STABILIZING MULTI-STEP DIFFUSION DISTILLATION WITH STUDENT PERTURBATION CONSISTENCY

Anonymous authors
Paper under double-blind review



Figure 1: 4-step generation results from Qwen-Image distilled with DMD w/ SPC.

ABSTRACT

Multi-step diffusion distillation often trains student transitions with local objectives but deploys them as a composition. Reducing per-step prediction error is therefore insufficient: residual error left at one step is amplified by the input sensitivity of every later step. Multi-step Distribution Matching Distillation (DMD) exposes this sensitivity twice. The exact gradient through the student rollout contains expensive and potentially ill-conditioned products of transition Jacobians, so practical methods detach intermediate states and train each step locally. Truncation stabilizes training, but the same Jacobians remain active at inference and propagate residual errors through the composed sampler. We propose Student Perturbation Consistency (SPC), a finite-difference regularizer that supplies this missing sensitivity supervision by matching endpoint predictions for clean and perturbed inputs at each supervised state. SPC approximates a squared Frobenius-norm penalty on the endpoint Jacobian, thereby controlling average transition gain without explicit Jacobians or rollout backpropagation. SPC does not reconstruct the discarded cross-step terms, and it targets controlled rather than minimal sensitivity, since over-regularization removes fine detail. On four-step Qwen-Image, SPC keeps the estimated squared Jacobian Frobenius norm near 1450 while the truncated baseline drifts above 2100, reduces endpoint amplification of injected perturbations from about 42 to 34, and also stabilizes full-path training without truncation. Across backbones, SPC consistently improves matched DMD baselines, benefits PCM as a plug-in regularizer, and matches or exceeds released accelerated models on several benchmarks.

1 INTRODUCTION

Diffusion distillation replaces dozens of teacher evaluations with one or a few learned transitions. Adding steps to a distilled generator should make it more accurate, yet in practice it also introduces

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.

new failure modes. End-to-end optimization can become unstable, and even when training is stable, defects introduced early in sampling can grow more pronounced downstream. Both problems are **structural**. The student is supervised locally, one transition at a time, but deployed as a composition of learned maps, so accuracy on each transition does not guarantee that errors stay controlled across the composition. We call this second property compositional stability, to distinguish it from the stability of optimization.

This distinction can be seen by decomposing the state error at step $i+1$ as

$$e_{i+1} \approx J_{\phi_i^\theta} e_i + \delta_i,$$

where δ_i is the current transition's approximation error and $J_{\phi_i^\theta} e_i$ propagates previous errors. Existing objectives primarily reduce δ_i through improved targets or state coverage. Multi-step quality, however, also depends on the gain $J_{\phi_i^\theta}$: even small residuals can be amplified by downstream Jacobian products.

Multi-step Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) exposes this tension particularly clearly. Its exact gradient backpropagates through the student rollout and contains products of intermediate Jacobians, which are memory intensive and can destabilize optimization. In practice, stop-gradient truncation, including stochastic gradient truncation strategy (SGTS) (Fan et al., 2025; Luo et al., 2025), removes these products by detaching intermediate states and retaining only local updates. This makes training tractable, but it does not change inference. The same endpoint sensitivities still enter the transition Jacobians that propagate residual errors through the composed sampler. The two effects therefore share one factor. The cross-step terms that truncation discards and the error amplification active at inference are both governed by the student Jacobian, so controlling that Jacobian addresses the inference-side amplification directly, without reconstructing the discarded terms. DMD's score-based distributional supervision *does not explicitly* constrain this local input sensitivity. *How, then, can we stabilize the learned composition without differentiating through the entire sampling chain?*

We introduce **Student Perturbation Consistency (SPC)**, a forward-only regularizer that supplies the missing sensitivity supervision. At a fixed timestep, SPC requires consistent endpoint predictions from a state and its small perturbation. A first-order expansion shows that this finite-difference objective implicitly penalizes Frobenius norm of the endpoint-map Jacobian, without constructing the Jacobian or backpropagating through the rollout. The penalty itself is classical. Our contribution is to aim it at the sensitivity that stop-gradient distillation leaves unsupervised. SPC thus preserves efficient stop-gradient training at the cost of one extra forward pass. Its target is controlled rather than minimal sensitivity, and an annealed radius keeps it from over-smoothing fine details.

Our experiments test each step of this mechanism. Full-path DMD exhibits Jacobian growth and training instability, and adding SPC alone makes it stable, which supports the shared-factor view. Under truncated training, SPC holds the measured sensitivity nearly constant while the baseline drifts upward, and it reduces the final-output amplification of controlled intermediate perturbations. It also consistently improves two-, and four-step DMD on Qwen-Image, transfers to Z-Image and Qwen-Image-Edit, and also improves PCM when used as a plug-in regularizer. It also improves the one-step student, suggesting that better local conditioning helps beyond composition.

Our contributions are:

- We formulate few-step generation through the lens of compositional stability and identify the student Jacobian as a shared factor in the cross-step gradient terms discarded by SGTS and the residual-error amplification active at inference.
- We propose SPC, a scalable finite-difference regularizer that controls local student sensitivity using only forward evaluations and is compatible with stop-gradient distillation.
- We validate the mechanism through training dynamics, Jacobian estimates, and controlled perturbations, and demonstrate consistent gains across model families, generation tasks, step budgets, and distillation objectives.

2 Related Work

Learning few-step generative mappings. Few-step diffusion generators are learned through trajectory regression (Salimans & Ho, 2022; Meng et al., 2023; Berthelot et al., 2023), consistency

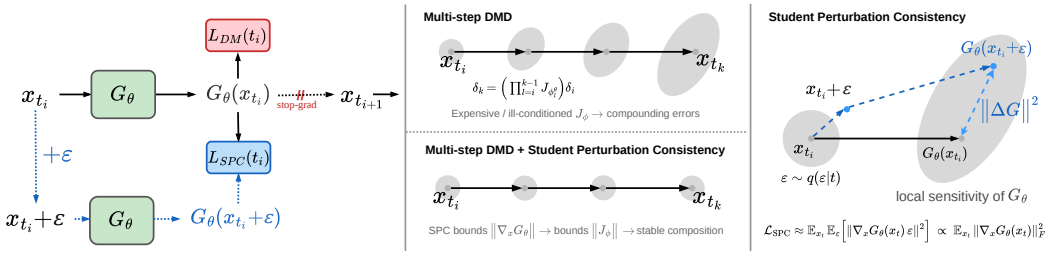


Figure 2: SPC enforces endpoint consistency around each supervised state (left), limiting residual-error amplification under composition (right).

objectives (Song et al., 2023; Song & Dhariwal, 2024; Luo et al., 2023; Wang et al., 2024; Kim et al., 2024), or flow-map and reparameterization approaches (Lipman et al., 2023; Liu et al., 2023; Geng et al., 2025; Cheng et al., 2026; Yang et al., 2026; Chen et al., 2026; Sun & Lin, 2026). Distribution-matching methods instead align output marginals: DMD (Yin et al., 2024b) uses a score-difference gradient, DMD2 (Yin et al., 2024a) adds student rollouts and adversarial training, and ADD/LADD (Sauer et al., 2024b;a) combine diffusion and adversarial objectives. Despite different targets, all multi-step variants deploy learned transitions whose errors interact under composition.

Error accumulation in multi-step distillation. Prior work primarily reduces the residuals introduced along a rollout. DDIL (Garrepalli et al., 2025) mixes forward and student-induced states, DMD2 and related backward-distillation methods train on student rollouts, and TDM (Luo et al., 2025) aligns teacher and student trajectories distributionally. Consistency objectives supervise shared endpoints along teacher trajectories; rCM (Zheng et al., 2025) further combines continuous-time consistency with score regularization. These methods improve targets or visited-state coverage but do not directly constrain the local gain that propagates remaining residuals. SPC supplies this complementary fixed-timestep neighborhood control, including when combined with PCM (Table 2).

