RETHINKING CLASSIFIER-FREE GUIDANCE IN REWARD-BASED DIFFUSION POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward-based diffusion post-training updates model parameters from task rewards, yet the role of classifier-free guidance (CFG) in this process remains unclear. We identify an *Inference–Learning Mismatch*: guidance can improve a fixed model's outputs while stronger guided collection weakens subsequent learning. This contrast motivates using CFG to organize reward measurements rather than prescribe a training direction. We construct local reward-query coordinates from the conditional–unconditional residual and compare symmetric same-state interventions to identify reward-preferred changes. The resulting reward differences provide directional credit tied to a known intervention, complementing the sample-level credit of individual images. To combine these signals in a parameter update, we introduce Fiber-Augmented Negative-Aware Fine-Tuning (FANFT), which pulls local responses into parameter space and fuses them with global NFT credit according to agreement between disjoint estimates. Frozen-policy tests show that these coordinates predict held-out preferences better than matched random coordinates, while training comparisons connect directional writing to improved learning on condition-sensitive rewards. Across tasks, the gains depend on how well the coordinate represents reward-improving directions. These findings place reward–coordinate alignment at the center of CFG's post-training role: the residual organizes local queries, and global reward supervision complements the directional information they provide.

1 INTRODUCTION

Reward-based diffusion post-training iteratively updates model parameters to maximize the reward of generated samples. DDPO and DPOK pursue this objective through denoising-policy optimization (Black et al., 2024; Fan et al., 2023), while DiffusionNFT learns from rewarded clean images through the forward process (Zheng et al., 2026). These advances make reward optimization a practical route for adapting pretrained generators. They also raise a question about how to use a central ingredient of diffusion inference during learning: *classifier-free guidance* (CFG).

CFG improves conditional generation by amplifying the conditional–unconditional residual (Ho & Salimans, 2022). It is therefore natural to collect training samples with stronger guidance: better current generations appear to offer better material for reward optimization. This expectation, however, links two different decisions—how to steer the current trajectory and how to improve future model parameters. Understanding that link requires asking: **what role should CFG play when the model itself is optimized by reward?**

Our controlled four-reward study exposes an *Inference–Learning Mismatch* between these decisions. Increasing inference CFG improves a frozen model on GenEval, OCR, HPSv2, and PickScore (Figure 1, column (a)), yet stronger CFG during collection yields weaker learning under a common evaluation protocol (Figure 1, column (b)). The scalar sweep likewise gives task-dependent collection preferences (Figure 1, column (c)). Together, these results separate the immediate effect of a chosen behavior field from the information that its rewarded samples provide for parameter updates.

This distinction suggests a different use of conditional structure: **CFG should define where to measure, not which way to move.** We use the residual $r = v_c - v_u$ to construct a compact local action space and compare opposite velocity interventions at the same prompt, latent, and solver state. Because the intervention is known, its reward contrast provides *directional credit*: the coordinate

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.

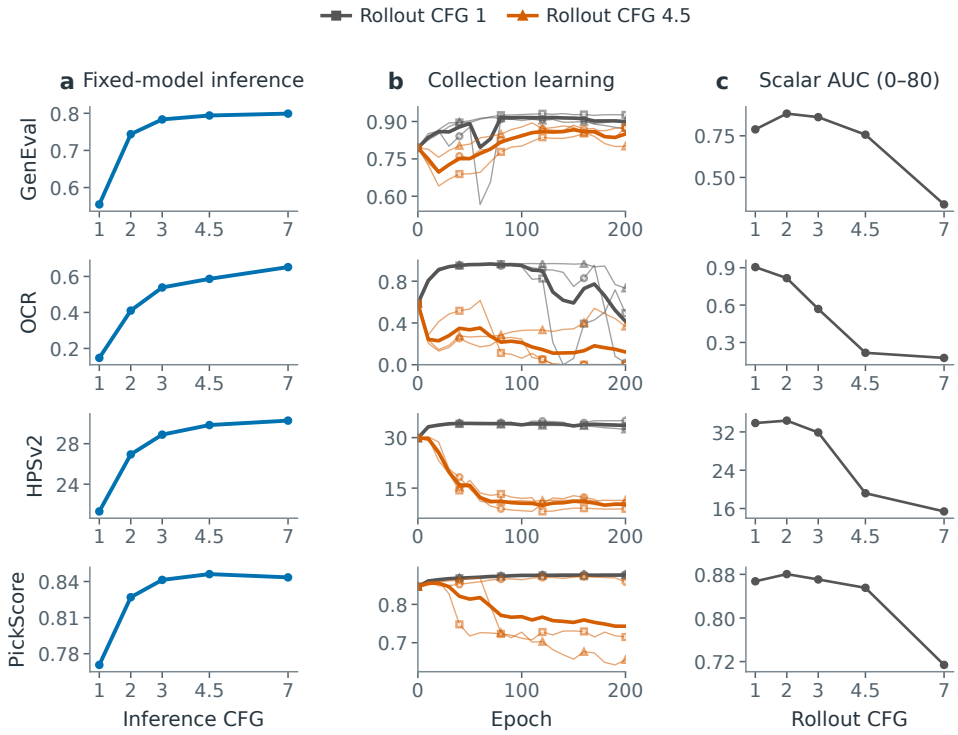

Figure 1: **The Inference–Learning Mismatch.** Columns show **(a)** fixed-model inference sweeps; **(b)** raw checkpoint evaluation at CFG 4.5 after collection at CFG 1 or 4.5, with individual trajectories in thin lines and their mean in thick lines over epochs 0–200; and **(c)** normalized AUC for five rollout scales over epochs 0–80, evaluated at CFG 4.5. Rows, from top to bottom, are GenEval, OCR, HPSv2, and PickScore in their native reward units.

comes from the model, but the preferred sign comes from reward. This information complements NFT *sample credit*, which assesses a completed image relative to its reward group.

We first test whether this directional information exists without changing model parameters. In frozen-policy experiments, residual-anchored coordinates predict held-out reward preferences better than norm- and dimension-matched random coordinates (Figure 2(d–f), Anchor), and current-state response coefficients outperform transferred ones (State/Prompt). These findings lead to **Fiber-Augmented Negative-Aware Fine-Tuning (FANFT)**, our implementation of the reward-query principle. FANFT pulls local responses into parameter space and uses agreement between disjoint estimates to control their contribution to the global NFT gradient, so every generated image retains its sample-level supervision.

We then examine the value and scope of writing this information into the update. Credit-channel comparisons show improved GenEval and OCR learning (Figure 3(c,d)); the four-reward system comparison connects these gains to condition-sensitive objectives and reveals the compact coordinate's boundary on broader preferences (Figure 4). Our contributions are threefold:

- **We identify the Inference–Learning Mismatch of CFG.** Fixed-model guidance can improve generation while stronger guided collection weakens subsequent learning (Section 2).
- **We introduce a reward-query role for CFG and instantiate it in FANFT.** Residual-based coordinates, symmetric reward probes, and agreement-controlled pullback connect local directional credit to global NFT supervision (Sections 3–4).
- **We characterize reward–coordinate alignment through four-reward evidence.** GenEval and OCR show the strongest gains and lower checkpoint-acquisition cost; HPSv2 remains close, while PickScore delineates the compact coordinate's scope (Figures 3(c,d), 4).

2 THE INFERENCE–LEARNING MISMATCH

2.1 BETTER INFERENCE GUIDANCE DOES NOT DEFINE A BETTER TRAINING POLICY

Let $v_c = v_{\bar{\theta}}(y, t, c)$ and $v_u = v_{\bar{\theta}}(y, t, \varnothing)$ be conditional and unconditional behavior-model velocities. CFG forms

$$r = v_c - v_u, \qquad v_b = v_u + wr, \tag{1}$$

Here w is fixed within each collection or evaluation protocol. With model parameters held fixed, the inference sweeps establish the benefit of stronger guidance (Figure 1, column (a)). We next examine whether this benefit carries over when guidance determines the samples used for learning.

Guidance changes the learning distribution. We compare rollout CFG 1 and 4.5 across compositional, text-rendering, and preference rewards, evaluating the resulting checkpoints with the same guidance (Ghosh et al., 2023; Wu et al., 2023; Kirstain et al., 2023). Rollout CFG 1 yields higher integrated evaluation reward in every task (Figure 1, column (b)). Since evaluation guidance is fixed, the curves reflect differences in the learned models. Inference guidance changes the outputs of the current model; collection guidance also changes the observations from which that model learns. This distinction motivates querying local changes around the behavior field and using their rewards to guide the next update (Appendix C).

2.2 A SCALAR SCALE CANNOT PROVIDE STATE-LOCAL DIRECTIONAL CREDIT

The scalar sweep does not reveal a universal collection setting across tasks (Figure 1, column (c)). Changing a rollout scale also changes entire trajectories, so their terminal rewards do not isolate the effect of a particular local action. A same-state comparison supplies that missing link: hold the prompt, latent, and solver state fixed, apply opposite interventions, and compare the resulting rewards.

This preserves the conditional structure exposed by CFG while separating the sampling decision from the direction learned from reward. The coordinate determines which changes to compare; the reward contrast determines their preferred sign. Before using that contrast in training, two questions must be answered: does the residual define an informative coordinate, and can its measured response be reused across states? We address both with a frozen policy, so the next section examines the measurements independently of parameter updates.

3 CFG RESIDUALS PROVIDE REWARD-QUERY COORDINATES

3.1 CONSTRUCTING HELD-OUT LOCAL ACTIONS

To test coordinate information directly, we change a selected velocity from v_b to $v_b + \delta$ and complete sampling with the unchanged behavior model. For a fixed prompt and state, $Q(\delta)$ denotes the resulting terminal reward. We organize these actions by projecting the residual into $v_b^{\perp}$ and completing a small orthonormal basis:

$$\widetilde{r}_i = r_i - \frac{\langle r_i, v_{b,i} \rangle}{\|v_{b,i}\|^2 + \varepsilon_{\text{num}}} v_{b,i},$$
$$U_i = \mathrm{QR}\left[\widetilde{r}_i, P_{v_{b,i}^{\perp}} q_{i,1}, \dots\right]. \tag{2}$$

The candidates $q_{i,j}$ are generated deterministically from the state, independently of reward, and degenerate axes are excluded. A unit coordinate a_i therefore specifies a known action $u_i = U_i a_i$, which we apply symmetrically:

$$v_{i,\pm}^{(0)} = v_{b,i} \pm \sigma_i u_i. \tag{3}$$

The two branches share the prompt, initial latent, and first-velocity intervention location, then continue under the behavior field at their respective states. Their terminal rewards provide the finite-difference response

$$\widehat{d}_i = \frac{R_{i,+} - R_{i,-}}{2\sigma_i}. \tag{4}$$

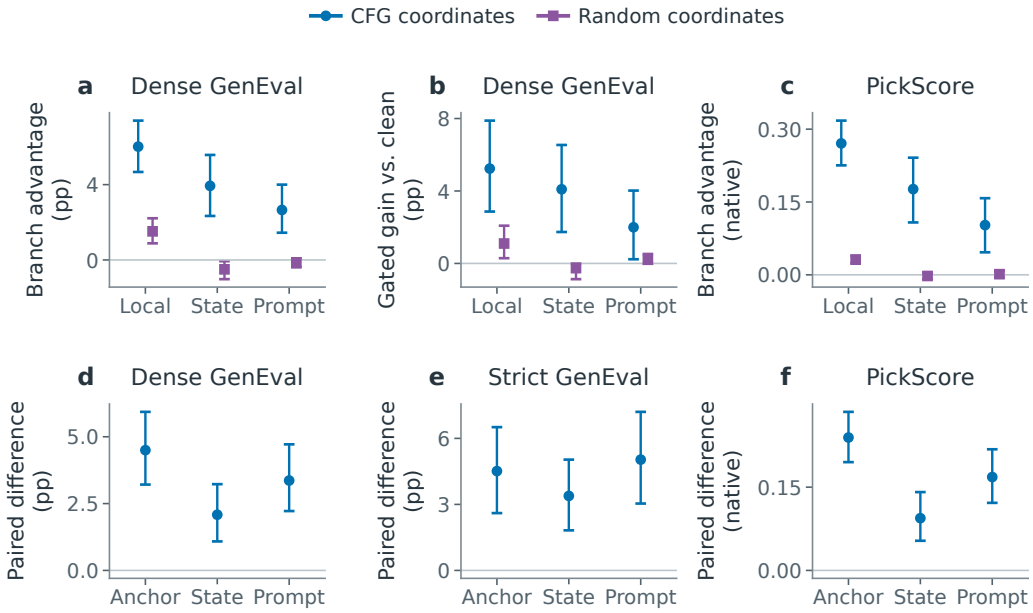

Figure 2: **CFG coordinates support local prediction and selective use.** (a,c) Branch advantage with local or transferred coefficients (State/Prompt). (b) Dense-GenEval clean-relative gain under agreement-based selection. (d–f) Paired contrasts: CFG minus random (Anchor), and local minus transferred coefficients within CFG coordinates (State/Prompt). GenEval uses percentage points; PickScore uses native units. Bars: 95% prompt-cluster bootstrap intervals. The paired contrasts isolate anchor information and response locality; the sampling protocol is in Appendix C.

