# How Many Images Are Enough to Compare Diversity? Adaptive Sample Sizes for Text-to-Image Evaluation

**Anonymous authors**
Paper under double-blind review

## Abstract

Diversity comparisons between text-to-image generators are often computed from a fixed number $k$ of images per prompt, yet there is no principled rule for choosing $k$. We study this question for the Vendi score, whose value keeps changing as more images are generated per prompt. However, a comparison between two generators need not be equally unstable: because the two scores tend to change by similar factors, their ratio can stabilize while each score is still changing. The number of images required for this stabilization differs widely across guidance settings, diversity interventions and model architectures. To address these issues, we introduce DivShift, a prospective procedure that determines the required number of images separately for each Vendi-based comparison. Given a user-specified reference size $K$ and a tolerance $\rho$, DivShift tracks the observed Vendi log-ratio, predicts how much the ratio may still change before $K$, accounts for variation across prompts, and either stops or requests additional images. At $\rho = 10\%$ and $K = 512$, the five-panel macro-average shows that DivShift saves $79\%$ of generations over evaluation at $K$, while only $0.10\%$ of stopping decisions remain outside the target tolerance. DivShift therefore provides a practical protocol for determining when a Vendi-based generator comparison has stabilized sufficiently to stop sampling, avoiding both premature stopping and unnecessary generations.

## 1 Introduction

Evaluating diversity in text-to-image generation requires deciding how many images to generate for each prompt. Recent studies commonly fix this number $k$ in advance and reuse it across their model or method comparisons, with reported choices ranging from only three to ten images per prompt (Sadat et al., 2024; Kim et al., 2026; Kwon et al., 2026; Jalali et al., 2025). In many cases, however, this choice is treated simply as part of the evaluation protocol rather than as a quantity that itself needs to be justified. A smaller set of works goes further by varying the sample count or repeating the evaluation at different budgets to check whether reported measurements stabilize (Basu et al., 2026; Yin et al., 2026). These checks can support the number $k$ chosen for the models and metric under study, but they do not provide a general criterion for choosing $k$ in a new generator comparison.

The choice of $k$ is particularly important for the Vendi score, which is widely used in recent work on generative diversity (Friedman & Dieng, 2023; Kwon et al., 2026; Jalali et al., 2025; Shi et al., 2026; Li et al., 2026). Vendi can be interpreted as the effective number of modes represented in a sample (Friedman & Dieng, 2023). Since additional samples can reveal previously unobserved modes or variations, the Vendi score of a single generator may continue to change as more samples are added (Ospanov & Farnia, 2025). Prior work studies this behavior for a single generator; we instead ask what it implies for the relative comparison between two generators. Although each Vendi score may keep changing with $k$, we find that two scores often grow by similar factors. Since this proportional growth largely cancels in their ratio, the comparison stabilizes substantially earlier than either score does on its own (Section 3.2; see also Appendix Fig. 7).

However, this cancellation is not equally strong for every pair: different generators can change at different rates, causing the reported ratio to keep shifting for some comparisons, as shown across



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.

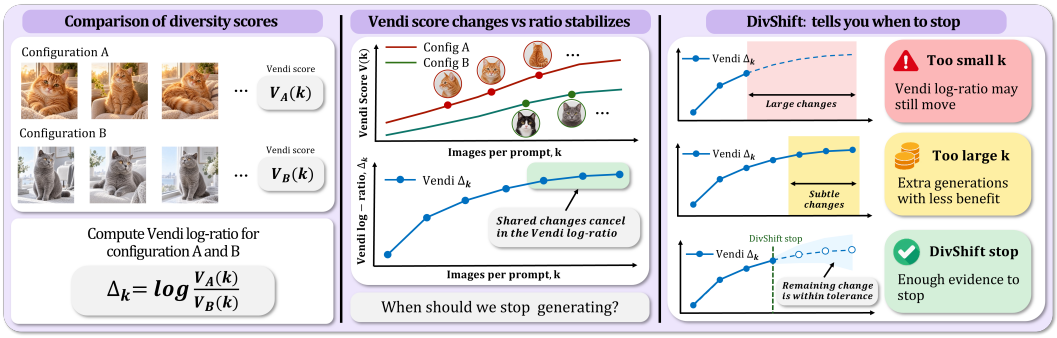


Figure 1: **DivShift overview.** Given two generation configurations, we compare their diversity through the Vendi log-ratio $\Delta_k$. Although the individual Vendi scores may continue to change as more images are generated, shared changes can cancel in the ratio, allowing the relative comparison to stabilize earlier. Rather than fixing the same $k$ for every comparison, DivShift sequentially determines whether the predicted remaining change toward a reference size $K$ is within a user-specified tolerance $\rho$, generating additional images only when needed.

guidance, intervention, and architecture comparisons in Appendix Fig. 8. Consequently, the number of images needed for the comparison to stabilize varies across generator pairs (Section 3.3). A small fixed $k$ can therefore report a ratio that still changes substantially with additional images, whereas a conservatively large $k$ wastes generations on comparisons whose ratio has already stabilized. Figure 1 summarizes this motivation and the resulting evaluation problem.

Motivated by these observations, we introduce **DivShift**, which treats the choice of $k$ as a sequential decision made separately for each Vendi-based generator comparison (Section 4). Rather than fixing the same number of images for every pair, DivShift uses the Vendi log-ratio observed so far to forecast how much the comparison may still change before a user-specified reference size $K$, while accounting for variation across prompts. If the predicted remaining change is within the user-specified tolerance, the current sample size is accepted; otherwise, DivShift estimates a larger sample size, generates only the required additional images, and repeats the decision while reusing all previously generated samples. This allows comparisons that stabilize earlier to stop earlier and allocates additional generations only when they are still needed. At $\rho = 10\%$ and $K = 512$, the macro-average across the five evaluation panels shows $79.3\%$ generation saving, while only $0.10\%$ of stopping decisions remain outside the target tolerance. Similar behavior holds across the other evaluated tolerances (Section 5.2).

In summary, our main contributions are as follows:

1. **Comparison stabilizes before the scores themselves.** We show that a Vendi-based comparison between two generators can stabilize well before either individual Vendi score does, revealing that reliable generator comparison does not require each score itself to have stabilized.
2. **The sufficient $k$ depends on the comparison.** The number of images needed for a stable Vendi ratio varies substantially across generator pairs: small fixed budgets can stop too early, while large ones waste generations on comparisons that have already stabilized.
3. **DivShift adapts $k$ to each comparison.** We introduce DivShift, an adaptive stopping procedure evaluated across a broad set of guidance, intervention, and architecture comparisons. At $\rho = 10\%$ and $K = 512$, it saves $79.3\%$ of generations with only $0.10\%$ of stopping decisions outside the target tolerance.

## 2 Related Work

**Diversity evaluation in text-to-image generation.** Text-to-image diversity has been studied through benchmarks that hold the prompt fixed and measure variation across generated outputs (Rassin et al., 2024; Teotia et al., 2025; Albuquerque et al., 2025). Generative-model evaluation has also developed metrics that disentangle sample quality from distributional coverage (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), while subsequent work has identified important reliability limitations in such evaluation metrics (Naeem et al., 2020; Stein et al., 2023). The Vendi score (Friedman & Dieng, 2023) is widely used for diversity evaluation, summarizing the similarity spectrum of a prompt's images as an effective number of distinct elements (Pasarkar & Dieng, 2024). Recent

sampling and guidance methods use Vendi to compare proposed methods with their baselines (Sadat et al., 2024; Kim et al., 2026; Kwon et al., 2026; Jalali et al., 2025; Nguyen et al., 2026; Ruan et al., 2025; Shi et al., 2026; Li et al., 2026; Askari Hemmat et al., 2024), while prompt-aware variants separate generator-induced variation from variation already present in the prompts (Jalali et al., 2024; Ospanov et al., 2024). Our focus is complementary: given a Vendi-based comparison, we study how many images per prompt are needed to support it.

**Fixed numbers of images per prompt.** Existing Vendi-based evaluations use substantially different sample sizes: four images in manifold-constrained noise optimization (Shi et al., 2026), five in cluster truncation (Nguyen et al., 2026), three to five depending on the backbone in contrastive noise optimization (Kim et al., 2026), ten in condition annealing (Sadat et al., 2024) and representation modulation (Kwon et al., 2026), sixteen in temperature sampling (Li et al., 2026), and 250 per entity–country pair in a geographical-diversity benchmark (Basu et al., 2026). These choices define reproducible protocols, but the appropriate sample size can depend on the comparison, motivating a criterion for selecting $k$ at the level of an individual generator comparison.

**Sample-count, stability, and efficient evaluation.** Empirical model comparisons can be sensitive to stochastic variation and finite-sample metric effects (Dodge et al., 2019; Bouthillier et al., 2021; Bińkowski et al., 2018; Chong & Forsyth, 2020). In text-to-image evaluation, prior work studies sampling sensitivity through repeated seed batches or independent runs (Nguyen et al., 2026; Albuquerque et al., 2025), as well as direct sample-count sweeps (Bhat et al., 2026; Dorfman et al., 2026; Yin et al., 2026). At the metric level, Ospanov & Farnia (2025) study how an individual Vendi score evolves with sample size and propose a truncated variant with convergence guarantees, while prompt-aware variants examine sample-size effects over prompt–image datasets (Jalali et al., 2024). Related work also reduces evaluation cost by selecting fewer prompts (Zhao et al., 2024) or using multi-round evaluation (Zhang et al., 2025). In contrast, DivShift asks when the Vendi log-ratio between two configurations has stabilized sufficiently and adapts the number of images per prompt accordingly, even when the individual scores continue to change.

## 3 When Does a Vendi-Based Comparison Stabilize?

### 3.1 Vendi-Based Diversity Comparison

We use *configuration* to denote a text-to-image generation setting, which may differ in model architecture, guidance setting, sampling strategy, or diversity intervention. Our focus is on pairwise diversity comparisons between such configurations.

