SIMPLE ALIGNMENT FOR PIXEL-SPACE JIT: MATCHED CONTROLS AND LOW-RANK TARGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Simple early-layer representation alignment is a strong baseline for pixel-space JiT on ImageNet-256. With a training-only MLP head, searched FID-50K improves from 4.80 ± 0.16 to 3.69 ± 0.06 across three matched training seeds, a 23.1% reduction. The benefit persists at every common guided CFG scale tested in two complete sweeps and is corroborated by one fixed-CFG pair in the released PixelREPA implementation. Attachment matters: the final-block endpoint is harmful in two seeds. Mean-preserving rank-32 PCA targets retain approximately 87% of the searched-FID gain over two seeds, despite retaining only 26.5% of centered teacher variance. A single random rank-32 basis reaches a similar searched endpoint with 4.2% variance, without establishing equivalence between subspaces. Analysis of the cosine objective separates variance retention, target direction, and conditional predictability; it supplies testable hypotheses rather than a causal account of generation quality. These results support including simple early alignment in matched JiT evaluations and show that much of its measured benefit survives restricted teacher targets.

1 INTRODUCTION

Representation alignment (REPA; Yu et al., 2025) supervises intermediate diffusion features with a frozen visual encoder. For pixel-space JiT, prior work already evaluates plain REPA, but reported outcomes vary across configurations. PixelREPA (Shin et al., 2026) reports that plain REPA worsens FID from 4.37 to 5.14 and introduces a masked transformer adapter. This raises a concrete question: can a simple training-only MLP head already provide a substantial, reproducible benefit under a controlled pixel-space configuration?

Our central finding. Simple early-layer alignment is a strong baseline for pixel-space JiT: its benefit replicates across training seeds, persists at common guidance scales, and largely survives severe restriction of the teacher target. In our ImageNet-256 study, early REPA reduces mean FID by 23.1% across three paired training seeds. The direction holds at each common guided scale in two complete sweeps and is corroborated by one matched pair in the released PixelREPA implementation. Obtaining these gains requires no specialized transformer adapter and adds no component to the inference model. These controls establish a baseline that should be tested before attributing a failure to an inherent incompatibility between REPA and pixel-space generation.

The target intervention makes the finding more informative. A mean-preserving rank-32 PCA target retains approximately 87% of the full-target searched-FID improvement over two seeds, despite restricting its centered variation to 32 of 768 feature directions. A random rank-32 target reaches a similar searched endpoint in one run while preserving much less centered variance. Using the full teacher target is therefore unnecessary for retaining most of the measured benefit in this configuration. To interpret this result, we analyze the actual cosine supervision: retained variance, normalized target direction, and predictability from the student's features are distinct quantities.

Contributions.

1. **A simple baseline with replicated gains and guidance robustness.** Three paired training seeds establish the early-REPA improvement in our implementation. Existing two-seed

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.

sweeps establish its positive sign at all common tested guided CFG values; a fixed-CFG control in the released implementation corroborates the direction (§3). These comparisons separate the alignment benefit from cross-paper baseline offsets and from selecting a different guidance scale for each arm.

2. **Configuration-specific evidence for beneficial and harmful alignment.** Replicated early and final-block endpoints show that alignment's sign depends on its configuration. The public attachment audit maps the released encoder output to our beneficial fourth-block output (§4). This locates the comparisons precisely and distinguishes our measured deep-endpoint harm from the unavailable published failure.
3. **Evidence that most of the gain survives low-rank target restriction.** The two-seed PCA result establishes substantial retained benefit without the full target. The random-subspace control, a conditional alignment-risk transfer bound, and an analysis of mean retention show why centered variance is an incomplete account of the supervision (§5–6). The analysis supplies explicit hypotheses for what makes a restricted target useful.

Ordinary REPA on JiT and alignment-depth sensitivity have precedents in iREPA, REPA, SRA^2, and PixelU (Singh et al., 2025; Yu et al., 2025; Wang et al., 2026; Guo et al., 2026). Our contribution is the combination of matched replication, a released-configuration audit, and controlled target restriction, with an analytical account of the resulting supervision. The claims concern the tested JiT/ImageNet-256 setting. The released-code and random-basis controls each have one training seed; neither establishes a multi-seed cross-implementation effect size or subspace equivalence. Mechanism and generalization limits are discussed in §7.

2 Experimental setup and comparison scope

Model and objective. Our independent JAX JiT stack uses bottleneck patch embedding, rotary position embeddings, QK normalization, and 32 in-context class tokens. The default model is JiT-B/16 with 12 blocks on ImageNet-256. Lk denotes the output after zero-based block index k: L3 is after the fourth block, and L11 after the twelfth. Class tokens enter at index 4; alignment excludes class tokens. For clean-image teacher patch features $z_i = f(x)_i$ and noisy-image activations $h_i^{(k)}$, the training objective is

$$\mathcal{L} = \mathcal{L}_{\text{JiT}} + 0.1\,\mathbb{E}_{x,t,\epsilon}\left[\frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{g(h_i^{(k)})^\top z_i}{\|g(h_i^{(k)})\|_2\|z_i\|_2}\right)\right]. \tag{1}$$

The frozen teacher is DINOv2-B (Oquab et al., 2024), using the released PixelREPA pipeline's patch-16 resampling at resolution 256. The projector g uses input LayerNorm, three linear layers of widths $2048, 2048, 768$, and SiLU after the first two layers. It is discarded at inference. JiT predicts clean pixels with the velocity-based training objective of Li & He (2025); the alignment weight and teacher are held fixed through the depth sweep. The displayed loss uses ideal cosine notation. Our implementation replaces each norm denominator by $\|v\|+10^{-8}$; Appendix A gives the corresponding stabilized gradient. The released-stack control uses its own MLP and normalization, as specified below.

Training and sampling. The default nominal budget is 200 epochs, effective batch 1024, learning rate 2×10^{-4}, five warmup epochs, Adam $\beta_2 = 0.95$, and constant EMA 0.9996. Training times follow a logit-normal distribution: $t = \text{sigmoid}(-0.8 + 0.8\epsilon)$ with $\epsilon \sim \mathcal{N}(0, 1)$ (the implementation names this schedule `lognorm`). Sampling uses the implementation's Heun-50 setting and CFG interval $[0.1, 1.0]$. These settings follow the public JiT recipe; they do not assert bitwise equivalence across stacks. Spot recovery replays parts of interrupted epochs, so 200 epochs is a nominal budget, not a claim of exactly 250,200 state steps. Appendix B gives recovered endpoint steps and evaluation exceptions. Comparisons match nominal training budgets, not FLOPs or wall-clock time: teacher inference and the auxiliary head add training work. We report endpoint quality rather than a compute-matched training-speed advantage.

Metric and uncertainty conventions. Generative performance is evaluated with FID (Heusel et al., 2017) on 50,000 generated images (FID-50K, lower is better), using the recorded ADM-compatible ImageNet-256 reference statistics. "S1–5" means the minimum over CFG $\{1, 2, 3, 4, 5\}$

Table 1: Matched controls, JiT-B/16 at nominal 200 epochs. Compare methods within each implementation. Our rows report mean ± sample SD over three matched training seeds; released-code rows use one matched seed and a plain-MLP patch. CFG notation is defined in §2; per-run results are in Appendix B.