For a smooth response, symmetry cancels even-order terms and gives second-order bias; for a discrete reward, the same contrast defines a finite branch preference. The transverse projection makes the probe a directional change rather than a first-order change in speed along v_b. Together, these choices separate the coordinate supplied by the model from the preference measured by reward.

A matched test of coordinate information. We isolate the residual anchor by replacing it with a random axis while keeping the policy, recipient state, dimension, shared auxiliary axis, and action norm fixed. Coefficients fitted on discovery probes predict reward preferences on directions excluded from fitting and selection. For a predicted sign $\widehat{s} \in \{-1, 0, 1\}$, the quality of this held-out prediction is measured by

$$B = \widehat{s}\,\frac{R_+ - R_-}{2}, \tag{5}$$

the selected branch's advantage over its symmetric-pair mean. Ties and abstentions contribute zero. Prompt-cluster intervals and the complete sampling protocol are specified in Appendix C.

3.2 Residual coordinates predict held-out reward preferences

The held-out test confirms that residual-anchored coordinates predict useful branch choices (Figure 2(a), Local). Their within-prompt advantage over random replacement is **4.50 percentage points [3.21, 5.93]** (Figure 2(d), Anchor). Since state, dimension, shared axis, and action norm are fixed, this difference isolates information supplied by the residual anchor. The reward measurements use that information to choose a branch, rather than inheriting a fixed preference for the positive residual direction.

This anchor advantage also appears under strict GenEval and independently scored PickScore (Figure 2(e,f), Anchor). Dense and strict GenEval vary scoring granularity within the same evaluator

stack, while PickScore supplies a separate readout of the same images. The agreement across these readouts supports using the residual to organize reward queries.

Proposition 2 provides context for this role of structure: an isotropic m-dimensional subspace of an n-dimensional working space captures an expected m/n fraction of local reward-gradient energy. The matched frozen test asks the corresponding empirical question—whether the chosen anchor supports predictable finite reward choices. Appendix B gives the geometric interpretation of this velocity-transverse working space.

3.3 Reward response is state-local

Having established anchor information, we ask whether its fitted response transfers across states. Keeping each recipient's held-out actions and rewards fixed, current-state coefficients outperform those transferred across latent states or prompts (Figure 2(a), Local/State/Prompt). The paired intervals are positive for dense GenEval and remain positive under the other readouts (Figure 2(d–f), State/Prompt).

Because donor coefficients are interpreted in the recipient's basis, this comparison tests the coordinate–response mapping. It distinguishes a reusable construction rule from a state-dependent reward preference: the residual specifies the local basis, but its coefficients must be measured where they will be used. FANFT follows this distinction when turning the frozen-query construction into an online learning procedure.

4 FANFT: Writing Directional Credit into Parameters

4.1 One collection supplies two credit channels

FANFT implements the reward-query principle as an additional credit channel, using NFT as the base objective. Every terminal image, including both branches of a symmetric pair, enters the NFT reward group. Prompt-relative rewards define negative-aware supervision and the global gradient $g_N = \nabla_\theta \mathcal{L}_{\mathrm{NFT}}$ (Zheng et al., 2026); Appendix A gives the normalization and reconstruction objective.

The same paired images also supply the response in Eq. (4). To write it into the model, set $h_i = m_i \widehat{d}_i$ for retained dimension m_i and project the current prediction onto the stored action:

$$p_i(\theta) = \langle v_\theta^{\mathrm{cfg}}(y_i, t_i, c_i; w), u_i \rangle,$$
$$\mathcal{L}_F^f = -\frac{1}{|\mathcal{I}_f|} \sum_{i \in \mathcal{I}_f} h_i p_i(\theta). \tag{6}$$

Collected states, directions, radii, and reward coefficients are stop-gradient quantities, so differentiation acts on the current velocity prediction. Gradient descent then increases its projection toward the reward-preferred branch, while NFT continues to supervise the completed image.

The channels are complementary because sample quality and local preference need not agree. Both images in a pair can have low reward relative to their prompt group, yet their difference can identify a preferred action. NFT keeps the sample-level assessment, while the directional channel associates that preference with the known intervention. Thus the pair adds information without changing its global labels. Antithetic forward-process noise views reuse completed images; symmetric rollout probes create the comparison itself (Appendix A).

4.2 Agreement controls gradient writing

Held-out evidence for agreement. A useful branch ordering does not by itself determine whether to intervene. We therefore also measure clean-relative gain: the reward change from the selected action compared with leaving the state unperturbed. Selecting on discovery agreement, and retaining the clean reference otherwise, gives positive held-out gain in CFG coordinates (Figure 2(b), Local), including strict GenEval and PickScore (Figure 10(b,c), Local). This motivates selective use of the measured response; during training, agreement is assessed after parameter pullback.

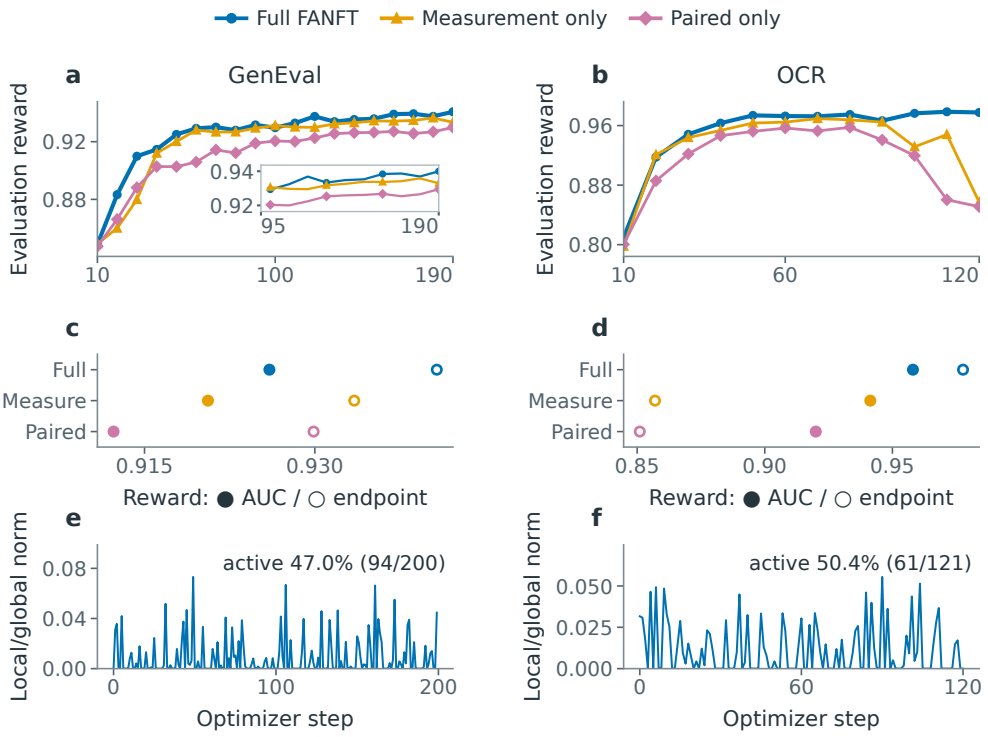

Figure 3: **Directional writing improves integrated reward and acts selectively.** Left: GenEval; right: OCR. (a,b) Raw evaluations over epochs 10–190 for GenEval and 10–120 for OCR. (c,d) Window-normalized AUC (filled) and the corresponding interval-end reward at epoch 190 / 120 (open). (e,f) Applied local/global gradient-norm ratios, including zero corrections; annotations give active fractions and valid-update counts. All configurations retain paired rewards and global NFT learning; the controls assign no local correction.

Parameter-space fusion. To assess agreement in the space where learning occurs, let g_A, g_B be gradients from two balanced disjoint folds and g_F their count-weighted mean. Define

$$S = \max(0, \langle g_A, g_B \rangle), \quad N = \tfrac{1}{4}\|g_A - g_B\|^2, \quad \kappa = S/(S + N + \varepsilon_{\text{num}}). \tag{7}$$

Here S measures their shared direction and N their disagreement. The resulting weight κ controls a complementary correction: project g_F orthogonally to g_N, cap its magnitude, and norm-match the fused gradient:

$$g_F^{\perp} = g_F - \frac{\langle g_F, g_N \rangle}{\|g_N\|^2 + \varepsilon_{\text{num}}} g_N, \qquad \alpha_F = \min\left(1, \frac{\rho\kappa\|g_N\|}{\|g_F^{\perp}\| + \varepsilon_{\text{num}}}\right),$$
$$g_{\text{raw}} = g_N + \alpha_F g_F^{\perp}, \qquad g_{\text{final}} = \frac{\|g_N\|}{\|g_{\text{raw}}\| + \varepsilon_{\text{num}}} g_{\text{raw}}. \tag{8}$$

Ignoring stabilizers, the local correction is orthogonal to g_N and bounded by $\rho\kappa\|g_N\|$, while norm matching preserves the input-gradient magnitude before optimizer preconditioning. Projection therefore removes the component already supplied by NFT, and agreement controls how much complementary direction is added. A zero correction recovers the global update. Proposition 1 formalizes this controlled rotation; Appendix A specifies the training hyperparameters.

Selective directional updates. The training trajectories show how this rule operates: the applied ratio $\|\alpha_F g_F^{\perp}\|/\|g_N\|$ varies across updates and includes zeros (Figure 3(e,f)). Directional writing is therefore intermittent and bounded, while every collected image continues to receive global supervision. These diagnostics describe when the extra channel is used; the following comparison tests its effect on learning. Table 11 summarizes norm ratios and alignment.

4.3 DIRECTIONAL WRITING IMPROVES LEARNING

To isolate directional writing, we keep paired collection and global NFT supervision in place and change how the local response is used (Table 3). **Measurement only** estimates the local gradient but gives it zero update weight; **Paired only** collects the paired rewards without estimating that gradient. All remaining settings match FANFT within each task (Appendix C.3).

We evaluate all three configurations without interventions over the same within-task window: epochs 10–190 for GenEval and 10–120 for OCR. Normalized AUC summarizes reward across the window, and the open marker reports its final checkpoint reward.

FANFT achieves higher window-normalized AUC than both no-write controls on GenEval and OCR (Figure 3(c,d), filled markers). The gains over measurement only are **0.54 and 1.67 percentage points**, respectively, with the underlying trajectories shown in panels (a,b). Since the controls retain the paired rewards and global objective, the comparison links the gain to using the measured direction in the update, rather than simply collecting paired samples.

This ordering also holds at the interval end: FANFT has the highest reward at GenEval epoch 190 and OCR epoch 120 (Figure 3(c,d), open markers; Table 10). We next examine how the complete method behaves across reward objectives.

5 DIRECTIONAL CREDIT IMPROVES REWARD-BASED LEARNING

5.1 CROSS-REWARD BEHAVIOR REVEALS THE OPERATING REGIME

We compare FANFT with standard NFT under matched within-task protocols, using checkpoint evaluation to assess the learned model and collection reward to describe its training samples (Figure 4(a–h)). Table 1 summarizes the evaluation scores and target-reaching ratios. FANFT improves GenEval in both readouts and reaches high OCR reward earlier (a–d), while HPSv2 remains close (e,f). PickScore separates the readouts: collection rewards are similar (h), but NFT achieves higher checkpoint evaluation (g).

We interpret this pattern through **reward–coordinate alignment**. The CFG residual arises from changing text conditioning, making it a natural source of condition-sensitive directions. GenEval and OCR directly reward such changes, whereas HPSv2 and PickScore also reflect broader visual and aesthetic preferences. The task differences therefore connect the benefit of local credit to how well the compact coordinate represents reward-improving changes, while motivating retention of the global reward signal.

5.2 ALIGNED REWARDS YIELD LOWER CHECKPOINT-ACQUISITION COST

The condition-sensitive tasks also let us measure the cost of reaching useful checkpoints. Figure 4(i,j) and Table 1 report the H200 GPU-hour budget at the first recorded checkpoint meeting each target; Appendix D.6 gives the corresponding image counts.

At GenEval reward 0.95, the **NFT/FANFT H200 GPU-hour budget ratio is 2.60**, with lower FANFT budgets at all four recorded targets (Figure 4(i)). OCR gives a **2.00-fold ratio** at both 0.90 and 0.92 (Figure 4(j)). Interpolated evaluation-curve crossings give ratios of 3.07 and 2.22 for GenEval and OCR, respectively (Figure 4(a,c)). GenEval also has higher normalized evaluation AUC over the shared epoch-10–130 interval (Appendix D.6).