Local sensitivity and Jacobian regularization. Input-sensitivity regularization is well established: double backpropagation, contractive autoencoders, and noise training induce derivative penalties (Drucker & Le Cun, 1992; Rifai et al., 2011; Bishop, 1995). Hoffman et al. (2019) provide a stochastic-projection Jacobian estimator but its extra computational cost is non-negligible, making it hard to apply on large diffusion models. Spectral-norm regularization and gradient penalties control related notions of gain (Yoshida & Miyato, 2017; Gulrajani et al., 2017), but act on layer weights or critic gradients rather than the student endpoint map. We therefore do not claim finite-difference Jacobian regularization itself as novel. Our contribution is to identify the endpoint-map Jacobian as the factor shared by SGTS's omitted cross-step terms and inference-time error propagation, specialize paired perturbation consistency to this setting without explicit Jacobians or JVPs, and validate the mechanism in large few-step diffusion students.

3 Method

3.1 Problem Setup

Multi-step diffusion distillation. We distill a pretrained diffusion teacher into a student endpoint predictor $G_\theta(x_t, t)$. For a K-step schedule $1 = t_1 > t_2 > \cdots > t_K > t_{K+1} = 0$, the first $K-1$ student evaluations define intermediate transitions

$$x^\theta_{t_{i+1}} = \alpha_i x^\theta_{t_i} + \beta_i G_\theta(x^\theta_{t_i}, t_i) =: \phi^\theta_i(x^\theta_{t_i}), \qquad i = 1, \ldots, K-1, \tag{1}$$

followed by the endpoint prediction $x^\theta_0 = G_\theta(x^\theta_{t_K}, t_K)$. The coefficients (α_i, β_i) are fixed by the discretization. Superscript θ denotes variables that remain connected to the full rollout graph; we write $x_t = \mathrm{sg}(x^\theta_t)$ for their stop-gradient counterparts.

Distribution Matching Distillation. DMD (Yin et al., 2024b) aligns diffused student and teacher distributions by minimizing $\mathcal{L}_{\mathrm{DMD}} = \int_0^1 D_{\mathrm{KL}}(p_{\mathrm{fake},t} \| p_{\mathrm{real},t})\,\mathrm{d}t$, whose gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \mathbb{E}\left[\Delta s(x_t) \cdot \nabla_\theta x^\theta_0\right], \qquad \Delta s(x_t) = s_{\mathrm{fake}}(x_t) - s_{\mathrm{real}}(x_t), \tag{2}$$

where the fake score is learned online and the real score comes from the teacher; endpoint regression provides an anchor (Yin et al., 2024b). These objectives reduce local residuals, whereas we study their propagation through the learned composition.

3.2 Compositional Error and the Hidden Jacobian Gap

Accuracy does not imply stable composition. Let $\phi_i^\star$ denote an ideal transition, $e_i = x_{t_i}^\theta - x_{t_i}^\star$ the state error, and $\delta_i = \phi_i^\theta(x_{t_i}^\star) - \phi_i^\star(x_{t_i}^\star)$ the local residual. A first-order expansion gives

$$e_{i+1} \approx J_{\phi_i^\theta} e_i + \delta_i, \qquad J_{\phi_i^\theta} = \alpha_i I + \beta_i \nabla_x G_\theta(x_{t_i}^\theta, t_i). \tag{3}$$

Reducing δ_i is insufficient when $J_{\phi_i^\theta}$ has high gain: a perturbation η_i introduced at step i reaches step k as

$$\eta_k \approx \left(\prod_{l=i}^{\overleftarrow{k-1}} J_{\phi_l^\theta} \right) \eta_i, \qquad \|\eta_k\| \le \left(\prod_{l=i}^{k-1} \|J_{\phi_l^\theta}\| \right) \|\eta_i\|. \tag{4}$$

Unrolling Eq. (3) makes this separation explicit:

$$e_K \approx \sum_{i=1}^{K-1} \left(\prod_{j=i+1}^{\overleftarrow{K-1}} J_{\phi_j^\theta} \right) \delta_i, \tag{5}$$

where an empty product is the identity. Better targets, distribution matching, and rollout-state supervision reduce δ_i but do not determine the gain of the multiplying Jacobian products. A student may therefore be accurate on supervised states yet amplify nearby deviations; we call control of this gain *compositional stability*.

The same Jacobian destabilizes full-path training. Because x_0^θ depends on every intermediate state, its exact derivative contains cross-step credit-assignment terms:

$$\nabla_\theta x_0^\theta = \frac{\partial G_\theta(x_{t_K}^\theta, t_K)}{\partial \theta} + \nabla_x G_\theta(x_{t_K}^\theta, t_K) \sum_{k=1}^{K-1} \left(\prod_{j=k+1}^{\overleftarrow{K-1}} J_{\phi_j^\theta} \right) \beta_k \frac{\partial G_\theta(x_{t_k}^\theta, t_k)}{\partial \theta}. \tag{6}$$

These ordered products make end-to-end optimization memory intensive and potentially ill-conditioned (full derivations in Section D). For two steps,

$$\nabla_\theta x_0^\theta = \underbrace{\frac{\partial G_\theta(x_{t_2}^\theta, t_2)}{\partial \theta}}_{\text{local update at step 2}} + \underbrace{\nabla_x G_\theta(x_{t_2}^\theta, t_2)\, \beta_1 \frac{\partial G_\theta(x_{t_1}^\theta, t_1)}{\partial \theta}}_{\text{credit assignment to step 1}}. \tag{7}$$

The second term coordinates the first transition with the endpoint objective but is multiplied by the final input Jacobian; for more steps, this becomes a Jacobian chain. High gain thus amplifies inference residuals in Eq. (5) and ill-conditions cross-step credit assignment.

Stop-gradient hides, rather than removes, sensitivity. SGTS (Yin et al., 2024a; Fan et al., 2025; Luo et al., 2025) detaches intermediate states and supervises sampled steps locally. At step t_i,

$$\nabla_\theta^{\text{sg}} \mathcal{L}_{\text{DMD}}(t_i) = \mathbb{E}\left[\Delta s \cdot \frac{\partial G_\theta(\text{sg}(x_{t_i}^\theta), t_i)}{\partial \theta} \right]. \tag{8}$$

This avoids storing or differentiating through preceding transitions, but removes the cross-step coordination in Eq. (6). For two steps, the omitted term is

$$\nabla_\theta \mathcal{L}_{\text{DMD}} - \nabla_\theta^{\text{sg}} \mathcal{L}_{\text{DMD}} = \mathbb{E}\left[\Delta s \cdot \beta_1 \nabla_x G_\theta(x_{t_2}^\theta, t_2) \frac{\partial G_\theta(x_{t_1}^\theta, t_1)}{\partial \theta} \right]. \tag{9}$$

The same Jacobian remains active during inference in Eq. (4). This *hidden Jacobian gap* removes an unstable factor from optimization without supervising its inference-time amplification. If $\|\nabla_x G_\theta\|_2 \le L$, then we have

$$\|J_{\phi_i^\theta}\|_2 \le |\alpha_i| + |\beta_i| L, \tag{10}$$

so local endpoint sensitivity bounds one factor in the SGTS gap and downstream amplification. Reducing L does not reconstruct the discarded term or make SGTS exact; it complements rollout supervision, which reduces δ_i, by controlling gain around the supervised states.

