PAST-SELF GUIDANCE: INTERNALIZING TEMPORAL CONTRAST IN FLOW MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training produces a history of predictors, but generation usually uses only its final model. At sampling time, adding the component of a current–past prediction difference perpendicular to the current prediction improves samples but not regression loss. We introduce Past-Self Guidance (PSG), which adds the full prediction difference to the flow-matching target. Both predictions use the same noisy input, one from the model's recent exponential moving average and one from an earlier ordinary-training checkpoint. Sampling needs only the trained model. On ImageNet 256×256, one setting improves FID-50k by 34–55% across four SiT scales at 400k updates, without classifier-free guidance (CFG). At XL/2, FID falls from 16.89 to 7.70 at similar recall. PSG also has lower FID than ordinary training that uses about twice as much compute. Applied only in the last 10% of a 2M-update run, PSG reaches CFG-free FID-50k 5.73, against 10.10 for ordinary training. It needs that run's earlier checkpoints and about 6% more training compute. Activated at 1M updates, it reaches FID-50k 1.99 with CFG at 2M. It combines with REPA and improves diffusion and unconditional generation. Ablations and sampling interventions show that the gain comes mainly from the direction of this difference and disappears when the difference is taken from another sample.

1 INTRODUCTION

Flow models learn to generate images by predicting velocities that transport noise toward data. Scaling the model and training it longer improve generation quality, but require substantial training compute (Ma et al., 2024). Guidance offers another route: it improves samples by contrasting predictions at sampling time. Classifier-free guidance (CFG) uses conditional and unconditional predictions (Ho & Salimans, 2021), while Autoguidance can contrast a model with an earlier checkpoint from the same training run (Karras et al., 2024a). These contrasts typically require an auxiliary model evaluation at every sampling step.

Training produces a sequence of predictors, but ordinary flow matching does not learn from their differences. At a fixed noisy input, the difference between a recent and a past prediction shows how the learned velocity has changed. Extrapolating along its angular component, perpendicular to the current prediction, improves samples but not regression loss. For four SiT models at 400k updates, FID-10k falls by 8.9–16.1 while the loss rises by 2–3% (Section 5.1). Can a model learn from this difference during training and keep the benefit when sampled alone?

Motivated by this observation, we introduce Past-Self Guidance (PSG), which adds the full prediction difference to the velocity target (Figure 1). Once activated after a phase of ordinary training, it uses two teachers that receive no gradients: a recent exponential moving average (EMA) of the student and an earlier checkpoint of ordinary training. The student does not copy either teacher; it learns the flow-matching target plus their scaled difference. Both teachers see the student's noisy input. Only the target changes, and sampling uses the final EMA alone. At an idealized stationary point, the learned field takes the form of guidance away from the past model (Section 3.3).

Guidance methods differ in where the contrast comes from. Model Guidance (MG; Tang et al., 2025) trains with conditional–unconditional predictions, while Internal Guidance (IG; Zhou et al., 2026) contrasts intermediate and final network outputs. PSG uses saved checkpoints instead, with no unconditional teacher or extra prediction head.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

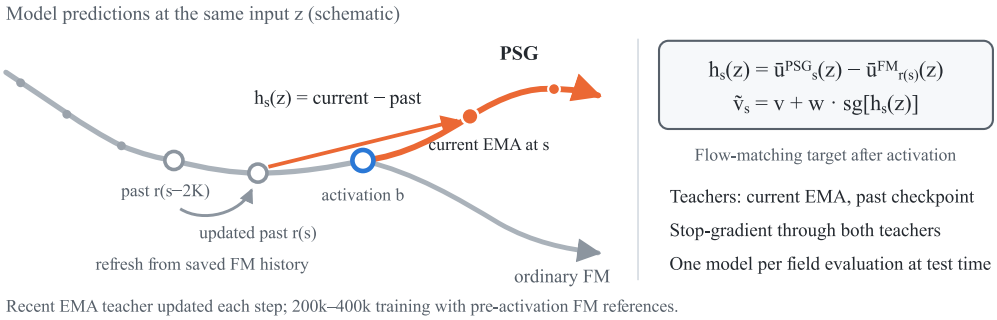

Figure 1: **PSG trains a temporal contrast into a single predictor.** The recent teacher (the current EMA) and the past teacher (an earlier ordinary-training EMA) are evaluated at the same input. Their difference, with gradients stopped, modifies the velocity target. The past teacher advances through saved ordinary-training checkpoints; sampling uses only the final EMA model. Here the recent EMA teacher is updated each step; in Table 1, it is a copy refreshed every 10k updates.

Our contributions are as follows:

- We introduce PSG, which adds the difference between a model's recent and past predictions to the flow-matching target and needs no extra model at inference.
- On ImageNet 256×256, one setting lowers CFG-free FID-50k by 34–55% across four SiT scales at 400k updates. At XL/2, PSG has lower FID than ordinary training that uses about twice as much compute. Applied only in the last 10% of a 2M-update run, it reaches CFG-free FID-50k 5.73, against 10.10 for ordinary training, at about 6% more training compute. It combines with REPA (Yu et al., 2025) and CFG, and also improves diffusion and unconditional generation.
- We trace most of the gain to the direction of the difference, computed for the sample it corrects. In short training continuations, the network's response to this direction is smaller and less aligned than its response to random ones.

2 Preliminaries

Flow matching learns a velocity field along paths between data and noise (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2023). For a data latent x, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and $t \sim \mathcal{U}[0, 1]$, the linear path is $x_t = (1 - t)x + t\epsilon$ with velocity target $v = \epsilon - x$. With $z = (x_t, t, c)$, where c is an optional class label, the network minimizes

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\|u_\theta(z) - v\|^2.$$

The target v depends on the sampled data–noise pair; regression estimates its conditional mean given z. Generation follows the learned field from $t = 1$ to $t = 0$, using an ODE or the corresponding SDE sampler (Ma et al., 2024).

Prediction-contrast guidance instead samples with a modified field of the form

$$u_\gamma(z) = u_{\mathrm{main}}(z) + \gamma\,[u_{\mathrm{main}}(z) - u_{\mathrm{ref}}(z)].$$

The two predictions use the same noisy state and time. For CFG, the reference omits the class condition; for temporal guidance, it comes from an earlier model. The guidance weight γ is zero without guidance; the conventional CFG scale is $1 + \gamma$. PSG instead adds a temporal contrast to the training target, with a separate training weight w. We use s for the optimizer update index, distinct from the interpolation time t: the teachers differ in training age, not in noise level.

3 Past-Self Guidance

PSG trains on the difference between the student's recent EMA and an earlier checkpoint of ordinary training (Figure 1). Both teachers receive the student's current noisy input.

3.1 A TRAINING TARGET FROM MODEL HISTORY

Bars denote predictions made with exponentially averaged weights, and superscripts name the run (PSG or ordinary flow matching, FM). At update s, the recent teacher is $\bar{u}^{\mathrm{PSG}}_{q(s)}$ and the past teacher $\bar{u}^{\mathrm{FM}}_{r(s)}$, with $q(s) \le s$ and $r(s) < s$ (Section 3.2).

After activation at step b, their difference h_s, which we call the temporal secant, modifies the target:

$$\begin{aligned} h_s(z) &= \bar{u}^{\mathrm{PSG}}_{q(s)}(z) - \bar{u}^{\mathrm{FM}}_{r(s)}(z), \\ \tilde{v}_s(z) &= v + w\,\mathrm{sg}[h_s(z)], \\ \mathcal{L}_{\mathrm{PSG}} &= \mathbb{E}\|u_\theta(z) - \tilde{v}_s(z)\|^2. \end{aligned} \tag{1}$$

The weight w sets the training dose, and sg stops gradients through both teachers. Before activation, the target is simply v. The default keeps the full prediction difference; its directional components are ablated in Section 4.4. Class labels enter the usual model input, with no conditional–unconditional contrast in the PSG target.

3.2 TEACHER SCHEDULE AND COST

The recent teacher is either the student's EMA itself ($q(s) = s$) or a copy refreshed every 10k updates ($q(s) = 10\mathrm{k}\lfloor s/10\mathrm{k} \rfloor$). The past teacher is an ordinary-training checkpoint at about ρ times the current step, $r(s) = K\lfloor \rho s/K \rfloor$, taken from checkpoints saved every K updates. We use $w = .5$, $\rho = .5$, $K = 10\mathrm{k}$, and $b = 200\mathrm{k}$ at all four SiT scales, so the past teacher is the model at about half the current step. Before b, the run is ordinary training, so its own checkpoints serve as past teachers. With $\rho = .5$, these stay at or before b as long as training ends by $2b$. Longer training needs ordinary-training checkpoints after b (Appendix A).

Training adds two teacher forward passes without backpropagation while PSG is active; at XL/2, an active update takes about 1.6 times as long as an ordinary update. A full XL/2 run is thus estimated to cost about 1.3 ordinary runs when PSG is active for half the updates, and about 1.06 when it covers only the last 10% (Appendix A). At inference, only the final EMA remains: SDE-250 uses 250 model calls, compared with 500 for CFG or a two-model contrast at every step.

3.3 A STATIONARY INTERPRETATION OF THE TARGET

Let $\mu(z) = \mathbb{E}[v \mid z]$ be the population minimizer of ordinary flow matching. Suppose the recent teacher equals the student u and the past teacher is fixed. A stationary point of the stop-gradient update in Eq. (1) then satisfies $u = \mu + w\,(u - u_{\mathrm{past}})$ pointwise, which for $0 \le w < 1$ gives

$$u^* = \mu + \frac{w}{1-w}\,(\mu - u_{\mathrm{past}}). \tag{2}$$

This has the form of sampling-time guidance away from the past model. Its weight is $w/(1-w)$ rather than w because the target feeds back through the student, as in MG; $w = .5$ gives weight 1. The idealization gives the form of the correction, not its FID effect: μ already generates the data distribution, and actual PSG uses moving teachers.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate SiT-S/2, B/2, L/2, and XL/2 (Ma et al., 2024) on ImageNet 256×256. Our models use the Stable Diffusion VAE (Rombach et al., 2022), with ft-EMA encoding and ft-MSE decoding. Training follows the SiT settings in Table 8. The main comparisons refresh the recent teacher every 10k updates and evaluate EMA weights with SDE-250 and 50,000 samples. The B/2–XL/2 dose sweeps and the past-checkpoint-spacing sweep update the recent teacher each step and use 10,000 samples; captions state all other schedules. Updating each step avoids a stored teacher copy and is recommended for new runs. The ADM evaluator (Dhariwal & Nichol, 2021) measures FID, sFID, Inception Score, and precision and recall on Inception features; evaluation seeds and reference statistics are shared within each controlled comparison (Appendix A). Guided FID is the best FID at the tested CFG scales above 1.

4.2 Scaling and training duration

The same PSG setting improves models from 33M to 675M parameters (Table 1). At 400k updates, CFG-free FID-50k falls by 34–55%, and recall changes by $-.029$ to $+.034$.

Table 1: **One PSG setting improves FID from S/2 through XL/2.** Every pair is trained for 400k updates on ImageNet 256×256 and evaluated with SDE-250, 50k samples, and no CFG. PSG refreshes its recent EMA teacher every 10k updates, using $w = .5$ and activation at 200k for all scales. Bold marks the better value within each matched pair.

Model	Training	Params	FID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
SiT-S/2	SiT	33M	59.63	9.83	23.9	0.393	0.597
	+ PSG		**39.28**	**6.69**	**36.9**	**0.502**	**0.631**
SiT-B/2	SiT	130M	36.31	6.38	41.4	0.519	**0.645**
	+ PSG		**19.43**	**5.29**	**70.9**	**0.636**	0.616
SiT-L/2	SiT	458M	20.75	5.12	66.6	0.624	**0.628**
	+ PSG		**9.32**	**4.60**	**111.1**	**0.718**	0.610
SiT-XL/2	SiT	675M	16.89	4.86	77.2	0.653	**0.620**
	+ PSG		**7.70**	**4.56**	**122.7**	**0.729**	0.614

Additional ordinary updates. Spending more compute on ordinary training does not close the gap. An ordinary XL/2 run of 480k updates, slightly cheaper than PSG at 400k (Section 3.2), reaches FID-50k 16.09 against 7.70 (Table 2). With about twice PSG's compute, 1M ordinary updates reach FID-10k 14.69, against 10.14 for PSG at 400k with the recent teacher updated each step (Table 16).

Late activation. Applied only in the last 10% of a 2M-update XL/2 run, PSG reaches CFG-free FID-50k 5.73, against 10.10 for ordinary training, at similar recall (Table 17). It needs the run's own earlier checkpoints and about 6% more training compute (Appendix A). Activating at 1M or 1.5M instead reaches FID-50k 5.35 or 5.76 at 2M. Figure 2 shows the three PSG runs with FID-10k.

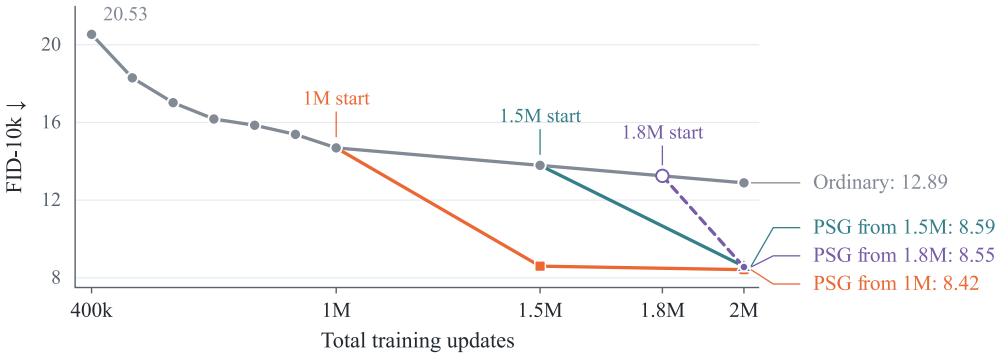

Figure 2: **PSG improves FID even when activated only for the last 10% of training.** XL/2, SDE-250, no CFG; FID-10k here, FID-50k in the text. PSG updates the recent teacher each step, with $w = .5$, $\rho = .5$, and past-checkpoint spacing $K = 10$k. Filled markers are measurements; straight lines only join them and do not trace FID between checkpoints; the axis counts updates, not compute. The open 1.8M starting point is not measured; it is placed by linear interpolation between the ordinary FID values at 1.5M and 2M. The dashed line connects it to the measured PSG endpoint.