For each prompt $p$, we generate $k$ images from configuration $m$ and compute their Vendi score. Given the eigenvalues $\lambda_1, \ldots, \lambda_k$ of the normalized similarity matrix, we compute the per-prompt

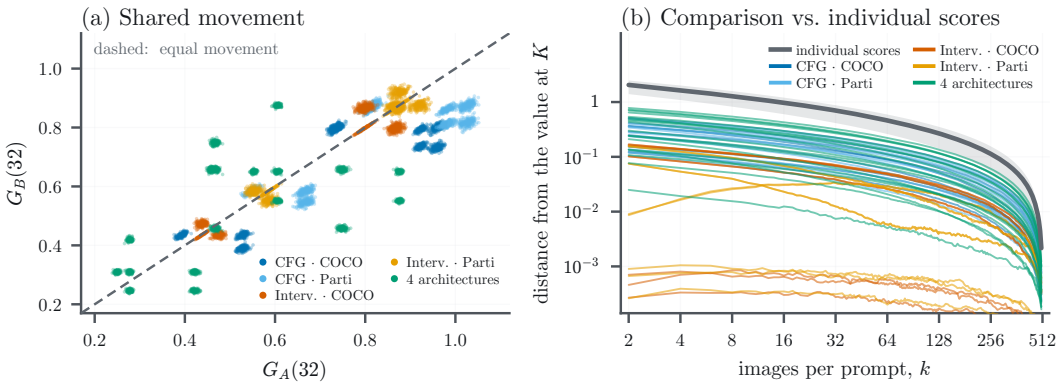


Figure 2: **Shared Vendi-score movement substantially reduces comparison-level change. (a)** Each point is one pairwise evaluation instance (configuration pair × representation × permutation replay; 3 representations, 200 replays). At $k = 32$, the two configurations often undergo similar log-scale changes toward $K = 512$; proximity to the dashed equal-movement line indicates cancellation in their log-ratio. **(b)** Across sample sizes, the remaining comparison-level change is much smaller than the movement remaining in the individual scores ($\sim 10\times$ smaller at $k = 32$), although the rate of decrease varies across comparisons.

and prompt-averaged scores as

$$V_{m,p}(k) = \exp\left(-\sum_{i=1}^{k} \lambda_i \log \lambda_i\right), \qquad V_m(k) = \frac{1}{P}\sum_{p=1}^{P} V_{m,p}(k). \tag{1}$$

Because $V_m(k)$ depends on the observed image set, it can continue to change as $k$ increases (Ospanov & Farnia, 2025). Our question is whether this continued movement of the individual scores necessarily carries over to their relative comparison.

### 3.2 Relative Comparisons Can Stabilize Before Individual Scores

Suppose configurations $A$ and $B$ are compared using Vendi. We represent their relative diversity at $k$ images per prompt by the log-ratio

$$\Delta_k = \log \frac{V_A(k)}{V_B(k)}. \tag{2}$$

A value of $\Delta_k = 0$ indicates equal reported diversity; its sign indicates which configuration has the larger Vendi score, and its magnitude captures the size of their relative difference.

To understand how this comparison changes with sample size, consider how much each individual Vendi score moves from the current size $k$ toward a reference size $K$. For configuration $m$, define

$$G_m(k) = \log \frac{V_m(K)}{V_m(k)}. \tag{3}$$

The change in the comparison is then

$$\Delta_K - \Delta_k = G_A(k) - G_B(k). \tag{4}$$

Equation 4 shows that the comparison depends not on how much the two scores move individually, but on how differently they move. Thus, even if both Vendi scores change substantially with additional samples, their relative comparison can change little when those movements are similar.

Figure 2 shows that this cancellation is substantial in our evaluation panels. At $k = 32$, the individual movements $G_A(32)$ and $G_B(32)$ have medians of $0.737$ and $0.754$ toward $K = 512$, whereas the median remaining comparison-level change, $|G_A(32) - G_B(32)|$, is only $0.086$ (Fig. 2(a)). Figure 2(b) shows the same pattern across sample sizes: the Vendi log-ratio remains much closer to its value at $K$ than either individual score does, with roughly a $10\times$ difference at $k = 32$. However, the comparison-level trajectories decrease at noticeably different rates across configuration pairs, suggesting that the sample size needed for stabilization is itself comparison-dependent.

### 3.3 The Sufficient $k$ Varies Across Comparisons

Because the stabilization rate differs across configuration pairs, we define when a particular comparison is sufficiently stable to report. Given a user-specified reference size $K$, we measure the remaining

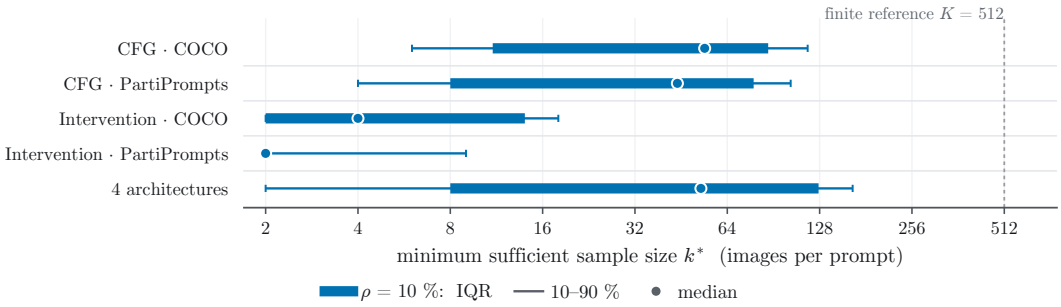


Figure 3: **The sufficient sample size varies substantially across Vendi comparisons.** Minimum sufficient sample size $k^\star_\rho(K)$ at $\rho = 10\%$ and $K = 512$, grouped by comparison family and prompt bank. The point denotes the median, the thick bar marks the interquartile range (IQR), and the thin line mark the 10th–90th percentile range. $k^\star_\rho(K)$ is a hindsight quantity used to characterize the problem and is not available to DivShift during evaluation. The evaluation panels, representations, and replay protocol follow Section 5.1.

comparison-level change at sample size $k$ as

$$R_{k\to K} = |\Delta_K - \Delta_k| \,. \tag{5}$$

We use a relative tolerance $\rho$, because the reported comparison is a ratio. Specifically, $\rho$ controls how much the Vendi ratio at $k$ may differ multiplicatively from its value at $K$. We call $k$ sufficient when

$$\frac{1}{1+\rho} \le \frac{V_A(k)/V_B(k)}{V_A(K)/V_B(K)} \le 1+\rho. \tag{6}$$

These reciprocal bounds are symmetric on the log scale. Defining $\varepsilon_\rho = \log(1+\rho)$, the criterion is equivalently

$$R_{k\to K} \le \varepsilon_\rho. \tag{7}$$

The corresponding hindsight minimum is

$$k^\star_\rho(K) = \min\{k : |\Delta_K - \Delta_k| \le \varepsilon_\rho\} \,. \tag{8}$$

Thus, $k^\star_\rho(K)$ is the smallest number of images per prompt whose reported comparison is within the target tolerance of the comparison observed at reference size $K$.

Figure 3 shows that this sufficient sample size varies substantially across comparison families and prompt banks. At $\rho = 10\%$ and $K = 512$, the panel medians range from only a few images per prompt for some intervention comparisons to roughly fifty images for the CFG and architecture comparisons, while the upper ranges extend beyond one hundred images for some settings. A single fixed $k$ therefore creates a two-sided problem: a small choice leaves some comparisons outside the target tolerance, whereas a conservatively large choice spends unnecessary generations on comparisons that have already stabilized.

## 4 Method: DivShift

The sufficient sample size $k^\star_\rho(K)$ defined in Section 3.3 is a hindsight quantity, since it requires observing the comparison at the reference size $K$. To determine whether the current sample size is sufficient during evaluation, we introduce DivShift, which makes this decision prospectively from the trajectory observed so far. DivShift consists of three steps: forecasting the remaining comparison-level change, accounting for uncertainty across prompts, and adaptively acquiring additional images when the stopping criterion is not satisfied.

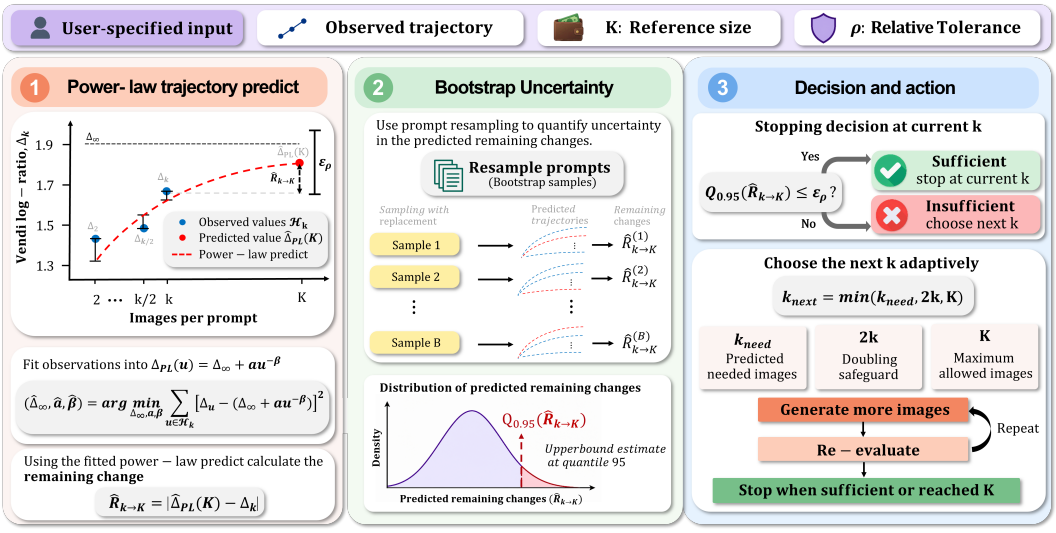


Figure 4: **How DivShift decides.** Given a user-specified reference size $K$ and relative tolerance $\rho$, DivShift uses the observed Vendi log-ratio trajectory to estimate the remaining change, quantifies uncertainty across prompts with bootstrap resampling, and decides whether the current sample size is sufficient. If not, it selects a larger sample size, generates only the required additional images, and repeats the procedure.

### 4.1 FORECASTING THE COMPARISON TRAJECTORY

At the current sample size $k$, the remaining change $R_{k\to K} = |\Delta_K - \Delta_k|$ cannot be observed because $\Delta_K$ is not yet available. DivShift therefore estimates it from the comparison trajectory observed so far. Let $\mathcal{H}_k$ denote the sample sizes at which the comparison has been evaluated so far, the largest of which is the current size $k$. We refer to these as the visited sample sizes, with $u \in \mathcal{H}_k$ indexing a visited sample size.

Representative comparison trajectories become progressively flatter as the sample size increases, while the rate of this stabilization varies across comparisons (Appendix Fig. 8). To capture this behavior, we model the observed comparison values $\{\Delta_u : u \in \mathcal{H}_k\}$ using the power-law trajectory

$$\Delta_{\mathrm{PL}}(u) = \Delta_\infty + a u^{-\beta}. \tag{9}$$

For $\beta > 0$, the trajectory approaches the finite asymptote $\Delta_\infty$ as $u$ increases. Moreover, from Eq. 9,

$$\Delta_{\mathrm{PL}}(2u) - \Delta_{\mathrm{PL}}(u) = a(2^{-\beta} - 1)u^{-\beta}, \tag{10}$$

so the magnitude of the change over each successive doubling is reduced by a factor of $2^{-\beta}$. Thus, smaller $\beta$ corresponds to slower stabilization, whereas larger $\beta$ corresponds to faster stabilization.

Because Eq. 9 has three free parameters, at least three observed sample sizes are required to fit the trajectory. We initialize DivShift with five nested observations,$\mathcal{H}_{32} = \{2, 4, 8, 16, 32\}$, before the first prospective decision. The sensitivity to the number and spacing of these initial observations is examined in Appendix C.1.

The parameters in Eq. 9 are estimated from all sample sizes visited so far by minimizing the least-squares objective

$$(\widehat{\Delta}_\infty, \widehat{a}, \widehat{\beta}) = \underset{\Delta_\infty, a,\, \beta \in [\beta_{\min}, \beta_{\max}]}{\arg\min} \sum_{u \in \mathcal{H}_k} \left[\Delta_u - \left(\Delta_\infty + a u^{-\beta}\right)\right]^2. \tag{11}$$

We use the fixed bounds $\beta_{\min} = 0.05$ and $\beta_{\max} = 3$ throughout our experiments. Details of the numerical fitting procedure are provided in Appendix B.1, and sensitivity to the bounds is reported in Appendix C.3.