5.3 REWARD–COORDINATE ALIGNMENT DEFINES FANFT'S SCOPE

The evidence distinguishes coordinate information from the scope of the resulting update. Frozen probes establish predictive local structure (Figure 2), no-write controls establish its value in learning (Figure 3), and the system comparison reveals reward-dependent gains (Figure 4). A compact coordinate can support useful local choices without representing every reward-improving change. FANFT therefore augments global NFT supervision: the local channel exploits conditional structure, while the base objective learns from the complete scalar reward. HPSv2 and PickScore help define where this complementary use of information is effective.

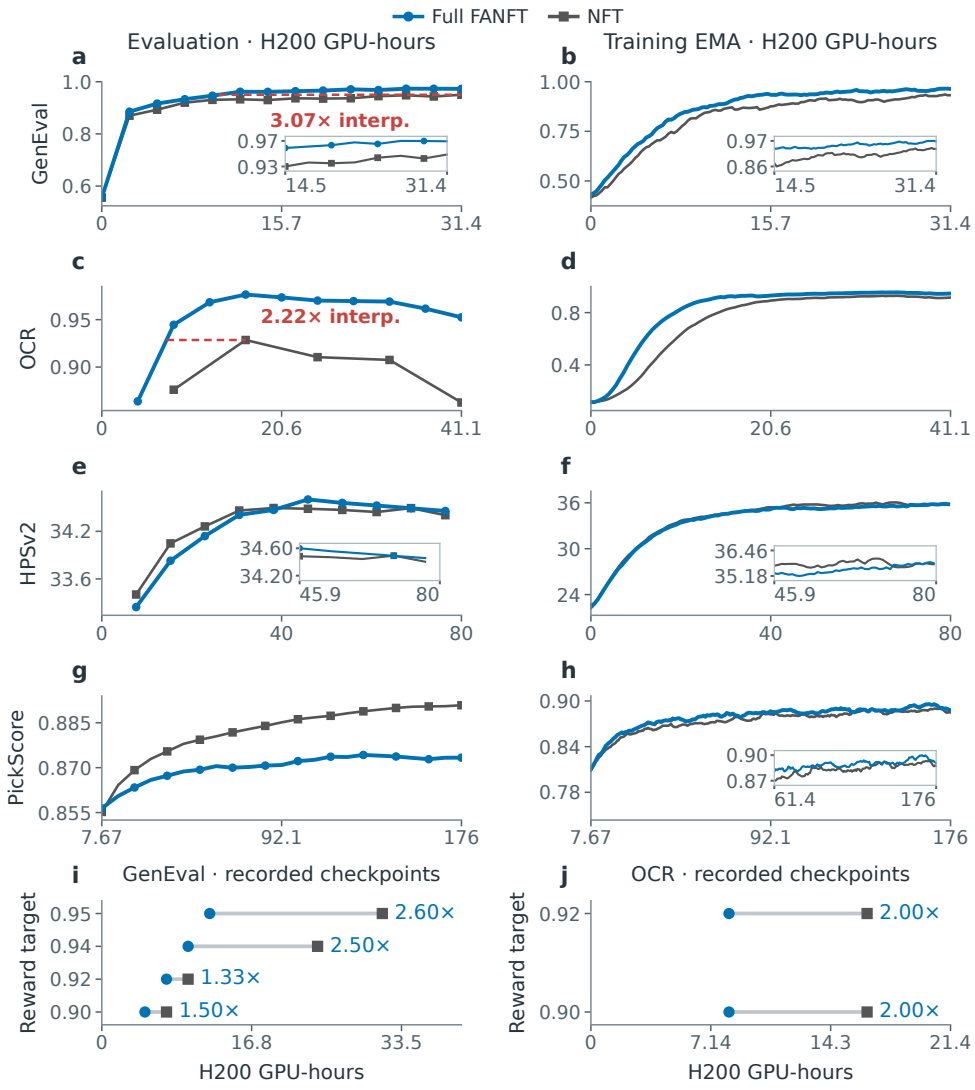

Figure 4: **Full-system learning and checkpoint acquisition in H200 GPU-hours.** (a,c,e,g) Evaluation on GenEval, OCR, HPSv2, and PickScore; (b,d,f,h) collection-reward EMA ($\alpha = 0.1$). All horizontal axes report H200 GPU-hours. Panels (a,c) annotate NFT/FANFT ratios at interpolated evaluation-curve crossings; panels (i,j) report H200 GPU-hour budgets and ratios at the first recorded qualifying checkpoints. Within-task non-method settings are shared (Table 4).

Table 1: **Numerical summary for Figure 4.** Eval entries are FANFT / NFT; ratios are NFT/FANFT.

Task	Eval (FANFT / NFT)	Interp. target / ratio	Recorded target(s) / ratio(s)
GenEval	0.97333 / 0.95008	0.95008 / 3.07×	0.90/0.92/0.94/0.95 / 1.50×/1.33×/2.50×/2.60×
OCR	0.97644 / 0.92855	0.92855 / 2.22×	0.90/0.92 / 2.00×/2.00×
HPSv2	34.60228 / 34.49438	–	–
PickScore	0.87425 / 0.89085	–	–

5.4 Residual anchoring matters beyond subspace dimension

To examine the factors behind the GenEval and OCR gains, Table 2 compares alternative coordinates, global-supervision variants, and scalar guidance. FANFT has the highest AUC on both tasks.

CFG 1D exceeds Random 2D on GenEval (0.91518 vs. 0.90939), while Random 2D is the closest alternative on OCR at 0.95696.

The controls vary how local information is obtained and used. CFG 1D versus Random 2D contrasts residual anchoring with an added dimension; Wrong-prompt 2D changes prompt coupling. Antithetic NFT and DiffusionNFT retain global credit without a local coordinate, while Varying CFG changes collection guidance without a same-state query. These comparisons place the residual-based update alongside extra noise views and scalar steering.

Table 2: **Coordinate choice has task-specific effects.** AUC on GenEval 10–190 and OCR 10–120; bold marks the maximum.

Configuration	GenEval AUC 10–190	OCR AUC 10–120
FANFT	**0.92600**	**0.95804**
Antithetic NFT	0.92170	0.94413
CFG 1D	0.91518	0.94474
Random 2D	0.90939	0.95696
Wrong-prompt 2D	0.91713	0.94054
DiffusionNFT	0.86135	0.92446
Varying CFG	0.52579	0.17965

The configuration study completes the evidence chain. Matched frozen replacements isolate anchor information (Figure 2(d–f)), and no-write controls isolate its use in the update (Figure 3(c,d)). Table 2 then evaluates the resulting learning procedures, connecting coordinate construction to reward response and parameter learning. Antithetic views reuse rewarded images to estimate the global objective, whereas paired interventions create a reward difference tied to a known action. These controls therefore address complementary parts of the learning procedure. Appendix C.3 gives the coordinate definitions and configuration differences.

6 Related Work

Reward-based diffusion post-training builds on denoising and score-based models and on latent or large-scale text-to-image systems (Ho et al., 2020; Song et al., 2021; Nichol & Dhariwal, 2021; Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2024; Esser et al., 2024), and reward or preference optimization (Black et al., 2024; Fan et al., 2023; Liu et al., 2025; He et al., 2026; Rafailov et al., 2023; Wallace et al., 2024; Li et al., 2024; Clark et al., 2024; Prabhudesai et al., 2024; Zheng et al., 2026; Sun et al., 2025). Learned evaluators cover semantic alignment, compositionality, and human preference (Radford et al., 2021; Xu et al., 2023; Wu et al., 2023; Kirstain et al., 2023; Huang et al., 2023; Ghosh et al., 2023). FANFT adds same-state directional credit to their global sample supervision.

Whereas these methods learn from rewards, guidance and controllable-generation methods use model structure at inference (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Hong et al., 2023; Shen et al., 2024; Kynkäänniemi et al., 2024; Sadat et al., 2025; Zhang et al., 2023; Hertz et al., 2023; Mokady et al., 2023; Haas et al., 2024); structured zeroth-order search and diffusion priors provide a complementary optimization view (Salimans et al., 2017; Maheswaranathan et al., 2019; Graikos et al., 2022). Gradient-combination methods (Yu et al., 2020; Liu et al., 2021) and fast-sampling distillation (Salimans & Ho, 2022; Song et al., 2023) address different update or inference objectives. FANFT instead uses the CFG residual as a local reward-query coordinate and lets reward determine the signed response.

7 Conclusion

The Inference–Learning Mismatch shows why improving fixed-model generation and improving parameter learning require different uses of CFG. By separating these roles, we use its residual to define local reward queries and let their measured responses determine the update direction. FANFT demonstrates this principle by combining same-state directional credit with a global NFT gradient. Frozen probes establish the coordinate's predictive information, and no-write controls connect its use to improved learning. The four-reward comparison then places these gains in context: GenEval and OCR benefit most, HPSv2 remains close, and PickScore favors NFT at checkpoint evaluation. These differences motivate retaining global supervision alongside the compact local channel. The resulting contribution is a modular reward-query principle whose value depends on reward–coordinate alignment, with FANFT providing its concrete instantiation on NFT.

AI USE STATEMENT

In preparing this work, we used generative AI tools to assist with writing and editing portions of the manuscript and with preliminary research, including literature discovery, summarization, and organization of relevant prior work. All AI-assisted material was reviewed and revised by the authors; citations and factual claims arising from AI-assisted research were checked against the original sources. We take full responsibility for the final content of this submission, including text and research artifacts produced with the assistance of generative AI.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S. Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2024. doi: 10.1109/FG59268.2024.10581912.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. TempFlow-GRPO: When timing matters for GRPO in flow models. In *International Conference on Learning Representations*, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4264–4273, 2019.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8162–8171, 2021.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9370–9379, 2024.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 32211–32252, 2023.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. DreamSync: Aligning text-to-image generation with image understanding feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5920–5945, 2025. doi: 10.18653/v1/2025.naacl-long.304.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026.

A Complete Training Objective and Update

A.1 Negative-aware global supervision

For each prompt, all independent generations and both branches of every paired intervention enter the same reward group. Let K be the number of terminal images in the group. We use

$$\overline{R}_c = \frac{1}{K}\sum_{j=1}^{K} R_{c,j}, \qquad s_R = \max\{\mathrm{Std}_{c,j}(R_{c,j}), \sigma_R\},$$
$$q_{c,j} = \frac{1}{2} + \frac{\mathrm{clip}((R_{c,j} - \overline{R}_c)/s_R, -C, C)}{2C}. \tag{9}$$

with a positive numerical reward floor σ_R and advantage clip C. Each pair supplies ordinary group-relative labels to the global objective and an antisymmetric reward difference to the local objective.

Given a terminal latent x_0, time t, and training noise $\xi \sim \mathcal{N}(0, I)$, construct the forward interpolation $x_t = (1-t)x_0 + t\xi$. Write $v_\theta = v_\theta(x_t, t, c)$ and $v_{\bar\theta} = v_{\bar\theta}(x_t, t, c)$. The implicit positive and negative velocities used by the negative-aware objective are

$$v_\theta^+ = \beta_{\mathrm{imp}} v_\theta + (1 - \beta_{\mathrm{imp}})\,\mathrm{sg}(v_{\bar\theta}), \tag{10}$$
$$v_\theta^- = (1 + \beta_{\mathrm{imp}})\,\mathrm{sg}(v_{\bar\theta}) - \beta_{\mathrm{imp}} v_\theta, \qquad \widehat{x}_0^{\pm} = x_t - t v_\theta^{\pm}. \tag{11}$$

For reconstruction penalty ℓ, the per-image objective is

$$\mathcal{L}_{\mathrm{NA}} = \frac{C}{\beta_{\mathrm{imp}}}\left[q\,\ell(\widehat{x}_0^+ - x_0) + (1-q)\,\ell(\widehat{x}_0^- - x_0)\right]. \tag{12}$$

The global gradient is taken from the average of this loss and the configured reference regularization,

$$\mathcal{L}_{\mathrm{NFT}} = \mathbb{E}[\mathcal{L}_{\mathrm{NA}}] + \lambda_{\mathrm{ref}}\mathbb{E}\|v_\theta(x_t, t, c) - v_{\mathrm{ref}}(x_t, t, c)\|^2, \qquad g_N = \nabla_\theta \mathcal{L}_{\mathrm{NFT}}. \tag{13}$$

Antithetic training views use $(\xi, -\xi)$ to estimate the forward-process loss. Each view has the prescribed marginal noise distribution. These views reuse the same rewarded terminal image and therefore do not double the rollout-image or reward-query budget. The global objective follows the forward-process negative-aware formulation of Zheng et al. (2026).

A.2 Action basis and guidance-cone radius

All direction construction uses behavior-model predictions and reward-independent, deterministically sampled candidate vectors. The first retained axis is the normalized projection of the CFG residual onto $v_b^\perp$. Additional candidates are orthogonalized against v_b and the previously accepted axes. A rank threshold handles degeneracy explicitly. The reported full configuration retains two dimensions, with four paired base states per prompt and one symmetric pair per base state.