4.3 Combination and comparison with other methods

Representation alignment. PSG modifies output targets, while REPA aligns hidden features. Adding REPA to PSG improves FID-50k at S/2, B/2, and XL/2 (Table 21). At XL/2, adding PSG to REPA lowers FID-10k from 10.49 to 6.08 without CFG (Table 16). With CFG, it lowers FID-50k from 2.84 to 2.29, each model at its best tested scale (Table 2). PSG and REPA+PSG also lower the CFG-free DINOv2 Fréchet distance relative to SiT at XL/2 (Table 24).

Table 2: **PSG improves quality at 400k and combines with REPA and CFG.** ImageNet 256×256, FID-50k, SDE-250 (250 field evaluations). PSG refreshes its recent EMA teacher every 10k updates, with $b = 200$k, $\rho = .5$, and $K = 10$k. Each row is one trained model, with its best tested CFG point on the right (grids in Table 19). The 480k row uses slightly less training compute than PSG (Appendix A). Bold: best displayed 400k value; dashes: not evaluated. Published REPA results are in Table 3.

	No CFG: 1 call/field		With CFG: 2 calls/field		
XL/2 at 400k	**FID ↓**	**Rec. ↑**	**Scale**	**FID ↓**	**Rec. ↑**
SiT	16.89	.620	1.50	5.15	.493
+ REPA	–	–	1.25	2.84	.589
+ PSG ($w = .5$)	7.70	.614	1.25	3.69	.538
+ REPA + PSG ($w = .5$)	**3.28**	**.638**	1.10	**2.29**	**.612**
SiT, 480k updates	16.09	.638	–	–	–

Training-time versus sampling-time contrast. Applying the contrast at sampling time instead reaches similar or lower FID at similar recall, but needs two model calls per field evaluation. This sampling-time composite projects the 400k–200k prediction difference perpendicular to the 400k prediction and adds α times the result to the ordinary 400k model (Appendix C). At $\alpha = 2$, chosen to match PSG's FID rather than taken from Eq. (2), it reaches FID-50k 7.45 at recall .617, against 7.70 and .614 for default PSG (Table 18; Figure 3). At $\alpha = 3$, its best measured setting, it reaches 6.37 at recall .616.

Conditional contrast. A conditional training contrast has far lower FID-10k than PSG at S/2 (18.23 against 42.01) but lower recall (.536 against .636; Table 15). It also needs class labels (Table 22).

Classifier-free guidance. A higher training dose leaves less room for CFG: CFG 1.25 improves PSG at $w = .5$ from 7.70 to 3.69 but worsens PSG at $w = .75$ from 3.94 to 4.67 (Figure 3).

System-level comparison. With CFG, PSG activated at 1M, 1.5M, or 1.8M reaches FID 1.99–2.11 at 2M (Table 3), comparable to the 2.06 reported for SiT-XL/2 after 7M updates under a different protocol.

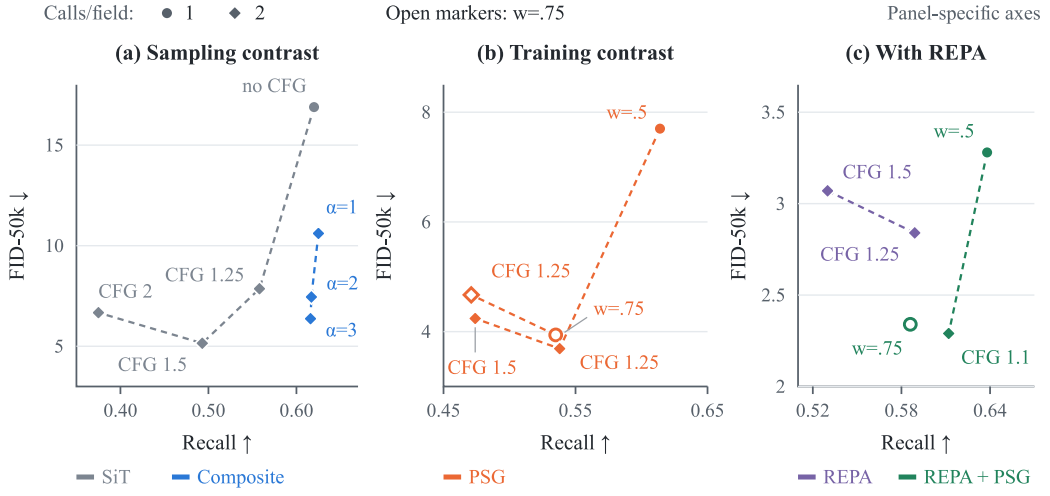

Figure 3: **Training dose and CFG both trade recall for fidelity, at different call costs.** XL/2 at 400k, SDE-250; FID-50k and recall under the same evaluator. (a) SiT and the sampling-time composite; (b) PSG; (c) REPA and REPA+PSG. Axis ranges differ. Circles use one model call per field evaluation and diamonds two (CFG or the composite). Open markers identify $w = .75$; default PSG uses $w = .5$. Dashed lines join sampling settings of the same model or fixed checkpoint pair, not different training doses.

Table 3: **Quality and coverage across published systems and training budgets.** ImageNet 256×256, 50k samples. Each row reports all metrics at one operating point; †values are transcribed from the cited papers. Our PSG rows use SDE-250, with training settings in Tables 2 and 17. *Includes interval guidance. IG uses internal guidance through two heads sharing one backbone, including in the no-CFG block. Training settings and SD-VAE configurations differ across published systems (Appendix C.3); n.r. denotes an unreported training budget.

Model (XL/2)	Training	CFG	FID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
No CFG: one backbone evaluation per field							
DiT† (Peebles & Xie, 2023)	7M	–	9.62	6.85	121.50	.67	.67
SiT† (Ma et al., 2024)	7M	–	8.26	6.32	131.65	.68	.67
REPA† (Yu et al., 2025)	400k	–	7.9	5.06	122.6	.70	.65
REPA†	4M	–	5.90	5.73	157.8	.70	.69
IG† (Zhou et al., 2026)	800 epochs	–	1.75	4.98	228.6	.80	.62
MG† (Tang et al., 2025)	n.r.	–	1.34	4.58	321.5	.81	.65
Self-Flow† (Chefer et al., 2026)	4M	–	5.70	4.97	151.40	.72	.67
PSG ($w = .5$)	400k	–	7.70	4.56	122.7	.729	.614
REPA + PSG ($w = .5$)	400k	–	3.28	4.59	177.0	.760	.638
With CFG: two backbone evaluations on guided steps							
DiT†	7M	1.50	2.27	4.60	278.24	.83	.57
SiT†	7M	1.50	2.06	4.49	277.50	.83	.59
REPA†	4M	1.80*	1.42	4.70	305.7	.80	.65
IG†	800 epochs	1.35*	1.46	4.79	265.7	.80	.64
PSG ($w = .5$)	400k	1.25	3.69	4.54	196.8	.818	.538
REPA + PSG ($w = .5$)	400k	1.10	2.29	4.53	215.5	.795	.612
PSG ($b = 1$M)	2M	1.25	1.99	4.45	246.4	.808	.605
PSG ($b = 1.5$M)	2M	1.25	2.02	4.38	243.3	.805	.606
PSG ($b = 1.8$M)	2M	1.25	2.11	4.38	239.0	.810	.604

4.4 Ablation studies

Which part of history matters? Most of the gain comes from the direction of the difference. As in adaptive projected guidance (Sadat et al., 2025), we split the difference into an angular component, which changes the direction of the predicted velocity, and a parallel component, which rescales it (Appendix D.1). We call the angular component the history direction. At S/2, angular-only training retains 89.8% of the full target's FID gain; the parallel component and a norm-matched random target retain little of it, and reversing the sign makes FID worse than ordinary training (Table 4).

Table 4: **The angular component retains 89.8% of PSG's FID improvement.** Matched S/2 interventions at 400k with 10k samples: activation at 200k, recent EMA teacher refreshed every 10k updates, $w = .5$, $\rho = .5$, and 10k past-checkpoint spacing. Reversing history increases FID; the norm-matched random direction leaves FID close to the ordinary baseline. Bold marks the best FID and recall.

Training target	FID-10k ↓	Recall ↑	FID gain retained
Ordinary	62.48	0.558	0%
Full PSG	**42.01**	**0.636**	100%
Angular component only	44.10	0.627	89.8%
Parallel component only	60.59	0.554	9.2%
Norm-matched random	62.11	0.546	1.8%
Reversed history	69.33	0.510	−33.5%

How does dose transfer across scales? At B/2, L/2, and XL/2, a higher dose lowers CFG-free FID at some cost in recall, and $w = .75$ also worsens guided FID (Table 5). At XL/2, raising w from .5 to .6 lowers CFG-free FID-10k from 10.14 to 7.77 and recall from .691 to .668, with guided FID

nearly unchanged (6.45 and 6.46). We keep $w = .5$ for comparisons and recommend $w = .6$, with the recent teacher updated each step, for CFG-free use when its recall cost is acceptable. A denser B/2 grid and a 50k evaluation support this choice (Appendix B).

Table 5: **Dose choice depends on sampling guidance.** B/2, L/2, and XL/2 checkpoints at 400k, with 200k activation, recent EMA teacher updated each step, and 10k samples. The right block reports the lowest measured FID at CFG scales greater than one. *Lowest tested scale; the XL/2 $w = .75$ grid is $\{1.25, 1.5\}$. Bold marks the best observed value within each model size.

		CFG-free		**Best measured CFG** > 1		
Model	w	**FID** $\downarrow$	**Rec.** $\uparrow$	**CFG**	**FID** $\downarrow$	**Rec.** $\uparrow$
B/2	.50	21.37	**.674**	1.50	8.75	**.573**
	.60	16.46	.659	1.50	**8.38**	.549
	.75	**11.95**	.617	1.25	9.29	.558
L/2	.50	11.56	**.679**	1.25	6.92	**.618**
	.60	8.89	.660	1.25	**6.85**	.597
	.75	**7.78**	.596	1.10*	7.72	.574
XL/2	.50	10.14	**.691**	1.25	**6.45**	.624
	.60	7.77	.668	1.10*	6.46	**.641**
	.75	**7.17**	.608	1.25*	8.57	.546

Spacing versus dose. Over the tested ranges at B/2, the past-checkpoint spacing K mainly moves CFG-free FID, while the dose also moves guided FID (Table 11). At $w = .5$, widening K from 10k to 100k lowers CFG-free FID from 21.37 to 16.69 with guided FID between 8.63 and 8.79; at $K = 100$k the past teacher stays at step 100k. With $K = 10$k, $w = .6$ reaches a similar CFG-free FID (16.46) with guided FID 8.38.

Teacher and activation settings. Older past checkpoints help up to a point. At S/2, a past teacher at a quarter of the current step ($\rho = .25$) rather than half improves FID-10k from 42.01 to 34.33, but one at a tenth ($\rho = .10$) is worse than ordinary training (Table 12). Appendix B also varies the recent teacher's averaging and the activation step.

Fixed teachers. A fixed-teacher target adds λ times the fixed 200k–100k history direction of ordinary training (Appendix B). At B/2 and XL/2, PSG at $w = .6$ has lower CFG-free FID and higher recall than every tested fixed-teacher setting; at B/2 it reaches FID-10k 16.46, against 19.29 for the best fixed setting. At B/2, L/2, and XL/2, fixed teachers with $\lambda \geq 2.5$ have lower CFG-free FID than PSG at $w = .5$, but lower recall.

4.5 Diffusion, unconditional generation, and learning-rate decay

The same temporal target also improves ϵ-prediction and unconditional generation (Table 22). For DiT-B/2 (Peebles & Xie, 2023), PSG modifies the noise-prediction target while retaining the learned-variance objective (Nichol & Dhariwal, 2021) and DDPM-250 sampler. FID-10k improves from 45.41 to 25.03 (recall .609 to .656). For unconditional SiT-B/2, FID-50k improves from 66.14 to 51.55 (recall .579 to .572). Finally, with cosine learning-rate decay at S/2, PSG improves FID-10k from 70.18 to 54.04 (Table 13).

5 What the temporal contrast contributes

5.1 Does history contain a useful sampling direction?

The history direction improves sampling without reducing the regression loss. The sampling-time composite of Section 4.3 adds the 400k–200k history direction d to a frozen ordinary 400k field u, giving $u_\alpha = u + \alpha d$. This changes the loss by exactly $2\alpha\,\mathbb{E}\langle u - v, d\rangle + \alpha^2\,\mathbb{E}\|d\|^2$ (Appendix D.1). Equal-norm directions have the same quadratic term, so the loss distinguishes directions only through the linear term. At $\alpha = 1.5$, the total loss increases by 2.1–2.8%, yet FID-10k improves by 8.9–16.1 across the four scales (Table 6). A random direction of the same norm raises

the loss by a similar 1.9–2.4% at $\alpha = 1.5$. At $\alpha = 1$, history lowers FID-10k by 6.9–11.3, while the random one changes it by at most .09.

Table 6: **History improves FID despite increasing loss.** Frozen 400k fields, 400k–200k history direction, $\alpha = 1.5$: FID-10k change and relative loss increases. Random directions are norm-matched and change FID-10k by at most .09 at $\alpha = 1$; endpoint controls are in Table 29.

		$\Delta L/L$	
Scale	**ΔFID ↓**	**History**	**Random**
S/2	−16.1	2.3%	2.1%
B/2	−13.5	2.1%	1.9%
L/2	−9.6	2.8%	2.4%
XL/2	−8.9	2.5%	2.3%

Table 7: **Reassigning history corrections removes the sampling gain.** XL/2, 400k, 10k samples; Heun-51 uses 51 field evaluations. History and donor corrections use $\alpha = 1$ at every field evaluation; random corrections match history's total FM-loss increase. Recomputed history uses the current state; replay uses the original trajectory's recorded corrections. Donors preserve class. No CFG; bold marks the lowest FID for each sampler.

	FID ↓	
Correction	**SDE-250**	**Heun-51**
None	20.54	21.30
Recomputed history	**13.29**	**14.53**
Same-trajectory replay	15.44	16.24
Fixed same-class donor	20.86	21.34
Per-evaluation donor	20.62	21.30
Matched random angular	20.56	21.30

5.2 What does the correction encode?

Reassigning history corrections to other samples of the same class removes the sampling gain (Table 7, Figure 5). Replaying corrections recorded on the same sample's uncorrected trajectory still keeps 70% of the gain from recomputing them at the current state (SDE-250). Shuffling differences across the batch during training likewise removes the gain at S/2, although the shuffle also changes their class and noise level (Appendix E.2). History directions from other model sizes still improve the XL/2 field when computed at its sampler's current state and matched in norm (Table 26).