Algorithm 1 Training with Distribution Matching and SPC Regularization

Require: Teacher G, student G_θ, fake-score estimator s_ψ, SPC weight λ, base perturbation scale γ

for $i = 1, \dots, N$ **do**
  Sample $t \sim \mathcal{T}$ and roll out $x_t^\theta \sim p_\theta$
  $x_t \leftarrow \mathrm{sg}(x_t^\theta)$
  Sample $x_{t'} \sim p_\theta^{\mathrm{sg}}(x_{t'} \mid x_t)$ and update ψ by score matching on $x_{t'}$
  $\mathcal{L}_{\mathrm{DM}} \leftarrow \|G_\theta(x_t, t) - \mathrm{sg}(G_\theta(x_t, t) - \Delta s(x_{t'}))\|_2^2$ ▷ Distribution-matching loss
  $r_t \leftarrow \gamma\, \textsc{Scheduler}(t, i)$; sample $\varepsilon \sim \mathcal{U}([-r_t, r_t]^d)$
  $\mathcal{L}_{\mathrm{SPC}} \leftarrow \|G_\theta(x_t, t) - G_\theta(x_t + \varepsilon, t)\|_2^2$ ▷ Student perturbation consistency loss
  Update θ using $\nabla_\theta(\mathcal{L}_{\mathrm{DM}} + \lambda \mathcal{L}_{\mathrm{SPC}})$
end for
return θ

3.3 Student Perturbation Consistency

We seek a local, forward-only sensitivity signal that preserves stop-gradient efficiency and is independent of the base target. *Student Perturbation Consistency* (SPC) couples a detached state x_t with $x_t + \varepsilon$ at the same timestep and matches their endpoint predictions:

$$\mathcal{L}_{\mathrm{SPC}}(t) = \mathbb{E}_{x_t} \mathbb{E}_{\varepsilon \sim \mathcal{U}([-r_t, r_t]^d)} \left[\|G_\theta(x_t, t) - G_\theta(x_t + \varepsilon, t)\|_2^2 \right]. \tag{11}$$

Unlike input-noise augmentation, the paired loss directly constrains neighborhood variation without introducing a new endpoint target. The clean prediction is reused, so SPC adds one student evaluation.

Forward-only Jacobian regularization. For sufficiently small ε,

$$G_\theta(x_t + \varepsilon, t) = G_\theta(x_t, t) + \nabla_x G_\theta(x_t, t)\varepsilon + O(\|\varepsilon\|^2). \tag{12}$$

Since the components of ε are independent with $r_t^2/3$, substituting Eq. (12) into Eq. (11) yields

$$\mathcal{L}_{\mathrm{SPC}}(t) \approx \frac{r_t^2}{3} \mathbb{E}_{x_t} \left[\|\nabla_x G_\theta(x_t, t)\|_F^2 \right]. \tag{13}$$

Thus SPC approximates average local-sensitivity regularization using paired forward evaluations, without explicit Jacobians or rollout backpropagation. Isotropic perturbations probe state-space directions, while the Frobenius penalty controls average rather than worst-case gain and does not require contractive transitions. We specialize this established finite-difference tool to the sensitivity left unsupervised by stop-gradient distillation.

The radius trades signal strength against locality: large perturbations can violate the Taylor approximation and suppress detail. We set $r_t = \gamma\, \textsc{Scheduler}(t)$ and anneal it from stronger early regularization to a smaller detail-preserving radius (Appendix B).

Training objective. We augment the stop-gradient base objective at every supervised step:

$$\min_\theta \sum_{i=1}^{K} \mathcal{L}_{\mathrm{DMD}}^{\mathrm{sg}}(t_i) + \lambda \sum_{i=1}^{K} \mathcal{L}_{\mathrm{SPC}}(t_i). \tag{14}$$

The first term reduces local residuals and the second controls their amplification. SPC leaves the target and rollout construction unchanged, so it also complements consistency objectives such as PCM (Section 4) rather than replacing distribution matching, rollout correction, or trajectory consistency.

4 Experiments

4.1 Experimental Setup

We evaluate SPC on Qwen-Image (Wu et al., 2025) and Z-Image (Z-Image Team, 2025) for text-to-image generation, and on Qwen-Image-Edit for instruction-based editing. Text-to-image benchmarks include DPG-Bench (Hu et al., 2024), GenEval (Ghosh et al., 2023), HPS-v2.1 (Wu et al., 2023), and OneIG-Bench (Chang et al., 2025); editing is evaluated with ImgEdit (Ye et al., 2025b) and

Table 1: Ablation study on $\mathcal{L}_{\text{SPC}}$ and stop-gradient truncation strategy.

SGTS	$\mathcal{L}_{\text{SPC}}$	DPG	GenEval	HPS-v2.1	OneIG-Bench				
					Alignment	Text	Diversity	Style	Reasoning
		86.36	0.84	29.76	0.853	0.823	0.114	0.399	0.273
✓		87.79	0.87	31.30	0.869	0.875	0.117	0.423	0.293
	✓	**88.44**	**0.88**	**32.15**	**0.885**	**0.903**	0.125	0.423	**0.302**
✓	✓	88.36	**0.88**	32.08	0.880	0.897	**0.128**	**0.425**	0.295

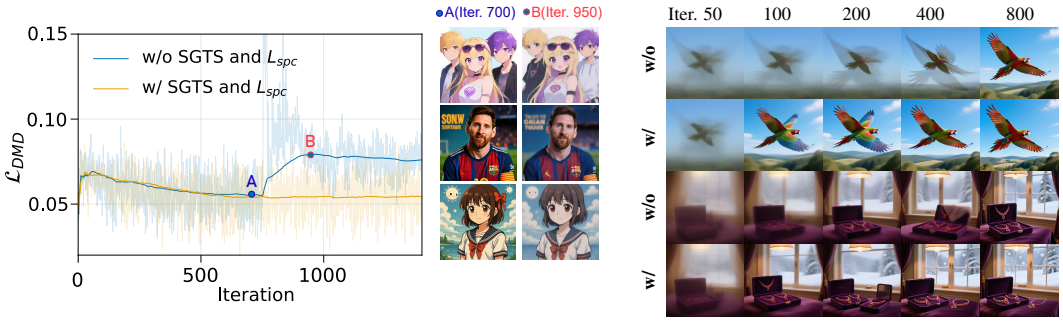


Figure 3: Training-side analysis. Left: full-path DMD (without SGTS or SPC) develops a sharp loss spike and quality collapse, whereas the practical SGTS+SPC configuration remains stable. Right: under the same SGTS training, examples without SPC (rows 1, 3) retain artifacts that SPC (rows 2, 4) progressively removes.

GEdit-EN (Liu et al., 2025). We compare with officially released accelerated models and recent few-step methods, including Qwen-Image-Lightning, Z-Image-Turbo, RCGM (Sun & Lin, 2026), TwinFlow (Cheng et al., 2026), ArcFlow (Yang et al., 2026), PiFlow (Chen et al., 2026), PCM (Wang et al., 2024), and DMD (Yin et al., 2024b;a). We evaluate released checkpoints when available and otherwise use reported results; DMD and PCM are reproduced under our setup.

DPG-Bench measures dense prompt following, GenEval tests compositional correctness, HPS-v2.1 approximates human preference, and OneIG-Bench separates Alignment, Text, Diversity, Style, and Reasoning. We retain this multidimensional evaluation because a stability regularizer may improve structural fidelity while over-smoothing details or contracting the output distribution.

Unless specified otherwise, we train LoRA rank-64 students with stop-gradient DMD and SPC ($\lambda = 1$, $\gamma = 0.01$), using a global batch size of 32. Text-to-image and editing models are trained for 7k and 2k iterations, respectively. Training details are in Appendix B.