After each new sample size is visited, it is added to $\mathcal{H}_k$ and the trajectory is refit using all visited observations accumulated so far(see Appendix Table 9 for a comparison with fitting every intermediate prefix). We denote the resulting fitted trajectory by

$$\widehat{\Delta}_{\mathrm{PL}}(u) = \widehat{\Delta}_\infty + \widehat{a} u^{-\widehat{\beta}}. \tag{12}$$

The fitted value at $K$ forecasts where the comparison would move if generation continued to the reference size. The predicted remaining change from the current observed comparison is therefore

$$\widehat{R}_{k\to K} = \left|\widehat{\Delta}_{\mathrm{PL}}(K) - \Delta_k\right|. \tag{13}$$

### 4.2 BOOTSTRAP UNCERTAINTY AND STOPPING

A point forecast alone is not sufficient for deciding whether to stop. The observed trajectory depends on the prompts used in the evaluation, and this variation propagates through the fitted power law and its extrapolation to $K$. DivShift therefore quantifies uncertainty in the predicted remaining change by bootstrap resampling over prompts (Efron, 1979).

For each of $B$ bootstrap replicates, prompts are sampled with replacement. The same resampled prompts are used for both configurations and at every visited sample size, preserving the paired comparison and the trajectory structure. For each replicate, the prompt-averaged Vendi scores, the complete comparison trajectory, the power-law fit, and the predicted remaining change are recomputed from scratch. Let $\widehat{R}^{(b)}_{k\to K}$ denote the resulting forecast in bootstrap replicate $b$.

We summarize these bootstrap forecasts by their upper quantile

$$U_k(q) = Q_q\left(\left\{\widehat{R}^{(b)}_{k\to K}\right\}_{b=1}^{B}\right). \tag{14}$$

DivShift stops at the current sample size when

$$\boxed{U_k(q) \le \varepsilon_\rho}. \tag{15}$$

Thus, stopping requires the upper bootstrap quantile of the predicted remaining change to lie within the target tolerance. When a study compares more than two configurations, DivShift advances them together. In particular, it stops only when every pair satisfies Eq. 15 at the same sample size. We use $q = 0.95$ throughout our main experiments; sensitivity to this choice is reported in Appendix C.2.

### 4.3 Adaptive Acquisition

When Eq. 15 is not satisfied, DivShift must decide how many additional images to generate. Rather than following a fixed sequence of sample sizes, DivShift uses the same bootstrap forecasts to estimate the smallest future sample size expected to satisfy the stopping criterion:

$$k_{\text{need}} = \min\left\{u : \; k < u \le K, \; Q_{0.95}\left(\left\{\left|\widehat{\Delta}^{(b)}_{\text{PL}}(K) - \widehat{\Delta}^{(b)}_{\text{PL}}(u)\right|\right\}_{b=1}^{B}\right) \le \varepsilon_\rho\right\}. \tag{16}$$

Thus, acquisition and stopping use the same criterion.

The initialization in Section 4.1 uses geometrically spaced sample sizes that double at each step. We retain the same maximum step size after initialization by allowing each acquisition to increase the current sample size by at most one doubling:

$$k_{\text{next}} = \min\left\{k_{\text{need}}, \; 2k, \; K\right\}. \tag{17}$$

Thus, DivShift adapts the next sample size to the predicted need while advancing by at most one doubling before refitting. Only the additional images required to reach $k_{\text{next}}$ are generated, and all previous images are reused. After each acquisition, we update $\mathcal{H}_{k_{\text{next}}} = \mathcal{H}_k \cup \{k_{\text{next}}\}$ and refit the power-law model using all visited sample sizes. This sparse acquisition avoids unnecessary intermediate evaluations, improving generation efficiency.

**Overall procedure.** Figure 4 summarizes the procedure. At each visited sample size, DivShift fits the Vendi log-ratio trajectory and uses prompt-bootstrap forecasts to test whether the predicted remaining change is within $\varepsilon_\rho$. If not, the same forecasts determine the next sample size, only the required additional images are generated, and the process repeats until stopping or reaching $K$.

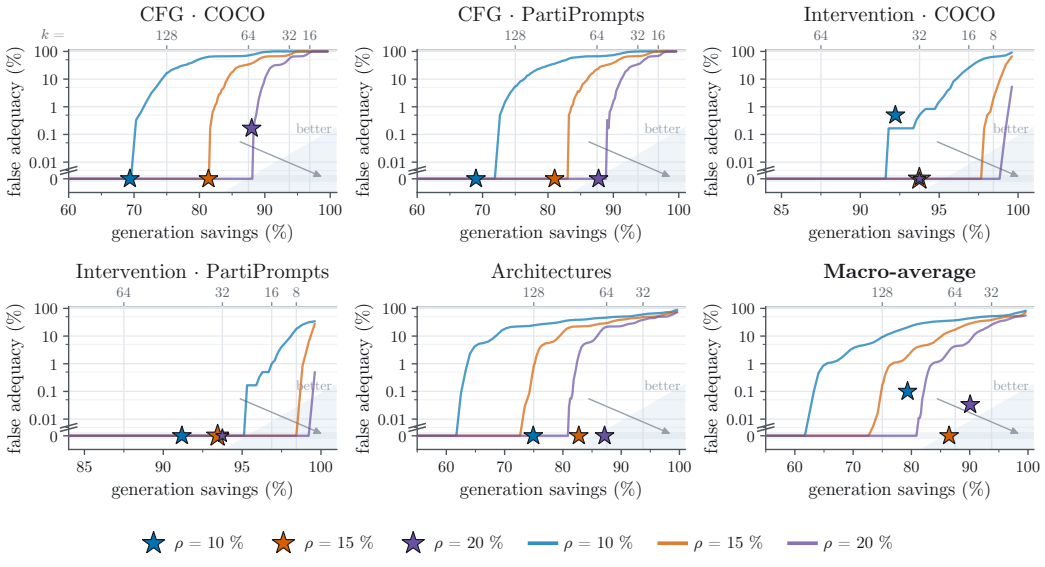


Figure 5: **No single fixed sample size is both safe and economical across all panels.** Curves show the false-adequacy–generation-saving trade-off of fixed sample sizes, while stars mark DivShift at the same tolerance. Lower and farther right is better; false adequacy is shown on a log scale, and the final panel reports the unweighted macro-average across the five panels.

# 5 EXPERIMENTS

## 5.1 EXPERIMENTAL SETUP

**Evaluation panels.** We evaluate DivShift across comparisons that vary in guidance, diversity intervention, and model architecture. On SD 1.4 (Rombach et al., 2022), we consider two families of three configurations each: guidance scales CFG $3/6/9$ (Ho & Salimans, 2022), and the unmodified SD 1.4 baseline (Base), CADS (Sadat et al., 2024), and CNO (Kim et al., 2026). Each family is evaluated separately on $64$ COCO captions (Lin et al., 2014) and $64$ PartiPrompts (Yu et al., 2022). We additionally evaluate four model architectures (SD 1.4, SDXL-LCM (Podell et al., 2024; Luo et al., 2023), FLUX.1-schnell (Black Forest Labs, 2024), and SD3 (Esser et al., 2024)) on $240$ constructed concept–attribute prompts (Rassin et al., 2024). We compute the standard Vendi score using DINOv2 (Oquab et al., 2024), OpenCLIP (Cherti et al., 2023), and DreamSim (Fu et al., 2023) representations of the generated images.

**Sampling protocol.** For each prompt and configuration, images follow a fixed nested seed order, so that the images available at a smaller sample size are always a prefix of those available at a larger sample size. Unless otherwise stated, we use $K = 512$ as the reference size. DivShift uses the initial trajectory defined in Section 4.1; after the first prospective decision, only sample sizes selected by the procedure are visited, and all previously generated images are reused. To evaluate the sequential procedure over different realizations of the sample order, we replay each rule on $200$ permutations of the fixed seed pool within each prompt. The same seed ordering is shared across compared configurations within a prompt. Additional generation and protocol details are provided in Appendix F.

**Evaluation criteria.** For a target relative tolerance $\rho$, a stopping decision at sample size $k$ is *falsely adequate* when $R_{k\to K} > \varepsilon_\rho$. We define *false adequacy* (FA) as the percentage of such stopping decisions. We also measure generation saving relative to evaluating every comparison at $K$:

$$\text{Saved} = 100\left(1 - \frac{\bar{k}}{K}\right), \tag{18}$$

where $\bar{k}$ is the mean stopping sample size across evaluation instances; comparisons that do not stop before $K$ are assigned $k = K$. Our primary setting is $\rho = 10\%$; we additionally report $15\%$ and $20\%$. Because the full image pools were generated before replaying each sequential rule, generation saving is evaluated retrospectively, while each rule itself uses only the observations available at its current sample size.

## 5.2 ADAPTIVE EVALUATION VERSUS FIXED SAMPLE SIZES

We first ask whether a single fixed sample size can provide both reliable and efficient evaluation across comparisons. Figure 5 shows that no single choice achieves both across all panels. At $\rho = 10\%$ and $K = 512$, a fixed $k = 128$ saves $75\%$ of generations but incurs $9.32\%$ false adequacy, whereas $k = 256$ eliminates false adequacy at only $50\%$ saving. In contrast, DivShift achieves $79.3\%$ saving with just $0.10\%$ false adequacy. At $\rho = 15\%$ and $20\%$, it saves $86.4\%$ and $90.1\%$ with $0.00\%$ and $0.03\%$ false adequacy, respectively.

The smaller gains on the intervention panels are consistent with their earlier stabilization. These comparisons exhibit the strongest cancellation of finite-sample Vendi-score movement (Appendix Table 3), and small fixed budgets already perform well on them (Appendix E). In contrast, CFG and

Table 1: **Which trend model should forecast the remaining change?** Macro-average over five panels. All variants differ only in the fitted trajectory model; MAE measures point-forecast error for the remaining comparison-level change.

| | | $\rho = 10\,\%$ (primary) | | | $\rho = 15\,\%$ | | $\rho = 20\,\%$ | |
|---|---|---|---|---|---|---|---|---|
| Trend model | Form | FA↓ | Save↑ | MAE↓ | FA↓ | Save↑ | FA↓ | Save↑ |
| Linear | $a + b\,u$ | 0.00 | 20.7 | 0.5753 | 0.00 | 29.7 | 0.00 | 35.1 |
| Logarithmic | $a + b\log u$ | 0.03 | 72.4 | 0.0384 | 0.00 | 82.1 | 0.00 | 87.2 |
| Exponential | $\Delta_\infty + a\,e^{-\beta u}$ | 51.65 | 90.9 | 0.0789 | 37.84 | 91.8 | 24.47 | 92.2 |
| Power-law **(DivShift)** | $\Delta_\infty + a\,u^{-\beta}$ | 0.10 | 79.3 | 0.0200 | 0.00 | 86.4 | 0.03 | 90.1 |

architecture comparisons require deeper evaluation, leaving more room for adaptive sampling. Thus, the variation in DivShift's gains across panels reflects the comparison-specific sample requirements that motivate the method.