At XL/2, history emphasizes components that changed rapidly during early training. Across 15 combinations of noise level and spatial frequency, the energy per coefficient of the 400k–200k history direction tracks the average angular change of the prediction per update over 50k–200k (Spearman .90). In the three lower-noise time bins, where sample displacement is nearly linear in α, corrections from the fastest-changing bands move samples more than size-matched corrections from the slowest-changing bands (Table 32).

5.3 How does the target turn history into model updates?

Because the teachers receive no gradients, the PSG target adds the gradient term $-2w\,\mathbb{E}[J_\theta(z)^\top h_s(z)]$ to ordinary flow matching, where $J_\theta = \partial u_\theta/\partial\theta$: the prediction difference reaches the weights only through the student's Jacobian. We ask how much of a target change the network adopts. Starting from the ordinary XL/2 400k checkpoint and its optimizer state, we train for 250 updates with and without a change to the target. We then compare the student's raw (unaveraged) predictions: how far they move, and how well the move aligns with the change. As a control, we form random target changes from parameter perturbations, project them perpendicular to the 400k prediction, and match their magnitude to the history direction (Appendix F). The history direction moves predictions 63% as far as this control, and the move aligns less with the target change (cosine .45 against .89; Table 33). Equal-sized target changes therefore produce different responses in this local test: their direction affects both the magnitude and alignment of the response.

6 Related work

Guidance at sampling and training time. CFG contrasts conditional and unconditional predictions (Ho & Salimans, 2021); Autoguidance contrasts stronger and weaker models, including checkpoints of one run (Karras et al., 2024a). MG moves a conditional contrast into the training target (Tang

et al., 2025). IG contrasts network depths at sampling time and also studies training with a depth contrast (Zhou et al., 2026). PSG instead modifies the training target using prediction differences across optimization time, retaining the base prediction head and sampler. This differs from distilling a fixed guided sampler (Meng et al., 2023). Adaptive projected, angle-domain, and interval guidance modify how a contrast is applied during sampling (Sadat et al., 2025; Jin et al., 2025; Kynkäänniemi et al., 2024), and momentum guidance uses sampling-time velocity history (Liao et al., 2026), whereas PSG uses optimization-time history.

Regression objectives and sample quality. Flow matching, rectified flow, and stochastic interpolants learn transport fields through pointwise objectives (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2023). Selective Underfitting studies spatial variation in score approximation (Song et al., 2025), and reweighting methods change which noise levels dominate optimization (Choi et al., 2022; Hang et al., 2023); PSG modifies the regression target rather than reweighting the loss.

Temporal teachers and representation learning. EMA and mean teachers stabilize predictions (Tarvainen & Valpola, 2017; Grill et al., 2020); LAWA, EDM2, Lookahead, and LCSC combine or interpolate checkpoint weights (Kaddour, 2022; Karras et al., 2024b; Zhang et al., 2019; Liu et al., 2025), and Neon extrapolates away from a model fine-tuned on its own samples (Alemohammad et al., 2026). PSG instead uses differences of predictions at the same input. REPA supplies a pretrained representation target (Yu et al., 2025; Oquab et al., 2024), and Self-Flow and Self-Transcendence use internal teachers (Chefer et al., 2026; Sun et al., 2026).

7 Conclusion and Limitations

Angular corrections derived from a model's prediction difference with an earlier checkpoint improve samples but not regression loss. PSG adds the full prediction difference to the training target, so a single model carries it at inference. One setting improves SiT at all four scales, and at XL/2, PSG has lower FID than ordinary training that uses about twice as much compute. Applied only in the last 10% of a 2M-update XL/2 run, PSG improves FID using that run's own earlier checkpoints. In sampling interventions, reassigning corrections across samples removes the gain.

Limitations. At XL/2, each active update takes about 1.6 times as long as an ordinary update, and training beyond twice the activation step needs ordinary-training checkpoints after activation. Higher doses trade Inception recall for fidelity. Our experiments cover ImageNet 256×256; future work can test other domains and reduce the storage needed for past checkpoints.

AI use statement

We used generative AI tools to implement and run the experiments, including the data-processing, training, evaluation, and plotting code. We also used them to give feedback on experimental comparisons, to help interpret results, to search for and summarize related literature, and to draft and edit the paper. We have not used them to generate synthetic data sets. We have reviewed all AI-assisted work, tested the code, and cross-checked the reported numbers against the experiment records and cited sources. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

Reproducibility statement

Appendix A lists the model and training configuration (Table 8), the teacher and past-checkpoint schedules, the sampling and evaluation protocol, and the training cost. Table captions state the sample count and, for PSG models, the teacher schedule; they name the sampler and CFG setting where these differ from Table 8. Appendix A also describes how evaluation seeds and reference statistics are shared within each comparison. PSG changes only the regression target (Eq. 1), so it can be added to existing SiT and DiT training code with two teacher forward passes per update.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Sina Alemohammad, Zhangyang Wang, and Richard G. Baraniuk. Neon: Negative extrapolation from self-training improves image generation. In *International Conference on Learning Representations*, 2026.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. In *International Conference on Machine Learning*, 2026.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Cheng Jin, Zhenyu Xiao, Chutao Liu, and Yuantao Gu. Angle domain guidance: Latent diffusion requires rotation rather than extrapolation. In *International Conference on Machine Learning*, 2025.

Jean Kaddour. Stop wasting my time! saving days of ImageNet and BERT training with latest weight averaging. *arXiv preprint arXiv:2209.14981*, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, 2024a.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Runlong Liao, Jian Yu, Baiyu Su, Chi Zhang, Lizhang Chen, and Qiang Liu. Momentum guidance: Plug-and-play guidance for flow models. *arXiv preprint arXiv:2602.20360*, 2026.

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

Enshu Liu, Junyi Zhu, Zinan Lin, Xuefei Ning, Shuaiqi Wang, Matthew B. Blaschko, Sergey Yekhanin, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Linear combination of saved checkpoints makes consistency and diffusion models better. In *International Conference on Learning Representations*, 2025.

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

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

Maxime Oquab, Timothée Darcet, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *International Conference on Learning Representations*, 2025.

Kiwhan Song, Jaeyeon Kim, Sitan Chen, Yilun Du, Sham Kakade, and Vincent Sitzmann. Selective underfitting in diffusion models. *arXiv preprint arXiv:2510.01378*, 2025.

Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Ruibin Li, Yujing Sun, Shuaizheng Liu, and Lei Zhang. Self-transcendence: Is external feature guidance indispensable for accelerating diffusion transformer training? In *European Conference on Computer Vision*, 2026.

Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, 2019.

Xingyu Zhou, Qifan Li, Xiaobin Hu, Hai Chen, and Shuhang Gu. Guiding a diffusion transformer with the internal dynamics of itself. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

A Implementation and evaluation details

We specify the shared training and evaluation procedures before examining target design (Appendix B), additional comparisons (Appendix C), and directional, sampling, and optimization analyses (Appendices D, E, and F).

A.1 Training configuration

Table 8: **Shared protocol for controlled training and evaluation.** Settings apply throughout unless a table or experiment explicitly overrides them. The DiT experiments use no gradient clipping.

Component	Setting
Model and data	
Architecture	SiT-S/2, B/2, L/2, XL/2; 33M, 130M, 458M, 675M parameters
Data	ImageNet-1k 256×256; SD-VAE latents (ft-EMA encode, ft-MSE decode); horizontal flips
Training	
Objective	Linear interpolant; velocity prediction; uniform time sampling
Optimizer	AdamW; learning rate 0.0001; weight decay 0; gradient clip 1.0
Batch / EMA	Global batch 256; EMA decay 0.9999
Past-Self Guidance	
PSG (default)	$w = 0.5$; activation 200k; recent EMA teacher refreshed every 10k updates (each step where a caption notes it); past teacher: ordinary EMA at $r = K\lfloor .5s/K \rfloor$, $K = 10$k
Sampling and evaluation	
Sampling	Euler–Maruyama SDE; 250 evaluations; EMA weights
Field measurements	8,192 fixed training inputs; independent 8,192-input validation repeat (Appendix D.1)
Diagnostics	10k samples; shared sampling seeds for paired interventions
Main evaluation	50k samples, CFG-free; CFG sweeps reported separately
DINO robustness	DINOv2-L/14 features; shared 10k real reference; coverage $k = 5$

Long-horizon training. For the late-activation comparisons in Figure 2 and Table 17, the ordinary run continues to 2M and PSG branches from its 1M, 1.5M, or 1.8M checkpoint. PSG updates its recent teacher each step; other settings follow Table 8, and PSG's past checkpoints come from the same ordinary run.

A.2 Representation-aligned PSG

For REPA+PSG, the training objective is

$$\mathcal{L}_{\mathrm{RP}} = \frac{1}{D}\mathbb{E}\,\|u_\theta - [v + w\,\mathrm{sg}(\bar{u}^{\mathrm{RP}}_{q(s)} - \bar{u}^{\mathrm{R}}_{r(s)})]\|^2 + 0.5\,\mathcal{L}_{\mathrm{REPA}}.$$

The regression term averages over $D = 4 \cdot 32^2$ latent coordinates. Here R denotes the REPA-only trajectory and RP the combined trajectory; both teachers receive the same interpolated state, time, and class-dropout mask. The combined model starts from the REPA-only checkpoint at 200k. Its recent teacher follows RP and is refreshed every 10k updates; every past teacher used through 400k comes from the preactivation R trajectory.

The REPA loss averages negative cosine similarities over examples and patch tokens. It aligns projected block-8 features of the noisy student input with frozen DINOv2-B/14 patch features of the VAE reconstruction of the clean training latent, resized to 224×224. The three-layer SiLU projector has two 2,048-dimensional hidden layers and a 768-dimensional output. The auxiliary projector and encoder are training-only; the deployed predictor and sampler are unchanged. All R past checkpoints come from our REPA-only reproduction; its 400k XL/2 model reaches FID-10k 10.49. Published results in Table 3 are identified separately from these controlled comparisons.

A.3 Training update and checkpoint schedule

The student EMA is updated after every optimizer step. The recent teacher either uses this EMA at every update, $q(s) = s$, or a copy refreshed every 10k updates, $q(s) = 10\text{k}\lfloor s/10\text{k} \rfloor$. The 400k scale-transfer (Table 1), S/2 dose and reference-ratio, directional-component ablation, REPA, and composite comparisons use copies refreshed every 10k updates (Table 10 uses 100k). The B/2–XL/2 dose sweeps, the past-checkpoint-spacing sweep, the S/2 shuffled-target control (Appendix E.2), activation at 300k, 1M, 1.5M, or 1.8M, cosine decay, and DiT update the recent teacher each step. Each caption states its schedule. Both schedules use the same training procedure:

1. Sample an interpolated state z, its flow-matching velocity v, and the base class-dropout mask.
2. Before activation b (200k by default), train with the standard target v.
3. Once PSG is active, evaluate both teachers on the student's state, time, label, and dropout mask, using $q(s)$ above and $r(s) = K\lfloor \rho s/K \rfloor$, $K = 10\text{k}$, $\rho = .5$.
4. Optimize Eq. (1) with the stop-gradient secant, then update the student EMA. Update the past teacher only when $r(s)$ changes.
5. At inference, use only the final student EMA and the unchanged SiT solver.

The past checkpoint is selected on a 10k grid. With $\rho = .5$, the past teacher advances by 10k every 20k student updates, half as often as the recent teacher when it is refreshed every 10k updates. For activation step b and final step T, all past teachers lie at or before activation whenever $\rho T \leq b$. The 200k–400k comparison satisfies this condition. Continuing until $r(s) > 200\text{k}$ requires the ordinary run to continue beyond 200k. The past teacher uses $\rho(T - b)/K$ saved ordinary-training EMA checkpoints (10 for the 400k runs and 50 for activation at 1M) and no optimizer state. The rule $b = T/2$ follows from this checkpoint availability rather than from tuning; activation at 300k recovers most of the B/2 gain at 400k (Appendix B).

In Table 11, varying K changes only the past-checkpoint spacing; the recent teacher is updated each step. Because $r(s)$ rounds down, the spacing also changes the past checkpoint's average age. Over 200k–400k, the past checkpoint advances from step 100k to 190k at $K = 10\text{k}$ but stays at step 100k at $K = 100\text{k}$.

A.4 Sampling and evaluation

The conditional models retain the base SiT null-class probability of 0.1; PSG never uses the conditional–unconditional contrast. Principal metrics use the official ImageNet-256 reference statistics and 50,000 generated samples. The paired 10,000-sample comparisons share generation seeds. Standard evaluations draw class labels uniformly at random; the full-path evaluations of Table 7 use 10 samples per class. CFG mixes conditional and unconditional predictions in all four latent channels. Different tables can use different seed sets: the ordinary S/2 400k model reaches FID-10k 62.48 in Tables 4, 13, and 15 and 61.80 in Tables 14 and 25, so smaller differences across seed sets are not read as rankings. Comparisons use the same sample count and reference statistics; 10k and 50k estimates are reported separately. Appendix E.1 specifies the integration grids and model-call counts.

DINO evaluation. Table 24 uses the final layer-normalized CLS vectors of DINOv2-L/14 (1,024 dimensions), without additional unit-norm scaling. We resize RGB images directly to 224×224 using antialiased bicubic interpolation, without cropping. After scaling pixels to $[0, 1]$, we normalize channels by means $(.485, .456, .406)$ and standard deviations $(.229, .224, .225)$. The shared real reference contains all 10,000 images in the ADM VIRTUAL_imagenet256_labeled.npz archive. Fréchet distance uses the empirical feature means and covariances. Coverage is the fraction of real feature vectors whose nearest generated neighbor is no farther away than their fifth-nearest other real neighbor, using Euclidean distances.

A.5 Training cost

Active PSG performs three forward passes: student, recent teacher, and past teacher. Only the student requires a backward pass. Using the current EMA as the recent teacher keeps three parameter

sets resident. Periodic teacher copies require a fourth set because the student EMA continues to update between copies. Let κ denote the cost of an active PSG update relative to an ordinary update. For XL/2 on eight A100 GPUs, ordinary training completes about 9.3k updates per hour and PSG with the recent teacher updated each step about 5.7k, so $\kappa \approx 1.6$. Refreshing the recent teacher every 10k updates instead performs the same forward passes. Ignoring checkpoint I/O and grid rounding, the relative training cost is

$$\mathcal{C}(T, b) = \frac{b + \kappa(T - b) + \max(0, \rho T - b)}{T}.$$

The denominator is the cost of an ordinary run of length T, and the last term accounts for ordinary training beyond the shared prefix. For $b = T/2$ and $\rho = .5$, this reduces to $(1 + \kappa)/2 \approx 1.3$; for activation at $b = .9T$, such as 1.8M of 2M updates, it is $.9 + .1\kappa \approx 1.06$.