Method	CFG	FID-50K ↓
Our implementation		
Vanilla	S1–5	4.80 ± 0.16
REPA L1	S1–5	4.22 ± 0.15
REPA L3	S1–5	$\mathbf{3.69 \pm 0.06}$
Released PixelREPA implementation		
Vanilla	F3.1	4.70
REPA, fourth-block output	F3.1	**4.12**

for each arm; "S4–8" denotes the L11 extension. "F3.1" and "F2.8" are fixed guidance settings. These are discrete search minima, not optima over continuous CFG. Search and reporting use the same evaluations; there is no independent held-out guidance-selection set. Means use the training seeds documented in Appendix B, and ± denotes sample standard deviation ($n - 1$ denominator), not standard error or a confidence interval. No uncertainty is assigned to single-seed points, and training-seed SD does not isolate finite-sample FID noise. Paired seeds mean matched training seed values, not identical optimization trajectories or paired evaluation samples across methods. For L3 seed 44, S1–5 was performed on H200 and the selected CFG 4 was confirmed on TPU; we report the TPU confirmation, not five TPU evaluations. FID values, differences, and SDs are displayed to two decimals; all aggregates and percentages use the archived precision.

3 Simple alignment is a strong baseline under matched controls

Independent-stack evidence. L3 improves every matched seed: vanilla $4.98/4.72/4.70$ becomes $3.71/3.75/3.63$ for seeds 42/43/44 (Table 1). The reduction in mean FID is 23.1%. L1 also improves all three pairs, with a 12.2% reduction in the mean. The gains are consistent across these endpoints, but three seeds remain insufficient to estimate tail behavior or precise uncertainty. The seed-42 training curve preserves the L3 advantage at each measured 40-epoch checkpoint (Appendix D). Seed 42 also supplies the exploratory depth and rank profiles. Excluding it leaves mean vanilla/L3 FID of $4.71/3.69$ over seeds 43/44, a paired improvement of 1.03 FID (21.8% of the baseline mean). Thus the replicated gain is not driven by the seed used for those profiles. This is a post-hoc sensitivity check, not a preregistered test or a held-out configuration-selection experiment.

The sign does not require separate guidance selection. Table 2 compares the same CFG across arms using the complete TPU sweeps for training seeds 42/43. L3 improves in all eight comparisons over CFG 2–5, with paired reductions from 0.71 to 2.01 FID. Thus its observed benefit is not confined to choosing CFG 3 for vanilla and CFG 4 for REPA. These are post-hoc summaries of existing checkpoints; they do not replace held-out guidance selection or constitute eight independent replicas. Seed 44 is omitted because its archived TPU confirmation covers only CFG 4. For the three searched endpoints, paired reductions are 1.27, 0.98, and 1.07 FID (mean 1.11, sample SD 0.15). Appendix C reports a leave-one-seed-out check; neither summary is a significance test.

Released-stack evidence. We pin the public PixelREPA code and replace its masked transformer adapter branch with a plain MLP, keeping the teacher, data loading, optimizer, and schedule. Both arms instantiate the same patch; the auxiliary weight is 0 for vanilla and 0.1 for REPA. At nominal 200 epochs, the stack's own evaluation at fixed CFG 3.1 yields 4.70 and 4.12, respectively, a 12.3% within-stack improvement. This is a single paired training seed, not a multi-seed replication of the published result. The artifact includes the patch, launch scripts, and attachment check. The released `build_mlp` has the same hidden widths and SiLU activations but omits our input LayerNorm; it uses `F.normalize` rather than our additive norm stabilizer. Thus the pair corroborates the benefit of simple alignment in that stack, without isolating any of these cross-implementation differences.

Table 2: Common-guidance sensitivity from archived TPU sweeps (training seeds 42/43). Both arms use the same CFG in each row, with clipping disabled. Means are over two training seeds; paired reductions report mean ± sample SD. These post-hoc evaluations reuse checkpoints and are not eight independent training replicas.

CFG	Vanilla FID	L3 FID	Paired reduction
2	8.83	7.41	1.42 ± 0.76
3	4.85	4.01	0.84 ± 0.18
4	5.06	3.73	1.33 ± 0.06
5	5.98	4.13	1.85 ± 0.23

What the code audit establishes. In the pinned release, the B/16 encoder executes `blocks[:in_context_start]` with `in_context_start=4`. Both the adapter and patched MLP consume this encoder output, which maps to L3 in our convention. The published failing plain-REPA branch was not available in the audited source, so this check cannot establish its implementation. It does establish that our released-stack control and independent-stack L3 control attach after the same number of blocks. Our harmful L11 intervention therefore does not explain PixelREPA's published $4.37 \rightarrow 5.14$ pair.

Cross-stack calibration and additional controls. Absolute performance differs across implementations. A recipe check moves our seed-42 vanilla from 4.98 to 4.67; reference-statistics and clipping checks do not account for the remaining discrepancy (Appendix E). We therefore make no claim of outperforming PixelREPA's method from cross-paper numbers. Within our JAX stack, the benefit also appears in single paired runs at 600 epochs (S1–5: $3.81 \rightarrow 3.19$) and for JiT-L/16 at 200 epochs (F2.8: $3.14 \rightarrow 2.32$, attachment index 6). These extend the observed scope but do not replicate it statistically across budgets or scales.

Why both controls matter. The three-seed experiment tests replication; the released-stack pair checks the direction in public code at fixed CFG. Both estimate a within-implementation change from vanilla to simple alignment at an audited attachment. Matched MTA and iREPA arms are absent, so the evidence does not rank the MLP against those adapters.

4 Replicated endpoints and an exploratory depth profile

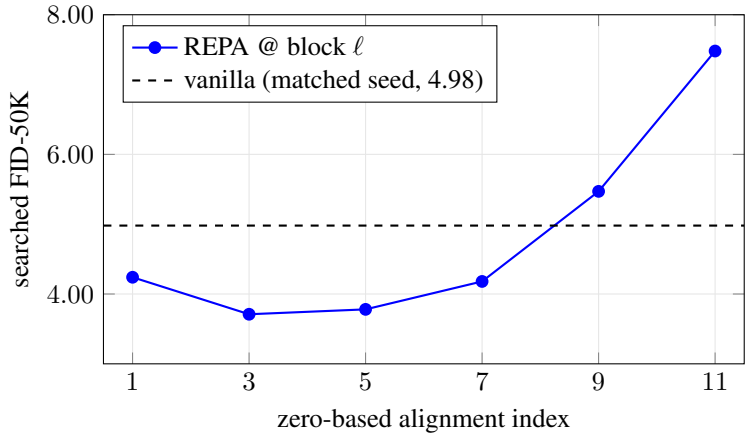


Figure 1: Seed-42 attachment profile, FID-50K minimized over CFG $\{1, 2, 3, 4, 5\}$. The dashed reference is the same-seed vanilla. Connected points are descriptive, with no intermediate-layer replication or error bars. L9 and the original L11 select the right edge; the separate L11 extension and replica appear in Table 3. The position of a population-level minimum or sign crossing is unresolved.

