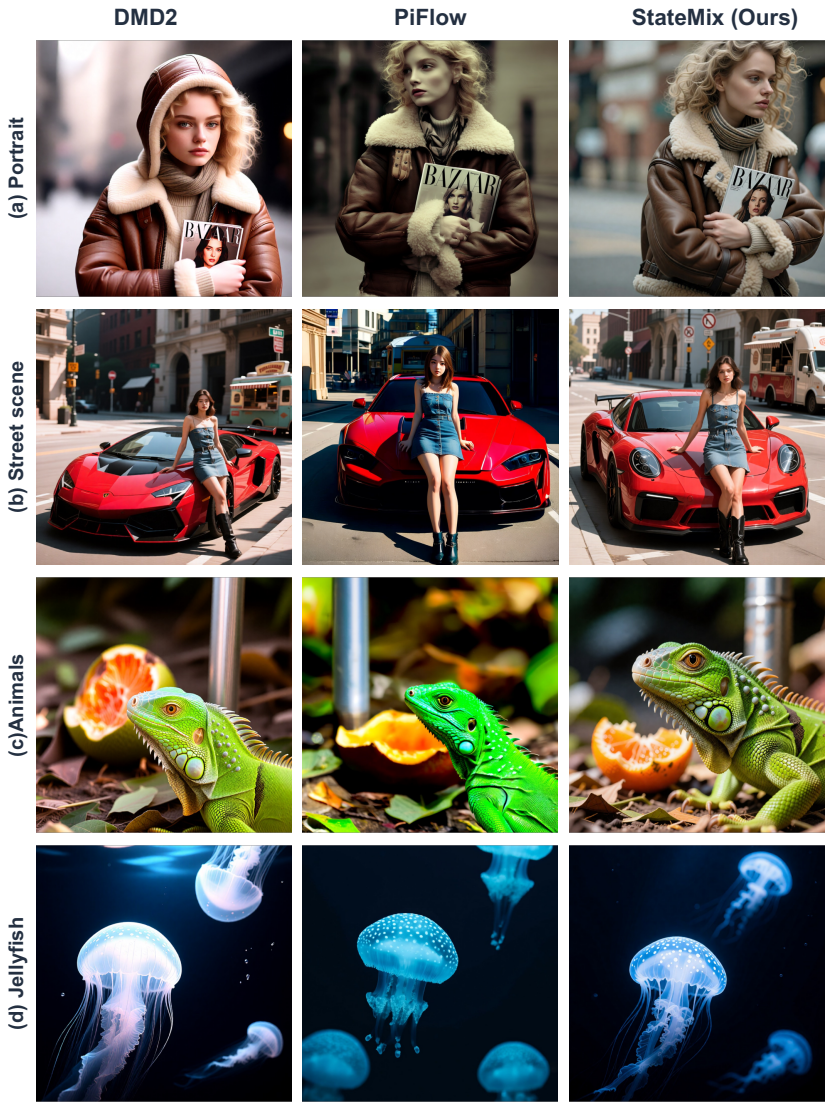

Figure 4: Qualitative comparison of 4-step Qwen-Image on HPSv3 prompts. StateMix produces more natural colors and more realistic textures in these examples.

HPSv2.1 (Wu et al., 2023), ImageReward (Xu et al., 2023) and HPSv3 (Ma et al., 2025) are automatic preference estimators. CLIP (Radford et al., 2021) measures image–text alignment. On HPSv2 prompts, FID (Parmar et al., 2022) measures the distributional distance between student- and teacher-generated images using global feature statistics (FID-T); lower values indicate closer agreement with the teacher's generated distribution. HPSv3 uses a separate prompt dataset and new generations. Full protocols are in Appendix B.

Preference gains and teacher-distribution similarity. StateMix leads the evaluated methods on HPSv2.1, ImageReward and HPSv3 for both backbones. On Qwen-Image, for example, HPSv2.1 increases from DMD2's 31.42 to 33.16, and ImageReward from 1.329 to 1.415. The agreement of preference estimators supports a practical benefit across the two backbones. FID and CLIP favor other methods: PiFlow retains the best Qwen-Image values, and SDXL also has different leaders. These differences show that preference scores and teacher-distribution similarity measure distinct aspects of the output. Category-wise preference scores are provided in Appendix D.1.

Qualitative comparison. Figure 4 presents qualitative comparisons across portrait, street-scene, animal and jellyfish prompts. Compared with DMD2 and PiFlow, StateMix produces more natural

Table 1: 4-step SDXL and Qwen-Image results on HPSv2 and HPSv3. StateMix leads the displayed methods on the three automatic preference estimators for both backbones. FID uses teacher-generated references; HPSv3 uses its own prompt dataset and generations.

Method	FID↓	CLIP↑	ImageReward↑	HPSv2.1↑	HPSv3↑
SDXL					
Flash (Chadebec et al., 2025)	28.10	29.18	0.584	25.77	8.09
PCM (Wang et al., 2024)	**20.51**	29.53	0.843	29.16	8.46
LCM (Luo et al., 2023)	30.10	29.41	0.647	26.28	7.49
SenseFlow (Ge et al., 2025)	25.79	**29.94**	<u>1.107</u>	<u>31.25</u>	9.32
DMD2 (Yin et al., 2024a)	<u>21.49</u>	<u>29.82</u>	1.026	31.10	<u>9.56</u>
Ours	23.21	29.61	**1.112**	**32.07**	**9.68**
Qwen-Image					
DMD2 (Yin et al., 2024a)	29.49	29.43	1.329	<u>31.42</u>	<u>9.57</u>
PiFlow (Chen et al., 2025)	**12.66**	**30.08**	<u>1.355</u>	31.02	9.43
Ours	<u>24.56</u>	<u>29.61</u>	**1.415**	**33.16**	**10.70**

Table 2: 8-step Qwen-Image-Edit results on CompBench and GEditBench v2. StateMix improves instruction following and content preservation.

	CompBench					GEditBench v2
Method	LC-T ↑	LC-I ↑	PSNR ↑	SSIM ↑	LPIPS ↓	VC Elo ↑
DMD2	<u>20.07</u>	**0.8425**	<u>26.41</u>	<u>0.8695</u>	<u>0.0690</u>	<u>989</u>
StateMix (Ours)	**20.33**	<u>0.8376</u>	**27.12**	**0.9041**	**0.0525**	**1011**

colors and more realistic textures in these examples, yielding a more lifelike overall appearance. Additional SDXL comparisons and Qwen-Image samples are provided in Appendix E.1.

5.2 Text-and-Image Conditioning under Application Requirements

Section 5.1 evaluates general text-to-image generation with text-only inputs. We further extend StateMix to generation conditioned on both text and images, addressing practical application requirements such as reference-guided content creation and instruction-based editing. We evaluate 8-step Qwen-Image-Edit on the public CompBench and GEditBench v2 benchmarks, together with human evaluation tailored to application-specific quality requirements.

Public benchmark evaluation. We compare StateMix with the DMD2 baseline. CompBench (Jia et al., 2026) measures edited-region instruction alignment and reference similarity (LC-T/LC-I), and preservation outside the edited region (PSNR/SSIM/LPIPS). GEditBench v2 (Jiang et al., 2026) provides a separate PVC-Judge visual-consistency (VC) comparison. Benchmark-specific settings are in Appendix B.4.

StateMix improves LC-T and all three non-edited-region preservation metrics (Table 2), showing that stronger instruction following can accompany better preservation. GEditBench's pair-relative VC Elo also favors StateMix. DMD2 retains higher LC-I, so edited-region reference similarity has a different ranking.

Human GSB evaluation for production acceptance. We evaluate 1,999 application-specific editing cases spanning text, products, people and animals, with three judgments per case and criterion against a 50-step teacher. Table 3 lists the eight criteria; aggregation details are in Appendix B.5. Let G, S, B count final case labels rated better than, indistinguishable from, or worse than the teacher:

$$\mathrm{GSB}_{\mathrm{human}} = \frac{G + S}{B + S}. \tag{7}$$

Scores above 1 favor the student and below 1 favor the teacher. The application's production standard requires an unrounded Human GSB strictly above 0.99 on *every* criterion.

StateMix exceeds DMD2 on all eight criteria and passes all eight gates, versus one for DMD2. Six StateMix scores exceed 1, indicating that human raters favor the 8-step StateMix student even over

Table 3: Human GSB from human-rater comparisons of 8-step students against a 50-step teacher.