B Training-target design and ablations

These experiments vary the strength, age, and source of the temporal target. We separate changes to the student-dependent PSG target from alternatives whose teachers remain on an ordinary-training trajectory.

B.1 Dose, past-checkpoint spacing, and teacher averaging

The S/2 sweeps vary the PSG weight w and the reference ratio ρ, which controls the age of the past checkpoint. At S/2, FID is lowest at $w = .75$ and at $\rho = .25$ (Table 12); the cross-scale dose and CFG comparison is in Table 5. The denser B/2 sweeps over dose and past-checkpoint spacing are in Table 11. With 50k samples and the recent teacher updated each step, raising w from .5 to .6 lowers B/2 CFG-free FID from 18.73 to 13.84 and recall from .615 to .594 (Table 9).

The effect of EMA smoothing differs between the two teacher roles. EMA smoothing lowers FID for the recent teacher, whose predictions depend on the student's training history, whereas raw past-teacher weights give similar or lower FID (by .46 and 1.75; Table 10). Here both teachers use checkpoints saved at 100k-update intervals, so the comparison examines EMA versus raw weights under a common checkpoint schedule.

Table 9: **Higher dose lowers FID and recall at 50k samples.** B/2, 400k, SDE-250; same checkpoints as Table 11(a), $w \in \{.5, .6\}$, with the recent EMA teacher updated each step. CFG 1.5 is fixed, not a 50k sweep minimum. Bold marks the better value in each column.

	CFG-free		CFG 1.5	
Dose w	**FID ↓**	**Rec. ↑**	**FID ↓**	**Rec. ↑**
.50	18.73	**.615**	5.95	**.469**
.60	**13.84**	.594	**5.58**	.438

Table 10: **EMA smoothing lowers FID for the recent teacher.** S/2 FID-10k; both teachers use checkpoints saved at 100k-update intervals in all four combinations. Bold marks the lower FID in each past-teacher column.

	Past teacher	
Recent teacher	**EMA**	**Non-EMA**
EMA	**44.98**	**44.52**
Non-EMA	48.19	46.44

Table 11: **Dose and past-checkpoint spacing affect guided and unguided FID differently.** SiT-B/2 at 400k, activation at 200k, recent EMA teacher updated each step, $\rho = .5$, SDE-250, and 10k samples. FID and recall are CFG-free; FID+CFG is the lowest measured FID over scales $\{1.1, 1.25, 1.5, 2.0\}$. The guided minima are at 1.25 for $w \in \{.70, .75, .80\}$ and 1.5 otherwise. Panels sweep w and K with all other settings fixed. Bold marks the best value within each panel.

(a) Training dose, $K = 10$k

w	FID ↓	Rec. ↑	FID+CFG ↓
.50	21.37	**.674**	8.75
.60	16.46	.659	**8.38**
.70	12.78	.639	8.94
.75	**11.95**	.617	9.29
.80	12.43	.587	10.44
.90	18.99	.528	16.35

(b) Past-checkpoint spacing, $w = .5$

K	FID ↓	Rec. ↑	FID+CFG ↓
10k	21.37	**.674**	8.75
20k	20.99	.665	8.79
50k	19.52	.667	**8.63**
100k	**16.69**	.655	8.74

Table 12: **FID is lowest at intermediate doses and reference ratios.** S/2 at 400k with 10k samples, recent EMA teacher refreshed every 10k updates, activation at 200k, and 10k past-checkpoint spacing. Defaults are $w = .5$ and $\rho = .5$; one quantity changes at a time. FID remains low near $w = .75$–$.80$. Moving the reference ratio from .25 to .10 sharply degrades every metric. Bold marks the best value within each panel.

(a) PSG weight

w	FID ↓	sFID ↓	Prec. ↑	Rec. ↑
0.25	54.43	22.07	0.439	0.591
0.50	42.01	19.84	0.501	**0.636**
0.75	**28.59**	**18.56**	0.598	0.605
0.80	28.97	20.67	**0.599**	0.592
0.85	31.55	25.62	0.573	0.556
0.90	37.95	34.63	0.509	0.496

(b) Reference ratio

ρ	FID ↓	sFID ↓	Prec. ↑	Rec. ↑
0.75	57.85	23.07	0.428	0.576
0.50	42.01	19.84	0.501	**0.636**
0.25	**34.33**	**18.02**	**0.542**	0.635
0.10	70.10	60.82	0.282	0.350

B.2 Activation and learning-rate schedule

Activating PSG at 300k also improves B/2, although by then a small forward step along the model's own 50k-lag angular history direction already raises its FM loss. This sign change occurs at 248k (Appendix D.4). FID-10k reaches 23.68 after 100k PSG updates, at 400k, compared with 37.99 for ordinary training (Table 16). It reaches 20.29 after 200k PSG updates, compared with 21.37 for activation at 200k followed by 200k PSG updates. The 20.29 and 21.37 models have the same PSG duration but different total training lengths, 500k and 400k. The benefit thus does not require activation at the default 200k point.

The S/2 improvement also persists under cosine learning-rate decay (Table 13). Under this schedule, the FM loss still favors moving backward along the ordinary run's 400k–200k direction. Measured between raw (non-EMA) checkpoints, its angular loss-minimizing coefficient is $-.0183$ ($-.0233$ for the unprojected direction), much smaller in magnitude than the constant-learning-rate value of $-.24$ (Appendix D).

Table 13: **PSG improves FID under both constant and cosine-decayed learning rates.** S/2 at 400k, with 200k activation and 10k samples. The recent EMA teacher is refreshed every 10k updates for constant LR and updated each step for cosine decay. The cosine configuration retains 79% of the constant-LR FID gain; because the teacher schedules also differ, this ratio does not isolate the learning-rate effect. Along the cosine ordinary run's 400k–200k direction between raw (non-EMA) checkpoints, the FM-loss-minimizing angular coefficient is $-.0183$, versus $-.0024$ for matched random angular directions; the optimum for the unprojected direction is $-.0233$. These measurements do not control for endpoint fluctuations.

	FID-10k ↓			
S/2 learning rate	**Ordinary**	**PSG**	**FID gain**	**Gain retained**
Constant 10^{-4}	62.48	42.01	20.47	100%
Cosine $10^{-4} \to 10^{-5}$	70.18	54.04	16.14	79%

B.3 Teachers independent of the student

The recent EMA in Eq. (1) makes the target depend on the student's training history. Two variants remove that dependence in 100k-update S/2 continuations from the same ordinary 400k checkpoint, using the 10k diagnostic protocol. *Fixed teachers* freeze both endpoints at the ordinary 400k/200k checkpoints and train toward $v + \lambda\,\mathrm{sg}[d]$ with an explicit dose λ and angular direction $d = d_{400\mathrm{k},200\mathrm{k}}$ (Appendix D.1). This has the form of the Appendix F objective, here at S/2 with $\lambda \in \{1, 2.5\}$ for 100k updates. *Refreshed ordinary teachers* use the same angular target and explicit dose λ, but both endpoints advance along the ordinary trajectory. The recent teacher is refreshed every 10k updates, and the past teacher advances by 10k every 20k updates. The direction therefore changes with ordinary training but not with the student (Table 14). The ordinary continuation shares this starting checkpoint. The final row instead continues the PSG model already trained from 200k to 400k, retaining its full-difference target.

Table 14: **Training with fixed or refreshed ordinary teachers improves generation.** S/2 FID-10k; calls are per field evaluation. Panel (a) reports 100k-update continuations; fixed and refreshed teachers both come from ordinary training and use the angular target. Rows above the divider start from ordinary 400k; the final row continues PSG 400k, activated at 200k, with its full-difference target; from 400k its recent teacher is updated each step. Panel (b) samples the original ordinary 400k field. Bold marks the best value within each panel.

(a) Training: 400k → 500k

S/2 continuation	**Calls**	**FID ↓**	**Rec. ↑**
Ordinary FM	1	59.31	.572
Fixed teachers, $\lambda = 1$	1	45.68	.632
Fixed teachers, $\lambda = 2.5$	1	33.65	.654
Refreshed teachers, $\lambda = 1$	1	45.48	.633
Refreshed teachers, $\lambda = 2.5$	1	**33.55**	**.657**
PSG continuation, full $w = .5$	1	38.12	.642

(b) Frozen inference: 400k

Angular	**Calls**	**FID ↓**	**Rec. ↑**
$\alpha = 0$ (ordinary)	1	61.80	.566
$\alpha = 1$	2	50.50	.615
$\alpha = 1.5$	2	45.75	.626
$\alpha = 2.5$	2	**40.69**	**.637**

Both student-independent variants improve on the ordinary continuation, from FID 59.31 to 45.48–45.68 at $\lambda = 1$ and 33.55–33.65 at $\lambda = 2.5$, with higher recall. At matched explicit doses, fixed and refreshed ordinary teachers differ by at most 0.2 FID: 45.68 versus 45.48 at dose 1, and 33.65 versus 33.55 at dose 2.5. The full-difference PSG continuation reaches 38.12. Because it starts from the PSG 400k model and keeps the full-difference target, its comparison with the fixed- and refreshed-teacher rows does not isolate feedback or estimate the field coefficient in Eq. (2). At $\lambda = 2.5$, fixed-teacher training reaches FID 33.65 and recall .654, compared with 40.69 and .637 for the angular sampling-time composite at $\alpha = 2.5$. The trained model receives 100k additional updates, over which ordinary training improves FID by 2.5 (61.80 to 59.31), less than the 7.04 difference.

Larger models. At B/2, L/2, and XL/2, the fixed-teacher models start from the ordinary 200k checkpoint, with fixed teachers at the ordinary 200k and 100k checkpoints. They train to 400k toward $v+\lambda\,\mathrm{sg}[d]$, with $d = d_{200\mathrm{k},100\mathrm{k}}$ and $\lambda = 2.5$ (also 2.0 and 3.0 at B/2). The B/2 model reaches FID-10k 19.82 at $\lambda = 2.5$ and 19.29 at $\lambda = 3.0$, compared with 21.37 for $w = .5$. The fixed-teacher recall values are .644 and .631, compared with .674 for PSG. The B/2 and XL/2 comparisons use PSG with the recent teacher updated each step; the L/2 FID-50k comparison refreshes it every 10k updates. The L/2 model reaches FID-50k 8.21, compared with 9.32 for $w = .5$. Its recall is .554, compared with .610 for PSG.

The XL/2 model reaches FID-10k 9.62 at recall .640, compared with 10.14 at .691 for PSG (Table 16). Across B/2, L/2, and XL/2, fixed-teacher angular targets with $\lambda \geq 2.5$ give lower CFG-free FID than PSG at $w = .5$, and PSG at $w = .5$ retains higher recall at every tested λ. At B/2 and XL/2, PSG at $w = .6$ (FID-10k 16.46 and 7.77, recall .659 and .668; Table 5) has both lower CFG-free FID and higher recall than every fixed-teacher setting. At B/2 it also has a lower guided FID (8.38). These configurations differ in projection, dose, and teacher dependence, so no run isolates the student-dependent teacher at matched dose. PSG needs only checkpoints saved before activation when $\rho T \leq b$, whereas refreshed ordinary teachers need an ordinary run continued to the current step. Appendix C also evaluates refreshed ordinary teachers with an unprojected target over a longer XL/2 training interval.

Interaction with CFG. The preferred target configuration also depends on inference-time guidance. At B/2, 400k, raising λ from 2.0 to 3.0 moves precision from .647 to .672 and recall from .650 to .631. At $\lambda = 2.0$, adding CFG at scale 2.0 moves precision from .647 to .888 and recall from .650 to .467. Raising λ and adding CFG both raise precision and lower recall.

At $\lambda = 2.0$, the fixed-teacher CFG sweep has an interior FID-10k minimum of 10.11 at scale 1.5. Its values at scales 1.1, 1.25, 1.5 and 2.0 are 16.90, 12.89, 10.11 and 10.55, while the $w = .5$ curve reaches 8.75 at the same minimizing scale. The dose ordering reverses with CFG: a larger dose gives lower CFG-free FID (21.45, 19.82, 19.29 for λ of 2.0, 2.5, 3.0), whereas a smaller dose gives lower guided FID (10.11, 10.29, 10.77) at CFG 1.5. On the same grid, full-difference PSG has a best guided FID 1.36 below the fixed-teacher minimum, but its CFG-free FID is 1.6 to 2.1 higher than the $\lambda = 2.5$ and $\lambda = 3.0$ models. Its no-CFG recall is .024 higher than the $\lambda = 2.0$ fixed-teacher point. At B/2, the default PSG target thus gives better guided FID and retains more recall without CFG.

B.4 Conditional rather than temporal contrast

A conditional–unconditional target replaces the temporal contrast with a different source of guidance (Table 15). From 200k onward, with the student's current EMA $\bar{u}_s$, the target is

$$\tilde{v}_{\mathrm{cond}} = v + 0.5\,\mathrm{sg}[\bar{u}_s(x_t, t, c') - \bar{u}_s(x_t, t, \emptyset)],$$

where c' is the label after the base 0.1 class-dropout mask. The contrast is zero on dropped examples. At 400k, the S/2 and B/2 models trained with this target have lower FID-10k and recall than the displayed PSG settings.

Table 15: **Models trained with conditional contrast have lower FID and recall than PSG.** 400k, 10k samples; each row is one model trained with the listed target. PSG updates its recent teacher each step at B/2 and refreshes it every 10k updates at S/2. Bold marks the best value within each scale.

Scale	Training target	FID-10k ↓	Recall ↑
S/2	Flow matching	62.48	.558
	PSG, $w = .5$	42.01	**.636**
	Conditional contrast	**18.23**	.536
B/2	Flow matching	37.99	.653
	PSG, $w = .5$	21.37	**.674**
	PSG, $w = .75$	11.95	.617
	Conditional contrast	**10.12**	.488

C Additional performance comparisons

The following comparisons supplement the main results with longer training, complete guidance sweeps, and alternative evaluation settings.

C.1 Training duration and later activation

PSG at 400k has lower FID than longer ordinary training under the 10k diagnostic protocol (Table 16). The table also distinguishes angular-target training with fixed teachers from training with refreshed ordinary teachers and an unprojected target. The refreshed-teacher run starts at 400k with explicit dose $\lambda = 2.5$. Its recent teacher is the ordinary EMA refreshed every 10k updates, and the past teacher uses $r(s) = K\lfloor .5s/K \rfloor$, $K = 10\text{k}$. FID-10k falls from 7.33 at 500k to 6.80 at 700k and 6.58 at 1M, while recall rises from .663 to .701; at 1M, ordinary training reaches FID 14.69 with recall .714. These ordinary teachers therefore provide useful temporal supervision without depending on the student. They need the ordinary run to reach the current step, so training to 1M costs about $1 + .6\kappa \approx 2.0$ ordinary 1M runs, compared with $\mathcal{C}(1\text{M}, 500\text{k}) \approx 1.3$ for PSG activated halfway (Appendix A).