### 5.3 Forecasting the Remaining Change

The stopping decision depends directly on predicting how much the Vendi log-ratio may still change before the reference size $K$. We therefore test whether the power-law trajectory used by DivShift is important for this decision by replacing it with linear, logarithmic, and exponential alternatives, while holding the observed trajectory, prompt bootstrap, tolerance, and acquisition rule fixed.

Table 1 shows that the shape of the forecast has a substantial effect on stopping behavior. At $\rho = 10\%$, the linear model is overly conservative, yielding zero false adequacy but only $20.7\%$ generation saving. The exponential model fails in the opposite direction: it flattens too quickly, saving $90.9\%$ but producing $51.65\%$ false adequacy. Models that allow a more gradual flattening perform substantially better. The logarithmic model achieves $0.03\%$ false adequacy with $72.4\%$ saving, while the power-law model used by DivShift reaches $0.10\%$ false adequacy with $79.3\%$ saving and the lowest forecast error among the tested models (MAE $= 0.0200$). These results suggest that the key modeling requirement is not a uniquely correct functional form, but capturing the slowly diminishing residual change of the comparison trajectory; among the tested models, the power law best balances forecast accuracy and generation savings.

| | $\rho = 10\,\%$ | | $\rho = 15\,\%$ | | $\rho = 20\,\%$ | |
|---|---|---|---|---|---|---|
| Stage | FA↓ | Save↑ | FA↓ | Save↑ | FA↓ | Save↑ |
| 1 Power-law forecast | 0.80 | 77.8 | 1.37 | 85.7 | 1.63 | 90.1 |
| 2 + Bootstrap | 0.10 | 71.3 | 0.00 | 82.9 | 0.03 | 87.4 |
| 3 + Adaptive acquisition **(DivShift)** | 0.10 | 79.3 | 0.00 | 86.4 | 0.03 | 90.1 |

Table 2: **Each component plays a distinct role.** The power-law forecast enables prospective stopping, the prompt bootstrap reduces false adequacy, and adaptive acquisition recovers generation savings. Macro-average over the five panels; the bootstrap quantile is $q = 0.95$.

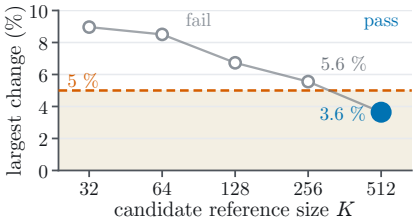


Figure 6: **Reference-size validation.** $K = 512$ is the smallest tested reference size for which one further doubling changes every evaluated comparison by less than $5\,\%$.

### 5.4 Ablations and Reference-Size Sensitivity

Table 2 shows distinct roles for DivShift's components: the power-law forecast enables prospective stopping, the prompt bootstrap reduces false adequacy, and adaptive acquisition recovers generation savings. At $\rho = 10\%$, DivShift achieves $0.10\%$ false adequacy and $79.3\%$ generation savings, with the same qualitative pattern at $\rho = 15\%$ and $20\%$. DivShift uses a user-specified finite reference rather than a universal $K$. We use $K = 512$ and validate it on the CFG–COCO panel with a one-more-doubling check: the largest comparison-level change falls from $5.6\%$ at $K = 256$ to $3.6\%$ at $K = 512$. Thus, $K = 512$ is stable for this panel under our $5\%$ criterion, not a universal choice; users should choose $K$ based on the desired reference fidelity and generation budget.

## 6 Conclusion

We showed that Vendi-based generator comparisons can stabilize well before their individual scores, while the sample size required for this stabilization varies substantially across comparisons. This makes the number of images per prompt not merely a fixed evaluation setting, but a comparison-specific quantity that should be justified. DivShift operationalizes this view by predicting the remaining comparison-level change, accounting for uncertainty across prompts, and allocating additional images only when they are still needed. At $\rho = 10\%$ and $K = 512$, DivShift saves 79.3% of generations with 0.10% false adequacy; at $\rho = 15\%$ and $20\%$, the savings increase to 86.4% and 90.1%, with 0.00% and 0.03% false adequacy, respectively. These results support treating sample size as a comparison-specific evaluation decision rather than a universal fixed constant.

## AI USE STATEMENT

Generative AI tools were used only to assist with editing and polishing the manuscript for clarity, grammar, and readability. They were not used to develop the research methodology, design or execute experiments, formulate mathematical claims, generate reported data, or interpret the experimental results. The authors reviewed all AI-assisted edits and take full responsibility for the final content of the paper.

## ETHICS STATEMENT

This work evaluates existing text-to-image models and proposes an evaluation procedure for measuring the stability of diversity comparisons; it introduces no new image-generation capability. The experiments use public prompt sources and generated images only to compute aggregate diversity statistics, and involve no human subjects. Our diversity measurements quantify perceptual variation in model outputs and should not be interpreted as measures of demographic representation or fairness.

## REPRODUCIBILITY STATEMENT

We provide detailed specifications of the evaluation panels, score computation, nested sample order, and sequential replay protocol in Appendix F. Appendix B describes the numerical power-law fitting, prompt-bootstrap procedure, and adaptive acquisition rule, while Appendix C reports the ablations used to fix the main design choices. All sequential evaluations use fixed nested image sets and frozen permutation replays so that comparisons across sample sizes reuse the same underlying samples.

## REFERENCES

Isabela Albuquerque, Ira Ktena, Olivia Wiles, Ivana Kajić, Amal Rannen-Triki, Cristina Vasconcelos, and Aida Nematzadeh. Benchmarking diversity in image generation via attribute-conditional human evaluation, 2025.

Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized Vendi score guidance. In *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2024. doi: 10.1007/978-3-031-73021-4_13. arXiv:2406.04551.

Abhipsa Basu, Mohana Singh, Shashank Agnihotri, Margret Keuper, and R. Venkatesh Babu. GeoDiv: Framework for measuring geographical diversity in text-to-image models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.22120.

Pradhaan S. Bhat, Rishubh Parihar, Abhijnya Bhat, and R. Venkatesh Babu. Don't settle at the mode! mitigating diversity collapse in pretrained flow models via feature self-guidance, 2026.

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

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, 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 A. Smola, A. Dimakis, and I. Stoica (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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)*, pp. 6070–6079, 2020.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Sara Dorfman, Maya Vishnevsky, Omer Dahary, Or Patashnik, and Daniel Cohen-Or. Semantic browsing: Controllable diversity for image generation, 2026.

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, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. arXiv:2210.02410.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems 36*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. A short version appeared in the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional Vendi score: Prompt-aware diversity evaluation for generative AI models and LLMs, 2024.

Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. SPARKE: Scalable prompt-aware diversity and novelty guidance in diffusion models via RKE score. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.10173.

Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. In *International Conference on Learning Representations (ICLR)*, 2026.

Dahee Kwon, Haeun Lee, and Jaesik Choi. Breaking the lock-in: Diversifying text-to-image generation via representation modulation. *arXiv preprint arXiv:2606.06813*, 2026. International Conference on Machine Learning (ICML).

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*, 2019.

Peizhuo Li, Emre Aksan, Alexandru-Eugen Ichim, Thabo Beeler, and Olga Sorkine-Hornung. Diversify diffusion with temperature sampling and variance-corrective time shifting, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10602-1_48.

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

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020.

Trang Nguyen, Shuang Wu, Runyan Tan, and Phillip Howard. Improving sample diversity in autoregressive text-to-image generation via cluster truncation, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. arXiv:2304.07193.

Azim Ospanov and Farzan Farnia. Do Vendi scores converge with finite samples? truncated Vendi score for finite-sample convergence guarantees, 2025. v3, 24 June 2025.

Azim Ospanov, Mohammad Jalali, and Farzan Farnia. Scendi score: Prompt-aware diversity evaluation via Schur complement of CLIP embeddings, 2024.

Amey P. Pasarkar and Adji Bousso Dieng. Cousins of the Vendi score: A family of similarity-based diversity metrics for science and machine learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3808–3816, 2024. arXiv:2310.12952.

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

Royi Rassin, Aviv Slobodkin, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. GRADE: Quantifying sample diversity in text-to-image models, 2024.

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)*, pp. 10684–10695, 2022.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, Yi-Lun Wu, and Hong-Han Shuai. PromptMoG: Enhancing diversity in long-prompt image generation via prompt embedding mixture-of-gaussian sampling, 2025.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations (ICLR)*, 2024.

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*, 2018.

Qitan Shi, Cheng Jin, Ziyuan Liu, and Yuantao Gu. Manifold-constrained noise optimization for diverse diffusion sampling, 2026.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric 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*, 2023.

Revant Teotia, Candace Ross, Karen Ullrich, Sumit Chopra, Adriana Romero-Soriano, Melissa Hall, and Matthew J. Muckley. DIMCIM: A quantitative evaluation framework for default-mode diversity and generalization in text-to-image generative models, 2025.

Xuanhua Yin, Chuanzhi Xu, Shunqi Mao, Wei Guo, and Weidong Cai. DefaultShift: Auditing semantic default shift in accelerated text-to-image models, 2026.

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. ISSN 2835-8856.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7561–7582, 2025. arXiv:2412.09645.

Lin Zhao, Tianchen Zhao, Zinan Lin, Xuefei Ning, Guohao Dai, Huazhong Yang, and Yu Wang. FlashEval: Towards fast and accurate evaluation of text-to-image diffusion generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16122–16131, 2024.

# A Additional Evidence for Comparison-Level Stabilization

This appendix supports §1 and §3.2: the two Vendi scores of a comparison move together, most of that movement cancels in their log-ratio, and the rate at which the remainder decays differs across comparisons. Every value here is computed on the frozen caches of Appendix F; no image, embedding or Vendi score is recomputed.

## A.1 Raw Score and Ratio Trajectories

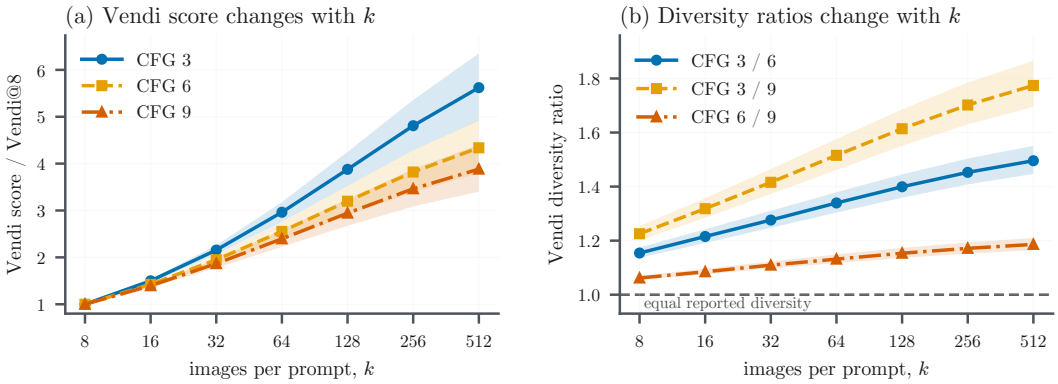


Figure 7: **Individual Vendi scores and the ratios formed from them, as the number of images per prompt grows.** Guidance panel, 64 COCO captions, DINOv2 representation. **(a)** Prompt-averaged Vendi score for CFG 3, 6 and 9, each normalized to its own value at $k = 8$ so that the three trajectories can be read on one axis. All three keep rising over the whole range of sample sizes. **(b)** The ratios formed from the same three trajectories. Pair orientation places the smaller guidance scale in the numerator (CFG $3/6$, $3/9$, $6/9$); the horizontal line at one marks equal reported diversity. Bands are prompt-bootstrap intervals. The ratios move over a range an order of magnitude narrower than the scores they are built from — the observation §3.2 makes precise.