The scalar guidance coordinate of an action is

$$\gamma(v) = \frac{\langle v - v_u, r\rangle}{\|r\|^2}, \qquad \gamma_0 = \gamma(v_b). \tag{14}$$

For a unit direction u, the symmetric pair has coordinates

$$\gamma(v_b \pm \sigma u) = \gamma_0 \pm \sigma\frac{\langle u, r\rangle}{\|r\|^2}. \tag{15}$$

Assuming the base field is inside the declared cone, $\gamma_0 \geq \gamma_{\min}$, both branches satisfy $\gamma \geq \gamma_{\min}$ whenever

$$\sigma \leq \sigma^{\mathrm{cone}} = \frac{(\gamma_0 - \gamma_{\min})\|r\|^2}{|\langle u, r\rangle|}. \tag{16}$$

When the denominator is zero, the direction imposes no cone-radius bound. Numerical implementations use stabilizers; an inadmissible base field or a degenerate direction supplies no local correction. The applied radius is the smaller of the requested radius and the cone bound.

FANFT uses a base velocity-ratio radius of 0.1, radius multipliers $(1, 1, 2, 2)$ across the four paired states, a cone floor of 0.5, and trust ratio $\rho = 0.1$. The action is applied to the first velocity prediction, with the initial latent shared by the paired branches. The frozen mechanism study uses a separate fixed-radius measurement protocol: fixed radius $0.1\|v_c\|$ for every probe, with no training radius schedule.

A.3 Parameter pullback and fusion

In Eq. (6), behavior states, basis vectors, actions, radii, and measured rewards are stop-gradient quantities. The derivative $\nabla_\theta p_i(\theta)$ pulls the first-action measurement into the current velocity model. It is a local action surrogate; the clean-sample objective continues to supervise the completed image. The behavior model follows the configured responsive averaging schedule, written generally as $\bar{\theta}_{k+1} = d_k\bar{\theta}_k + (1-d_k)\theta_{k+1}$.

The reliability coefficient uses disjoint, balanced collections of probe pairs. An idealized model $g_A = s + \epsilon_A$, $g_B = s + \epsilon_B$ with independent zero-mean fold noise gives

$$\mathbb{E}\langle g_A, g_B\rangle = \|s\|^2, \qquad \mathbb{E}\left[\tfrac{1}{4}\|g_A - g_B\|^2\right] = \tfrac{1}{4}\big(\mathbb{E}\|\epsilon_A\|^2 + \mathbb{E}\|\epsilon_B\|^2\big). \tag{17}$$

This motivates common-signal versus disagreement weighting. The coefficient κ is an admission weight computed from shared gradient signal and cross-fold disagreement, including variation across prompts and states.

Proposition 1 (Controlled gradient rotation)**.** *For nonzero g_N, ignore numerical stabilizers and let $h = \alpha_F g_F^\perp$ in Eq.* (8)*. Then*

$$\langle h, g_N\rangle = 0, \qquad \|h\| \le \rho\kappa\|g_N\|, \qquad \|g_{\text{final}}\| = \|g_N\|, \qquad \cos(g_{\text{final}}, g_N) \ge \frac{1}{\sqrt{1+(\rho\kappa)^2}}. \tag{18}$$

Proof. Orthogonal projection gives the first identity and the definition of α_F gives the norm bound. Since $\|g_N + h\|^2 = \|g_N\|^2 + \|h\|^2$, norm matching gives the third identity and the cosine equals $[1 + \|h\|^2/\|g_N\|^2]^{-1/2}$. □

Writing $q = \|h\|/\|g_N\|$, the proof gives the exact expression

$$g_{\text{final}} = \frac{g_N + h}{\sqrt{1+q^2}}, \qquad \frac{\|g_{\text{final}} - g_N\|^2}{\|g_N\|^2} = 2\left(1 - \frac{1}{\sqrt{1+q^2}}\right). \tag{19}$$

Hence $g_{\text{final}} = g_N + h - \frac{1}{2}q^2 g_N + O(q^3\|g_N\|)$: the complementary component enters at first order, while its normalization changes the global component at second order. If an unpreconditioned step is $\Delta\theta = -\eta g_{\text{final}}$, the first-order NFT loss change is

$$g_N^\top \Delta\theta = -\eta\frac{\|g_N\|^2}{\sqrt{1+q^2}}, \qquad (g_F^\perp)^\top\Delta\theta = -\eta\frac{\alpha_F\|g_F^\perp\|^2}{\sqrt{1+q^2}}. \tag{20}$$

These identities characterize the two descent components of the norm-matched gradient supplied to the optimizer. They show how a first-order complementary component is combined with a second-order change to the global component. Adam subsequently applies coordinate-wise preconditioning and moment updates; checkpoint rewards evaluate the resulting training behavior.

B Geometric Interpretation and Local Estimation

B.1 State-indexed affine action spaces

For a fixed prompt, latent, and solver state, let U be the retained orthonormal basis in Eq. (2). FANFT measures actions in

$$\mathcal{T}_{y,t,c} = v_b^\perp, \qquad \mathcal{A}_{y,t,c} = \{v_b + Ua : a \in \mathbb{R}^m\} \subset v_b + \mathcal{T}_{y,t,c}. \tag{21}$$

The implemented probes occupy a bounded neighborhood in this affine space, with radius set by the velocity scale and guidance cone. The base state indexes the local space; the reward response supplies its directional preference. This is the operational meaning of the fiber-augmented construction.

The transverse constraint removes the component parallel to the current behavior velocity. For a unit action $u \perp v_b$, $\|v_b + \sigma u\|^2 = \|v_b\|^2 + \sigma^2$, so the action changes the direction at first order while changing the speed at second order. The behavior velocity and the orthonormal action basis fully specify this state-indexed velocity chart. Matched interventions in the chart provide the reward contrasts used to evaluate its predictive structure.

The first-velocity intervention, radius, basis, and recipient state fix the query before reward evaluation. The following analysis describes measurement in this chart.

Algorithm 1 FANFT: global sample credit with local directional credit

Require: Current model θ, behavior model $\bar{\theta}$, prompt batch, reward evaluator, image budget, trust ratio ρ.

for each prompt c **do**
  Allocate the image budget to independent images and same-state symmetric pairs.
  for each paired base state i **do**
    Compute v_c, v_u and form the residual-anchored transverse basis U_i.
    Select a balanced direction u_i and compute the admissible radius σ_i.
    Perturb the first velocity by $\pm\sigma_i u_i$; complete both branches with the behavior model.
  end for
  Score all terminal images and assign their ordinary prompt-group labels.
end for
Compute the global negative-aware gradient g_N from the complete image collection.
Compute paired finite differences and the two local fold gradients g_A, g_B.
Compute κ, project the local gradient orthogonally to g_N, and apply the trust cap.
Norm-match the fused gradient; use g_N when no admissible local correction is available.
Apply the configured optimizer and update the behavior model.

B.2 Random-subspace benchmark

Proposition 2 (Isotropic exploration benchmark)**.** *Let $\mathcal{T} \subseteq \mathbb{R}^d$ be an n-dimensional action working space, $g \in \mathcal{T}$ a local reward gradient, and U a uniformly distributed orthonormal m-frame in $\mathcal{T}$. Then*

$$\mathbb{E}\|U^\top g\|^2 = \frac{m}{n}\|g\|^2. \tag{22}$$

For a fixed orthonormal frame U_F and a direction a uniform on the unit sphere in $\mathbb{R}^m$, the mean squared first-order directional response is $\|U_F^\top g\|^2/m$.

Proof of Proposition 2. In an orthonormal coordinate system for $\mathcal{T}$, rotational invariance implies $\mathbb{E}[UU^\top] = \lambda I_n$. Taking traces gives $m = \lambda n$, hence $\lambda = m/n$ and

$$\mathbb{E}\|U^\top g\|^2 = g^\top \mathbb{E}[UU^\top] g = \frac{m}{n}\|g\|^2. \tag{23}$$

For a unit coordinate a uniform on the sphere, $\mathbb{E}[aa^\top] = I_m/m$, so $\mathbb{E}[(g^\top U_F a)^2] = \|U_F^\top g\|^2/m$. □

The frozen random-replacement control is more specific than the isotropic benchmark. It shares one auxiliary axis with the CFG space and excludes the residual from its replacement axis. Its purpose is to exchange the residual coordinate while holding the dimension, shared direction, state, and perturbation norm fixed. The isotropic benchmark quantifies dimension-dependent gradient-energy capture; the shared-axis experiment quantifies the reward-prediction advantage of the residual anchor under these controls.

B.3 Symmetric finite differences

Suppose the third directional derivative of Q is bounded in magnitude by M along $[-\sigma u, \sigma u]$, with $\|u\| = 1$. Taylor expansion gives

$$\widehat{d}(u) = \frac{Q(\sigma u) - Q(-\sigma u)}{2\sigma} = \langle \nabla Q(0), u\rangle + e(u), \qquad |e(u)| \le \frac{M\sigma^2}{6}. \tag{24}$$

For $u = Ua$ with a uniform on the unit sphere in $\mathbb{R}^m$,

$$\mathbb{E}[m\widehat{d}(u)u] = UU^\top \nabla Q(0) + b_\sigma, \qquad \|b_\sigma\| \le \frac{mM\sigma^2}{6}. \tag{25}$$

A deterministic design with $P^{-1}\sum_{i=1}^P a_i a_i^\top = I_m/m$ gives the same first-order projection for a common local linear response. When training probes come from different base states, this identity motivates the local estimator at each state; fold agreement measures their compatibility after parameter pullback. For quantized evaluators, the branch preference itself remains defined even when differentiability is unavailable.

B.4 Analytical roles of the design parameters

Training uses radius scale 0.1, trust ratio $\rho = 0.1$, the first velocity evaluation, four paired states per prompt, and a two-dimensional basis. Frozen coordinate identification uses a fixed radius; training applies the guidance-cone bound.

The estimator gives separate roles to these choices. Under the smoothness assumption above, central-difference bias is $O(\sigma^2)$. If the two rewards have additive noise with variances s_+^2, s_-^2 and covariance c_{+-}, then

$$\mathrm{Var}(\widehat{d}) = \frac{s_+^2 + s_-^2 - 2c_{+-}}{4\sigma^2}. \tag{26}$$

Under this additive-noise model, variance scales with $1/\sigma^2$, while central-difference bias scales with σ^2. The radius controls this bias–variance balance. Shared states control nuisance variation; nonzero radii define finite responses for discrete rewards.

The trust ratio instead bounds rotation by $\varphi \le \arctan(\rho\kappa)$; zero admission recovers global NFT on the paired collection. More pairs allocate more images to measurement, while a larger basis spreads measurements across more coefficients. Table 2 compares specified constructions; Proposition 2 isolates dimensional dependence. These relations identify how radius, trust ratio, and pair allocation control the local estimator and its contribution to the update.

C Experimental Setup

C.1 Model, rewards, and evaluation scales

Experiments use Stable Diffusion 3.5 Medium: training updates LoRA adapters, and frozen evaluation uses the unadapted model. Collection and frozen evaluation use deterministic DPM2 with 10 steps at 512×512 resolution. Trained models are evaluated with 40 steps at CFG 4.5, except in the guidance sweep, which explicitly varies the evaluation scale.

Training is evaluated using GenEval, OCR, HPSv2, and PickScore. The frozen study uses both dense GenEval rewards and strict binary correctness from the same detector and evaluation pipeline, allowing us to test sensitivity to scoring granularity. PickScore provides an independent scoring function on the same images. Frozen PickScore results use the raw score, whereas training comparisons use normalized PickScore; these scales are analyzed separately. HPSv2 is reported as the underlying score multiplied by 100. These readouts test local-response predictions, while checkpoint evaluation measures reward-aligned adaptation.

Evaluation rewards are indexed by the training epoch of the evaluated model. Collection-reward curves use exponential smoothing, $s_t = 0.1r_t + 0.9s_{t-1}$, initialized at the first reward. Evaluation curves remain unsmoothed. Learning-curve AUC, interval-end rewards, and first qualifying checkpoints are computed from these evaluation rewards; interpolated threshold crossings are defined separately in Appendix D.6.

Guidance and credit-channel schedules. The collection-policy comparison uses rollout CFG 1 or 4.5 with remaining training settings matched, learning rate 3×10^{-4}, and evaluation at epochs $0, 10, \ldots, 200$. The five-scale collection sweep uses epochs 0–80 and fixed evaluation CFG 4.5. Credit-channel configurations generate 576 images and perform one optimizer update per epoch, reusing each image under two antithetic forward-process noise views. Intervention-free evaluation is reported every ten epochs: 19 evaluations over epochs 10–190 for GenEval and 12 over epochs 10–120 for OCR (Figure 3; Table 2). AUC divides the trapezoidal integral by 180 for GenEval and 110 for OCR. Open markers in Figure 3(c,d) report the GenEval epoch-190 and OCR epoch-120 rewards, respectively.

C.2 Frozen confirmation protocol

The frozen evaluation comprises 144 prompt–latent states from 48 prompts balanced across six GenEval categories. Discovery and validation directions are determined by the prompt and initial latent before reward evaluation.