4.2 Does SPC Stabilize the Learned Composition?

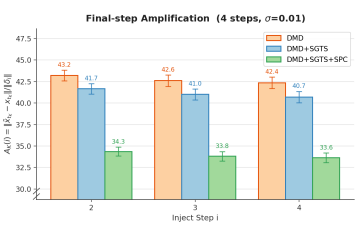
Our analysis makes three linked predictions: the hidden Jacobian should destabilize full-path optimization, SPC should suppress local student sensitivity under SGTS, and this suppression should reduce downstream perturbation amplification without forcing the Jacobian toward zero. We test each prediction on four-step Qwen-Image before comparing benchmark scores.

We distinguish three configurations throughout this analysis. Full-path DMD differentiates through the complete rollout and exposes the Jacobian products in Eq. (6); DMD with SGTS truncates those products; and SGTS+SPC keeps the same truncation while adding local sensitivity supervision. SPC therefore neither replaces SGTS nor restores full-path differentiation. This separation lets the experiments test the optimization benefit of truncation and the conditioning benefit of SPC independently.

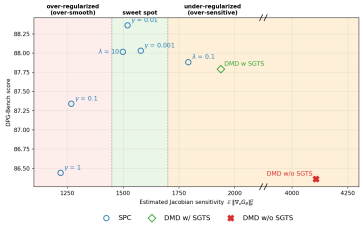
Training-side hidden Jacobian gap. Full-path DMD becomes unstable around iteration 750 (Fig. 3, left), consistent with ill-conditioned cross-step Jacobian products. SGTS restores stable optimization by truncating these products, while SPC directly controls the underlying local sensitivity. Accordingly,

SGTS	$\mathcal{L}_{\text{SPC}}$	500	1k	2k
–	–	2092	4342	4096
✓	–	1965	2069	2123
✓	✓	1447	1511	1462

(a) Jacobian sensitivity.



(b) Perturbation amplification.



(c) Quality vs. sensitivity.

Figure 4: Mechanism analysis on four-step Qwen-Image-20B. (a) Estimated squared Jacobian Frobenius norm during training $||\nabla_x G_\theta(x_t, t)||_F^2$. (b) Endpoint amplification after intermediate-state perturbations $\|x_0^{\text{pert}} - x_0\|_2/\|\varepsilon\|_2$. (c) DPG-Bench follows an inverted-U with student sensitivity.

adding SPC also stabilizes full-path training even without SGTS and achieves the best performance in Table 1. This result confirms that SPC can address the instability itself, and suggests that retaining complete cross-step gradients is beneficial once their conditioning is controlled. We nevertheless use SGTS by default because full-path differentiation incurs substantially greater computation and memory costs. SGTS with SPC therefore provides a more practical tradeoff between efficiency, stability, and performance.

The qualitative trajectories further show that SPC progressively removes persistent structural and texture artifacts. Overall, SGTS offers an efficient route to stable optimization, whereas SPC directly regularizes transition composition and remains effective with or without gradient truncation.

SPC reduces measured student sensitivity. We estimate $\mathbb{E}\|\nabla_x G_\theta(x_t, t)\|_F^2$ with five Hutchinson probes on a fixed set of 64 samples. The SGTS baseline rises from 1965 to 2123, showing that stable optimization alone does not prevent the endpoint map from becoming more sensitive. Full-path training is more severe, increasing from 2092 to 4342 before collapse. By contrast, SPC maintains a substantially lower and nearly constant estimate near 1450 (Fig. 4a), consistent with the finite-difference interpretation in Eq. (13).

Lower sensitivity coincides with lower downstream amplification. We inject equal-norm perturbations at each intermediate step and measure their effect on the final output $\|x_0^{\text{pert}} - x_0\|_2/\|\varepsilon\|_2$. In Fig. 4b, the SGTS baseline amplifies them by roughly 41–43, whereas SPC reduces amplification to about 34 at every injection location. The largest absolute reduction occurs for early perturbations that traverse the longest remaining composition, as predicted by Eq. (4). Because each comparison begins from the same sample and differs only by a controlled latent perturbation, this test is independent of a benchmark evaluator. Importantly, the objective is controlled rather than minimal sensitivity: Fig. 4c exhibits an inverted-U, with high-gain models amplifying errors and over-regularized models losing fine detail. The annealed SPC configuration occupies the favorable intermediate regime.

Additional controls show that the gains are not reproduced by simpler alternatives. Perturbed-state DMD and output-norm regularization fail to improve the baseline, while a one-sided SPC variant provides no clear gain (Table 8, Appendix C).

Taken together, the training trace, Jacobian estimate, amplification test, and inverted-U support the narrower claim that SGTS makes learning feasible, while SPC controls the local sensitivity that amplifies residual errors under composition.

4.3 Main Results

Qwen-Image. Table 2 shows that SPC provides the best aggregate results with two and four nfe evaluations. With two nfe evaluations, it improves DMD from 87.76 to 88.25 DPG, 0.86 to 0.88 GenEval, and 28.88 to 31.27 HPS-v2.1. With four nfe evaluations, SPC reaches 88.43 DPG, 0.88 GenEval, and 31.72 HPS-v2.1, while leading Alignment, Text, Style, and Reasoning. Its Diversity remains close to the matched DMD baseline. The gap from ArcFlow and PiFlow therefore reflects the diversity profile of DMD rather than suppression caused by SPC. SPC controls transition sensitivity and compound error, not output diversity.

Table 2: Quantitative comparisons on Qwen-Image. The NFE of Qwen-Image is recorded as 50×2 since it uses CFG. † indicates results reported in the original paper.

Model	NFE	DPG	GenEval	HPS-v2.1	OneIG-Bench				
					Alignment	Text	Diversity	Style	Reasoning
Baselines									
Qwen-Image	50×2	88.32†	0.87†	31.67	0.882†	0.891†	0.197†	0.418†	0.306†
Qwen-Image-Lightning	4	88.35	0.88	32.55	0.881	0.902	0.105	0.407	0.307
2-Step Generation									
RCGM	2	84.09†	0.82†	–	–	–	–	–	–
TwinFlow	2	86.50	0.82	29.23	0.863	0.810	0.127	0.364	0.266
ArcFlow	2	87.63	0.85	30.08	0.872	0.845	**0.182**	0.418	0.280
PiFlow	2	86.51	0.83	29.98	0.832	0.631	0.179	0.382	0.257
DMD	2	87.76	0.86	28.88	0.859	0.861	0.113	0.414	0.277
SPC (Ours)	2	**88.25**	**0.88**	**31.27**	**0.876**	**0.877**	0.119	**0.423**	**0.287**
4-Step Generation									
TwinFlow	4	86.86	0.83	30.88	0.857	0.872	0.126	0.356	0.265
ArcFlow	4	88.01	0.86	30.83	0.873	0.871	**0.183**	0.422	0.291
PiFlow	4	88.07	0.85	30.76	0.867	0.865	0.178	0.408	0.273
DMD	4	87.86	0.87	30.42	0.862	0.864	0.117	0.411	0.290
SPC (Ours)	4	**88.43**	**0.88**	**31.72**	**0.882**	**0.911**	0.118	**0.423**	**0.309**
Combine With Other Method									
PCM	4	87.58	0.86	28.01	0.852	0.861	0.176	0.405	0.267
PCM+SPC (Ours)	4	**88.02**	**0.87**	**30.96**	**0.876**	**0.871**	**0.182**	**0.413**	**0.272**

Table 3: Quantitative comparisons on Z-Image.

Model	NFE	DPG	GenEval	HPS-v2.1	OneIG-Bench				
					Alignment	Text	Diversity	Style	Reasoning
Baselines									
Z-Image	50×2	88.14†	0.87†	29.15	0.881†	0.987†	0.194†	0.387†	0.280†
Z-Image-Turbo	4	84.94	0.76	29.90	0.840	0.985	0.118	0.392	0.287
4-Step Generation									
DMD	4	85.36	0.73	29.12	0.836	0.908	0.165	0.430	0.276
SPC (Ours)	4	**86.97**	**0.77**	**30.10**	**0.862**	**0.956**	**0.177**	**0.459**	**0.287**

Increasing from two to four evaluations further improves DPG, HPS-v2.1, Alignment, Text, and Reasoning, while GenEval, Style and Diversity remain largely stable. The four step student also matches or exceeds the teacher on DPG and GenEval.