Figure 7 shows one panel in raw units, before any of the transformations used elsewhere in the paper. It is the visual form of Eq. 4: the three score trajectories in (a) rise together, and what survives in (b) is only the part of that rise on which they disagree.

## A.2 Comparison Trajectories Across All Evaluated Comparisons

Figure 8 is the evidence behind the trend model of §4.1. Panel (b) is the informative one: on log–log axes a power-law trajectory is a straight line whose slope is $-\beta$, the curves are close to straight over the range that matters, and their slopes differ by roughly a factor of five across panels. A single fitted exponent per comparison therefore provides a compact model of the observed flattening while allowing the stabilization rate to vary across comparisons.

## A.3 How Much of the Individual Movement Cancels

Table 3 is the source of the numbers quoted in §3.2. At $k = 32$ — the sample size at which DivShift takes its first prospective decision — the macro-average median change still ahead of each individual score is $0.737$ and $0.754$ in log units, about eight times $\varepsilon_{10\%} = 0.0953$, while the median change still ahead of the comparison built from them is $0.086$, inside one tolerance. The two configurations of a comparison move in the *same direction* in $100.0\,\%$ of all $10{,}800$ evaluation instances and in every panel, so the cancellation is not an averaging artifact of mixed signs.

The retained share varies across panels in the direction the rest of the paper predicts: it is smallest on the intervention panels, whose comparisons are nearly stable at every sample size, and largest on the architecture panel, whose two configurations differ most in how their scores scale. That spread, not the macro-average, is what makes the required sample size comparison-specific.

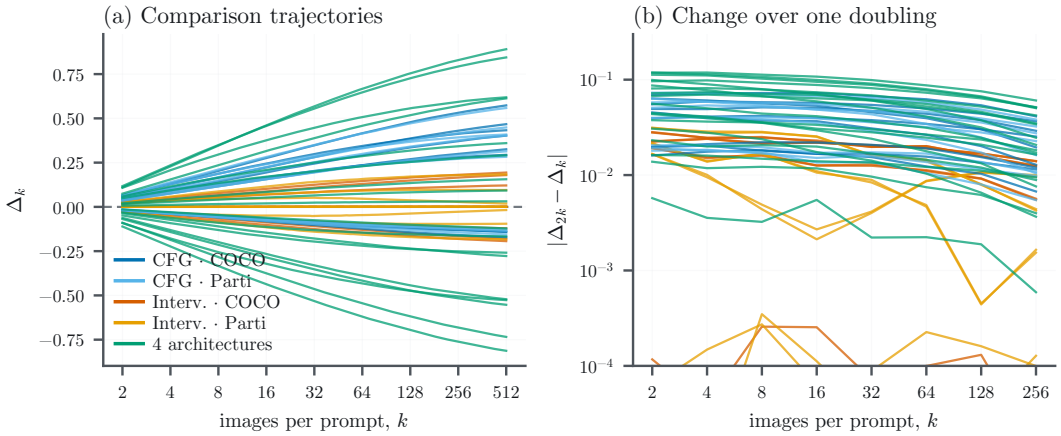


Figure 8: **Comparison trajectories flatten with sample size, at rates that differ across comparisons.** All 54 evaluated comparisons (group × prompt bank × representation × pair); each curve is the median over that cell's 200 permutation replays. **(a)** The Vendi log-ratio $\Delta_k$ against $k$ on a logarithmic sample-size axis. Curves fan out from near zero at $k = 2$ and bend toward a limit at different sample sizes. **(b)** The change over one doubling, $|\Delta_{2k} - \Delta_k|$, on log–log axes. A power-law trajectory $\Delta_{\mathrm{PL}}(u) = \Delta_\infty + a\,u^{-\beta}$ predicts a straight line of slope $-\beta$ here (Eq. 9), and the spread of slopes is the spread of stabilization rates the method must accommodate. Six series — the three CNO/Base representations in each of the two prompt banks — fall below the axis: their comparison is at most $8 \times 10^{-4}$ in log units even at $K = 512$, so their doubling increments sit at the numerical floor and measure nothing about stabilization.

Table 3: **How much of the individual movement cancels, at $k = 32$.** $G_m(32) = \log[V_m(512)/V_m(32)]$ is the movement still ahead of configuration $m$; $R_{32\to K} = |G_A(32) - G_B(32)|$ is the movement still ahead of the comparison. The retained share is the per-instance ratio $R_{32\to K}\big/\frac{1}{2}(|G_A(32)| + |G_B(32)|)$. An instance is one (group, prompt bank, representation, pair, replay); the macro row is the unweighted mean of the five panel medians, never a pooled median. For scale, $\varepsilon_{10\%} = 0.0953$.

| | | median individual | | median | retained | same |
|---|---|---|---|---|---|---|
| panel | $n$ | $\lvert G_A(32)\rvert$ | $\lvert G_B(32)\rvert$ | $R_{32\to K}$ | share | sign (%) |
| CFG · COCO | 1800 | 0.739 | 0.739 | 0.126 | 0.175 | 100.0 |
| CFG · Parti | 1800 | 0.820 | 0.820 | 0.114 | 0.135 | 100.0 |
| Intervention · COCO | 1800 | 0.794 | 0.794 | 0.044 | 0.073 | 100.0 |
| Intervention · Parti | 1800 | 0.863 | 0.863 | 0.016 | 0.019 | 100.0 |
| 4 architectures | 3600 | 0.470 | 0.553 | 0.128 | 0.294 | 100.0 |
| **Macro (5 panels)** | 10,800 | 0.737 | 0.754 | 0.086 | 0.139 | 100.0 |

# B DivShift Fitting and Implementation Details

This appendix gives the implementation behind §4 in enough detail to reproduce it. Every stage runs on CPU from cached per-prompt Vendi scores; no part of the method reads an image.

## B.1 Numerical Solution of the Power-Law Fit

The objective of Eq. 11 is nonlinear in $\beta$ but *linear* in $\Delta_\infty$ and $a$ once $\beta$ is fixed: with $x_u = u^{-\beta}$ the model is an ordinary straight line $\Delta_\infty + a\,x_u$. We exploit this by profiling out the two linear parameters rather than running a general nonlinear optimizer.

1. **Grid over $\beta$.** $\beta$ is evaluated on a fixed grid of 60 values spaced geometrically over $[\beta_{\min}, \beta_{\max}] = [0.05, 3]$, that is $\beta_j = \exp\big(\log 0.05 + \frac{j}{59}(\log 3 - \log 0.05)\big)$ for $j = 0, \ldots, 59$. The grid is geometric because $\beta$ acts as an exponent, so equal *ratios* of $\beta$, not equal differences, correspond to comparable changes in the fitted trajectory.

2. **Closed-form conditional least squares.** For each $\beta_j$, $\Delta_\infty$ and $a$ are obtained from the two-parameter normal equations in closed form,

$$\hat{a}(\beta_j) = \frac{\sum_{u \in \mathcal{H}_k} (x_u - \bar{x})(\Delta_u - \bar{\Delta})}{\sum_{u \in \mathcal{H}_k} (x_u - \bar{x})^2}, \qquad \hat{\Delta}_\infty(\beta_j) = \bar{\Delta} - \hat{a}(\beta_j)\, \bar{x}, \tag{19}$$

with $\bar{x}$ and $\bar{\Delta}$ the means over the visited sample sizes. No iteration and no starting value are involved, so the fit cannot fail to converge.

3. **Selection.** The residual sum of squares is evaluated as $\mathrm{SSE}(\beta_j) = S_{\Delta\Delta} - \hat{a}(\beta_j)^2 S_{xx}$ and the grid point with the smallest SSE is selected; ties are broken by the first index, so the procedure is deterministic.
4. **Bounds.** $\beta$ cannot leave $[0.05, 3]$: a trajectory whose best unconstrained exponent lies outside the interval is fitted at the nearer endpoint. Both bounds are protocol constants fixed in advance and their effect is measured in Appendix C.3. Over all decisions at $\rho = 10\,\%$ the selected exponent has a macro-average median of $0.171$ and sits at the lower bound in $37.1\,\%$ and at the upper bound in $3.6\,\%$ of point fits.
5. **Degenerate fits.** A fit is recorded as a failure if $\hat{\Delta}_\infty$ is not finite, which happens only when the visited sample sizes produce a degenerate regressor ($S_{xx} = 0$). The rule then treats the stopping criterion as not satisfied and continues sampling, so a failed fit can never cause an early stop. Across all panels, tolerances and decisions the observed fit-failure rate is $0.0\,\%$.

The same routine fits the three alternative trend models of §5.3. The linear and logarithmic models are single regressions on $u$ and $\log u$ and need no grid; the exponential model $\Delta_\infty + a\, e^{-\beta u}$ uses the identical profile-and-grid procedure with a 60-point geometric grid over $\beta \in [2^{-10}, 0.5]$, chosen so that the implied half-life spans the range of sample sizes.

### B.2 The Sequential Procedure

1. Initialize the visited sample sizes $\mathcal{H} \leftarrow \{2, 4, 8, 16, 32\}$ and $k \leftarrow 32$. The five entries are nested prefixes of the same frozen 32-image sample, so the initial trajectory costs 32 images per prompt and not 62.
2. Form $\Delta_u$ for every $u \in \mathcal{H}$ and fit Eq. 9 by Appendix B.1, once on the observed prompt set and once inside each of the $B$ bootstrap replicates (Appendix B.3).
3. Compute $U_k(0.95)$ from Eq. 14. If $U_k(0.95) \le \varepsilon_\rho$, report at $k$ and stop.
4. Otherwise choose $k_{\mathrm{next}}$ by Eq. 17 (Appendix B.4), generate only the images needed to extend every prompt from $k$ to $k_{\mathrm{next}}$, set $\mathcal{H} \leftarrow \mathcal{H} \cup \{k_{\mathrm{next}}\}$ and $k \leftarrow k_{\mathrm{next}}$, and return to step 2.
5. If $k$ reaches the reference size $K$ without the criterion being satisfied, the evaluation is charged the full $K$ when generation savings are scored.

**Cells with more than one pair.** A configuration panel compares three configurations and therefore three pairs at once. The rule advances the whole cell together: it reports only when *every* pair of the cell satisfies Eq. 15 at the same sample size, and the acquisition target is the deepest of the per-pair targets. False adequacy is scored the same way, so a cell counts as falsely adequate if *any* of its pairs has moved by more than $\varepsilon_\rho$ between the stopping sample size and $K$. The architecture panel is scored per pair, because its six pairs are not evaluated as one study. This is the source of the different instance counts in Appendix F.4; it is a property of the evaluation unit, not of the rule.