Table 3: Replicated harmful endpoint. All rows report mean ± sample SD over the same two training seeds. L11 uses S4–8 with clipping enabled; its original unclipped S1–5 evaluation was 7.48 for one run. The extension is an evaluation-setting check, not solely a wider search. Per-run values and settings are in Appendix B.

Method	CFG	FID-50K ↓
Vanilla	S1–5	4.85 ± 0.18
REPA L3	S1–5	3.73 ± 0.03
REPA L11	S4–8	7.70 ± 0.33

The seed-42 depth profile evaluates every other block with all training settings otherwise fixed (Figure 1). On the common S1–5 grid, FID is $4.24, 3.71, 3.78, 4.18, 5.47, 7.48$ at L1/L3/L5/L7/L9/L11; vanilla is 4.98. This sampled profile is non-monotone. L3 and L5 differ by only 0.07, and L5 is single-seed; we cannot identify an exact optimum. L9 selects CFG 5 at the right boundary, so its reported value may overestimate the best achievable with a wider search. In particular, the observed L7-to-L9 sign change on this grid is not a localized or universal crossing.

The deep endpoint remains harmful. An independent seed-42 L11 evaluation on CFG $\{4, 5, 6, 7, 8\}$, with clipping enabled, returns $7.66, 7.46, 8.36, 10.24, 13.35$. The minimum is now interior and within 0.02 of the original 7.48. A second training seed reaches 7.93 at CFG 5 on the extended grid, compared with its matched vanilla 4.72. Relative to the vanilla endpoints under their recorded policies, the two degradations are approximately 50% and 68%. This supports a harmful endpoint across two seeds under the stated search policies, without establishing a common-grid multi-seed curve or ruling out all other sampling settings.

Depth is a practical intervention, not an isolated mechanism. L3 and L11 differ in both position and their relation to the class-token insertion boundary at index 4. The sweep holds the model architecture fixed, but cannot separate depth from conditioning structure. At fixed CFG 3.1, the first measured L11 checkpoint at epoch 40 is already worse than vanilla, and the relative gap grows through epoch 200 (Appendix D). This differs from demonstrating a late reversal following an initial benefit.

5 Most of the gain survives low-rank target restriction

We compute a global PCA basis over 25.6 million teacher patch tokens, shared across images and spatial positions. With mean μ and top-r eigenvectors V_r, the modified target is

$$z^{(r)} = \mu + V_r V_r^\top (z - \mu). \tag{2}$$

The output remains 768-dimensional and the MLP is unchanged; rank refers to the centered variation of the target, not the projector width or a compressed inference model. Retained variance is the fraction of centered PCA eigenvalue mass. Thus the intervention also preserves the teacher mean and spatially indexed patch targets. We still compute the full teacher features before projection; the experiment tests target restriction and does not implement a smaller teacher or a cheaper training pipeline.

What rank 32 preserves. Rank 32 reaches 3.84 and 3.90 for seeds 42/43. Define retained gain using the same seeds in all terms as $G = (\overline{\text{FID}}_{\text{vanilla}} - \overline{\text{FID}}_{r=32})/(\overline{\text{FID}}_{\text{vanilla}} - \overline{\text{FID}}_{\text{full}})$. This gives approximately 87%; the per-seed ratios are approximately 90% and 85%. Thus a target restricted to 32 centered principal directions plus the mean retains most of the measured benefit. A recovered random-subspace control replaces V_r with a fixed orthonormal basis from Gaussian QR (basis seed 0), preserving the mean, rank, L3 attachment, weight, and nominal budget. It retains 4.151% of centered teacher variance yet reaches FID 3.84, compared with top-PCA's 3.84 at the same training seed 42. This searched endpoint does not resolve an advantage for either subspace. The random arm ran on v5e rather than the rank-32 baseline's v5p, and has only one training seed and one basis seed. Matching endpoints after rounding to two decimals are not an equivalence result. Mean retention, spatial correspondence, and other target properties remain possible contributors. We cannot infer that retained variance determines alignment benefit.

Table 4: Low-rank evidence at L3, nominal 200 epochs, S1–5. Panel A uses seed 42 throughout; Panel B uses seeds 42/43 throughout. Variance refers to centered teacher features. Random rank 32 uses one basis. Panel B reports mean ± sample SD.

A: Target and objective comparisons		
Method	Retained variance	FID-50K ↓
Vanilla	—	4.98
Dispersive, teacher-free	—	4.53
VICReg-style variance/covariance	—	5.49
Top-PCA rank 8	13.7%	4.01
Top-PCA rank 32	26.5%	3.84
Random rank 32	4.2%	3.84
Top-PCA rank 384	80.1%	3.65
Full teacher	100%	3.71
B: Replicated comparisons		
Method	Retained variance	FID-50K ↓
Vanilla	—	4.85 ± 0.18
Top-PCA rank 32	26.5%	3.87 ± 0.04
Full teacher	100%	3.73 ± 0.03

The retained-gain ratio depends on guidance. The 87% ratio summarizes the separately searched endpoints, not a fraction of teacher information transferred. Applying the same ratio at common CFG 2, 3, 4, and 5 over seeds 42/43 gives 92%, 106%, 89%, and 83%, respectively. A value above 100% simply means rank 32 has lower FID than the full target at that setting. Similarly, for seed 42, random-minus-PCA FID is $+0.23$, $+0.16$, less than 0.01 in magnitude, and -0.13 across CFG 2–5. Their ordering changes with guidance; the near tie of searched endpoints therefore cannot establish either a universal PCA advantage or guidance-independent equivalence. All these comparisons reuse the archived sweeps (Appendix C).

Rank 384 and the full target differ by 0.06 in seed 42. This single-seed difference supplies neither an equivalence test nor evidence of a precise saturation threshold. CFG spacing does not define a FID error bar. Only the rank-32 endpoint, not the rank profile, has been replicated.

Teacher-free comparisons. An InfoNCE-ℓ_2 dispersive objective (Wang & He, 2025) at the same activations reaches 4.53, retaining approximately 35% of the seed-42 full-REPA gain. A VICReg-style variance/covariance objective (Bardes et al., 2022) reaches 5.49 at nominal 200 epochs, worse than the same-seed vanilla. It omits paired-view invariance and uses outer weight 0.1, variance weight 1, and covariance weight 0.04. The completed budget, rather than its earlier 185-epoch evaluation, is reported. These are individual configured objectives with one seed each, not exhaustive or equally optimized representatives of teacher-free learning. Their results do not establish that teacher information is necessary or separate information from geometry in general.

6 An Analytical View of Target Restriction and Alignment

The low-rank intervention acts on a teacher target, while FID measures the output of a trained sampler. To connect these levels without equating them, we analyze the intervening cosine objective. The elementary results below identify properties of the alignment signal; they do not establish convergence. Appendix A provides derivations and a counterexample.