SPC also improves PCM from 87.58 to 88.02 DPG and from 28.01 to 30.96 HPS-v2.1, together with all five OneIG dimensions. Appendix C further evaluates diversity metrics like Vendi (Friedman & Dieng, 2023), Precision, Recall, Density, Coverage, and DreamSim (Fu et al., 2023). SPC improves DMD Recall (0.224→0.252) and Coverage (0.724→0.783), while keeping the other diversity metrics comparable. These results show no systematic diversity loss from SPC.

Transfer to Z-Image. With four evaluations, SPC improves its DMD baseline on every reported metric (Table 3) and outperforms the official Z-Image-Turbo checkpoint on DPG-Bench, GenEval, HPS-v2.1, Alignment, Diversity, and Style. Although it does not surpass the full-step teacher on every dimension, it is the strongest overall four-step trained variant in the comparison. The result suggests that the same local sensitivity objective can improve a different teacher–student pair without redesigning the distillation target.

Image editing. On Qwen-Image-Edit, SPC improves DMD on ImgEdit and all GEdit-EN dimensions, reaching an ImgEdit score of 4.32 (Table 4). Its perceptual-quality score also exceeds the teacher, although semantic consistency remains below full-step inference, as is also observed for Qwen-Image-Lightning. This remaining semantic gap is informative: SPC is intended to limit the amplification of errors but cannot recover information already lost by the local distillation target. The result is consistent with local sensitivity control also being useful when the endpoint map must preserve an input image while applying an instruction.

Across these settings, the consistent comparison is SPC against the same base objective, not a claim that one regularizer dominates every specialized accelerator. The improvements transfer across two

Table 4: Quantitative comparisons. SC: semantic correctness, PQ: perceptual quality, O: overall.

Model	NFE	ImgEdit	GEdit-EN		
			SC	PQ	O
Baselines					
Qwen-Image	50×2	4.27†	8.00†	7.86†	7.56†
Qwen-Image-Lightning	4	4.15	5.49	8.03	5.49
4-Step Generation					
DMD	4	4.18	5.42	8.01	5.43
SPC (Ours)	4	**4.32**	**5.71**	**8.11**	**5.61**

Table 5: Quantitative comparisons on Qwen-Image (1-step generation).

Model	NFE	DPG	GenEval	HPS-v2.1
Baselines				
Qwen-Image	50×2	88.32†	0.87†	31.67†
Qwen-Image-Lightning	1	87.55	0.85	29.12
RCGM	1	59.50†	0.52†	–
TwinFlow	1	86.27	0.82	26.64
DMD	1	84.63	0.84	27.74
SPC (Ours)	1	**87.34**	**0.87**	**29.35**

Table 6: Ablation study on the weighting factor λ, the perturbation radius γ, and the radius scheduler. The default setting is $\lambda = 1$, $\gamma = 0.01$ with the scheduler.

Metric	λ			γ				w/o scheduler
	0.1	1	10	0.001	0.01	0.1	1	
DPG	87.88	**88.36**	88.01	88.03	**88.36**	87.34	86.44	87.89
GenEval	0.87	**0.88**	0.87	**0.88**	**0.88**	0.87	0.86	**0.88**
HPS-v2.1	**32.22**	32.08	32.07	31.79	**32.08**	31.34	30.68	31.54

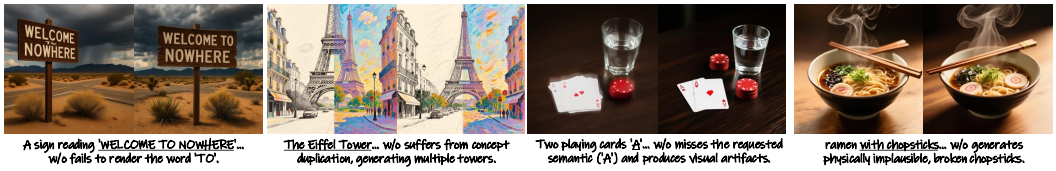


Figure 5: Same-prompt comparison of DMD baseline with and without SPC on Qwen-Image.

generative backbones, a conditional editing task, two- and four-step schedules, and both distribution-matching and consistency training. This breadth is useful because the method changes neither the teacher target nor the rollout states; the common intervention is local conditioning of the student map.

4.4 Additional Analyses

Step count and hyperparameters. Although motivated by multi-step composition, SPC also improves the one-step DMD baseline and reaches 87.34 DPG, 0.87 GenEval, and 29.35 HPS-v2.1 (Table 5). This gain suggests a broader regularization effect: beyond controlling error propagation under composition, SPC may also improve the local conditioning of an individual student map. The λ and radius studies in Table 6 support the controlled-sensitivity interpretation: weak regularization approaches the DMD baseline, whereas a large radius over-smooths the model. A radius of 0.01 and cosine annealing give the best balance.

Qualitative comparison. In Fig. 5, across repeated objects, thin structures, fine textures, and text, SPC more often removes artifacts that persist in the DMD student while preserving the prompt content. These examples illustrate the reported trend.

5 Conclusion

We presented Student Perturbation Consistency (SPC), a forward-only regularizer for few-step distillation. By constraining the local sensitivity of the endpoint map, SPC stabilizes multi-step composition and consistently improves few-step text-to-image generation and image editing across multiple benchmarks and model families. On four-step Qwen-Image, SPC reduces perturbation amplification from about 42 to 34 and keeps estimated Jacobian sensitivity stable. Ablations further show that SPC stabilizes full-path training without truncation. Given its minimal overhead and compatibility with existing stop-gradient pipelines, we believe SPC can serve as a plug-and-play component for accelerating multi-step distillation of large-scale generative models.

AI USE STATEMENT

We used generative AI tools (large language models) in the following ways. (i) Writing: to improve the clarity, grammar, and presentation of text written by the authors; AI was not used to draft any section of the paper from scratch. (ii) Retrieval and discovery: to help identify and summarize related work during the literature search; every cited work was subsequently read and verified by the authors. (iii) Mathematical derivations: to assist in deriving and checking intermediate steps of some proofs and equations; the formulation of all mathematical claims, the overall proof strategies, and the final proofs are the authors' own, and every AI-assisted step was independently re-derived and verified by the authors. (iv) Implementation: to assist in writing the code (e.g., evaluation scripts, baseline implementations, and plotting utilities). The method itself, including its design and algorithmic details, was developed by the authors; AI tools were used to translate the authors' specifications into code. All AI-assisted code was reviewed line by line, tested against hand-verified cases, and validated by reproducing the reported results before use.

Generative AI was not used to generate or clean datasets, to develop the theoretical framework or research hypotheses, to design the experiments, to perform qualitative or thematic analysis, to interpret results, or for translation. The research questions, methodology, experimental design, results, and conclusions are the authors' own. We take full responsibility for the content of this paper, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human participants, user studies, personal or sensitive data, or newly collected datasets. Our experiments use publicly available academic datasets and pretrained models released under licenses that permit academic research. We follow the applicable terms of use and do not redistribute restricted assets. To the best of our knowledge, this work raises no additional ethical concerns.

REPRODUCIBILITY STATEMENT

The main paper specifies the proposed method, assumptions, algorithms, and evaluation protocol. The appendix provides the complete derivations and proofs, implementation and training details, hyperparameters, dataset and model information, and additional experimental results. An anonymized implementation, including configuration files and evaluation scripts, will be available at `anonymous.4open.science/r/ICLR2027-SPC`. Together, these materials are intended to support reproduction of all main and supplementary results.