Criterion	Ours		DMD2	
	Human GSB↑	Production gate	Human GSB↑	Production gate
Overall (whole image)	**1.093**	**Pass**	0.921	**Fail**
Appearance consistency	**1.109**	**Pass**	0.968	**Fail**
Text/logo details	**1.024**	**Pass**	0.990	**Fail**
Physical plausibility and interactions	**0.999**	**Pass**	0.983	**Fail**
Anatomical correctness	**0.999**	**Pass**	0.989	**Fail**
Text rendering	**1.017**	**Pass**	1.001	**Pass**
Realism	**1.029**	**Pass**	0.974	**Fail**
Aesthetics	**1.062**	**Pass**	0.985	**Fail**

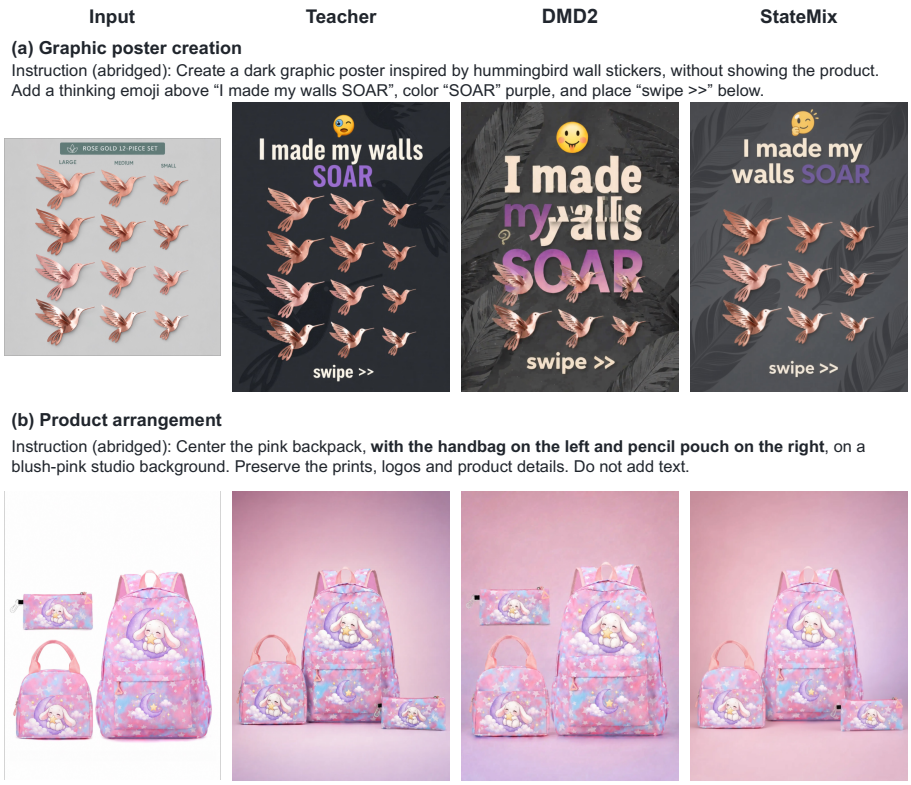

Figure 5: Qualitative image-editing comparison on graphic poster creation and product arrangement. Columns show the input, teacher, DMD2, and StateMix. StateMix produces more legible text and better follows the requested product arrangement than DMD2.

the 50-step teacher on these dimensions. Physical plausibility and anatomical correctness remain slightly below 1. Meeting this application-specific gate demonstrates practical utility, not statistical equivalence to the teacher.

Qualitative comparison. Figure 5 presents qualitative examples of graphic poster creation and product arrangement, illustrating more legible text and better adherence to the requested spatial relationships with StateMix than with DMD2.

5.3 Ablation Studies

State-source balance. To further investigate the effect of the training-state mixture ratio, we vary the rollout-start probability $p \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ on Qwen-Image with 4-step inference. The endpoints $p = 0$ and $p = 1$ correspond to noise-start-only and rollout-start-only training, respec-

tively. We use the same HPSv2 and HPSv3 prompt datasets as the main experiments and report FID, CLIP, ImageReward, HPSv2.1 and HPSv3 in Table 4.

Table 4: Training-state mixture ratio ablation on 4-step Qwen-Image.

Training states	p	FID↓	CLIP↑	ImageReward↑	HPSv2.1↑	HPSv3↑
Noise-start states	0.0	35.59	29.35	1.115	28.26	7.85
StateMix	0.2	36.71	29.29	<u>1.409</u>	<u>31.90</u>	<u>10.02</u>
StateMix	0.4	24.56	29.61	**1.415**	**33.16**	**10.70**
StateMix	0.6	**21.10**	**29.89**	1.310	31.15	9.99
StateMix	0.8	<u>23.98</u>	<u>29.70</u>	1.323	31.62	9.93
Rollout-start states (DMD2)	1.0	47.17	28.85	1.000	28.03	8.33

All evaluated mixed-source settings outperform both pure-source endpoints on HPSv2.1, ImageReward and HPSv3, showing that the preference gains extend across multiple mixture ratios. Different metrics favor different balances: $p = 0.4$ achieves the highest preference scores, whereas $p = 0.6$ gives the best FID and CLIP. At $p = 0.2$, FID and CLIP remain worse than noise-start-only training despite higher preference scores. These results show that adjusting the training-state mixture ratio allows a flexible trade-off between human preference scores, closeness to the teacher's generated distribution and image–text alignment.

Detailed settings and a visualization of the preference–teacher-distance trade-off are provided in Appendix C.1. Appendix D.2 further compares sample diversity across noise-start training, rollout-start training and StateMix, with quantitative metrics and qualitative examples.

6 Related work

Few-step diffusion distillation uses trajectory supervision or distribution matching. The former compresses sampling paths through progressive distillation (Salimans & Ho, 2022), enforces consistency across times (Song et al., 2023; Kim et al., 2024; Wang et al., 2024), or learns policies from teacher velocities (Chen et al., 2025). DMD matches distributions using teacher and learned fake scores (Yin et al., 2024b); DMD2 improves fake-score tracking and adds student rollouts (Yin et al., 2024a), while Decoupled DMD separates matching from CFG augmentation (Liu et al., 2025). CrossDistill assigns trajectory and distribution objectives to different noise intervals (Liu et al., 2026), and DDIL combines forward-noised data with teacher/student trajectories (Garrepalli et al., 2024). StateMix instead varies the input source under shared DMD2 supervision, mixing fixed-prior noise at intermediate target times with student rollout states to study conditional response and state transmission.

7 Conclusion

We introduced StateMix, a mixed-state training strategy that combines our newly introduced noise-start states with the original rollout-start states under the same DMD2 supervision, without changing the inference procedure. Controlled toy experiments reveal complementary conditional response and input-variation transmission. StateMix achieves leading human preference scores on SDXL and Qwen-Image and improves instruction following and content preservation on Qwen-Image-Edit. Human GSB evaluation further shows that the editing student meets all eight production quality criteria of the target application.

Limitations. We validate StateMix for 4-step generation and 8-step editing; other sampling budgets remain unexplored. In the single-step limit, both sources reduce to the same Gaussian prior, so state mixing adds no input-distribution difference over the corresponding DMD2 baseline.

AI use statement

Generative AI was used to assist with literature organization, manuscript revision, and toy-experiment code development and checking.

Ethics statement

This draft reports synthetic toy experiments, automated image-generation evaluation and supplied human-evaluation aggregates for image editing. All participants in the human evaluation provided informed consent. Faster generation can amplify misleading imagery, unwanted likeness use and biases inherited from teachers, training data or preference models. Preference and distribution scores do not establish safety, fairness or permission to redistribute data. Any release or deployment should preserve third-party model and dataset restrictions and undergo the relevant safety and license review.

Reproducibility statement

Algorithm 1 specifies the state-source update. The paired toy training and interventions, aggregation rules, secondary checks and negative results are detailed in Appendix A. The implementation retains trained weights and per-seed/per-pair results. Appendix B separates verified evaluation metadata from unavailable real-model configurations and explains the metric scales and teacher-reference FID. No public code link is asserted before an anonymous release is available.

References

Black Forest Labs. FLUX.2: Frontier visual intelligence. https://bfl.ai/blog/flux-2, 2025. November 25, 2025.

Clément Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2406.02347.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. pi-Flow: Policy-based few-step generation via imitation distillation. *arXiv preprint arXiv:2510.14974*, 2025. URL https://arxiv.org/abs/2510.14974.

Risheek Garrepalli, Shweta Mahajan, Munawar Hayat, and Fatih Porikli. DDIL: Diversity enhancing diffusion distillation with imitation learning. *arXiv preprint arXiv:2410.11971*, 2024. URL https://arxiv.org/abs/2410.11971.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. SenseFlow: Scaling distribution matching for flow-based text-to-image distillation. *arXiv preprint arXiv:2506.00523*, 2025. URL https://arxiv.org/abs/2506.00523.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, Lei Bai, Wanli Ouyang, Lin Chen, Fei Zhao, Yao Hu, Zihan Wang, Yuan Xie, and Shaohui Lin. CompBench: Benchmarking complex instruction-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2505.12200.

Zhangqi Jiang, Zheng Sun, Xianfang Zeng, Yufeng Yang, Xuanyang Zhang, Yongliang Wu, Wei Cheng, Gang Yu, Xu Yang, and Bihan Wen. GEditBench v2: A human-aligned benchmark for general image editing. *arXiv preprint arXiv:2603.28547*, 2026. URL https://arxiv.org/abs/2603.28547.

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 *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02279.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled DMD: CFG augmentation as the

spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025. URL https://arxiv.org/abs/2511.22677.

Yuxi Liu, Haoyu Li, Yixiang Cai, Tengxu Sun, Zekun Zhang, Baole Ai, Ang Wang, Jiamang Wang, Lin Qu, Kun Yuan, and Kai Zhang. CrossDistill: Balancing quality and diversity via trajectory-level hybrid few-step distillation. *arXiv preprint arXiv:2609.14725*, 2026. URL https://arxiv.org/abs/2609.14725.

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

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025. URL https://arxiv.org/abs/2508.03789. Version 2.

NVIDIA. Qwen-Image-Flash. https://huggingface.co/nvidia/Qwen-Image-Flash, 2026. Model card. Released July 23, 2026.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. *arXiv preprint arXiv:2104.11222*, 2022. URL https://arxiv.org/abs/2104.11222. Version 3.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

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. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00512.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2311.17042.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.18407.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.05977.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024a. URL https://arxiv.org/abs/2405.14867.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. URL https://arxiv.org/abs/2311.18828.