For each state, let $b = v_c/\|v_c\|$ and let e_1 be the normalized residual after projection against b. A deterministically sampled Gaussian candidate is orthogonalized against b and e_1 to form a shared axis e_2. A second candidate is orthogonalized against b, e_1, e_2 to form e_0. The compared bases are

$$U_{\mathrm{CFG}} = [e_1, e_2], \qquad U_{\mathrm{random}} = [e_0, e_2]. \tag{27}$$

Each basis is orthonormal and transverse to the conditional behavior velocity. The maximum relative action-norm error is 1.19×10^{-7}.

In each space, discovery folds A and B each contain two orthogonal coordinate directions, with an independently determined rotation for each fold. The held-out fold C contains four additional coordinate directions. For normalized radius $r_0 = 0.1$, the actual velocity action is $r_0\|v_c\|Ua$. A discovery fold fits a two-dimensional coefficient vector through

$$A_f \widehat{b}_f = \frac{\boldsymbol{R}_f^+ - \boldsymbol{R}_f^-}{2r_0}, \qquad \widehat{b} = \frac{\widehat{b}_A + \widehat{b}_B}{2}. \tag{28}$$

The use of r_0 defines coefficients in normalized velocity-action coordinates. A C-direction prediction is $\widehat{s} = \mathrm{sign}(\widehat{b}^\top a_C)$, with numerical zero treated as abstention. No held-out reward enters basis construction, fitting, radius choice, or discovery confidence.

The discovery confidence uses Eq. (7) with the two *action-response coefficient vectors* in place of parameter gradients. A gated clean-relative outcome equals the selected branch reward minus the clean reward when confidence is positive, and zero otherwise. A confidence-weighted outcome multiplies that selected-branch gain by the confidence. Their denominators retain all recipient states and held-out directions.

For the other-state control, donor coefficients are averaged over the other latent states of the same prompt. For the other-prompt control, a fixed cyclic permutation assigns a different donor prompt while keeping the initial latent fixed. These coefficients are interpreted in the recipient's coordinate convention. The residual basis changes with the state; the mean cosine of its first axis across latent states is 0.598. Accordingly, donor transfer tests the combined coordinate–response mapping rather than transfer of an unchanged ambient vector.

Uncertainty. Each prompt contributes the average of its 12 held-out C-direction outcomes. Means and paired differences are bootstrapped over the 48 prompts, preserving all measurements inside a sampled prompt. We use 2,000 resamples, the 2.5th and 97.5th percentiles. The same held-out outcomes are evaluated across coordinate spaces, coefficient-transfer conditions, and scoring functions. Uncertainty is therefore determined by the 48 prompt clusters, with latent states and directions nested within each cluster. These prompt-cluster intervals quantify frozen-response variability across prompts. Normalized AUC and interval-end rewards summarize observed checkpoint rewards, and empirical interquartile ranges summarize the distribution of per-update measurements.

C.3 Coordinate-choice and credit-channel configurations

Table 3 summarizes the coordinate choices and credit-channel ablations. Each uses 24 prompt groups, 576 generated images per epoch, learning rate 3×10^{-4}, and a 24-image prompt group. The full design reserves eight images for four symmetric pairs and keeps global supervision for the complete collection. Full FANFT, CFG 1D, and the 2D replacement configurations use two antithetic forward-process views; antithetic NFT also uses two views, while ordinary NFT uses one. Both studies use the same FANFT reference in each task.

FANFT perturbs the first velocity symmetrically at each shared state and estimates directional responses from paired reward differences. Balanced deterministic folds provide the gradient-agreement estimate. Global NFT advantages, orthogonal local projection, and norm matching follow Appendix A, with trust ratio $\rho = 0.1$. The two-task ablations share the same NFT and behavior-averaging settings. The full-system OCR comparison uses its own training configuration and 1,152 images per epoch; it is analyzed separately from these 576-image ablations.

Wrong-prompt 2D constructs the residual from a cyclically shifted batch prompt at the recipient state, projects it against the original behavior velocity, and retains the original reward and training condition. Random 2D uses random exploration directions and disables CFG-based exploration.

Table 3: **Training configurations and local-credit ablations.** Views $2\pm$ denotes antithetic forward-process noise views of the same image. Pairs indicates symmetric rollout probes; Estimate and Update indicate whether the local gradient is computed and used. All configurations retain global NFT supervision.

Method	Coordinates	Views	Pairs	Estimate	Update
NFT	—	1	No	No	No
Antithetic NFT	—	$2\pm$	No	No	No
Varying CFG	—	1	No	No	No
FANFT	Residual-anchored	$2\pm$	Yes	Yes	Yes
CFG 1D	CFG 1D	$2\pm$	Yes	Yes	Yes
Random 2D	Random 2D	$2\pm$	Yes	Yes	Yes
Wrong-prompt 2D	Donor residual	$2\pm$	Yes	Yes	Yes
Measurement only	Residual-anchored	$2\pm$	Yes	Yes	No
Paired only	Residual-anchored	$2\pm$	Yes	No	No

These training comparisons test the resulting configurations, while the frozen shared-axis experiment isolates residual-anchor information with dimension and action norm fixed.

The credit-channel ablations preserve paired collection, reward observations, and global learning. Measurement-only computes the local gradient with zero update weight; paired-collection-only omits its estimation. The comparison uses the GenEval epoch-10–190 grid and the OCR epoch-10–120 grid (Section 4.3; Table 10).

C.4 Matched full-system training protocol

NFT and FANFT share the within-task protocol in Table 4; FANFT adds the local-credit construction of Section 4 to global sample supervision.

Parameter and behavior-model averaging use shared rules within each comparison. Display smoothing applies to collection rewards. Global and local gradients are fused before each optimizer step.

Within-task matching applies separately to the system comparison and the 576-image credit-channel study. Appendix D.6 gives the GenEval and OCR acquisition budgets.

Table 4: **Shared protocol for within-task NFT/FANFT comparisons.** Method-specific collection and credit components are defined separately in Section 4. The reward changes across tasks; non-method training conditions are shared within each comparison.

Training component	Shared within each task
Model and data	SD3.5 Medium, adapter parameterization, prompt distribution
Generation budget	Equal number of terminal images per collection round
Collection	512×512; 10-step DPM2; rollout CFG 1
Optimization	Learning rate 3×10^{-4}; shared optimizer and update schedule
Parameter averaging	Same averaging rule and schedule for both methods
Behavior model	Same averaging rule for collection
Evaluation	Same prompts and checkpoints; 40 steps; CFG 4.5
Training outputs	All terminal images retain global NFT supervision

D Additional Experiments

D.1 Guidance role separation across collection and evaluation scales

Figure 1 presents all four rewards under three distinct protocols. Tables 5 and 6 retain the exact fixed-policy scores and AUCs. The long-window study compares two collection policies, evaluating checkpoints at inference CFG 1 and 4.5 over epochs 0, 10, . . ., 200. The five-scale study evaluates nine checkpoints over epochs 0–80.

Table 5: **Exact guidance-role summaries.** Fixed-policy scores and training AUCs are distinct protocols. Training AUC spans epochs 0–200 at evaluation CFG 4.5; reported values average the learning trajectories.

	Fixed-policy inference		Training AUC, eval CFG 4.5	
Reward	CFG 1	CFG 4.5	Rollout CFG 1	Rollout CFG 4.5
GenEval	0.5545	0.7941	0.8869	0.8121
OCR	0.1473	0.5866	0.8127	0.2174
HPSv2	21.3090	29.8432	33.8435	13.8428
PickScore	0.7707	0.8462	0.8722	0.7851

Table 6: **Scalar operating-point AUCs.** Epochs 0–80, evaluation CFG 4.5; native reward units.

Reward	CFG 1	CFG 2	CFG 3	CFG 4.5	CFG 7
GenEval	0.7886	0.8833	0.8621	0.7569	0.3379
OCR	0.9048	0.8163	0.5690	0.2170	0.1760
HPSv2	33.8529	34.3625	31.8837	19.1958	15.3974
PickScore	0.8673	0.8805	0.8706	0.8553	0.7139

Table 7: **Integrated reward under both evaluation protocols.** Each row pairs collection-policy trajectories with the remaining training settings matched. Each value integrates 21 raw evaluations over epochs 0–200 and divides by 200.

	Inference CFG 1		Inference CFG 4.5	
Reward	Rollout 1	Rollout 4.5	Rollout 1	Rollout 4.5
GenEval	0.8719	0.2761	0.8570	0.8169
	0.8802	0.2104	0.9086	0.7836
	0.8761	0.4251	0.8951	0.8357
OCR	0.9080	0.0056	0.8199	0.1253
	0.6909	0.0426	0.6919	0.1974
	0.9377	0.0347	0.9265	0.3294
HPSv2	34.8803	11.3248	34.2917	13.9672
	35.4322	9.5532	33.7222	13.4531
	35.5384	12.5588	33.5166	14.1082
PickScore	0.8859	0.7548	0.8734	0.8633
	0.8831	0.7315	0.8710	0.7451
	0.8845	0.7155	0.8721	0.7467

Each rollout-guidance comparison matches prompts and the remaining training settings. The upper panels separate evaluation at CFG 1 and 4.5 (Figure 5(a,b), Figure 6(a,b), Figure 7(a,b), and Figure 8(a,b)); the lower panels report the five-scale sweeps under those respective protocols. Together, these sweeps show that scalar collection guidance does not define a single task-independent training policy.

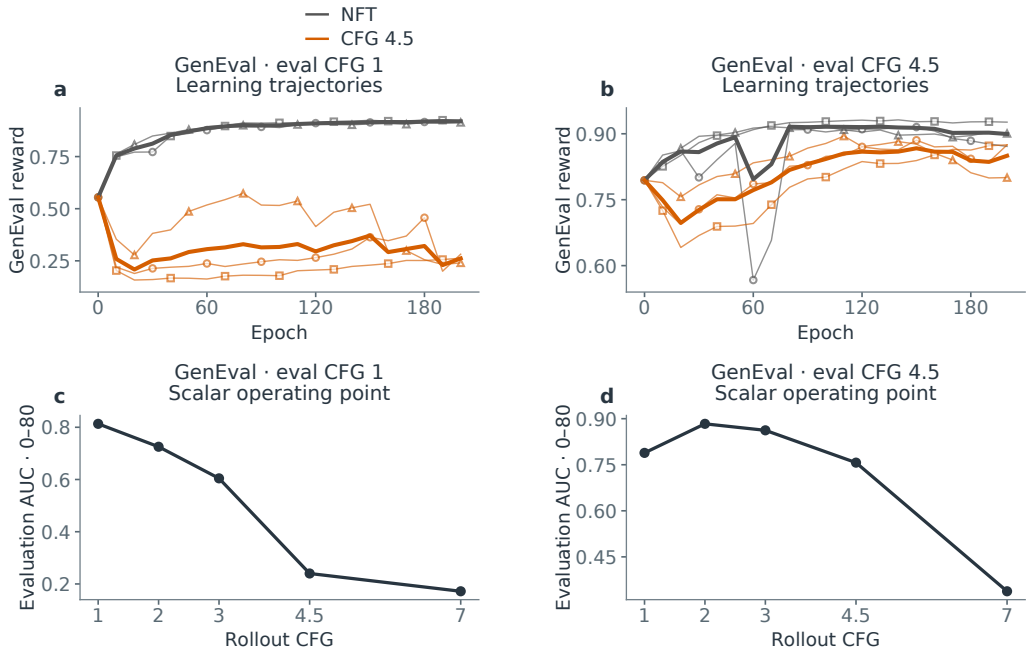

Figure 5: **GenEval: separate collection policy from evaluation guidance.** Top: raw evaluations of rollout CFG 1 (NFT) and 4.5 at each evaluation scale; thin/thick curves show individual trajectories/their means. Bottom: five-scale AUC over epochs 0–80 under the corresponding evaluation protocol.

Figure 9 characterizes rewarded collection with EMA coefficient 0.1; the raw task-specific checkpoint evaluations above measure learning.

D.2 Frozen scoring robustness, ties, and coverage

Figure 2(d–f), Anchor, shows positive paired coordinate advantages under all three scoring readouts; Table 8 gives their exact values. Locality also transfers across scoring functions: the raw-PickScore differences against other-state and other-prompt coefficients are 0.094 [0.053, 0.141] and 0.168 [0.122, 0.218] (Figure 2(f), State/Prompt).

Table 8: **Three scoring readouts for the frozen comparison.** Marginal held-out advantages and within-prompt differences with 95% prompt-bootstrap intervals. GenEval uses percentage points; PickScore uses its raw analysis scale.

