VISUALIZING DISTRIBUTION COVERAGE IN GENERATIVE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion distillation is widely adopted to accelerate sampling, and the resulting few-step models are broadly believed to match or even surpass their multi-step teachers in generation. However, standard evaluations such as GenEval2 typically draw only one sample per prompt, so improved scores may fail to reveal losses in distribution coverage. We therefore revisit whether distilled models truly match their teachers beyond single-draw performance using **pass@k**, which measures the probability that at least one of k independent samples satisfies a quality criterion. At k=1, pass@k reduces to standard single-draw evaluation. As k grows, the curve reveals whether additional draws find genuinely different successes or merely revisit the same modes, directly exposing how broadly a model covers the space of valid outputs. We first show that classifier-free guidance (CFG), whose quality–coverage tradeoff is well established, is the clearest case: higher guidance improves pass@1, but its advantage shrinks and reverses at larger k. Applying pass@k to few-step distilled models, we find the same tradeoff splits along training objectives: distribution-matching objectives concentrate the student's output distribution, boosting early-hit rates while eroding large-budget coverage, whereas consistency and trajectory-based objectives better preserve the teacher's coverage even at large k. We further show that this tradeoff extends to few-step causal video generation. Our findings reveal a previously overlooked cost of diffusion distillation: across both image and video generation, the choice of training objective fundamentally determines whether a few-step model inherits its teacher's distribution coverage or trades it away for single-draw quality.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021a;b) generate high-quality images and videos but require iterative denoising, making sampling slow and expensive. Distillation addresses this by training a student model to reproduce the teacher's output distribution in fewer steps, reducing inference cost by an order of magnitude or more while retaining the teacher's pretrained knowledge (Luo et al., 2023; Song et al., 2023). The resulting speedups, often 10–50$\times$ fewer function evaluations, have made distilled models the default choice for latency-constrained diffusion deployment (Chen et al., 2025; Sauer et al., 2024; Zhou et al., 2025; Yin et al., 2024).

The resulting few-step models are broadly believed to match or even surpass their multi-step teachers in generation quality. Indeed, on compositional benchmarks such as GenEval (Ghosh et al., 2023) and GenEval2 (Kamath et al., 2025), distilled models routinely match or exceed their teachers' scores (Sauer et al., 2024; Chen et al., 2025). However, these benchmarks are typically evaluated with only one sample per prompt, so a model can score better simply by concentrating probability on a narrower set of high-scoring outputs, even as its coverage of other valid modes declines. In practice, distilled models are often observed to produce less varied outputs than their teachers, yet no established metric quantifies this coverage gap or identifies its source.

Fig. 1 illustrates what these metrics miss. For two prompts, we draw 256 independent samples from a 50-step teacher and its few-step student. The student's pixel-wise standard deviation is visibly lower: it concentrates probability on fewer modes while the teacher spreads mass more broadly. Both models can produce individually convincing samples, yet the student's distribution is narrower. In practice, drawing multiple samples per prompt is the norm: a user may browse a batch to select a preferred composition, or a best-of-N pipeline may score candidates and return the strongest. The value of this extra budget depends on each draw exploring a different valid interpretation of the

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.



Fig. 1: 256 independent samples from Qwen-Image (50-step teacher) and its few-step student Qwen-Lightning for "a photo of a couch and a horse" (top) and "a photo of a pink car" (bottom). Each row shows the pixelwise mean of the 256 samples, the per-pixel standard deviation, and four samples. The student produces sharp individual samples but its lower standard deviation reveals a narrower output distribution: additional draws yield redundant variants rather than new valid compositions.

prompt. A narrow distribution undercuts this: additional samples converge toward similar outputs, and the marginal gain from each extra draw shrinks.

To quantify this effect, we adopt pass@k from code-generation evaluation (Chen et al., 2021). For each prompt, let $q_\theta(x)$ denote the probability that one sample satisfies a benchmark success criterion. Pass@1 reports the mean of $q_\theta(x)$; the pass@k curve reveals how that success probability is distributed across prompts as the sampling budget grows. When a model's advantage at small k shrinks or reverses at large k, it signals that the model has concentrated its distribution rather than expanded coverage.

We first validate pass@k in the setting of classifier-free guidance (CFG) (Ho & Salimans, 2022), where the tradeoff between stronger single-sample quality and reduced diversity is well established. This provides a natural sanity check: if pass@k reflects distribution coverage, it should recover the known contraction induced by strong guidance. Indeed, high CFG improves pass@1, but its advantage shrinks and eventually reverses at large k.

