THE FID LOTTERY: QUANTIFYING HIDDEN RANDOMNESS IN GENERATIVE MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Fréchet Inception Distance (FID) is the de facto arbiter of image generation, yet most papers report just a single number from a single trained model using a single sampling seed. How reproducible is that number if we retrain the model, or merely resample from it? In this paper, we treat FID as a random variable on a two-axis panel of training and generation seeds, and measure its variance directly on several hundred SiT networks trained on class-conditional ImageNet 256×256. We report surprising findings: (a) Retraining the model using the same recipe with a different seed moves FID $3.2\times$ more (in Inception feature space) than redrawing samples from a fixed network. (b) That gap is driven by three factors: random initialisation, data ordering, and the per-step Gaussian noise of the flow-matching loss. (c) Increasing compute or model size barely tightens the spread, holding the FID coefficient of variation (CoV) inside a 1–2% band. (d) Per-cell classifier-free-guidance tuning halves the spread but perturbs which seeds rank best, and a lucky training seed reaches the same FID with up to $2\times$ less compute than an unlucky one. (e) The lottery extends beyond ImageNet and Inception features: on a text-to-image model retrained from 10 seeds and scored under 10 evaluation seeds each, the training seed carries 76% of the GenEval and 85% of the DPG-Bench variance. Based on these findings, we recommend a new FID evaluation protocol: evaluate under per-cell optimal guidance, treat any gap between two single runs below $\approx4\%$ of the baseline FID as inconclusive (the threshold implied by the $\approx1.3\%$ CoV floor we measure for SiT/DiT-family latent flow matching), and report an error bar over several training seeds rather than a single FID number.

1 INTRODUCTION

> *"If the lottery is an intensification of chance, a periodic infusion of chaos into the cosmos, would it not be desirable for chance to intervene at all stages of the lottery and not merely in the drawing?"*
>
> — Jorge Luis Borges, *The Lottery in Babylon*

Open any recent image generation paper and the central claim usually rests on a single trained model, scored with a single number, the Fréchet Inception Distance (FID). FID has the reputation of usually (though not always) correlating well with perceived visual quality, so it has become the closest the image generation community has to an universal arbiter. A decade of recipes has been justified by single-FID-unit gains, and budgets in the low millions of GPU-hours hinge on which architecture lands a few decimals lower. But behind every trained model sits a chain of pseudo-random draws (parameter initialisation, minibatch order, per-step Gaussian noise injected by the training loss, hardware stochasticity and the initial noise drawn at sampling time), any of which could have produced a different model and score had the seed been different. Conventional wisdom considers this seed-to-seed variance to be negligible, especially for well-trained models. In this paper we show that this reproducibility gap of generative models is real and a serious concern.

Each time one trains and evaluates a generative model, two lotteries are played (Figure 1). The *training lottery* runs once during training: it draws an initialisation, a data ordering, and the per-step noise that the loss injects at every gradient step, and what comes out is one trained network among many that the same recipe could have produced. The *generation lottery* runs on the trained network: one draws an initial noise $x_T \sim \mathcal{N}(0, I)$ to seed the sampler, generates a sample set, and scores

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.

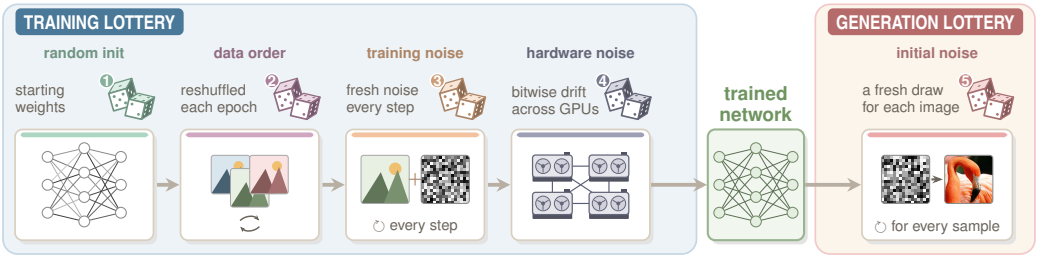

Figure 1: **All sources of randomness behind a generative model.** Training and then sampling a generative model is a chain of pseudo-random draws. They fall into two lotteries. The *training lottery* (left) is drawn from four sources: the random weight initialisation ❶, the order in which examples are visited ❷, the fresh Gaussian noise the flow-matching loss injects at every gradient step ❸, and the bitwise non-determinism of multi-GPU execution ❹. This draws a single trained network out of the many networks the same recipe could have produced. The *generation lottery* (right) then draws a fresh initial noise $x_T \sim \mathcal{N}(0, I)$ for every sampled image ❺. Common practice accounts for only ❺, reporting error bars over sampling seeds on one fixed network. In this work we study all five.

it. Practitioners have learned to mitigate the second lottery by redrawing the initial noise across several seeds and reporting an error bar or an averaged score (Chong & Forsyth, 2020; Parmar et al., 2022). However, no amount of resampling a single trained network says anything about where a retrained network would have landed. The training variability stays hidden behind the one ticket we actually drew. Diffusion makes the problem worse: the flow-matching (Lipman et al., 2023) or score-matching (Song et al., 2021b) loss redraws a fresh Gaussian $\varepsilon \sim \mathcal{N}(0, I)$ at every gradient step, a permanent random injection that an independent run resolves differently, not transient noise that longer training averages out. Nor does scale help by default: neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) describe how the mean loss falls with parameters and tokens, not the seed-induced *spread around* that mean. Zhang et al. (2024) recently showed that independently trained diffusion networks converge to nearly the same noise-to-image mapping. But do nearly identical mappings also yield models of the same quality?

The two lotteries define two axes of variation in model quality: a *training axis* of N independent training runs and a *generation axis* of K sampling seeds per run. To measure both, we score the resulting $N \times K$ panel of FID evaluations. Figure 2 renders the panel for a converged SiT-B/2: every violin is one trained network, every dot is one FID evaluation, and the spread along the training axis visibly exceeds the spread along the generation axis. On this panel we first decompose the training lottery into its independent random sources, then sweep three practitioner-controlled axes (classifier-free guidance (Ho & Salimans, 2022), compute, and model size) to test whether any of them tightens it. Across several hundred SiT networks from S through XL on ImageNet 256×256, the training axis dominates the generation axis at every scale we probe and none of the three knobs closes the gap. A comparison between two single training runs therefore sits on a noise floor that "SoTA" improvements regularly fall below.

This noise floor lives in the models themselves. A training recipe defines a distribution of models, and a paper reports a single draw from it. We measure this distribution directly: how much the quality of a trained generative model depends on its seed. FID is our main instrument, as it is where the field selects models, and the same picture emerges under DINOv2 FID, a sliced Wasserstein distance, and two text-to-image benchmarks scored by detectors and vision-language models (Sec. 4.3).

Main Findings:

- **Retraining moves FID $3.2\times$ more than resampling.** Treating the quality of a trained model as a random variable over training and sampling seeds, we score several hundred SiT networks. The relative spread stays in a 1–2% band across model sizes and training budgets. This floor makes any gap below $\approx 4\%$ of FID between two single training runs inconclusive (Sec. 4.1, Sec. 4.5).
- **The lottery comes from the intended random draws, not from hardware.** The per-step flow-matching noise is the largest single source, initialisation and data order add comparable amounts, and bitwise hardware non-determinism alone barely moves FID (Sec. 4.2).

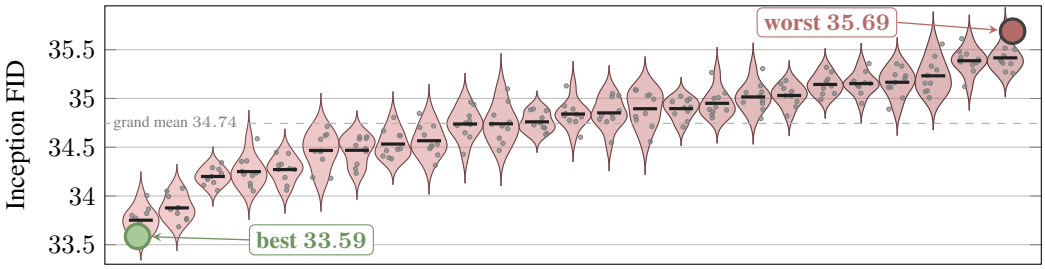

Figure 2: **The FID lottery in SiT-B/2 at 400k steps.** Each violin is one of 25 independently trained SiT-B/2 models, sorted by per-seed mean Inception FID. Small dots are the 250 individual sampling-seed evaluations and short black ticks are per-seed means. The two highlighted markers pick out the single best (**33.59**) and worst (**35.69**) FID across the panel, a **2.10**-point gap produced purely by changing seeds. Each violin is short ($\sigma_{\text{within}} \approx 0.14$, generation lottery). The per-seed means stagger over a $3\times$ wider range ($\sigma_{\text{between}} = 0.44$, training lottery).

- **The lottery is a property of the models, not of FID.** The asymmetry holds under DINOv2 FID and a non-Gaussian distance, and on a text-to-image model retrained from 10 seeds the training seed carries 76% of the GenEval and 85% of the DPG-Bench variance (Sec. 4.3).
- **Per-cell optimal guidance halves the noise floor but perturbs rankings.** Scoring every (training, sampling) seed pair at its own optimal classifier-free guidance (CFG) scale halves the relative spread, but the best unguided seed is not reliably the best guided one (Sec. 4.4). A golden-section search finds this scale in a logarithmic number of FID evaluations (Appendix J).
- **A lucky seed saves up to half the compute.** The luckiest training seed reaches the FID of the unluckiest with up to $2\times$ less compute (Sec. 4.6, Figure 8).

2 Related Work

FID and its limitations. FID (Heusel et al., 2017) compares Gaussian moments of Inception-V3 (Szegedy et al., 2016) features and remains the dominant image-generation metric (Borji, 2022) despite known flaws: it is a biased finite-sample estimator whose rankings can flip under sampling noise (Chong & Forsyth, 2020), it is sensitive to image processing (Parmar et al., 2022) and to ImageNet class information (Kynkäänniemi et al., 2023), and it disagrees with human raters on diffusion models (Stein et al., 2023). Alternatives change the distance or the feature space (Bińkowski et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020; Oquab et al., 2024; Jayasumana et al., 2024). These works study FID for a *fixed* model. We study how it varies when the model is retrained.

Seed variance and reproducibility. Run-to-run variance is well documented in reinforcement learning (Henderson et al., 2018), NLP (Reimers & Gurevych, 2017; Dodge et al., 2020) and vision (Picard, 2021; Summers & Dinneen, 2021), and Bouthillier et al. (2021) decompose benchmark variance into data sampling, initialisation, data order and more. Evidence for generative models is scarce: Lucic et al. (2018) find that equalised hyperparameter budgets and random restarts erase most reported GAN gains. We adopt the statistical toolkit of this literature (random-effects models, multiplicity control, meta-analysis; Demšar, 2006; Dror et al., 2018; DerSimonian & Laird, 1986).

Closest work. Bench & Thomas (2025) obtain a distribution over FID-like scores by Monte Carlo dropout in the feature extractor, which measures the uncertainty of the evaluator rather than of the generator, and Banerjee et al. (2024) test for seed-induced variability with a Kolmogorov–Smirnov statistic on discriminative models. Zhang et al. (2024) show that independently trained diffusion models learn nearly the same noise-to-image map, and Xu et al. (2025) study the inference-time seed of text-to-image models. None varies training seeds, architectures and checkpoints jointly, or separates the training lottery from the generation lottery. We show that near-identical noise-to-image maps still yield percent-level FID fluctuations (details in Appendix B).

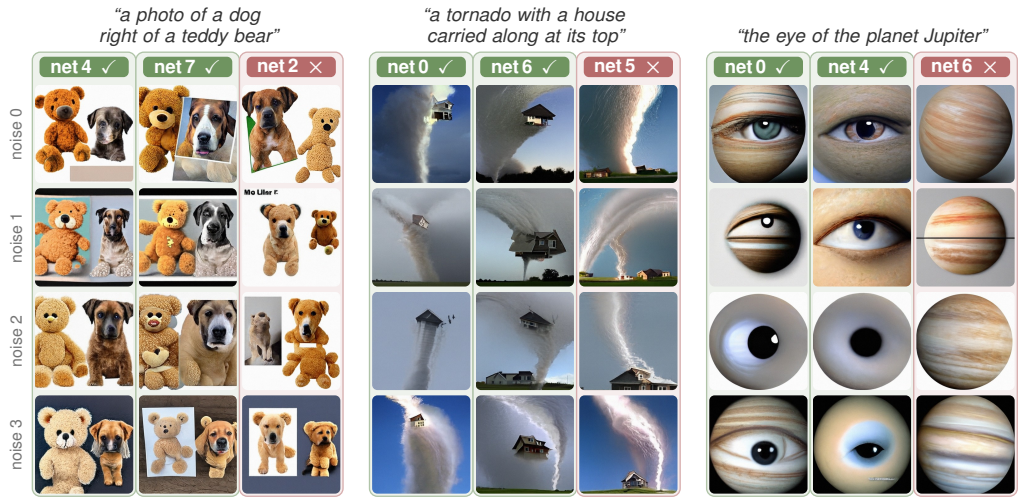

Figure 3: **Some training seeds follow a prompt from every noise, others never do.** Per prompt, three of the 10 text-to-image models of Sec. 4.3, which differ only in their training seed, sampled from the same four noises (rows; CFG 7, 5 for Jupiter). Green columns succeed on all four noises (dog on the right, house lifted by the tornado, eye on Jupiter), red columns on none (net 2 never puts the dog right of the bear). Success is prompt-specific: net 6 lifts the house but never draws the eye.

3 Experimental Setup

All experiments in Sec. 4 share one setting: an $N \times K$ panel of FID evaluations, the random number generators that fill it, and three statistics that summarise it.

Experimental Setting. All experiments train Scalable Interpolant Transformers (SiT) (Ma et al., 2024) at four widths (S/2, B/2, L/2, XL/2) on class-conditional ImageNet 256×256 (Deng et al., 2009) under latent conditional flow matching (Lipman et al., 2023). The loss redraws a Gaussian $\varepsilon \sim \mathcal{N}(0, I)$ at every gradient step, one of three training-time random number generators (Sec. 4.2). FID compares 50k generated samples with the ImageNet training-set statistics in Inception-V3 feature space (Szegedy et al., 2016; Heusel et al., 2017), on one shared pipeline. As a Fréchet distance between feature *distributions*, not a per-image score, it can shift on differences too small to perceive (Figure 27 in Appendix L). Sampling uses a fixed deterministic ODE solver and step count. Classifier-free guidance (Ho & Salimans, 2022) is off by default, and *guided* metrics (Sec. 4.4) use the per-(training, sampling)-seed optimal CFG scale. Per-experiment N, K and step counts are in Appendix C.

The two-axis panel. Every §4 experiment produces an $N \times K$ panel: N independently trained models (*training seeds*) and, for each, K generations under different *sampling seeds*. A training seed drives parameter initialisation, data-loader order, and the per-step flow-matching noise. A sampling seed drives the initial noise drawn at generation time. Crossing the two seeds lets a random-effects model, $\mathrm{FID}_{ij} = \mu + b_i + e_{ij}$ with a training-seed effect b_i and a sampling residual e_{ij}, separate the two lotteries (Appendix E).

Notation. Three statistics summarise a panel. σ_{within}, the spread across the K sampling evaluations of one training seed (averaged over seeds), measures FID noise for a fixed model (the *generation lottery*). σ_{between}, the spread of the N per-seed means, measures the *training lottery*. The coefficient of variation $\mathrm{CoV} = \sigma/\mu$ (%) compares panels whose FID differs tenfold (e.g. unguided vs. guided). Statistical tests (random-effects ANOVA, intervals, multiplicity control, meta-analysis) are in Appendix E.

Other metrics and a second task. The same panels are rescored with DINOv2 FID and Inception precision/recall/density/coverage (definitions in Sec. I.1). As an out-of-family check we also retrain a 260M-parameter text-to-image flow-matching transformer with JiT-style blocks (Li & He, 2025) in pixel space (512^2, CC12M (Changpinyo et al., 2021)) from 10 seeds for 250k steps (≈4k H100-hours) and score it with GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024) (samples in Figure 3, details in Appendix H).

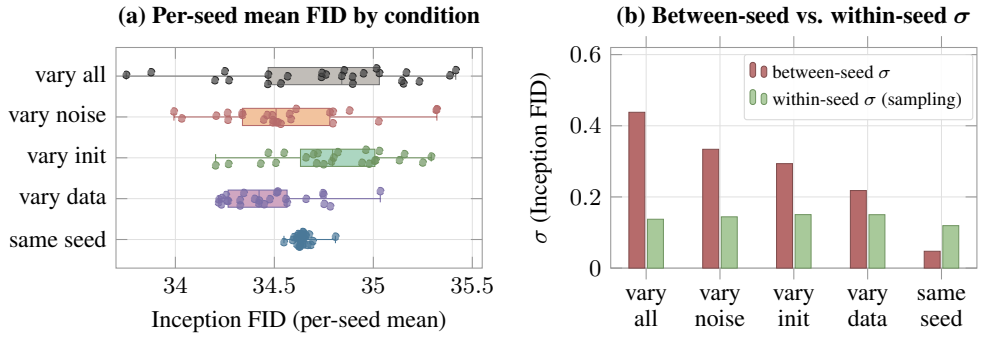

Figure 4: **Variance decomposition of the training-seed lottery (SiT-B/2, 400k, no CFG).** (a) Per-seed mean Inception FID under each single-source condition, the fully random baseline (*vary all*) and the *same seed* control. Dots are training seeds, boxes the 25/50/75 percentiles. (b) Between-seed σ (coral) and within-seed sampling σ (sage). Between-seed σ decreases monotonically over the four random conditions while sampling σ stays flat. The *same seed* control, with only multi-GPU non-determinism left, falls below the sampling floor.

4 Experiments

We establish the asymmetry between the two lotteries (Sec. 4.1), trace it to its sources (Sec. 4.2), test it beyond Inception FID (Sec. 4.3), then turn to guidance, compute and scale (Sec. 4.4–4.6). The appendix adds the μP sweep (Appendix K) and two seed-ranking analyses (Sec. G.2, G.3).

4.1 Training Variability Dominates Evaluation Variability

Takeaway

Resampling a fixed model gives the small error bar. Retraining the same recipe moves FID 3.2× more than redrawing samples, so most of the variance hides in which training run you drew.

The training lottery has a 3.2× larger effect than the generation lottery. On the converged SiT-B/2 panel of Figure 2 ($N = 25$ training seeds, $K = 10$ sampling seeds, 400k steps, no CFG), the spread across columns (training seeds) visibly exceeds the spread within a column (sampling seeds). The between-seed $\sigma_{\text{between}} = 0.438$ (CoV $\approx 1.3\%$) is 3.2× the within-seed $\sigma_{\text{within}} = 0.137$ (CoV $\approx 0.4\%$), and per-seed means span $33.75 \to 35.42$ around a grand mean of 34.74. An exact random-effects F-test rejects $\sigma_b \le \sigma_w$ at $p = 1.5 \times 10^{-21}$ (Appendix E). The variance a benchmark cares about lives in *which model* was trained, not in *which samples* are drawn from it.

Sampling-seed confidence intervals (CIs) misreport the spread. The 25 per-seed means give a 95% Student's-t interval of 34.74 ± 0.18 for the grand mean, but a single run has a one-σ spread of 0.44 FID, more than the headline gain of many recent papers. Ten times more samples on one run shrink the within-seed jitter by $\sqrt{10} \approx 3.2$ but leave this 0.44 envelope untouched. Only more training seeds reduce the dominant source of FID variance. Comparing two single runs doubles this variance: a gap must exceed $\approx 4\%$ of the baseline FID to keep the false-positive rate at 5% (Appendix F).

4.2 Flow-Matching Noise Leads Init and Data Order

Takeaway

The per-step noise of the flow-matching loss is the largest single source of training variance, but init and data order add comparable amounts, and the three overlap rather than sum.

Three training-time sources. A SiT run draws from three independent generators: initialisation, data-loader order, and the per-step flow-matching noise ε (*training noise*). Each condition fixes two and varies the third (VARY-NOISE/INIT/DATA), and VARY-ALL is the panel of Sec. 4.1.

Per-source contributions. Between-seed σ orders monotonically (Figure 4b): VARY-ALL (0.438) > VARY-NOISE (0.334) > VARY-INIT (0.294) > VARY-DATA (0.218). Noise alone reproduces 76% of

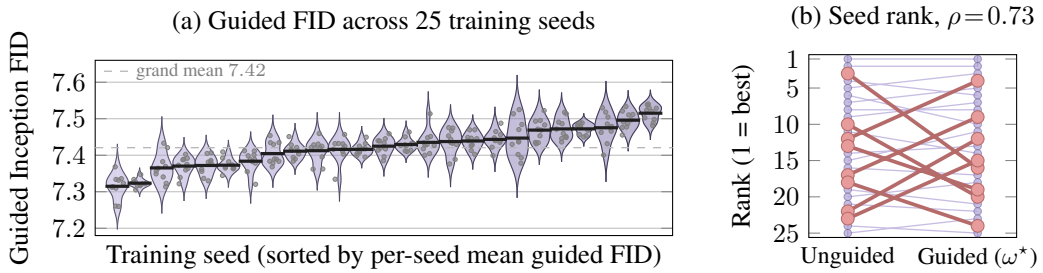

Figure 5: **Per-cell guidance tuning halves the relative seed-induced FID spread, but perturbs the seed ranking near the top.** Guided FID (FID at the per-(training, sampling)-seed optimal CFG scale) of 25 SiT-B/2 training seeds (400k steps, 10 sampling seeds each). (a) Per-seed violins, sorted by mean: CoV $= 0.67\%$, about half the 1.26% unguided. (b) Unguided and guided ranks correlate at Spearman $\rho = 0.73$, but the $8/25$ coral seeds move by $|\Delta\text{rank}| \geq 5$ (lavender: < 5).

the baseline σ, init 67%, and data order 50%. Since the sources do not add (below), each VARY-X condition only upper-bounds the share of X, and $76\% > 67\%$ does not rule out init, or its interactions, as a major factor. The ordering still contradicts the informal "different seeds mean different inits": init is not the largest single source. The within-seed σ stays flat across conditions (0.137–0.150, sage bars), so each σ_{between} measures the spread of the trained models, not of the scoring.

The data lottery is shape-different, not just smaller. VARY-DATA has a tight bulk and a long upper tail (skew $+0.74$), unlike the symmetric VARY-INIT (-0.24) and the broader VARY-NOISE ($+0.62$, Figure 4a). With init and noise fixed, most runs agree and a few resemble training failures.

The sources combine sub-additively. The quadrature sum $\sqrt{\sigma^2_{\text{noise}} + \sigma^2_{\text{init}} + \sigma^2_{\text{data}}} \approx 0.50$ overshoots the observed $\sigma_{\text{vary-all}} = 0.44$ by 13%. The sources share variance through the trained weights, so one-at-a-time ablations overestimate how much variance fixing any single source recovers.

Numerical non-determinism is not the source. A SAME-SEED control fixes init, data order *and* training noise across $N = 24$ retrains, leaving only the bitwise non-determinism of multi-GPU (DDP) execution, a floating-point reduction-order effect known to perturb training (Summers & Dinneen, 2021) and inference (He & Thinking Machines Lab, 2025). The EMA weights end $\approx$5–6% of their norm apart (one run, 33%), yet FID barely moves: the between-seed $\sigma = 0.047$ falls *below* the within-seed floor (0.119), a ratio of $0.4\times$ instead of $3.2\times$ (Figure 4b, right). The lottery comes from the *intended* draws of init, data and noise, not from numerical noise.

4.3 BEYOND INCEPTION FID

Takeaway

On a text-to-image model, the training seed carries 76–85% of the GenEval and DPG-Bench variance, without Inception features. On ImageNet, DINOv2 FID is even more training-dominated than Inception FID, and MIND, a sliced Wasserstein distance, shows the same asymmetry.

A second task, architecture and metric family. The text-to-image models of Sec. 3 are scored with GenEval (object-detector based) and DPG-Bench (VLM based), each under the same 10 evaluation seeds (a crossed 10×10 design, Figure 6a,b). The spread across models ($\sigma_{\text{model}} = 0.013$ GenEval, 0.64 DPG) is $1.8\times$ and $2.4\times$ the sampling spread within a model: the training seed carries 76% and 85% of the variance, the shared evaluation seed only 4% and 2% (samples in Figure 3). Averaging 10 evaluation seeds thus narrows the 2σ band of one run (± 0.030 GenEval, ± 1.4 DPG) by only 12% and 7%. Compositional GenEval skills are the most training-dominated (position 83%, simple skills 26–33%, Appendix H). As on ImageNet, a good seed does not transfer: the rankings of the three metrics are uncorrelated (all $|\rho| \leq 0.47$, n.s.), and FID rankings do not persist across checkpoints. MS-COCO FID ends at a CoV of 1.4%, inside the ImageNet band (Figure 6c). This is an existence check across architecture, task, data and metric family, not a claim that ImageNet constants transfer.

Other metrics on ImageNet. The balance between the two lotteries is metric-specific (Appendix I). DINOv2 FID is more training-dominated than Inception FID ($\sigma_{\text{between}}/\sigma_{\text{within}} = 4.8\times$), the real-

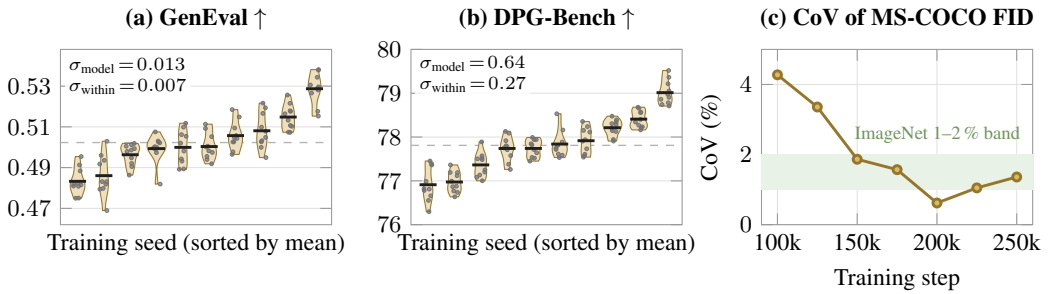

Figure 6: **The training lottery on a text-to-image model.** Ten 260M-parameter flow-matching transformers (512^2 pixels, CC12M) that differ only in their seed. (a, b) GenEval and DPG-Bench at 250k steps: each violin is one model scored under 10 shared evaluation seeds (dots), ticks are means. As for FID on ImageNet (Figure 2), violins are short and means stagger: the training seed carries 76% and 85% of the variance. (c) CoV of MS-COCO FID across the 10 seeds falls from 4.3% at 100k to 0.6–1.4%, near the ImageNet band. The early decrease plausibly reflects the CC12M/MS-COCO gap. This online FID mixes both lotteries (no within-model control).

anchored k-NN metrics sit near parity (1.1–1.4×), and recall, whose k-NN balls are redrawn with every sampling seed, inverts (0.28×). We therefore restrict "training-dominated" to the Fréchet metrics. This is not an artefact of the Gaussian approximation of FID: MIND (Berthet et al., 2026), a sliced Wasserstein-2 distance on the same images, gives ratios of 2.5× (Inception) and 4.5× (DINOv2), a between-seed CoV of 1.1–1.5%, and the same seed ranking as FID ($\rho \geq 0.99$). At 5k instead of 50k images, sampling noise buries the training lottery (Sec. I.7).

4.4 Guided FID Halves the Floor but Perturbs Rankings

Takeaway

Tuning CFG per seed nearly halves the relative FID spread (CoV : 1.26% → 0.67%) but perturbs the seed ranking ($\rho = 0.73$): the best unguided seed is not reliably the best guided one.

Guided FID. We score every (training, sampling) seed pair at its own FID-optimal CFG scale $\omega^\star$, instead of one shared, hand-picked scale, and call the result *guided FID* (other guided metrics in Sec. I.4). This mirrors practice, where each paper tunes the scale for the one model it reports. As a side contribution, a golden-section search finds $\omega^\star$ at a cost logarithmic in the required precision, instead of linear for a dense sweep (procedure, convergence and unimodality of $\text{FID}(\omega)$ in Appendix J). The numbers below are search minima on the samples that selected $\omega^\star$, so they are optimistically biased. Our recommended protocol re-evaluates at $\omega^\star$ on fresh noise (Sec. J.1).

Guided FID halves the relative noise floor, but sampling jitter takes a larger share. Per-seed mean guided FID spans $[7.31, 7.52]$ around 7.42, with $\sigma_{\text{between}} = 0.050$ and $\sigma_{\text{within}} = 0.027$ (Figure 5a). The grand mean itself falls from ≈34.7 to ≈7.4, so absolute ranges overstate the gain. The dimensionless CoV is the fair comparison: it drops to 0.67%, half the 1.26% unguided. This floor still sits above the ≈0.4% pure-sampling floor, and the between-to-within ratio falls only from 3.2× to 1.87×. Tuning does not remove the seed lottery, and sampling jitter takes a larger share of what remains, so multi-sampling-seed reporting matters *more* under guided FID. The optima concentrate tightly ($\sigma_\omega \approx 0.045$), so a $\pm\mathbf{0.05}$ error on the scale injects FID noise comparable to the within-seed floor: any CFG number must come with its search tolerance.

Guidance tuning perturbs the seed ranking. Under mean, min and median criteria, Spearman ρ between the unguided and guided rankings of the 25 seeds is 0.73 ($p < 10^{-4}$), a strong global correlation. Yet $8/25$ seeds move by at least five places (Figure 5b), and the top of the leaderboard mixes seeds that unguided FID placed mid-table. A model selected on unguided FID is thus not guaranteed to be best under guided FID, and the two protocols should not be compared across papers. Guided FID also changes the estimand to the best FID of each model under guidance: its tighter spread is a property of that protocol, not a more precise estimate of fixed-ω FID.