Scorer	CFG anchor	Random replacement	Paired difference
GenEval dense (pp)	6.02 [4.67, 7.40]	1.52 [0.88, 2.21]	4.50 [3.21, 5.93]
GenEval strict (pp)	6.08 [3.91, 8.25]	1.56 [0.52, 2.78]	4.51 [2.60, 6.51]
PickScore (raw)	0.271 [0.225, 0.318]	0.031 [0.024, 0.039]	0.239 [0.195, 0.286]

Dense GenEval produces fewer reward ties than the strict readout in both coordinate spaces (Table 9, Tied/all). CFG has 189 eligible dense-GenEval pairs, compared with 55 for the random replacement (Figure 11(d), Local; Table 9). Conditional correctness and informative coverage are therefore reported together, while branch advantage retains ties and abstentions in its denominator.

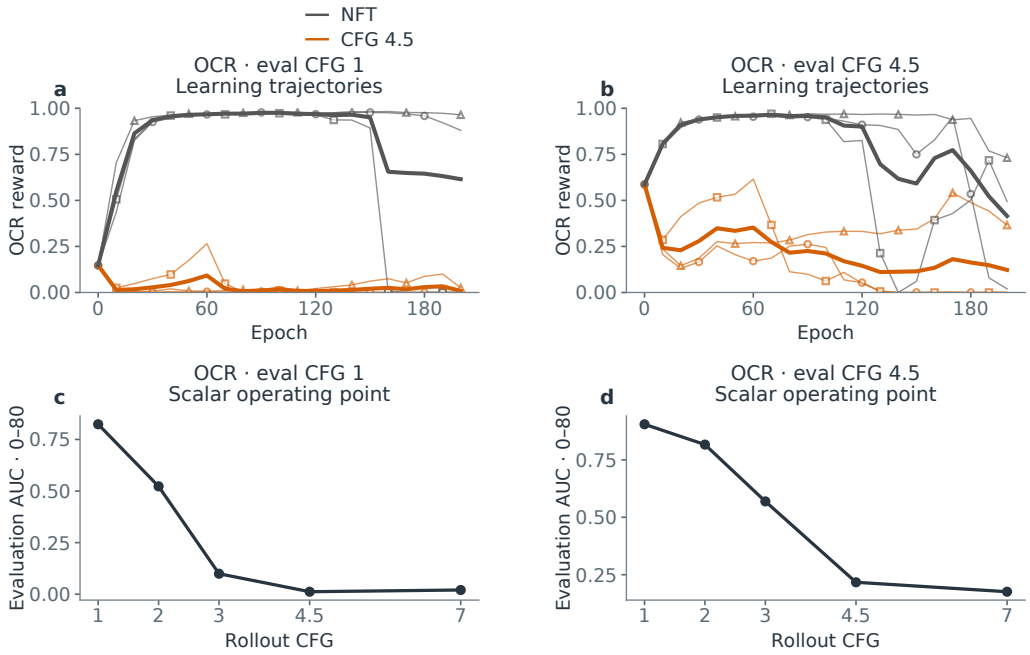

Figure 6: **OCR: separate collection policy from evaluation guidance.** Top: raw evaluations of rollout CFG 1 (NFT) and 4.5 at each evaluation scale; thin/thick curves show individual trajectories/their means. Bottom: five-scale AUC over epochs 0–80 under the corresponding evaluation protocol.

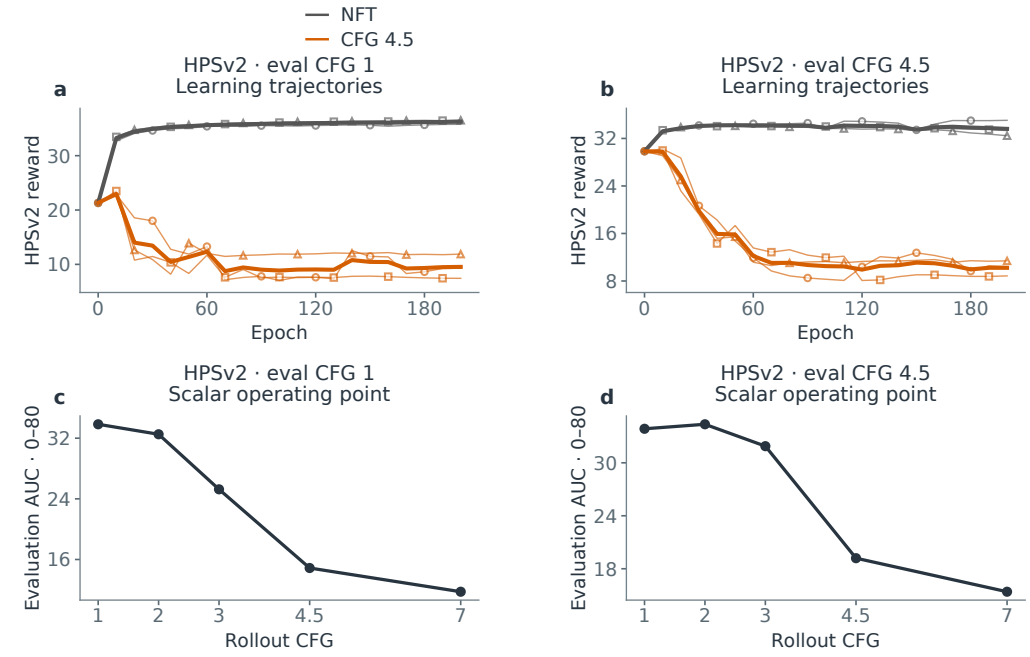

Figure 7: **HPSv2: separate collection policy from evaluation guidance.** Top: raw evaluations of rollout CFG 1 (NFT) and 4.5 at each evaluation scale; thin/thick curves show individual trajectories/their means. Bottom: five-scale AUC over epochs 0–80 under the corresponding evaluation protocol.

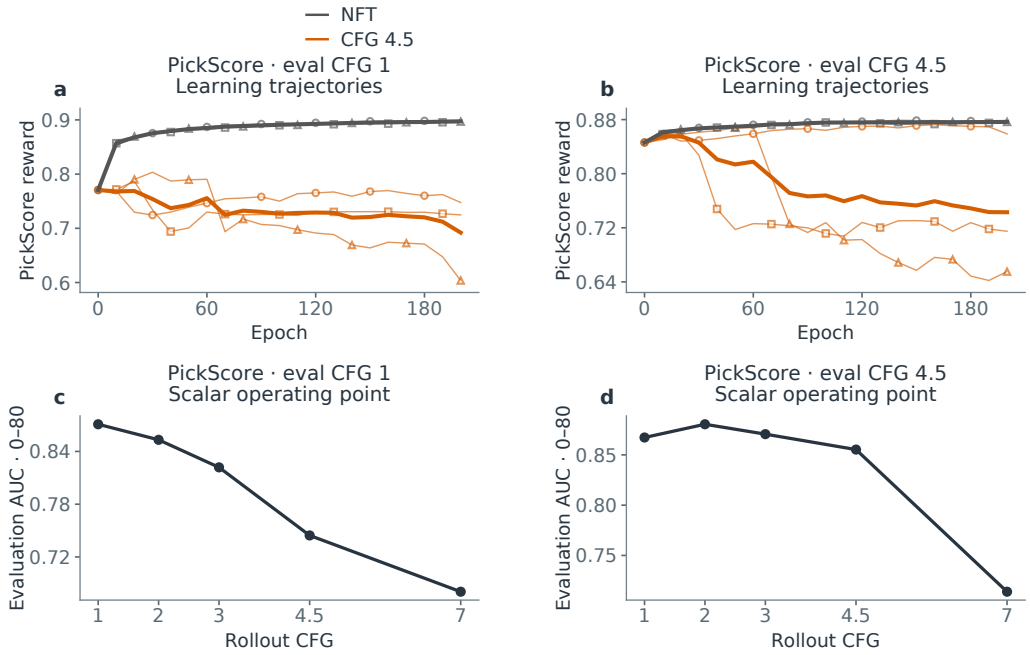

Figure 8: **PickScore: separate collection policy from evaluation guidance.** Top: raw evaluations of rollout CFG 1 (NFT) and 4.5 at each evaluation scale; thin/thick curves show individual trajectories/their means. Bottom: five-scale AUC over epochs 0–80 under the corresponding evaluation protocol.

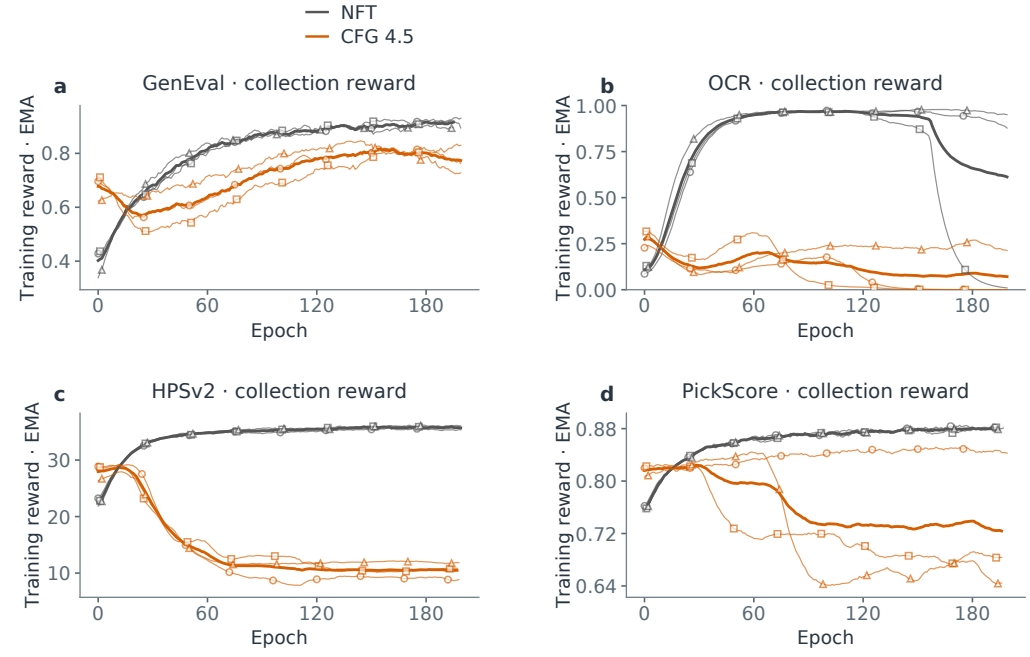

Figure 9: **Collection reward across four tasks.** Thin lines show individual EMA-0.1 reward trajectories; thick lines average shared observed epochs. HPSv2 follows its $\times 100$ display convention. Collection reward complements the checkpoint evaluations in Figure 1.

Table 9: **Sign accuracy with its coverage.** Each local predictor has 576 held-out pairs (144 prompt–latent states × 4 directions). Eligibility requires a non-tied reward and a non-abstained sign prediction. The main branch-advantage statistic retains all pairs, including ties and abstentions.

Scorer	Space	Tied / all	Correct / eligible	Accuracy
GenEval dense	CFG	385/576	174/189	92.1%
GenEval dense	Random	511/576	50/55	90.9%
GenEval strict	CFG	496/576	73/76	96.1%
GenEval strict	Random	554/576	19/20	95.0%
PickScore (raw)	CFG	0/576	493/576	85.6%
PickScore (raw)	Random	0/576	432/576	75.0%

The dense-GenEval gated result in Figure 2(b) extends to positive local clean-relative gains across all three readouts (Figure 10(a–c), Local); confidence weighting gives a complementary use of the same measurements (Figure 10(d–f), Local). Both retain every held-out direction in their denominators. Conditional sign accuracy (Figure 11(a–c)) is paired with eligible counts (Figure 11(d–f)). Fixed-bin accuracy and coverage are shown separately in Figure 12(a–c) and Figure 12(d–f); empty-bin accuracies remain undefined. The same held-out frozen outcomes support complementary measurements of branch choice, clean-relative gain, sign accuracy, and informative coverage.

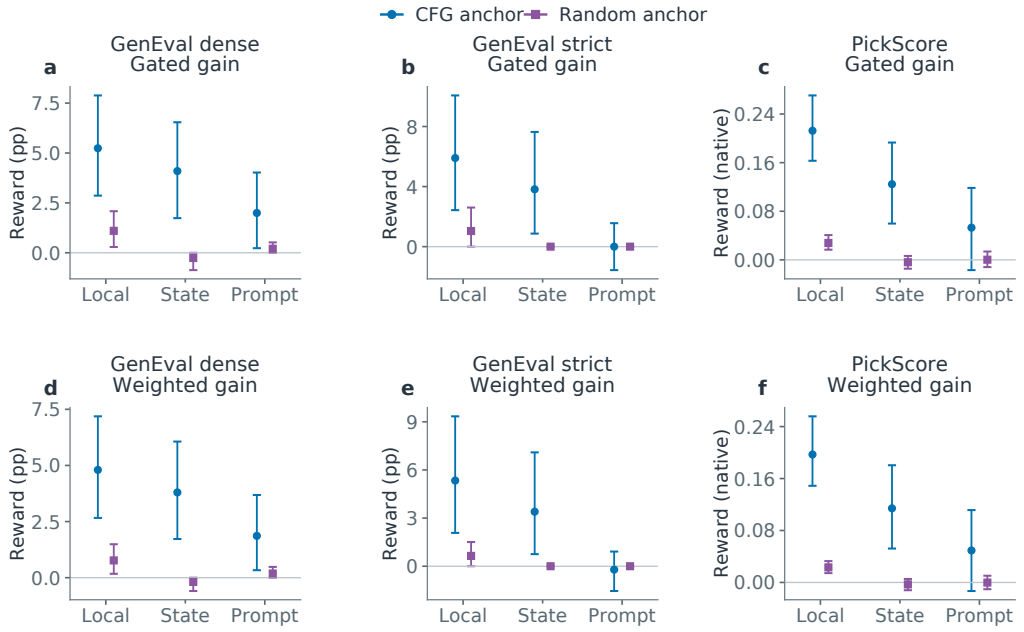

Figure 10: **Selection and weighting of the same held-out response.** Top: clean-relative gain when discovery confidence is positive, retaining the clean reference otherwise. Bottom: confidence-weighted clean-relative gain. Both retain all states and directions. Columns use dense GenEval, strict GenEval, and raw PickScore; bars are 95% prompt-bootstrap intervals. Local/State/Prompt denote the fitted or transferred predictor.