REFERENCES

David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. TRACT: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

Christopher M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. In *NeurIPS*, 2025.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. π-flow: Policy-based few-step generation via imitation distillation. In *ICLR*, 2026.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation, 2025. URL `https://arxiv.org/abs/2506.18095`.

Zhenglin Cheng, Peng Sun, Jianguo Li, and Tao Lin. TwinFlow: Realizing one-step generation on large models with self-adversarial flows. *ICLR*, 2026.

Harris Drucker and Yann Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992. doi: 10.1109/72.165600.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals, 2025. URL https://arxiv.org/abs/2510.27684.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023.

Risheek Garrepalli, Shweta Mahajan, Munawar Hayat, and Fatih Porikli. Ddil: Diversity enhancing diffusion distillation with imitation learning, 2025. URL https://arxiv.org/abs/2410.11971.

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

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

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

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 *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

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. *ICLR*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

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

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching, 2025. URL https://arxiv.org/abs/2503.06674.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. *CVPR*, 2023.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 833–840, 2011.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia*, SA '24, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687625. URL https://doi.org/10.1145/3680528.3687625.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, pp. 87–103, 2024b.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *ICLR*, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Peng Sun and Tao Lin. Any-step generation via n-th order recursive consistent velocity field estimation. *ICLR*, 2026.

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 *NeurIPS*, 2024.

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, 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.

Zihan Yang, Shuyuan Tu, Licheng Zhang, Qi Dai, Yu-Gang Jiang, and Zuxuan Wu. ArcFlow: Unleashing 2-step text-to-image generation via high-precision non-linear flow distillation. *arXiv preprint arXiv:2602.09014*, 2026.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *https://arxiv.org/abs/2508.09987*, 2025a.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025b.

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 *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.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

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

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency, 2025. URL https://arxiv.org/abs/2510.08431.

A Training Data

For text-to-image distillation, we aggregate prompts from the publicly available ShareGPT-4o-Image Chen et al. (2025) (excluding its editing split), Echo-4o Ye et al. (2025a), and PiFlow Chen et al. (2026) datasets, and randomly select 500k prompts from the resulting pool. For Qwen-Image-Edit, we use the editing split of ShareGPT-4o-Image.

B Implementation and Evaluation Details

Model and optimization. The student is initialized from the pretrained teacher. We fine-tune rank-64 LoRA adapters (Hu et al., 2022), with scaling factor $\alpha = 64$, on all attention projection layers of the diffusion transformer. We use AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and weight decay 10^{-2}, without learning-rate warmup or decay. Following DMD training practice (Yin et al., 2024a), the fake-score estimator is initialized from the teacher and updated five times per generator update. All runs use bf16 on 8 GPUs, a per-GPU batch size of 4, and a global batch size of 32. Text-to-image and editing students are trained for 7k and 2k generator iterations, respectively.

SPC schedule. We use $\lambda = 1$ and set the per-step perturbation radius to $r_t = \gamma \,\textsc{Scheduler}(t, i)$ with $\gamma = 0.01$. For training iteration i out of N,

$$\textsc{Scheduler}(t, i) = \frac{1}{2}\left[1 + \cos\left(\frac{i}{N}\pi\right)\right] f(t), \qquad f(t) = t. \tag{15}$$

The perturbation is sampled independently in every latent-state coordinate from $\mathcal{U}([-r_t, r_t])$. The schedule supplies stronger neighborhood regularization at noisier timesteps and early in training, then reduces the radius as the student approaches its final solution.

Evaluation. We use the complete DPG-Bench (Hu et al., 2024) set and its official pipeline, the standard GenEval (Ghosh et al., 2023) prompts and scoring script, and the official HPS-v2.1 (Wu et al., 2023) model. OneIG-Bench (Chang et al., 2025) reports Alignment, Text, Diversity, Style, and Reasoning on its corresponding prompt subsets. ImgEdit (Ye et al., 2025b) and GEdit-EN (Liu et al., 2025) are evaluated with their official codebases. Main-table captions identify values quoted from prior work; otherwise, released checkpoints are evaluated through these pipelines. DMD and PCM are reproduced under the same training and evaluation setup.

C Additional Ablation Results

Regularization weight, radius, and schedule. Table 6 varies the weight balancing $\mathcal{L}_{\mathrm{DMD}}^{\mathrm{sg}}$ and $\mathcal{L}_{\mathrm{SPC}}$. A small weight approaches the stop-gradient DMD baseline, whereas a very large weight slightly reduces DPG-Bench and GenEval despite remaining competitive on HPS-v2.1. Table 6 varies the maximum neighborhood radius and the annealing schedule. A very small radius supplies little sensitivity supervision; a large radius moves beyond the intended local regime and can suppress fine detail. The intermediate radius $\gamma = 0.01$ with cosine annealing gives the best overall balance, consistent with the inverted-U analysis in Fig. 4c.

Comparison with generic regularizers. Input perturbation (Input Pert.) applies the same noise to the DMD input without matching paired endpoint predictions; output-norm shrinkage (Output Norm) directly penalizes prediction magnitude. We additionally consider a one-sided SPC variant that stops gradients through the perturbed prediction $G_\theta(x_t + \epsilon, t)$. This treats the perturbed branch as a fixed pseudo-target rather than symmetrically regularizing the local neighborhood. Input Pert. and Output Norm generally reduce few-step quality, while the one-sided variant remains comparable to the baseline and full SPC consistently improves it. These controls show that the symmetric pairing in Eq. (11), rather than perturbation, output shrinkage, or a one-sided target correction alone, is essential.

Table 7: Diversity & coverage on Qwen-Image (higher is better). SPC does not noticeably reduce diversity.

Method	Vendi$_i$	Vendi$_g$	Pre.	Rec.	Den.	Cov.	DSim
Teacher	2.089	274.84	1.000	1.000	1.000	1.000	0.189
DMD	1.521	241.58	0.941	0.224	1.060	0.724	0.117
+SPC	**1.562**	**247.18**	**0.951**	**0.252**	**1.078**	**0.783**	**0.128**
PCM	1.787	**275.45**	**0.940**	**0.495**	**1.053**	0.831	0.176
+SPC	**1.814**	275.21	0.934	0.492	1.051	**0.836**	**0.182**

Table 8: Generic regularizers vs. SPC: input perturbation, output-norm shrinkage and stop-gradient all hurt.

Model	NFE	DPG	GE	Aln.	Txt	Div.	Sty.	Rsn.
DMD	2	87.76	0.86	0.859	0.861	0.113	0.414	0.277
Input Pert.	2	85.02	0.85	0.837	0.726	0.104	0.374	0.285
Output Norm	2	85.93	0.87	0.854	0.800	0.107	0.392	0.281
sg$(G_\theta(x_t+\varepsilon,t))$	2	87.75	0.87	0.861	0.859	0.113	0.416	0.284
SPC	2	**88.25**	**0.88**	**0.876**	**0.877**	**0.119**	**0.423**	**0.287**
DMD	4	87.86	0.87	0.862	0.864	0.117	0.411	0.290
Input Pert.	4	86.15	0.85	0.857	0.863	0.116	0.388	0.293
Output Norm	4	86.58	0.86	0.859	0.873	0.115	0.397	0.289
sg$(G_\theta(x_t+\varepsilon,t))$	4	87.99	0.88	0.868	0.871	0.115	0.409	0.295
SPC	4	**88.43**	**0.88**	**0.882**	**0.911**	**0.118**	**0.423**	**0.309**

Diversity and distribution coverage. Because local regularization can in principle trade coverage for precision, Table 7 reports complementary distribution-level metrics. Adding SPC to DMD improves both Vendi variants (Friedman & Dieng, 2023), Precision, Recall, Density, Coverage, and DreamSim (Fu et al., 2023). The PCM pair is essentially unchanged: some metrics move slightly downward and others upward. These results do not support the hypothesis that SPC's quality gains are obtained through additional mode contraction, but they also do not imply that every diversity statistic must improve.