With this validation in hand, we use pass@k to study few-step distillation. Across several widely used model families, distilled students often outperform their teachers at small k but fall behind at large k, revealing a previously obscured loss of coverage. A controlled comparison under a shared teacher and evaluation protocol further traces this behavior to the distillation objective: distribution-matching losses concentrate probability on a narrower set of outputs, improving early-hit performance at the expense of large-budget coverage, whereas consistency- and trajectory-based losses better preserve the teacher's coverage. A complementary CLIP-space analysis shows that this contraction is accompanied by reduced within-prompt semantic diversity. Together, these results identify the distillation objective as a key factor in whether a few-step model preserves its teacher's distribution coverage or trades it for stronger single-draw performance.

We finally ask whether this finding extends beyond image generation. In few-step causal video generation (Huang et al., 2025; Zhao et al., 2026b), we observe the same objective-dependent pattern: distribution-matching methods exhibit diminishing gains as the sampling budget grows, while consistency-style alternatives better preserve coverage.

Our contributions can be summarized as follows: (1) we introduce pass@k as a diagnostic of distribution coverage for diffusion models, validating it on CFG where the coverage–quality tradeoff is independently known; (2) we identify an objective-level divide in diffusion distillation, showing that distribution-matching losses systematically contract coverage while consistency-based losses preserve it; and (3) we demonstrate that this divide holds across both image and video generation.

2 Related Work

Text-to-image benchmark evaluation. Recent text-to-image benchmarks evaluate increasingly fine-grained aspects of prompt following. GenEval (Ghosh et al., 2023) emphasizes object-centric compositional correctness, GenEval2 (Kamath et al., 2025) provides softer and more fine-grained atomic evaluation, DPG-Bench (Hu et al., 2024) targets dense prompt following with VQA-style checks, GenAI-Bench (Li et al., 2024) evaluates semantic text-to-image generation across diverse instruction categories, and T2I-CompBench (Huang et al., 2023) evaluates disentangled compositional skills such as attribute binding, spatial relations, and counting. These benchmarks are usually reported as single-sample or aggregated scores, which retain only the mean of the prompt-wise success probabilities. We use pass@k on top of these same verifiers to reveal how that success probability is distributed across prompts as the sampling budget grows.

Distribution coverage evaluation. Prior work has addressed the quality–coverage tradeoff through diversity metrics such as the Vendi score (Friedman & Dieng, 2022), precision–recall decompositions for generative models (Kynkäänniemi et al., 2019), and distributional distances such as FID (Heusel et al., 2017). These metrics provide useful ImageNet (Deng et al., 2009)-level diagnostics of realism and diversity, but they are not directly tied to whether a model can solve a particular text prompt under a benchmark verifier; moreover, representation-based distributional metrics can be directly optimized during training (Yang et al., 2026), limiting their reliability as independent evaluators. Our approach instead builds on prompt-level verifiers to measure coverage in terms of benchmark success, complemented by CLIP-space analysis of within-prompt semantic diversity.

3 Pass@k as a Lens for Distribution Coverage

Pass@k on Text-to-image Generation Benchmarks. We borrow the pass@k metric from code generation (Chen et al., 2021), where it measures the probability that at least one of k independently sampled completions passes a verifier. To extend it to text-to-image generation, we treat each benchmark instance as a problem and the benchmark's automatic evaluator—e.g., the object, attribute, and spatial checks of GenEval (Ghosh et al., 2023), the DSG-style VQA questions of DPG-Bench (Hu et al., 2024), or the semantic instruction-following evaluations of GenAI-Bench (Li et al., 2024)—as the scoring oracle. A benchmark instance may be a single prompt or a prompt–category pair, depending on the benchmark's native protocol; in all cases, one sample denotes one independently generated image or image set scored by the corresponding evaluator.

For each benchmark instance, we draw $n \geq k$ independent samples from the model, let c denote the number judged correct by the evaluator, and let $q_\theta(x)$ denote the population one-draw success probability of instance x. We report the unbiased estimator of (Chen et al., 2021) together with its population form:

$$\text{pass@}k = \mathbb{E}_{\text{instances}}\left[1 - \binom{n-c}{k} \Big/ \binom{n}{k}\right], \qquad \text{Pass}_\theta(k) = \mathbb{E}_x\left[1 - (1 - q_\theta(x))^k\right]. \tag{1}$$

The estimator equals the probability that a uniformly random size-k subset of the n samples contains at least one success, averaged over instances, and $1 - \text{Pass}_\theta(k)$ is the kth moment of the prompt-wise failure probability $1 - q_\theta(x)$: pass@1 retains only the mean success probability, whereas the pass@k curve provides a sequence of nonlinear summaries of how success probability is distributed across benchmark instances.