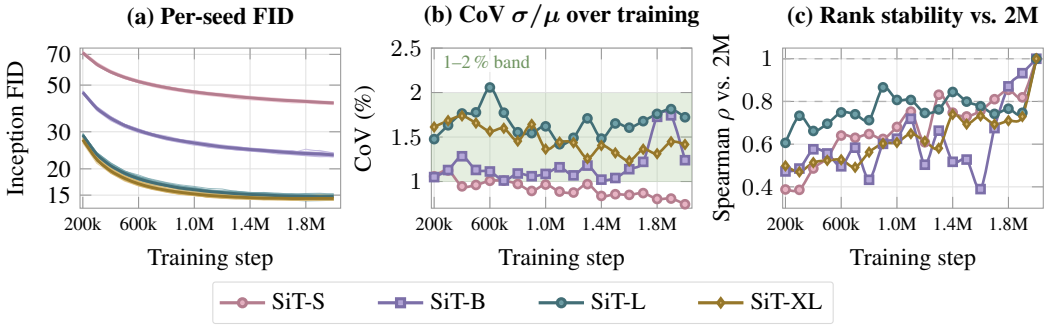

Figure 7: **The seed lottery across compute and model size.** (a) Inception FID over training: thin lines are training seeds, bold lines per-step means. The spread stays wide as training converges. (b) CoV σ/μ: all four models stay near a 1–2% band, and bigger models are not tighter. (c) Spearman ρ between the seed ranking at step t and at 2M: below 0.9 until 1.8M.

4.5 The 1–2 % CoV Floor Survives Scale and Compute

Takeaway

More compute and larger models do not reduce the variance: the CoV stays in a 1–2% band (median 1.30%, pooled 1.21%) at every checkpoint and size.

The relative floor is scale-invariant. Across the SiT-S/B/L/XL panel ($N = 25$, $K = 10$ at every 100k-step checkpoint to 2M), mean Inception FID drops $\approx 2\times$ from 200k to 2M while σ_{between} shrinks at most $2.4\times$: the seed spread trails the mean down rather than collapsing at convergence (Figure 7a). The CoV stays inside $[0.74\%, 2.06\%]$ across all 76 cells (median 1.30%, Figure 7b), and a random-effects meta-analysis pools it at 1.21% with a 95% prediction interval of $[0.79, 1.84]\%$ for an unseen cell (diverged runs screened, Appendix E). *The FID noise floor is 1–2% of the mean FID the model has reached, at every compute budget and scale on this family.* Appendix F turns this floor into a decision threshold for comparing two runs.

More parameters does not mean less variance. The CoV at 2M is non-monotonic in scale: SiT-S (0.74%) and SiT-B (1.24%) sit *below* SiT-XL (1.42%) and SiT-L (1.72%). This matches Sec. 4.2: the dominant per-step noise is redrawn every batch, so it neither fades with compute nor averages out across width. Reproducibility is a property of the metric and the loss, not of compute or scale.

Rank stability is weakest early in training. Spearman ρ between the seed ranking at step t and at 2M is 0.39–0.61 at 200k and 0.65–0.81 by 1.1M (Figure 7c). The later values are strong correlations, but selecting a seed at 200k and reusing it for a long final run is close to picking at random. The spread is Gaussian, not heavy-tailed: $(\max - \min)/\sigma \in [3.1, 5.0]$ at every step, bracketing the 3.5–3.9 expected for a Gaussian sample of size $n \in [19, 25]$. This licenses the σ-based tools of this paper ($2\sigma/\sqrt{N}$ thresholds, F-tests, Appendix F), which heavy tails would invalidate, and means the spread is the bulk of the distribution, not a few failed runs that a screen could remove.

4.6 The Luck of the Draw: A 1.2–2.0× Convergence Gap

Takeaway

Cherry-picking the luckiest of many training seeds makes training look up to $2.0\times$ faster without changing the recipe, from $1.25\times$ on SiT-S/B to $2.0\times$ on SiT-XL.

An extra 100k training steps and a seed swap move FID by similar amounts. Past $\approx$1.5M steps, the smallest single-seed gain clearing 2σ ($\approx$0.5–0.8 FID at 2M) equals 200–500k extra steps. With N seeds the gap shrinks as $2\sigma/\sqrt{N}$: $N = 5$ brings it to $\approx$0.25 FID, enough to detect 100k-step gains.

The lucky–unlucky gap is 400k–1M steps depending on size. We ask when the luckiest seed first reaches the FID the unluckiest reaches at 2M (Figure 8): $1.25\times$ sooner on SiT-S/B, $1.82\times$ on SiT-L, and $2.0\times$ on SiT-XL, where the lucky seed reaches at 1M what the unlucky one attains at 2M. An

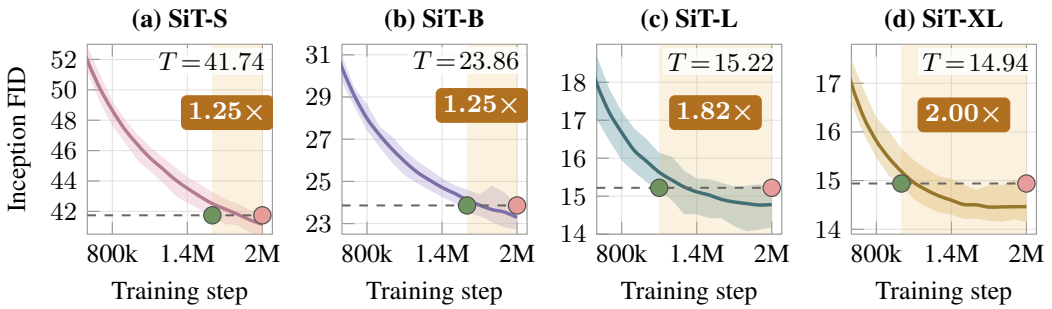

Figure 8: **The luck of the draw: a 1.2–2.0× convergence gap.** The dashed line marks the target T, the FID the *unluckiest* of 19–25 seeds reaches at 2M. The green dot marks when the *luckiest* seed first crosses T, the coral dot the unlucky seed reaching it at 2M, and the amber band the compute wasted in between. Shaded: min–max across seeds. Bold: mean.

unlucky seed thus wastes a fifth to a half of the budget. Any single-seed paper claiming a $\approx 1.3\times$ speedup here competes with what the seed lottery delivers without changing a line of code.

Advice for Practitioners

1. More evaluations cannot replace more training runs. Resampling a fixed network leaves the training variance untouched (Sec. 4.1). Only multi-seed training beats the $\approx 1.3\%$ floor.

2. Treat any gap between two single runs below $\approx 4\%$ as inconclusive. The threshold shrinks as $1/\sqrt{N}$ with N seeds per method (Appendix F).

3. Guided and unguided FID disagree on best seeds and hyperparameters. Seed rankings agree only up to $\rho \approx 0.73$ (Sec. 4.4). On learning rates, guided FID picks a stable optimum and unguided FID the unstable edge of training (Appendix K). Tune with the FID you report.

4. The lottery also holds for text-to-image benchmarks. The training seed carries 76–85% of the GenEval and DPG-Bench variance, and seed rankings transfer neither across checkpoints nor across metrics (Sec. 4.3). Report means over training seeds, not a "good" seed.

5. Report guided FID at the per-cell optimal CFG scale. Golden-section search finds it in a logarithmic number of evaluations (Appendix J). Select $\omega^\star$ on a proxy budget or a dedicated seed, and report a fresh evaluation at $\omega^\star$.

5 Discussion

Limitations. We study only SiT, flow matching, class-conditional ImageNet 256×256 and Inception-V3 FID. As diffusion architectures differ fundamentally, the $\approx 1.3\%$ CoV is a calibration target for SiT/DiT-family latent flow matching, not a universal constant. Other backbones, objectives, pixel diffusion, or Fréchet variants (FDD, CMMD, KID) may sit on different floors, and the lottery balance is metric-specific (Appendix I). The dimensionless machinery (relative flip rates, the $1/\sqrt{N}$ decision threshold) is general, and a lab can estimate its own floor with $N \approx 5$ trainings. The text-to-image study (one configuration, checkpoint and guidance scale) is an existence check, not a calibration. The panel is finite (20–25 training seeds, 10 sampling seeds, up to SiT-XL and 2M steps), and all floors assume non-divergence (8 of 50 SiT-S/B runs diverge and are screened). Without a human study, we cannot say which of these gaps are perceptible, only that such gaps routinely decide ablations.

Outlook. We read these findings less as a verdict on FID than as a proposal for how to measure it: a two-axis panel of training and sampling seeds, summarised by a per-source variance decomposition. The object of study is the reproducibility of a stochastic training recipe, and FID is the instrument because that is where selection happens. A pinned, disclosed guidance search removes one undocumented degree of freedom from guided numbers. Extending the study beyond images is a natural next step. The two-axis panel does not depend on the modality, and could be run on video generators evaluated with FVD (Unterthiner et al., 2018), audio generators with FAD (Kilgour et al., 2019), and text generators with MAUVE (Pillutla et al., 2021) or benchmark scores. Whether the 1–2% floor and the balance between the two lotteries carry over to other training methods, model families and modalities, and whether cheaper proxies can predict them, remain open.

AI USE STATEMENT

We used generative AI tools to assist with experimental design and with monitoring the training and evaluation runs, and to help write and edit the paper. All experimental choices, results and analyses were checked by the authors, and all AI-assisted text was reviewed and revised by the authors. We take full responsibility for the content of this work, including any text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and releases no new dataset. All models are trained on ImageNet (Deng et al., 2009) and CC12M (Changpinyo et al., 2021) under their existing licences. The findings are dual-use in a benign sense: knowing the seed-induced floor saves compute by avoiding sub-floor retraining, but the same numbers reveal how large a headline a few extra seeds can manufacture (Sec. 4.6). We state this temptation explicitly so that it can be priced into peer review.

REPRODUCIBILITY STATEMENT

Every experiment follows the public SiT recipe (Ma et al., 2024); per-experiment panel sizes, training budgets and deviations are listed in Appendix C. The statistical procedures (random-effects ANOVA, exact F-tests, confidence intervals, multiplicity control, meta-analysis and false-positive calibration) are specified in Appendix E, the golden-section protocol and its proofs in Appendix J, the multi-metric definitions in Sec. I.1, and the text-to-image study in Appendix H.

REFERENCES

Sinjini Banerjee, Tim Marrinan, Reilly Cannon, Tony Chiang, and Anand D. Sarwate. Measuring training variability from stochastic optimization using robust nonparametric testing. *arXiv preprint arXiv:2406.08307*, 2024.

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

Ciaran Bench and Spencer Angus Thomas. Quantifying the uncertainty of model-based synthetic image quality metrics. *arXiv preprint arXiv:2504.03623*, 2025.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Quentin Berthet, Yu-Han Wu, Clement Crepy, Romuald Elie, Klaus Greff, and Michael Eli Sander. MIND: Monge inception distance for generative models evaluation. *arXiv preprint arXiv:2605.06797*, 2026.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *6th International Conference on Learning Representations (ICLR)*, 2018.

Ali Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. doi: 10.1016/j.cviu.2018.10.009.

Ali Borji. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. doi: 10.1016/j.cviu.2021.103329.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3568, 2021.

Min Jin Chong and David Forsyth. Effectively unbiased FID and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Matt Crane. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252, 2018. doi: 10.1162/tacl_a_00018.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Lucas Degeorge, Arijit Ghosh, Nicolas Dufour, David Picard, and Vicky Kalogeiton. How far can we go with ImageNet for text-to-image generation? *arXiv preprint arXiv:2502.21318*, 2025.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7 (3):177–188, 1986.

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

Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3): 189–228, 1996. doi: 10.1214/ss/1032280214.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi: 10.1162/089976698300017197.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP-IJCNLP*, 2019.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, 2018.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. doi: 10.1007/s11263-014-0733-5.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/, 2025. Thinking Machines Lab blog, 10 September 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

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 30 (NIPS 2017)*, 2017.

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 33 (NeurIPS)*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Keller Jordan. On the variance of neural network training with respect to test sets and distributions. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

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

M. G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939. doi: 10.1214/aoms/1177732186.

J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953. doi: 10.1090/S0002-9939-1953-0055639-3.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

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 32 (NeurIPS)*, 2019.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of ImageNet classes in Fréchet inception distance. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Doudou LaLoudouana and Mambobo Bonouliqui Tarare. Data set selection. *Journal of Machine Learning Gossip*, 2003.

Howard Levene. Robust tests for equality of variances. In Ingram Olkin (ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pp. 278–292. Stanford University Press, 1960.

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

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers. *arXiv preprint arXiv:2410.08184*, 2024.

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

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

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

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

Pranava Madhyastha and Rishabh Jain. On model stability as a function of random seed. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations (ICLR)*, 2018.

Stanislav Morozov, Andrey Voynov, and Artem Babenko. On self-supervised image representations for GAN evaluation. In *9th International Conference on Learning Representations (ICLR)*, 2021.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations (ICLR)*, 2021.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2020.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Edward Raff. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

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 (CVPR)*, 2022.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

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 29 (NIPS 2016)*, 2016.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35 (NeurIPS), Datasets and Benchmarks Track*, 2022.

D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *6th International Conference on Learning Representations (ICLR), Workshop Track*, 2018.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations (ICLR)*, 2021b.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28.

Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Yuli Wu, Fucheng Liu, Rüveyda Yilmaz, Henning Konermann, Peter Walter, and Johannes Stegmaier. A pragmatic note on evaluating generative models with Fréchet inception distance for retinal image synthesis. *arXiv preprint arXiv:2502.17160*, 2025.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3024–3034, 2025. doi: 10.1109/WACV61041.2025.00299.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Supplementary Material

A Overview

The supplementary material extends the main paper along four axes: (i) the inferential statistics behind the main claims, (ii) further empirical analyses of the seed lottery on Inception FID, (iii) a metric-robustness stress test that replicates every main-paper claim under DINOv2 FID, the four Inception PRDC metrics, and GenEval/DPG-Bench on a text-to-image model, and (iv) the golden-section search that computes guided FID efficiently, with its theory.

Supplementary findings at a glance

- **The asymmetry is statistically robust.** On the headline panel, $\sigma_b/\sigma_w = 3.12$ (95% CI $[2.38, 4.39]$, $p = 1.5\times10^{-21}$), and $\sigma_b > \sigma_w$ holds in all 78 (model, step) cells after FDR control (Appendix E).
- **A gap must exceed $\approx 4\%$ to beat seed luck.** Two single runs of one recipe differ by more than 4% only 5% of the time, so certifying a gain g takes $N \approx (4/g)^2$ seeds per method. On the μP sweep, single-seed comparisons with a true gap of 1–4% come out backwards 17–27% of the time (Appendix F).
- **Seed rankings are stable, but no init seed is universally good.** The mean, minimum, and maximum over sampling seeds rank the training seeds alike ($\rho \geq 0.97$), and so do Inception and DINOv2 ($\rho \geq 0.94$). In a 10×15 factorial, the best init wins only 5 of 15 (data, noise) pairings (Sec. G.2, G.3).
- **The seed-induced CI is a fixed fraction of FID.** For a mean FID F over N training seeds, the 95% CI is $\pm 2.5\%\, F/\sqrt{N}$. One model at $F \approx 15$ carries ±0.38, and a ±0.1 interval needs $N \geq 15$ (Sec. G.4).
- **Retraining shifts the whole feature distribution.** Across training seeds, the mean and covariance terms of FID move together ($\rho = 0.97$). Resampling one model mostly moves the mean term, which carries 67% of the within-seed variance (Sec. G.5).
- **Text-to-image models face the same lottery.** The training seed carries 76% of the GenEval and 85% of the DPG-Bench variance, so averaging 10 evaluation seeds shrinks the spread of one run by only 7–12%. Compositional skills (position, color attribution) depend most on the seed, and the CFG scale moves GenEval by $\approx 16\times$ the seed σ (Appendix H).
- **The claims replicate on DINOv2 FID and on precision, density, and coverage.** DINOv2 FID widens the training-to-sampling ratio to $4.8\times$, and the three PRDC metrics bring it close to $1\times$. The noise $>$ init $>$ data order, the scale-invariant floor, and the 1.2–$2.9\times$ lucky-seed speedup all carry over. Recall is the exception: it builds its k-NN balls on the generated samples, so its sampling spread is over three times its training spread (Sec. I.2–I.6).
- **Guidance tightens only the Inception-FID floor.** Per-seed CFG halves the CoV of Inception FID, leaves DINOv2 FID at 0.9%, and moves PRDC to a high-precision, low-recall operating point (Sec. I.4).
- **The lottery is not an artefact of the Gaussian fit.** MIND, a sliced Wasserstein distance, reproduces the asymmetry ($2.5\times$ on Inception, $4.5\times$ on DINOv2) and ranks seeds like FID ($\rho = 0.99$). More images shrink only σ_w, never σ_b, so at 5k images the training lottery hides in sampling noise (Sec. I.7).
- **Guided FID costs ≈ 14 FID evaluations per cell.** Golden-section search replaces a $\approx$200-point CFG grid. Under an equal-covariance Gaussian feature model, FID(CFG) is a strictly convex quadratic, so the search finds the global optimum (Appendix J).
- **μP transfers a $1.7\times$ learning-rate window, not a point.** Three adjacent learning rates tie within the seed spread at every size. Without CFG, the sweep points to the largest learning rate, where some seeds diverge (Appendix K).
- **Guidance hides FID gaps from the eye.** Unguided samples degrade steadily as FID rises, while guided samples look clean even at the worst checkpoint (Appendix L).

Roadmap.

- Appendix B – extended related work, moved from Sec. 2.
- Appendix C – per-experiment configurations (N, K, training-step budget, deviations from the shared protocol) for every subsection of Sec. 4.
- Appendix D – companion tables consolidating the headline numbers behind Sec. 4.2 and Sec. 4.5 for one-stop lookup.
- Appendix E – the random-effects model, exact tests and intervals, multiplicity control, meta-analysis of the floor, data integrity, and the calibration behind Appendix F.
- Appendix F – the $\approx 4\%$ threshold for comparing two runs, and how often single-seed comparisons flip on the μP sweep.
- Appendix G – five additional empirical analyses on Inception FID: panel-by-panel violins for the source decomposition (Sec. G.1), rank stability across summary statistics (Sec. G.2), a 10×15 factorial test of init-seed universality (Sec. G.3), a practitioner-facing FID$\rightarrow$CI lookup (Sec. G.4), and the split of FID's variance into its mean and covariance terms (Sec. G.5).
- Appendix H – the 10-seed text-to-image study behind Sec. 4.3.
- Appendix I – DINOv2 FID and Inception PRDC replication of every Sec. 4 subsection, row-for-row with the main-paper tables, and a MIND (sliced Wasserstein-2) rescoring of the converged SiT-B/2 panel at 5k and 50k images (Sec. I.7).
- Appendix J – the guided-FID protocol, convergence of Algorithm 1 and unimodality of FID(CFG) under a Gaussian feature model.
- Appendix K – the μP learning-rate sweep across SiT-S/B/L/XL (moved from the main text).
- Appendix L – per-class galleries of samples ordered by FID, with and without classifier-free guidance, that visualise how perceptual quality tracks (or fails to track) FID.

B Extended related work

Sec. 2 keeps only the positioning needed to read the main paper. This section describes the surrounding literature in more detail and states, for each line of work, what it measures and how it differs from the two-lottery measurement of this paper.

FID and its known fragilities. The Fréchet Inception Distance (Heusel et al., 2017) was introduced together with the two time-scale update rule for GANs. It fits a Gaussian to the Inception-V3 (Szegedy et al., 2016) pool features of real and generated images and reports the Fréchet distance between the two Gaussians. It replaced the Inception Score (Salimans et al., 2016), which ignores the real data and which Barratt & Sharma (2018) show to be sensitive to the exact Inception weights and implementation, and to be easy to game. A series of works then showed that FID itself is less stable than its use as a leaderboard metric suggests. Chong & Forsyth (2020) prove that FID computed from a finite sample is biased, and that the bias depends on the generator, so two models compared at the same sample size can be mis-ranked; they propose FID_∞, extrapolated to an infinite sample from a fit in $1/n$. This is the closest analysis of what we call the generation lottery, but it keeps the trained generator fixed. Parmar et al. (2022) show that implementation details of the evaluation pipeline (aliased resizing, quantisation, JPEG compression) move FID by amounts comparable to published gains, and release Clean-FID; this is a deterministic pipeline effect that we remove by scoring every panel with one shared pipeline. Kynkäänniemi et al. (2023) show that Inception features encode ImageNet class information, so that matching the top-class histogram of the generated set to that of the reference set lowers FID without a perceptible change in quality. Stein et al. (2023) run a large human study and find that no existing metric tracks human judgement well, that FID unfairly penalises diffusion models, and recommend a DINOv2-ViT-L/14 feature space instead. Wu et al. (2025) find, on retinal image synthesis, that small FID gaps do not predict downstream utility. Surveys by Borji (2019; 2022) catalogue these and other GAN-era measures. None of these works retrains the generator: they study the estimator, pipeline or feature space of a fixed model, whereas we measure how much the number moves when the model is retrained under the same recipe.

Alternative metrics and text-to-image benchmarks. KID (Bińkowski et al., 2018) replaces the Gaussian fit with an unbiased polynomial-kernel MMD estimate on the same features. Precision and recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) split FID's single number into a fidelity and a diversity term, and density and coverage (Naeem et al., 2020) make these k-NN estimates robust to outliers in the real set. Other work changes the feature space, using self-supervised representations (Morozov et al., 2021) such as DINOv2 (Oquab et al., 2024), or both the features and the distance, as CMMD (Jayasumana et al., 2024) does with CLIP embeddings and a Gaussian-kernel MMD that drops the normality assumption. We rescore our panels with DINOv2 FID and the four PRDC metrics (Appendix I) and find that the balance between the two lotteries is metric-specific, so our conclusions about which lottery dominates apply to the Fréchet metrics. For text-to-image models, GenEval (Ghosh et al., 2023) checks object presence, count, colour and position with an object detector, and DPG-Bench (Hu et al., 2024) scores long, dense prompts with a vision-language model. Both are reported from a single trained model in practice; our text-to-image study (Appendix H) shows that they are subject to the same training lottery.

Seed variance and reproducibility in machine learning. Empirical machine learning has long been under scrutiny, prefigured by the satire of LaLoudouana & Tarare (2003) on choosing data sets so that one's method wins. In reinforcement learning, Henderson et al. (2018) show that nominally identical agents diverge across seeds, to the point that two groups of five seeds of the same algorithm yield significantly different learning curves. In NLP, Reimers & Gurevych (2017) show that sequence-tagging scores change significantly with the seed and argue for reporting score distributions, Melis et al. (2018) find that well-tuned LSTM baselines match newer language-model architectures once tuning and seed variance are controlled, and Crane (2018) and Madhyastha & Jain (2019) report seed-driven variability in question answering and classification. The fine-tuning studies are the closest in spirit to our decomposition: Dodge et al. (2020) fine-tune BERT with many seeds and separate the effect of weight initialisation (of the classification head) from that of data order, finding both comparable, and Mosbach et al. (2021) attribute fine-tuning instability to optimisation difficulties and late-training differences in generalisation rather than to small data sets per se. Systemic critiques (Sculley et al., 2018; Gundersen & Kjensmo, 2018; Raff, 2019; Pineau et al., 2021) survey reporting practice and reproduction attempts and call for stronger standards, echoed by the Pascal VOC retrospective (Everingham et al., 2015), which uses bootstrapping to ask which challenge rankings are statistically meaningful. Dodge et al. (2019) argue that the hyperparameter-search budget must be disclosed, since expected performance depends on it. Bouthillier et al. (2019) distinguish kinds of reproducibility and show that conclusions can change with the sources of variation that are held fixed, and Bouthillier et al. (2021) decompose benchmark variance into data sampling, initialisation, data order, augmentation and hyperparameter choice, recommending that several sources be randomised when comparing methods. Our three-generator decomposition (Sec. 4.2) is the generative-model analogue, with one source that has no discriminative counterpart: the per-step Gaussian noise of the diffusion loss.

Numerical non-determinism and instability. Pham et al. (2020) show that identical training runs of deep learning software differ because of non-deterministic kernels and that many practitioners are unaware of it. Summers & Dinneen (2021) show that every source of non-determinism, down to a single-bit change in the initial weights, produces run-to-run variance of similar size, because training is unstable to small perturbations. Picard (2021) scan thousands of seeds on CIFAR-10 and ImageNet and find that the best and worst seeds differ by amounts comparable to published improvements. Our SAME-SEED control (Sec. 4.2) tests the generative analogue: multi-GPU non-determinism alone moves the weights by several percent of their norm, yet leaves FID variance below the sampling floor, so in our setting numerical noise does not drive the training lottery.

Why independently trained networks differ. The lottery-ticket hypothesis (Frankle & Carbin, 2019) and linear mode connectivity (Frankle et al., 2020) show that the outcome of training depends on the initialisation and, early in training, on the data order, after which runs become stable to SGD noise. Fort et al. (2019) show that independently initialised networks land in different modes of the loss landscape and make different predictions, which is why deep ensembles help. Initialisation schemes (Glorot & Bengio, 2010; He et al., 2015), adaptive optimisers (Kingma & Ba, 2015) and architectures (He et al., 2016; Vaswani et al., 2017; Dosovitskiy et al., 2021) all shape this landscape. Nagarajan & Kolter (2019) show that uniform-convergence bounds cannot explain generalisation in overparameterised networks, one reason theory gives little guidance on run-to-run gaps. Run-to-run

diversity is exploited for robustness and uncertainty estimation by ensembles over hyperparameters and seeds (Wenzel et al., 2020). Jordan (2024) show that for image classification most of the test-set variance between runs is finite-sample noise of the test set, with little variance on the underlying distribution. Our finding is the opposite for FID: the variance on a fixed reference set is dominated by the trained model, not by the finite sample of generated images. Beyond the test set, Recht et al. (2019) show that accuracy drops on a freshly collected ImageNet test set, and D'Amour et al. (2022) show that models that are equivalent in distribution can differ widely off distribution.

Statistical methodology. The classical toolkit for comparing learning algorithms includes paired tests and 5×2 cross-validation (Dietterich, 1998), the Friedman and Nemenyi tests across data sets (Demšar, 2006), their Bayesian counterparts (Benavoli et al., 2017), bootstrap confidence intervals (Efron, 1979; DiCiccio & Efron, 1996), Welch's unequal-variance t-test (Welch, 1947), family-wise (Dunn, 1961) and false-discovery-rate (Benjamini & Hochberg, 1995) multiplicity control, random-effects meta-analysis (DerSimonian & Laird, 1986), and rank statistics (Spearman, 1904; Kendall & Babington Smith, 1939). NLP has adopted these tools (Dror et al., 2018; Dodge et al., 2019); generative modelling largely has not. Appendix E applies the random-effects model, exact tests, multiplicity control and meta-analysis to our panels. Banerjee et al. (2024) propose a robust nonparametric (Kolmogorov–Smirnov) test for training variability from stochastic optimisation, which targets the same question for discriminative models.

Diffusion models, scaling and generative reproducibility. Diffusion and flow models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021a;b; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Kingma et al., 2021; Karras et al., 2022; Ho & Salimans, 2022; Rombach et al., 2022; Lipman et al., 2023; Liu et al., 2023; Karras et al., 2024; Ma et al., 2024; Esser et al., 2024; Li et al., 2024) define the current state of the art on ImageNet (Deng et al., 2009), FFHQ (Karras et al., 2019) and LAION-5B-scale data (Schuhmann et al., 2022), and are ranked almost exclusively by single-run FID. Scaling laws (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; Henighan et al., 2020), including for diffusion transformers (Peebles & Xie, 2023; Liang et al., 2024), describe how the mean loss or FID improves with compute, which encourages treating seed variance as a residual that vanishes at scale; we measure it within each scale and find that it does not (Sec. 4.5). In generative modelling specifically, Lucic et al. (2018) run a large-scale GAN study and find that, with equal hyperparameter budgets and random restarts, most models reach similar FID, so that many reported gains fall within the spread of a single model. Degeorge et al. (2025) attack the data side of reproducibility with a text-to-image protocol trained on ImageNet only. At the function level, Zhang et al. (2024) show that diffusion models trained independently, even with different architectures, produce nearly the same image from the same initial noise, and Kadkhodaie et al. (2024) show that networks trained on disjoint subsets of the data converge to nearly the same score once the data set is large enough. We add that such near-identical noise-to-image maps still yield percent-level FID fluctuations when scored as a distribution. Xu et al. (2025) study the complementary inference-time lottery in text-to-image diffusion, showing that some sampling seeds consistently produce better images.

Closest precursors. Three works measure uncertainty in generative metrics directly. Chong & Forsyth (2020) vary the number of generated samples for a fixed generator. Bench & Thomas (2025) apply Monte Carlo dropout to the feature extractor to obtain a distribution over FID-like scores, which measures the uncertainty of the evaluator rather than of the generator. Lucic et al. (2018) report FID distributions over random restarts of GANs, but do not separate the sources of variance or relate the spread to compute and scale. None of them varies training seeds, architectures and checkpoints jointly, separates the training lottery from the generation lottery, or studies how per-cell guidance tuning interacts with seed variance, which are the contributions of this paper.

C Per-experiment configurations

This appendix gathers the per-experiment N, K, training-step budget, and any deviations from the shared protocol of Sec. 3. The default panel uses $K = 10$ sampling seeds per trained model and the shared FID evaluation pipeline (Inception-V3 features, 50 000 generated samples, ImageNet train-set reference statistics, deterministic ODE sampler). Each entry below is referenced from the corresponding subsection of Sec. 4.

Implementation details. Training follows the SiT recipe of Ma et al. (2024) on class-conditional ImageNet 256×256. The only deliberate deviation is that we train $N \approx 25$ independent runs per cell to populate the two-axis panel rather than a single run per configuration. The aggregate compute footprint of the experiments reported in Sec. 4 (including preliminary runs that did not make it into the final paper) is approximately 100 000 H100 GPU-hours.

Sec. 4.1 (training- vs. sampling-seed asymmetry). A converged SiT-B/2 (class-conditional ImageNet 256×256, 400k training steps, no classifier-free guidance) generates 50k samples and we compute FID. Repeating with $K = 10$ different sampling seeds on each of $N = 25$ independently trained SiT-B/2 networks yields the 250 FID evaluations of Figure 2: every dot is one evaluation, every violin is one trained model.

Sec. 4.2 (variance decomposition across the three training-time RNGs). Each single-source ablation fixes two of the three training-time random number generators (parameter initialisation, data-loader order, per-step flow-matching noise) and varies the third across 24–25 SiT-B/2 runs trained to 400k steps with $K = 10$ sampling seeds per run, yielding the conditions VARY-NOISE, VARY-INIT, and VARY-DATA. The fully-random baseline (VARY-ALL) is the 25×10 panel of Sec. 4.1.

Sec. 4.4 (guided FID). The CFG sweep operates on the same 25×10 converged SiT-B/2 panel. For every (training, sampling) seed pair, golden-section search runs over the bracket $[\omega_{\min}, \omega_{\max}] = [0, 2]$ at tolerance $\varepsilon = 0.01$, costing ≈ 14 FID evaluations per cell, and we report the FID at the recovered optimum (Algorithm 1, convergence and unimodality in Appendix J).