A Toy Experiments and Supporting Analysis

A.1 Training and evaluation protocol

The clean distribution contains six Gaussian modes with means on a circle of radius 3. Each covariance has tangential and radial standard deviations $s_{\tan} = 0.45$ and $s_{\mathrm{rad}} = 0.12$. The student and fake score are MLPs with three hidden layers of width 128 and SiLU activations; inputs are the two-dimensional state, sinusoidal noise-level features and a one-hot condition. The real score is analytic. The DMD2 surrogate uses the fake-minus-real clean-estimate direction, normalized by the batch mean absolute difference between the student output and real clean estimate (minimum denominator 10^{-4}). Score re-noising is uniform on $[0.02, 0.98]$. The fake score uses velocity flow matching, with two fake updates per student update. No CFG or auxiliary GAN/regression objective is used. Adam learning rates are 2×10^{-4} for the student and 4×10^{-4} for the fake score; batch size is 512 and each arm receives 3,000 student updates. The 4-step Euler grid is $\sigma \in \{1, 0.75, 0.5, 0.25, 0\}$.

The main results use new training seeds 3–7. Each seed has independently trained students using only rollout-start states, only noise-start states or their mixture ($p = 0.5$). Initialization is identical across arms within a seed. Purpose-specific RNG streams pair the conditions, target indices, generator noise, DMD2 re-noising and fake re-noising across arms; the branch draw has its own stream. Replaying all 15,000 training-step input sets verified exact input pairing, including both fake updates. The mixed $p = 0$ and $p = 1$ endpoints reproduce their pure arms exactly in code tests. Generated clean predictions and the fake-score trajectories naturally differ across arms; the experiment does not isolate a single gradient term from the coupled training dynamics.

Evaluation uses independent seed 2027, all 30 ordered pairs (c_A, c_B) with $c_A \neq c_B$, and 256 samples per pair. For each training seed we obtain common states from each of the three students at $k = 1, 2, 3$ and feed every state to every target student. Analytic marginal states are an additional check and are not pooled into the main student-source average. We record both one-forward clean predictions and complete remaining Euler trajectories; the main figure uses the latter. At $k = 3$ these readouts coincide because only 1 step remains. All settings and source-specific rows are retained.

Measurements. Let $Y_\theta(z, k, c)$ denote the output after completing the remaining grid. For a condition switch we record the final identity nearest to μ_{c_B}, retention of μ_{c_A}, and assignment to any other mode. We also record the continuous response

$$G_c = \mathbb{E}\left[\frac{\langle Y_\theta(z, k, c_B) - Y_\theta(z, k, c_A),\, \mu_{c_B} - \mu_{c_A}\rangle}{\|\mu_{c_B} - \mu_{c_A}\|^2}\right]. \tag{8}$$

For a within-mode direction e_c with true clean standard deviation s_e, we keep the condition and all other input factors fixed and use $z^{\pm} = z \pm (1 - \sigma_k) h s_e e_c$. The response gain is

$$G_e = \mathbb{E}\left[\frac{\langle Y_\theta(z^+, k, c) - Y_\theta(z^-, k, c), e_c\rangle}{2 h s_e}\right]. \tag{9}$$

Both tangential and radial directions are evaluated at $h = 0.25, 0.5$. The primary comparison is $k = 3$, with $h = 0.5$ for variation transmission; these choices were recorded before the new interventions were run, informed by the earlier low-noise role observation. A gain measures response, not a quality score or a required sample-wise coupling to a teacher trajectory.

Aggregation and uncertainty. All pairs receive equal sample counts. We first average the three student state sources within a training seed, then compute the mean and sample standard deviation across the five seeds. Sources, condition pairs and inference samples are not independent training replicates. Error bars are not confidence intervals, and no statistical-significance claim is made.

A.2 Conditional response and variation transmission

Point-level illustration and selection. Figure 3 is a post-hoc visualization of the existing primary setting, not a new training experiment. We export points from the saved students using evaluation seed 2027, $k = 3$, $h = 0.5$, and the same 256 samples per ordered condition pair. Among the 450 combinations of five seeds, three student state sources and 30 condition pairs, we select the case closest to the coordinate-wise cohort median of six measurements: new-condition accuracy

and tangential gain for each of the three arms. Distance is the mean absolute deviation after dividing each coordinate by its 10th–90th percentile range, with a floor of 0.01. The rule treats all arms equally and does not maximize a StateMix improvement. It selects seed 3, rollout-source states, and condition $3 \to 5$.

The figure shows all 256 points per cloud, with 12 uniformly indexed pairs connected in the second row. Coordinates are rotated and translated identically across each row; there is no per-model rescaling or point filtering. Blue contours are analytic clean Gaussian densities, not fitted output densities. Their levels are 0.02, 0.1, 0.3 and 0.6 of peak density. Gray dashed contours mark the old condition. In the second row, circles and crosses identify negative and positive input perturbations, respectively. The representative case has new-condition accuracies of 100%, 0% and 26.95% for students trained on noise-start states, rollout-start states and their mixture, respectively; the corresponding tangential gains are 0.051, 0.579 and 0.412. These case values are not the all-seed means. The saved candidate gallery includes a representative per seed and the least favorable mixed-minus-rollout condition contrast. Replayed measurements agree with the original per-pair report within 2×10^{-5}; no model was retrained.

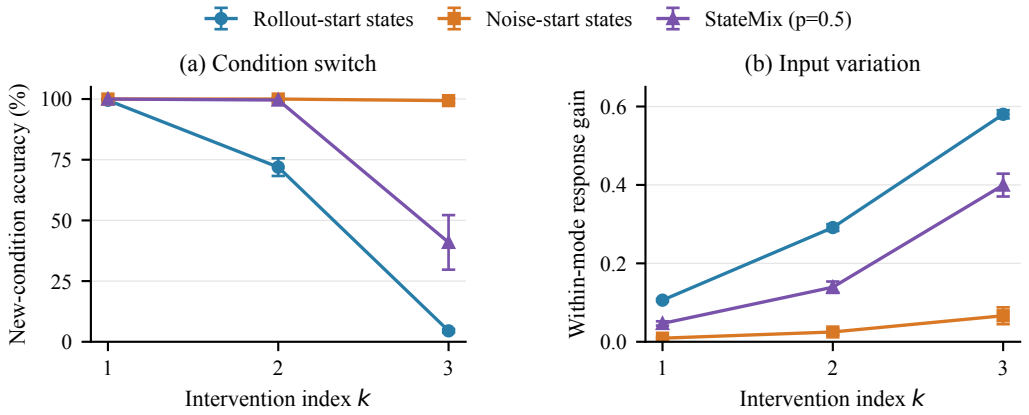

Figure 6: Toy condition and variation responses across five training seeds. Error bars show standard deviations.

Aggregate results. Table 5 gives the primary values underlying Figure 6. Relative to training on rollout-start states, StateMix improves new-condition accuracy by 36.44 ± 10.69 percentage points across seeds. Relative to training on noise-start states, it increases tangential gain by 0.333 ± 0.024. Each contrast is positive in every training seed and every state-source family, including the analytic marginal check.

Table 5: Toy condition-response and variation-transmission results at $k = 3$ (mean $\pm$ standard deviation over five training seeds).

Training states	New condition (%)	Inherited identity (%)	Tangential gain
Rollout-start states	4.51 ± 0.72	66.19 ± 3.13	0.580 ± 0.010
Noise-start states	99.34 ± 0.98	0.00 ± 0.00	0.066 ± 0.021
StateMix ($p = 0.5$)	40.94 ± 11.24	12.22 ± 6.56	0.400 ± 0.029

At $k = 2$, new-condition accuracy is $71.95 \pm 3.64\%$, 100% and $99.59 \pm 0.89\%$ for students trained on rollout-start states, noise-start states and their mixture; the corresponding tangential gains are 0.291, 0.025 and 0.139. At $k = 1$, accuracy is near ceiling. Reducing h to 0.25 leaves the $k = 3$ gains at approximately 0.5801, 0.0662 and 0.3996, respectively. Mean identity accuracy under the unchanged condition, including the detail interventions, remains near 100%. Cross-index trends also change the remaining inference horizon and the input scaling; we do not interpret their slopes as the location where structure is formed.

The new-condition accuracy contrast is not positive in every individual condition pair. Across 600 seed/source/pair cells, StateMix minus rollout is positive in 484, equal in 114 and negative in two; these cells are not independent repetitions. Both negative cells use seed 3 and condition $4 \to 3$:

for rollout-source states the rates are 8.984% versus 10.156%, and for analytic states 4.688% versus 5.859%. Tangential gain exceeds that of training on noise-start states in all 600 cells. Outputs assigned to neither requested identity are not counted as successful rewriting: at $k = 3$ their mean proportions are 29.31%, 0.66% and 46.83% for students trained on rollout-start states, noise-start states and their mixture.

A.3 Ordinary generation and distributional trade-offs

We also generate 1,280 samples per condition from standard Gaussian noise with the full 4-step sampler. Conditional sliced W_2^2 uses 128 fixed unit projections, averages sorted-sample squared discrepancies within the *requested* condition, and then averages conditions. Incorrect-mode samples are not reassigned or discarded. The reference is the known clean mixture, not a teacher image set. Spread ratios divide the generated standard deviation by its true tangential or radial value and have ideal value 1. Table 6 shows both benefits and costs. Values are means $\pm$ standard deviations over five training seeds; identity accuracy is reported as a mean only.

Table 6: 4-step toy generation across five paired training seeds.

Training states	Identity (%)	Tangential ratio	Radial ratio	Sliced W_2^2
Rollout-start states	99.982	0.745 ± 0.020	2.726 ± 0.097	0.0710 ± 0.0126
Noise-start states	100	0.081 ± 0.017	0.202 ± 0.062	0.2868 ± 0.0794
StateMix ($p = 0.5$)	100	0.324 ± 0.038	1.136 ± 0.128	0.1391 ± 0.0438

StateMix reduces the variation collapse seen with noise-start states and the radial over-dispersion seen with rollout-start states, but remains tangentially contracted and does not achieve the latter's lower overall distribution distance. Saturated identity prevents this toy from establishing an ordinary-generation structural advantage. These costs are retained rather than defining success using only the intervention metrics.