D.3 Credit-channel and coordinate comparisons

Table 10 reports the AUC and interval-end markers in Figure 3(c,d), using the same FANFT reference as Table 2. For checkpoint scores R_i at ordered evaluation epochs t_i, our interval-average statistic is

$$\mathrm{AUC}_{[t_0,t_m]} = \frac{1}{t_m - t_0} \sum_{i=1}^{m} \frac{R_{i-1} + R_i}{2} (t_i - t_{i-1}).$$

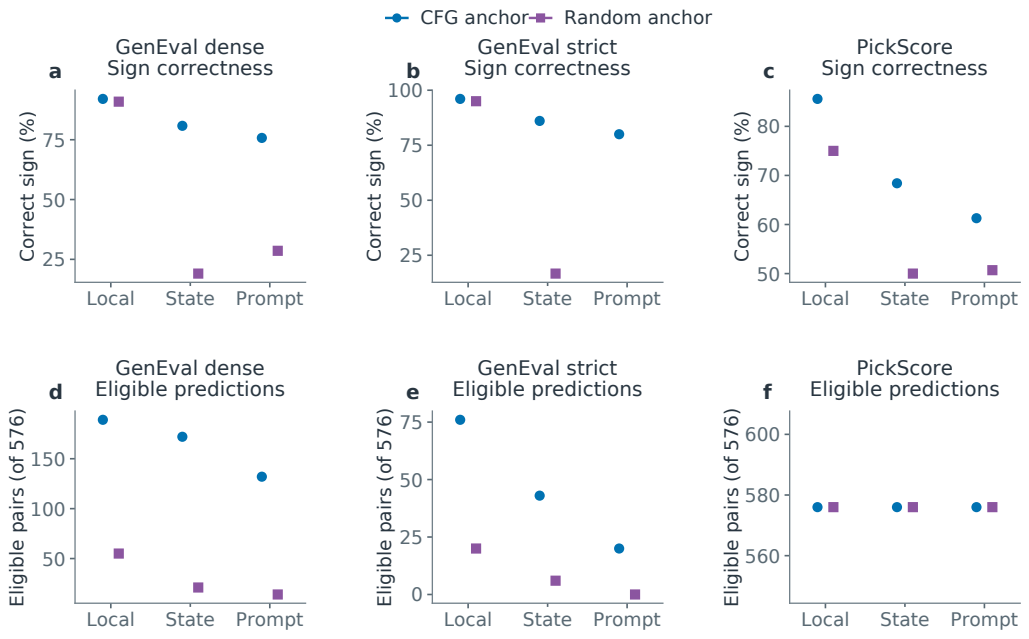

Figure 11: **Directional accuracy with its eligible sample count.** Top: sign accuracy among non-tied rewards with a non-abstained prediction. Bottom: eligible pairs out of 576 per condition. Empty eligible sets have no reported accuracy. The prompt is the sampling unit, with all within-prompt measurements retained as a cluster.

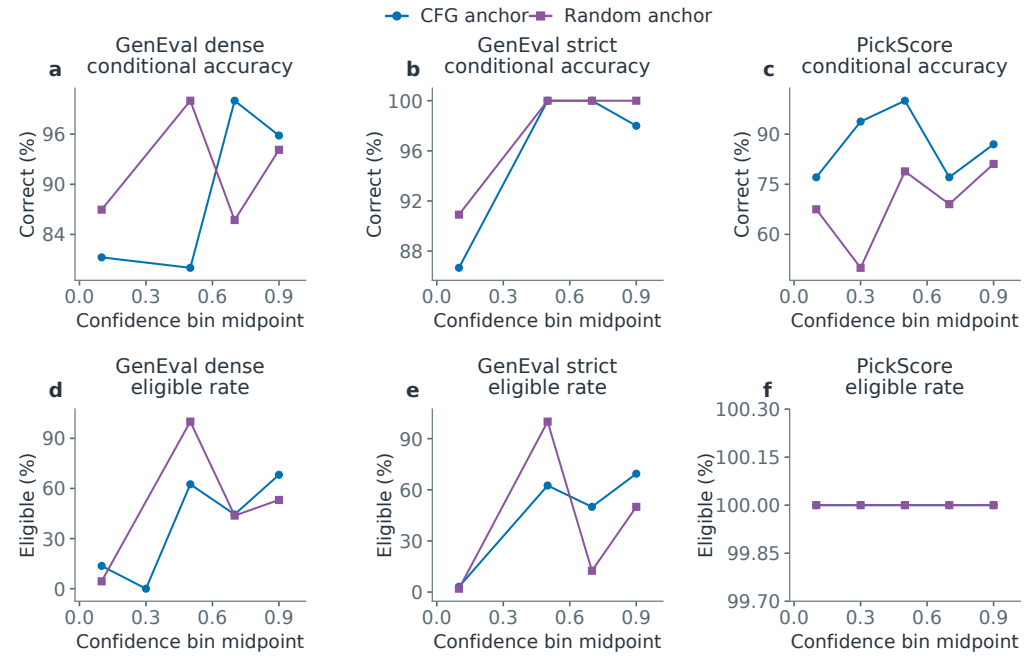

Figure 12: **Confidence bins separate conditional accuracy from coverage.** Top: eligible-case sign accuracy by fixed discovery-confidence bin. Bottom: eligible coverage in the same confidence bins. Each column uses one readout, and CFG/random use the same bins. These panels relate discovery confidence to held-out sign accuracy and eligible coverage; zero-denominator accuracy is left undefined.

It retains the input score scale and weights adjacent evaluations by their epoch spacing. The separate interval-end statistic is R_m. For OCR, these statistics summarize epochs 10–120 and the reward at epoch 120, respectively (Table 10).

Table 10: **Credit-channel AUC and interval-end reward.** Window-normalized AUC and the last checkpoint reward, corresponding to Figure 3(c,d). Intervals: epochs 10–190 for GenEval and 10–120 for OCR. Bold identifies each column maximum.

	GenEval (10–190)		OCR (10–120)	
Configuration	AUC	Epoch 190	AUC	Epoch 120
Full FANFT	**0.92600**	**0.94074**	**0.95804**	**0.97766**
Measurement-only	0.92059	0.93347	0.94135	0.85701
Paired-collection-only	0.91228	0.92989	0.91998	0.85104

The pairwise evaluations show how differences accumulate within each interval. FANFT leads GenEval AUC over epochs 10–190 (Figure 13(a–f)). Over OCR epochs 10–120, its reported AUC exceeds the six controls in Figure 14(a–f). Each annotation compares the two methods over the interval named in its caption.

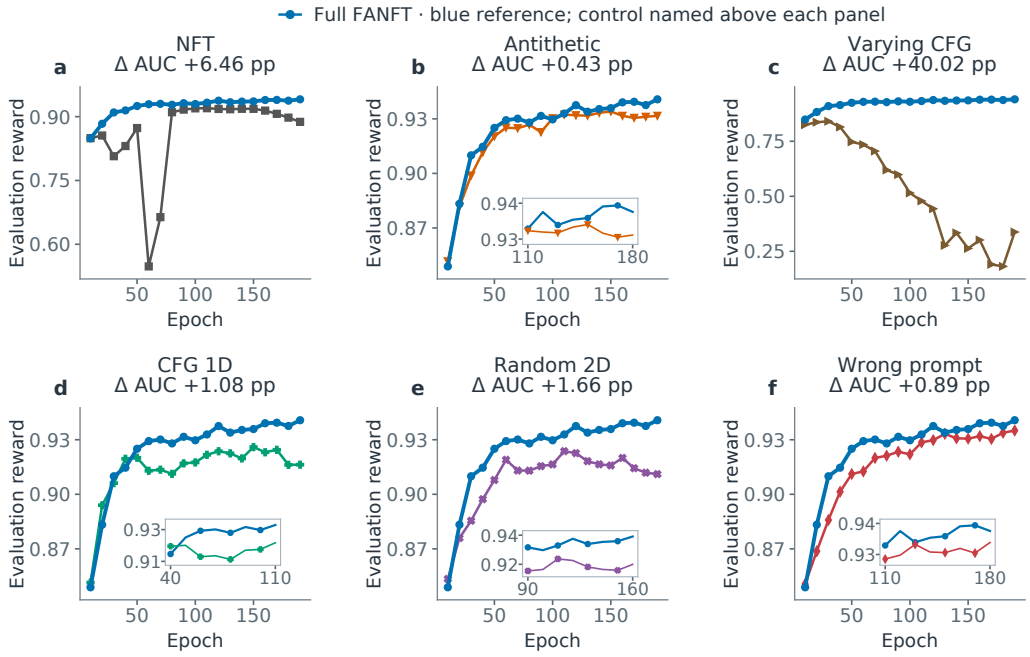

Figure 13: **GenEval: full trajectories for the training-configuration comparison.** Each panel compares full FANFT with one named control; all raw evaluations are retained. Insets magnify the indicated epoch intervals. The annotation is FANFT's normalized AUC minus the control's normalized AUC over the shared epoch-10–190 window, in percentage points. Each panel uses its own reward-axis range, shown by the labeled ticks.

D.4 Selective credit admission and training dynamics

The local correction is selectively activated (Figure 3(e,f)), with its magnitude controlled by fold agreement, projection, and shrinkage (Figure 15(c,d); Eqs. (7)–(8)). Global NFT supervision continues throughout training, and the credit-channel comparison relates directional writing to integrated evaluation reward (Figure 3(c,d)). Figure 15 resolves the same selective behavior over time and reports the nonzero-correction fractions for the no-write controls.

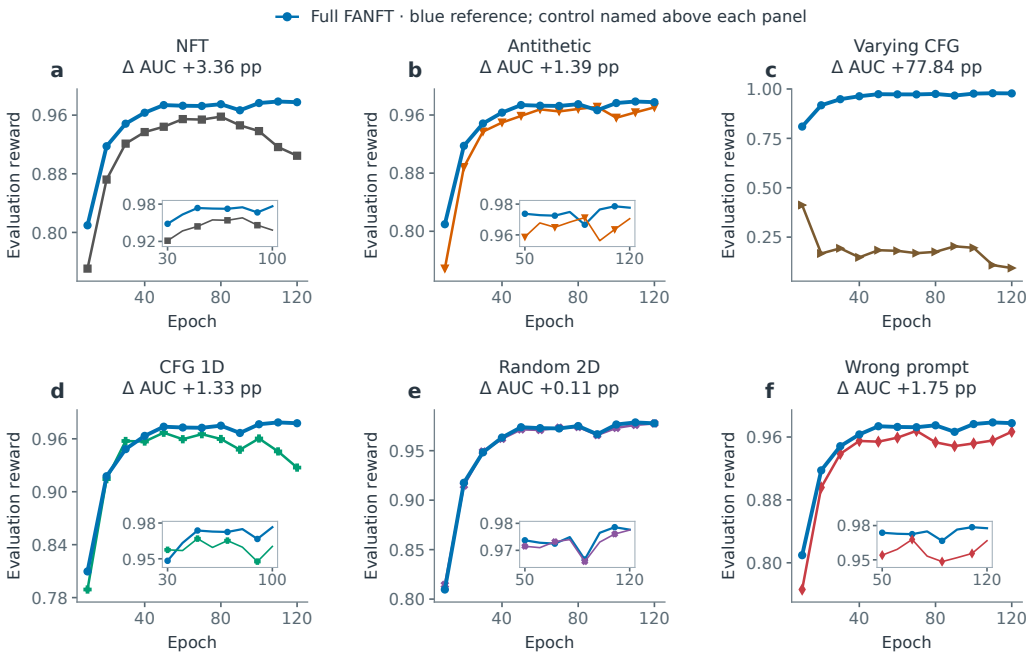

Figure 14: **OCR: configuration trajectories over epochs 10–120.** Each panel compares FANFT with one named control using 12 checkpoint evaluations. Annotations report 100 times the difference between FANFT's window-normalized AUC and the control's. Insets magnify the indicated intervals; each panel uses its labeled reward-axis range.