Table 16: **PSG at 400k has lower FID than longer ordinary training.** SDE-250, no CFG, 10k samples. PSG updates its recent EMA teacher each step; REPA+PSG refreshes its recent EMA teacher every 10k updates. Both fixed- and refreshed-teacher rows use explicit dose $\lambda = 2.5$ rather than PSG's $w = .5$: fixed teachers use the angular 200k–100k direction from 200k, and refreshed ordinary teachers use their full prediction difference from 400k. Activation at 1M, 1.5M, and 1.8M is reported in Table 17. Dashes denote checkpoints not included in this comparison.

XL/2: FID-10k ↓ **Training setting**	**400k**	**1M**	**2M**
Ordinary	20.53	14.69	12.89
PSG ($b = 200\text{k}$)	10.14	–	–
Fixed teachers, angular ($b = 200\text{k}$)	9.62	–	–
Refreshed ordinary teachers ($b = 400\text{k}$)	–	6.58	–
REPA	10.49	9.52	9.61
REPA + PSG ($b = 200\text{k}$)	6.08	–	–
B/2: FID-10k ↓			
Ordinary	37.99	29.75	–
PSG ($b = 200\text{k}$)	21.37	–	–

Activating PSG at 1M improves both the 1.5M and 2M endpoints (Table 17). At 2M, the FID-50k reduction is 47.0% relative to ordinary training at the same update count. Here the recent teacher is updated each step (Appendix A.1). For this 2M model, a 10k-sample sweep gives FID 6.02, 5.02, and 6.40 at CFG 1.1, 1.25, and 1.5; with 50k samples, CFG 1.25 reaches 1.99 (Table 3), against

2.08 at 1.20 and 2.06 at 1.30. CFG lowers FID-50k from 5.35 to 1.99, while recall falls from .660 to .605. Activation at 1.5M or 1.8M reaches FID-50k 5.76 or 5.73 at 2M without CFG, against 10.10 for ordinary training. With CFG 1.25, the best scale in each model's 10k sweep, these models reach 2.02 or 2.11.

Table 17: **PSG activated as late as 1.8M still improves FID at 2M updates.** PSG branches from one ordinary run at the listed activation b (1M, 1.5M, or 1.8M). PSG updates its recent EMA teacher each step, with $w = .5$, $\rho = .5$, and $K = 10$k; past checkpoints remain at or before the listed activation b. Both panels use SDE-250 without CFG. Comparisons match updates, not compute. For $b = 1$M, the FID-50k advantage is 5.04 at 1.5M and 4.75 at 2M.

(a) 10k samples XL/2	Updates	FID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
Ordinary	1M	14.69	18.76	100.2	.680	.714
Ordinary	1.5M	13.79	18.85	108.0	.679	.718
PSG ($b = 1$M)	1.5M	8.60	18.63	148.5	.722	.722
Ordinary	2M	12.89	18.92	114.3	.687	.729
PSG ($b = 1$M)	2M	8.42	18.78	153.2	.719	.736
PSG ($b = 1.5$M)	2M	8.59	18.58	150.5	.716	.732
PSG ($b = 1.8$M)	2M	8.55	18.48	149.7	.721	.729
(b) 50k samples						
Ordinary	1M	12.08	5.42	99.9	.673	.647
Ordinary	1.5M	10.86	5.51	108.3	.679	.646
PSG ($b = 1$M)	1.5M	5.82	5.24	145.7	.717	.660
Ordinary	2M	10.10	5.60	113.8	.681	.663
PSG ($b = 1$M)	2M	5.35	5.38	152.8	.715	.660
PSG ($b = 1.5$M)	2M	5.76	5.28	147.9	.714	.660
PSG ($b = 1.8$M)	2M	5.73	5.12	145.9	.715	.660

C.2 Training-time and sampling-time guidance

The full and angular sampling-time composites apply history at inference rather than during training (Table 18). At $\alpha = 2$, the angular and full composites reach FID 7.45 and 7.43, with recall .617 and .611, respectively. The angular sweep further reduces FID to 6.37 at $\alpha = 3$. Both require two model calls per field evaluation; PSG incorporates the temporal contrast during training and uses one predictor.

Table 18: **PSG halves inference calls relative to the sampling-time composite, with a dose-dependent fidelity–recall trade-off.** XL/2 with 50k samples; PSG refreshes its recent EMA teacher every 10k updates. Default PSG and the $\alpha = 2$ angular composite differ by .25 FID at recall .614 versus .617. The full composite uses the unprojected temporal difference; angular rows remove its component parallel to the current prediction. At similar recall, the best measured composite ($\alpha = 3$, 6.37) has lower FID than default PSG but uses two calls; higher-dose PSG reaches 3.94 at lower recall. Training weight w and inference dose α are different interventions. Dashes denote unreported metrics.

XL/2 sampler at 400k	Calls/field	FID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
Full composite, $\alpha = 2$	2	7.43	–	–	.720	.611
Angular composite, $\alpha = 1$	2	10.61	4.59	104.8	.693	**.625**
Angular composite, $\alpha = 2$	2	7.45	–	–	.716	.617
Angular composite, $\alpha = 3$	2	6.37	5.01	135.8	.723	.616
PSG, $w = .5$	1	7.70	**4.56**	122.7	.729	.614
PSG, $w = .75$	1	**3.94**	6.30	**180.1**	**.810**	.535

At XL/2 (400k, 50k samples), increasing the PSG dose from $w = .5$ to .75 without CFG lowers FID and recall. Adding representation alignment raises recall at either dose. For PSG alone, CFG 1.25

and 1.5 lower FID at $w = .5$, whereas CFG 1.25, the only scale tested at $w = .75$ with 50k samples, raises it. Table 19 gives the complete XL/2 metrics behind the principal comparisons, including PSG activated at 1M, 1.5M, or 1.8M and evaluated at 2M, and REPA trained to 1M.

Table 19: **Training dose and CFG both trade recall for fidelity, at different call costs.** All rows use 50k samples and SDE-250. Panels (a) and (b) evaluate the 400k models with one and two model calls per field evaluation; in these panels, PSG and REPA+PSG refresh their recent EMA teachers every 10k updates. Panel (c) reports later checkpoints; its PSG models update the recent teacher each step. The 400k minima of PSG ($w = .5$), REPA, and REPA+PSG lie at the lowest displayed scale; among the displayed scales, the CFG curves of PSG ($b = 1$M) and REPA (1M) are lowest at 1.25, the only scale shown for $b = 1.5$M and $b = 1.8$M. Dashes denote unreported metrics; bold marks the best values within each 400k panel.

XL/2 training	Sampling	FID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
(a) 400k, one model call per field evaluation						
PSG, $w = .5$	no CFG	7.70	**4.56**	122.7	0.729	0.614
PSG, $w = .75$	no CFG	3.94	6.30	180.1	**0.810**	0.535
REPA + PSG, $w = .5$	no CFG	3.28	4.59	177.0	0.760	**0.638**
REPA + PSG, $w = .75$	no CFG	**2.34**	6.40	**234.8**	0.808	0.586
(b) 400k with CFG, two model calls per field evaluation						
SiT	CFG 1.25	7.86	4.84	132.3	0.757	0.558
SiT	CFG 1.5	5.15	5.46	187.3	0.828	0.493
SiT	CFG 2.0	6.67	7.77	268.7	**0.896**	0.375
PSG, $w = .5$	CFG 1.25	3.69	4.54	196.8	0.818	0.538
PSG, $w = .5$	CFG 1.5	4.24	5.34	255.2	0.868	0.474
PSG, $w = .75$	CFG 1.25	4.67	6.12	253.2	0.869	0.471
REPA	CFG 1.25	2.84	**4.52**	212.2	0.800	0.589
REPA	CFG 1.5	3.07	4.86	**285.8**	0.856	0.530
REPA + PSG, $w = .5$	CFG 1.1	**2.29**	4.53	215.5	0.795	**0.612**
REPA + PSG, $w = .5$	CFG 1.25	2.35	4.74	265.8	0.831	0.580
(c) Later checkpoints						
PSG, $b = 1$M (2M)	no CFG	5.35	5.38	152.8	0.715	0.660
PSG, $b = 1$M (2M)	CFG 1.2	2.08	4.53	229.3	0.794	0.627
PSG, $b = 1$M (2M)	CFG 1.25	1.99	4.45	246.4	0.808	0.605
PSG, $b = 1$M (2M)	CFG 1.3	2.06	4.41	262.6	0.820	0.597
PSG, $b = 1.5$M (2M)	no CFG	5.76	5.28	147.9	0.714	0.660
PSG, $b = 1.5$M (2M)	CFG 1.25	2.02	4.38	243.3	0.805	0.606
PSG, $b = 1.8$M (2M)	no CFG	5.73	5.12	145.9	0.715	0.660
PSG, $b = 1.8$M (2M)	CFG 1.25	2.11	4.38	239.0	0.810	0.604
REPA (1M)	no CFG	6.85	5.38	–	0.700	0.669
REPA (1M)	CFG 1.1	4.03	4.93	–	0.741	0.639
REPA (1M)	CFG 1.25	2.26	4.55	233.1	0.792	0.611

C.3 System-level comparison

Table 3 compares published results from different training setups. The SiT entry uses the revised paper's SDE result (8.26 without CFG at 7M updates); the corresponding older value 8.30 is quoted by several later papers. REPA's 400k and 4M no-CFG metrics come from Table 8 of the REPA paper; the guided results come from its Table 4 and use CFG with a guidance interval. IG (internal guidance) contrasts two output heads at different depths of a shared backbone, so its "no CFG" endpoint still uses guidance; IG can restrict this contrast to high- and middle-noise levels. Its guided endpoint adds CFG with a guidance interval. Its Stable Diffusion VAE is ft-EMA, whereas our evaluation decodes with ft-MSE. MG incorporates a conditional contrast during training and samples without CFG; the training budget of its headline 1.34 endpoint is not specified in the reported table. Self-Flow's displayed endpoint uses the SD-VAE setting; its RAE results use a different representation and sampler. The table therefore compares entire systems, not isolated target changes.

C.4 Representation alignment and alternative generation settings

PSG improves unconditional generation without a conditional prediction pair (Table 20). Adding REPA lowers PSG's FID-50k by 4.42–4.48 at S/2, B/2, and XL/2 (Table 21).

Table 20: **PSG improves unconditional generation.** B/2, 400k, 50k samples, SDE-250; PSG uses $w = .5$ and refreshes its recent EMA teacher every 10k updates, with no conditional–unconditional contrast. Recall decreases by .007.

Metric	Ordinary	PSG
FID ↓	66.14	**51.55**
sFID ↓	7.71	**6.27**
IS ↑	16.2	**19.2**
Prec. ↑	0.393	**0.491**
Rec. ↑	**0.579**	0.572
ΔFID ↓	0.00	**−14.59**

Table 21: **Representation alignment improves PSG across scales.** Class-conditional ImageNet 256×256, 400k updates, SDE-250, 50k samples, no CFG; recent EMA teacher refreshed every 10k updates, $w = \rho = .5$, 200k activation. ΔFID is REPA+PSG relative to PSG alone; recall is that of REPA+PSG.

	FID ↓			
Scale	PSG	+REPA	ΔFID	Rec. ↑
S/2	39.28	**34.80**	−4.48	0.632
B/2	19.43	**14.98**	−4.45	0.638
XL/2	7.70	**3.28**	−4.42	0.638

The DiT noise-prediction experiment retains the learned-variance objective and DDPM sampler while modifying the regression target. PSG lowers DiT-B/2 FID-10k from 45.41 to 25.03 and raises recall from .609 to .656 (Table 22).

Table 22: **PSG improves diffusion and unconditional generation.** ImageNet 256×256, B/2 at 400k updates, $w = .5$, activation at 200k, and no CFG. DiT uses DDPM-250 and updates its recent EMA teacher each step; unconditional SiT uses SDE-250 and refreshes its recent EMA teacher every 10k updates. Sample counts are shown per row; bold compares ordinary training and PSG within each setting.

		FID ↓		Recall ↑	
Setting (B/2)	Samples	Ordinary	+ PSG	Ordinary	+ PSG
Diffusion (ϵ-prediction)	10k	45.41	**25.03**	.609	**.656**
Unconditional flow matching	50k	66.14	**51.55**	**.579**	.572

C.5 Sampling budget and feature-space evaluation

PSG improves FID with as few as 25 field evaluations under both stochastic and deterministic integration (Table 23). The benefit persists across the tested solvers without an additional predictor at inference.

To test whether the Inception-FID ranking of the XL/2 operating points is specific to Inception features, we reuse the same 50k generated samples and compute a Fréchet distance in frozen DINOv2-L/14 features, denoted $\mathrm{FD}_{\mathrm{DINO}}$. We also report 5-nearest-neighbor coverage in that feature space; Appendix A.4 gives the reference set and preprocessing. Every PSG and REPA+PSG operating point has lower $\mathrm{FD}_{\mathrm{DINO}}$ and higher coverage than SiT, and only the top two points, both with REPA, swap order. Because REPA aligns hidden features with DINOv2-B/14 during training (Appendix A), this feature space is independent of the training objective only for the rows without REPA. DINO coverage, unlike Inception recall, rises with training dose: from .867 to .956 for PSG and from .979 to .993 for REPA+PSG. Elsewhere in the paper, coverage in the fidelity–coverage trade-off refers to Inception recall.

Table 23: **PSG improves both stochastic and deterministic solvers across evaluation budgets.** The XL/2 400k pair of Table 1, whose PSG model refreshes its recent teacher every 10k updates, is evaluated across all budgets. SDE columns use the displayed evaluation counts; Heun uses 13, 26, and 51 grid points, corresponding to 25, 51, and 101 field evaluations. Columns group the nearest evaluation budgets. Bold marks the better value within each solver.