A.4 Failure outcomes and state-source sensitivity

Table 7 separates changed-condition outputs into the requested new condition, the inherited condition, and all other modes. We use the clean readout at $k = 3$, averaging over the rollout-, noise- and mixed-student state sources within each training seed before summarizing the five seeds. Analytic states are excluded from this primary aggregate and retained in the separate source checks.

Table 7: Outcomes after switching the condition at $k = 3$ in the paired toy experiment. Percentages are means $\pm$ one standard deviation over five training seeds, after averaging the three student state sources within each seed.

Training states	New condition	Inherited condition	Other modes
Rollout-start states	4.51 ± 0.72	66.19 ± 3.13	29.31 ± 2.49
Noise-start states	99.34 ± 0.98	0.00 ± 0.00	0.66 ± 0.98
StateMix ($p = 0.5$)	40.94 ± 11.24	12.22 ± 6.56	46.83 ± 5.34

Relative to rollout-start training, StateMix reduces inherited-condition retention from 66.19% to 12.22% and raises new-condition accuracy from 4.51% to 40.94%. However, its other-mode rate also rises from 29.31% to 46.83%. Reduced retention therefore does not translate entirely into correct responses to the new condition. Noise-start training gives the highest new-condition accuracy, while its weaker variation transmission is documented in Appendix A.2.

The full toy crossing retains both supportive and non-supportive evidence. The changed-condition contrast is positive in each seed/source aggregate but not every condition pair; the two pair-level counterexamples are listed in Appendix A.2. Normal generation shows variation contraction for StateMix and the student trained on noise-start states; the student trained on rollout-start states over-disperses radially but retains the lowest overall conditional distribution distance. A separate isotropic-noise probe did not establish a general robustness advantage. These observations delimit the functional-complementarity claim.

Interpretive boundary. The toy supports complementary learned dependencies under controlled interventions. It does not equate a posterior mean with the unique correct generator output, identify image texture with a two-dimensional residual, or establish the causal origin of image-model preference-score gains. Because changing student outputs also changes the fake-score training distribution, it does not isolate the effect of one gradient pathway from the coupled training dynamics.

A.5 Supporting derivations

DMD2 objective, fake-score fitting and guidance For a student training prediction y, the forward corruption is

$$x_\tau = \alpha_\tau y + \sigma_\tau \xi, \qquad \xi \sim \mathcal{N}(0, I). \tag{10}$$

Write $p_{\theta,\tau}(\cdot \mid c)$ for the distribution of these re-noised predictions and $p_{T,\tau}(\cdot \mid c)$ for the teacher distribution. The former includes the mixture of sampled training times and state sources; it need not equal the distribution of final few-step inference outputs. The unguided objective is a noise-averaged reverse KL (Yin et al., 2024b):

$$\mathcal{L}_{\mathrm{DM}}(\theta) = \mathbb{E}_{c,\tau}[w(\tau) D_{\mathrm{KL}}(p_{\theta,\tau}(\cdot \mid c) \,\|\, p_{T,\tau}(\cdot \mid c))]\,. \tag{11}$$

With $s_\phi \approx \nabla_x \log p_{\theta,\tau}$, the estimated clean-prediction gradient is the fake-minus-teacher field

$$g^{\mathrm{DM}}_{k,\tau} = \mathrm{sg}\big[w(\tau)\alpha_\tau \big(s_\phi(x_\tau, \tau, c) - s_T(x_\tau, \tau, c)\big)\big]\,. \tag{12}$$

Applying the selected forward's Jacobian to this field gives Equation 2, with the prepared input held fixed. Denoised-prediction implementations express this score difference using fake and teacher clean estimates. Practical guidance, weighting and normalization belong to the chosen injected field $g_{k,\tau}$. The gradient is implemented with

$$\mathcal{L}^{\mathrm{proxy}}_G = \tfrac{1}{2}\|y - \mathrm{sg}(y - g_{k,\tau})\|_2^2, \qquad \nabla_y \mathcal{L}^{\mathrm{proxy}}_G = g_{k,\tau}. \tag{13}$$

This proxy reproduces the chosen update, not the numerical KL value. Batch and dimension averaging are omitted.

Fake-score fitting uses detached samples $\bar{y} = \mathrm{sg}(y)$ and independent corruption draws $\bar{x}_\tau = \alpha_\tau \bar{y} + \sigma_\tau \xi$. In clean-prediction form,

$$\mathcal{L}_{\mathrm{fake}}(\phi) = \mathbb{E}\big[\lambda(\tau)\|D_\phi(\bar{x}_\tau, \tau, c) - \bar{y}\|_2^2\big]\,. \tag{14}$$

Noise- or velocity-prediction losses use corresponding targets and timestep weights. The fake-update ratio is a recipe choice; DMD2 motivates multiple fake updates per generator update (Yin et al., 2024a).

For linear text CFG, the teacher becomes $s^\gamma_T = s^\varnothing_T + \gamma(s^c_T - s^\varnothing_T)$, where $\varnothing$ is the null-prompt branch (Ho & Salimans, 2022). This modifies the unguided update in the DMD2 training framework. Norm rescaling, score-time schedules, original DMD's paired regression and DMD2's auxiliary adversarial term remain backbone-specific choices. The controlled toy omits CFG and these auxiliary terms.

Training predictions and final inference Fix a condition c and a snapshot $\bar{\theta}$ for input preparation. Let $\pi_{\mathrm{roll}} = p$, $\pi_{\mathrm{noise}} = 1 - p$, and $\mu^{\mathrm{noise}}_{\bar{\theta},k} = \nu$. The clean-prediction law for one group $i = (k, r)$ is the pushforward

$$Q^0_{\theta|\bar{\theta},k,r}(\cdot \mid c) = \big(F_\theta(\cdot, t_k, c)\big)_{\#}\mu^r_{\bar{\theta},k}(\cdot \mid c), \qquad Q^0_{\theta|\bar{\theta}} = \sum_{k,r} q(k)\pi_r Q^0_{\theta|\bar{\theta},k,r}. \tag{15}$$

The pushforward denotes the distribution obtained by applying the clean predictor to the selected input law. In contrast, $P^{\mathrm{infer}}_\theta(\cdot \mid c)$ is induced by running the complete inference sampler from its initial noise, including any sampler randomness. Training selects a clean prediction at a sampled target time and source; inference composes the configured sampling transitions. There is no general identity between these laws. If the sampler's terminal transition equals the clean predictor, its final rollout component at $\bar{\theta} = \theta$ can coincide with $P^{\mathrm{infer}}_\theta$; matching the pooled mixture still does not enforce matching that component. Detaching the rollout freezes input preparation for the current derivative.

Let $\mathcal{C}_\tau$ be the corruption kernel $x_\tau = \alpha_\tau y + \sigma_\tau \xi$, with Gaussian ξ. If score time is sampled independently of (k, r), its conditional prediction law is $\sum_{k,r} q(k)\pi_r \mathcal{C}_\tau Q^0_{\theta|\bar{\theta},k,r}$. More generally, define

$$a_i(c,\tau) = \Pr(i \mid c, \tau), \qquad Q_i = \mathcal{L}(x_\tau \mid i, c, \tau), \qquad Q = \sum_i a_i(c,\tau) Q_i. \tag{16}$$

This conditional definition also covers group-dependent score-time schedules, so distinct generation and score times do not require statistical independence. All subsequent identities are pointwise in (c, τ).

Pooled matching and component matching

Proposition. Let $a_i > 0$ sum to one, let $Q = \sum_i a_i Q_i$, and let P_T be one common teacher law. Assume the components have densities with respect to a common dominating measure and that $\sum_i a_i D_{\mathrm{KL}}(Q_i \| P_T) < \infty$. Then Equation 6 holds. Zero-weight groups at $p = 0$ or $p = 1$ are omitted.

Proof. Write the densities as ρ_i, $\rho = \sum_i a_i \rho_i$, and ρ_T. Since $\rho \geq a_i \rho_i$, the component-to-mixture KL terms are finite. Splitting the log ratio and summing yields

$$\sum_i a_i \int \rho_i \log \frac{\rho_i}{\rho_T} = \sum_i a_i \int \rho_i \log \frac{\rho_i}{\rho} + \sum_i a_i \int \rho_i \log \frac{\rho}{\rho_T} \tag{17}$$

$$= \sum_i a_i D_{\mathrm{KL}}(Q_i \| Q) + \int \rho \log \frac{\rho}{\rho_T}. \tag{18}$$

The final integral is $D_{\mathrm{KL}}(Q \| P_T)$, proving the identity. The component-to-mixture sum is the weighted Jensen–Shannon divergence and is nonnegative. Equality of pooled KL and mean component KL requires every positive-weight component to equal Q almost everywhere.

Consequence. For distinct smooth Q_1, Q_2 and $P_T = (Q_1 + Q_2)/2$, pooled KL vanishes although both component KLs remain positive. This analytical example shows that matching Q does not imply matching any particular source/time. The actual inference sampler must also be evaluated separately.

When the fake score represents this mixture. A fake network conditioned on (x_τ, τ, c) without the group label has the pooled score $\nabla_x \log Q$ as its ideal denoising target when its fitting samples follow the same conditional mixture and its loss has the corresponding score-matching interpretation. In particular,

$$\nabla_x \log \rho(x) = \sum_i \frac{a_i \rho_i(x)}{\rho(x)} \nabla_x \log \rho_i(x). \tag{19}$$

Its component weights depend on the observed state; it is generally not the constant-weight mean of component scores. Independent generator/fake draws are compatible with this law; unequal source/time sampling changes the fake's target. Finite capacity and imperfect score tracking introduce additional approximation.