Table 11: **Frequency and magnitude of directional updates on GenEval.** n counts valid local/global norm-ratio measurements; Active is the nonzero-correction fraction. Means include zero corrections, with each statistic using its valid-measurement denominator. OCR diagnostics through epoch 120 appear in Figures 15(b,d) and 17.

Reward	Configuration	n	Active	Norm ratio	κ	Cosine
GenEval	FANFT full	200	47.0%	0.00871	0.0872	0.999849
GenEval	CFG 1D	199	55.3%	0.01369	0.1371	0.999729
GenEval	Random 2D	199	33.2%	0.00259	0.0320	0.999977
GenEval	Wrong-prompt 2D	197	46.7%	0.00872	0.0873	0.999852

The local signal is measured at four distinct stages: cross-fold agreement (Figure 16(a); Figure 17(a)), discovery/validation agreement (Figure 16(b); Figure 17(b)), admission confidence (Figure 16(c); Figure 17(c)), and applied norm ratio (Figure 16(d); Figure 17(d)). Points show update-wise means; intervals span the empirical 25th–75th percentiles of valid update measurements. Sparse admission can place a positive mean above an all-zero interquartile interval.

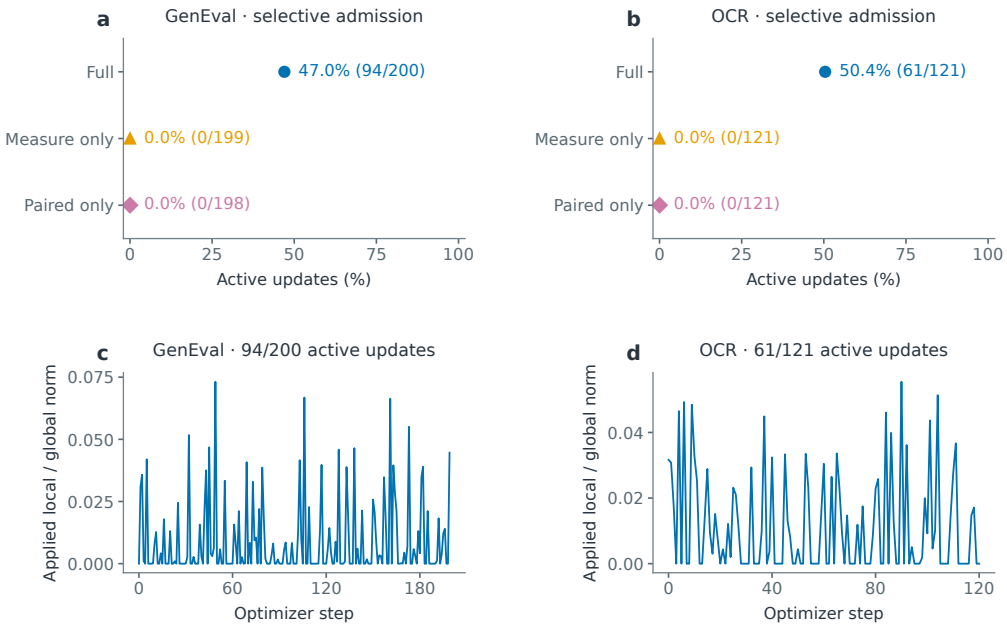

Figure 15: **Selective activation of directional credit during training.** Top: nonzero-correction fractions and valid counts. Bottom: applied local/global gradient-norm ratios. GenEval includes the full displayed history; OCR includes updates through epoch 120, as in Figure 3(e,f). All three configurations retain paired rewards.

Figure 18 compares FANFT and antithetic NFT through collection reward (a) and checkpoint evaluation (b). The former describes the rewarded collection, while the latter measures the quality of the updated model.

D.5 Computational cost

Table 12 compares GenEval per-round runtime within the paired-collection pipeline. Each configuration generates 576 images and performs one optimizer update per epoch. The three configurations take 171.51–174.00 seconds per epoch. These measurements place local estimation and writing in the context of the complete paired training round; image counts and GPU-hours describe the stated training intervals.

Table 12: **Measured GenEval cost of the paired credit-channel controls.** Images and GPU-hours are totals over the stated intervals; seconds per epoch is their average. Hardware is shared across configurations, which vary local estimation and writing within the paired-collection pipeline.

Reward	Configuration	Training epochs	Images	s/epoch	GPU-hours
GenEval	Full FANFT	0–199	115,200	171.51	76.23
GenEval	Measurement-only	0–198	114,624	172.89	76.46
GenEval	Paired collection	0–197	114,048	174.00	76.56



Figure 16: **GenEval: local-signal diagnostics by method.** Fold cosine, discovery/validation cosine, confidence, and applied local/global norm ratio. Points show update-wise means; horizontal intervals span the empirical 25th–75th percentiles of valid update measurements.

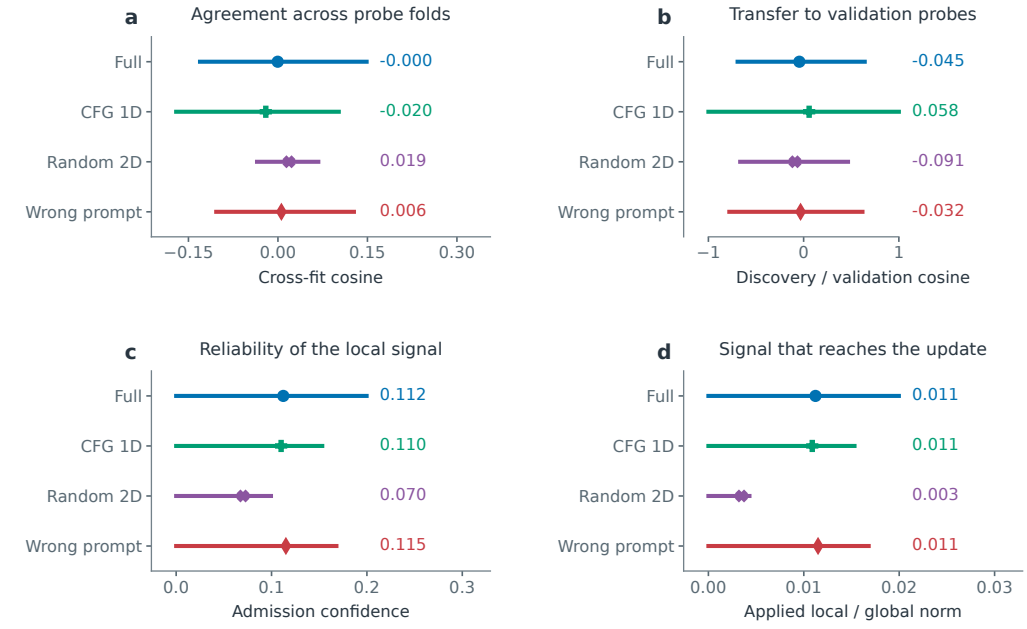

Figure 17: **OCR: local-signal diagnostics through epoch 120.** Fold cosine, discovery/validation cosine, confidence, and applied local/global norm ratio. Points show means; horizontal intervals span the empirical 25th–75th percentiles of valid update measurements in this interval.

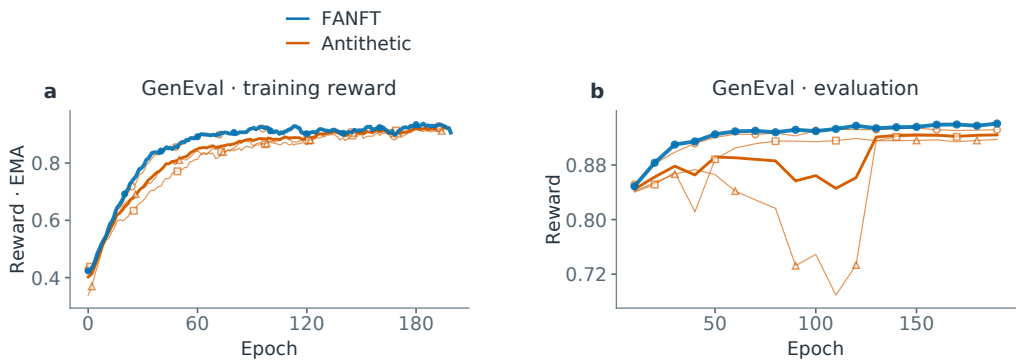

Figure 18: **Collection and evaluation dynamics with antithetic NFT.** EMA-0.1 collection reward and raw checkpoint evaluation are shown separately. Thin lines show individual trajectories; thick lines show their shared-epoch mean.

Reward queries and model computation. For N generated training images, both branches included, N objective-reward scores serve global and directional credit. Forward-process noise views reuse the same images and rewards. Evaluation images and their reward calls are counted separately from training-image acquisition.

Network cost includes rollout and basis-construction forwards, $n_v N$ global sample-view evaluations and their backward passes for n_v noise views, and local velocity differentiation. Fold accumulation and fusion add gradient operations. Total network computation is determined by these rollout, basis-construction, global-view, and local-pullback operations after accounting for batching, CFG-branch sharing, and graph reuse.

D.6 Full-system evaluation and checkpoint acquisition

Figure 4 combines evaluation and collection rewards (a–h) with checkpoint acquisition (i,j), using H200 GPU-hours as the budget coordinate under the within-task protocol in Table 4.

Recorded checkpoint acquisition. Let e_τ be the first recorded evaluation epoch with reward at least τ. Its image budget is $N_\tau = be_\tau$, with 576 images per epoch for GenEval and 1,152 for OCR, shared within each NFT/FANFT comparison. Table 13 reports these image counts and their NFT/FANFT ratios, counting each image once across its noise views. Figure 4(i,j) reports the H200 GPU-hour budgets for these reward targets.

Table 13: **Image counts at first recorded qualifying checkpoints.** Targets match those in Figure 4(i,j), which reports GPU-hour budgets. Ratios in this table are NFT/FANFT image counts.

Reward	Target	FANFT images	NFT images	NFT / FANFT
GenEval	0.90	11,520	17,280	1.50×
GenEval	0.92	17,280	23,040	1.33×
GenEval	0.94	23,040	57,600	2.50×
GenEval	0.95	28,800	74,880	2.60×
OCR	0.90	23,040	46,080	2.00×
OCR	0.92	23,040	46,080	2.00×

GenEval shared-window reward. For shared evaluation epochs $e_1 < \cdots < e_K$, normalized AUC is

$$\mathrm{AUC}_{[e_1,e_K]} = \frac{1}{e_K - e_1} \sum_{j=1}^{K-1} \frac{R(e_j) + R(e_{j+1})}{2} (e_{j+1} - e_j). \tag{29}$$

The GenEval comparison uses the same 13 checkpoints at epochs 10–130, giving FANFT/NFT AUCs of 0.95523/0.92997 and epoch-130 rewards of 0.97269/0.95008 (Figure 4(a)). The corresponding image counts imply first recorded 0.95 checkpoints at epochs 50 and 130 for FANFT and NFT, respectively (Table 13); Figure 4(i) reports their H200 GPU-hour budgets. Thus lower acquisition cost accompanies higher average evaluation reward on the shared horizon.

Piecewise-linear reference crossings. Figure 4(a,c) shows interpolated evaluation crossings in H200 GPU-hours. Table 14 retains the epoch-indexed evaluation and collection-reward reference calculations, with EMA coefficient $\alpha = 0.1$ for collection reward. For each readout in Table 14, the reference target τ_* is the maximum NFT reward within its shared epoch interval. Let (e_{j-1}, R_{j-1}) and (e_j, R_j) bracket its first upward crossing. The interpolated crossing index and NFT/FANFT ratio are

$$e^{\text{int}}_{\tau_*} = e_{j-1} + \frac{\tau_* - R_{j-1}}{R_j - R_{j-1}}(e_j - e_{j-1}), \qquad q_{\text{int}} = \frac{e^{\text{int}}_{\tau_*,\text{NFT}}}{e^{\text{int}}_{\tau_*,\text{FANFT}}}. \tag{30}$$

Table 14 locates the reference targets between adjacent epoch-indexed samples. Interpolation on the GPU-hour axis uses the same expression with GPU-hours in place of epochs. Recorded-checkpoint acquisition uses the first evaluated model reaching the target (Figure 4(i,j)).

Table 14: **Reference targets and interpolated crossing epochs.** Targets are NFT maxima over each shared interval, defined separately for evaluation and collection-reward EMA ($\alpha = 0.1$). Ratios are NFT/FANFT.

Reward	Readout	Target	FANFT epoch	NFT epoch	Ratio
GenEval	Eval	0.950083	42.350	130.000	3.07×
OCR	Eval	0.928547	18.010	40.000	2.22×
GenEval	Train EMA	0.935746	62.357	128.000	2.05×
OCR	Train EMA	0.928929	43.323	85.000	1.96×