6.1 Centered Variance is Not the Dimension of All Supervision

Let $z \in \mathbb{R}^d$ have mean μ and covariance C, with finite second moment. For a rank-r orthogonal projector P, write $q_P = \mu + P(z - \mu)$; $d = 768$ here. The reconstruction residual and retained centered variance satisfy

$$\mathbb{E}\|z - q_P\|^2 = \operatorname{tr}((I - P)C), \qquad \rho(P) = \frac{\operatorname{tr}(PC)}{\operatorname{tr}(C)}, \tag{3}$$

where $\mathrm{tr}(C) > 0$. Top-PCA minimizes this residual among rank-r projectors. For a uniformly random rank-r subspace independent of C, $\mathbb{E}_P \rho(P) = r/d$. Thus $32/768 \simeq 4.17\%$ is a reference expectation for an isotropic random basis, close to the measured 4.151%; it is not a new fitted explanation of its FID. The expectation is over bases, whereas the experiment uses one fixed basis.

Mean retention changes the interpretation of rank. Although $\mathrm{Cov}(q_P) = PCP$ has rank at most r, every target lies in $\mu + \mathrm{range}(P)$. After cosine normalization, its direction lies in the linear space $\mathrm{span}(\mu, \mathrm{range}(P))$, of dimension at most $r + 1$. Normalization generally removes the affine-subspace description; it does not remove this linear-span bound. Consequently, "rank 32" restricts centered variability while retaining an additional possible mean direction and the association between each patch and its target. It neither removes the teacher nor makes the supervision equivalent to 32 independent labels. This distinction motivates mean-removal and patch-shuffling controls if one seeks to identify which retained structure is responsible for the gain.

6.2 Cosine supervision depends on direction and predictability

Result 1: a local perturbation bound. For nonzero head output a, teacher z, and projected target q_P, let $\widehat{v} = v/\|v\|$ and $\ell(a, z) = 1 - \widehat{a}^\top \widehat{z}$. Then

$$\nabla_a \ell(a, z) = -\frac{1}{\|a\|}(I - \widehat{a}\widehat{a}^\top)\widehat{z}, \tag{4}$$

$$\|\nabla_a \ell(a, z) - \nabla_a \ell(a, q_P)\| \le \frac{\|\widehat{z} - \widehat{q}_P\|}{\|a\|} \le \frac{2\|(I - P)(z - \mu)\|}{\|a\|\,\|z\|}. \tag{5}$$

These statements describe exact cosine normalization away from zero norms; numerical stabilization is handled explicitly in Appendix A. Multiplication by the head/encoder Jacobian transfers the perturbation to parameter space and can amplify or suppress it. A scalar retained-variance fraction therefore does not determine either the per-token angular change or its effect on shared parameters. PCA's reconstruction optimality gives no ordering of the actual gradient discrepancies, let alone final FID. The bound can also be loose when a large residual changes target magnitude more than direction.

Result 2: attainable alignment risk depends on available features. Fix an encoder and let H be the information available to its alignment head for a random token. Suppose $Q = \widehat{q}_P$ is defined almost surely. Among all measurable unit-vector predictors $u(H)$, the minimum expected cosine loss is

$$\inf_{\|u(H)\|=1} \mathbb{E}[1 - u(H)^\top Q] = 1 - \mathbb{E}\|\mathbb{E}[Q \mid H]\|. \tag{6}$$

Where $m(H) = \mathbb{E}[Q \mid H] \neq 0$, an optimal direction is $u^*(H) = m(H)/\|m(H)\|$. This fixed-encoder optimum lower-bounds a finite MLP's risk, without characterizing joint training.

The conditional quantity in Eq. (6) differs from the unconditional covariance in Eq. (3). Low-variance directions may be predictable from noisy features, while high-variance directions need not be. Projection and normalization can change this predictability in either direction. Appendix A gives a two-dimensional example where retaining the principal direction yields worse optimal cosine loss than retaining the lower-variance direction. This establishes that no universal monotonic relation follows from variance retention alone; it does not identify which directions help JiT. Our random/PCA endpoints do not measure this conditional quantity or identify the example's mechanism. Moreover, a nonzero mean-only target ($P = 0$) has a constant direction and optimal risk zero in Eq. (6), while carrying no image-dependent variation. Minimizing this auxiliary risk therefore supplies no ranking of targets by their denoising utility. Mean-only, mean-removal, and patch-shuffling interventions would help distinguish the retained structures; none is measured here.

6.3 When restricted-target alignment transfers to the full target

We now bound a restricted-target head's excess alignment risk on the full target at fixed H. Let $Q = \widehat{z}$ and $Q_P = \widehat{q}_P$ be defined almost surely. For measurable unit predictors, write $R_Q(u) = \mathbb{E}[1 - u(H)^\top Q]$ and let R_Q^* denote the unrestricted optimum in Eq. (6). Define

$$\delta_H(P) = \mathbb{E}\big\|\mathbb{E}[Q - Q_P \mid H]\big\| \le \mathbb{E}\|Q - Q_P\|. \tag{7}$$

It compares the target directions that are predictable from the same features, rather than the total energy discarded by projection.

Result 3: transfer of alignment risk. For any such predictor u_P, let $e_P = R_{Q_P}(u_P) - R^*_{Q_P} \geq 0$ be its excess risk on the restricted target. Then

$$|R^*_Q - R^*_{Q_P}| \leq \delta_H(P), \qquad R_Q(u_P) - R^*_Q \leq e_P + 2\delta_H(P). \tag{8}$$

Proof. Conditional expectation and Cauchy–Schwarz give, uniformly in u, $|R_Q(u) - R_{Q_P}(u)| = |\mathbb{E}[u(H)^\top \mathbb{E}[Q_P - Q \mid H]]| \leq \delta_H(P)$. Taking infima proves the first bound. Adding the restricted-target excess risk and the two changes of target proves the second. Conditional Jensen gives Eq. (7). □

Thus small restricted-target excess risk transfers when the predictable direction changes little. This is relative to R^*_Q, which need not be small. The term e_P retains finite head capacity and incomplete optimization; it is not assumed to vanish. For a more conservative sufficient condition, if $\|z\| \geq b > 0$ almost surely, the normalization inequality used in Result 1 and Cauchy–Schwarz imply

$$\delta_H(P) \leq \mathbb{E}\|Q - Q_P\| \leq \frac{2}{b}\sqrt{\operatorname{tr}((I-P)C)}. \tag{9}$$

PCA minimizes this last upper bound at fixed rank, but not necessarily $\delta_H(P)$ or the actual transfer error. In particular, a variance fraction supplies neither the norm scale b nor the conditional discrepancy. Our rank-32 results do not measure these quantities, so Eq. (8) is a testable sufficient condition, not a certificate for the observed 87% retained FID gain.

Result 4: an easy mean-dominated target need not be informative. For $\mu \neq 0$, a constant predictor $u(H) = \widehat{\mu}$ gives

$$R^*_{Q_P} \leq \min\left\{1, \frac{2\operatorname{tr}(PC)}{\|\mu\|^2}\right\}. \tag{10}$$

Proof. Since unit vectors satisfy $1 - \widehat{\mu}^\top Q_P = \frac{1}{2}\|\widehat{\mu} - Q_P\|^2$ and $\|\widehat{\mu} - Q_P\| \leq 2\|q_P - \mu\|/\|\mu\|$, the constant predictor has risk at most $2\mathbb{E}\|P(z-\mu)\|^2/\|\mu\|^2$. Eq. (3) evaluates this expectation, and Result 2 gives the additional upper bound of one. □

When retained variation is small relative to the mean, low auxiliary loss is possible without using any image-dependent feature. This extends the mean-only example to nonzero-rank targets. It neither establishes mean dominance in our cached features nor rules out useful variation in the retained subspace. Together, Results 3–4 separate preserving alignment to the full teacher from merely making a restricted target easy to predict. Measuring the conditional discrepancy and testing mean removal address different questions; neither alone would establish improved denoising or FID.