D Detailed Derivations

D.1 Multi-Step Endpoint Gradient

For the transition in Eq. (1), differentiation with respect to θ gives

$$\nabla_\theta x^\theta_{t_{i+1}} = J_{\phi_i^\theta} \nabla_\theta x^\theta_{t_i} + \beta_i \frac{\partial G_\theta(x^\theta_{t_i}, t_i)}{\partial \theta}, \qquad J_{\phi_i^\theta} = \alpha_i I + \beta_i \nabla_x G_\theta(x^\theta_{t_i}, t_i). \tag{16}$$

The initial noise is independent of θ, so $\nabla_\theta x^\theta_{t_1} = 0$. Unrolling through the last intermediate state yields

$$\nabla_\theta x^\theta_{t_K} = \sum_{k=1}^{K-1} \left(\prod_{j=k+1}^{\overleftarrow{K-1}} J_{\phi_j^\theta} \right) \beta_k \frac{\partial G_\theta(x^\theta_{t_k}, t_k)}{\partial \theta}, \tag{17}$$

where an empty product is the identity. Applying the chain rule to $x_0^\theta = G_\theta(x^\theta_{t_K}, t_K)$ and substituting Eq. (17) gives Eq. (6).

D.2 Two-Step SGTS Gap

For $K = 2$, $x^\theta_{t_2} = \alpha_1 x^\theta_{t_1} + \beta_1 G_\theta(x^\theta_{t_1}, t_1)$ and $x_0^\theta = G_\theta(x^\theta_{t_2}, t_2)$. Hence

$$\nabla_\theta x_0^\theta = \frac{\partial G_\theta(x^\theta_{t_2}, t_2)}{\partial \theta} + \nabla_x G_\theta(x^\theta_{t_2}, t_2)\, \beta_1 \frac{\partial G_\theta(x^\theta_{t_1}, t_1)}{\partial \theta}. \tag{18}$$

SGTS retains the local update at whichever step is supervised and removes the cross-step credit-assignment term from that update. Multiplying the omitted term by the DMD score difference and taking its expectation gives Eq. (9). This derivation identifies the Jacobian-mediated term absent from optimization; it does not claim that SPC reconstructs the direction of the full-path gradient.

D.3 Transition Bound and Error Propagation

Using the spectral norm and the triangle inequality,

$$\|J_{\phi_i^\theta}\|_2 \le |\alpha_i| + |\beta_i| \, \|\nabla_x G_\theta(x_{t_i}^\theta, t_i)\|_2. \tag{19}$$

Thus Eq. (10) follows when the endpoint-map spectral norm is bounded by L. SPC targets the Frobenius norm, which provides an upper bound because $\|A\|_2 \le \|A\|_F$. We state the transition inequality for the spectral norm specifically; writing the same final equality for the Frobenius norm would introduce the dimension-dependent factor $\|I\|_F = \sqrt{d}$.

For a perturbation η_i at step i, first-order expansion gives $\eta_{i+1} \approx J_{\phi_i^\theta}\eta_i$. Recursion then produces Eq. (4). The norm of the perturbation is controlled by the product of the remaining transition gains. A single factor above one need not cause exponential growth; sustained high gain across many remaining transitions can. This is why the main claim is control of amplification, not strict contraction of every step.

D.4 SPC as Implicit Jacobian Regularization

Let $J = \nabla_x G_\theta(x_t, t)$. The Taylor expansion in Eq. (12) gives

$$\|G_\theta(x_t, t) - G_\theta(x_t + \varepsilon, t)\|_2^2 \approx \varepsilon^\top J^\top J \varepsilon. \tag{20}$$

For independent $\varepsilon_j \sim \mathcal{U}([-r_t, r_t])$, $\mathbb{E}[\varepsilon\varepsilon^\top] = (r_t^2/3)I$. Therefore

$$\mathbb{E}_\varepsilon[\varepsilon^\top J^\top J \varepsilon] = \frac{r_t^2}{3} \operatorname{tr}(J^\top J) = \frac{r_t^2}{3}\|J\|_F^2, \tag{21}$$

which yields Eq. (13). The equivalence is first-order and local: for a large radius, higher-order terms matter and the loss may suppress valid variation rather than only ill-conditioned sensitivity.

E More Analysis

E.1 Jacobian and Amplification Diagnostics

Squared Frobenius estimator. The full Jacobian is too large to materialize. For M independent probes $v_m \sim \mathcal{N}(0, I)$, we use

$$\|\nabla_x G_\theta(x_t, t)\|_F^2 \approx \frac{1}{M} \sum_{m=1}^{M} \|\nabla_x G_\theta(x_t, t) v_m\|_2^2. \tag{22}$$

Each term is evaluated with a Jacobian–vector product. We use $M = 5$ probes per sample, average over a fixed set of 64 samples, and measure every 500 generator iterations. Accordingly, Fig. 4a reports an estimate of the *squared* Frobenius norm, not the Frobenius norm itself or a spectral-norm certificate.

Interpretation. The SGTS baseline remains trainable while its estimate grows from 1965 to 2123, demonstrating that optimization stability does not by itself bound local sensitivity. Full-path training rises from 2092 to 4342 before its observed collapse. SGTS+SPC instead remains near 1450. These measurements support the proposed sensitivity mechanism but should not be read as proving that Jacobian size alone determines perceptual quality; Fig. 4c explicitly shows the adverse effect of excessive regularization.

For perturbation amplification, we add the same-norm random perturbation to a chosen intermediate latent, continue the remaining student transitions, and report $\mathrm{PA} = \|x_0^{\mathrm{pert}} - x_0\|_2 / \|\varepsilon\|_2$. PA is an unsquared norm ratio, with no additional squaring or square-root operation. Repeating the measurement at each injection location probes the composed gain that is active at inference, complementary to the local Jacobian estimate above.