Training	Solver	FID-10k ↓ by evaluation budget		
		25	**≈50**	**≈100**
Ordinary	SDE	31.27	23.15	21.13
PSG	SDE	**19.04**	**12.48**	**11.11**
Ordinary	Heun	22.50	21.24	21.02
PSG	Heun	**12.75**	**11.72**	**11.41**

Table 24: **PSG and REPA+PSG gains persist in DINOv2 features.** XL/2 at 400k, with 50k generated samples and no CFG unless indicated; PSG and REPA+PSG refresh their recent EMA teachers every 10k updates. The CFG setting gives the lowest displayed Inception FID but not the lowest $\mathrm{FD}_{\mathrm{DINO}}$. Only the top two points swap order between the two feature spaces. DINO statistics use one shared 10k real-image reference and are not compared with other protocols.

XL/2 operating point	Inception FID ↓	$\mathrm{FD}_{\mathrm{DINO}}$ ↓	DINO coverage ↑
SiT	16.89	482.9	.635
PSG, $w = .5$	7.70	298.9	.867
PSG, $w = .75$	3.94	168.5	.956
REPA + PSG, $w = .5$	3.28	162.9	.979
REPA + PSG, $w = .75$	2.34	**101.0**	**.993**
REPA + PSG, $w = .5$, CFG 1.1	**2.29**	137.9	.987

D FROZEN TEMPORAL DIRECTIONS AND REGRESSION LOSS

A frozen history direction raises regression loss yet lowers FID. This appendix gives the loss algebra, dose and scale transfer, lag dependence, and endpoint controls.

D.1 ANGULAR PROJECTION AND LOSS ALONG A DIRECTION

For frozen EMA fields u_s and u_r at updates $s > r$, the temporal secant is $h_{s,r}(z) = u_s(z) - u_r(z)$. Its angular component removes velocity rescaling:

$$d_{s,r} = P_{u_s}^{\perp,\eta} h_{s,r} = h_{s,r} - \frac{\langle h_{s,r}, u_s \rangle}{\|u_s\|^2 + \eta} u_s, \qquad \eta = 10^{-12}. \tag{3}$$

All fields receive the same z. Norms and inner products are per sample over channel and spatial axes. Reported losses and FM drives (negative directional derivatives of FM loss, defined below) are divided by $D = 4 \cdot 32^2$, the number of latent coordinates. Field measurements use 8,192 fixed training inputs, with times stratified across five bins. Each input fixes the latent, class label, noise draw, and interpolation time. Table 30 repeats the loss-optimum measurements and the crossing measurements of Appendix D.4 on 8,192 independent validation inputs. We write $\mathbb{E}_{\mathcal{P}}$ for the average over the input set used; the loss calculations below use this empirical average in $\mathcal{L}_{\mathrm{FM}}$. For $u = u_{400\mathrm{k}}$ and $d = d_{400\mathrm{k},200\mathrm{k}}$, the angular sampling-time composite is $u_\alpha = u + \alpha d$, where α is the frozen-field dose. Positive values extrapolate along history; negative values move backward. The loss change is exactly

$$\mathcal{L}_{\mathrm{FM}}(u + \alpha d) - \mathcal{L}_{\mathrm{FM}}(u) = 2\alpha\, \mathbb{E}_{\mathcal{P}} \langle u - v, d \rangle + \alpha^2\, \mathbb{E}_{\mathcal{P}} \|d\|^2. \tag{4}$$

Equal-norm directions have equal quadratic cost. Their alignment with the prediction residual determines the loss-minimizing coefficient, provided $\mathbb{E}_{\mathcal{P}} \|d\|^2 > 0$:

$$\alpha^*_{\mathrm{FM}} = -\frac{\mathbb{E}_{\mathcal{P}} \langle u - v, d \rangle}{\mathbb{E}_{\mathcal{P}} \|d\|^2}. \tag{5}$$

Across the four model scales, the positive linear term at $\alpha = 1.5$ adds to the quadratic cost shared by equal-norm directions (1.9–2.4% for norm-matched random directions), producing a 2.1–2.8% relative loss increase (Table 6). The Gaussian controls in this appendix use seed 1234, with a fresh direction at each evaluation, angularized against the recipient output and norm matched per sample to history. Table 25 gives higher-dose 10k-sample results; the 50k comparison is in Table 18.

Table 25: **The angular sampling-time composite lowers FID at all four model scales.** FID-10k, 400k, SDE-250, no CFG; each model's angular 400k–200k history direction is added at dose α, and $\alpha = 0$ is ordinary sampling. For S/B/L, the lowest tested FID is at the largest tested dose. XL/2 FID rises to 10.23 at $\alpha = 3.5$ and 11.05 at $\alpha = 4$.

Scale	$\alpha = 0$	2	2.5	3
S/2	61.80	42.93	40.69	**39.77**
B/2	37.99	22.41	20.70	**19.86**
L/2	22.89	12.00	11.54	**11.44**
XL/2	20.53	10.54	9.84	**9.66**

D.2 Transfer across model scales

Histories from every tested model size improve sampling FID when added to the same XL/2 field after angular projection and norm matching (Table 26). Each source supplies predictions at the recipient's current sampling state. Transfer occurs although the per-input cosines between these histories and XL/2's own are only .020–.175. A useful correction direction therefore need not originate from the recipient model's own training trajectory; in these tests its per-sample norm still comes from the recipient's 400k–200k history.

Table 26: **Histories from every model scale reduce XL/2 FID by 4.3–7.1.** Sources use ordinary 400k–200k EMA pairs. Cosines compare the angular components defined relative to each source's own 400k field, averaging per-input cosines over the 8,192 fixed training inputs. For sampling, we project each prediction difference perpendicular to the XL/2 400k field and match its norm to the recipient's angular history, then apply it at $\alpha = 1$ with SDE-250 and no CFG. The matched random row is norm-matched in the same way.

History source	**Cosine to XL/2**	**FID-10k ↓**	**ΔFID ↓**	**Recall ↑**
None	–	20.53	0.00	.685
Matched random	≈0	20.53	0.00	.685
S/2	.020	16.21	−4.32	.673
B/2	.091	14.31	−6.22	.682
L/2	.175	**13.47**	**−7.06**	**.703**
XL/2	1.000	13.59	−6.94	.701

D.3 Sensitivity to checkpoint lag

For lag τ, unrelated to the generation time τ of Appendix E, let $d_\tau^{\mathrm{nat}} = d_{400\mathrm{k},400\mathrm{k}-\tau}$ be the angular direction before norm matching, and define its norm-matched version $d_\tau^{\mathrm{match}}(z) = d_\tau^{\mathrm{nat}}(z)\,\|d_{400\mathrm{k},200\mathrm{k}}(z)\|/(\|d_\tau^{\mathrm{nat}}(z)\| + \eta)$. The unscaled direction determines the FM drive and optimum. FID interventions match norms per sample; optimizer interventions match global RMS magnitudes as specified in Appendix F. We define

$$G(d_\tau^{\mathrm{nat}}) = -\frac{1}{D}\,\frac{\partial \mathcal{L}_{\mathrm{FM}}(u + \alpha d_\tau^{\mathrm{nat}})}{\partial \alpha}\bigg|_{\alpha=0}, \tag{6}$$

$$T_{\mathrm{FID}}(d_\tau^{\mathrm{match}}) = \frac{\mathrm{FID}(u - 0.25\,d_\tau^{\mathrm{match}}) - \mathrm{FID}(u + 0.25\,d_\tau^{\mathrm{match}})}{0.5}.$$

Positive T_{FID} means that the forward perturbation has lower FID than the backward perturbation. This is a centered finite-difference sensitivity, distinct from the terminal-state response in Appendix E. The optimizer-response quantities A_{norm} and C are defined in Appendix F. Longer lags increase both the FID sensitivity and the optimizer-response magnitude (Table 27). The loss columns use unscaled directions that share the 400k endpoint, so they are not controlled for endpoint fluctuations and do not compare preference strength across lags (Appendix D.4).

Table 27: **Longer checkpoint lags increase FID sensitivity and optimizer-response magnitude.** XL/2 at 400k; G and α^*_{FM} use the unscaled direction. T_{FID} measures FID-10k sensitivity (SDE-250, no CFG) along the norm-matched direction. A_{norm} and C measure the normalized magnitude and alignment after 250 optimizer updates at $\lambda = .25$, with global RMS-matched target perturbations (Appendix F). Bold marks the largest value in each of the last three columns.

Lag τ	**Field and sampler diagnostics**			**Optimizer response**	
	$G(d^{\mathrm{nat}})$	α^*_{FM}	$T_{\mathrm{FID}}(d^{\mathrm{match}})$	$10^3 A_{\mathrm{norm}}$	C
10k	-0.00045	-0.35	$+0.99$	1.27	.32
50k	-0.00126	-0.25	$+2.31$	1.36	.27
100k	-0.00108	-0.14	$+3.37$	1.71	.33
200k	-0.00082	-0.05	$\mathbf{+8.19}$	**2.33**	**.45**

D.4 Robustness to checkpoint endpoints

A negative loss-minimizing coefficient along a model's own angular history direction favors reversing that prediction change. This resembles the benefit of averaging recent weights (Kaddour, 2022; Karras et al., 2024b). We test whether such backward optima reflect fluctuations at the two checkpoints, which weight averaging would suppress. We shift or average the endpoints after the 50k-lag FM-drive sign changes s^* of S/2, L/2, and XL/2, defined below (Table 28), and for the 400k–200k direction (Table 29).

We first locate each model's sign change s^* and test whether the sampling benefit changes sign there. Let $d_s = d_{s,s-50\mathrm{k}}$ and $\hat{d}_s = d_s / \max(\|d_s\|, 10^{-12})$, with normalization per sample. The coordinate-averaged FM drive is $G_{\mathrm{FM}}(s) = -\frac{2}{D}\mathbb{E}_{\mathcal{P}}\langle u_s - v, \hat{d}_s\rangle$. Positive values favor forward motion along that direction. Unlike G in Table 27, G_{FM} uses a unit-norm direction per sample, so the two drives differ by the direction norm. We evaluate the drive on each scale's 10k checkpoint grid. For the first consecutive points $s_1 < s_2$ with $G_{\mathrm{FM}}(s_1) > 0 > G_{\mathrm{FM}}(s_2)$, linear interpolation gives $s^* = s_1 + (s_2 - s_1)G_{\mathrm{FM}}(s_1)/(G_{\mathrm{FM}}(s_1) - G_{\mathrm{FM}}(s_2))$. These crossing curves (Figure 4a) use a 50k lag; the 400k endpoint analysis uses a 200k lag. FID interventions use checkpoints $\{100, 150, 200, 300, 400\}$k for S/B and $\{100, 200, 300, 400\}$k for L/XL. FID corrections add the unit 50k-lag direction $\hat{d}_s$ at dose 1, SDE-250, and no CFG. At each sampling state, its norm is matched per sample to $\|P^{\perp,\eta}_{u_s}(u_{400\mathrm{k}} - u_{200\mathrm{k}})\|$: the same scale's 400k–200k difference is projected against the current base field u_s. All 18 measured FID gains are positive on both sides of each crossing (Figure 4b); the smallest, 2.01, is L/2 at 400k.

The crossings do not occur at a common quality level or where FM loss stops falling. Log-linear interpolation between measured baseline FID-10k checkpoints gives FID values of 31.70–63.72 at the crossings (Table 30b), rather than a common FID. FM loss on the fixed training inputs continues to fall, from 0.7603 near the XL/2 crossing to 0.7522 at 400k. The crossing step decreases with model size; the 109k S–B separation exceeds the 10k checkpoint grid, whereas the 9k L–XL difference does not.

For each row of Table 28, all directions are angularized against the field at the listed step s. The lag τ is 100k for L/2 and XL/2 and 200k for S/2. The original checkpoint pair is $(s, s-\tau)$. The earlier pair shifts both checkpoints to $(s-\delta, s-\tau-\delta)$, with $\delta = 30$k for L/2 and XL/2 and 50k for S/2. The averaged weights use the five EMA checkpoints from $s - 40$k to s and from $s - \tau - 40$k to $s - \tau$, at 10k intervals. We average weights before evaluating predictions. The noise proxy is the angular component of the current field minus the prediction from the recent averaged weights.

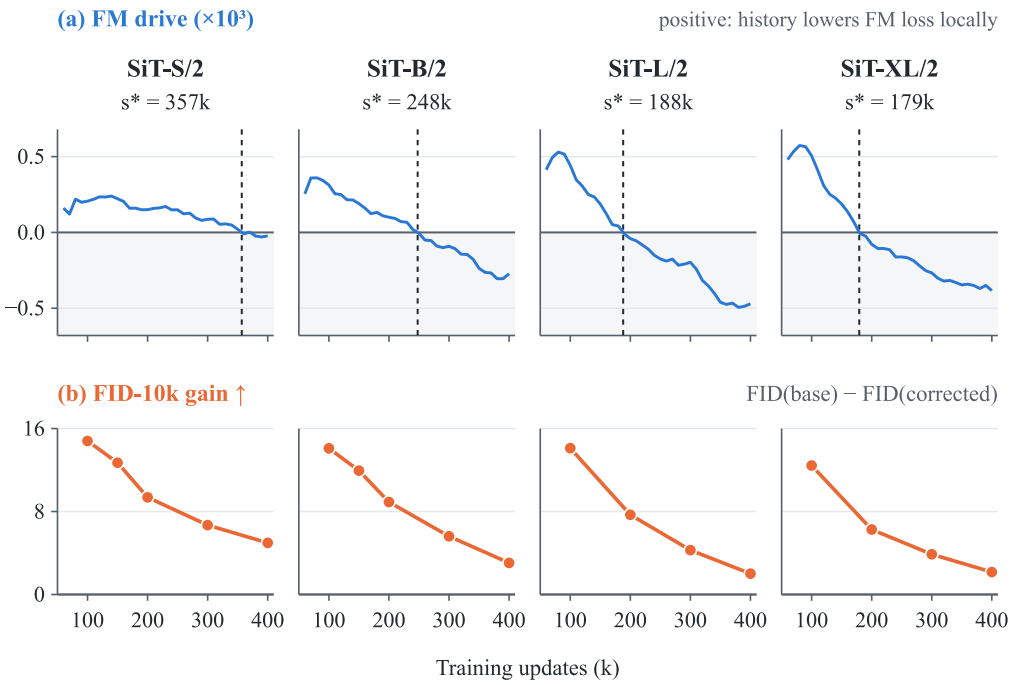

Figure 4: **FID improves on both sides of the FM-drive sign change.** (a) Signed FM drive along each model's own unit-normalized 50k-lag angular direction. Positive drive means that moving forward along history locally lowers FM loss. (b) Paired FID-10k gain from adding the same direction at dose 1 with SDE-250 and no CFG. Norm matching uses the same model's 400k–200k difference projected against the base field at each checkpoint. Axes are shared within each row; markers identify measured FID gains. Dashed lines in (a) mark interpolated crossings at 357k, 248k, 188k, and 179k for S/2, B/2, L/2, and XL/2. All 18 measured gains are positive; the smallest, 2.01, is L/2 at 400k. These crossings use each model's own history. Endpoint controls do not support interpreting every crossing as the onset of backward preference (Table 28).