6.4 Why attachment can alter the effect of the same target

At attachment k, both the available features H_k and the Jacobian carrying the auxiliary signal to shared parameters change. Even an easily predicted teacher target need not produce a useful denoising update. To separate these issues, let $g_D = \nabla_\theta L_D$ and $g_A = \nabla_\theta L_A$ at a fixed parameter state, with smooth denoising loss L_D. For the idealized gradient step $\theta^+ = \theta - \eta(g_D + \lambda g_A)$,

$$L_D(\theta^+) - L_D(\theta) = -\eta\|g_D\|^2 - \eta\lambda\langle g_D, g_A\rangle + O(\eta^2). \tag{11}$$

The auxiliary term helps the immediate denoising decrease if the inner product is positive and reduces that decrease if it is negative. A negative inner product alone does not imply an increase in loss: the first-order change becomes positive only if $\lambda\langle g_D, g_A\rangle < -\|g_D\|^2$. Gradient interference has established precedents in multi-task optimization (Yu et al., 2020); we use this local identity to formulate a hypothesis, not to claim its discovery or its measurement in our runs.

In our depth experiment, a plausible explanation is a change in the compatibility of the auxiliary signal with denoising, together with changes in target predictability and conditioning. The L3/L11 endpoints motivate this hypothesis but do not distinguish its components. Eq. (11) does not model Adam's evolving state, long-horizon optimization, or FID. A direct test would measure timestep-resolved gradient compatibility on matched checkpoints and vary the class-token boundary independently of attachment. Our generated-image and noised-real probes cannot substitute for those measurements.

7 Related work and limitations

Alignment in latent and pixel space. REPA studies early/interior attachment in latent diffusion transformers (Yu et al., 2025; Peebles & Xie, 2023; Ma et al., 2024). SRA^2 explicitly reports a deep SiT attachment that is worse than vanilla (Wang et al., 2026); general depth-dependent harm is therefore prior evidence. HASTE studies stopping alignment over training time (Wang et al., 2025), a different intervention from holding the training schedule fixed while changing attachment. iREPA includes REPA-versus-iREPA comparisons on JiT and emphasizes spatial target structure (Singh et al., 2025). Its reported unguided JiT setting differs from our guidance-searched control. Our emphasis is the within-stack comparison and public attachment audit. PixelU's four-depth ablation improves over its no-alignment baseline at every tested depth (Guo et al., 2026). V-Co studies interacting pixel and feature denoising streams (Lin et al., 2026). Our comparisons instead hold a single-branch architecture fixed and examine the released PixelREPA configuration alongside matched controls. Self-representation guidance (Jiang et al., 2025) also goes beyond the two teacher-free objectives tested here.

Mechanism remains unresolved. Exploratory diagnostics show lower classifier confidence in the deep arm, but neither this correlation nor the spectral and conditioning probes establishes a cause (Appendix F). Generated-image diagnostics use only 560 images per arm; the real reference subset has 2,000 images from the first 200 sorted training classes. These are separate from FID-50K evaluations and cannot rule out population-level mode collapse. Conditioning probes use noised real images rather than sampler trajectories. The probes do not reconstruct PixelREPA's published bucket-level test.

Generalization and remaining uncertainty. The study covers one model family and ImageNet-256. The released-stack pair and random target each have one training seed, with one random basis. Class-token structure and evaluation policy limit the depth interpretation. Complete checkpoint provenance and some guidance sweeps remain unavailable (Appendix B).

8 Conclusion

Simple early-layer REPA improves all three matched training seeds, preserves its sign at every common guided scale tested in two sweeps, and is corroborated in the released implementation. The training-time MLP adds no inference component. Replicated harmful final-block endpoints show that attachment is an essential part of this baseline.

Most of the measured gain also survives mean-preserving rank-32 target restriction. The random control and cosine analysis show why centered variance alone is insufficient to characterize the supervision. Determining which retained structures cause the gain requires further interventions.

For JiT evaluations in this setting, our results support including a simple early-MLP baseline and reporting both common-guidance comparisons and searched endpoints, with the attachment output and head specified. The evidence establishes this baseline's utility; ranking it against MTA or iREPA requires their own matched comparisons.

Reproducibility statement

The supplement provides arm configurations, an anonymous claim ledger, training and evaluation commands, the pinned released-code patch and attachment check, PCA construction, and diagnostic code. Appendix B records endpoint and artifact limitations. Minimal CPU checks reproduce the reported aggregate and sensitivity calculations and test the analytical identities without installing the training stack. Checkpoints, ImageNet images, and full sample archives are not distributed.

AI use statement

Generative AI assisted drafting, restructuring, literature discovery and checking, code and experimental workflows, numerical audits, analytical formulation and derivation, and simulated reviews.

Reported generation results come from experimental records. The authors take responsibility for the text, evidence, citations, and code.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Zipeng Guo, Lichen Ma, Yu He, Xiaolong Fu, Jingling Fu, Junshi Huang, and Yan Li. PixelU: A u-shaped transformer for efficient end-to-end pixel diffusion. *arXiv preprint*, 2026. arXiv:2606.27760.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.08500.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint*, 2025. arXiv:2505.02831.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint*, 2025. arXiv:2511.13720.

Han Lin, Xichen Pan, Zun Wang, Yue Zhang, Chu Wang, Jaemin Cho, and Mohit Bansal. V-Co: A closer look at visual representation alignment via co-denoising. *arXiv preprint*, 2026. arXiv:2603.16792.

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. *arXiv preprint*, 2024. arXiv:2401.08740.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2304.07193.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Jaeyo Shin, Jiwook Kim, and Hyunjung Shim. Representation alignment for just image transformers is not easier than you think. *arXiv preprint*, 2026. arXiv:2603.14366.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *arXiv preprint*, 2025. arXiv:2512.10794.

Mengmeng Wang, Dengyang Jiang, Liuzhuozheng Li, Yucheng Lin, Guojiang Shen, Xiangjie Kong, Yong Liu, Guang Dai, and Jingdong Wang. SRA 2: Variational autoencoder self-representation alignment for efficient diffusion training. *arXiv preprint*, 2026. arXiv:2601.17830v2.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint*, 2025. arXiv:2506.09027.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, Kai Wang, and Yang You. REPA works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *arXiv preprint*, 2025. arXiv:2505.16792.

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 (ICLR)*, 2025.

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 (NeurIPS)*, 2020.

A Derivations and an illustrative counterexample

These derivations concern the simplified objectives specified in §6. They introduce no additional trained models or measured generation results. The supplement's `tools/check_theory.py` checks algebraic identities and the finite-distribution example on CPU; numerical checks complement, rather than replace, the derivations.