### B.3 Prompt-Bootstrap Implementation

We use $B = 1000$ replicates. For each evaluation cell an index matrix of shape $B \times P$ is drawn once, with entries uniform on $\{1, \ldots, P\}$, from a generator seeded by the first 32 bits of SHA-256 of a master string and the cell's identifiers (panel, prompt bank, representation, functional); the master constant is fixed in the released code. The draw is therefore reproducible bitwise and independent of run order or machine.

Replicate $b$ reuses its single index vector for *both* configurations of a comparison and at *every* visited sample size. This preserves two structures at once: the paired comparison, because both

Table 4: **Initial-trajectory ablation.** The observed trajectory is initialized with nested sample sizes before the first prospective decision; everything else – accumulated-history power-law fitting, $\beta \in [0.05, 3]$, $Q_{0.95}$ prompt bootstrap, $k_{\text{next}} = \min(k_{\text{need}}, 2k, K)$, $K = 512$ – is identical. Longer initial trajectories slightly reduce saving but substantially improve safety and forecast accuracy; $\{2, 4, 8, 16, 32\}$ is selected. MAE and the under-prediction rate are measured at the first prospective decision, where the candidates differ most.

| | | $\rho = 10\,\%$ | | $\rho = 15\,\%$ | | $\rho = 20\,\%$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| Initial trajectory | first dec. | FA | Save | FA | Save | FA | Save | MAE↓ | under (%) |
| $\{2, 4, 8\}$ | $k = 8$ | 1.63 | 80.0 | 0.30 | 87.9 | 0.20 | 92.0 | 0.0503 | 52.1 |
| $\{2, 4, 8, 16\}$ | $k = 16$ | 0.97 | 79.8 | 0.00 | 87.6 | 0.10 | 91.5 | 0.0331 | 41.2 |
| $\{2, 4, 8, 16, 32\}$ **(selected)** | $k = 32$ | 0.10 | 79.3 | 0.00 | 86.4 | 0.03 | 90.1 | 0.0232 | 34.5 |

configurations see the same resampled prompts, and the nested trajectory, because a replicate's trajectory is the trajectory of one fixed prompt multiset. Within a replicate the prompt-averaged score, the log-ratio trajectory, the power-law fit and the predicted remaining change are all recomputed from scratch; nothing is carried over from the observed fit.

Two details bound what the envelope covers. The resampling is over prompts only, so it quantifies prompt sampling and not seed sampling. And on the architecture panel each replicate additionally draws one of the 200 permutation replays at random, with the reference-size column pinned to a fixed replay so that the replicate trajectory and its reference endpoint stay consistent; the configuration panels need no such step because they are evaluated one replay at a time.

### B.4 Adaptive Acquisition

When the stopping criterion is not satisfied, the predicted required sample size $k_{\text{need}}$ of Eq. 16 is read off the *same* $Q_{0.95}$ prompt-bootstrap envelope that the stopping rule uses, evaluated at candidate future sample sizes instead of at the current one. Using one criterion for both decisions is what makes the target self-consistent: a sample size is requested only if, on arrival, the evidence there would be judged sufficient by the test that will actually be applied.

**Candidate grid.** The sample sizes available to the procedure in our evaluation are the multiples of eight below the reference size, $\mathcal{G} = \{8, 16, 24, \ldots, 504\}$, because those are the sample sizes present in the cached evaluation. A requested $k_{\text{next}}$ is rounded up to the next member of $\mathcal{G}$ strictly greater than the current $k$. This is a property of the cache used for the replay, not of the method: a live evaluation can generate any sample size.

**The one-doubling cap.** Without $\min(\cdot, 2k)$ in Eq. 17 a single early fit — made from five points, the deepest of which is $k = 32$ — could request a sample size close to $K$ and forfeit the opportunity to refit. The cap costs generation savings only when the fit was already right, and the component ablation isolates what it is worth: stage 3 of Table 2 differs from stage 2 *only* in which sample sizes are visited, and recovers 8.0 points of generation savings at $\rho = 10\,\%$ at an unchanged false-adequacy rate.

## C DivShift Ablations and Efficiency Checks

Each ablation below changes exactly one element of the locked procedure and holds everything else fixed: the same cached images, the same pairs, the same representations, the same 200 permutation replays, the same $B = 1000$ prompt bootstrap, the same reference size $K = 512$ and the same false-adequacy criterion. Macro-averages are the unweighted mean of the five panels, which are never pooled.

### C.1 Initial Trajectory: How Many Sample Sizes, and How Spaced

**How many.** Table 4 compares the initial trajectories $\{2, 4, 8\}$, $\{2, 4, 8, 16\}$ and $\{2, 4, 8, 16, 32\}$; the last element of each is the sample size at which the first prospective decision is taken. Shorter initial trajectories decide earlier and save marginally more, but they pay for it in safety and in forecast quality: at $\rho = 10\,\%$ false adequacy falls from $1.63\,\%$ to $0.97\,\%$ to $0.10\,\%$ as the initial trajectory

Table 5: **Initial fitting-grid sensitivity: does DivShift depend on powers-of-two prefix sample sizes?** DivShift initializes its trajectory with geometrically spaced nested prefixes $\{2, 4, 8, 16, 32\}$; subsequent adaptively visited sample sizes are appended to the accumulated fitting history. This ablation varies *only* that initialization – every variant is a set of prefixes of the same frozen 32-image sample, the first prospective decision is at $k = 32$ for all of them, and everything after it is identical. Macro-average over the five panels at $\rho = 10\,\%$; MAE and the under-prediction rate are for the point forecast against the realized remaining change. Sample sizes 3, 5 and 7 are absent from every cache, so "dense" means every cached sample size $\leq 32$.

| Initial grid | # points | FA↓ (%) | Save↑ (%) | MAE↓ | under-pred. (%) |
|---|---|---|---|---|---|
| $\{2, 4, 8, 16, 32\}$ geometric **(current)** | 5 | 0.10 | 79.33 | 0.0200 | 26.4 |
| $\{8, 16, 24, 32\}$ step 8 | 4 | 0.30 | 79.60 | 0.0194 | 38.6 |
| $\{4, 8, \ldots, 32\}$ step 4 | 8 | 0.17 | 79.40 | 0.0181 | 31.1 |
| $\{2, 4, 6, 8, 9, \ldots, 32\}$ dense | 28 | 0.13 | 78.54 | 0.0202 | 24.0 |
| $\{2, 10, 17, 25, 32\}$ linear | 5 | 0.13 | 79.60 | 0.0187 | 28.5 |

Table 6: **The bootstrap quantile is a monotone safety knob.** Macro-average over the five panels; DivShift fits all observed trajectory points accumulated up to the current decision; the observed trajectory is initialized with the nested sample sizes $\{2, 4, 8, 16, 32\}$ before the first prospective decision. only the quantile of the prompt-bootstrap envelope changes. $q = 0.95$ was fixed in advance and is not selected from this sweep.

| | $\rho = 10\,\%$ | | $\rho = 15\,\%$ | | $\rho = 20\,\%$ | |
|---|---|---|---|---|---|---|
| Envelope quantile $q$ | FA↓ | Save↑ | FA↓ | Save↑ | FA↓ | Save↑ |
| 0.50 | 2.27 | 83.4 | 2.33 | 89.0 | 2.80 | 91.8 |
| 0.80 | 0.27 | 81.3 | 0.27 | 87.7 | 0.40 | 90.9 |
| 0.90 | 0.17 | 80.2 | 0.03 | 87.0 | 0.10 | 90.4 |
| 0.95 **(DivShift)** | 0.10 | 79.3 | 0.00 | 86.4 | 0.03 | 90.1 |
| 0.99 | 0.10 | 77.5 | 0.00 | 85.4 | 0.00 | 89.4 |

lengthens, while generation savings vary only between $80.0\,\%$ and $79.3\,\%$. The mechanism is in the last two columns, both measured at the first prospective decision where the candidates differ most: the mean absolute error of the predicted remaining change falls from $0.0503$ to $0.0232$, and the rate at which it *under*-predicts the realized remaining change falls from $52.1\,\%$ to $34.5\,\%$. A three-point fit of a three-parameter model is exactly determined and carries no residual information about its own reliability; five points is the shortest initial trajectory in this set that does. $\{2, 4, 8, 16, 32\}$ is the selected setting.

**How spaced.** Table 5 holds the number of images used to build the initial trajectory fixed — every variant is a set of prefixes of the *same* frozen 32-image sample, and the first prospective decision is at $k = 32$ for all of them — and varies only which prefixes are fitted. Geometric spacing $\{2, 4, 8, 16, 32\}$, arithmetic spacing at steps of 8 and of 4, a linear five-point grid $\{2, 10, 17, 25, 32\}$ and a 28-point dense grid all give false adequacy between $0.10\,\%$ and $0.30\,\%$ and generation savings between $78.5\,\%$ and $79.6\,\%$. DivShift therefore does not depend on powers-of-two spacing. Sample sizes 3, 5 and 7 are absent from every cache, so "dense" means every cached sample size at most 32.

### C.2 Bootstrap-Quantile Sensitivity

The envelope quantile $q$ of Eq. 14 is the only safety parameter of the rule. Table 6 sweeps it over $\{0.50, 0.80, 0.90, 0.95, 0.99\}$ with everything else fixed. It behaves as a monotone knob in the expected direction at every tolerance: raising $q$ lowers false adequacy and lowers generation savings. At $\rho = 10\,\%$, $q = 0.50$ — the bootstrap median, which discards the uncertainty entirely — gives $2.27\,\%$ false adequacy for $83.4\,\%$ savings, while $q = 0.99$ gives $0.10\,\%$ for $77.5\,\%$. The curve is flat near the operating point: between $q = 0.90$ and $q = 0.99$ false adequacy moves by $0.07$ points and savings by $2.7$.

$q = 0.95$ **was fixed in advance and is not selected from this sweep.** It is the conventional one-sided $95\,\%$ envelope, declared before the sweep was run; the sweep is reported to show what the choice costs and buys, not to justify it after the fact.