For continuous evaluators, best-of-k measures the expected maximum score among k independent generations. Given $n \geq k$ samples per prompt, we estimate it by averaging the maximum over all uniformly chosen size-k subsets, computed exactly from the ascending order statistics $s_{p,(1)} \leq \cdots \leq s_{p,(n)}$:

$$\widehat{B}(k) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{j=k}^{n} \frac{\binom{j-1}{k-1}}{\binom{n}{k}} s_{p,(j)}. \tag{2}$$

The weight is the probability that the jth ordered sample is the maximum of the subset. For a fixed pool of independent, identically distributed samples, this estimator is unbiased and reduces to Eq. 1 for binary scores. It probes access to the upper tail of continuous evaluator scores as the budget grows.

Experiment Setup. We use pass@k and best-of-k as unified diagnostics of distribution coverage. For each model and experimental condition, we generate n independent samples per prompt using the same prompt set, resolution, benchmark split, and evaluator. Unless otherwise noted, all models are sampled with their default `diffusers` (von Platen et al., 2022) inference pipelines and hyperparameters. We then compute the appropriate statistic for $k \in \{1, 2, 4, \ldots, n\}$, producing a coverage profile rather than a single score. The k=1 endpoint recovers standard single-sample performance, while the growth of the curve measures how much additional success becomes accessible at larger k.

All comparisons hold the sampling budget fixed across methods and differ only in the intervention being studied: the guidance scale for CFG experiments, or the student checkpoint for distillation experiments. The curves distinguish three budget-dependent behaviors. An improvement across the measured curve indicates greater access to successful generations at those budgets. A gain at k=1 that saturates or reverses indicates that the one-draw advantage does not persist with additional samples. Conversely, weaker one-draw performance followed by stronger large-k performance indicates lower early-hit reliability but greater measured benefit from additional sampling. Because automatic evaluators have nonzero false-positive rates that accumulate at large budgets, we base claims on crossovers at moderate budgets and verify in Appendix D that all reported comparisons survive an FP-immune variant of pass@k and a direct calibration of evaluator false-positive rates.

4 Pass@k Profiles Distribution Coverage under CFG

Background. Classifier-free guidance (Ho & Salimans, 2022) is one of the most widely used test-time techniques for improving prompt alignment in text-to-image diffusion models. During sampling, CFG combines the conditional and unconditional model predictions to amplify the direction associated with the text condition. In a common noise-prediction parameterization, the guided prediction is written as:

$$\epsilon_{\text{cfg}}(x_t, c) = \epsilon_\theta(x_t, \varnothing) + w\,(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing))\,, \tag{3}$$

where c is the text prompt, $\varnothing$ denotes the empty condition, and w is the guidance scale.

Larger guidance scales often improve prompt adherence and visual fidelity at the single-sample level. However, CFG also changes the sampling distribution by reweighting the reverse process toward samples that are more strongly favored by the condition. This motivates a coverage question: does guidance improve success broadly across prompts, or is its advantage concentrated at small sampling budgets?

Experimental Analysis. Fig. 2 presents results for SD3.5-Large on GenEval2. GenEval2 decomposes each prompt into atomic requirements and assigns each a VQA-based soft score; Soft-TIFA GM is the geometric mean of these scores and serves as the prompt-level correctness measure (Kamath et al., 2025). Panel (a) shows that CFG= 7 achieves the strongest best-of-1 score, but its advantage erodes as additional samples are drawn. CFG= 3.5 overtakes it at a small budget and remains strongest thereafter, while CFG= 1 eventually overtakes CFG= 7. Panel (b) explains this crossover through the distribution of per-prompt success rates. As guidance increases, probability mass moves