Table 28: **Endpoint controls remove the 300k backward optima at L/2 and XL/2 but deepen the S/2 optimum at 500k.** α^*_{FM} is measured along directions from the original pair $(s, s-\tau)$, the earlier pair $(s-\delta, s-\tau-\delta)$, or averaged weights (Appendix D.4). The lag is 100k for L/2 and XL/2 and 200k for S/2. Depth is the reduction from the unperturbed loss to its minimum along a direction. Parentheses and the last column report depth relative to the original-pair direction; dashes mark near-zero denominators. Matched random directions have loss-minimizing coefficients near zero in every tested setting.

	α^*_{FM}			
Checkpoint	**Original pair**	**Earlier pair**	**Averaged weights**	**Noise-proxy depth / original-pair depth**
XL/2 230k (crossing + 51k)	+.003	+.037	+.042	–
XL/2 300k	−.057	+.045	+.018	2.4×
L/2 240k (crossing + 52k)	+.001	+.047	+.049	–
L/2 300k	−.053	+.031	+.008	1.5×
S/2 500k (crossing + 143k)	−.078	−.098 (1.96×)	−.088 (1.31×)	.25×

For the 400k comparison in Table 29, all directions are angularized against the base field $u = u_{400k}$. The original pair is 400k–200k; the earlier pair uses $u_{350k} - u_{150k}$. For the averaged-weight direction, we first average EMA *weights* over $\{160, 170, 180, 190, 200\}$k and $\{360, 370, 380, 390, 400\}$k, obtaining $\bar{\theta}_{\text{early}}$ and $\bar{\theta}_{\text{late}}$. We then evaluate $u_{\bar{\theta}_{\text{late}}} - u_{\bar{\theta}_{\text{early}}}$. This differs from averaging prediction fields.

Loss depth $D(d)$, distinct from the coordinate count D, permits comparisons across directions with different norms:

$$D(d) = \mathcal{L}_{\text{FM}}(u) - \min_{\alpha} \mathcal{L}_{\text{FM}}(u + \alpha d) = \frac{[\mathbb{E}_{\mathcal{P}}\langle u - v, d\rangle]^2}{\mathbb{E}_{\mathcal{P}}\|d\|^2}.$$

The parenthesized ratios in Table 29 compare this depth with that of the original 400k–200k direction. The noise proxy is $n = P_u^{\perp,\eta}(u_{400k} - u_{\bar{\theta}_{\text{late}}})$. Removing the component of d aligned with this proxy gives $d - \langle d, n\rangle n/\|n\|^2$, with inner products over direction vectors concatenated across the fixed inputs. The residual-depth ratio measures the remaining loss depth relative to the original direction. This does not isolate noise, because the recent-window difference can also contain training progress.

Table 29: **Endpoint-modified controls remove the XL/2 backward optimum but retain it at S/2.** Panel (a) compares the original 400k–200k pair, the earlier 350k–150k pair, and weights averaged over two five-checkpoint windows (Appendix D.4). Parentheses give loss depth relative to the original-pair direction. Panel (b) reports noise-proxy and residual depths relative to the same direction, then applies S/2, B/2, and L/2 histories to XL/2. Dashes denote cross-scale comparisons that were not measured. Random controls satisfy $|\alpha^*_{\text{FM}}| \leq .009$ in every tested setting.

(a) Endpoint controls: α^*_{FM}

Scale	Original pair	Earlier pair	Averaged weights
S/2	−.083	−.085 (1.46×)	−.086 (1.19×)
B/2	−.089	−.042 (.29×)	−.057 (.45×)
L/2	−.136	−.035 (.08×)	−.076 (.32×)
XL/2	−.053	+.021	.000

(b) Noise proxy and other-scale histories

Scale	Noise proxy	Residual	Other-scale α^*_{FM}
S/2	.001	1.10	–
B/2	1.1	.49	–
L/2	1.2	.48	–
XL/2	6.1	.07	+.014, +.093, +.235

The endpoint controls remove the XL/2 backward optimum, weaken the B/2 and L/2 optima, and leave the S/2 optimum intact. At XL/2, the earlier-pair direction has a positive loss-minimizing coefficient, and the averaged-weight direction's coefficient is near zero. At B/2 and L/2, both controls keep a negative coefficient but reduce its loss depth to .29–.45× and .08–.32×; at S/2 they increase it to 1.19–1.46×. After removing the noise-proxy component, the residual retains only 7% of the original 400k–200k direction's loss depth at XL/2, compared with 48–49% at B/2 and L/2 and 110% at S/2. Histories from S/2, B/2, and L/2, which share no checkpoints with XL/2, have positive loss-minimizing coefficients on the XL/2 field. With raw (non-EMA) weights, the loss-minimizing coefficients along the 400k–200k direction are $-.24$ at S/2 and $-.25$ at XL/2, compared with $-.083$ and $-.053$ with EMA weights, consistent with a contribution from endpoint fluctuations.

In Table 28, both controls remove the 300k backward optima at L/2 and XL/2 and deepen the S/2 optimum at 500k (1.96× and 1.31× depth). At 240k and 230k, about 50k after the L/2 and XL/2 crossings, the 100k-lag original pairs have near-zero loss-minimizing coefficients (+.001 and +.003), although the 50k-lag drive in Figure 4a is already negative. All controlled directions at these steps favor forward motion. Endpoint controls and a longer lag therefore do not support reading every 50k-lag sign crossing as the onset of backward preference.

To test whether the negative optima reflect fitting of the fixed training inputs, we repeat the loss-optimum and crossing measurements on validation inputs. The negative optimum signs and the crossing order persist. The validation optima are more negative at every scale, and the B/2 and XL/2 crossings occur about 20k earlier (Table 30); the endpoint controls above use training inputs.

Uncertainty estimates use 1,000 paired bootstrap resamples of the 8,192 validation inputs. For each crossing replicate, we reuse the two checkpoints $s_1 < s_2$ bracketing the observed crossing rather than repeat the full crossing search. The interpolated replicate is clipped to $[2s_1 - s_2, 2s_2 - s_1]$

before taking the 2.5th and 97.5th percentiles. The resulting range is conditional on that checkpoint pair.

Table 30: **Negative loss-minimizing coefficients and crossing order persist on validation inputs.** Panel (a) uses a 200k lag at 400k; brackets are 95% bootstrap intervals. Panel (b) uses a 50k lag; brackets give clipped 95% bootstrap percentile ranges conditional on the observed crossing's checkpoint pair (Appendix D.4), in thousands of updates. At B/2 both limits, and at S/2 the upper limit, equal the clipping bounds; XL/2's upper limit, 169.6k, lies just inside its bound. FIDs at s^*_{train} use log-linear interpolation between measured checkpoints and span a $2.01\times$ range. The crossings therefore do not occur at a common FID. On validation inputs, random angular controls satisfy $|\alpha^*_{\text{FM}}| \le 0.0004$. These measurements along each model's own history do not remove endpoint noise.

(a) Loss-minimizing coefficient

Scale	$\alpha^*_{\text{FM,train}}$	$\alpha^*_{\text{FM,val}}$ [95%]
S/2	$-.083$	$-.093$ $[-.102, -.085]$
B/2	$-.089$	$-.116$ $[-.126, -.106]$
L/2	$-.136$	$-.170$ $[-.179, -.161]$
XL/2	$-.053$	$-.098$ $[-.107, -.090]$

(b) Crossing order and FID

Scale	s^*_{train}	s^*_{val} [95%]	FID-10k
S/2	357k	360k $[341, 370]$	63.72
B/2	248k	228k $[210, 240]$	46.28
L/2	188k	182k $[174, 188]$	34.83
XL/2	179k	159k $[153, 170]$	31.70

E Input assignment and sampling response

We test whether a temporal correction is useful because of its magnitude, its assignment to the input, or its propagation through the sampler. Sampling interventions keep model weights fixed; a separate training intervention permutes the temporal secants in the training target.

E.1 Sampler dynamics and evaluation counts

We use generation time $\tau = 1 - t$ and initialize $X_0 \sim \mathcal{N}(0, I)$. Heun integrates $\dot{x} = f^{\text{H}}_u$, while Euler–Maruyama uses $dX_\tau = f^{\text{S}}_u\, d\tau + \sqrt{2(1-\tau)}\, dW_\tau$, with

$$f^{\text{H}}_u(x, \tau) = -u(x, 1-\tau, c), \qquad f^{\text{S}}_u(x, \tau) = -(1+\tau)\, u(x, 1-\tau, c) - x.$$

Here W_τ is Brownian motion, and the superscripts H and S denote the Heun and SDE drifts. Both processes run from noise ($\tau = 0$) to data ($\tau = 1$). Replacing u with $u + \gamma d$, where the dose γ plays the role of α in Appendix D, changes the drift by $\gamma a(\tau) d$, with $a_{\text{H}} = -1$ and $a_{\text{S}}(\tau) = -(1+\tau)$. For the SDE, we condition on the same Brownian path. The first-order response and its terminal value then obey

$$\dot{\delta x}_\tau = J_x f_u(x_\tau, \tau)\, \delta x_\tau + a(\tau)\, d(x_\tau, 1-\tau), \qquad \delta x_0 = 0, \tag{7}$$

$$\delta x_1 = \int_0^1 \Phi(1, \tau, x_\tau)\, a(\tau)\, d(x_\tau, 1-\tau)\, d\tau. \tag{8}$$

Here $\delta x_\tau = \partial x_\tau(\gamma)/\partial\gamma|_{\gamma=0}$ is the first-order trajectory displacement per unit dose, and $J_x f_u$ is the drift Jacobian. The state-transition operator $\Phi(1, \tau, x_\tau)$ propagates a perturbation from x_τ to the terminal state. The terminal response therefore depends on the correction and its location along the trajectory. This expression is a local derivative at zero dose; we test finite-dose additivity below.

The SDE-250 grid is $\tau_j = .96j/249$, $j = 0, \dots, 249$, with $\Delta = .96/249$. For $j < 249$ the Euler–Maruyama update is

$$X_{j+1} = X_j + \Delta f^{\text{S}}_u(X_j, \tau_j) + \sqrt{2(1-\tau_j)\Delta}\, \xi_j, \qquad \xi_j \sim \mathcal{N}(0, I),$$

followed by the drift-only endpoint step $X_{\tau=1} = X_{249} + .04 f^{\text{S}}_u(X_{249}, .96)$. Heun-51 uses $\tau_j = .96j/25$, $j = 0, \dots, 25$, and $\Delta = .96/25$. For $j = 0, \dots, 24$,

$$k_1 = f^{\text{H}}_u(x_j, \tau_j), \qquad k_2 = f^{\text{H}}_u(x_j + \Delta k_1, \tau_{j+1}), \qquad x_{j+1} = x_j + \Delta(k_1 + k_2)/2,$$

followed by $x_{\tau=1} = x_{25} + .04 f_u^{\mathrm{H}}(x_{25}, .96)$. SDE-250 and Heun-51 use 250 and 51 field evaluations, respectively. Each field evaluation requires one model call for ordinary or PSG sampling and two for CFG or a sampling-time composite. The corresponding total model-call counts are 250/500 and 51/102. Paired stochastic comparisons share X_0, labels, and every ξ_j.

E.2 Trajectory assignment controls

For the bin-wise analysis in Section 5.2, we partition model time t into $B_0, \ldots, B_4$: $[0, .2)$, $[.2, .4)$, $[.4, .6)$, $[.6, .8)$, and $[.8, 1]$. The symmetric terminal response is

$$\delta x_{1,i}^{(k)} = \frac{x_{1,i}(+\gamma d \text{ on } B_k) - x_{1,i}(-\gamma d \text{ on } B_k)}{2\gamma}.$$

This centered difference estimates terminal displacement per unit dose when the correction is active only in bin B_k. Both signs use the same recorded corrections and donor-to-recipient transformations.

Comparing replayed corrections under their original trajectory assignments (same-trajectory replay) and shuffled assignments tests whether input assignment matters. Recomputed history evaluates the correction at the current perturbed state; replay uses corrections recorded on the unperturbed baseline trajectory. For full-path FID evaluation, the recomputed, replayed, and shuffled corrections use dose 1 at every field evaluation; the random control uses the loss calibration below. We record u_e^i and d_e^i along trajectory i's unperturbed baseline. Index e distinguishes field evaluations, including separate Heun predictor and corrector calls at the same model time. For a donor j and recipient i at field evaluation e, we transform the recorded correction before replaying it:

$$\tilde{d}_e^{j \to i} = P_{u_e^i}^{\perp,\eta} \, d_e^j \, \frac{\|d_e^i\|}{\|P_{u_e^i}^{\perp,\eta} \, d_e^j\|}.$$

The projector removes the component of donor d_e^j parallel to recipient output u_e^i; the norm ratio restores the recipient's original correction magnitude. We replay this correction without reprojecting it at the perturbed state. A fixed same-class donor permutation assigns a different donor to each recipient throughout sampling; a per-evaluation donor permutation redraws the assignment at every field evaluation. Near-zero projected donors are redrawn before norm matching. Both shuffled controls preserve class, solver time, angularity relative to the recorded baseline output, and per-sample correction norm. Only the assignment of corrections to recipients changes.

Full-path FID evaluations draw 10 samples for each of the 1,000 classes, so every recipient has same-class donors. Permutation and Gaussian-control base seeds are 314159 and 271828, respectively, offset by sample-group index. The Gaussian control is rescaled so that its FM-loss increase equals the total, linear plus quadratic, increase from the recomputed history correction. The shuffled corrections are norm matched, but not matched in total FM-loss change.

Reassignment during training. The S/2 comparison in Section 5.2 starts at 200k and trains to 400k with the full secant, $w = .5$, $\rho = .5$, and past-checkpoint spacing 10k. The recent teacher is updated each step. Both models use global batch 256 and seed 41. At every active update, the shuffled-target model applies the same fixed-point-free permutation to both teacher outputs within each of eight 32-example subsets of the batch before taking their difference. This preserves the batch's set of full secants, but not their class and time association or each recipient's secant norm. Student inputs and original flow-matching velocities are unchanged. Unlike the sampling intervention above, this control does not isolate within-class, equal-time assignment.

FID uses 10k samples, SDE-250, and no CFG. Samples are generated in eight groups, using the same seeds $1000 + i$ for $i = 0, \ldots, 7$ for both models. The shuffled-target model reaches 62.12, against 40.55 for the matched PSG model and 61.80 for ordinary training (Table 25).