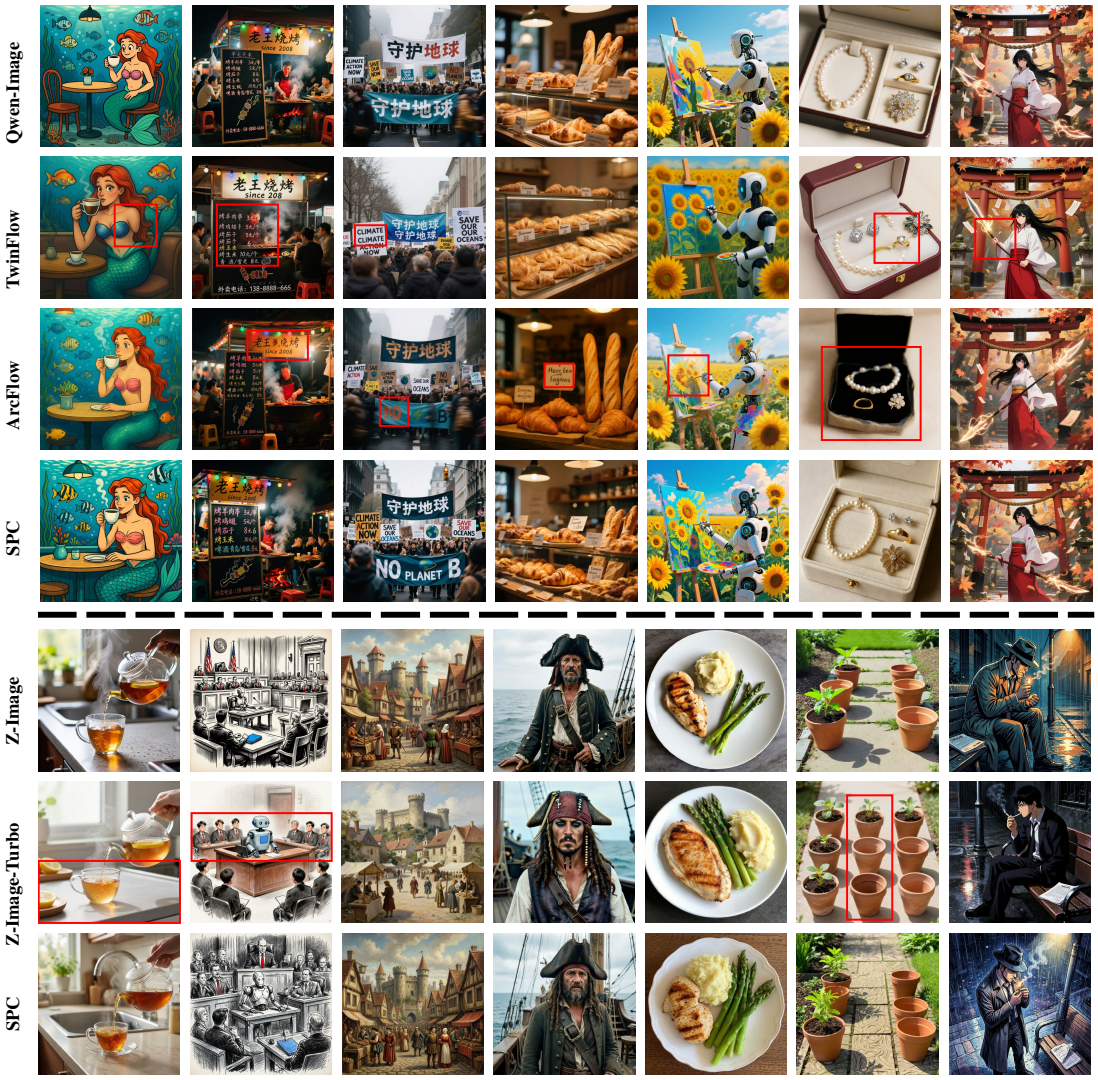


Figure 6: Visual comparison of SPC against baselines on Qwen-Image-20B (NFE=2, top four rows) and Z-Image (NFE=4, bottom three rows). Red rectangles show manifest artifacts or inconsistency.

E.2 Integration with Phased Consistency Models

PCM Wang et al. (2024) partitions the probability-flow trajectory into phases and supplies supervision between states associated with that trajectory. SPC adds a different signal: fixed-timestep sensitivity in isotropic local state-space directions. We integrate it without changing the PCM target or rollout construction:

$$\mathcal{L}_{\text{PCM+SPC}} = \mathcal{L}_{\text{PCM}} + \lambda \sum_{i=1}^{K} \mathcal{L}_{\text{SPC}}(t_i). \tag{23}$$

We retain all PCM hyperparameters and use the same $\lambda = 1$, $\gamma = 0.01$, and cosine schedule as for DMD. On four-step Qwen-Image, SPC improves PCM from 87.58 to 88.02 DPG and from 28.01 to 30.96 HPS-v2.1, while improving all five OneIG dimensions (Table 2). The result shows that local neighborhood conditioning is compatible with a trajectory-consistency objective; it does not imply that the two objectives are interchangeable.

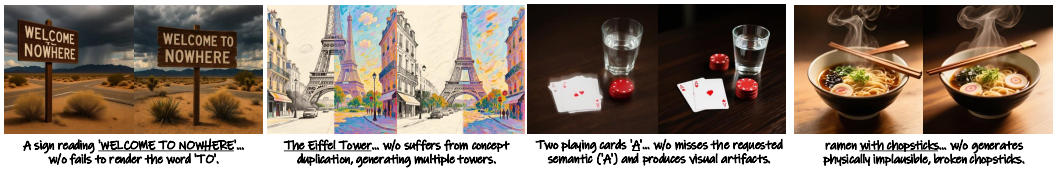


Figure 7: Same-prompt comparison of the stop-gradient DMD baseline with and without SPC on Qwen-Image.

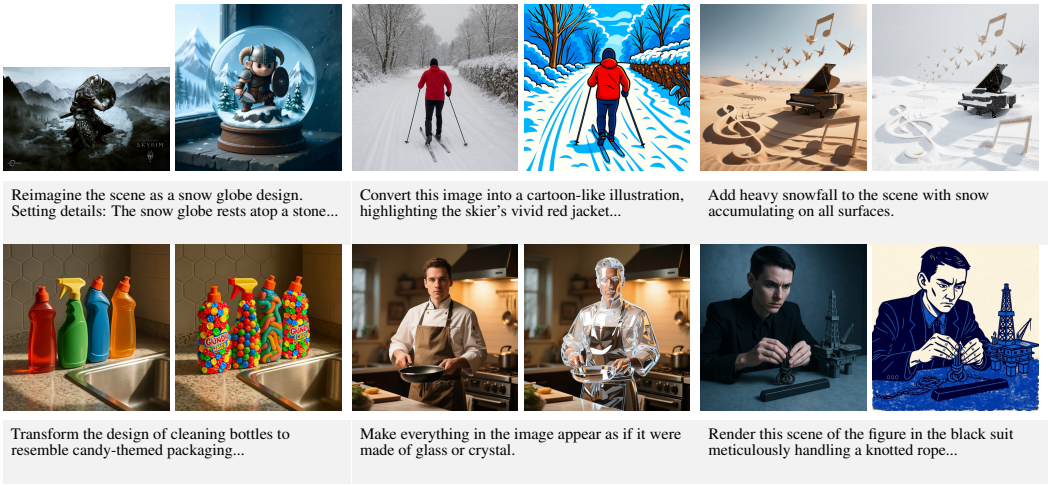


Figure 8: Image editing results of SPC on Qwen-Image-Edit (NFE=4). Each triplet shows the source image, editing instruction, and output.

F Additional Qualitative Results

Fig. 7 compares the same prompts with and without SPC. The remaining pages (Fig. 9) provide additional four-step Qwen-Image examples.

In Fig. 7, SPC reduces structural defects, object blending, and illegible text relative to TwinFlow and ArcFlow on Qwen-Image, while remaining visually closer to the teacher. These errors are consistent with the proposed mechanism: a malformed thin structure, repeated object, or broken character introduced upstream can be reinforced by later high-gain transitions.

Moreover, we also provide qualitative result on image editing in Fig. 8.

G Limitations and Broader Impact

Limitations. SPC introduces a neighborhood scale γ that must be calibrated. Large perturbations weaken the local Taylor interpretation and may over-smooth fine detail, while very small perturbations provide a weak signal. Although the diversity study finds no additional mode contraction in our settings, local regularization always creates a potential quality–coverage trade-off and should be re-evaluated for new backbones and domains. Our main experiments cover DMD and PCM on Qwen-Image, Z-Image, and Qwen-Image-Edit; validation on progressive or adversarial distillation (Salimans & Ho, 2022; Sauer et al., 2024b) would further test generality. Finally, the Frobenius estimator measures average local sensitivity rather than the worst-case spectral gain of the complete nonlinear rollout.

Broader impact. Few-step generation can reduce deployment cost and energy use, but faster image generation also lowers the cost of misuse, including misinformation and deceptive media. Accelerated generators should be deployed together with appropriate provenance, watermarking, and content-safety mechanisms.



Figure 9: Additional four-step text-to-image results on Qwen-Image (Part I).



Figure 9: Additional four-step text-to-image results on Qwen-Image (Part II).



Figure 9: Additional four-step text-to-image results on Qwen-Image (Part III).