Sec. 4.5 / 4.6 (compute and scale). We train $N = 25$ networks per model size for 2M steps and compute $K = 10$ Inception FID evaluations per checkpoint every 100k steps, yielding 76 (model, step) cells across SiT-S/B/L/XL. Sec. 4.6 reads the lucky-seed speedup off the same panel.

Appendix K (μP learning-rate sweep). We sweep 10 μP-coordinated learning rates log-spaced in $[5 \times 10^{-5}, 5 \times 10^{-4}]$ with $N = 10$ seeds per cell across SiT-S/B/L/XL, train each cell to 100k steps, and score with both unguided FID and guided FID (≈ 400 trained networks total, Figure 26).

D COMPANION SUMMARY TABLES

The figures of the main paper carry the qualitative argument. The two tables below consolidate the headline numbers behind them so they can be looked up in one place.

Table 1: **Variance decomposition of the training-seed lottery (SiT-B/2, 400k, no CFG).** Companion to Figure 4. Each row fixes two of the three training-time random sources and varies the third, and VARY-ALL is the fully-stochastic baseline. σ_{between} is the standard deviation across the N per-seed means. σ_{within} is the within-seed sampling-noise σ, averaged across the N training seeds. $\text{CoV}_{\text{between}}$ is $\sigma_{\text{between}}/\mu$. The rightmost column reports the share of the baseline σ_{between} reproduced by each single-source ablation. The naive sum $\sqrt{\sigma^2_{\text{noise}} + \sigma^2_{\text{init}} + \sigma^2_{\text{data}}} \approx 0.50$ overshoots the observed $\sigma_{\text{vary-all}} = 0.44$ by 13%: the sources are not independent.

Condition	N	σ_{between}	σ_{within}	$\text{CoV}_{\text{between}}$ (%)	range	% of vary-all
VARY-ALL	25	0.438	0.137	1.26	1.66	100.0
VARY-NOISE	25	0.334	0.144	0.97	1.33	76.3
VARY-INIT	24	0.294	0.150	0.84	1.09	67.0
VARY-DATA	25	0.218	0.150	0.63	0.82	49.8

Table 2: **The seed lottery across compute and model size.** Companion to Figure 7. Numbers at the start (200k) and the end (2M) of the scaling sweep. Mean FID drops by $\approx$1.7–2$\times$ over 1.8M extra steps, while σ_{between} shrinks 1.7–2.4$\times$, so $\text{CoV}_{\text{between}}$ stays near a 1–2% band at every checkpoint.

Model	Step	N	mean FID	σ_{between}	$\text{CoV}_{\text{between}}$ (%)	σ shrink
SiT-S/2	200k	19	71.2	0.75	1.06	—
SiT-S/2	2M	19	41.1	0.31	0.74	2.4$\times$
SiT-B/2	200k	20	46.2	0.48	1.05	—
SiT-B/2	2M	20	23.3	0.29	1.24	1.7$\times$
SiT-L/2	200k	24	28.9	0.43	1.48	—
SiT-L/2	2M	24	14.8	0.25	1.72	1.7$\times$
SiT-XL/2	200k	25	27.5	0.44	1.61	—
SiT-XL/2	2M	25	14.5	0.20	1.42	2.2$\times$

E INFERENTIAL STATISTICS

This appendix gives the statistical backing of Sec. 4.1, Sec. 4.5 and Appendix F. All tests are exact or Monte-Carlo calibrated and are run over 78 (model, step) cells, 19,291 FID evaluations in total: the 76 cells of the scaling sweep plus two 100k checkpoints that pass the integrity screen below.

Model. Each cell is a balanced one-way random-effects ANOVA, $\text{FID}_{ij} = \mu + b_i + e_{ij}$ for training seed $i \in \{1, \dots, N\}$ and sampling seed $j \in \{1, \dots, K\}$, with $b_i \sim \mathcal{N}(0, \sigma_b^2)$ and $e_{ij} \sim \mathcal{N}(0, \sigma_w^2)$ independent. With MS_b and MS_w the between- and within-seed mean squares, $\mathbb{E}[\text{MS}_b] = \sigma_w^2 + K\sigma_b^2$ and $\mathbb{E}[\text{MS}_w] = \sigma_w^2$, so the statistic $F = \text{MS}_b/\text{MS}_w$ satisfies $F/(1+Kr^2) \sim F(N-1, N(K-1))$ when $\sigma_b = r\,\sigma_w$. This gives an exact test of $H_0 : \sigma_b \le r\,\sigma_w$ for any ratio r, and exact confidence intervals for σ_b/σ_w by inversion. The raw σ_{between} of the main text is the standard deviation of the per-seed means, whose expectation is $\sqrt{\sigma_b^2 + \sigma_w^2/K}$; this is why the raw ratio (3.19$\times$ on SiT-B/2) exceeds the variance-component ratio (3.12$\times$).

The headline cell. On the 25×10 SiT-B/2 panel of Figure 2 (400k steps): $F = 98.04$ on $(24, 225)$ degrees of freedom; $\sigma_b \le \sigma_w$ rejected at $p = 1.5 \times 10^{-21}$; $\sigma_b/\sigma_w = 3.12$, 95% CI $[2.38, 4.39]$, exact one-sided 95% lower bound 2.48; ICC 0.907 $[0.850, 0.951]$. A permutation test of the seed labels agrees ($p < 10^{-4}$, 20,000 permutations). Homogeneity of the within-seed variances across the 25 seeds (Levene (Levene, 1960) $p = 0.93$, Bartlett $p = 0.88$) and normality of the residuals (Shapiro–Wilk $p = 0.56$) are not rejected. $N = 25$ leaves σ_b itself imprecise, which is why we quote interval bounds: the claim $\sigma_b > 2\sigma_w$ survives the least favourable end of the interval.

Multiplicity across cells. Testing $\sigma_b \le \sigma_w$ in each of the 78 cells, 78/78 rejections survive Benjamini–Hochberg control at 5% FDR and 67/78 survive Bonferroni. The 11 Bonferroni failures are late checkpoints with ratios 1.29–1.57. The median ratio across cells is 2.0$\times$. We therefore claim $\sigma_b > \sigma_w$ for the whole family and larger multiples only cell by cell.

Meta-analysis of the floor. Pooling $\text{CoV}_{\text{between}}$ across the 78 cells with a DerSimonian–Laird random-effects model (DerSimonian & Laird, 1986) gives 1.21% (95% CI $[1.14, 1.28]$). The 95% prediction interval for the CoV of an unseen cell of this family is $[0.79, 1.84]$%, and the 95/95 upper tolerance bound is 2.0%.

Data integrity. Two cells (SiT-B and SiT-L at 100k) whose evaluations are identical to those of the 2M checkpoint of the same model, a checkpoint mix-up, are dropped. 19 diverged entries are screened: 15 (cell, training seed) entries of the scaling panel, i.e. 150 evaluations, whose mean FID exceeds 1.5$\times$ the cell median (the count is the same for any threshold between 1.1$\times$ and 1.8$\times$), coming from 8 of the 50 SiT-S/B runs, and 4 of the 400 μP runs (3 SiT-S and 1 SiT-XL at the largest LR). The per-checkpoint figures and tables of Sec. 4.5 use a stricter, run-level screen that keeps a training run only if it is valid at every checkpoint, which leaves $N = 19$, 20, 24 and 25 runs for SiT-S/B/L/XL (Table 2). All reported CoVs are conditional on non-divergence, and we report divergence counts alongside them.

False-positive calibration of the decision rule. We draw 3,000 pairs of single runs from the same cell with different training seeds, so any declared winner is a false positive, and take the 95th

percentile of their relative FID gap as the threshold at a 5% false-positive rate: 4.0% for single evaluations and 3.6% for runs averaged over their 10 sampling seeds. Both sit slightly above the Gaussian approximation $1.96\sqrt{2}\,\mathrm{CoV} = 3.4\%$ at the pooled CoV because the CoV varies across cells. Appendix F turns this into a decision rule.

Flip rates on the μP sweep. For Table 3, the sweep (4 sizes $\times$ 10 LRs $\times$ 10 training seeds $\times$ 10 sampling seeds, 4 diverged runs screened) is read as the 180 pairs of LRs within a size. The reference ordering is that of the 10-seed means, the true gap is $|\Delta\mathrm{FID}|$ relative to the larger of the two means, and the flip rate is the probability that 1 (or 5) evaluations per side invert it. One per side is computed exhaustively over all evaluations of each LR; five per side draws evaluations with replacement from that pool. The two smallest bins contain only 4 pairs each.

Sampling-only intervals. For each metric on the 25×10 SiT-B/2 panel, the share of the combined (training plus sampling) spread covered by an interval built from the sampling seeds of one network is 30% (Inception FID), 20% (DINOv2 FID), 66% (precision), 58% (density), 67% (coverage) and 96% (recall). The share is $\sigma_{\text{within}}/\sqrt{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$.

F From noise floor to decision rule

Comparing two single runs, a gap must exceed $\approx$4% of the baseline FID to beat seed luck at a 5% false-positive rate, not the 1.3% floor itself. Between 1% and 4%, single-seed comparisons from a real learning-rate sweep are backwards one time in four to six.

Two runs, not one. The CoV floor describes one run around the recipe mean, but practitioners compare two runs, each carrying its own seed luck, so the relevant scale is $\sqrt{2}\,\sigma_b$. Calibrated on 3,000 same-configuration pairs drawn from the panel, the rule "declare a winner above the $\approx$1.3% floor" carries a false-positive rate of $\approx$50%. The threshold that holds the false-positive rate at 5% is $\approx$4% **of the baseline FID** for single evaluations (3.6% if each run is averaged over 10 sampling seeds, Appendix E). With N training seeds per method it shrinks as $4\%/\sqrt{N}$, so certifying a relative gain g at 5% false-positive rate needs $N \approx (4/g)^2$ seeds per method: one for 4%, four for 2%, nine for $\approx$1.3%. Sec. G.4 applies the same rules at any FID level.

Table 3: **How often a seed-limited comparison inverts the 10-seed ordering**, on the μP sweep read as 180 ablation-style LR pairs, binned by the true relative gap between the two LRs.

true gap (% of baseline)	# pairs	1 seed / side	5 seeds / side
$< 1\%$	4	52%	44%
1–2%	4	27%	12%
2–4%	17	17%	6.4%
4–5%	11	8.3%	1.2%
5–10%	40	3.2%	0.4%
$> 10\%$	104	0.4%	0.0%

How often single-seed conclusions flip. To measure the price of ignoring the threshold, we read the μP sweep of Appendix K as 180 ablation-style decisions, every pair of learning rates within a model size, with the 10-seed means as ground truth (Appendix E). A single-seed comparison, the one a paper would report, inverts the ordering for 52% of pairs whose true gap is below 1%, 17–27% between 1% and 4%, and 0.4% above 10% (Table 3). Below 1% five seeds per side barely help (44%): the ordering is not identifiable at that scale. Between 1% and 4% they cut the flip rate to 6–12%. At the decision level the damage is larger: a single-seed LR sweep selects a different LR than the 10-seed sweep in 31%, 19%, 43% and 18% of draws (SiT-S/B/L/XL) and picks 4 to 6 distinct LRs from the same grid. Same grid, same code, different seed, different published recommendation. The sweep is one family, not the literature; doing the same for published methods would mean retraining each several times, the cost barrier this paper documents.

G Additional analyses on Inception FID

This appendix collects five further analyses of the seed lottery on Inception FID, each tied to a subsection of Sec. 4.

G.1 PER-TRAIN-SEED VIOLIN PANELS BY VARIABILITY SOURCE

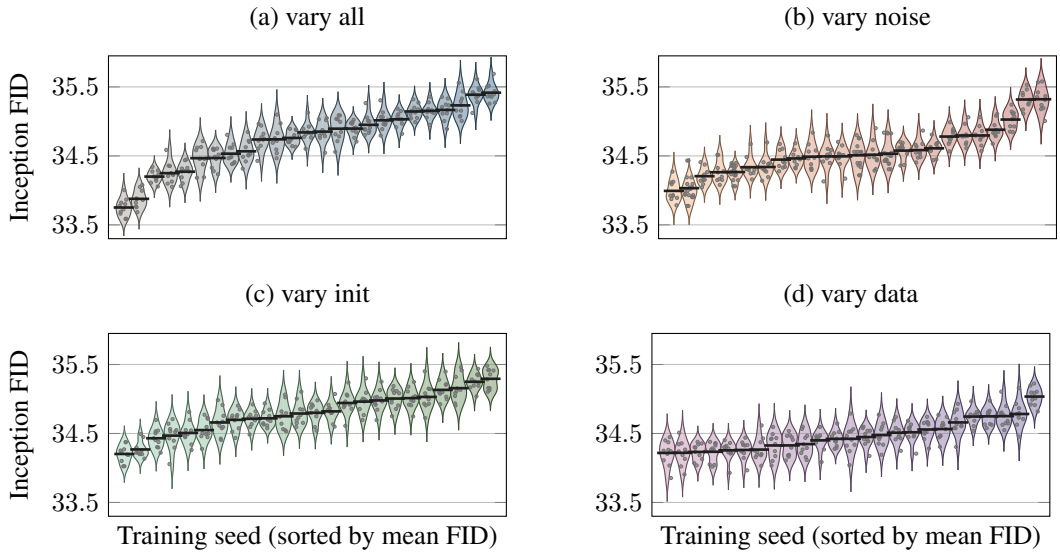

Figure 9: **Per-train-seed violins for the four single-source conditions** (SiT-B/2, 400k). (a) Vary all: the panel of Figure 2 ($\sigma_{\text{between}} = 0.438$). (b) Vary noise: init and data order fixed (0.334, $\approx$76% of (a)). (c) Vary init (0.294, $\approx$67%). (d) Vary data (0.218, $\approx$50%). Every violin is one training seed (Gaussian KDE of its 10 sampling-seed evaluations), dots are individual evaluations, and the black tick is the per-seed mean. All panels share one y-range.

Figure 9 repeats the layout of Figure 2 for each single-source condition of Sec. 4.2: violin height tracks the within-seed σ of one training run, and the spread of the black mean ticks tracks the between-seed σ of the condition. The violins keep the same height in every panel ($\sigma_{\text{within}} \in [0.137, 0.150]$), so what is varied during training does not leak into the evaluation jitter. The mean ticks contract from panel (a) to panel (d), their range halving from ≈ 1.7 to under ≈ 0.9 FID: the lottery shrinks along the training axis only. Within every panel the ticks are spread evenly along the sorted axis rather than bunched around a few outliers, consistent with the well-behaved $(\max - \min)/\sigma$ ratios of Sec. 3.

G.2 THE TRAINING-SEED RANKING IS ROBUST TO THE SUMMARY STATISTIC

Even within a single feature space, the way the ten sampling seeds per training run are summarised determines which seed is declared best. Reusing the 25×10 SiT-B/2 panel of Sec. 4.1, we collapse the within-seed dimension three ways: per-seed mean, per-seed minimum, and per-seed maximum FID. We then ask whether the resulting orderings of the 25 training seeds agree.

The three rankings largely agree (Figure 10a). Crossings are local (at most 4 ranks), the same seed is best under all three criteria, and the rankings correlate at $\rho \geq 0.97$. This follows from Sec. 4.1: the within-seed spread ($\sigma_{\text{within}} \approx 0.14$) is small next to the between-seed spread ($\sigma_{\text{between}} \approx 0.44$), so reporting the best of k sampling seeds instead of their average shifts every training seed by a similar amount and rarely reorders them.

The feature extractor matters as little: Inception and DINOv2 agree on which training seed is good under each of the three criteria (Spearman $\rho \geq 0.94$, Figure 10b). The seed ranking is a property of the trained model, not of how its sampling seeds are summarised.

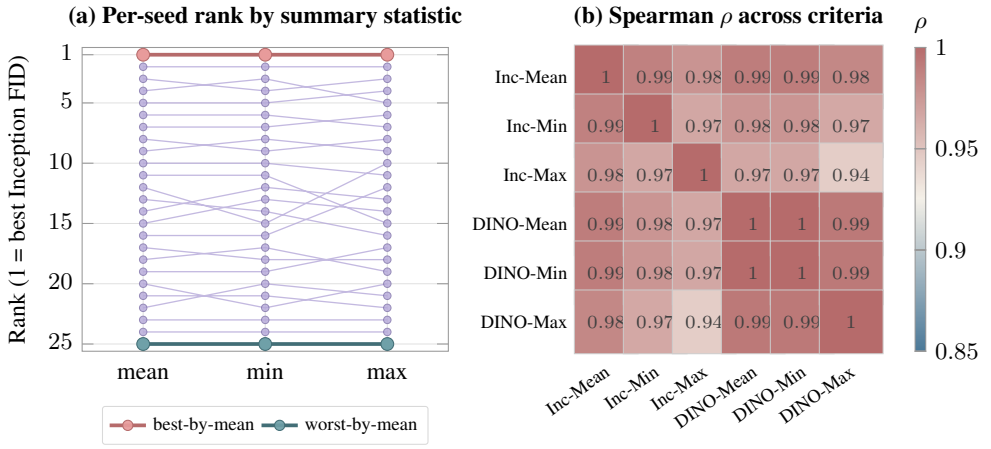

Figure 10: **Rank stability of the 25 training seeds (SiT-B/2, 400k).** (a) Bump chart: each line is one training seed traced through three ranking criteria — mean, min, and max FID across its 10 sampling seeds. Crossings are local (at most 4 ranks): the best and worst seeds by *mean* are also best and worst by *min* and *max*, and the three Inception rankings correlate at $\rho \geq 0.97$. Coral and teal highlight the seeds that are best- and worst-by-mean to make their rank trajectories visible. (b) Spearman ρ between rankings under all six combinations of {Inception, DINOv2} $\times$ {mean, min, max}. Inception and DINOv2 agree strongly on the *mean* ranking ($\rho = 0.99$) but the agreement weakens for the *max* criterion ($\rho = 0.94$), i.e. the two metrics disagree more often on which seed had the worst-case sampling.

G.3 Are "good" init seeds universal?

Sec. 4.2 showed that initialisation contributes $\approx 67\%$ of the baseline between-seed σ on its own, which makes "find a good init" the cheapest and most tempting shortcut a practitioner can take to a low FID. We test whether such an init exists by running a full factorial of 10 init seeds against 15 independent (data, noise) pairings on SiT-B/2 at 400k, with ≈ 11 sampling seeds per cell, totalling $\approx 1\,600$ FID evaluations (two of the 150 cells failed to train and are dropped). Figure 11 shows the resulting 10×15 FID heatmap and the bump chart of init ranks across pairs.

If we average across the 15 (data, noise) pairs, the 10 init grand means span only $[34.49, 35.27]$ Inception FID, a 0.78-FID range slightly tighter than the 1.66 between-seed range of Sec. 4.1. After enough marginalisation, init seeds are nearly interchangeable. The within-cell spread, on the other hand, lands almost exactly on the generation lottery of Sec. 4.1: the median per-cell range is 0.51 and the median per-cell σ is 0.149, against 0.137 in the 25×10 baseline panel. The cross-experiment match is a useful sanity check (two completely separate sweeps recover the same within-seed sampling-noise scale), and it tells us the generation lottery is set by the converged SiT-B/2 weights, not by which init or which data pairing produced them. In aggregate, an init looks like a constant offset.

The aggregate picture is misleading. Computing the 10-init ranking *independently* for each of the 15 pairs and asking how concordant the rankings are gives a Kendall's W of 0.41 under the mean criterion, 0.45 under the min, and 0.36 under the max. The average pairwise Spearman ρ between pair-rankings is 0.36. These values are significant ($p < 0.01$), so "init quality" is not pure noise, but they are also far from concordant: the same init can be top-3 under one (data, noise) pair and bottom-3 under another. The complementary 15-pair ranking computed per init has Kendall's W in $[0.11, 0.13]$, statistically indistinguishable from random, so knowing which (data, noise) pair was good for one init says almost nothing about another init.

The bump chart in Figure 11(b) makes this visceral: most init seeds visit both the top three and the bottom three across the 15 pairings. The best init by grand mean (90682) takes top-1 in only 5 of 15 pairings. The worst init (96273) is bottom-1 in 8 of 15. Some inits are weakly worse than the rest, but no init is universally best, and no practitioner who reports a "best of 10 inits" run is reporting a transferable artefact. This is also why fixing the init alone in Sec. 4.2 only removes 33% of the baseline variance: the init lottery is not a constant per-init offset to the FID, but interacts strongly

with the (data, noise) draw the init is paired with. The init lottery and the data lottery are entangled, and disentangling them by varying one at a time understates the seed lottery practitioners face.

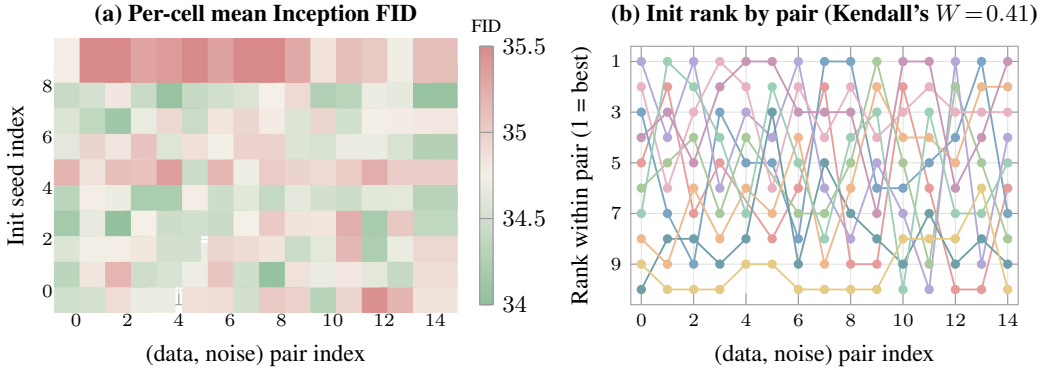

Figure 11: **Seed optimality: are "good" init seeds universal? (SiT-B/2, 400k, no CFG.)** (a) Heatmap of mean Inception FID over a 10×15 grid of init seeds (rows) and (data, noise) pairings (columns). Cells span $\sim$34.0–35.8. No init row is consistently the greenest across columns. White cells are the 2 missing runs. (b) Bump chart of the same data: each line is one init seed traced through its rank within each (data, noise) pair. Heavy crossings make the rank instability visceral. Kendall's $W = 0.41$ on the init rankings (mean criterion) shows significant agreement, but far from concordance. The "best" init by grand mean wins only $5/15$ pairs.

G.4 What CI should I expect at my reported FID?

A practitioner-facing lookup distilled from Table 2 and Figure 22: given a measured FID, the seed-induced 95% CI is a fixed fraction of the FID, set by the scale-invariant CoV floor.

Setup. The between-seed coefficient of variation $\mathrm{CoV} = \sigma_{\text{between}}/\mu$ on Inception FID stays in a tight band across the 76 (model, step) cells of the scaling sweep: median 1.30%, p_{10}–p_{90} 0.88–1.73% (Figure 22, Table 2). For a mean FID computed from N independently trained models with $K = 10$ sampling seeds each, the normal-approximation 95% CI half-width on the seed-mean is

$$\mathrm{CI}_{95}(F, N) \;=\; z_{0.975}\,\frac{\mathrm{CoV}\,F}{\sqrt{N}} \;\approx\; \frac{0.0254\,F}{\sqrt{N}}, \qquad z_{0.975} \approx 1.96, \tag{1}$$

under $\mathrm{CoV} = 1.30\%$. The within-seed contribution $\sigma^2_{\text{within}}/K$ is absorbed by the displayed CoV band: at $K = 10$ the within-CoV is $\approx 0.4\%$ (Table 6) and adds less than 5% to the variance of the seed-mean.

Reading the figure. Figure 12 plots CI_{95} against the reported mean FID for $N \in \{1, 5, 10, 25\}$. Three operating points to keep in mind. At $F \approx 15$ (typical converged SiT-L), a single trained model carries a ± 0.38 FID seed-only 95% CI. At $F \approx 35$ (SiT-B/2 baseline panel) the CI is ± 0.89. At $F \approx 70$ (early-training regime) it is ± 1.78. The shaded band traces the $p_{10} - p_{90}$ envelope of the CoV across the panel, so a cell at the noisy end of the floor inflates these numbers by a further $\approx 1.3\times$.

Implications for benchmarking. The CI shrinks only as $1/\sqrt{N}$. Pulling the CI on a $F \approx 15$ benchmark below ± 0.1 – the gap that routinely separates published methods – requires $N \geq 15$ training seeds at the median CoV ($N \geq 26$ at p_{90}). Against the single-seed ± 0.38 band at this FID, a 0.1-FID gain is indistinguishable from noise. It describes one mean around the recipe mean; to compare two methods, each with N seeds, the gap must clear $\sqrt{2}$ times this half-width, i.e. $\approx 3.4\%/\sqrt{N}$ of the baseline at the pooled CoV, or $\approx 4\%/\sqrt{N}$ measured empirically on single evaluations (Appendix F, Appendix E). For a metric other than Inception FID, substitute the matching CoV band from Table 10.

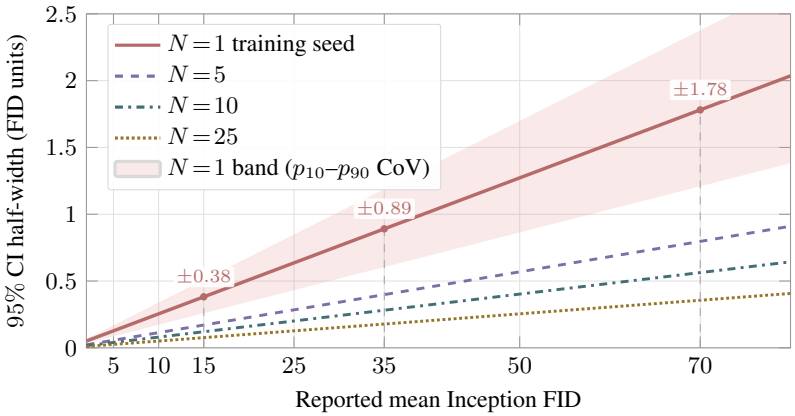

Figure 12: **Seed-induced 95% confidence interval as a function of the reported Inception FID.** For a mean FID computed from N independently trained models with $K=10$ sampling seeds each, the normal-approximation half-width is $\mathrm{CI}_{95} = 1.96\,\mathrm{CoV}\,F/\sqrt{N}$, where $\mathrm{CoV} = \sigma_{\text{between}}/\mu$ is the scale-invariant noise floor reported for Inception FID across the 76 (model, step) cells of the scaling sweep (Table 10, with median 1.30% and $p_{10}-p_{90}=0.88-1.73\%$). The solid coral curve uses the median CoV. The shaded band traces the $p_{10}-p_{90}$ envelope at $N=1$. At FID ≈ 15 (typical converged SiT-L), a single trained model carries a ± 0.38 FID seed-only CI – four times the ± 0.1 gaps that routinely separate published methods. Pulling that CI under ± 0.1 takes $N \geq 15$ training seeds at the median CoV ($N \geq 26$ at p_{90}). Within-seed (sampling-only) contributions are absorbed into the displayed band: at $K=10$ the within-CoV is $\approx 0.4\%$ on this panel and contributes under 5% of the seed-mean variance.

G.5 Mean and covariance terms of FID

FID is exactly the sum of a mean term and a covariance term,

$$\mathrm{FID} = \underbrace{\|\mu_r - \mu_g\|^2}_{M} + \underbrace{\mathrm{Tr}\,\Sigma_r + \mathrm{Tr}\,\Sigma_g - 2\,\mathrm{Tr}\big(\Sigma_r\Sigma_g\big)^{1/2}}_{C\,=\,\mathrm{Bures}(\Sigma_r,\Sigma_g)},$$

so its variance splits exactly as $\mathrm{Var}(\mathrm{FID}) = \mathrm{Var}(M) + \mathrm{Var}(C) + 2\,\mathrm{Cov}(M,C)$. We attribute it additively as $\mathrm{share}(M) = [\mathrm{Var}(M) + \mathrm{Cov}(M,C)]/\mathrm{Var}(\mathrm{FID})$ and likewise for C, so that the two shares sum to one. To estimate both levels of variance with many sampling draws, we generate 1M images from each of 25 SiT-B/2 networks at 400k steps (the seeds of the scaling panel of Sec. 4.5), split them into 20 disjoint 50k-image subsets, and compute (M, C, FID) on every (training seed, subset) cell against the ImageNet train reference. This panel reproduces the headline spread ($\sigma_{\text{between}} = 0.428$, $\mathrm{CoV}_{\text{between}} = 1.23\%$; Table 4).

Neither moment explains the training lottery alone. Across training seeds the mean term carries 59% of the variance and the covariance term 41% (Figure 13b). The dominant structure is co-movement: the per-seed means of the two terms correlate at $\rho = +0.97$ (Figure 13a), and the cross term $2\,\mathrm{Cov}(M,C)$ alone accounts for 48% of $\mathrm{Var}(\mathrm{FID})$, against 35% for $\mathrm{Var}(M)$ and 17% for $\mathrm{Var}(C)$. A lucky seed is thus closer to the reference in mean and better aligned in covariance at the same time, which makes the training lottery a coherent change in the generated distribution rather than an artefact of one moment. A counterfactual that does not rely on additivity agrees: freezing the covariance at its across-seed average and letting only the mean vary recovers 35% of the between-seed FID variance, and the converse recovers 16%. The covariance term does not vary through the overall feature energy: $\mathrm{Tr}\,\Sigma_g$ is nearly seed-constant ($\mathrm{CoV}_{\text{between}} = 0.15\%$, an order of magnitude below FID's), so C moves through the alignment term $\mathrm{Tr}(\Sigma_r\Sigma_g)^{1/2}$.

The generation lottery is more mean-driven. Across the 20 subsets of one network, the mean term carries 67% of the variance and the two terms correlate only at $\rho = 0.69$, a second qualitative difference between the two lotteries: resampling mostly moves the estimated mean, whereas retraining moves both moments together. The same holds when the networks are sampled without class conditioning (Figure 13c): 62%/38% across training seeds ($\rho = 0.89$), 78%/22% within a seed ($\rho = 0.38$).

Table 4: **Variance components of FID and of its two terms** (SiT-B/2, 400k steps, 25 training seeds $\times$ 20 disjoint 50k-image subsets, Inception features, train reference). Bottom: exact additive shares of $\mathrm{Var}(\mathrm{FID})$ and the correlation between the two terms at each level.