E.3 Coherence and terminal response

We measure coherence and terminal amplification with Heun-51 on 50 classes $\times$ 20 trajectories, separately from the full-path FID evaluations. The fixed and per-evaluation permutations use seeds 314159 and 314160, respectively; Gaussian directions use seed 271828. We assess local linearity using the unnormalized symmetric displacement $2\gamma\delta x_{1,i}^{(k)}$. Doubling the dose should approximately

double this displacement without rotating it. When γ increases from 0.5 to 1, displacements concatenated across trajectories within each bin have norm ratios within 1.7–2.3 and cosines of at least 0.95 in model-time bins $\mathcal{K} = \{0, 1, 2\}$ for same-trajectory replay and the fixed same-class donor permutation. Bins B_3 and B_4 do not meet these criteria and are excluded from the linear analysis. For same-trajectory replay, simultaneous perturbations in the retained bins agree with their summed individual responses (global cosine 0.994, norm ratio 1.03).

With $\widehat{\delta x}_{1,i}^{(k)} = \delta x_{1,i}^{(k)} / \|\delta x_{1,i}^{(k)}\|$, coherence is measured by $C_{\text{align}} = \mathbb{E}_i[\|\sum_{k\in\mathcal{K}} \widehat{\delta x}_{1,i}^{(k)}\|^2 / |\mathcal{K}|]$. Normalizing the responses removes their magnitudes, so this statistic measures whether corrections in different time bins move the same trajectory in aligned terminal directions. Perfect alignment gives $|\mathcal{K}| = 3$, independent directions average near one, and the matched random control measures .998. All four conditions match the recipient history norm, so the reported terminal amplification $M_k = \mathbb{E}_i\|\delta x_{1,i}^{(k)}\|$ is comparable across them. The fixed same-class donor is slightly more coherent than same-trajectory replay (1.780 versus 1.722). Replay has larger terminal amplification in all three retained bins, most in B_2, where it is 3.1 times the donor's (4.07 versus 1.31), and only replay improves FID over the uncorrected 20.54 (Tables 7 and 31; Figure 5). Coherence therefore does not separate the two assignments; amplification and FID do.

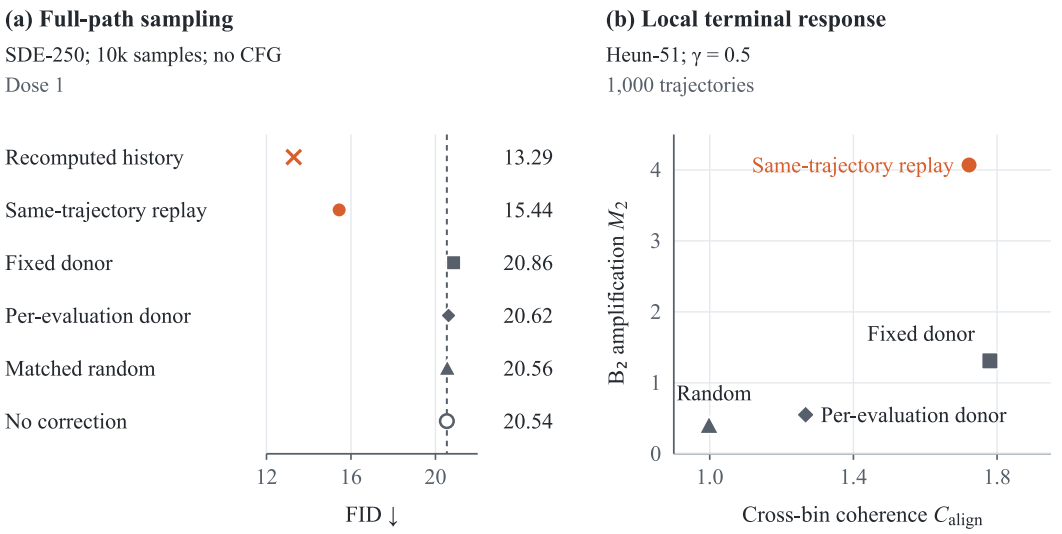

Figure 5: **Input assignment matters beyond directional coherence.** (a) Replaying corrections on their original trajectories retains most of the FID gain; same-class reassignment removes it. The dashed line marks the uncorrected baseline. FID uses full-path corrections at dose 1, SDE-250, 10k samples, and no CFG. (b) In separate Heun-51 measurements at $\gamma = .5$ on 1,000 trajectories, the fixed same-class donor has higher cross-bin coherence C_{align} than same-trajectory replay, but replay has 3.1$\times$ its terminal amplification M_2. Coherence uses bins B_0–B_2; amplification is shown for B_2, where the replay–donor gap is largest. Random controls match the FM-loss increase in (a) and the recipient history norm in (b). Both panels use XL/2 at 400k; Tables 7 and 31 give the corresponding measurements.

Table 31: **Greater coherence need not imply greater amplification or FID improvement.** XL/2 at 400k. Response statistics use Heun-51 at $\gamma = .5$; FID comes from the separate full-path SDE-250 experiment in Table 7. C_{align} measures alignment of terminal responses from model-time bins B_0–B_2; M_2 is terminal amplification in $B_2 = [.4, .6)$. The fixed same-class donor is more coherent than same-trajectory replay but has only one third of its B_2 amplification. The random row matches the recipient history norm for C_{align} and M_2 and the total FM-loss increase for FID.

Correction	C_{align}	M_2	**SDE-250 FID ↓**
Same-trajectory replay	1.722	4.07	**15.44**
Fixed same-class donor	1.780	1.31	20.86
Per-evaluation donor	1.267	.55	20.62
Matched random angular	.998	.40	20.56

E.4 Time and spatial-frequency components

We compare the components that changed during early training with those present in the later temporal contrast. We use the ordinary XL/2 field and its 400k–200k history, with the same five model-time bins. A 32×32 orthonormal DCT-II is applied independently to each latent channel. The 1,024 spatial frequencies are ordered by radius $\sqrt{k_x^2 + k_y^2}$, with flattened-index order resolving ties, and split into three nearly equal-count bands. Including four channels, their degrees of freedom are 1,368, 1,364, and 1,364.

Let $\mathcal{P}_i$ contain the fixed training inputs whose interpolation time lies in bin i, and let P_j select frequency band j with D_j coefficients. The history energy and early rate of angular change are

$$\bar{H}_{ij} = \frac{\mathbb{E}_{\mathcal{P}_i} \|P_j\, d_{400\text{k},200\text{k}}\|^2}{D_j}, \qquad \bar{V}^{\text{ang}}_{ij,\text{early}} = \frac{1}{3} \sum_{s \in \mathcal{S}} \frac{1}{50\text{k}} \sqrt{\frac{\mathbb{E}_{\mathcal{P}_i} \|P_j\, d_{s,s-50\text{k}}\|^2}{D_j}},$$

where $\mathcal{S} = \{100\text{k}, 150\text{k}, 200\text{k}\}$. For $u = u_{400\text{k}}$, $d = d_{400\text{k},200\text{k}}$, and example $n \in \mathcal{P}_i$, we define the per-example FM drive $g_n^j = -2\langle u - v, P_j d\rangle$ and its absolute mean-to-standard-deviation ratio $S_{ij} = |\mathbb{E}_{\mathcal{P}_i} g_n^j| / (\text{Std}_{\mathcal{P}_i}(g_n^j) + 10^{-12})$. All means and standard deviations are conditional on the time bin.

Table 32: **History energy tracks faster-changing components.** ρ denotes Spearman rank correlation (unrelated to the reference ratio), and S is the absolute mean-to-standard-deviation ratio of the per-example FM drive. All measurements use XL/2; correlations are across 15 time-bin–frequency combinations. $\bar{V}_{\text{early}}$ is the early rate of angular change and $\bar{H}$ is history energy per degree of freedom, each within a time-bin–frequency combination. L_{term} measures the response to replayed corrections derived from slow- or fast-changing bands of the recorded history and reprojected perpendicular to the recorded baseline output, in model-time bins B_0–B_2. The negative difference indicates a larger response for the fast-derived correction; reprojection need not preserve frequency support. Brackets are 95% bootstrap intervals: resampling fixed inputs within each time bin for $\rho(\bar{V}_{\text{early}}, \bar{H})$, and paired class-hierarchical trajectory resampling for the terminal response.

Diagnostic	**Measurement**
$\rho(\bar{V}_{\text{early}}, \bar{H})$	$+.900$ $[.88, .94]$
$\rho(S, \bar{H})$ (alternative)	+.204
$L_{\text{term,slow}} - L_{\text{term,fast}}$	-12.3 $[-13.2, -11.5]$

Pooled across the 15 time–frequency combinations, later history energy correlates positively with the early rate of angular change (Spearman .90). As an alternative predictor, S_{ij} measures how consistently the per-example FM drive favors or opposes band j of the history within time bin i. If history accumulated where this drive is inconsistent, S_{ij} would correlate negatively with history energy; the observed correlation is .204 (Spearman, 95% interval $[-.011, .425]$), so the predicted negative correlation is not supported. To compare the terminal responses of slow- and fast-changing

components, we select the slowest and fastest bands in each time bin by $\bar{V}^{\mathrm{ang}}_{\mathrm{early}}$, breaking ties from low to high frequency. We retain the selected frequency components of the 400k–200k history recorded on each baseline trajectory and reapply $P^{\perp,\eta}$ against the recorded output. We replay the result after matching each direction to the full history's global RMS over all 51 baseline Heun evaluations. Angular reprojection can add components outside the selected band, so these are band-derived, not strictly band-limited, corrections. The correction is active only in model-time bins B_0–B_2 ($t \in [0, .6)$), which met the linearity criteria for full replayed corrections.

On 50 classes $\times$ 20 trajectories, using Heun-51 and dose $\gamma = .5$, the normalized terminal response is

$$L_{\mathrm{term}}(d) = \mathbb{E}_i \left[\|\delta x_{1,i}(d)\| \Big/ \sqrt{\tfrac{1}{D|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d} \|d_i(e)\|^2} \right].$$

Here i indexes trajectories rather than the time bins above, and $\mathcal{E}_d$ contains the field evaluations at which the correction is applied. The displacement $\delta x_1 = [x_1(+\gamma d) - x_1(-\gamma d)]/(2\gamma)$ is measured at the data endpoint, $\tau = 1$. The denominator is the correction's coordinate RMS, not a vector L2 norm. Corrections derived from fast-changing bands produce larger normalized terminal responses than those derived from slow-changing bands (Table 32).

F Local response to a modified training target

A target perturbation specifies a change in predictions, but learning acts on model parameters through the student Jacobian and optimizer. We measure the resulting prediction change by comparing short continuations from the same XL/2 400k student and Adam state, with and without a target perturbation.

Matched target perturbations. Starting from the ordinary XL/2 400k model and its Adam state, each perturbed continuation trains toward $v + \lambda\,\mathrm{sg}[d]$ for 250 updates at $\lambda = .25$, with d the RMS-matched target perturbation defined below. The paired unperturbed continuation uses ordinary flow matching with the same data and random draws. Teacher EMA weights remain fixed, and the target perturbation is recomputed on each minibatch. Each direction is rescaled once by a single factor to match the 400k–200k angular history direction's root-mean-square magnitude across all inputs and latent coordinates. The evaluation set $\mathcal{P}$ contains 8,192 fixed training inputs, constructed as in Appendix D.1. In addition to temporal angular and parallel directions, the controls include random parameter-tangent directions, smooth random fields, and histories from other model sizes.

The parameter-tangent control computes eight exact Jacobian–vector products at the fixed 400k EMA teacher. Each parameter perturbation $\delta\theta$ has the same tensor shapes as the teacher's parameters, is sampled with independent standard Gaussian entries, and is globally rescaled to $\|\delta\theta\| = 10^{-4}\|\theta\|$ (seeds 0–7).

The smooth controls use two fixed random-convolution fields (seeds 0–1). Each takes the four-channel latent plus a constant channel containing time t, with no class input. Three 3×3 convolutions with padding one have channel widths $5 \to 64 \to 64 \to 4$, with GELU after the first two. Weights are independent $\mathcal{N}(0, 1/\mathrm{fan_in})$ and biases $\mathcal{N}(0, .1^2)$. Both control families are projected perpendicular to the fixed 400k EMA output before matching RMS magnitudes over $\mathcal{P}$; Table 33 reports their median responses.

Response magnitude and alignment. Let d denote the RMS-matched target perturbation and Δu the difference between perturbed and unperturbed student predictions after N paired updates with target dose λ. Predictions use the student weights without EMA averaging. We measure

$$A_{\mathrm{norm}} = \frac{\sqrt{\mathbb{E}_{\mathcal{P}}\|\Delta u\|^2}}{\lambda N \sqrt{\mathbb{E}_{\mathcal{P}}\|d\|^2}}, \qquad C = \frac{\mathbb{E}_{\mathcal{P}}\langle \Delta u,\, d\rangle}{\sqrt{\mathbb{E}_{\mathcal{P}}\|\Delta u\|^2}\sqrt{\mathbb{E}_{\mathcal{P}}\|d\|^2}}.$$

A_{norm} is the dimensionless field-response gain per unit dose and optimizer update. The cosine C measures the alignment of the learned response with the target perturbation: one indicates alignment, zero orthogonality, and negative values opposition. For the projected random parameter-tangent control, the interquartile range of A_{norm} is $.23 \times 10^{-3}$ around its median of 3.70×10^{-3}, so the angular perturbation's 2.33×10^{-3} lies well outside this spread.

Table 33: **Angular history produces a smaller, less aligned response than the random controls.** Equal-RMS target perturbations at $\lambda = .25$, applied for 250 updates from the ordinary XL/2 400k model. A_{norm} is normalized response magnitude and C is alignment with the target perturbation; random rows report medians over seeds, and ranges summarize histories from the three other model sizes.

Target perturbation	$10^3 A_{\mathrm{norm}}$	C
Temporal angular, 200k lag	2.33	.448
Temporal parallel, 200k lag	3.43	.839
Random smooth	3.33	.856
Projected random parameter-tangent	3.70	.889
Other-model histories	1.14–1.34	.228–.255

The temporal angular perturbation produces 63% of the response magnitude of the projected random parameter-tangent control, with cosine .45 versus .89. The parallel component and smooth random field have higher alignment, .839 and .856, respectively. Histories from other model sizes produce smaller, less aligned responses than those to the XL/2 temporal angular direction. Equal target-space magnitudes therefore do not lead to equal changes in the learned field. These measurements describe the local response at this checkpoint and optimizer state, rather than the effect accumulated over full PSG training.