Relating the distributional identity to the DMD2 update requires differentiability, integrability, exact scores and fixed input preparation. Actual CFG, rescaling, gradient normalization and auxiliary losses modify the recipe. The identity does not make the full implemented update the gradient of an unmodified fixed KL or guarantee convergence.

Noise-start states and flow parameterization Distribution matching acts on re-noised clean predictions. Its definition does not require inputs to follow the diffusion marginal at the nominal target time; optimization and expressivity remain separate questions.

For the flow clean predictor $F_\theta(z, \sigma_k, c) = z - \sigma_k v_\theta(z, \sigma_k, c)$, representing a specified map $G_k(z, c)$ at $\sigma_k > 0$ requires

$$v_\theta(z, \sigma_k, c) = \frac{z - G_k(z, c)}{\sigma_k}. \tag{20}$$

If the displacement does not decrease with σ_k, small target noise levels require larger velocity magnitudes. This is a pointwise parameterization relation, not a condition-number, stability or convergence result. It applies to this flow form and is not substituted for SDXL's parameterization. At $\sigma_k = 0$ the formula does not apply.

Local mixed gradients At fixed student and fake-score parameters, total expectation gives

$$\mathbb{E}[u_{\text{StateMix}}] = p\,\mathbb{E}[u_{\text{roll}}] + (1-p)\,\mathbb{E}[u_{\text{noise}}]. \tag{21}$$

Equation 21 holds at fixed current student and fake parameters. Both conditional terms use these models, not separately trained endpoints. Adam's nonlinear moment normalization prevents interpreting it as a corresponding identity for parameter updates. The moving rollout distribution and fake score also preclude reading it as a fixed regularization penalty.

Detached-prefix gradient paths The following identities concern the selected sample's surrogate gradient, not the numerical KL value or the full alternating optimization dynamics. Hold the condition, noise, target index, grid and detached clean-space field $g_{k,\tau}$ fixed. For the composed prediction $F_\theta(R_{\theta,<k}(\epsilon, c), t_k, c)$, the chain rule gives

$$u_k^{\text{cur}} = J_\theta F_\theta(z_k, t_k, c)^\top g_{k,\tau}, \tag{22}$$

$$u_k^{\text{pre}} = J_\theta R_{\theta,<k}(\epsilon, c)^\top J_z F_\theta(z_k, t_k, c)^\top g_{k,\tau}, \tag{23}$$

$$\nabla_\theta^{\text{full}} \ell = u_k^{\text{cur}} + u_k^{\text{pre}}, \qquad \widetilde{\nabla}_\theta \ell_{\text{roll}} = u_k^{\text{cur}}. \tag{24}$$

The first Jacobian differentiates the target forward with its input held fixed; the second term differentiates through the prefix. A stop-gradient on z_k deletes only the latter path. It does not remove the teacher/fake guidance acting on the target clean prediction. Shared parameters can still change the prefix on subsequent updates, which is different from backpropagating this sample's error through that prefix.

For a flow-velocity parameterization $F_\theta(z, \sigma, c) = z - \sigma v_\theta(z, \sigma, c)$,

$$J_\theta F_\theta = -\sigma J_\theta v_\theta, \qquad J_z F_\theta = I - \sigma J_z v_\theta. \tag{25}$$

Under Euler transitions $z_{j+1} = z_j + \Delta\sigma_j v_\theta(z_j, \sigma_j, c)$, the prefix sensitivity $A_j = \partial z_j / \partial\theta$ satisfies

$$A_{j+1} = (I + \Delta\sigma_j J_z v_\theta) A_j + \Delta\sigma_j J_\theta v_\theta, \qquad A_0 = 0. \tag{26}$$

Thus the empty prefix has no parameter sensitivity. The toy's paired two-backward diagnostic checks this anchor numerically. These identities do not specify the magnitude, sign or image-level consequence of the removed term.

Probe scope. The historical toy probe covers all trainable parameters in FP32, with the same 256-sample batch, rollout noise and score re-noising draws in both passes. Unused gradients are zero. We report

$$\rho_k = \frac{\|u_k^{\text{pre}}\|}{\|u_k^{\text{cur}}\| + \|u_k^{\text{pre}}\| + \varepsilon}, \qquad \cos(u_k^{\text{cur}}, u_k^{\text{pre}}). \tag{27}$$

We summarize across training seeds, not parameters. No image-model ρ_k is reported; applying this probe to image models requires declaring the parameter scope, precision and stochastic recomputation.

B Implementation and Evaluation Protocols

B.1 StateMix training procedure

StateMix mixes noise-start and rollout-start states under the same DMD2 supervision. Target time t_k and score re-noising time τ remain separate. The image-edit implementation synchronizes routing across distributed ranks and uses a separate seeded route when generating examples for fake-score updates; it does not require both optimizers to reuse the same branch. A reproduction must retain its backbone's conditioning, trainable scope, scheduler grid and shift, CFG convention, precision, optimizer and update ratio. The generic mixing rule does not determine those settings, and their

presence in a current script is not evidence of an evaluated historical run. The editing baseline is denoted DMD2 throughout; its complete historical supervision configuration and the human-study checkpoint identity are not recovered by citing the DMD2 paper. Figure 2 uses the same target index k as the text and separates source routing from score-time sampling.

Algorithm 1 StateMix mixed-state training: one student update

Require: Shared student F_θ, teacher score s_T, fake-score model s_ϕ
Require: Target-time grid $\{t_k\}_{k=0}^{K-1}$, index distribution $q(k)$, rollout-start probability p
Sample condition c and target index $k \sim q(k)$.
Sample state-source indicator $b \sim \mathrm{Bernoulli}(p)$.
if $b = 1$ **then**
    Sample $\epsilon \sim \mathcal{N}(0, I)$.
    Form rollout-start state $z_k^{\mathrm{roll}} \leftarrow \mathrm{sg}[R_{\theta,<k}(\epsilon, c)]$.
    Set $z_k \leftarrow z_k^{\mathrm{roll}}$.
else
    Sample fresh $\epsilon' \sim \mathcal{N}(0, I)$.
    Form noise-start state $z_k^{\mathrm{noise}} \leftarrow \epsilon'$ and set $z_k \leftarrow z_k^{\mathrm{noise}}$.
end if
Predict clean sample $y \leftarrow F_\theta(z_k, t_k, c)$ with gradients.
Sample score time τ following the DMD2 recipe and $\xi \sim \mathcal{N}(0, I)$.
Re-noise the prediction: $x_\tau \leftarrow \alpha_\tau y + \sigma_\tau \xi$.
Evaluate $s_T(x_\tau, \tau, c)$ and $s_\phi(x_\tau, \tau, c)$.
Update θ with the DMD2 objective, holding s_T and s_ϕ fixed.

B.2 Model and training configurations

The paired toy configuration is complete in Appendix A. Its implementation uses independent purpose-specific random streams so that branch selection cannot shift the conditions, target times or noises seen by another arm. Recorded outputs include all five training seeds, all state-source crossings, both one-forward and full-tail readouts, and both intervention strengths. No old model was overwritten when creating the paired replication.

For image models, the available records have different levels of detail. The sampling budgets are 8 steps for the Qwen-Image-Edit students and 4 steps for the SDXL and Qwen-Image text-to-image comparisons. The editing human-evaluation reference uses 50 steps. These are sampler step counts, not measured latency ratios or CFG-adjusted network-evaluation counts. The SDXL archive contains six complete score reports with per-sample data and common evaluator metadata, but not the full generation or training configurations. Qwen-Image HPSv2 results are supplied aggregates, including a named FastGen multi-step DMD2 checkpoint, using its DMD2 implementation (Yin et al., 2024a) and is denoted DMD2 throughout. The subsequently supplied HPSv2.1 reports identify the selected StateMix run and PiFlow as the data-free release. The HPSv3 values are separately supplied means on the HPSv3 prompt dataset. These records do not establish matched training budgets, identical seeds or complete checkpoint-selection windows across all methods. We do not fill missing historical fields from current launcher defaults.

Table 8: Training recipe settings specified in the repository launchers and their inherited configurations. Environment overrides can change these values; the table distinguishes implementation defaults from run-specific settings.

Setting	SDXL	Qwen-Image	Qwen-Image-Edit
Student sampling steps	4	4	8
Student learning rate	5×10^{-7}	10^{-5}	8×10^{-6}
Fake-score learning rate	5×10^{-7}	10^{-5}	8×10^{-6}
Per-device batch size	2	2	1
Generator update interval	5	5	4
Precision	FP16 option	BF16	BF16
Teacher guidance	CFG 8	CFG 4	Text 4.5 / image 0.3
Checkpoint interval	500	500	500

The SDXL recipe uses the mixed DMD2 launcher, with FSDP and 4-step backward simulation. The Qwen-Image recipe inherits FastGen's Qwen-Image DMD2 configuration and enables FSDP through the integration configuration; its computation precision is BF16 and its FSDP precision setting is FP32. The 8-step image-editing launcher specifies gradient checkpointing, a cosine learning-rate schedule with 200 warmup steps, and one gradient accumulation step. Its text/image guidance values in Table 8 are training settings, distinct from the zero-guidance inference settings used for CompBench below. For the Qwen-Image runs, the author confirms eight B200 GPUs, a per-device batch size of 2 and one accumulation round, giving a global batch size of 16. For the other recipes, per-device batch sizes do not by themselves specify a global batch size: world size and any run-level accumulation override must also be accounted for. The evaluated text-to-image runs use 1024×1024 generation, while the Qwen-Image integration permits a resolution override. In particular, its default resolution and the launchers' default mixing probabilities are not substituted for the recorded main-experiment settings.