Class-conditional	mean	σ_{between}	$\mathrm{CoV}_{\text{between}}$	σ_{within}	$\mathrm{CoV}_{\text{within}}$
FID	34.80	0.428	1.23%	0.190	0.55%
mean term M	18.58	0.254	1.37%	0.133	0.72%
covariance term C	16.21	0.176	1.09%	0.072	0.45%
$\mathrm{Tr}\,\Sigma_g$	190.53	0.286	0.15%	0.257	0.13%
	share M / C		$\rho(M, C)$		
training lottery (between seeds)	59% / 41%		0.97		
generation lottery (within seed)	67% / 33%		0.69		
single evaluation (pooled)	61% / 39%		0.92		
Unconditional sampling					
FID	70.66	0.433	0.61%	0.187	0.26%
training lottery	62% / 38%		0.89		
generation lottery	78% / 22%		0.38		

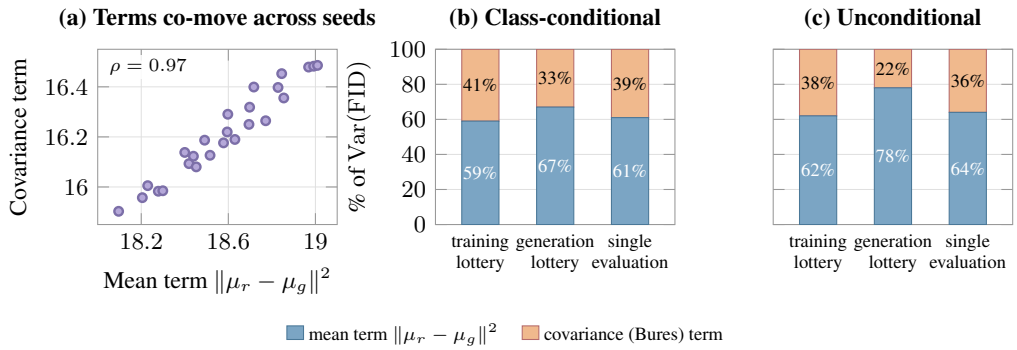

Figure 13: **Both moments of FID carry the training lottery.** SiT-B/2 at 400k steps, 25 training seeds, each scored on 20 disjoint 50k-image subsets of a shared 1M noise pool. (a) Per-seed mean of the two terms of $\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Bures}(\Sigma_r, \Sigma_g)$: a lucky seed is closer in mean *and* better aligned in covariance. (b,c) Additive share of $\mathrm{Var}(\mathrm{FID})$ carried by each term, $[\mathrm{Var}(X) + \mathrm{Cov}(M, C)]/\mathrm{Var}(\mathrm{FID})$, across training seeds (training lottery), across subsets of one seed (generation lottery) and per evaluation, for conditional and unconditional sampling of the same networks.

H TEXT-TO-IMAGE STUDY

Setup. We train a 260.2M-parameter multimodal transformer with JiT-style blocks (Li & He, 2025) and patch size 32, directly in RGB pixel space at 512×512 with flow matching, conditioned on a frozen FLAN-T5-Large encoder (Chung et al., 2024). Training data is CC12M (Changpinyo et al., 2021) (10.97M pairs, all with synthetic recaptions). Ten runs differ only in their seed: batch 1024, 250k steps (256M samples, $\approx$23 epochs), 16 H100 per run ($\approx$25.5 h), checkpoints every 25k steps, $\approx$4k H100-hours in total. Evaluation uses EMA weights (decay 0.99995), 100 Euler steps and CFG scale 5.0 unless stated.

Design. Each of the 10 models is scored under the same 10 evaluation seeds (4 images per prompt), a crossed 10×10 grid. Evaluation seed e fixes the initial noise of every model (common random numbers), so we fit the random-effects model $y_{me} = \mu + a_m + b_e + r_{me}$ with a training effect a_m (σ_{model}), an evaluation-seed effect b_e shared by all models (σ_{seed}) and a residual r_{me}, the part of the sampling noise specific to one model (σ_{resid}). The components are the standard two-way ANOVA estimators, and the sampling spread within a model is $\sigma^2_{\text{within}} = \sigma^2_{\text{seed}} + \sigma^2_{\text{resid}}$.

Training versus sampling. On both benchmarks the training seed dominates (Table 5): the model effect is significant (F-test, $p < 10^{-25}$) and σ_{model} is $1.8\times$ (GenEval) and $2.4\times$ (DPG-Bench) σ_{within}, against $3.2\times$ for Inception FID on ImageNet. The shared evaluation-seed effect is small (4% and 2% of the variance) but significant ($p = 0.003$ and 0.01), so scoring all models under the same evaluation seed, as we do, removes it from comparisons. Because the variance sits in the model, extra evaluation

Table 5: **Training-seed variance on the text-to-image model** (10 training × 10 evaluation seeds, 250k steps, CFG 5.0). Training share is $\sigma^2_{\text{model}}/(\sigma^2_{\text{model}} + \sigma^2_{\text{seed}} + \sigma^2_{\text{resid}})$; $\bar{\rho}$ is the mean Spearman correlation of the model ranking between two evaluation seeds. GenEval tasks are sorted by training share.

metric	mean	σ_{model}	σ_{seed}	σ_{resid}	training share	$\bar{\rho}$
GenEval	0.502	0.0131	0.0031	0.0067	76%	0.73
DPG-Bench	77.81	0.636	0.098	0.250	85%	0.84
position	0.158	0.0335	0.0030	0.0150	83%	0.81
color attribution	0.391	0.0236	0.0038	0.0158	68%	0.74
two objects	0.555	0.0219	0.0092	0.0210	48%	0.41
counting	0.344	0.0143	0.0058	0.0195	33%	0.34
single object	0.886	0.0097	0.0085	0.0126	29%	0.36
colors	0.679	0.0093	0.0064	0.0141	26%	0.29

2σ band of one training run at one evaluation seed: GenEval 0.030, DPG 1.38, FID (EMA) 0.42
2σ band of re-scoring one checkpoint ($2\sigma_{\text{within}}$): GenEval 0.015, DPG 0.54
difference of two runs at a shared evaluation seed ($2\sqrt{2(\sigma^2_{\text{model}} + \sigma^2_{\text{resid}})}$): GenEval 0.042, DPG 1.93

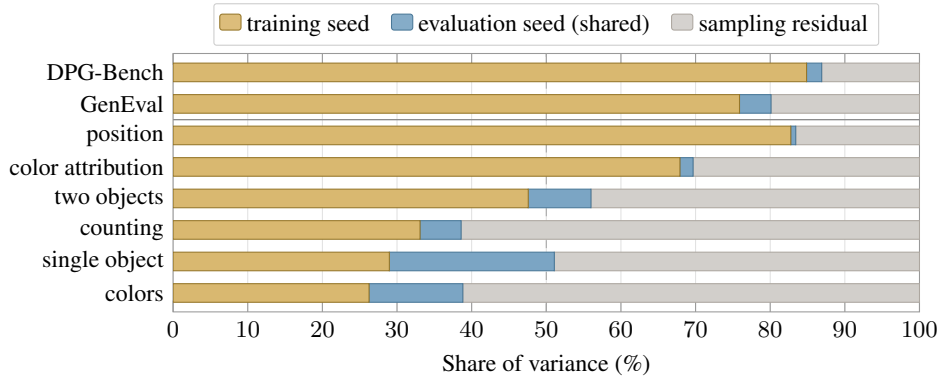

Figure 14: **Where the text-to-image variance comes from.** Variance decomposition of the 10×10 grid of training and evaluation seeds (Table 5) for the two benchmarks (top) and the six GenEval tasks (bottom, sorted by training share). The training seed dominates both benchmarks and the compositional tasks (position, color attribution); the easy tasks are sampling-dominated. The evaluation seed shared across models contributes little except on single object and colors.

seeds buy little: the standard deviation of one run's score falls from 0.0150 to 0.0133 GenEval (-12%) and from 0.69 to 0.64 DPG (-7%) when averaging 10 evaluation seeds, and would fall by at most 13% and 8% with infinitely many. Only more training seeds shrink it. The split across GenEval tasks (Figure 14) is not uniform. Single object, colors and counting are sampling-dominated (26–33% training share, model rankings barely reproducible, $\bar{\rho} \leq 0.36$). Position and color attribution, the compositional skills, are training-dominated (83% and 68%, $\bar{\rho} \geq 0.74$): whether a model binds attributes and places objects correctly is decided by its training seed. Position also has the largest relative spread across models (CoV 21%, per-model means 0.107–0.209), so per-task claims on compositional skills from a single training run are the least reliable.

Prompt-level consistency. The same holds prompt by prompt. For each prompt we test whether the 10 models differ in success rate across the 10 evaluation seeds (one-way F-test, Benjamini–Hochberg at 5%). The model effect is significant on 226 of the 519 GenEval prompts that are neither always solved nor always failed (≈ 26 expected by chance) and on 713 of the 1 065 DPG-Bench prompts. It is reproducible: a model's success rate on a prompt over evaluation seeds 0–4 predicts it over seeds 5–9 (median Spearman $\rho = 0.47$ GenEval, 0.54 DPG). The effect is mostly graded rather than all-or-nothing: no GenEval prompt is solved on all 40 images by one model and on none by another, and 8 prompts separate a model above 75% from one below 25% (0.1 expected under permutation).

FID over training. MS-COCO FID (10k images, CFG 2.0, logged online) on EMA weights has a CoV of 17.7% at 50k steps, 4.3% at 100k, 1.9% at 150k and 0.6–1.4% from 200k to 250k, ending at 15.33 ± 0.21. Raw (non-EMA) weights are $\approx 3\times$ noisier at the same checkpoint (4.0% at 250k), so EMA weights should be reported. The online FID draws its initial noise from the training RNG, so it

has no within-model control and mixes the two lotteries; extrapolating the 76–85% training share of GenEval and DPG would give $\sigma_{\text{model}} \approx 0.18$–$0.19$ FID, an inference rather than a measurement.

No lucky seeds. Within one checkpoint and one metric the model ranking is stable across evaluation seeds ($\bar{\rho} = 0.73$ GenEval, 0.84 DPG), so a model's advantage there is real, not benchmark noise. It does not transfer, however. Seed rankings by FID across the 9 checkpoints have mean $|\rho| = 0.25$ over the 36 checkpoint pairs, with $1/36$ significant, and at 250k the three metrics disagree on which seed is best, even with GenEval and DPG averaged over 10 evaluation seeds (GenEval vs. DPG $\rho = 0.47$, GenEval vs. FID $|\rho| = 0.31$, DPG vs. FID 0.18, all n.s.). As on ImageNet early in training, best-of-N seed selection does not generalise.

Guidance dominates. On seed 0, GenEval rises monotonically with the CFG scale (0.33, 0.43, 0.48, 0.51, 0.52, 0.54 at $\omega = 2, \ldots, 7$) without saturating, a shift of $\approx 16\times$ the training-seed σ. The guidance scale must therefore be fixed and reported before any seed-level comparison.

Caveats. The study covers one configuration, one checkpoint (250k) and one guidance scale for the full grid. The configuration is an ablation-scale $B/32$ model, and GenEval is scored with an adapted mmdet 3.x detector, so absolute scores are not comparable to published GenEval numbers. About 31.5% of the source images have a short side below 512 and are upscaled.

Qualitative variability. Figure 15–17 show samples from nine of the ten models on PartiPrompts (Yu et al., 2022) prompts, with the initial noise shared across models, so that any difference along a row comes from the training seed alone. On 100 prompts (4 noises each), the mean pairwise pixel distance (at 32^2) between models at a fixed noise is $0.47\times$ (median; range 0.25–0.81) the distance between noises for a fixed model: the noise sets the composition. Yet which models succeed on a prompt is not random: reading each column of Figure 16–17 down its four noises, some training seeds succeed on every noise while others never do on these prompts. Nets 0, 6 and 8 draw a smiling banana from all four noises and nets 3 and 4 from none; net 4 always draws the eye of Jupiter and net 6 always the bare planet; nets 0, 6 and 7 fill the volcano plume with fish every time. Despite the large visual variability, some models are simply worse at some tasks because of their training seed, and no choice of noise fixes it: the qualitative counterpart of the prompt-level analysis above, where over 40 images the effect is usually graded rather than all-or-nothing.

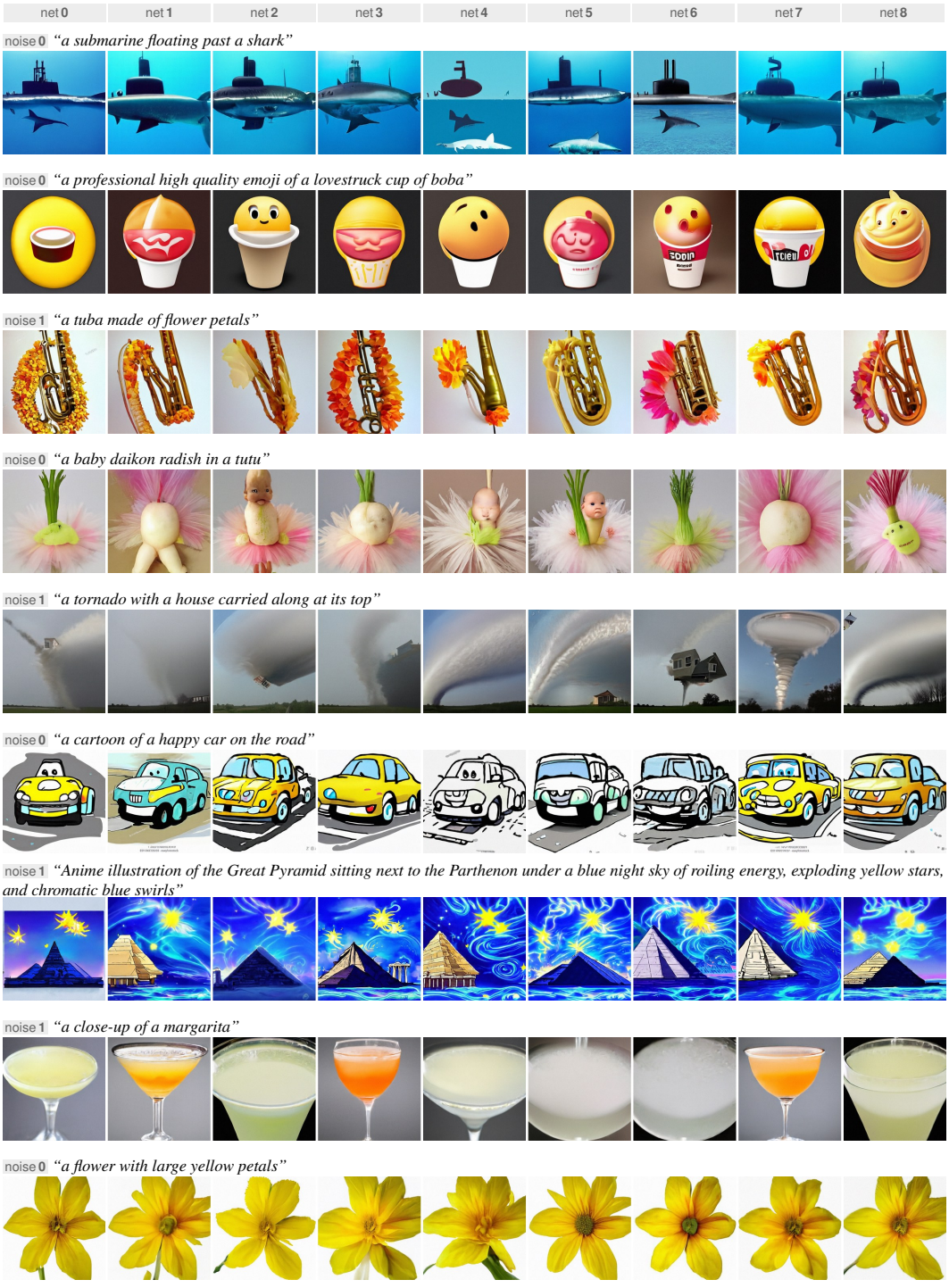

Figure 15: **Same noise, different training seed.** Each row is one PartiPrompts prompt sampled from the same initial noise by the 9 of the 10 text-to-image models of Appendix H, which differ only in their training seed (250k steps, EMA, 100 Euler steps, CFG 5.0). The noise fixes the layout and palette, but the training seed decides what is drawn: whether the shark is a separate animal or fused with the hull, whether the tuba is made of petals or merely garlanded, whether the radish gets a baby face, whether the house is carried by the tornado, and whether the margarita is pale or orange. The last row is a prompt on which the nine models agree.

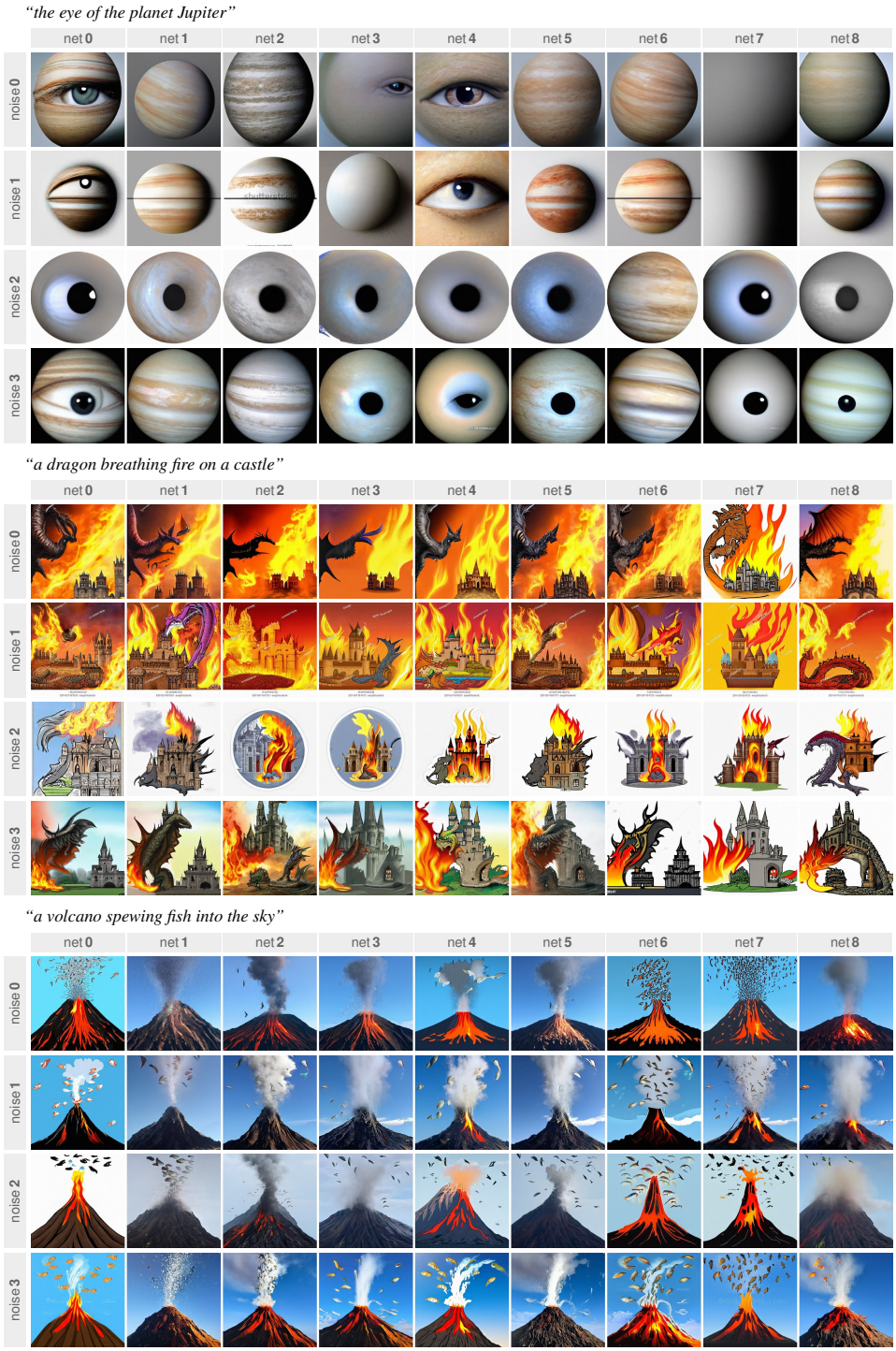

Figure 16: **Noise seed × training seed.** Rows are 4 initial noises (shared by every column), columns the 9 training seeds. Moving down a column (new noise, same model) changes the composition; moving along a row (same noise, new model) keeps the composition but can change the reading of the prompt. For *"the eye of the planet Jupiter"*, nets 0, 3 and 4 draw an eye from noise 0 while the others draw the planet (net 7 only a gradient), and from noise 2 every model but net 6 draws an eyeball. Reading down a column, net 4 draws an eye from all four noises and net 6 from none. For the dragon, noise 2 gives a sketch, a circular badge or a clip-art sticker depending on the model. From noise 0, only nets 0, 6 and 7 turn the plume itself into fish, and they do so from every noise.

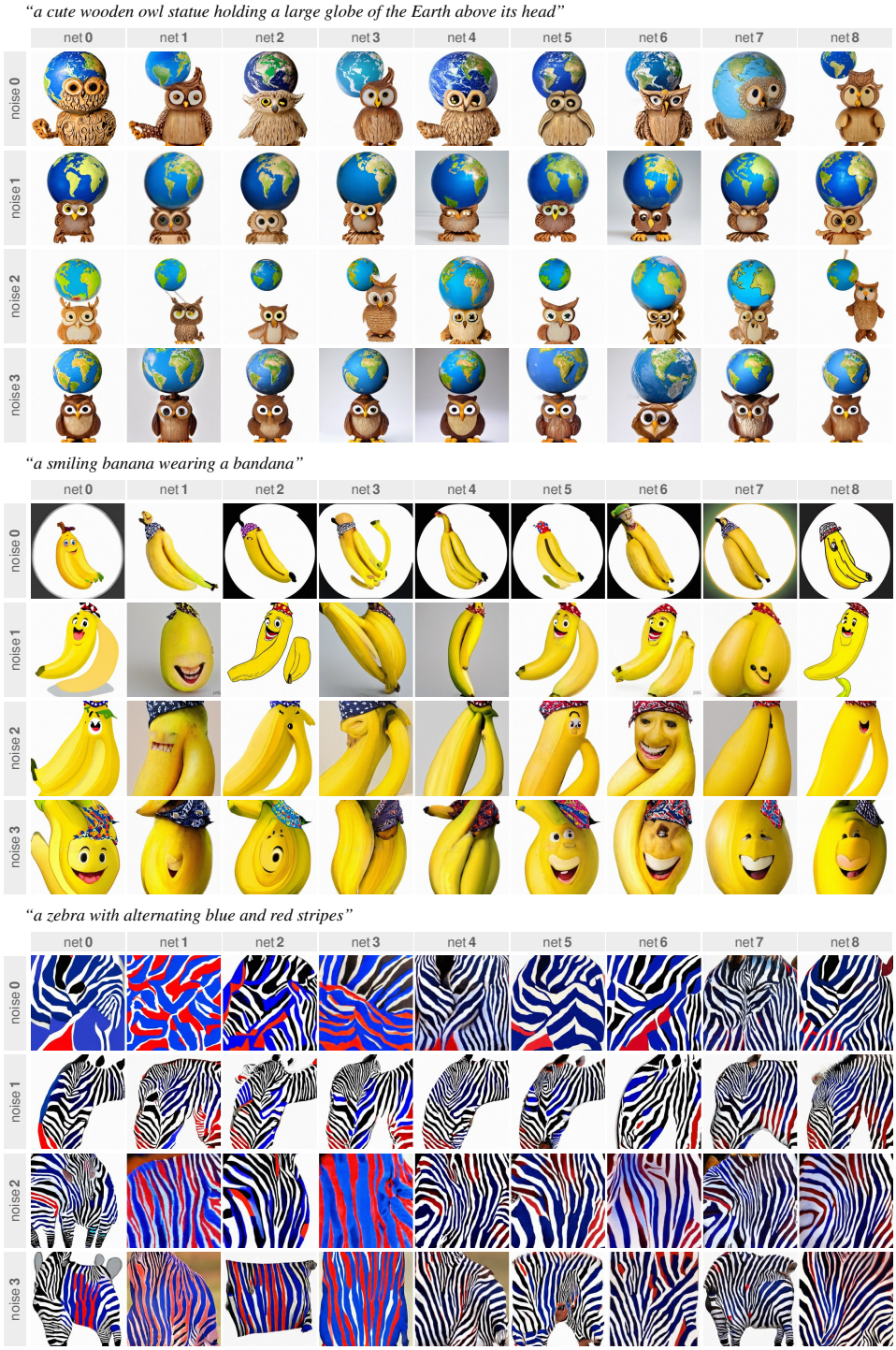

Figure 17: **Noise seed × training seed (continued).** Same layout as Figure 16. Prompt adherence is a property of the model rather than of the noise: nets 0, 6 and 8 give the banana a smiling face from all four noises, whereas nets 3 and 4 never do. Likewise, from noises 0, 2 and 3, nets 1 and 3 paint the whole zebra blue and red, while nets 4–8 keep black-and-white stripes with colour accents.

I Replication across DINOv2 FID and Inception PRDC

The main paper anchors the seed-lottery analysis on Inception FID. This appendix re-runs each of the five Sec. 4 questions on five complementary metrics: DINOv2 FID (Stein et al., 2023) and Inception precision, recall, density, coverage (Kynkäänniemi et al., 2019; Naeem et al., 2020). The same panels of Sec. 3 are reused, so the columns of Table 6, Table 8, Table 10 and Table 11 parallel the main-paper tables row-for-row. Each subsection rescores the panel of the matching main-text subsection, with σ_{between} the standard deviation of the per-seed means and σ_{within} the average per-seed standard deviation. Sec. 4.3 summarises these results in the main text.

I.1 Metric definitions

Notation. R and G are the real and generated feature sets (50k each; R is the same ImageNet-train subset on every evaluation), embedded by Inception-V3 (Inception FID and PRDC) or DINOv2 (DINOv2 FID). $\rho_k^S(x)$ is the distance from x to its k-th nearest neighbour in S, with $k=3$ as in Naeem et al. (2020), and $B(x,\rho)$ is the closed ball of radius ρ around x.

Metrics. Both FIDs fit a Gaussian to each feature set and take the Fréchet distance (Heusel et al., 2017) $\|\mu_R-\mu_G\|^2+\operatorname{tr}\big(\Sigma_R+\Sigma_G-2(\Sigma_R\Sigma_G)^{1/2}\big)$; DINOv2 FID (Stein et al., 2023) is roughly ten times larger than Inception FID at matched quality, and lower is better for both. The PRDC metrics (Kynkäänniemi et al., 2019; Naeem et al., 2020) replace the Gaussian fit by k-NN balls:

$$\text{Precision}=\tfrac{1}{|G|}\textstyle\sum_g \mathbf{1}\big[\exists r: g\in B(r,\rho_k^R(r))\big],\qquad \text{Recall}=\tfrac{1}{|R|}\textstyle\sum_r \mathbf{1}\big[\exists g: r\in B(g,\rho_k^G(g))\big],$$
$$\text{Density}=\tfrac{1}{k|G|}\textstyle\sum_g\sum_r \mathbf{1}\big[g\in B(r,\rho_k^R(r))\big],\quad \text{Coverage}=\tfrac{1}{|R|}\textstyle\sum_r \mathbf{1}\big[\exists g: g\in B(r,\rho_k^R(r))\big].$$

Precision and density measure fidelity, recall and coverage diversity; higher is better for all four (density is unbounded above, the others lie in $[0,1]$).

Where the balls live. Precision, density and coverage test points against balls built on the fixed real set R. Recall alone builds its balls on the generated set G, which is redrawn at every sampling seed (Figure 18); Naeem et al. (2020) introduced coverage precisely as the real-anchored counterpart of recall. This asymmetry explains recall's anomalous behaviour throughout this appendix (Sec. I.2).

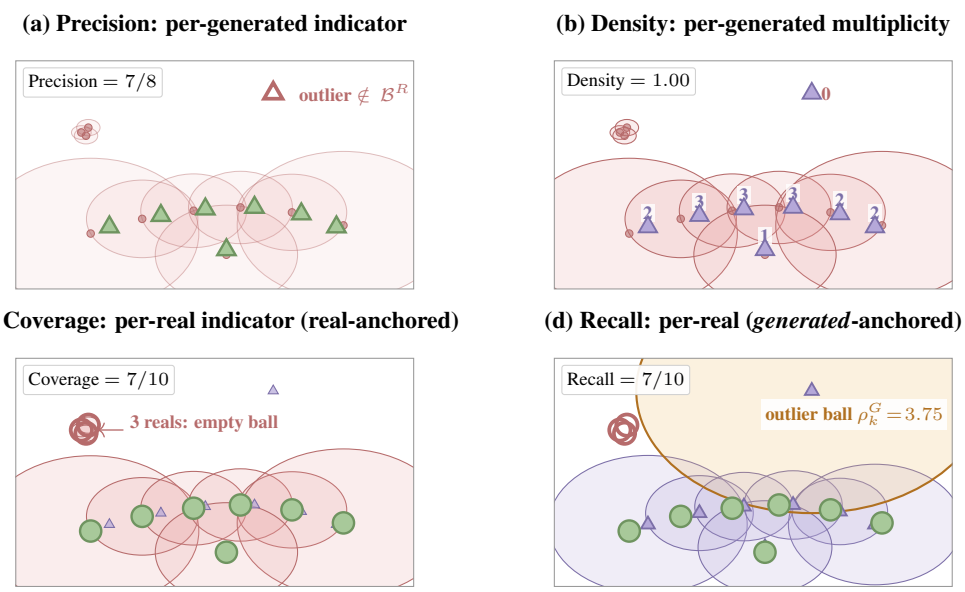

Figure 18: **The four PRDC metrics on a 2D toy.** *Marker convention:* circles (●) are *real* points and triangles (▲) are *generated* points. The larger marker in each panel is the entity the metric tests. In panels (a, c, d) the metric is a per-point indicator, drawn green (● / ▲ = counted / covered) or red and hollow (○ / △ = not counted / uncovered). Lavender (▲) marks generated points whose coverage status is not relevant in that panel. Coral discs are real-anchored k-NN balls and lavender discs are generated-anchored balls. The amber disc in (d) is the outlier generated point's enlarged ball. *Scenario:* reals sit on a curved manifold plus an isolated mini-mode of three tightly grouped points in the upper-left. The generated points cover the manifold but miss the isolated mode and add one outlier far above. (a) *Precision* = $7/8$: each generated point is tested against the union of real-anchored balls, and the outlier triangle is the only hollow one. (b) *Density* = 1.00: same balls, but each generated point is labelled with the count of real-balls it lies in (a multiplicity, not an indicator). Cluster points score 1–3, while the outlier scores 0. (c) *Coverage* = $7/10$: each real point is tested against the same real-anchored balls. The three isolated reals have tiny k-NN balls ($\rho_k^R \approx 0.2$) that no generated point reaches, so they appear as red hollow circles. (d) *Recall* = $7/10$: real points are tested against generated-anchored balls instead. The outlier generated point's ball ($\rho_k^G = 3.75$, amber) is $1.8\times$ larger than any cluster ball and engulfs much of the manifold. The three isolated reals remain hollow, so recall coincides with coverage *on this draw*, but *by a different mechanism*: in (c) the balls are fixed across draws of G, while in (d) every sampling seed redraws the entire ball geometry, inflating the within-seed σ above the binomial floor (Table 7).