Projection identities and the random-basis expectation. Write $x = z - \mu$. Since $P = P^\top = P^2$, $\mathbb{E}\|(I-P)x\|^2 = \operatorname{tr}((I-P)C)$ and $\operatorname{Cov}(q_P) = PCP$. Maximizing $\operatorname{tr}(PC)$ selects the top r eigenvectors of C, giving the PCA reconstruction property. For a uniformly random subspace, rotational invariance implies $\mathbb{E}_P P = \alpha I$; taking traces yields $\alpha = r/d$. For fixed C, linearity gives $\mathbb{E}_P \operatorname{tr}(PC) = (r/d)\operatorname{tr}(C)$. Gaussian QR supplies such a subspace for independent isotropic Gaussian entries; orientation signs do not change P. This expectation does not provide an error bar for our single measured basis or predict its generative performance.

The target lies in $\mu + \operatorname{range}(P)$, hence in the linear span of μ and that range. Dividing a nonzero target by its norm preserves membership in that linear span, of dimension at most $r+1$. Its normalized covariance need not have rank r. In the rank-zero case with $\mu \neq 0$, all target directions are constant; retaining a mean does not by itself preserve image-dependent semantic variation. Our measurements do not include this mean-only control.

Derivation of Result 1. Differentiating $a/\|a\|$ gives $D_a\widehat{a} = (I - \widehat{a}\widehat{a}^\top)/\|a\|$, establishing Eq. (4). The matrix in parentheses is an orthogonal projector with operator norm at most one. For nonzero z, q,

$$\begin{aligned}\|\widehat{z} - \widehat{q}\| &\le \left\|\frac{z-q}{\|z\|}\right\| + \left\|q\left(\frac{1}{\|z\|} - \frac{1}{\|q\|}\right)\right\| \\ &= \frac{\|z-q\| + |\|q\| - \|z\||}{\|z\|} \le \frac{2\|z-q\|}{\|z\|}.\end{aligned}$$

Substitution of $q = q_P$ proves Eq. (5). At a common parameter state with $a = a(\theta)$ and fixed teacher targets, the parameter-gradient discrepancy is bounded by $\|J_a\|_{\mathrm{op}}\|\widehat{z} - \widehat{q}_P\|/\|a\|$. This comparison holds at the same a and Jacobian, not between differently trained final checkpoints. Without control of these norms and residual directions, a variance ratio supplies no tight gradient estimate.

The additive stabilizer used by our implementation. In `jit_jax/diffusion.py`, define $n_\varepsilon(v) = v/(\|v\| + \varepsilon)$ with $\varepsilon = 10^{-8}$ and $\ell_\varepsilon(a, z) = 1 - n_\varepsilon(a)^\top n_\varepsilon(z)$. For $s = \|a\| > 0$,

$$Dn_\varepsilon(a) = \frac{I}{s+\varepsilon} - \frac{aa^\top}{s(s+\varepsilon)^2}, \qquad \nabla_a \ell_\varepsilon(a, z) = -Dn_\varepsilon(a) n_\varepsilon(z). \tag{12}$$

The radial eigenvalue is $\varepsilon/(s+\varepsilon)^2$ and each tangential eigenvalue is $1/(s+\varepsilon)$, so the operator norm is at most $1/(s+\varepsilon)$. The same decomposition as above, using $\|q\|/(\|q\| + \varepsilon) \le 1$, gives

$$\|\nabla_a \ell_\varepsilon(a, z) - \nabla_a \ell_\varepsilon(a, q_P)\| \le \frac{\|n_\varepsilon(z) - n_\varepsilon(q_P)\|}{s+\varepsilon} \le \frac{2\|z - q_P\|}{(s+\varepsilon)(\|z\| + \varepsilon)}. \tag{13}$$

As a mathematical function, n_ε is differentiable at $a = 0$ with derivative I/ε; an autodiff implementation of the norm may need a separate zero-input convention. This identity does not certify that numerical case. The released PyTorch branch uses `F.normalize`, whose denominator is $\max(\|v\|, \varepsilon)$; Eq. (12) describes our additive implementation only. The fixed-encoder unit-direction analysis in Result 2 remains an idealized cosine result, not the exact optimum of either stabilized objective.

Derivation of Result 2. Conditioning on H gives $\mathbb{E}[1 - u(H)^\top Q \mid H] = 1 - u(H)^\top m(H)$. Cauchy–Schwarz bounds $u(H)^\top m(H) \le \|m(H)\|$, attained by normalizing $m(H)$ when nonzero. If $m(H) = 0$, every unit direction attains conditional loss one. Taking expectations yields Eq. (6). Since $\|Q\| = 1$, $0 \le \|m(H)\| \le 1$, so the optimal expected loss lies in $[0, 1]$. Parameter sharing, finite head capacity, and training dynamics may prevent the implemented head from attaining this unrestricted optimum.

Scope and edge cases of Results 3–4. The proofs of Eqs. (8)–(10) are in the main text. Both risks in Result 3 use the same joint distribution and the same fixed features H; the theorem does not compare the encoders of two separately trained checkpoints. Conditional expectations of unit vectors always exist, so the transfer bound requires no density assumption or lower norm bound beyond targets being nonzero almost surely. The stronger assumption $\|z\| \ge b > 0$ is used only for Eq. (9). Explicitly, the normalization inequality gives $\mathbb{E}\|Q - Q_P\| \le (2/b)\mathbb{E}\|(I - P)(z - \mu)\|$; Cauchy–Schwarz and Eq. (3) give the displayed square root. When this bound is large, the universal bound $\delta_H(P) \le 2$ still holds, but need not give a useful transfer guarantee. No such norm floor is verified for our teacher.

In Result 4, the bound by $2\,\mathrm{tr}(PC)/\|\mu\|^2$ follows from the particular constant predictor, whereas the bound by one follows from the unrestricted optimum; the constant predictor itself need not have risk at most one. The result requires $\mu \neq 0$ and cannot be applied to centered targets by substituting a zero mean. At $P = 0$ its risk is zero; at $P = I$ the transfer discrepancy in Result 3 is zero. These bounds concern ideal unit-normalized cosine loss, as does Result 2. The additive stabilizer's gradient was treated separately above; no exact optimum for the stabilized training objective is asserted here.

The accompanying CPU check evaluates 360 finite-distribution/feature cases, combining three mean scales, four ranks (including zero and full rank), ten random distributions, and three feature partitions. It checks both arbitrary and optimal unit predictors, the risk-transfer inequalities, the residual bound, and the constant-predictor bound. All inputs are synthetic, and these checks supply no estimate of the conditional discrepancy or mean dominance of the trained JiT features.

A counterexample to a variance-only ordering. Let independent S, T be equally likely -1 or $+1$, and set $z = (AS, bT)$ with $A > b > 0$. The mean is zero, $C = \mathrm{diag}(A^2, b^2)$, and suppose $H = T$. The rank-one PCA target is $(AS, 0)$; its normalized direction is independent of H with conditional mean zero, giving optimal expected loss one. The lower-variance target $(0, bT)$ has normalized direction $(0, T)$, known exactly from H, giving optimal loss zero. For the full target the optimum is $1 - b/\sqrt{A^2 + b^2}$. Changing the available information to $H = S$ reverses the two rank-one optima while leaving the covariance unchanged.