B.3 TEXT-TO-IMAGE EVALUATION

SDXL HPSv2 score protocol. The first four metric columns in the SDXL block of Table 1 come from six completed reports on 3,200 HPSv2 prompts. Each report contains 3,200 unique sample IDs and no recorded scoring errors. Ordered IDs, prompt texts, categories and teacher-reference filenames match across reports. Anime, concept art, paintings and photographs each contribute 800 samples, making category-macro and sample-micro means equivalent up to rounding. The table uses the sample-micro means; these are not training-seed averages.

CLIP uses OpenCLIP `ViT-L-14-336-quickgelu` with OpenAI weights in FP16. RGB images are resized with antialiased bicubic interpolation to short side 336 and center-cropped to 336×336. Text is the original prompt without a prefix. The score is the dot product of normalized image and text embeddings, without the model's logit scale. ImageReward uses `ImageReward-v1.0`; preference scoring uses HPSv2.1. Raw ImageReward is displayed to three decimals; CLIP and HPSv2.1 are multiplied by 100 only for presentation and are not percentages.

Teacher-reference FID uses Clean-FID's Inception extractor in clean mode. Before its own 299-pixel resizing, images are center-square-cropped and resized to 512×512 using LANCZOS. Features from all eight extraction workers are concatenated before global means, covariances and FID are computed on CPU. The result is not a mean of shard FIDs or a paired pixel-reconstruction error. The reference is teacher-generated imagery, not a real-image dataset. Only the supplied `fid_t` is displayed; unused patch-FID fields are retained in the original score records.

The SDXL StateMix run uses $p = 0.8$. The supplied report identifiers establish the selected runs, not equal optimization budgets or a complete search history.

Separate HPSv3 evaluation. HPSv3 (Ma et al., 2025) is evaluated on its own prompt dataset, using new generations rather than re-scoring the HPSv2 image set. Its raw means are not multiplied by 100 and are not bounded percentage scores. The supplied SDXL values cover all six table methods. Qwen-Image reports StateMix (10.70), DMD2 (9.57) and PiFlow (9.43), ranked in that order.

The HPSv3 aggregates were provided separately from the HPSv2 archive. Their per-sample records, sample counts, image-generation settings and checkpoint identities are not independently available in the current record. In particular, the 3,200-image HPSv2 count and its matched filenames must not be assigned to the HPSv3 column. Rankings in that column describe the reported method means without an uncertainty or significance claim.

Qwen-Image score records. The Qwen-Image block of Table 1 uses the supplied `fid_t`, `clip_piflow_x100`, `imagereward_raw` and `hpsv2.1_x100` fields, with the same display conventions as SDXL. The supplied evaluation workflow points to original Qwen-Image teacher generations with an empty negative prompt and to the common scoring code. However, aggregate consistency does not independently establish per-sample matching or exact protocol equivalence. The latest StateMix aggregate supersedes an earlier reported run; the earlier values are retained as history, not mixed into the table.

The main Qwen-Image DMD2 HPSv2 metrics and category-wise results use the same selected model. The probability ablation uses a separate fixed-budget comparison. The DMD2 HPSv3 merged mean is 9.567282284269732.

Best and runner-up formatting uses the unrounded supplied means among the three displayed Qwen-Image methods. The distinction is a ranking convention, not a claim of a statistically meaningful gap.

B.4 IMAGE-EDITING EVALUATION

Table 2 keeps automatic scores separate from human preference scores. Its CompBench (Jia et al., 2026) values are transcribed from the two supplied terminal screenshots: the first is StateMix (the author's "mix" label) and the second DMD2. The evaluator prints the sample-micro means; category-macro means are stored separately and are not the values shown in this table. LC-T and PSNR are displayed to two decimals, and LC-I, SSIM and LPIPS to four; rankings use the unrounded values. PSNR is measured in dB, and neither CLIP similarity nor SSIM is an accuracy percentage.

The supplied rerun workflow uses eight sampling steps for both methods, with image and prompt CFG parameters set to zero. CompBench inputs are directly resized to 512×512 with bicubic interpolation, generated on that canvas, and scored with the corresponding source, target and nearest-neighbor-resized mask. The earlier fixed portrait-padding workflow did not preserve the same coordinate mapping and its scores are excluded. Since generation resolution, reference token counts and their associated dynamic shift also changed, the before/after comparison does not isolate padding as a single causal factor. The screenshots show completion but do not expose sample/error counts or per-sample scores; the full reports remain needed to independently verify coverage and uncertainty.

The supplied GEditBench v2 (Jiang et al., 2026) summary records complete PVC-Judge evaluation of visual consistency on 1,200 pairs: StateMix wins 638 and DMD2 wins 562, with no ties, giving preference rates of 53.17% and 46.83%. The reported VC Elo values are 1011 and 989, each with the supplied rounded bootstrap interval string $-10/+10$. These ratings belong only to this candidate pair and judge; they are not comparable to another arena, and instruction following and visual quality were not separately scored. The source labels are "mixed" and "multi_step"; the latter is denoted DMD2 for the corresponding automatic comparison. The canonical pairing places mixed at Image A and multi-step at Image B. No swapped-order evaluation was supplied, so position-bias robustness is not established. No relationship to the human study's checkpoint identities is assumed.

B.5 HUMAN EVALUATION PROTOCOL

The internal dataset and human-rating rubric are designed around the requirements of a specific application. The dataset contains 1,999 editing cases in four categories: text, products, people and animals. Unlike a single automatic preference score, the rubric separately assesses realism and aesthetics alongside overall quality, appearance consistency, text/logo details, text rendering, physical plausibility and anatomical correctness. Requiring every dimension to pass makes the acceptance rule stricter than relying on an aggregate preference score that can obscure individual weaknesses. The human-evaluation protocol uses three-fold annotation redundancy, with 8-step students evaluated against the 50-step teacher and with three judgments per case. The number of distinct annotators across the full study and the distribution of cases across categories have not been specified.

We report eight human-evaluation criteria, three-decimal Human GSB scores and count triples for StateMix (Ours) and DMD2. For each case and criterion, let g, s, b count the three G/S/B votes. Discard S votes and assign G if $g > b$, B if $b > g$, and S if $g = b$, including three S votes. Aggregate these final labels into counts G, S, B and apply Equation 7. S votes are ignored only for within-case voting; final S labels contribute to both numerator and denominator of the dataset-level score. Each method is evaluated relative to a 50-step teacher: Good (G) means better, Bad (B) means worse, and Same (S) means indistinguishable in human preference. Higher scores are better for all criteria, including anatomical correctness. Since $\mathrm{GSB}_{\mathrm{human}} = 1 + (G - B)/(B + S)$, values above 1 indicate net preference for the student, values below 1 indicate net preference for the teacher, and 1 means balanced G/B counts. Values just below 1 express a slight deficit in this aggregate human assessment. These scores come from human judgments, not a learned preference scorer. For this

application, the production quality standard requires a Human GSB score strictly above 0.99 on each of the eight criteria. All eight reported StateMix values exceed this threshold, so the 8-step student meets this standard. DMD2 falls below it on multiple criteria, including overall quality and appearance consistency, and thus fails the acceptance rule. This is a model-level quality gate for the specified application, not a fraction of individual test cases that pass or a universal threshold. Detailed counts and their composition are provided in Appendix D.3.

C Ablations and Training Analysis

C.1 Mixing-probability ablation details

This section supplements the mixing-probability ablation in Table 4. We vary the rollout probability $p \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ for 4-step Qwen-Image generation. At each sampled target time, the student receives a rollout-start state with probability p and a noise-start state with probability $1 - p$. Thus, $p = 0$ and $p = 1$ are the two pure-source endpoints, while $p = 0.4$ is the main StateMix setting. The $p = 0.4$ entry matches Table 1; the $p = 1$ ablation uses the rollout-only model from the fixed-budget ablation rather than the selected main-table model. For the noise-only, main mixed and rollout-only training runs, all other training variables are held fixed, as confirmed by the experiment author. The comparison uses the same 4-step inference budget; equal training update counts do not imply equal computational cost because rollout prefixes require additional student evaluations.

Metrics and aggregation. FID, CLIP, ImageReward and HPSv2.1 are reported for all six settings. FID is measured against teacher-generated images (FID-T), rather than real images or image patches. CLIP and HPSv2.1 are displayed after multiplication by 100, while ImageReward retains its raw scale. The supplied CLIP, ImageReward and HPSv2.1 results at $p = 0, 0.2, 0.6, 0.8$ use category-macro averaging. The HPSv2.1 reports at $p = 0.4, 1$ confirm four equally sized categories of 800 samples, making macro and micro means equivalent for these HPSv2.1 evaluations. The inherited CLIP and ImageReward aggregates retain their original reporting conventions. Full-precision values are retained where available, with rounding applied only for display.

HPSv3 uses its separate prompt dataset and raw score scale. Results are available for all six settings; the noise-start endpoint ($p = 0$) scores 7.8496200927595305, displayed as 7.85. The $p = 0.4$ value matches the main table; $p = 1$ retains the original 8.33 result rather than the selected main-table score. The $p = 0.2$ value is author-corrected to 10.018089697708687, displayed as 10.02; the updated $p = 0.6, 0.8$ scores are 9.994385379046202 and 9.926492227524518, displayed as 9.99 and 9.93. Exact checkpoint selection and repeated-training uncertainty are not established by these aggregate score records.