I.2 The training lottery dominates for distributional metrics, not for recall

The $3.2\times$ training-vs-sampling asymmetry of Sec. 4.1 grows to $4.8\times$ under DINOv2 FID, shrinks to $\approx 1\times$ under Inception precision/density/coverage, and inverts to $0.28\times$ under Inception recall.

Table 6: **Sampling vs. training lottery on the 25×10 SiT-B/2 panel.** Companion to Figure 2 and the headline of Sec. 4.1. σ_{between} is the standard deviation of the 25 per-seed means. σ_{within} is the within-seed sampling standard deviation, averaged across the 25 training seeds. $\text{CoV}_{\text{between}} = \sigma_{\text{between}}/|\mu|$. Higher-better metrics are marked ↑, lower-better ↓. Bold marks the largest ratio (most training-lottery-dominated metric) and the smallest CoV in each direction column.

Metric	μ	σ_{between}	σ_{within}	$\text{CoV}_{\text{between}}$ (%)	$\sigma_{\text{between}}/\sigma_{\text{within}}$
Inception FID ↓	34.74	0.438	0.137	1.26	3.19×
DINOv2 FID ↓	708.6	6.106	1.268	**0.86**	**4.82**×
Inception Precision ↑	0.485	3.24e-3	2.84e-3	0.67	1.14×
Inception Recall ↑	0.314	1.21e-3	4.38e-3	**0.39**	0.28×
Inception Density ↑	0.442	5.26e-3	3.74e-3	1.19	1.41×
Inception Coverage ↑	0.224	1.75e-3	1.56e-3	0.78	1.12×

DINOv2 FID amplifies the asymmetry. The between-to-within ratio rises from $3.19\times$ on Inception FID to $4.82\times$ on DINOv2 FID (Table 6). The relative floor tightens from $\mathrm{CoV}_{\text{between}} = 1.26\%$ to 0.86%. Figure 19 visualises the consequence: the black per-seed mean ticks stagger over a wider multiple of the violin height than under Inception FID, making the dominance of the training lottery even more visible. Switching the feature extractor strengthens the main-paper conclusion.

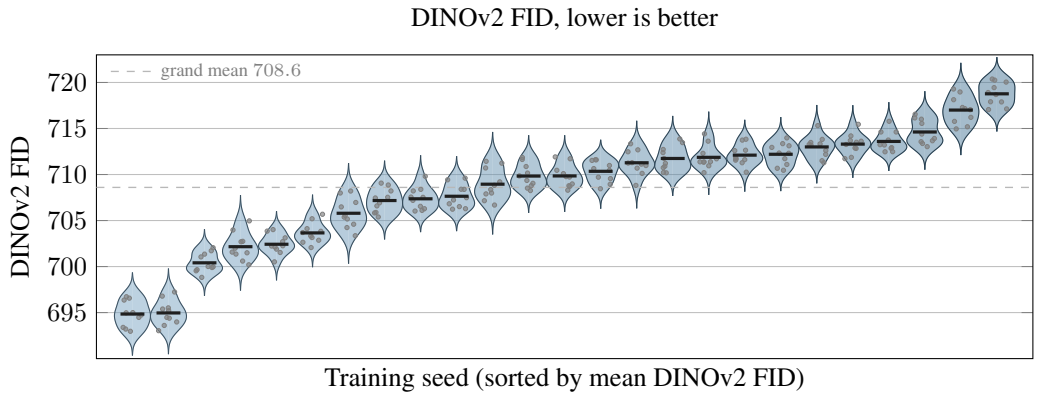

Figure 19: **Seed lottery on the same SiT-B/2 panel under DINOv2 FID.** Each violin is one of 25 trained models, sorted left-to-right by per-seed mean. Violin shape traces the within-seed sampling distribution. The black tick is the per-seed mean. The training-seed mean ticks span 694.8 to 718.8 around grand mean 708.6. The between-to-within ratio is $\sigma_{\text{between}}/\sigma_{\text{within}} = 4.82\times$, against $3.19\times$ for Inception FID, so the mean ticks span a wider multiple of the violin height than in Figure 2.

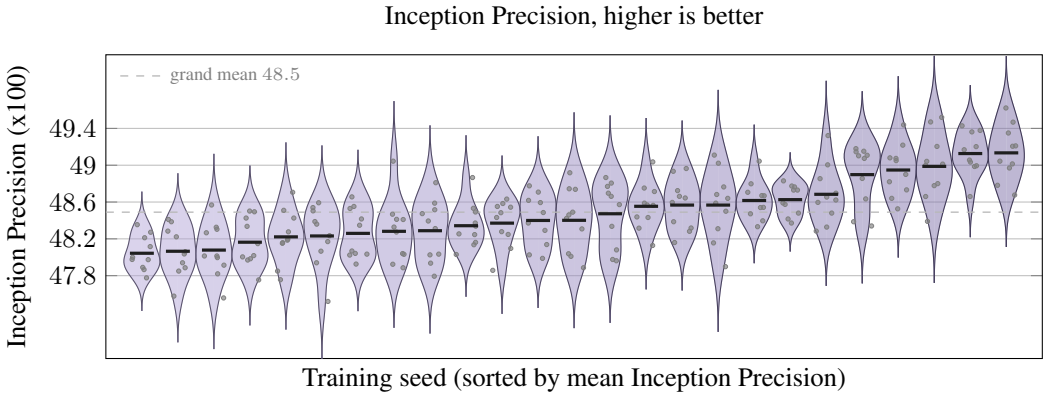

Figure 20: **Seed lottery under Inception Precision.** Y-axis values are multiplied by 100 for readability. Per-seed means span 0.480 to 0.491 around grand mean 0.485. Violin height (within-seed sampling spread) and the staircase of mean ticks (between-seed training spread) are comparable, since $\sigma_{\text{between}}/\sigma_{\text{within}} = 1.14\times$. A multi-sampling-seed CI therefore covers roughly half of the seed-induced envelope on precision, in contrast with FID where it covers a quarter.

Precision, density, and coverage put the two lotteries near parity. The ratio drops to roughly $1\times$ on the three real-anchored PRDC metrics (1.14 for precision, 1.41 for density, 1.12 for coverage, Table 6). Multi-sampling-seed CIs cover roughly half of the true seed-induced envelope on these three metrics, so the sampling-only confidence interval is no longer asymptotically irrelevant the way it is for FID. Figure 20 shows the parity: violin heights and per-seed-mean staircases occupy comparable vertical extents.

Recall inverts the asymmetry. On Inception recall the ratio is $0.28\times$: the within-seed sampling σ is over three times the between-seed training σ. Figure 21 shows individual violins taller than the spread of the per-seed means. Recall is the only metric in the bundle for which the sampling-seed CI is the right CI to report on a fixed trained model, and the one metric where the main claim inverts.

Why recall inverts: the k-NN balls move with the sampling seed. Precision, density and coverage test the 50k generated points against balls built on the fixed real set, so a new sampling

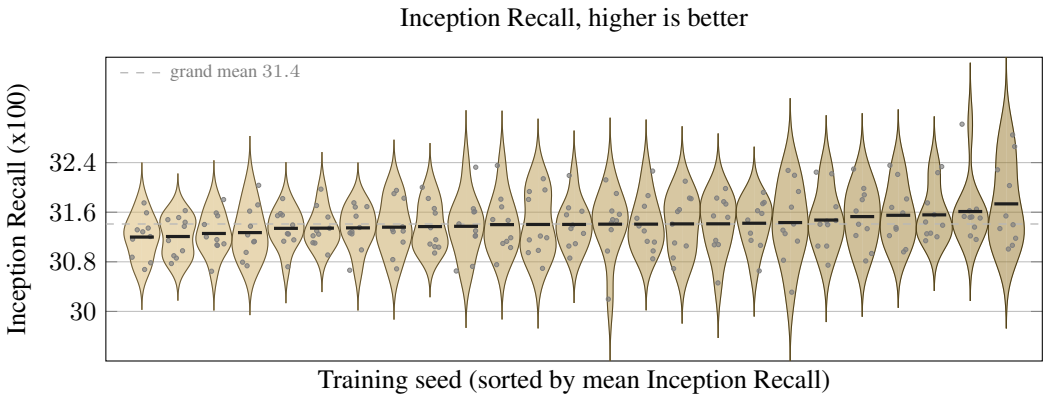

Figure 21: **Seed lottery under Inception Recall (the inversion case).** Y-axis values are multiplied by 100 for readability. Per-seed means span 0.312 to 0.317 around grand mean 0.314. Each violin is taller than the staircase of mean ticks: $\sigma_{\text{between}}/\sigma_{\text{within}} = 0.28\times$, an inversion of the FID asymmetry. On Inception recall the right CI to report on a fixed trained model is the sampling-only one. This is the opposite recommendation from FID and from the other PRDC metrics (Table 6).

seed only redraws nearly independent indicators and the within-seed σ sits near the binomial floor $\sqrt{p(1-p)/N}$. Recall builds its balls on the generated set: a new sampling seed moves the whole ball geometry, many real points flip together, and the effective sample size is far below 50k. Table 7 confirms it: the three real-anchored metrics sit at 0.84–1.68$\times$ the floor, recall at 2.11$\times$. With a saturated recall axis ($\mu = 0.314$, similar for all converged networks, hence a small σ_{between}), this produces the inversion.

Table 7: **Within-seed σ versus the binomial estimator floor for each PRDC metric.** The floor is $\sqrt{p(1-p)/N}$ with $N = 50\,000$ and $p = \mu$. Only recall, the one generated-anchored metric, sits well above it.

Metric	μ	σ_{within}	$\sqrt{p(1-p)/N}$	ratio (obs / floor)
Inception Precision ↑	0.485	0.00284	0.00224	1.27× (real-anchored)
Inception Recall ↑	0.314	0.00438	0.00208	**2.11× (generated-anchored)**
Inception Density ↑	0.442	0.00374	0.00222	1.68× (real-anchored)
Inception Coverage ↑	0.224	0.00156	0.00186	0.84× (real-anchored)

I.3 Flow-matching noise dominates for the fidelity metrics

The noise > init > data *hierarchy of Sec. 4.2 carries over to DINOv2 FID and Inception precision, density and coverage, with the same sub-additivity. Recall is too sampling-dominated to disentangle.*

The hierarchy holds for the four well-behaved metrics. Inception FID, DINOv2 FID, Inception Precision, Inception Density, and Inception Coverage all give the same ranking VARY-NOISE > VARY-INIT > VARY-DATA. The flow-matching noise alone reproduces about three-quarters of the baseline σ on every metric. Init alone recovers about two-thirds, and data order alone about half (per-metric numbers in Table 8). Adding the three single-source σ values in quadrature overshoots $\sigma_{\text{vary-all}}$ by 7–19%, mirroring the 13% overshoot on Inception FID. Three takeaways follow. (i) The noise dominance is a property of the loss formulation, not of the feature extractor. (ii) The sub-additivity is a stable two-digit number across feature spaces. (iii) Init is the second-cheapest knob in every case.

Recall fails to decompose. On Inception recall, three of the four single-source conditions give a between-seed σ *larger* than the baseline (112–126%) and the quadrature sum overshoots by 90%: with σ_{between} below σ_{within} (Sec. I.2), every estimate is dominated by the same sampling jitter, so the decomposition does not apply.

Table 8: **Variance decomposition across feature spaces.** Companion to Table 1. Per-condition σ_{between} as a percentage of the fully-stochastic $\sigma_{\text{vary-all}}$, plus the naive sum-of-squares overshoot $\sqrt{\sigma^2_{\text{noise}} + \sigma^2_{\text{init}} + \sigma^2_{\text{data}}}/\sigma_{\text{vary-all}} - 1$. Underlines mark the leading single source per metric.

Metric	vary-noise (%)	vary-init (%)	vary-data (%)	vary-noisedata (%)	overshoot (%)
Inception FID ↓	<u>76.3</u>	67.0	49.8	59.2	+13.1
DINOv2 FID ↓	<u>76.9</u>	72.4	55.1	59.7	+19.2
Inception Precision ↑	<u>66.0</u>	64.4	54.9	57.9	+7.3
Inception Density ↑	<u>73.3</u>	64.0	53.1	57.7	+10.9
Inception Coverage ↑	<u>72.3</u>	69.2	59.4	56.1	+16.4
Inception Recall ↑	124.1	90.2	112.6	126.1	+90.3

I.4 Guided metrics beyond Inception FID

The CoV-halving claim of Sec. 4.4 is specific to Inception FID. On DINOv2 FID the relative noise floor is already low (0.86% unguided) and stays at 0.94% at the Inception-FID-optimal CFG. The recovered CFG also shifts every PRDC metric's operating point by a large amount.

Setup. The search of Algorithm 1 (Sec. J.1) optimises Inception FID on the same panel; the other metrics are *read out* at the Inception-FID-optimal CFG, not optimised separately.

DINOv2 FID's floor is not improved by Inception-FID guidance. Guidance tuning drops Inception FID's $\text{CoV}_{\text{between}}$ from 1.26% to 0.67% (Sec. 4.4). On DINOv2 FID, $\text{CoV}_{\text{between}}$ moves from 0.86% to 0.94% on the same CFG-selected samples, statistically indistinguishable. The 0.86% unguided DINOv2-FID floor already sits below the 1–2% band, so the extra FID evaluations per cell that guided FID costs do not buy a tighter DINOv2-FID estimate.

The recovered CFG sets the PRDC operating point. Classifier-free guidance trades recall for fidelity, and the Inception-FID-optimal CFG picks an operating point that strongly favours fidelity. On Inception precision, density, and coverage the mean roughly doubles or triples once guided, while recall drops by more than a factor of two (Table 9). Guided PRDC numbers therefore measure a *different operating point* than the unguided numbers (a high-fidelity, low-recall regime) and should not be compared with unguided PRDC across studies.

Table 9: **Effect of per-cell optimal CFG on each metric.** Companion to Sec. 4.4. "Unguided" is the 25×10 panel. "Guided" is the same panel rescored at the per-(training, sampling) CFG that minimises Inception FID. Means shift sharply on PRDC because guidance reweights fidelity vs. recall. The CoV column shows that the relative noise floor changes substantially only for Inception FID.

Metric	μ unguided	μ guided	CoV unguided (%)	CoV guided (%)
Inception FID ↓	34.74	7.42	1.26	**0.67**
DINOv2 FID ↓	708.6	289.8	0.86	0.94
Inception Precision ↑	0.485	0.843	0.67	0.24
Inception Recall ↑	0.314	0.135	0.39	1.16
Inception Density ↑	0.442	1.320	1.19	0.50
Inception Coverage ↑	0.224	0.295	0.78	0.17

I.5 The relative noise floor is metric-dependent but scale-invariant

The 1–2% CoV band of Sec. 4.5 covers DINOv2 FID, Inception precision, density, and coverage at every checkpoint of every model size. Recall keeps a steady 1–1.8% CoV on SiT-B/L.

Setup. Figure 22 and Table 10 rescore the 200k–2M sweep with the seed sets of Table 2.

DINOv2 FID is tighter than Inception FID at every scale. The DINOv2 band sits inside the Inception band across the 76 cells with similar σ-shrink factors, so DINOv2 FID is the tighter benchmark at every model size and every checkpoint past 200k.

Precision, density, and coverage hit the lowest floors at scale. Inception precision, density, and coverage on SiT-XL all settle below 0.4% CoV by 1M steps. The median CoVs of precision and coverage on SiT-XL drop below 0.3%, with maxima below 0.7%. Three fidelity-axis metrics on a

Table 10: **Per-(metric, model) CoV bands across the 200k–2M sweep.** Companion to Table 2. Each cell gives the min/median/max $\mathrm{CoV}_{\text{between}}$ across the 19 checkpoints and the σ-shrink $\sigma(200\mathrm{k})/\sigma(2\mathrm{M})$. Shrink factors below 1 (**bold**) mark metrics whose spread *grows* with compute.

Metric	Model	min CoV (%)	median CoV (%)	max CoV (%)	σ-shrink
Inception FID ↓	SiT-S	0.74	0.89	1.12	2.46×
	SiT-B	1.01	1.13	1.74	1.68×
	SiT-L	1.42	1.65	2.06	1.68×
	SiT-XL	1.23	1.44	1.74	2.16×
DINOv2 FID ↓	SiT-S	0.48	0.57	0.66	1.69×
	SiT-B	0.57	0.73	1.39	1.07×
	SiT-L	1.11	1.47	1.59	1.29×
	SiT-XL	1.25	1.43	1.51	1.59×
Inception Precision ↑	SiT-S	0.52	0.64	1.18	1.30×
	SiT-B	0.40	0.53	0.79	1.18×
	SiT-L	0.29	0.42	0.71	1.35×
	SiT-XL	0.21	0.30	0.67	2.27×
Inception Recall ↑	SiT-S	0.34	0.42	0.73	1.94×
	SiT-B	1.17	1.40	1.83	**0.85**×
	SiT-L	1.04	1.36	1.74	**0.86**×
	SiT-XL	0.31	0.45	0.50	1.03×
Inception Density ↑	SiT-S	0.85	1.06	1.73	**0.87**×
	SiT-B	0.69	0.96	1.21	**0.81**×
	SiT-L	0.63	0.82	1.33	1.17×
	SiT-XL	0.40	0.58	1.24	2.25×
Inception Coverage ↑	SiT-S	0.63	0.83	1.66	1.28×
	SiT-B	0.57	0.71	1.19	1.03×
	SiT-L	0.35	0.46	0.92	1.85×
	SiT-XL	0.18	0.21	0.66	2.80×

large model produce the tightest seed-induced spread on this panel, sitting an order of magnitude below the 3–4% headline gain typical of recipe-comparison studies.

Recall does not respond to compute. On SiT-B and SiT-L, recall σ *grows* between 200k and 2M (σ-shrink below $1\times$) and on SiT-XL it stays flat, so the 1–1.8% recall band is a property of the metric, not of under-training. Its rankings are unstable too: Spearman ρ with the 2M ranking stays below 0.6 at every earlier checkpoint and dips below zero on SiT-S/L/XL early in training (Figure 23).

The Gaussian shape carries over to every metric. $(\max-\min)/\sigma$ stays in $[2.9, 5.5]$ across all 76 cells of each metric, bracketing the 3.5–3.9 a Gaussian sample of size $n \in [19, 25]$ predicts. For every metric the seed lottery is a smooth spread, not a tail of training failures.

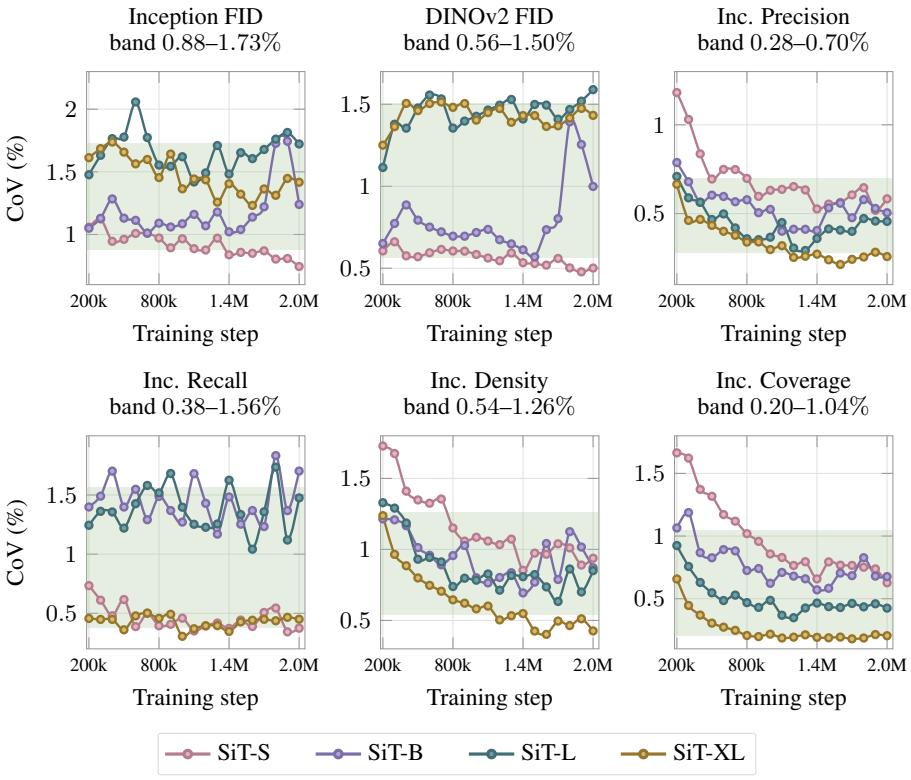

Figure 22: **Coefficient of variation across compute and scale, for six metrics.** Each panel plots between-seed $\mathrm{CoV} = \sigma/\mu$ versus training step on the variance-over-training panel for SiT-S/B/L/XL with the clean seed sets of Table 2. The shaded green stripe in each panel is that metric's empirical $[p_{10}, p_{90}]$ envelope of CoV across its 76 $(\mathrm{model}, \mathrm{step})$ cells, reported in the panel title. Inception FID's envelope is 0.88–1.73%, matching the 1–2% headline of Figure 7b. DINOv2 FID (0.56–1.50%) and Inception precision/coverage (0.28–0.70%, 0.20–1.04%) have lower envelopes than Inception FID. Recall (0.38–1.56%) and density (0.54–1.26%) also sit lower. Each metric therefore has its own scale-invariant relative noise floor. The floor is wider for recall because σ does not shrink with compute on SiT-B/L (Table 10).

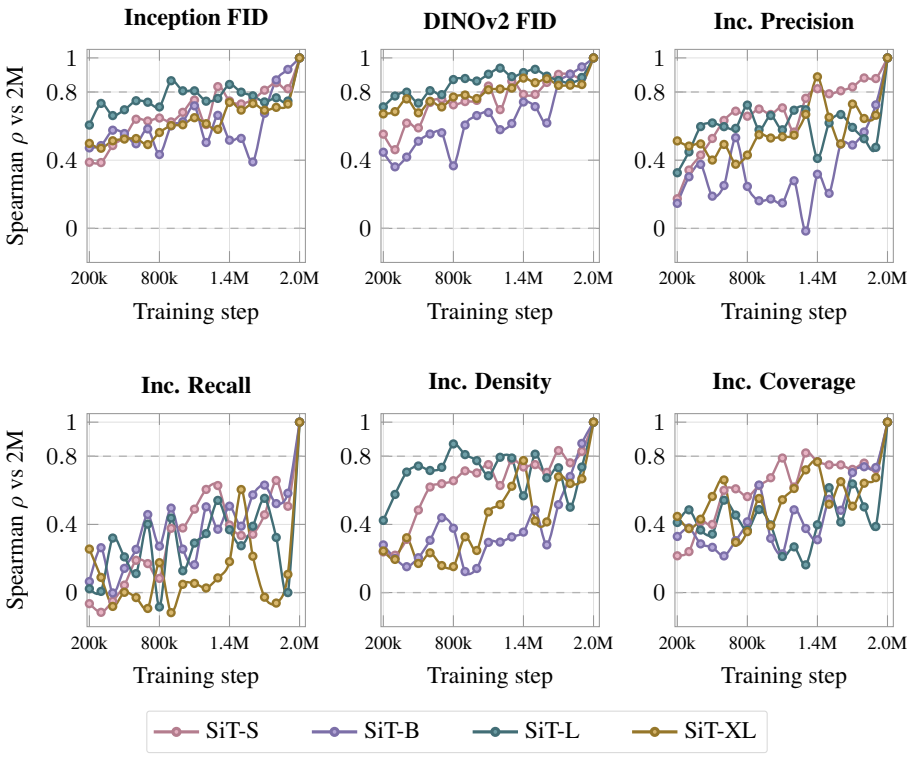

Figure 23: **Rank stability across compute, for six metrics.** Each panel plots Spearman ρ between the seed ranking at training step t and at 2M, one curve per model size, mirroring the layout of Figure 7c. Dashed horizontal lines mark $\rho = 0.8$ (the high-stability target) and $\rho = 0$ (random). Inception FID and DINOv2 FID approach $\rho \approx 0.8$ by ~ 1.4M on most models. Inception precision, density and coverage stabilise more slowly. Inception Recall is the outlier: ρ stays below 0.7 at every pre-final checkpoint on every model and dips negative early in training, so the recall ranking at any intermediate checkpoint is essentially uninformative about the eventual recall ranking.

I.6 CHERRY-PICKING SAVES **1.2–2.9**× ON EVERY WELL-BEHAVED METRIC

The lucky-seed speedup of Sec. 4.6 generalises across DINOv2 FID and Inception precision, density, and coverage, with the largest savings (2.0–2.9× on SiT-L/XL) on the fidelity PRDC metrics. Recall is compute-invariant on this panel, so the framing does not apply.

Setup. As in Figure 8, the lucky and unlucky seeds are the best and worst at 2M for the given metric (direction flipped for higher-better metrics), and the speedup is 2M divided by the first checkpoint at which the lucky seed matches the unlucky seed's 2M value.

The speedup transfers and grows on the fidelity axes. DINOv2 FID gives speedups in the same band as Inception FID (1.18–1.67× vs 1.18–1.82×, Table 11), so the lucky-seed effect is not specific to one feature extractor. Inception precision, density, and coverage push past it on the larger models, reaching 2.0–2.9× on SiT-L and SiT-XL. The "free training-time speedup" of Sec. 4.6 therefore strengthens when the benchmark is precision, density or coverage rather than FID.

Recall is compute-invariant. Recall's nominal 6–10× speedups are degenerate: the lucky seed already beats the unlucky seed's 2M value at 200k–300k, because recall barely responds to compute (Sec. I.5).

Table 11: **Lucky-seed speedup across metrics.** Companion to Figure 8. Each cell reports the speedup factor $2\mathrm{M}/t^\star$, where $t^\star$ is the first checkpoint at which the lucky seed reaches the unlucky-seed-at-2M target. Recall entries are degenerate (compute-invariant, see Sec. I.5).

Metric	SiT-S	SiT-B	SiT-L	SiT-XL
Inception FID ↓	1.25×	1.18×	1.82×	1.82×
DINOv2 FID ↓	1.18×	1.25×	1.67×	1.54×
Inception Precision ↑	1.25×	1.43×	**2.50**×	**2.50**×
Inception Density ↑	1.25×	1.33×	**2.86**×	2.00×
Inception Coverage ↑	1.18×	1.33×	**2.86**×	2.22×
Recall (degenerate, lucky seed already exceeds unlucky-at-2M target by 200–300k):				
Inception Recall ↑	(6.67×)	(10.0×)	(10.0×)	(10.0×)

I.7 A NON-GAUSSIAN DISTANCE: MIND

FID compares only the first two moments of the feature distributions, so one could worry that the training lottery is an artefact of its Gaussian approximation or of covariance estimation. We therefore rescore the converged SiT-B/2 panel of Sec. 4.1 ($N=25$ training seeds, $K=10$ sampling seeds, 400k steps, no CFG) with MIND (Berthet et al., 2026), the sliced squared Wasserstein-2 distance between generated and real features scaled by $3d$ (d the feature dimension). Each one-dimensional W_2^2 is computed exactly by sorting, over 1000 fixed random projections, so MIND makes no Gaussian assumption and estimates no covariance. We compute it in Inception-V3 and DINOv2 feature space, against the ImageNet train and validation references, and at 5k and 50k generated images per evaluation. FID is computed on the *same* generated images, so the MIND and FID columns of Table 12 differ only in the distance functional. All statistics follow Appendix E.

The lottery survives the change of distance. At the standard 50k images (Table 12, Figure 24a), MIND reproduces the asymmetry: $\sigma_{\text{between}}/\sigma_{\text{within}}=2.54\times$ in Inception space and $4.51\times$ in DINOv2 space, with one-sided 95% lower bounds of 2.02 and 3.60, and $\sigma_b \le \sigma_w$ is rejected at $p \le 5\times10^{-14}$. The ratio is somewhat smaller than for FID on the same images ($2.94\times$ and $6.75\times$), but the ordering across feature spaces is the same: DINOv2 is more training-dominated than Inception under both distances. The between-seed CoV of MIND is 1.10–1.49%, inside the 1–2% band of Sec. 4.5. The two distances also agree on *which* seeds are lucky: Spearman ρ between the per-seed means of MIND and FID is 0.99 in every feature space and reference set (Figure 24b,c). The training lottery is thus a property of the trained models, not of the Fréchet estimator.

The two lotteries scale differently with the evaluation budget. Because MIND and FID were computed at both 5k and 50k images on the same checkpoints and sampling seeds, we can separate how each variance component depends on the number of evaluated images (Table 13, Figure 24d). σ_{within} shrinks by 3.2–4.4× for 10× more images, close to the $3.16\times$ of a $1/\sqrt{n}$ estimator (the

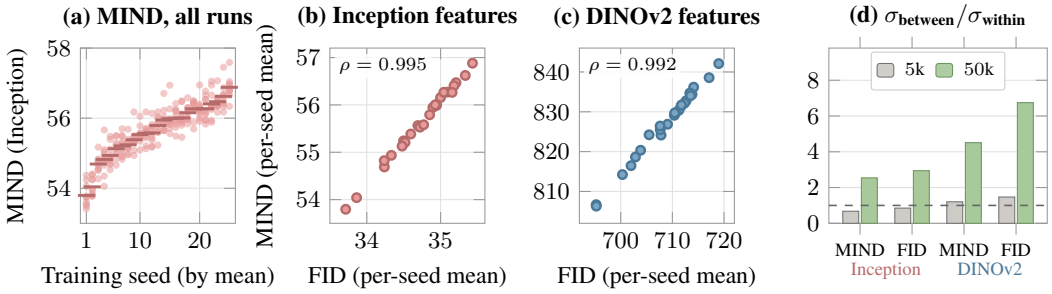

Figure 24: **The training lottery under MIND.** Converged SiT-B/2 panel (25 training seeds $\times$ 10 sampling seeds, 400k steps, no CFG, ImageNet train reference). (a) MIND in Inception space for all 250 evaluations at 50k images, one column per training seed sorted by its mean (ticks): as for FID (Figure 2), the spread across columns exceeds the spread within them. (b,c) Per-seed mean MIND against per-seed mean FID computed on the same images, in Inception and DINOv2 space: the two distances rank the 25 retrained models almost identically. (d) Ratio of between- to within-seed standard deviation at 5k and 50k generated images per evaluation. $10\times$ more images shrinks σ_{within} by $\approx\sqrt{10}$ and leaves σ_{between} unchanged, so the training lottery, below parity (dashed) in Inception space at 5k, dominates at 50k.