Table 7: **$\beta$-bound sensitivity under accumulated-history fitting.** Macro-average at $\rho = 10\,\%$; each half sweeps one bound with the other at its locked value. Recomputed from scratch for this method: the accumulated trajectory includes very small sample sizes, which changes boundary behavior relative to any earlier engine. "hit" is the percentage of point fits pinned at the swept bound.

| $\beta_{\max}$ | FA | Save | hit (%) | MAE | $\beta_{\min}$ | FA | Save | hit (%) | MAE |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.10 | 79.3 | 6.4 | 0.0200 | 0.01 | 0.10 | 77.7 | 26.0 | 0.0221 |
| 1.5 | 0.10 | 79.3 | 5.0 | 0.0199 | 0.025 | 0.10 | 78.3 | 29.3 | 0.0214 |
| 2 | 0.10 | 79.3 | 4.4 | 0.0199 | 0.05 **(locked)** | 0.10 | 79.3 | 37.1 | 0.0200 |
| 3 **(locked)** | 0.10 | 79.3 | 3.6 | 0.0200 | 0.1 | 0.37 | 81.8 | 56.3 | 0.0159 |
| 4 | 0.10 | 79.3 | 3.3 | 0.0200 | | | | | |
| 6 | 0.10 | 79.3 | 3.1 | 0.0199 | | | | | |

Table 8: **Evaluation-side compute of the complete sequential replay.** Wall clock for replaying every rule over all 2,418 evaluation units of the five panels, single-threaded (`OMP/MKL/OPENBLAS_NUM_THREADS= 1`), CPU only, on a shared machine; it covers every trajectory fit the rule performs, including the $N_B = 1000$ prompt-bootstrap refit at every decision, and excludes image generation entirely. Absolute values are machine-dependent; the ratios between rows are the transferable quantity.

| trend model | $\rho = 10\,\%$ | | $\rho = 15\,\%$ | | $\rho = 20\,\%$ | |
|---|---|---|---|---|---|---|
| | total (s) | per instance (ms) | total (s) | per instance (ms) | total (s) | per instance (ms) |
| Linear | 41.9 | 17.3 | 41.2 | 17.0 | 43.3 | 17.9 |
| Logarithmic | 32.8 | 13.5 | 24.2 | 10.0 | 20.3 | 8.4 |
| Exponential | 28.9 | 12.0 | 28.1 | 11.6 | 35.4 | 14.6 |
| Power-law **(DivShift)** | 66.6 | 27.5 | 52.8 | 21.8 | 42.2 | 17.4 |

### C.3 Power-Law Exponent Bounds

The bounds $[\beta_{\min}, \beta_{\max}] = [0.05, 3]$ keep the fitted exponent inside a range in which extrapolation to $K$ is meaningful. Table 7 sweeps each bound with the other held at its locked value.

The upper bound does nothing: false adequacy, generation savings and forecast error are unchanged to the reported precision for every $\beta_{\max} \in \{1, 1.5, 2, 3, 4, 6\}$, and the share of point fits pinned at it falls from $6.4\,\%$ to $3.1\,\%$ without any consequence. A trajectory that wants a very large exponent has already flattened, and the forecast it produces is the same either way.

The lower bound is the one that matters, because it controls how slowly a trajectory may be judged to decay. Tightening it to $\beta_{\min} = 0.01$ or $0.025$ leaves false adequacy at $0.10\,\%$ and costs $1.6$ and $1.0$ points of generation savings; loosening it to $0.1$ raises savings to $81.8\,\%$ but also raises false adequacy to $0.37\,\%$, and pins $56.3\,\%$ of point fits at the bound. At the locked value $37.1\,\%$ of point fits sit at $\beta_{\min}$. That this share is large is expected rather than alarming: the near-tie comparisons of Appendix A.2 have essentially flat trajectories at the numerical floor, for which no exponent is identified, and the bound is what keeps their predicted remaining change finite and conservative.

### C.4 Evaluation-Side Compute

Table 8 reports what the decision procedure itself costs, separately from the image generation it is trying to avoid. The dominant term is the $B = 1000$ prompt-bootstrap refit performed at every prospective decision. The power-law model is the most expensive of the four trend models because it evaluates a 60-point exponent grid, and it is still under 30 milliseconds per evaluation instance for a complete sequential replay — against the hundreds of images per prompt per configuration that the decision avoids generating. That is why the paper reports generation savings rather than a combined cost. The absolute numbers come from a single-threaded run on a shared machine and are not a benchmark; the ratios between rows are the transferable quantity, and the measurement covers the replay of the rule over cached Vendi scores, excluding representation extraction and the Vendi computation a live evaluation would also pay at each new sample size.

Table 9: **Fitting every integer prefix does not improve the decision, at any tolerance.** Macro-average over the five panels. *Sparse visited trajectory* fits only the sample sizes the evaluation has actually visited, which is what DivShift does; *true dense history* re-forms every integer prefix $k = 2, \ldots, k_{\mathrm{cur}}$ after each acquisition. Everything else – images, pairs, representations, bootstrap draws, $\beta$ grid, criterion and acquisition rule – is identical. FA is false adequacy (%), Save is generation savings (%), and MAE is the error of the predicted remaining change against the realized one. Fit failures are zero for both. Neighboring integer prefixes correlate at 0.988 across evaluation instances, so the observations dense fitting adds are close to duplicates.

| | $\rho = 10\,\%$ | | $\rho = 15\,\%$ | | $\rho = 20\,\%$ | | MAE ($\rho = 10\,\%$) | | | $\beta$ hit (%) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| History | FA | Save | FA | Save | FA | Save | all | first | later | lower | upper |
| Sparse visited-history **(current)** | 0.10 | 79.3 | 0.00 | 86.4 | 0.03 | 90.1 | 0.0200 | 0.0232 | 0.0185 | 37.1 | 3.6 |
| True fully dense history | 0.13 | 79.0 | 0.00 | 86.1 | 0.03 | 89.8 | 0.0197 | 0.0238 | 0.0177 | 36.4 | 1.6 |

### C.5 Sparse Visited Trajectory against a Dense Integer History

DivShift fits the power law only to the sample sizes the evaluation has actually visited. An evaluation holding the images could instead fit every integer prefix $k = 2, \ldots, k_{\mathrm{cur}}$, re-formed after each acquisition. Table 9 reports the comparison across all three tolerances. At none of them does the dense history improve the decision. At $\rho = 10\,\%$ it is slightly worse on both axes ($0.133\,\%$ against $0.100\,\%$ false adequacy, $78.99\,\%$ against $79.33\,\%$ generation savings); at $15\,\%$ both are exact; at $20\,\%$ they tie on false adequacy and the dense history saves $0.24$ points less. Neither policy records a fit failure. The dense policy also fits far more points to reach that same decision: $31$ rather than $5$ at the first decision, and $95.4$ rather than $6.7$ on average thereafter.

The reason is that the added points are not independent observations. Consecutive nested prefixes of the same permutation share all but one image, so $\Delta_k$ and $\Delta_{k+1}$ are almost the same number: across evaluation instances the adjacent-prefix correlation is $0.988$ on average and $0.995$ at the median, identically on all five panels. Adding every intermediate sample size therefore re-weights the least-squares objective toward the most densely sampled part of the trajectory without adding information about where it is going, and it costs $5.9\times$ the forecast-fit time at the first decision and $10.1\times$ at later ones. The sample sizes the dense policy needs were recomputed from the sealed per-prompt distance matrices with the sealed functional and the same frozen permutation replays — **no new image was generated and no embedding was recomputed** — and seven already-cached sample sizes recomputed alongside them matched bitwise in all 35 checks.

## D Reference-Size Sensitivity

Every result in this paper is defined against a finite reference size $K$, and the main experiments use $K = 512$. This appendix asks what changes when the reference size is doubled. **Neither value is treated as a population limit**: $1024$ is the deepest sample size for which a frozen cache exists, and it enters only as a second finite reference against which the first can be checked.

**Setup.** The extension exists for one panel: the guidance comparisons CFG $3/6/9$ on the $64$ COCO captions, in all three representations. Every sample size is a nested prefix of a single $1024$-image permutation per (prompt, replay), so $k = 512$ is an exact subset of $k = 1024$ and nothing below differs because of a redraw; this was verified by re-deriving the permutation and reproducing $54$ cached scores bitwise, and the extended cache is bitwise equal to the independently frozen $1024$-image tensor at all eight doubling sample sizes. The $K = 1024$ cache stores no sample size below $8$, so wherever the method is replayed at $K = 1024$ its initial trajectory is the available prefix $\{8, 16, 32\}$, and **the matched** $K = 512$ **control uses the same** $\{8, 16, 32\}$ **initial trajectory**. Those two rows are therefore comparable with each other but not with §5.2, which uses the full five-point initial trajectory.

**The comparison itself barely moves.** The top half of Table 10 measures how far the Vendi log-ratio moves between the two reference sizes. The median $|\Delta_{1024} - \Delta_{512}|$ is $0.0114$ and the maximum $0.0274$, that is at most $28.8\,\%$ of $\varepsilon_{10\%} = 0.0953$; **the sign of the comparison is preserved on all** $9$ **pairs** and no ratio crosses one. The matched pairwise control, a functional whose expected value does not move with sample size, moves by a median of $4 \times 10^{-5}$ under the identical procedure — $0.25\,\%$ of the same tolerance — so what is being measured is a property of the spectral functional

Table 10: **Reference-size validation,** $K = 512 \to 1024$**.** CFG 3/6/9 $\times$ COCO64, three representations. Top: how far the effect moves between the two reference sizes, as a fraction of each tolerance. Bottom: the final method replayed with each reference size. The frozen $K = 1024$ extension caches no sample size below 8, so the observed trajectory is initialized with the available prefix $\{8, 16, 32\}$ and the same prefix is used for the matched $K = 512$ control, making the two rows comparable.

| Movement $\lvert\Delta_{1024} - \Delta_{512}\rvert$ | median | max | as % of $\varepsilon_{10\%}$ (max) |
|---|---|---|---|
| Vendi (spectral) | 0.0114 | 0.0274 | 28.8% |
| Pairwise (fixed-target control) | 0.00004 | 0.00024 | 0.25% |
| Sign of the effect preserved | | | 9 / 9 pairs |
| **Final method replay** | **FA (%)** | **Save (%)** | **mean $k_{\text{stop}}$** |
| $K = 512$, $\rho = 10\,\%$ | 0.33 | 69.8 | 154 |
| $K = 512$, $\rho = 15\,\%$ | 0.50 | 81.6 | 94 |
| $K = 512$, $\rho = 20\,\%$ | 0.50 | 87.8 | 62 |
| $K = 1024$, $\rho = 10\,\%$ | 0.17 | 72.7 | 279 |
| $K = 1024$, $\rho = 15\,\%$ | 0.00 | 83.6 | 168 |
| $K = 1024$, $\rho = 20\,\%$ | 0.33 | 89.6 | 106 |

Table 11: **Fixed sample sizes against DivShift.** FA is the percentage of reported evaluations whose effect still moves by more than $\varepsilon = \log(1+\rho)$ between the stopping sample size and the reference size $K = 512$; Save is the mean generation savings, $100\,(1 - k_{\text{stop}}/K)$. Macro is the unweighted mean of the five panels; 600 instances per configuration panel and 3,600 for the architecture panel. DivShift fits all visited sample sizes accumulated up to the current decision, starting from the initial trajectory $\{2, 4, 8, 16, 32\}$. The fixed rows evaluate every comparison at one sample size and are scored by the identical criterion.