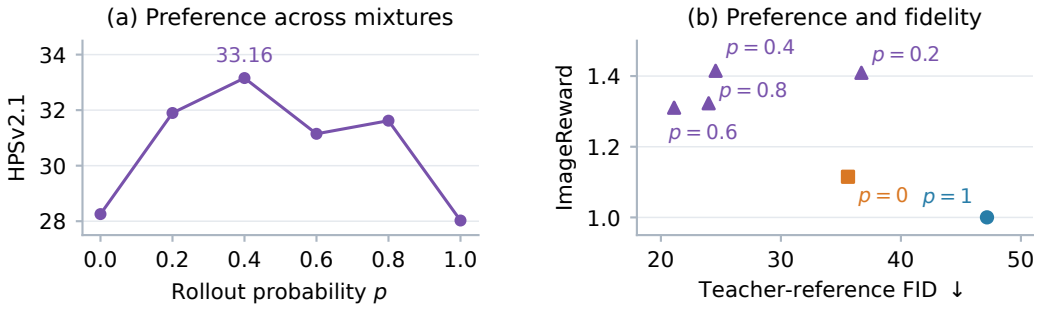

Figure 7: Supplementary visualization of the main-text probability ablation (Table 4). HPSv2.1 favors $p = 0.4$, while teacher-reference FID favors $p = 0.6$. These are descriptive checkpoint aggregates with no repeated-training uncertainty estimate.

Figure 7 visualizes the preference–fidelity trade-off in the existing ablation. It introduces no additional experiment and does not measure training stability or convergence speed. Training time is measured separately in Appendix C.2.

C.2 TRAINING COST

We compare noise-start training ($p = 0$), StateMix ($p = 0.4$) and rollout-start training ($p = 1$) on Qwen-Image. All other training settings are held fixed, as confirmed by the experiment author. We use the common iteration interval $(1000, 4000]$, which contains 30 consecutive timing records at 100-iteration spacing for each run. The logged `profiler/avg_iteration_time` measures elapsed wall-clock time between logging events divided by the logging interval. We average these equally sized windows and sum their durations for the interval time. We retain every window, including periodic spikes; these measurements may include checkpointing, evaluation and waiting costs within the interval. They are seconds per training iteration, not isolated generator-update latencies or time to reach a target quality.

Table 9: Recorded Qwen-Image training time over iterations 1,000–4,000 under matched settings. All logged intervals, including periodic timing spikes, are retained. Relative time uses rollout-start training as the baseline.

Training states	p	Time (s/iter)	Relative time	Time (min)
Noise-start states	0	2.681	0.870×	134.05
StateMix	0.4	2.848	0.924×	142.39
Rollout-start states	1	3.081	1.000×	154.03

StateMix takes 2.848 seconds per iteration compared with 3.081 for rollout-start training, a 7.56% reduction in recorded time over this interval. Noise-start training takes 2.681 seconds per iteration. This reduction is consistent with the lower frequency of student rollout generation during training. Rollout-start training always constructs the input state through the preceding student denoising steps, whereas StateMix with $p = 0.4$ selects this path only 40% of the time. The remaining updates draw noise-start states directly from the Gaussian prior, avoiding the rollout prefix. The resulting reduction in total iteration time is smaller than the reduction in rollout frequency because the target-step forward and backward passes, teacher/fake-score computation and optimizer updates remain necessary. These results measure the cost of the three state-source recipes over the same update interval; they do not establish a convergence-speed gain.

Training configuration. The runs use NVIDIA B200 GPUs with a generator-to-fake update ratio of $1 : 5$, as confirmed by the experiment author. The author confirms eight GPUs, a per-device batch size of 2 and one gradient accumulation round, giving a global batch size of 16. Computation uses BF16, with FP32 specified for FSDP precision.

D ADDITIONAL QUANTITATIVE EVALUATION

D.1 CATEGORY-WISE PREFERENCE SCORES

Table 10 expands the SDXL HPSv2.1 comparison into the four benchmark categories, using the archived per-category reports. Equal category sizes make category-macro and sample-micro averages equivalent. Scores are multiplied by 100 for display. Table 11 gives the corresponding Qwen-

Table 10: SDXL category-wise HPSv2.1 (×100). Each category contains 800 prompts. The overall score averages all 3,200 samples.

Method	Anime	Concept art	Paintings	Photo	Overall
LCM	27.80	26.36	27.14	23.81	26.28
PCM	30.56	29.26	29.52	27.29	29.16
Flash	27.23	25.59	26.24	24.00	25.77
SenseFlow	33.07	31.56	30.92	29.45	31.25
DMD2	32.57	31.52	31.47	28.85	31.10
StateMix	**33.56**	**32.42**	**32.43**	**29.88**	**32.07**

Image results. All three reports use HPSv2.1 and the same HPSv2 manifest, with 3,200 successful samples, zero scoring errors and 800 samples in each category. StateMix and DMD2 use the same selected models as the main-table HPSv2 comparison; PiFlow uses the data-free release. Category-macro and sample-micro means coincide for these equally sized categories. StateMix leads all

Table 11: Qwen-Image category-wise HPSv2.1 ($\times 100$). Each category contains 800 prompts. Best and second-best scores are bold and underlined, respectively.

Method	Anime	Concept art	Paintings	Photo	Overall
DMD2	<u>33.00</u>	<u>31.64</u>	<u>31.22</u>	29.81	<u>31.42</u>
PiFlow	32.70	30.81	30.42	<u>30.15</u>	31.02
StateMix	**34.33**	**33.70**	**33.53**	**31.07**	**33.16**

four categories on both SDXL and Qwen-Image, so the overall preference gain is not confined to one prompt category. On Qwen-Image, its largest margin over PiFlow is in paintings (33.53 versus 30.42), while photographs show the smallest margin (31.07 versus 30.15). These comparisons describe aggregate means and do not establish statistical significance. HPSv3 remains a separate evaluation and is not assigned the HPSv2 category labels.

D.2 Sample diversity

We compare the three Qwen-Image state-source recipes using 4-step, 1024×1024 generation. We sample eight prompts from each HPSv2 category (selection seed 2026) and use the same 16 inference seeds (42–57) for every prompt and model, yielding 512 images per model. Each prompt contributes all $\binom{16}{2} = 120$ unordered image pairs; we average within prompts and then equally across 32 prompts. LPIPS uses AlexNet (version 0.1) with bicubic resizing to 256×256. CLIP diversity is the mean $1 - \cos$ distance between normalized OpenAI ViT-L/14 image embeddings with standard preprocessing. The reported standard deviations describe variation across prompts, not uncertainty across independent training runs.

Table 12: Qwen-Image diversity using 32 prompts and 16 seeds per prompt. Values are means $\pm$ standard deviations across prompt-level averages. Higher values indicate greater dispersion in the respective feature space, not necessarily higher image quality.

Training states	LPIPS diversity	CLIP diversity
Noise-start states ($p = 0$)	0.2219 ± 0.0617	0.0669 ± 0.0573
StateMix ($p = 0.4$)	0.2498 ± 0.0720	0.0504 ± 0.0490
Rollout-start states ($p = 1$)	0.2894 ± 0.0724	0.0733 ± 0.0446

Table 12 shows a metric-dependent trade-off. StateMix's LPIPS diversity (0.2498) exceeds noise-start training (0.2219) but remains below rollout-start training (0.2894). Its CLIP diversity (0.0504) is lower than both noise-start (0.0669) and rollout-start (0.0733). Thus, the preference gains reported elsewhere do not imply uniformly increased sample diversity: mixing recovers some LPIPS variation relative to the noise-start endpoint while reducing dispersion in CLIP feature space. These averages alone do not establish the presence or absence of mode collapse.

Qualitative examples. Figure 8 illustrates two selected prompts from this diversity evaluation, with five outputs per method. Within each block, rows are noise-start ($p = 0$), DMD2 / rollout-start ($p = 1$) and StateMix ($p = 0.4$), as labeled. The supplied seed subsets differ between methods: columns are not seed-matched. These examples do not replace the full quantitative evaluation over sixteen inference seeds per prompt.



Figure 8: Diversity examples. Top block: “A squirrel is riding a skateboard.” Bottom block: “An elderly man is sitting on a couch.” Each block contains five samples per method; columns are not seed-matched across methods.

D.3 HUMAN EVALUATION BREAKDOWN

Table 3 reports Human GSB computed from the author-verified final case counts, using $(G + S)/(B + S)$ and rounding only for display. Production-gate decisions use the unrounded ratio and the strict threshold 0.99. StateMix passes all eight criteria; DMD2 passes only text rendering. DMD2's text/logo ratio is $1895/1915 \approx 0.989556$, which displays as 0.990 but remains below the threshold.

Table 13 preserves the accompanying triples. The author-verified dataset contains 1,999 cases, matching the sum of each final G/S/B count triple. The counts are retained unchanged. The exact reference checkpoint, annotator recruitment, blinding, detailed sample selection and per-annotation records are also unavailable, so no uncertainty estimate or statistical-significance claim is reported.

Table 13: Aggregated case-level G/S/B counts underlying Human GSB, from human-rater comparisons against the 50-step teacher.