Table 12: **MIND and FID on the same generated images** (SiT-B/2, 400k steps, 25×10 panel, 50k images per evaluation, no CFG). σ_b/σ_w is given with its 95% confidence interval; p tests $\sigma_b \le \sigma_w$ (Appendix E). ρ is the Spearman correlation between the per-seed means of MIND and FID in the same feature space and reference set.

Metric (reference)	mean	σ_b	σ_w	σ_b/σ_w [95% CI]	CoV_b	p	ρ
MIND Inception (train)	55.64	0.770	0.303	**2.54** [1.94, 3.59]	1.38%	5.3×10^{-14}	0.995
MIND Inception (val)	48.43	0.722	0.283	2.55 [1.94, 3.60]	1.49%	4.0×10^{-14}	0.996
FID Inception (train)	34.75	0.433	0.147	2.94 [2.25, 4.15]	1.25%	3.1×10^{-19}	–
MIND DINOv2 (train)	827.1	9.08	2.01	**4.51** [3.45, 6.35]	1.10%	7.8×10^{-41}	0.992
MIND DINOv2 (val)	820.8	9.04	2.00	4.53 [3.46, 6.37]	1.10%	4.8×10^{-41}	0.992
FID DINOv2 (train)	708.8	6.07	0.90	6.75 [5.17, 9.49]	0.86%	4.2×10^{-69}	–

Table 13: **The two lotteries scale differently with the evaluation budget.** Same images as Table 12, at 5k and 50k images. A $1/\sqrt{n}$ estimator shrinks by $\sqrt{10}\approx3.16\times$ for $10\times$ more images.

	σ_w		σ_b		σ_b/σ_w	
Metric (train ref.)	5k $\to$ 50k	shrink	5k $\to$ 50k	shrink	5k	50k
MIND Inception	1.034 $\to$ 0.303	3.41$\times$	0.698 $\to$ 0.770	0.91$\times$	0.68	2.54
FID Inception	0.475 $\to$ 0.147	3.23$\times$	0.405 $\to$ 0.433	0.93$\times$	0.85	2.94
MIND DINOv2	7.22 $\to$ 2.01	3.59$\times$	8.68 $\to$ 9.08	0.96$\times$	1.20	4.51
FID DINOv2	3.96 $\to$ 0.90	4.40$\times$	5.82 $\to$ 6.07	0.96$\times$	1.47	6.75

additional $O(1/n)$ bias of FID's covariance term plausibly explains the factors above 3.16): it is estimator noise. σ_{between} does not shrink at all (0.91–0.96$\times$, well within its confidence interval): it is a property of the trained models. At 5k images the ordering therefore inverts in Inception space: MIND gives $\sigma_b/\sigma_w = 0.68$ (95% CI $[0.48, 1.00]$, $p = 0.98$ for $\sigma_b \le \sigma_w$) and FID 0.85 ($p = 0.80$), so the training lottery is buried in sampling noise at that budget. Our training-dominance claims are thus specific to the standard 50k-image protocol. This strengthens rather than weakens the recommendation of Sec. 4.1: at small budgets the training lottery is not even detectable from one run, and at any budget more images reduce only σ_{within}, never the σ_{between} that limits comparisons.

J Computing guided FID efficiently: golden-section search

This appendix supports Sec. 4.4. It describes how guided FID is computed, i.e. how the per-cell optimal CFG scale $\omega^\star$ is found in a logarithmic number of FID evaluations. The golden-section

procedure is stated formally, its convergence is stated, and FID(CFG) is shown to be unimodal under a Gaussian feature model so that the search returns the global optimum. Notation: let $f : [a, b] \to \mathbb{R}$ be the objective, $\varphi = (1 + \sqrt{5})/2$ the golden ratio, and $\rho = 1/\varphi = \varphi - 1 \approx 0.618$.

J.1 Protocol

Guided FID is the FID at the CFG scale $\omega^\star$ that minimises it, selected separately for every (training, sampling) seed pair. A dense grid over the bracket $[\omega_{\min}, \omega_{\max}] = [0, 2]$ at the resolution we need (0.01) would cost ≈ 200 FID evaluations per cell, i.e. 50 000 on the 25×10 panel. Here ω parameterises the guided prediction as $(1+\omega)\, D_{\text{cond}} - \omega\, D_{\text{uncond}}$, so $\omega = 0$ is unguided and the recovered optima ($\omega^\star \in [0.92, 1.13]$) correspond to a guidance scale of ≈ 2 in the $w\, D_{\text{cond}} - (w-1) D_{\text{uncond}}$ convention. Instead, we run golden-section search (Kiefer, 1953) on the CFG scale over the same bracket at tolerance 0.01: each iteration queries two interior probes and discards the half-bracket with the larger FID, reusing one probe so that every iteration after the first costs a single FID evaluation. The cost is set by the tolerance ε and grows only as $\log(1/\varepsilon)$ (Lemma 2); at our $\varepsilon = 0.01$ it is ≈ 14 evaluations per cell, a $\approx 14\times$ saving over the grid; any other budget can be chosen by setting ε. Pseudocode is given in Algorithm 1 and one step is illustrated in Figure 25. Reporting the search minimum on the samples that selected it is optimistically biased, so $\omega^\star$ should be selected on a proxy budget or a dedicated sampling seed and the reported FID computed on fresh initial noise.

(a) **Algorithm 1:** Golden-section search on FID(ω) (Kiefer, 1953).

Require: Bracket $[a, b]$, tolerance $\varepsilon > 0$

$\rho \leftarrow (\sqrt{5} - 1)/2$
$x_1 \leftarrow b - \rho(b - a)$; $x_2 \leftarrow a + \rho(b - a)$
$f_i \leftarrow \text{FID}(x_i)$ for $i \in \{1, 2\}$
while $b - a > \varepsilon$ **do**
  if $f_1 \le f_2$ **then** ▷ keep $[a, x_2]$
    $(b, x_2, f_2) \leftarrow (x_2, x_1, f_1)$
    $x_1 \leftarrow b - \rho(b - a)$; $f_1 \leftarrow \text{FID}(x_1)$
  else ▷ keep $[x_1, b]$
    $(a, x_1, f_1) \leftarrow (x_1, x_2, f_2)$
    $x_2 \leftarrow a + \rho(b - a)$; $f_2 \leftarrow \text{FID}(x_2)$
  end if
end while
return $(a + b)/2$

(b) One golden-section step on FID(ω)

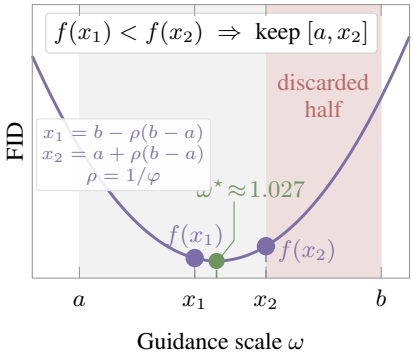

Figure 25: **Golden-section search on FID(ω).** (a) Algorithm 1: pseudocode for the bracket-contraction loop. (b) Two interior probes $x_1 = b - \rho(b - a)$ and $x_2 = a + \rho(b - a)$ with $\rho = 1/\varphi \approx 0.618$ split the bracket $[a, b]$. The side with the larger f-value is discarded.

J.2 Convergence

Figure 25 states Algorithm 1 and illustrates one golden-section step. We restate the standard convergence results (Kiefer, 1953) and instantiate them on our setup; both follow from each iteration shrinking the bracket by the factor ρ.

Lemma 1 (Interval contraction). *Let L_n denote the bracket width after n iterations of Algorithm 1, with $L_0 = b - a$. Then $L_n = \rho^n L_0$.*

Lemma 2 (Evaluations to tolerance). *Reaching tolerance ε from initial bracket width L_0 takes $N = 2 + \lceil \log(L_0/\varepsilon)/\log\varphi \rceil$ FID evaluations: two for the initial probes, one per iteration thereafter.*

For our setup, $L_0 = 2$ (CFG bracket $[0, 2]$) and $\varepsilon = 0.01$ yield $\log(200)/\log\varphi \approx 11.0$, hence $N = 14$, the measured median across the 250 runs of Sec. 4.4.

J.3 Unimodality of FID(CFG) under a Gaussian feature model

Algorithm 1 returns the global minimum only when f is strictly unimodal on $[a, b]$. The next proposition derives that $F(w) := \text{FID}(\text{CFG} = 1 + w)$ satisfies this condition under a Gaussian model of the Inception features, the model that underlies FID itself (Heusel et al., 2017).

Setup. Conditional flow matching learns scores $s_\theta(x,c) \approx \nabla_x \log p(x|c)$ and $s_\theta(x,\emptyset) \approx \nabla_x \log p(x)$. Classifier-free guidance samples from $p_w(x|c) \propto p(x|c)^{1+w}\, p(x)^{-w}$, equivalently using the guided score $\hat{s}(x,c) = (1+w)s_\theta(x,c) - w\, s_\theta(x,\emptyset)$. Let $p_d = \mathcal{N}(\mu_d, \Sigma_d)$ denote the data distribution in Inception feature space.

Assumption (A1). Both $p(x|c) = \mathcal{N}(\mu_c, \Sigma)$ and $p(x) = \mathcal{N}(\mu_0, \Sigma)$ are Gaussian with the same covariance Σ in feature space.

Proposition 1 (Unimodality of FID(CFG))**.** *Under assumption (A1), with $\mu_c \neq \mu_0$, $F(w)$ is a strictly convex quadratic on $\mathbb{R}$, hence has a unique global minimum*

$$w^\star = -\frac{\langle \mu_c - \mu_d,\, \mu_c - \mu_0 \rangle}{\|\mu_c - \mu_0\|^2}, \tag{2}$$

and is strictly unimodal on every interval.

Proof. Under (A1), $p_w(x|c) \propto \exp\!\big(-\tfrac{1}{2}(1+w)(x-\mu_c)^\top \Sigma^{-1}(x-\mu_c) + \tfrac{1}{2} w\,(x-\mu_0)^\top \Sigma^{-1}(x-\mu_0)\big)$. Collecting quadratic and linear terms in x, the precision is $(1+w)\Sigma^{-1} - w\Sigma^{-1} = \Sigma^{-1}$ (independent of w) and the natural mean is $\Sigma^{-1}\big((1+w)\mu_c - w\mu_0\big)$, so

$$p_w(x|c) = \mathcal{N}\big((1+w)\mu_c - w\mu_0,\, \Sigma\big). \tag{3}$$

The Fréchet distance between p_w and p_d in feature space is

$$F(w) = \big\|(1+w)\mu_c - w\mu_0 - \mu_d\big\|^2 + \underbrace{\operatorname{tr}\Big(\Sigma + \Sigma_d - 2\big(\Sigma\Sigma_d\big)^{1/2}\Big)}_{\text{constant in } w}. \tag{4}$$

With $u = \mu_c - \mu_0$ and $b = \mu_c - \mu_d$, the squared norm becomes $\|b + wu\|^2 = \|u\|^2 w^2 + 2\langle b, u\rangle w + \|b\|^2$, a strictly convex quadratic with leading coefficient $\|u\|^2 > 0$ since $\mu_c \neq \mu_0$. Convex quadratics are strictly unimodal, with unique global minimum at $w^\star = -\langle b, u\rangle / \|u\|^2$, which gives (2). Convexity on $\mathbb{R}$ implies unimodality on every subinterval. □

Sanity check. If the conditional matches the data ($\mu_c = \mu_d$), then $b = 0$ and $w^\star = 0$: guidance is unnecessary. The empirical mean $w^\star \approx 0.027$ from Sec. 4.4 implies that the conditional model places its mass within $\approx 3\%$ of the guidance direction $\mu_c - \mu_0$ of the data mean.

Beyond (A1) and evaluation noise. The proposition assumes equal covariances and noise-free FID evaluations. Empirically, all 250 searches converge to a single tight cluster ($\sigma_{w^\star} = 0.045$), consistent with unimodality beyond (A1) and with a search that is robust to the sampling noise of each evaluation.

K μP Transfers a 1.7× LR Window, Not a Point

With CFG, several nearby LRs all give similar FID, a 1.7×-wide window. Without CFG, the best LR moves to the edge of training stability.

μP transfers, but to a window, not a point. The sweep covers 10 LRs coordinated by μP (Yang et al., 2022), a parameterisation that makes the optimal LR width-invariant, log-spaced in $[5{\times}10^{-5}, 5{\times}10^{-4}]$ across SiT-S/B/L/XL with $N = 10$ seeds per cell, 100k steps, and both unguided FID and guided FID ($\approx$400 networks, Appendix C). The guided-FID panel of Figure 26 shows flat-bottomed valleys near an optimal LR of 2–3$\times 10^{-4}$ at every size, with highlighted minima within one log-step. The two LRs flanking each minimum sit inside its seed envelope, so three adjacent LRs share the per-size best FID, a 1.7× LR window per size. μP does transfer the optimum across widths, but the object it transfers is a window, not a single number. The bootstrap credible set of the optimal LR (strips in Figure 26) puts 83–97% of its mass on one LR for SiT-S/B/L, but spreads over three adjacent LRs on SiT-XL (51%, 28% and 21% on $\{3, 3.9, 5\}\times 10^{-4}$), one of them at the edge of stability. A recipe-comparison study sweeping a few LRs and reporting the per-recipe best is out-resolved by the seed lottery, exactly like the single-LR practitioner of Sec. 4.5.

The seed CoV does not dip at the optimal LR. A flatter loss-landscape region should also be lower-variance, but Figure 26c refutes this: every size's argmin CoV lies 1–3 brackets *below* its

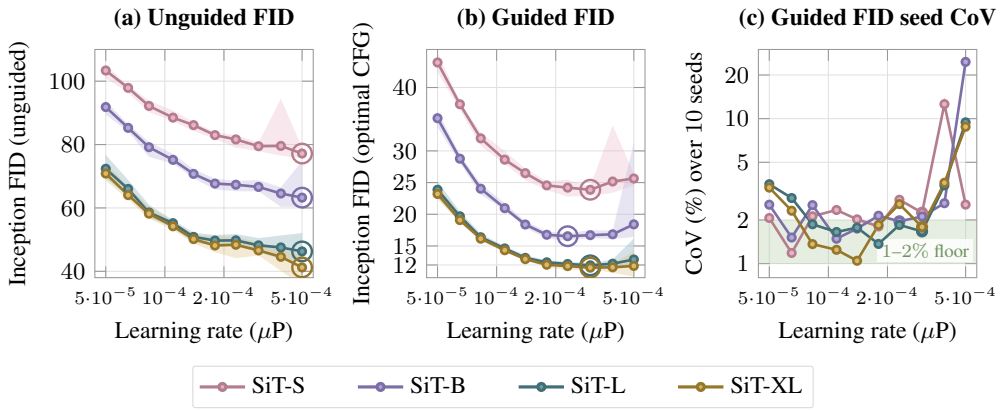

Figure 26: ***μP*-coordinated LR sweep at 100k for SiT-S/B/L/XL.** Solid lines are the per-LR mean Inception FID across 10 training seeds, the shaded envelope is the per-LR seed min–max, and the open ring on each curve circles each size's best-mean-FID dot. (a) Unguided FID is near-monotone in LR for every size, so the highlighted dot sits at 5×10^{-4} (the edge of training stability) for all four. (b) Guided FID has flat-bottomed valleys around an optimal LR of 2–3×10^{-4}. Within seed noise, the LR on either side is indistinguishable from the highlighted dot, giving a $1.7\times$ LR window per size. (c) Between-seed $\sigma_{\text{between}}/\mu$ on log scale. The green band is the 1–2% floor of Sec. 4.5. Open rings mark each size's optimum LR from (b): the CoV does *not* dip there.

argmin mean, in the under-trained regime where seeds have not moved far from initialisation. At the actual optimum LR (open rings, panel c) the CoV is 1.7–2.3%, inside the 1–2% floor of Sec. 4.5. The floor is therefore the variance at the LR a practitioner would actually pick: under-training squashes seeds toward initialisation, but the practitioner gives that regime up when they tune for FID. Across the central seven well-conditioned LRs the between-seed CoV stays in a 1–3% band on every size under both protocols (Table 14), so μP transfers a stable noise floor across widths, not just a stable mean.

Table 14: ***μP* sweep at 100k: per-size noise floor.** Companion to Figure 26. Range of between-seed coefficient of variation across the central seven well-conditioned LRs $[8\times10^{-5}, 4\times10^{-4}]$, broken out by protocol (guided versus unguided FID). The argmin LR per size is the modal bootstrap optimum. CFG-tuning halves the absolute FID but the relative spread sits in the same 1–3% band as the long-run scaling sweep (Table 2).

Size	argmin LR (guided)	Guided CoV range (%)	Unguided CoV range (%)	Guided % at argmin
SiT-S/2	3.0×10^{-4}	1.18–2.76	0.82–1.49	1.7
SiT-B/2	2.3×10^{-4}	1.48–2.61	1.06–2.05	1.9
SiT-L/2	3.0×10^{-4}	1.37–3.53	1.58–3.46	2.0
SiT-XL/2	3.0×10^{-4}	1.05–3.34	1.55–3.84	2.3

Without CFG, the sweep points to the unstable edge. The unguided panel is monotone in LR at 100k, so the lowest mean sits at the right edge (5×10^{-4}), the same LR where $3/10$ SiT-S and $1/10$ SiT-XL seeds diverge. Single-seed unguided LR selection therefore lands on an unstable point with a confident-looking number, while guided FID returns the $1.7\times$ window around 3×10^{-4}. At 5×10^{-4} the guided-FID CoV jumps an order of magnitude on SiT-B (one seed at FID 31.3 against a cluster around 17) and 3–4× on SiT-L/XL, and SiT-S diverges on three of ten seeds (FID >300, excluded from the curve); unguided FID hides this under larger absolute values (CoV below 7%), so a single point at this edge can mix a successful and a divergent run. The two protocols also disagree per seed: the per-seed argmin differs on $8/10$, $6/8$ and $5/10$ seeds of SiT-S/B/L, and on SiT-B/L the unguided argmin is 7.7–11.5% worse in guided FID than the guided optimum, with a 5–12× wider between-seed envelope.

L WHAT DOES FID LOOK LIKE?

The figures in this appendix complement the quantitative results of Sec. 4 with a purely visual diagnostic: how does the appearance of generated samples change as the FID of the underlying checkpoint moves across its range, and does that change look the same with and without CFG? Figure 27 condenses three of the galleries into one row each across the full FID range.

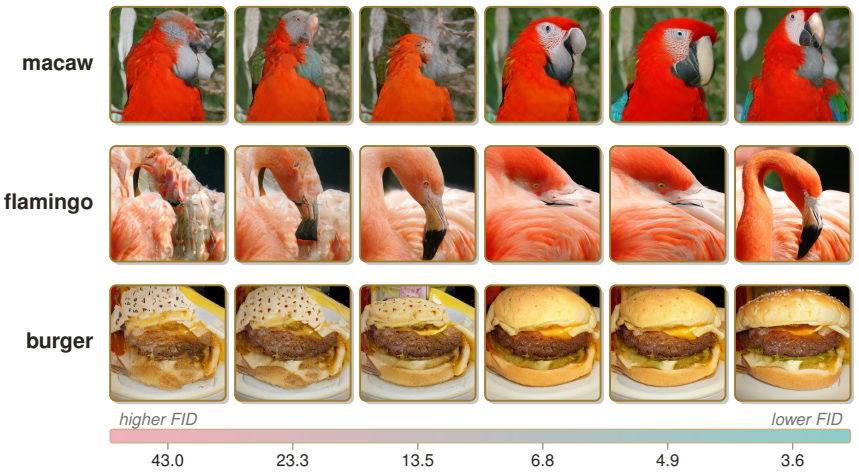

Figure 27: **What does the FID spectrum look like?** Each row is one scene rendered by SiT checkpoints whose FID falls from **43** (left, unguided) to **3.6** (right, guided), a **12**× range. FID is a Fréchet distance between Gaussians fit to the Inception features of ImageNet and of 50,000 generated images: it is a property of the whole *set*, never of a single image (per-class galleries in Figure 28–47).

Setup. We pick 10 ImageNet classes (golden retriever, tabby cat, macaw, flamingo, cheeseburger, ice cream, volcano, alp, geyser, daisy) and 10 fixed initial-noise tensors. The figures show the first 6 of those noise seeds to keep each panel within page bounds. Every (class, noise) pair is decoded with a 50-step Euler flow sampler at 16 FID levels. For the unguided panel, the levels are sampled along the full SiT training trajectory and span FID $\in [13.5, 87.4]$ (SiT-XL fully trained $\rightarrow$ SiT-S undertrained). We uniformly subsample these 16 from the 32 available checkpoints. For the guided panel, the levels are SiT-XL checkpoints evaluated at the golden-section-selected best CFG, spanning FID $\in [3.55, 11.3]$. The 16 FID-ordered checkpoints are split into two stacked half-panels of 8 columns each, so each figure is 8 columns wide and 12 rows tall: the top half holds the lower-FID octave and the bottom half the higher-FID octave. Each half-panel carries its own pastel colorbar, but both share a single global FID scale, so column colours are comparable across the two halves.

What the galleries show. Holding initial noise fixed across columns isolates the effect of training-time / guidance changes on the rendered image. The unguided panels show the visual quality degrading monotonically along the FID axis – a smooth, well-formed object near the leftmost columns collapses into class-correct but textureless blobs by the rightmost columns. The guided panels show that even the highest-FID (worst) checkpoint produces visually clean, recognisable samples once an appropriate CFG is applied – consistent with the Sec. 4.4 observation that classifier-free guidance compresses the entire FID range it is computed over.

The two condition-specific axes are shown back-to-back per class: Figure 28–37 for the guided panels and Figure 38–47 for the unguided panels.

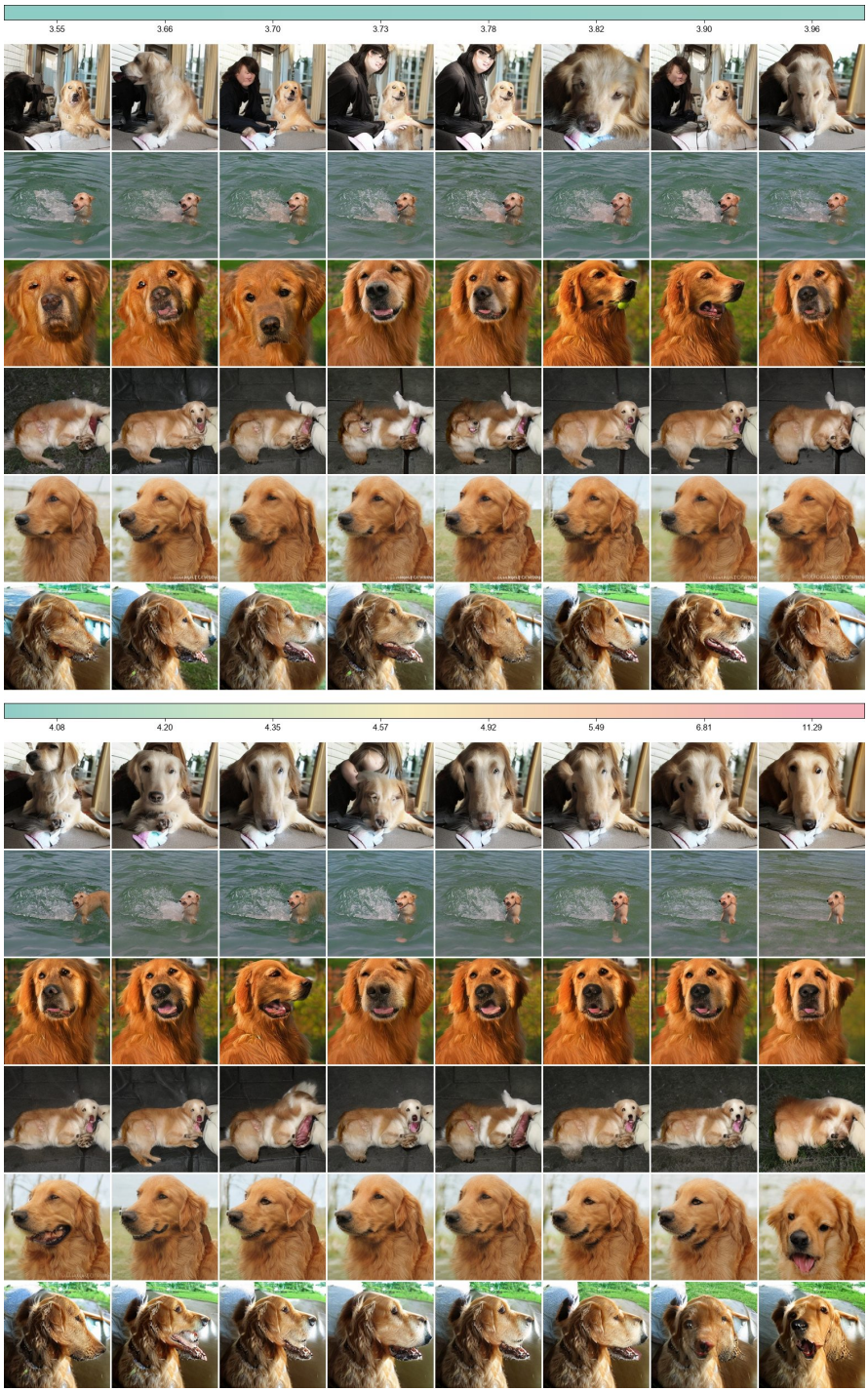

Figure 28: **Guided FID gallery – golden retriever.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

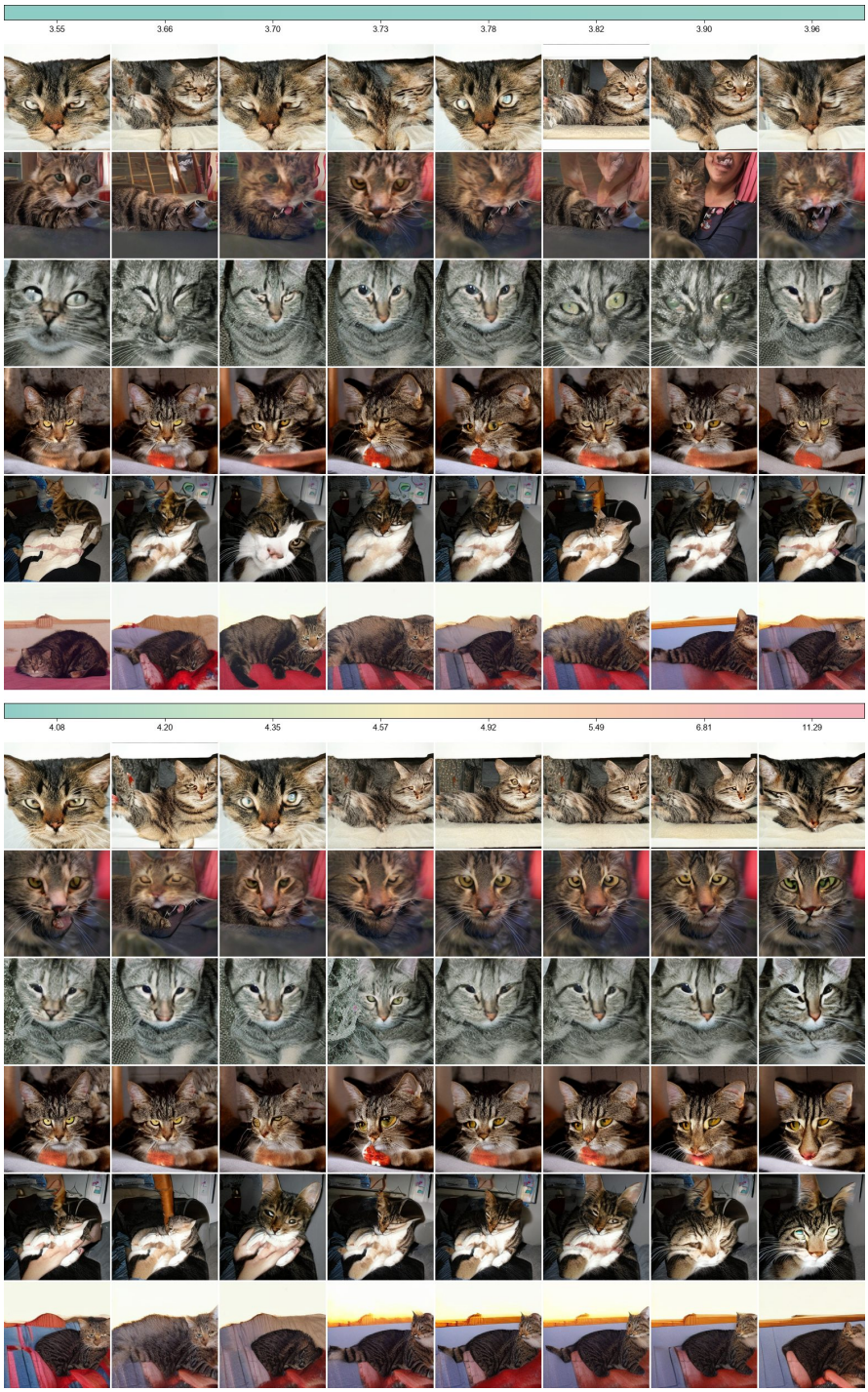

Figure 29: **Guided FID gallery – tabby cat.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