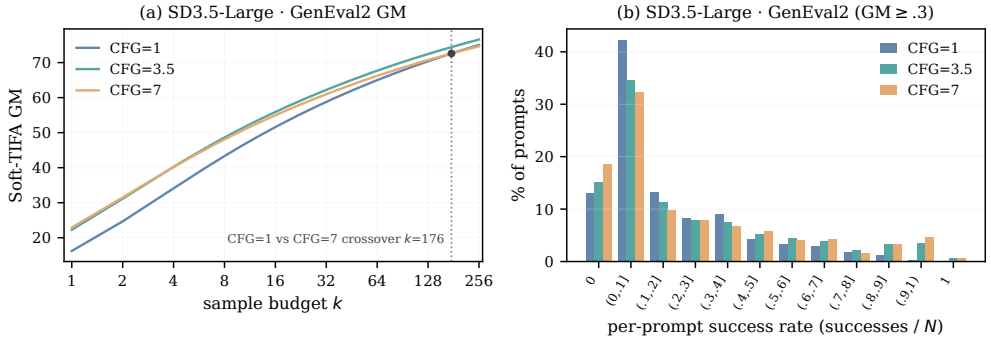

Fig. 2: **SD3.5-Large CFG sweep experiment on GenEval2.** (a) Best-of-k GenEval2 Soft-TIFA GM for CFG $\in \{1, 3.5, 7\}$. The dotted line marks where CFG= 1 overtakes CFG= 7. (b) Distribution of per-prompt success rates, where success is Soft-TIFA GM$\geq$ 0.3 over N=256 samples—a lenient cutoff, shared by all GenEval2 per-prompt diagnostics, that credits any mostly-correct sample (Appendix A.2).

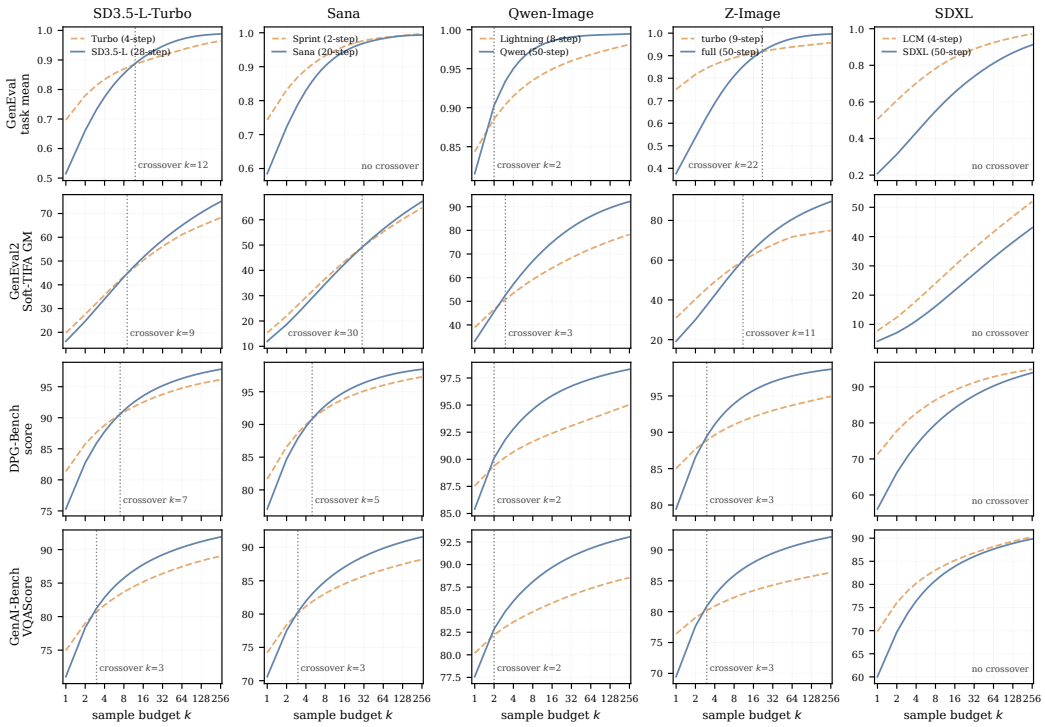

Fig. 3: **Teacher–student performance across sampling budgets.** Pass@k on GenEval and expected best-of-k on GenEval2, DPG-Bench, and GenAI-Bench for five distillation families (N=256 per prompt). Solid and dashed curves denote the teacher and student, respectively; dotted lines mark the first teacher–student crossover.

away from prompts with low-but-nonzero success rates and toward both extremes: more prompts are solved reliably, but more remain unsolved throughout the measured sample pool. Together, these results provide evidence of prompt-level coverage contraction under high CFG—stronger guidance improves early-hit success but yields smaller marginal gains from additional sampling—which we term benchmark-conditioned mode collapse. Two caveats delimit this claim: CFG serves here as a controlled knob for demonstrating that pass@k separates behaviors indistinguishable at k=1, not as evidence that guidance is harmful; and since extreme guidance also introduces perceptual artifacts that a VQA evaluator may reject, the curves measure benchmark-conditioned coverage, the quantity a verifier-based best-of-k pipeline consumes, with mechanism-level diversity analyzed directly in Section 5. The complete CFG sweep and corresponding per-prompt distributions are provided in Appendix A.