Criterion	Ours	DMD2
Overall (whole image)	612/904/483	500/881/618
Appearance consistency	457/1253/289	318/1310/371
Text/logo details	110/1824/65	84/1811/104
Physical plausibility and interactions	76/1845/78	56/1853/90
Anatomical correctness	48/1901/50	39/1899/61
Text rendering	93/1845/61	79/1842/78
Realism	260/1530/209	198/1556/245
Aesthetics	380/1340/279	344/1287/368

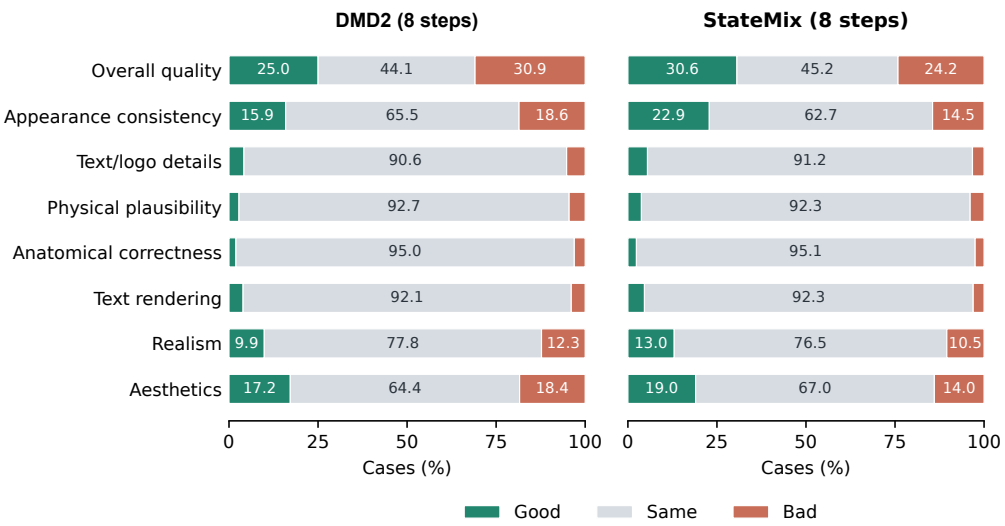

Figure 9: Final case-level human judgments over 1,999 editing cases per criterion. Good, Same and Bad compare each 8-step student separately with the 50-step teacher. Bar labels are percentages; the count table retains exact values. These are not direct StateMix-versus-DMD2 votes.

Figure 9 shows how the aggregate ratios arise. For overall quality, StateMix receives 612 Good and 483 Bad labels, compared with 500 Good and 618 Bad for DMD2. Appearance consistency also combines more Good labels with fewer Bad labels. Same judgments dominate text/logo details, physical plausibility, anatomical correctness and text rendering, making the composition useful for interpreting ratios close to one. The two models are each compared with the teacher, so differences between these bars do not estimate a direct pairwise student win rate.

E Additional Qualitative Results

E.1 Text-to-image examples

The Qwen-Image comparison in main Figure 4 presents portrait, street-scene, animal and jellyfish cases, with methods in columns. The rows correspond to HPSv3 characters_00015, transportation_00000, animals_00065 and animals_00012, respectively. Full prompts and source-image mappings are retained with the figure records. Images are displayed whole without enhancement. These selected examples are not a random-sample preference or anatomy-error estimate.

Additional SDXL comparisons. Figure 10 shows five supplied comparison cases, matched by HPSv3 sample ID, with the SDXL teacher, LCM, PCM, Flash, DMD2, SenseFlow and StateMix. These selected examples illustrate appearance and scene composition, not aggregate win rates.

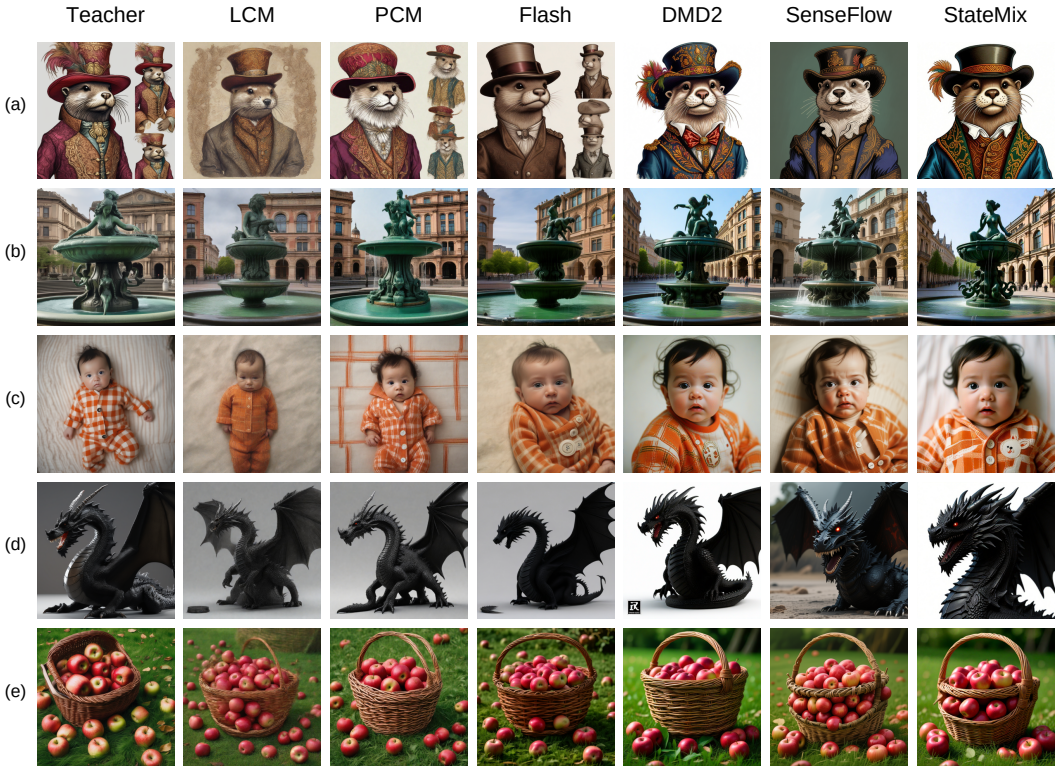

Figure 10: Additional SDXL comparisons. Columns: teacher, LCM, PCM, Flash, DMD2, SenseFlow and StateMix. Rows (a)–(e): HPSv3 animals_00848, architecture_00368, characters_00253, design_00225 and food_00370. All methods in a row share the same sample ID; identical initial noise is not established by the supplied files. Full images are shown without enhancement.

Additional Qwen-Image samples. Figure 11 shows the complete supplied 20-image StateMix Qwen-Image showcase, spanning portraits, animals, architecture, illustration, products and transportation. Images are retained in full; this selected gallery does not replace benchmark measurements.

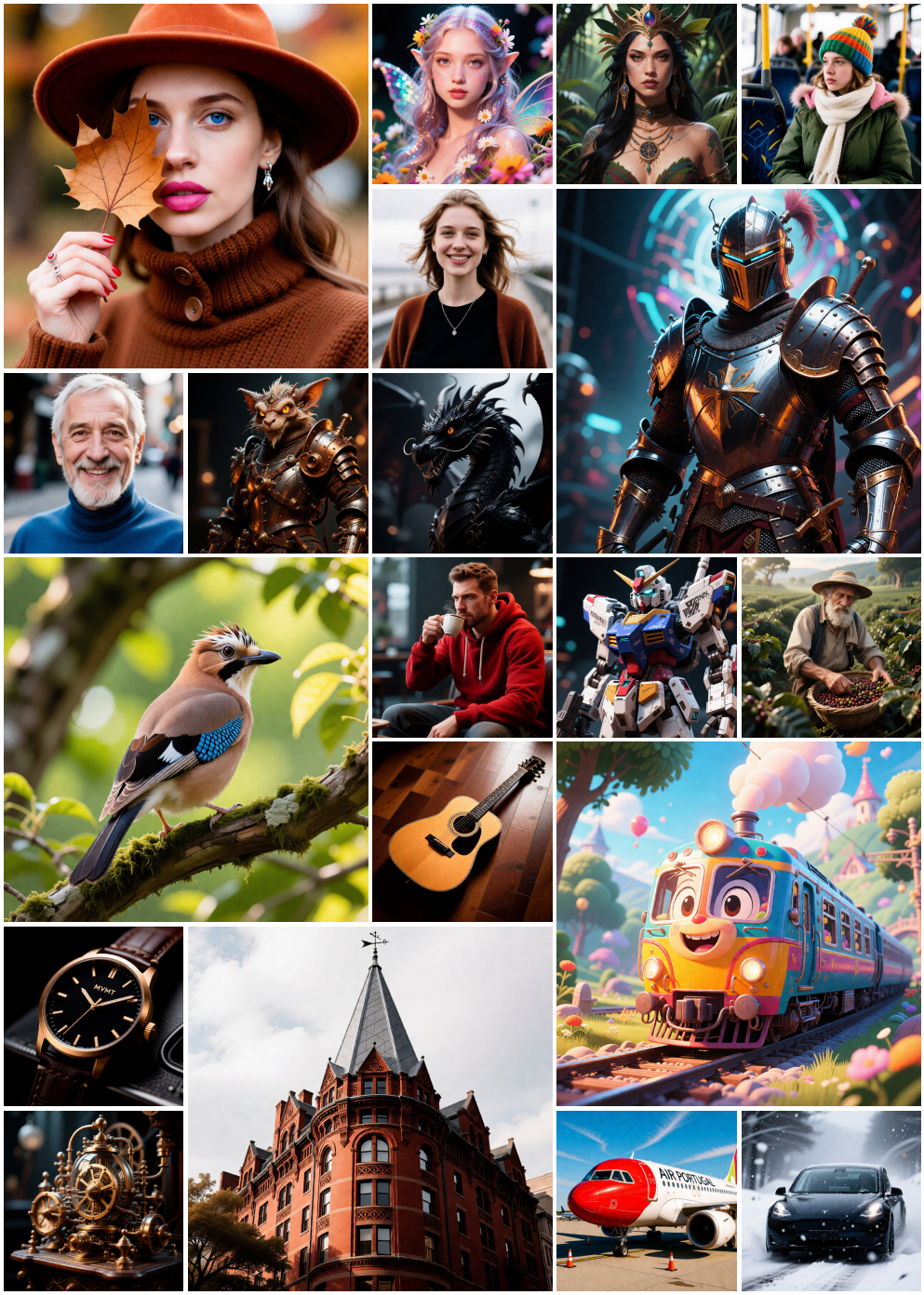

Figure 11: Twenty additional StateMix Qwen-Image examples on HPSv3 prompts. Five larger tiles and fifteen smaller tiles show all supplied examples; source sample IDs and layout positions are retained in the figure records. No image is cropped, retouched or enhanced.