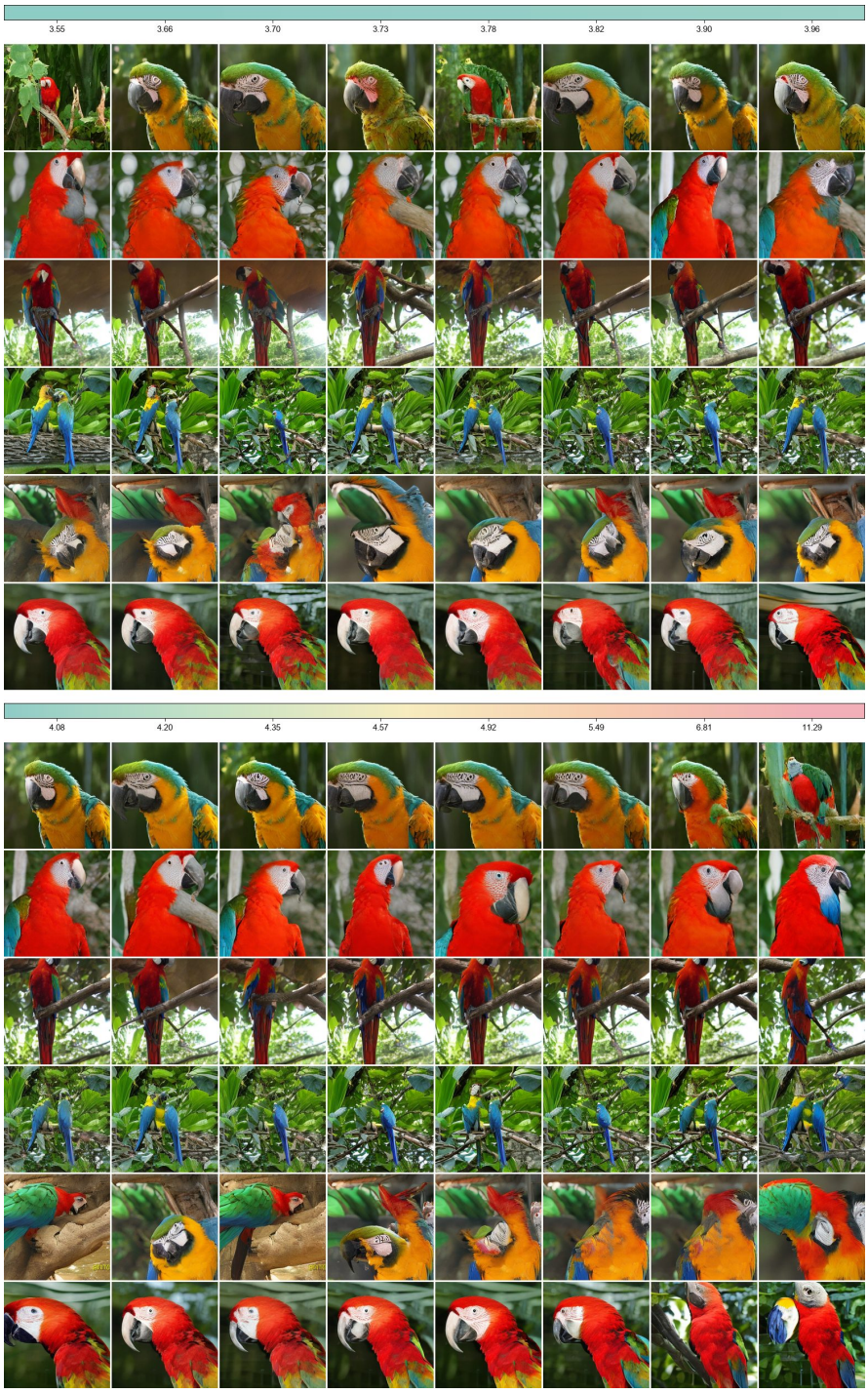

Figure 30: **Guided FID gallery – macaw.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

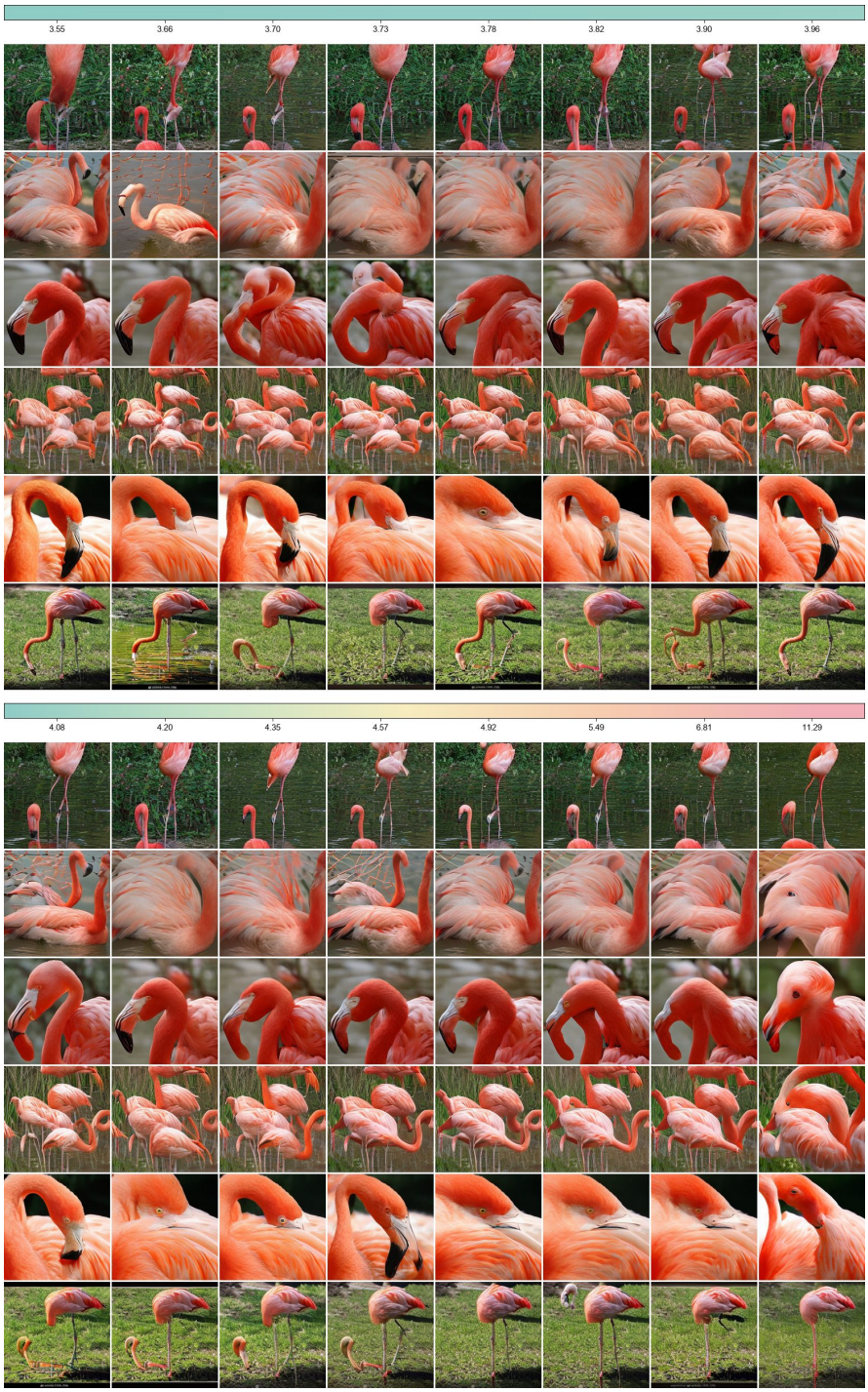

Figure 31: **Guided FID gallery – flamingo.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

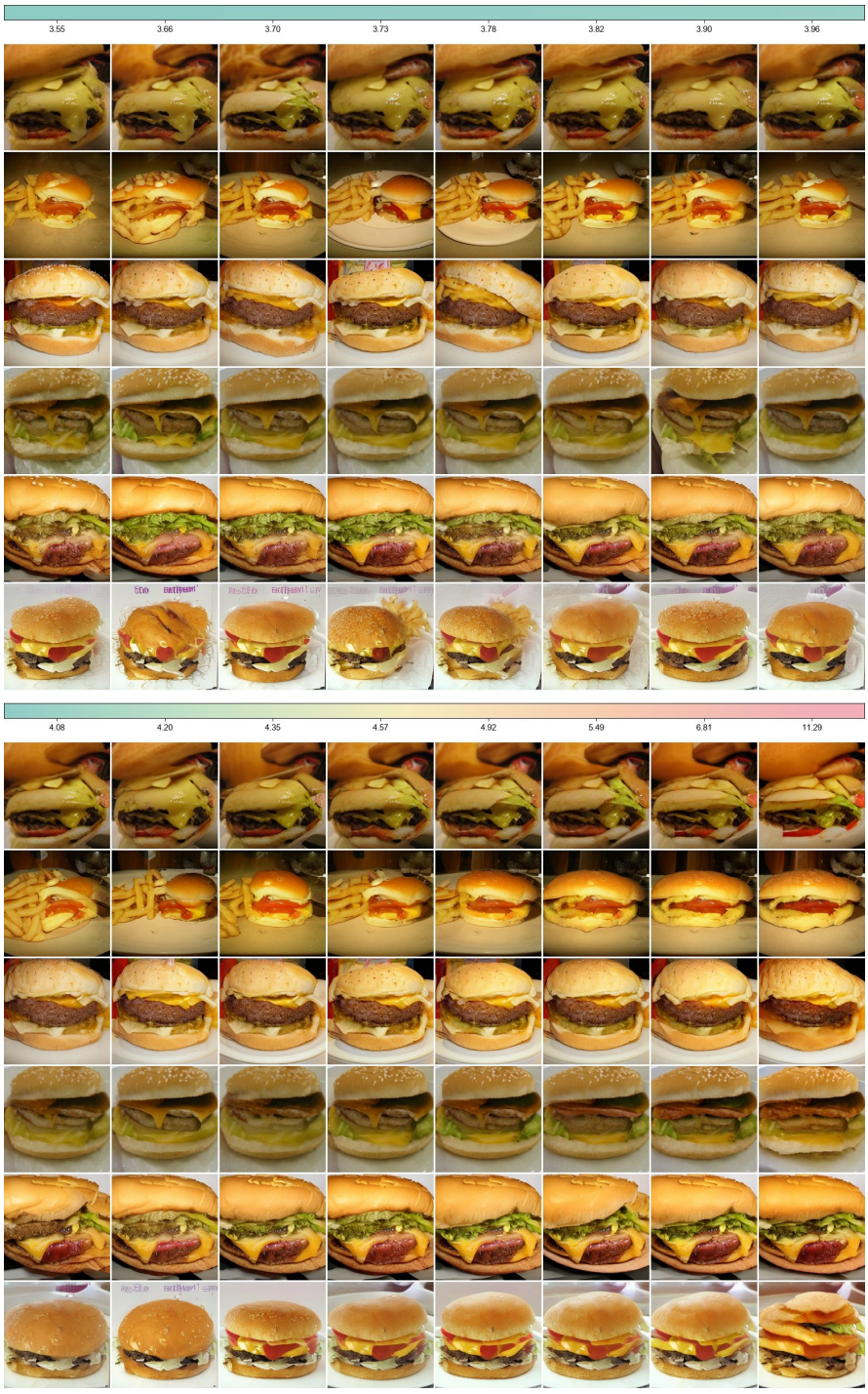

Figure 32: **Guided FID gallery – cheeseburger.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

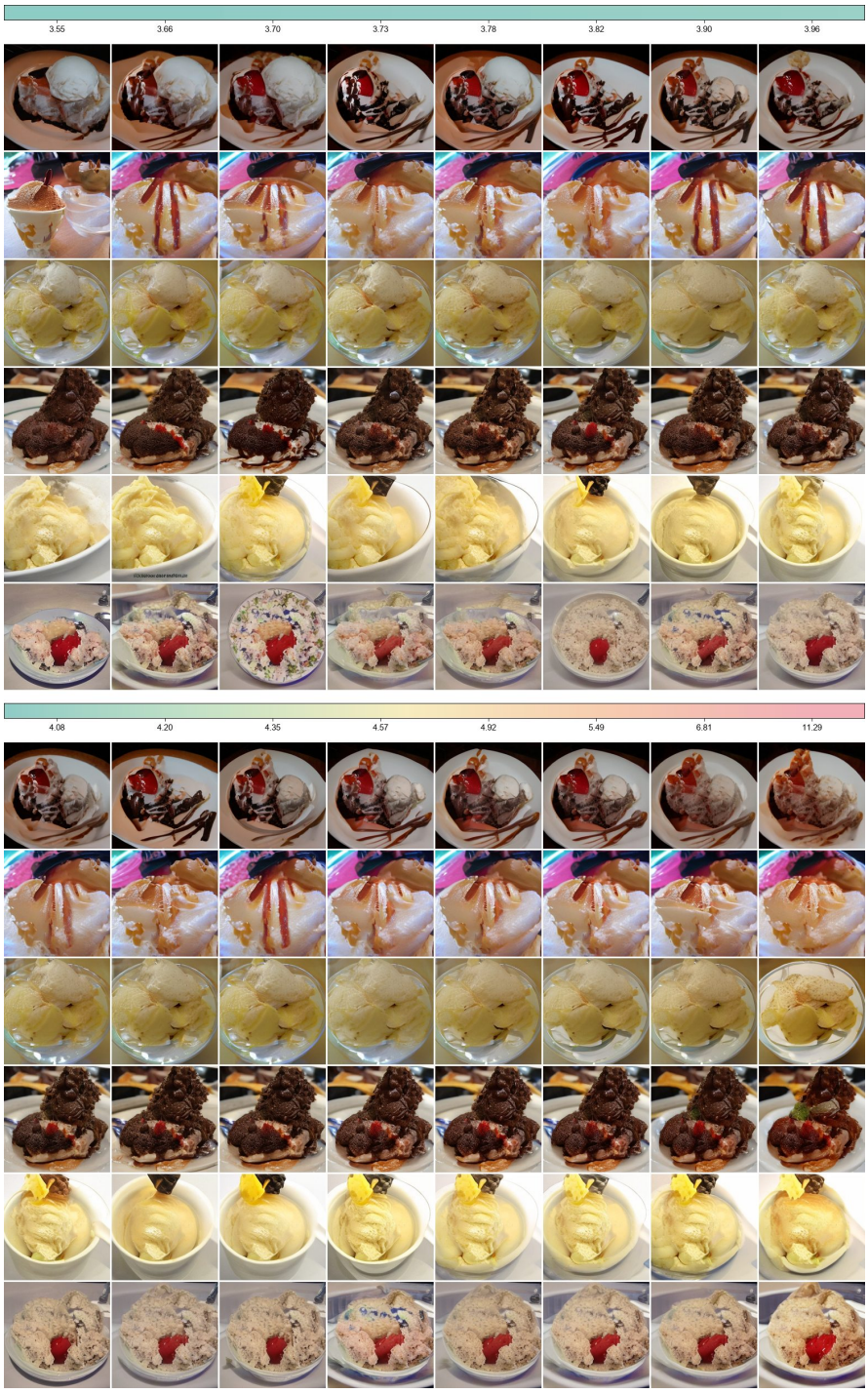

Figure 33: **Guided FID gallery – ice cream.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

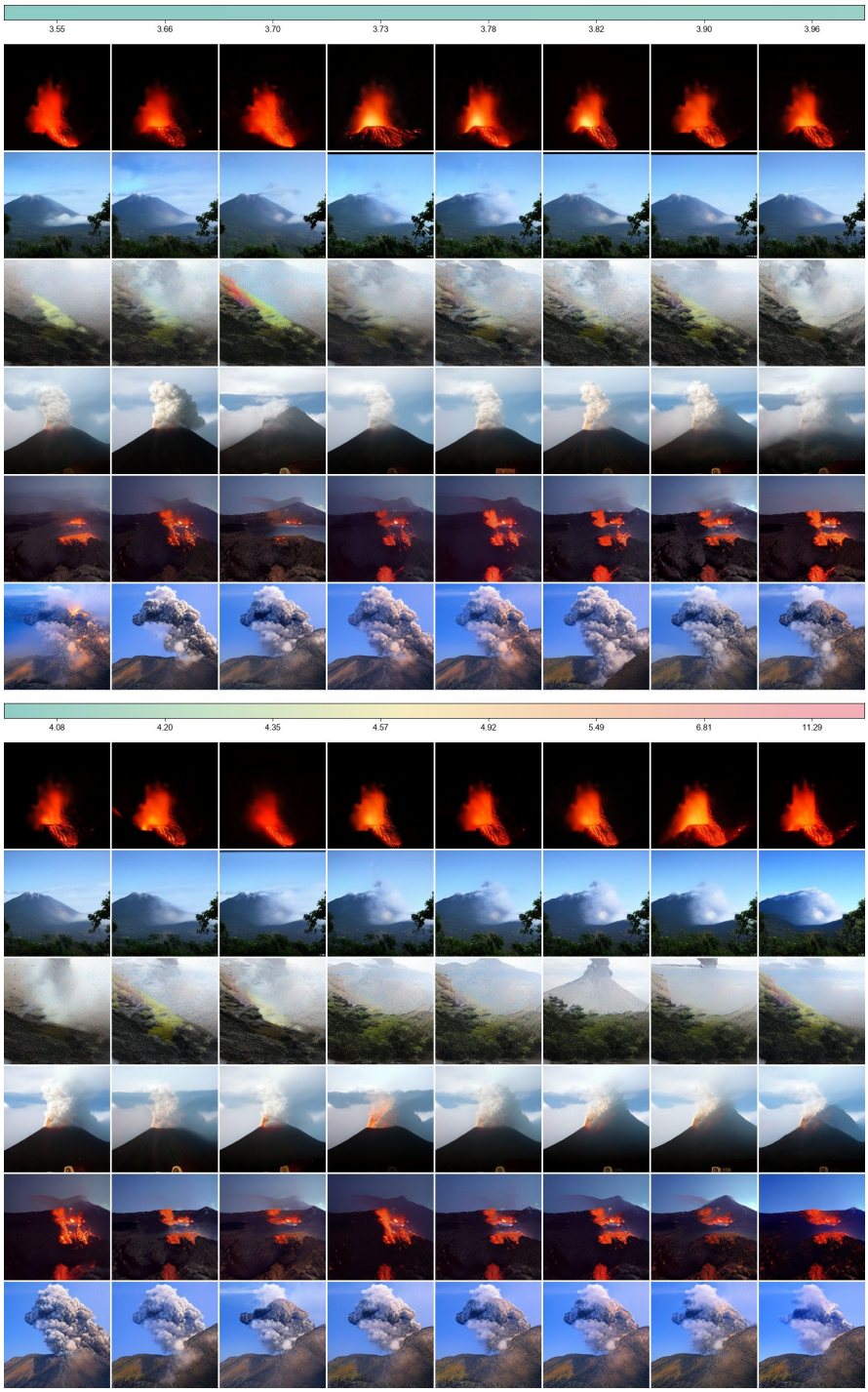

Figure 34: **Guided FID gallery – volcano.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

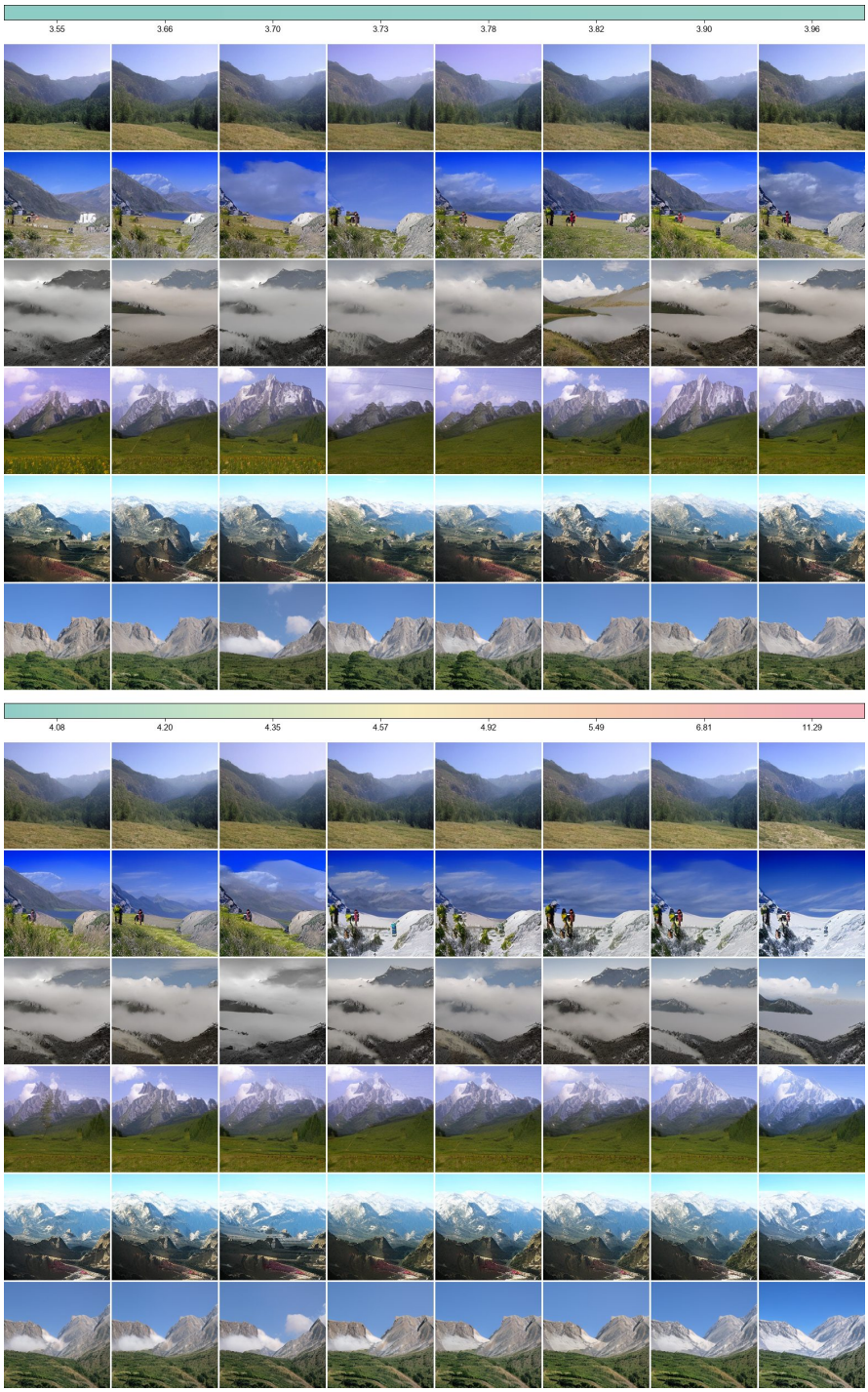

Figure 35: **Guided FID gallery – alp.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

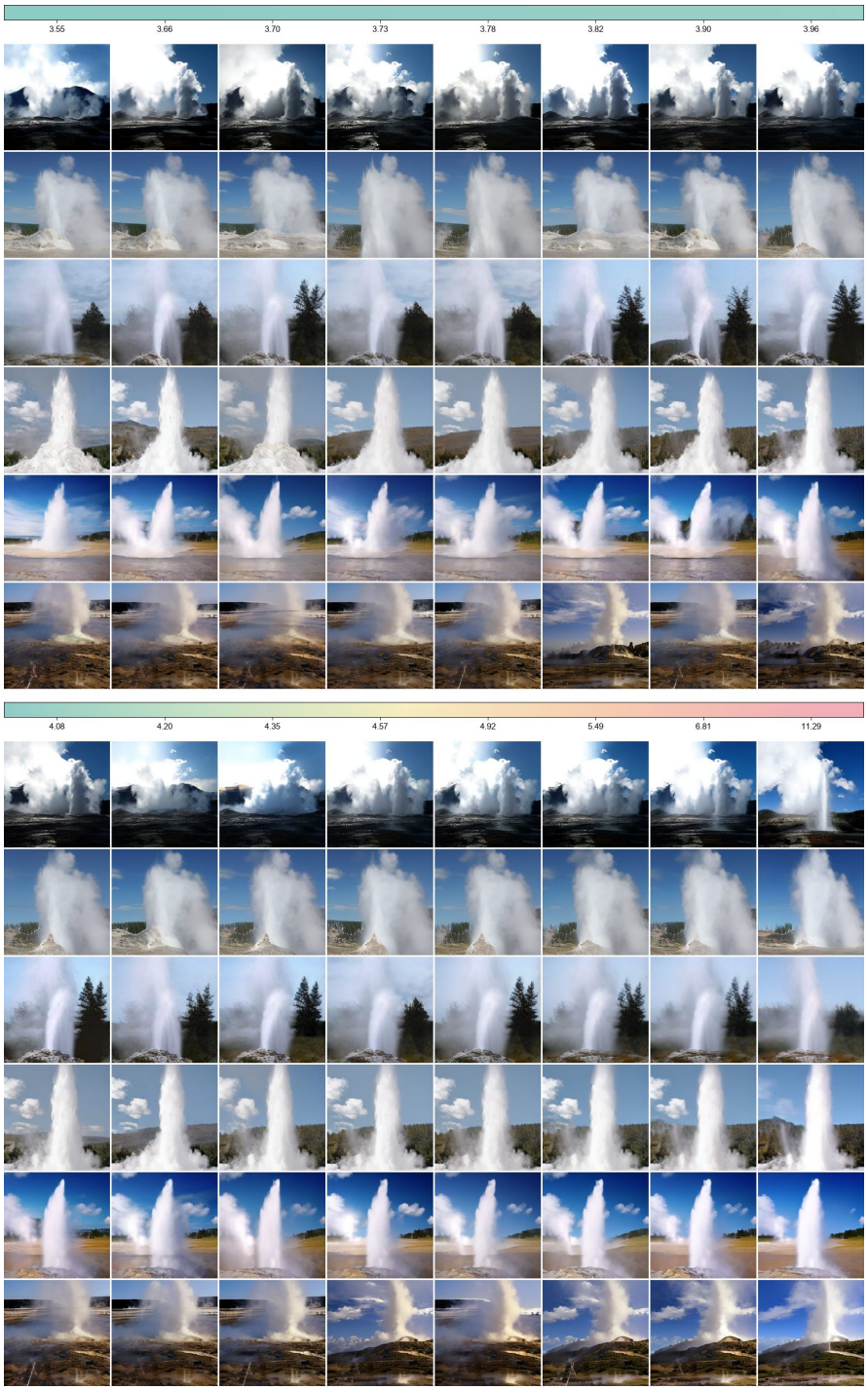

Figure 36: **Guided FID gallery – geyser.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.