| | | Configuration panels | | | | | | | | Architecture | | Macro | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CFG · COCO | | CFG · Parti | | Intervention · COCO | | Intervention · Parti | | 4 arch. | | 5 panels | |
| $\rho$ | Reporting rule | FA↓ | Save↑ | FA↓ | Save↑ | FA↓ | Save↑ | FA↓ | Save↑ | FA↓ | Save↑ | FA↓ | Save↑ |
| 10% | Fixed $k = 32$ | 100.00 | 93.8 | 99.67 | 93.8 | 0.50 | 93.8 | 0.00 | 93.8 | 61.08 | 93.8 | 52.25 | 93.8 |
| | Fixed $k = 64$ | 71.33 | 87.5 | 66.83 | 87.5 | 0.00 | 87.5 | 0.00 | 87.5 | 46.36 | 87.5 | 36.91 | 87.5 |
| | Fixed $k = 128$ | 16.50 | 75.0 | 5.50 | 75.0 | 0.00 | 75.0 | 0.00 | 75.0 | 24.61 | 75.0 | 9.32 | 75.0 |
| | Fixed $k = 256$ | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 |
| | **DivShift** | 0.00 | 69.4 | 0.00 | 69.0 | 0.50 | 92.2 | 0.00 | 91.2 | 0.00 | 74.9 | 0.10 | 79.3 |
| 15% | Fixed $k = 32$ | 75.17 | 93.8 | 67.33 | 93.8 | 0.00 | 93.8 | 0.00 | 93.8 | 49.06 | 93.8 | 38.31 | 93.8 |
| | Fixed $k = 64$ | 34.33 | 87.5 | 22.50 | 87.5 | 0.00 | 87.5 | 0.00 | 87.5 | 29.33 | 87.5 | 17.23 | 87.5 |
| | Fixed $k = 128$ | 0.00 | 75.0 | 0.00 | 75.0 | 0.00 | 75.0 | 0.00 | 75.0 | 0.86 | 75.0 | 0.17 | 75.0 |
| | Fixed $k = 256$ | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 |
| | **DivShift** | 0.00 | 81.4 | 0.00 | 81.0 | 0.00 | 93.7 | 0.00 | 93.4 | 0.00 | 82.7 | 0.00 | 86.4 |
| 20% | Fixed $k = 32$ | 55.00 | 93.8 | 34.00 | 93.8 | 0.00 | 93.8 | 0.00 | 93.8 | 38.50 | 93.8 | 25.50 | 93.8 |
| | Fixed $k = 64$ | 0.00 | 87.5 | 0.00 | 87.5 | 0.00 | 87.5 | 0.00 | 87.5 | 21.47 | 87.5 | 4.29 | 87.5 |
| | Fixed $k = 128$ | 0.00 | 75.0 | 0.00 | 75.0 | 0.00 | 75.0 | 0.00 | 75.0 | 0.00 | 75.0 | 0.00 | 75.0 |
| | Fixed $k = 256$ | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 | 0.00 | 50.0 |
| | **DivShift** | 0.17 | 88.0 | 0.00 | 87.7 | 0.00 | 93.8 | 0.00 | 93.7 | 0.00 | 87.2 | 0.03 | 90.1 |

and not of the cache or the extension. No conclusion a reader draws from a comparison reported at $K = 512$ is reversed by doubling the reference size.

**The method behaves consistently at both reference sizes.** The lower half replays the complete sequential procedure with each $K$, under the matched initial trajectory above. Doubling the reference size roughly doubles the mean stopping sample size ($154 \to 279$ at $\rho = 10\,\%$, $94 \to 168$ at $15\,\%$, $62 \to 106$ at $20\,\%$), which is what a criterion defined against $K$ should do, while generation savings *rise* — from $69.8\,\%$ to $72.7\,\%$ at $\rho = 10\,\%$ — because the denominator grows faster than the stopping sample size. False adequacy stays at or below $0.50\,\%$ in every row and is not systematically worse at the larger reference size. The evidence covers one comparison family on one prompt bank, the only panel for which a 1024-image cache exists.

# E FULL FIXED-$k$ AGAINST DIVSHIFT RESULTS

Table 11 gives the panel-level values behind Figure 5 and behind the macro-average numbers quoted in §5.2. Each fixed row evaluates every comparison at one sample size and is scored by the identical criterion as DivShift, so the rows are directly comparable; the four fixed sample sizes shown are the ones marked on the top axis of Figure 5, and the full frontier in that figure is the same computation over every cached sample size.

Reading the table by column is what makes the fixed choice a two-sided problem rather than a tuning problem. At $\rho = 10\,\%$ a fixed $k = 64$ is falsely adequate in $71.3\,\%$ and $66.8\,\%$ of evaluations on the two guidance panels and $46.4\,\%$ on the architecture panel, while on the two intervention panels the same $k = 64$ is already exact, so every image beyond it is spent for nothing. The sample size that repairs the worst panel over-spends on the best by a factor of four. DIVSHIFT is not uniformly better than every fixed sample size on every panel — a fixed $k = 256$ has zero false adequacy everywhere — but it reaches $0.10\,\%$ false adequacy at $79.3\,\%$ generation savings, where $k = 256$ saves $50.0\,\%$ and $k = 128$ saves $75.0\,\%$ at $9.32\,\%$ false adequacy.

# F EXPERIMENTAL AND REPLAY PROTOCOL

This appendix records what was generated, how the scores are computed, what one evaluation instance is, and how a sequential rule is replayed and scored. It is the reference for §5.1 and is restricted to the experiments reported there.

## F.1 PANELS

Five evaluation panels are reported, and they are never pooled.

- **Guidance, two panels.** SD 1.4 (Rombach et al., 2022) at classifier-free guidance scales 3, 6 and 9 (Ho & Salimans, 2022), giving three pairwise comparisons, evaluated separately on 64 COCO captions (Lin et al., 2014) and on 64 PartiPrompts (Yu et al., 2022).
- **Intervention, two panels.** The unmodified SD 1.4 baseline, CADS (Sadat et al., 2024) and CNO (Kim et al., 2026) on the same backbone and checkpoint, so that only the sampling intervention differs, again three pairwise comparisons, on the same two prompt banks of 64.
- **Architectures, one panel.** SD 1.4, SDXL-LCM (Podell et al., 2024; Luo et al., 2023), FLUX.1-schnell (Black Forest Labs, 2024) and SD3 (Esser et al., 2024) on 240 constructed concept–attribute prompts (Rassin et al., 2024), giving all six unordered pairs.

A macro-average over panels is always their **unweighted mean**, so the architecture panel does not dominate the two-panel families by instance count.

## F.2 SCORES

Every image is embedded with DINOv2 (Oquab et al., 2024), OpenCLIP (Cherti et al., 2023) and DreamSim (Fu et al., 2023), and the three representations are reported separately; they are computed on the same images and are correlated conditions, not independent replications. For each prompt and configuration we compute the standard Vendi score, Eq. 1, from the eigenvalues of the normalized similarity matrix $K^{(k)}/k$ with unit trace, evaluated in the closed form $\exp(-\sum_i \lambda_i \log \lambda_i)$ with an eigenvalue tolerance of $10^{-12}$. The configuration average is formed over prompts *first* and the log-ratio second; per-prompt ratios are a different estimand and are used nowhere.

The DreamSim artifact is a complete pairwise distance matrix whose producing code computes $D = 1 - \tilde{E}\tilde{E}^\top$ on row-$\ell_2$-normalized embeddings, so $K = 1 - D$ is exactly the normalized Gram matrix and the spectrum is recoverable in closed form without loading a checkpoint. The reconstruction agrees with the canonical DreamSim forward pass to a maximum absolute error of $2.384 \times 10^{-7}$. DreamSim distances are stored at half precision, so its small eigenvalues are resolved less precisely than DINOv2's and OpenCLIP's, whose half-precision embeddings are squared into a double-precision Gram.

### F.3 Sample Order and Generation

Each prompt has its own frozen list of seeds, drawn without replacement from that prompt's private block using `SeedSequence` keyed by the SHA-256 of a master string and the prompt id, then permuted. Disjoint blocks make a cross-prompt collision impossible by construction, and Python's built-in `hash()` is never used because it is not stable across runs. Within a prompt the compared configurations use the *identical* list, $S_{p,A} = S_{p,B}$, which preserves common random numbers so that a comparison stays a paired contrast; across prompts the lists differ.

**Sample sizes are nested.** The images available at $k$ are the first $k$ entries of that prompt's list, so they are an exact prefix of those available at any larger sample size and an evaluation that grows from $k$ to $2k$ reuses every image it already has. Generation settings were reused verbatim across configurations — checkpoint, revision, pipeline, scheduler, step count, guidance scale, resolution, dtypes, library versions and the 8-image generation chunk — and the work queue is keyed by prompt with each block carrying every configuration, so no device is confounded with a configuration.

### F.4 Replay Construction and the Unit of Evaluation

Every sequential result in this paper is a *replay*: the full image pools were generated first, and each reporting rule is then re-run using only the images that would have been available to it at each of its own decisions. No rule ever reads a sample size deeper than the one it has reached, and the reference size $K$ enters only after the fact, when a decision is scored. Generation savings are therefore measured retrospectively while every decision is made prospectively.

**Permutation replays.** Within each prompt the seed pool is fixed, and 200 random permutations of it are drawn once and frozen. A replay evaluates the rule along one permutation, so $k$ always indexes a nested prefix of that permutation. The same permutation is used for both configurations of a comparison within a prompt, so a comparison remains paired at every sample size. The 200 replays are realizations of the sample order, not independent experiments: they share the same images and the same prompts, so counts over replays are descriptive and are never converted into $p$-values.

**What one evaluation instance is.** The unit differs between the two panel types, and the difference is a property of the evaluation rather than of the rule.

- **Configuration panels** (guidance and intervention, each on two prompt banks) compare three configurations, hence three pairs, as one study. One instance is a (representation $\times$ replay) pair and the rule reports only when *all three* pairs satisfy the criterion at the same sample size, giving $3 \times 200 = 600$ instances per panel.
- **The architecture panel** scores its six pairs separately, so one instance is a (pair $\times$ representation $\times$ replay) triple, giving $6 \times 3 \times 200 = 3{,}600$ instances.

Summed over the five panels this is 6,000 evaluation instances at each tolerance.

**Scoring a decision.** A decision to report at sample size $k$ is *falsely adequate* when the realized remaining change exceeds the declared tolerance, $R_{k\to K} = |\Delta_K - \Delta_k| > \varepsilon_\rho$, evaluated on the same replay. For a configuration-panel instance the criterion is applied to the worst of the cell's three pairs, so an instance counts as falsely adequate if any of its pairs has moved too far. False adequacy is the percentage of instances that are falsely adequate; an evaluation that reaches $K$ without stopping is charged the full reference size when generation savings are computed and cannot be falsely adequate.

**Sample-size grids.** Two grids appear in the paper and they are not the same. The sequential rules choose among the cached sample sizes $\mathcal{G} = \{8, 16, 24, \ldots, 504\}$, the multiples of eight below $K$. The descriptive quantity $k^\star_\rho(K)$ of §3.3 and Figure 3 is instead evaluated on a denser 188-point grid that includes every integer from 8 to 128, and its unit is one (representation $\times$ replay $\times$ pair). A value of $k^\star$ read off the dense grid is therefore not directly comparable with a stopping sample size chosen from $\mathcal{G}$, and the two are never placed in the same table.

**Resampling.** Unless stated otherwise, intervals and envelopes resample **prompts**, $B = 1000$ draws, with one index vector per draw shared across every configuration, sample size and representation in the cell, so that the paired comparison and the nested trajectory both survive (Appendix B.3).