For $A = 3, b = 1$, PCA and the other direction retain 90% and 10% of centered variance. These are analytical example values, not ImageNet measurements. An easier target need not improve denoising: an uninformative constant target can be perfectly predictable. The example only rules out a universal inference from greater variance retention to better directional predictability. It does not model a random rank-32 JiT run.

The local optimization identity. For $v = g_D + \lambda g_A$, Taylor expansion yields $L_D(\theta - \eta v) = L_D(\theta) - \eta\langle g_D, v\rangle + O(\eta^2\|v\|^2)$, assuming a bounded Hessian locally. Holding v fixed gives Eq. (11). The comparison to a denoising-only step is first-order, not a guarantee at an arbitrary finite learning rate. With a fixed symmetric positive-definite preconditioner M, the terms instead involve $g_D^\top M g_D$ and $g_D^\top M g_A$; compatibility may differ from the Euclidean case. Adam also carries evolving moment estimates, so neither simplification proves its long-run behavior. No step maps training loss to FID or identifies the cause of the measured deep-endpoint degradation.

B PROTOCOL, ENDPOINTS, AND ARTIFACT SCOPE

The supplement's `RUN_CONTRACT.yaml` records training defaults and external-asset hashes. `ledger/claim_matrix.csv` maps arms to configurations; the measured release also supplies endpoint records with explicit precision and missing-field status. These are not complete checkpoint provenance: most per-run checkpoint content hashes and some full guidance sweeps are unavailable in the curated bundle. External teacher, PCA, and reference-statistics hashes cannot substitute for per-run checkpoint hashes. Rounded legacy values remain labeled as rounded rather than being padded into purported exact values.

Table 5: Training-seed endpoints used for aggregates. FID-50K is rounded to two decimals; aggregates use the unrounded endpoint logs preserved in the supplement. Blank cells are unmeasured. L11 values use the extended search.

Arm	Seed 42	Seed 43	Seed 44
Vanilla	4.98	4.72	4.70
L1	4.24	4.35	4.06
L3	3.71	3.75	3.63
L11	7.46	7.93	—
Top-PCA rank 32	3.84	3.90	—
Random rank 32	3.84	—	—

Recovery endpoints. Recovered vanilla steps are 250,200/250,200/250,652 for seeds 42/43/44; L3 steps are 251,443/251,259/250,677. Documented state steps are 250,262 and 251,362 for L1 seeds 43/44; 250,677 for L3 seed 44; 251,124 and 250,220 for L11 seeds 42/43; 250,532 for rank 32 seed 43; and 250,590 for the VICReg-style endpoint. Top-PCA rank 32 seed 42 and random rank 32 both end at 250,200. The VICReg-style continuation replays 390 steps. These are close to the nominal 250,200-step budget but not exactly equal. Missing step records for other arms are identified in the endpoint ledger rather than assumed to be 250,200.

Search and device exceptions. L3 seed 44 was evaluated at CFG 1–5 on H200, selecting CFG 4 at FID 3.64. The selected CFG alone was re-evaluated on TPU from the same committed checkpoint, giving 3.63, a difference of 0.02. The confirmation changes neither the selected CFG nor the three paired improvement directions, but it does not measure a full TPU search grid. The vanilla seed-44 checkpoint was trained with seed 44, but its evaluation config uses seed 42; therefore training-seed pairing is not sample pairing. Both L11 extended-search records use clipping enabled; the seed-42 common-grid record used clipping disabled. This distinction is retained in the ledger. The seed-42 extension has an interior minimum that agrees with the original endpoint, but does not isolate the search-range change from clipping. The separate vanilla clipping check in Appendix E is not a proof of clipping invariance for every arm.

Table 6: Archived evaluation sweeps, FID-50K. Each row is one evaluation setting, not another training seed. L11 uses CFG 4–8; the other rows use CFG 1–5. All FID values are displayed to two decimals; the supplement preserves the available archive precision.

Arm	CFG grid	FID in increasing CFG order				
L11 seed 42	4–8	7.66	7.46	8.36	10.24	13.35
L11 seed 43	4–8	8.62	7.93	8.29	9.42	11.46
Rank 32 seed 43	1–5	28.24	7.45	3.95	3.90	4.49
Random rank 32	1–5	28.42	7.83	4.13	3.84	4.25
VICReg-style	1–5	37.11	11.64	6.08	5.49	6.04
L3 seed 44 (GPU)	1–5	25.84	7.07	3.83	3.64	4.09

Released-code control. The anonymous patch pins PixelREPA commit `5181608` (full revision in the supplement). The included attachment verifier checks depth 12, encoder boundary 4, and

that the adapter and patched MLP consume the same encoder tensor. This is a check of the pinned release, not a claim that the current upstream tree or the unavailable failing branch is identical.

C GUIDANCE SENSITIVITY AND PAIRED SUMMARIES

The supplemental script `tools/analyze_guidance.py` reads the exact archived `ledger/measured_endpoints.json` and generates both Table 2 and `results/guidance_sensitivity.json`. It uses CFG 2–5, all guided values in the original integer grid; CFG 1 is unguided. Both complete training-seed 42/43 sweeps use clipping disabled. We exclude seed 44 from the complete-grid comparison because its selected-CFG TPU confirmation does not supply the other TPU grid points.

For each fixed CFG c and training seed s, the paired reduction is $\Delta_s(c) = \mathrm{FID}_{\mathrm{vanilla},s}(c) - \mathrm{FID}_{\mathrm{L3},s}(c)$. Means average training seeds, not CFG values. The three-seed paired summary instead uses each arm's separately selected endpoint, with the seed-44 device exception stated above. The leave-one-seed-out range is 1.03–1.17 FID, computed from the three two-seed mean reductions; it is neither a confidence interval nor an additional experiment. These reanalyses were selected after inspecting existing results and should be interpreted as descriptive sensitivity checks. For the specific exclusion of seed 42, the baseline mean is 4.71 and the L3 mean is 3.69; computing their relative difference before rounding gives 21.8%. The script records this subset explicitly. It also checks a fixed-configuration comparison at CFG 3, selected by both complete-sweep vanilla runs: the two paired L3 improvements are 0.97 and 0.71 FID. This uses the existing evaluations, with no new selection holdout.

For low-rank targets, the fixed-CFG retained-gain ratio replaces each searched FID in G by its value at c, using the same two seeds in all terms. FID ratios need not be bounded by one and have no direct information-theoretic interpretation. The random-minus-PCA comparison uses only training seed 42 and basis seed 0, preserving its accelerator-generation and replication limitations. No additional trained model or generated sample is introduced.

Table 7: Common-guidance sensitivity from archived TPU sweeps (training seeds 42/43). Both arms use the same CFG in each row, with clipping disabled. Means are over two training seeds; reductions show each pair separately. These post-hoc evaluations reuse checkpoints and are not eight independent training replicas.