Figure 37: **Guided FID gallery – daisy.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

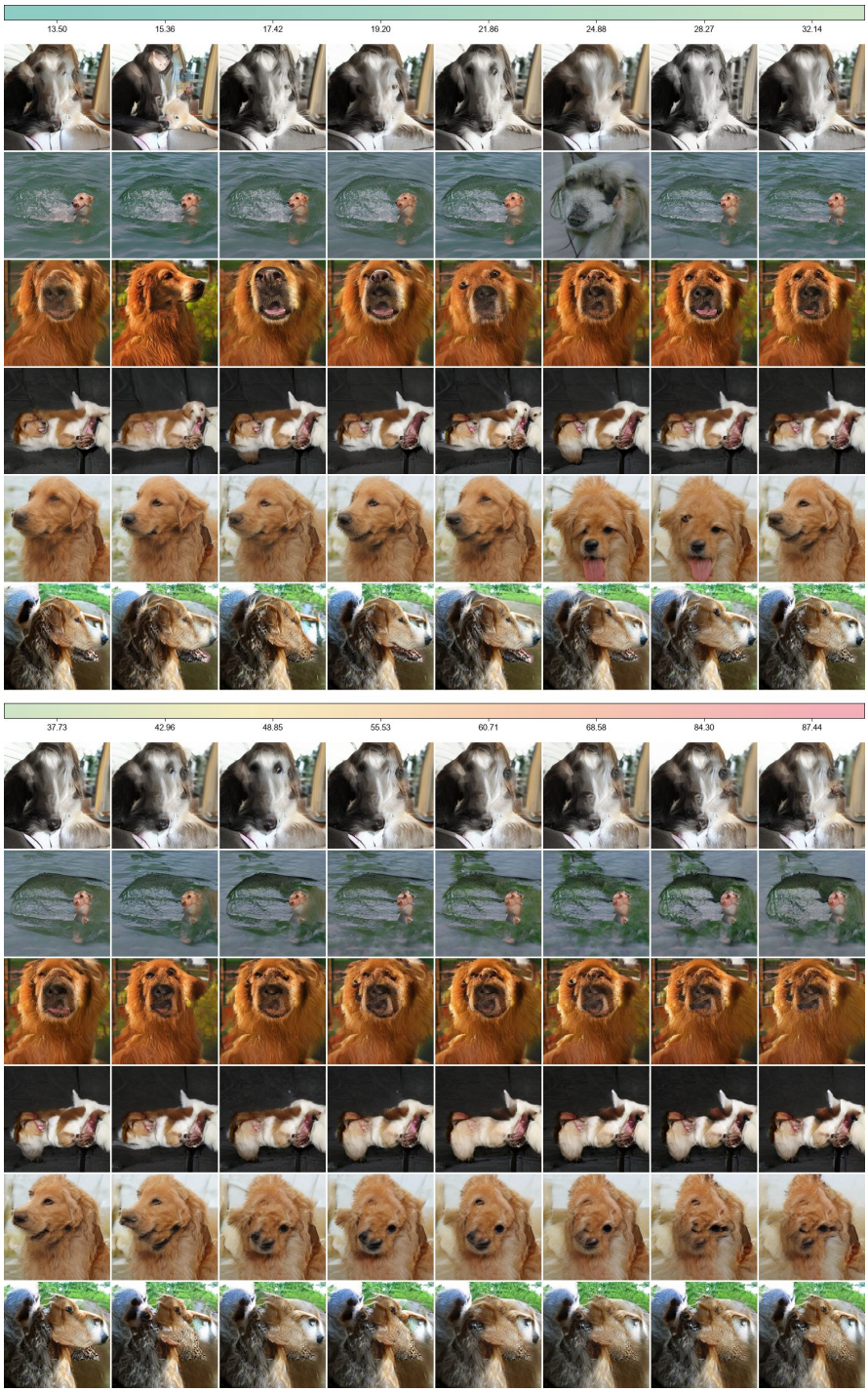

Figure 38: **Unguided FID gallery – golden retriever.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

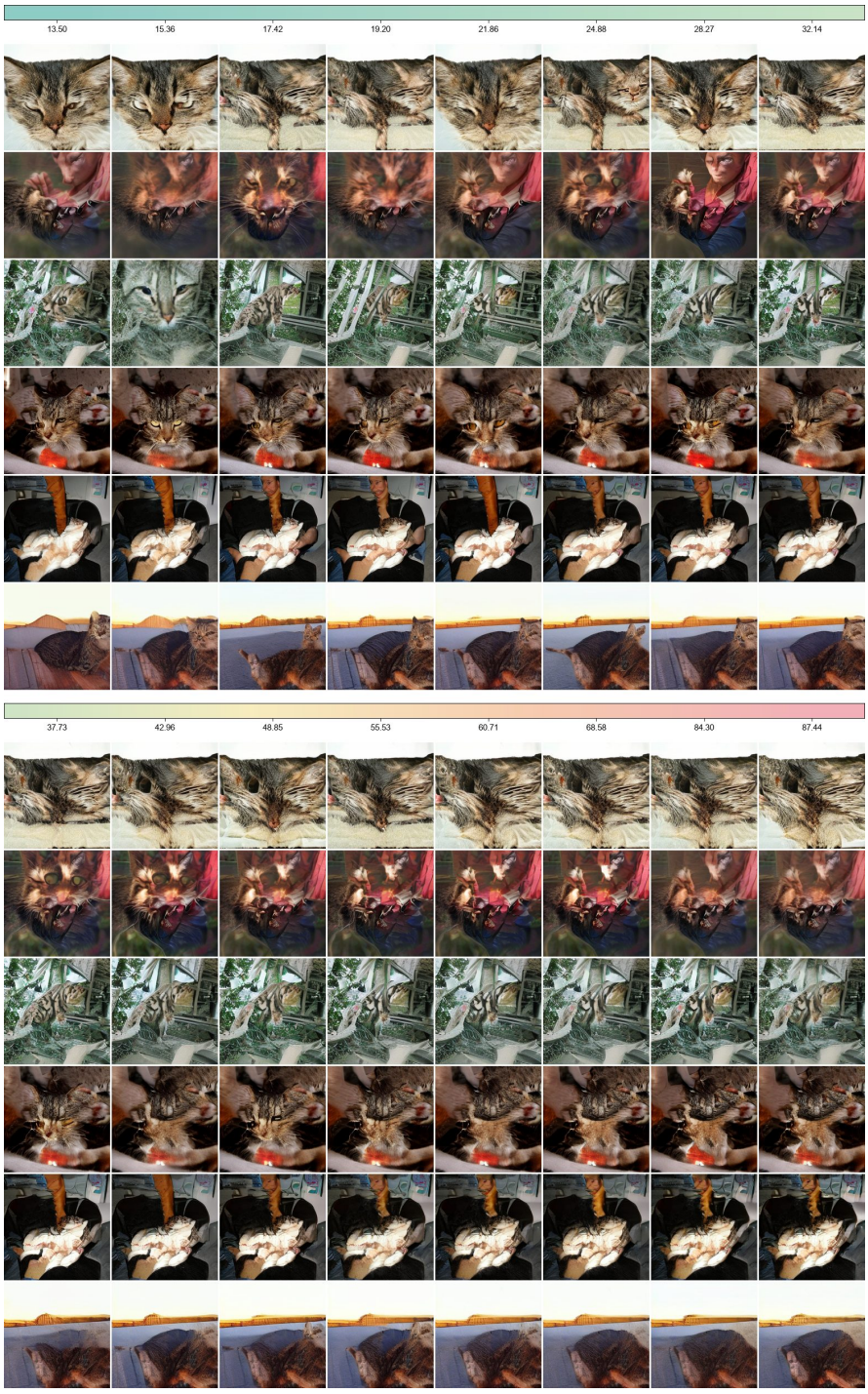

Figure 39: **Unguided FID gallery – tabby cat.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

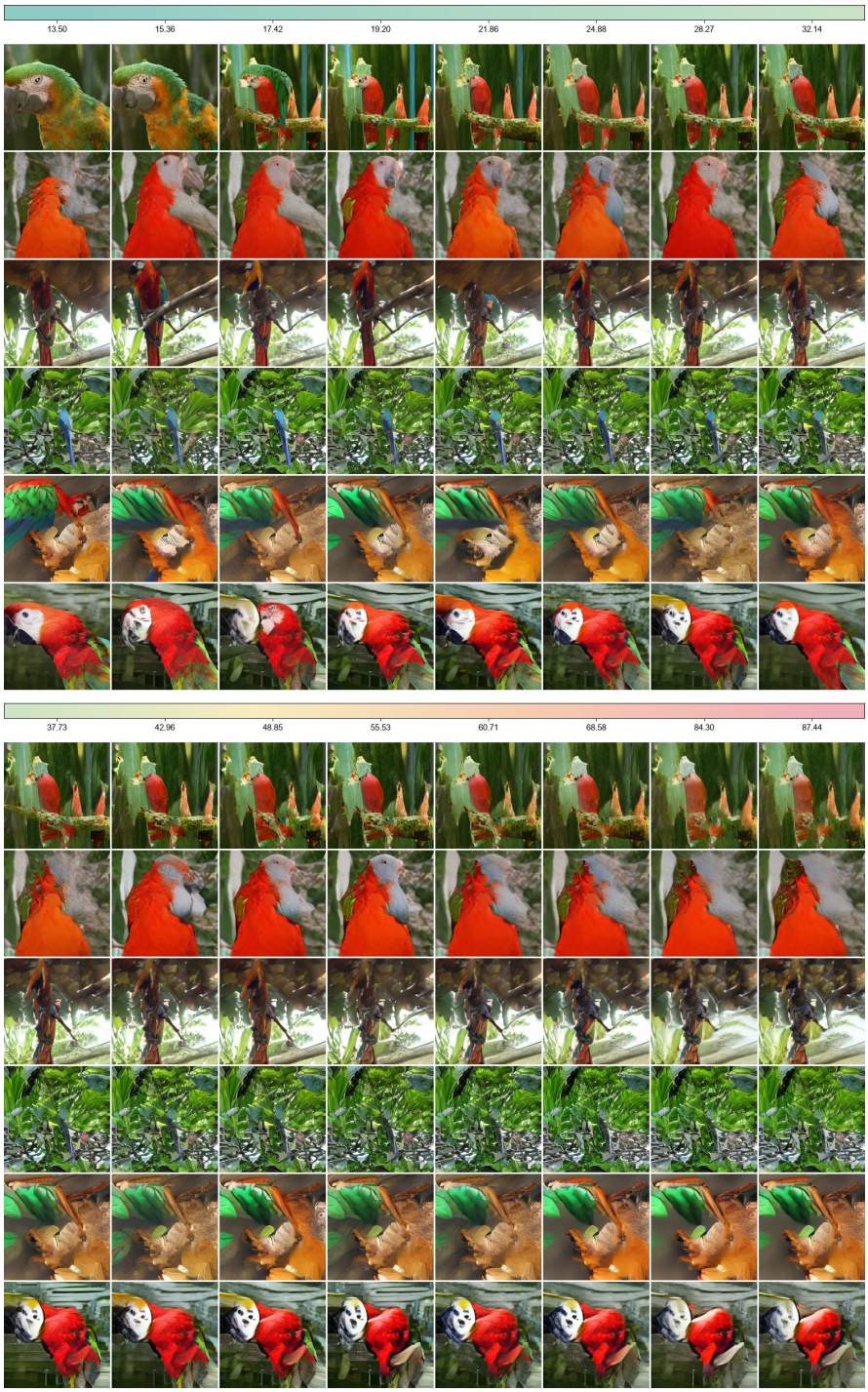

Figure 40: **Unguided FID gallery – macaw.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

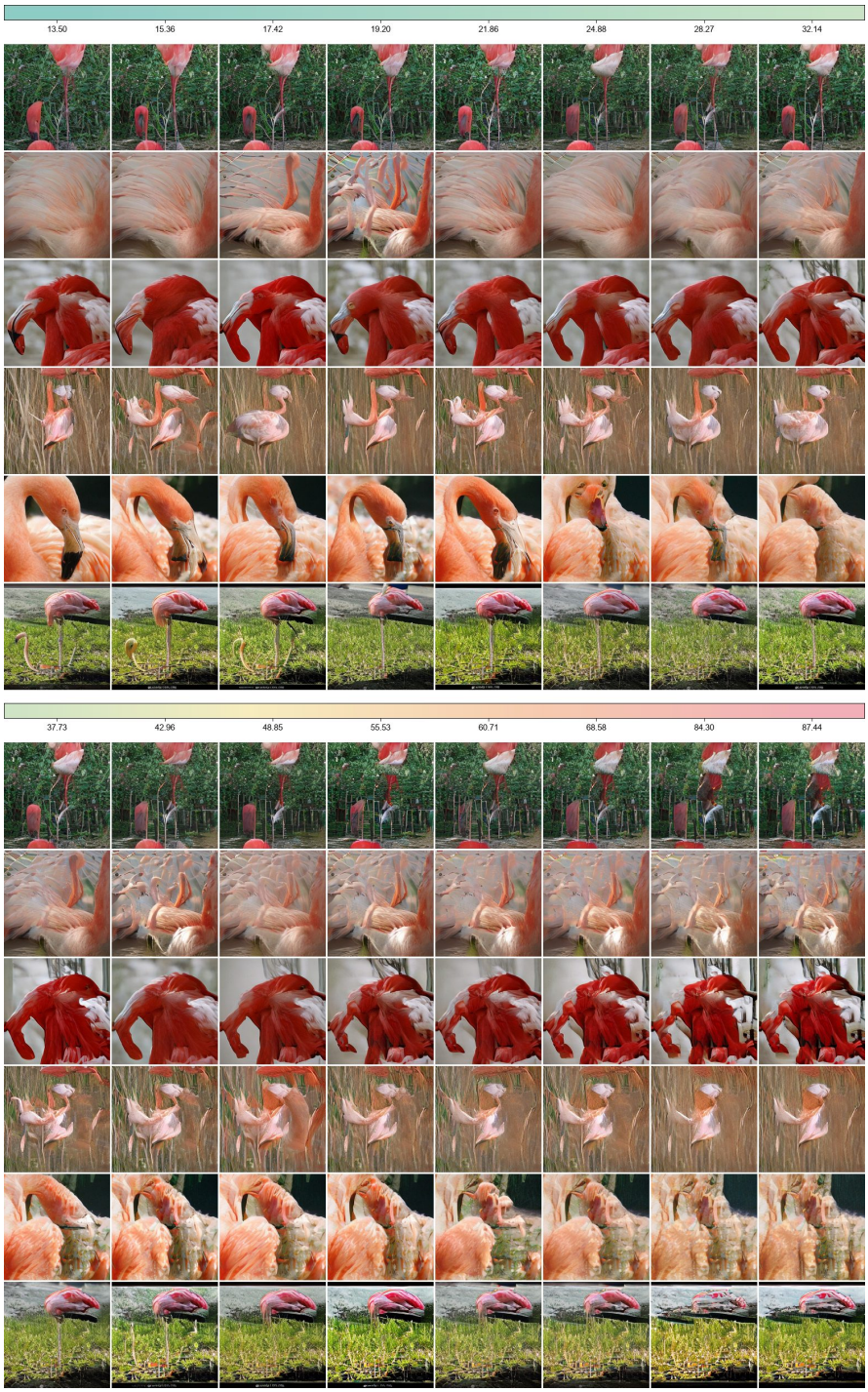

Figure 41: **Unguided FID gallery – flamingo.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

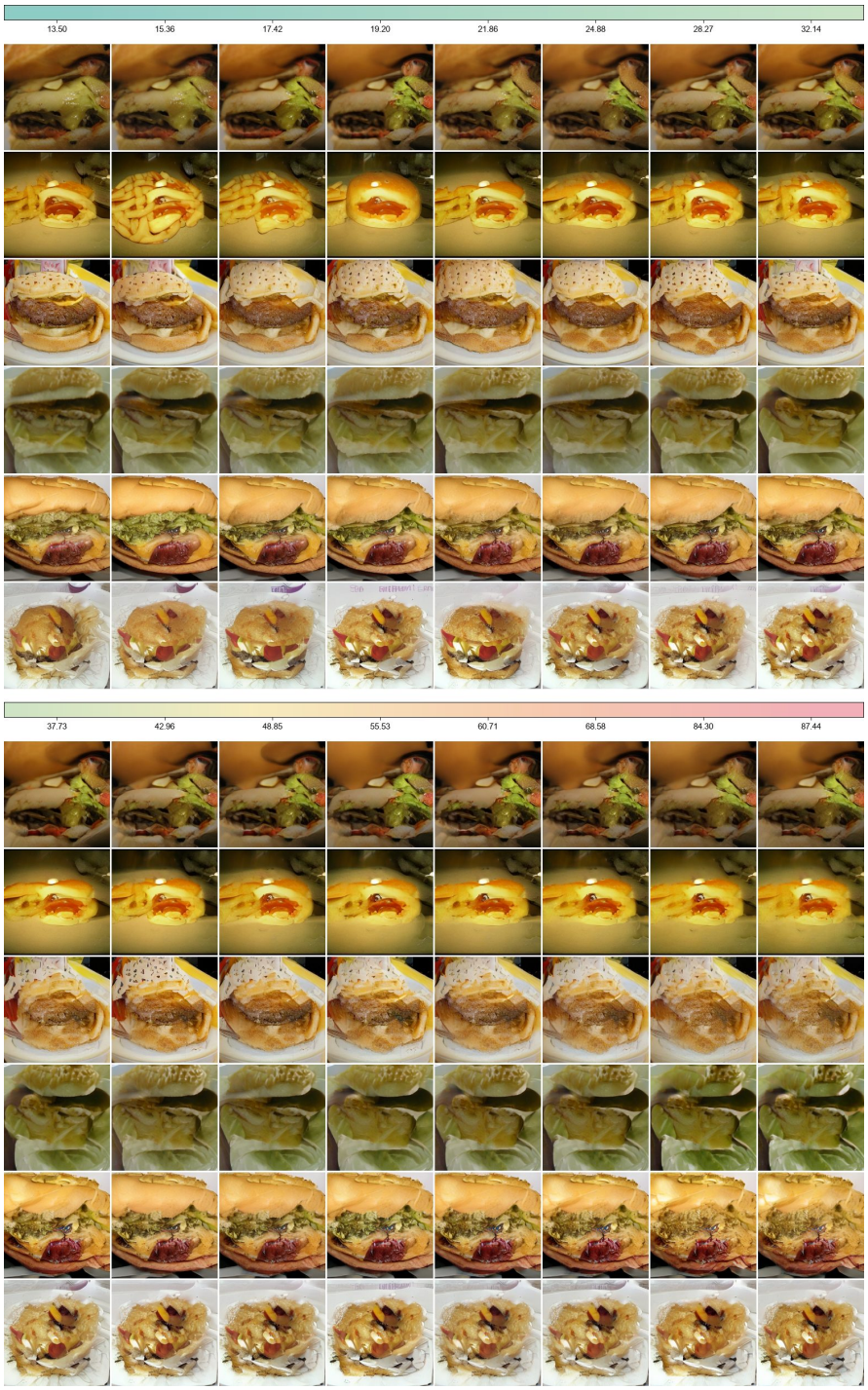

Figure 42: **Unguided FID gallery – cheeseburger.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

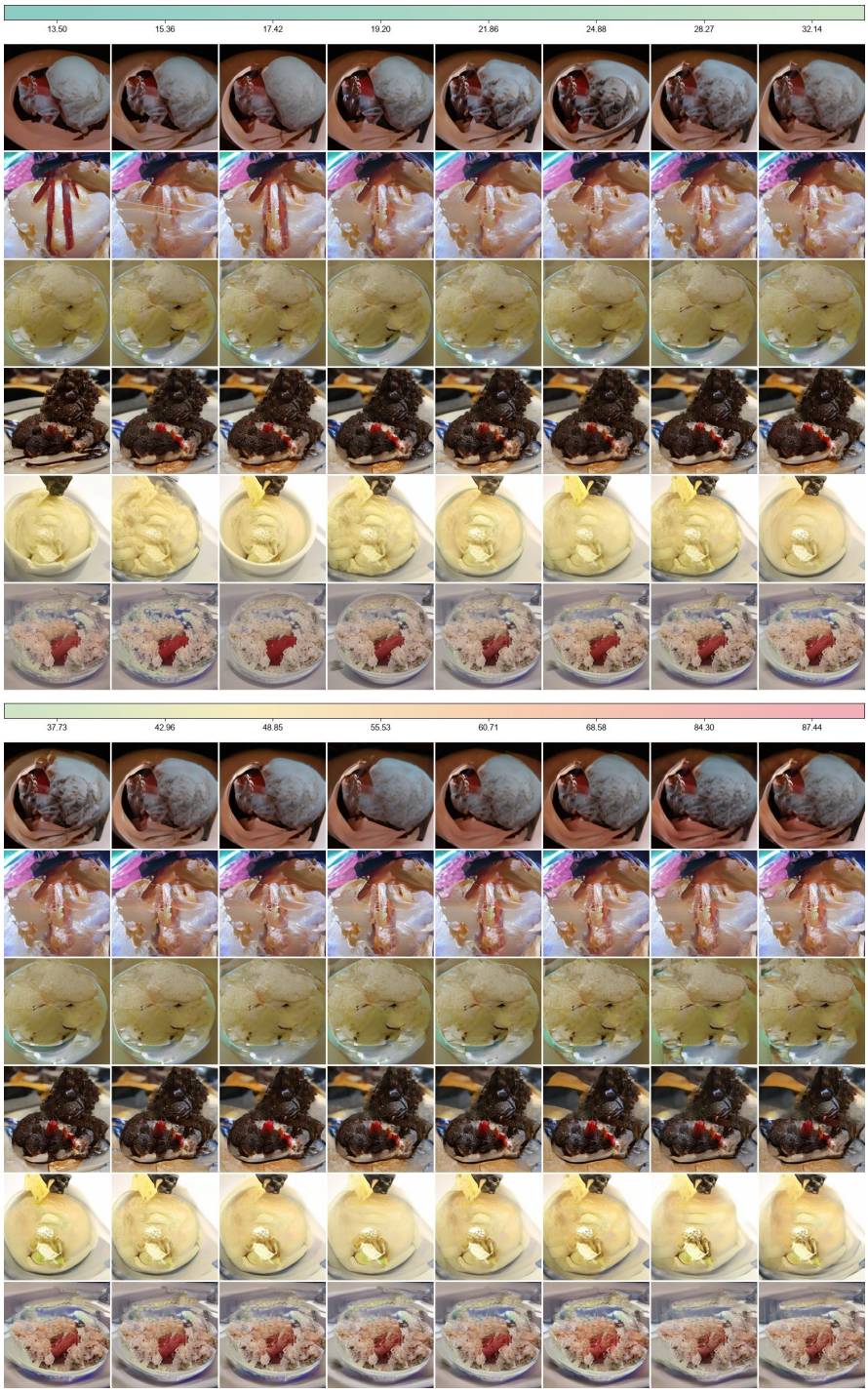

Figure 43: **Unguided FID gallery – ice cream.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

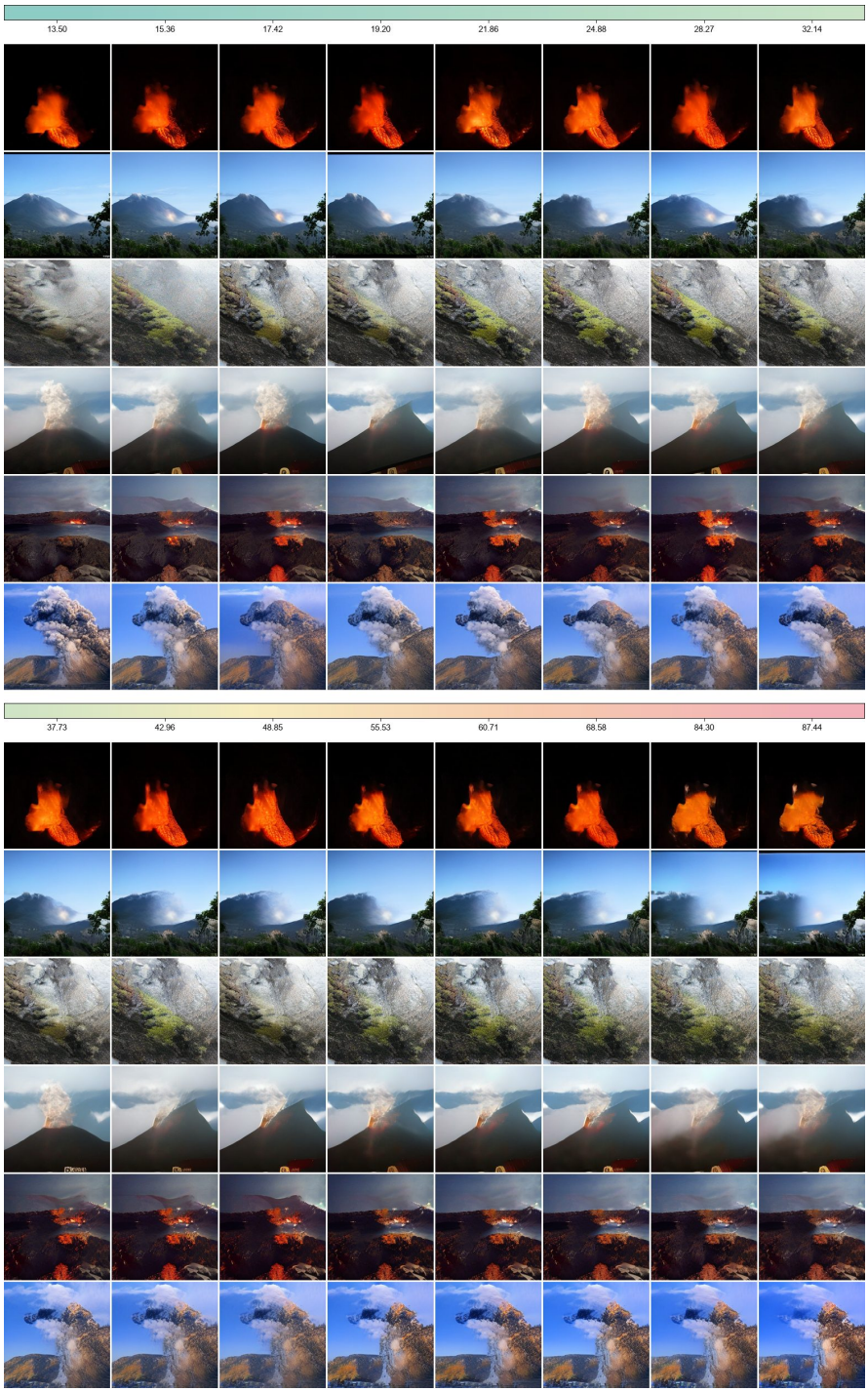

Figure 44: **Unguided FID gallery – volcano.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

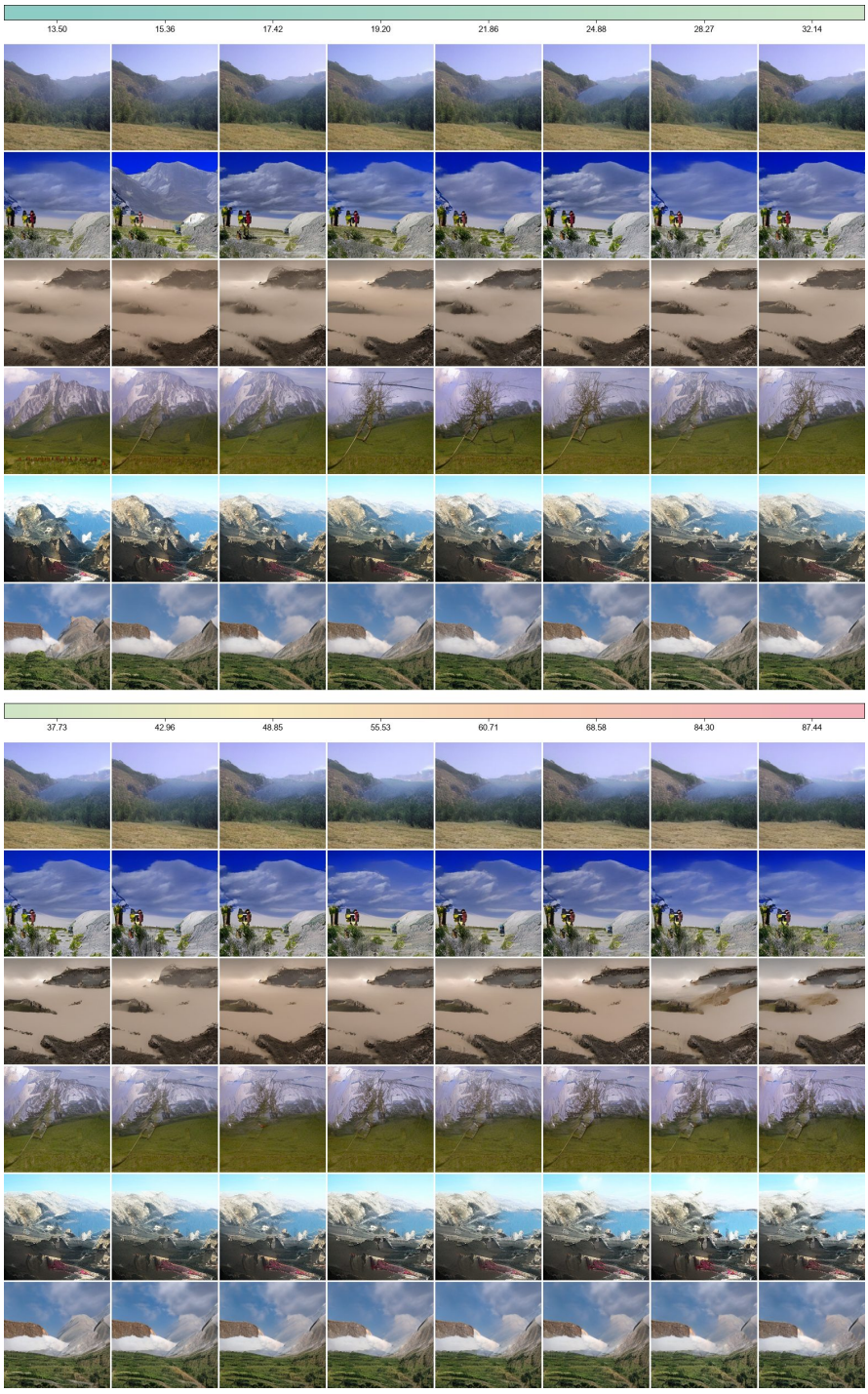

Figure 45: **Unguided FID gallery – alp.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.

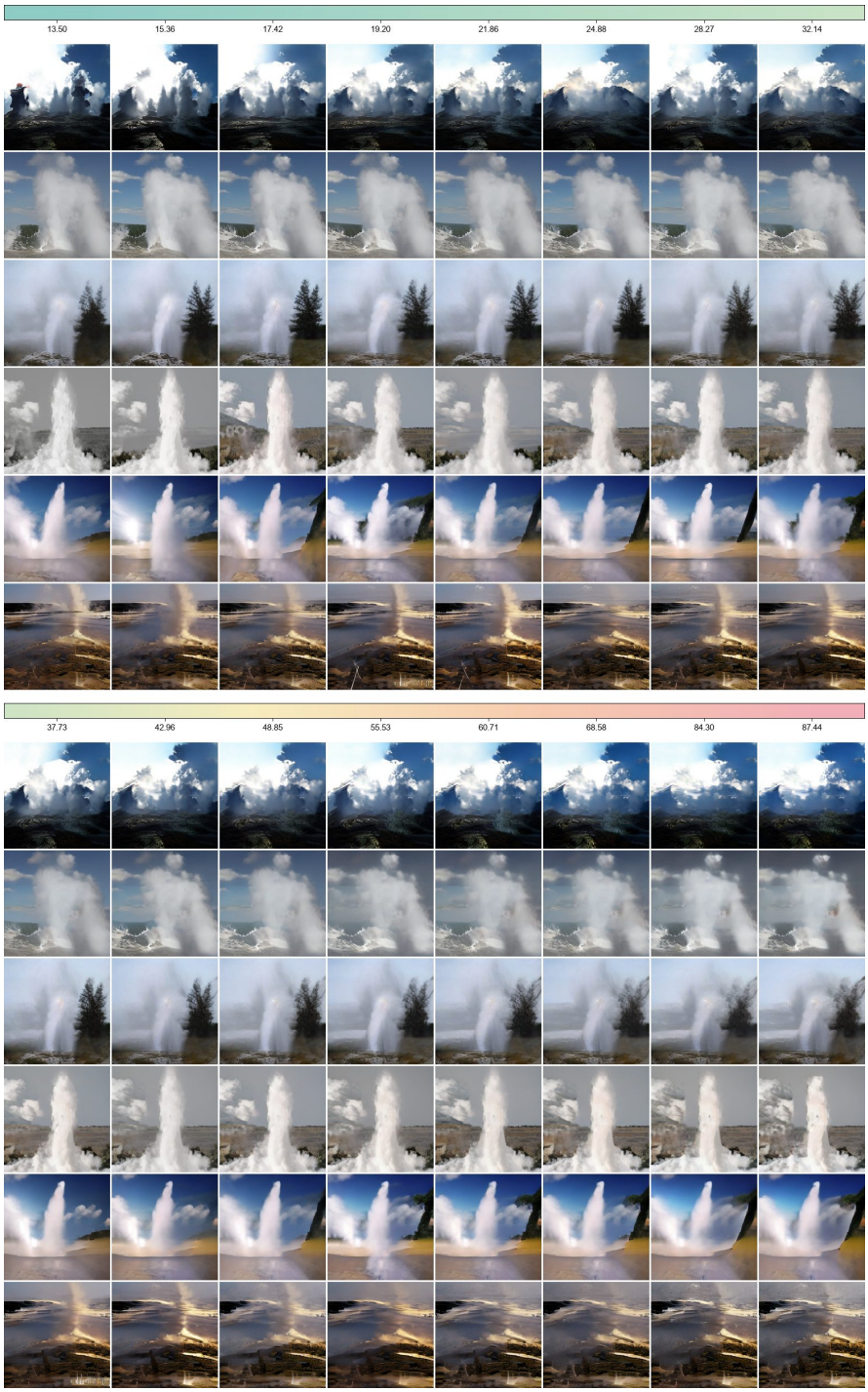

Figure 46: **Unguided FID gallery – geyser.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.



Figure 47: **Unguided FID gallery – daisy.** Rows are 6 fixed initial-noise seeds, and columns step through 8 FID-ordered checkpoints. The two halves continue along the same FID axis. Pastel colorbars report per-column FID. See Appendix L for the full setup.