CFG	Vanilla mean	L3 mean	Reduction, seed 42	Reduction, seed 43
2	8.83	7.41	1.96	0.88
3	4.85	4.01	0.97	0.71
4	5.06	3.73	1.29	1.37
5	5.98	4.13	1.69	2.01

D TRAINING CHECKPOINTS AND SECONDARY CONTROLS

Table 8: Secondary within-stack comparisons, one paired seed per setting. Fixed and searched columns are separate evaluation policies; a discrete search need not beat an off-grid fixed value. REPA uses L3 for B/16 and index 6 for L/16.

Setting	Fixed CFG		S1–5	
	Vanilla	REPA	Vanilla	REPA
B/16, 200 epochs (F3.1)	4.89	3.90	4.98	3.71
B/16, 600 epochs (F3.1)	3.89	3.13	3.81	3.19
L/16, 200 epochs (F2.8)	3.14	2.32	3.19	2.37

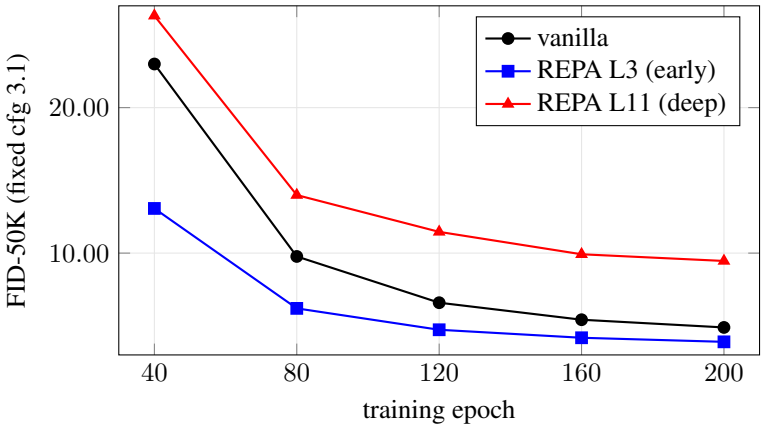


Figure 2: Seed-42 training checkpoints at fixed CFG 3.1, separate from the searched endpoint comparisons. L3 improves over vanilla at each measured checkpoint; L11 is worse already at epoch 40. These observations do not demonstrate an early benefit followed by reversal.

E Absolute calibration checks

Published absolute FID values are not matched controls for this stack. Comparing reference statistics found small differences (relative mean-norm difference 0.3% and covariance difference 0.5%), insufficient to account for the observed gap in the recorded check. Enabling predicted-x_0 clipping for the same vanilla checkpoint changes FID from 4.98 to 4.97. Rerunning with the reference implementation's recipe (including no warmup and its Euler-100 setting) gives 4.67, versus 4.98 under the main recipe. This joint recipe change does not isolate any individual ingredient. The residual to a historical reference value 4.34 is 0.33, similar in scale to the 0.28 range over our three vanilla seeds, but similarity of scale does not identify seed variation as its cause. The reference checkpoint is no longer available for direct evaluation. These checks support using within-stack differences and leaving cross-stack offsets unresolved.

F Exploratory diagnostics: domains and sample scope

Generated-image probes. Spectral and distribution probes use cached final generated images at fixed CFG 3.1: 560 per arm, with 2,000 real center crops. The real loader takes ten images from each of the first 200 sorted ImageNet training-class directories, not a balanced sample of all 1,000 classes. Cached Inception features have shape $(560, 2048)$ per generated arm and $(2000, 2048)$ for real images; logits have shape $(560, 1008)$. These limited diagnostic samples are entirely separate from the headline FID-50K evaluation.

Table 9: Exploratory generated-image statistics (560 images per arm, fixed CFG 3.1). Precision/recall use a 2,000-image, 200-class real subset and $k = 3$ neighborhoods. IS averages ten splits of 56 images. Predicted-class counts use the 1008-output classifier. These estimates are not FID-50K companion metrics or evidence excluding mode collapse.

Arm	IS	Precision	Recall	Classes	Confidence	Agreement
Vanilla	34.17	0.227	0.750	423	0.761	—
L3	35.72	0.195	0.7795	431	0.781	0.7875
L11	30.77	0.186	0.7615	426	0.704	0.716

IS follows Salimans et al. (2016); precision/recall follow the neighborhood construction of Kynkäänniemi et al. (2019). Confidence is mean maximum softmax probability. Agreement is predicted-class agreement with the paired vanilla image, not accuracy against the conditioning label. Matched-index feature cosine (0.78–0.80) exceeds a shuffled comparison (0.43–0.45), supporting the recorded shared class/noise pairing. Lower confidence and agreement for L11 are descriptive

correlations. The small, class-restricted real subset and limited generated sample prohibit conclusions about population coverage, diversity, or a causal semantic failure mechanism.

Radially averaged power spectra in the upper half of the frequency range are -1.39 dB for L11 and -1.51 dB for vanilla relative to real images. The deep arm does not exhibit the predicted reduction in this aggregate high-frequency statistic; extra high-frequency power can also reflect artifacts and does not establish fidelity.

Conditioning probes on noised real images. These probes evaluate denoiser responses on perturbed real inputs, not generated images or sampler-visited trajectories. The relative conditional/unconditional prediction gap is 0.0297 for L11 versus 0.0213 for vanilla, and its cosine with the ground-truth update direction is 0.0169 versus 0.0075. Thus the recorded statistics do not support a smaller gap or lower cosine in this evaluation domain. They do not rule out conditioning errors along sampling trajectories, establish a cause of FID degradation, or validate the published baseline's proposed mechanism.

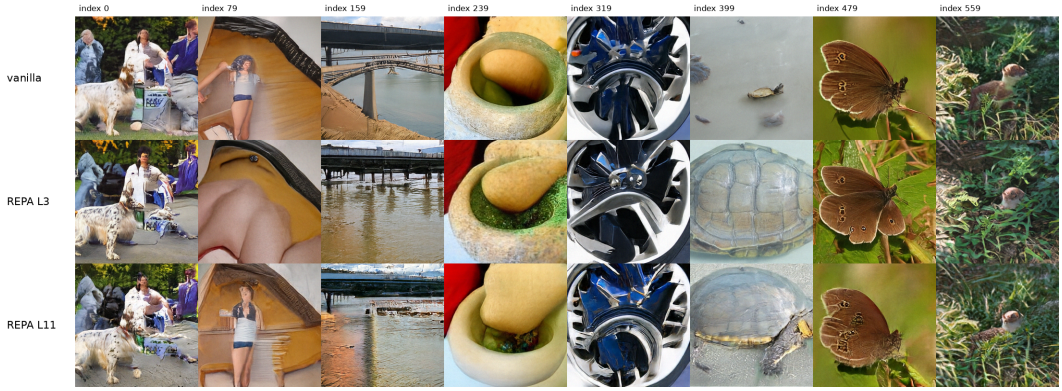


Figure 3: Paired examples from the 560-image diagnostic caches. Each column shares class and initial noise across vanilla, L3, and L11. Eight indices were set by evenly spaced selection from 0–559 before visual inspection. These 24 images illustrate the cache and do not substitute for aggregate FID-50K or establish a failure mechanism.