5 Distribution Coverage in Few-Step Distillation

Background. Diffusion distillation compresses the iterative reverse process of a multi-step teacher into a student that generates in one or a few denoising steps. Recent methods have substantially narrowed the fidelity gap to their teachers, making one-step synthesis increasingly practical (Yin et al., 2024; Chen et al., 2025; Wang et al., 2025; Geng et al., 2025; Song et al., 2023). Diffusion distillation can be divided into two broad paradigms base on training objectives: distribution matching and consistency distillation. Practical recipes may combine either principle with adversarial or regression losses (Chen et al., 2025; Yin et al., 2024). For example, Sana-Sprint combines consistency and latent adversarial distillation.

Distribution matching. Let $p_\theta(x \mid c)$ and $p_\phi(x \mid c)$ denote the conditional output distributions of the teacher and student. Distribution-matching distillation directly reduces a discrepancy between them, commonly expressed as

$$\mathcal{L}_{\mathrm{DM}}(\phi) = \mathbb{E}_c[D_{\mathrm{KL}}(p_\phi(\cdot \mid c) \,\|\, p_\theta(\cdot \mid c))]\,. \tag{4}$$

Because the output densities are implicit, this gradient is estimated using teacher and student score functions and is often supplemented by adversarial or regression losses (Yin et al., 2024; Sauer et al.,

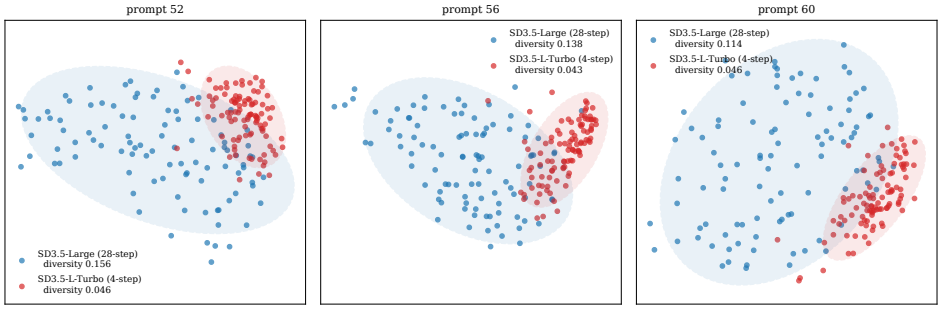

Fig. 4: **Distilled samples are more concentrated on matched-correct prompts.** Joint top-2 PCA projections of CLIP embeddings from SD3.5-Large and SD3.5-Large-Turbo on three matched-correct prompts (100 samples per model, per-model 2σ ellipses). Each panel reports the mean pairwise cosine distance in the original CLIP space.

2024). The student is thus trained to match the teacher distribution without reproducing a specific teacher sampling trajectory.

Consistency distillation. Consistency methods instead require the student to make the same endpoint prediction from different states on a teacher probability-flow ODE trajectory (Luo et al., 2023; Chen et al., 2025; Geng et al., 2025). Let $\hat{x}_s$ be obtained by advancing x_t from time t to $s < t$ with the teacher solver, and let f_ϕ map a noisy state to the trajectory endpoint. A standard objective has the form

$$\mathcal{L}_{\mathrm{CD}}(\phi, \bar{\phi}) = \mathbb{E}\big[\lambda(t)\, d\big(f_\phi(x_t, t, c), \mathrm{sg}\big[f_{\bar{\phi}}(\hat{x}_s, s, c)\big]\big)\big]\,, \tag{5}$$

where $\bar{\phi}$ is an exponential-moving-average target, sg stops gradients, and d is a regression distance. Once trained, the endpoint map can be evaluated once or composed for a small number of steps. We group our teacher-supervised MeanFlow variant under consistency-style distillation because its average-velocity identity corresponds to a differential consistency condition on the induced flow map (Geng et al., 2025).

Although one- and few-step students now produce sharp, high-scoring samples, visual fidelity does not guarantee that they preserve the teacher's conditional distribution. A student can concentrate probability on fewer valid outcomes, yielding more homogeneous samples even when its one-sample score improves. Compared with fidelity and sampling speed, this loss of prompt-conditioned diversity has received much less systematic evaluation. Moreover, it cannot be resolved from FID or a one-sample benchmark score alone (Ravishankar et al., 2026). We therefore ask whether the student preserves the teacher's distribution coverage as the sampling budget increases.

5.1 Coverage Contraction and Conditional Mode Collapse

Fig. 3 presents five distilled students with their corresponding multi-step baselines: SD3.5-Large-Turbo (Sauer et al., 2024) with SD3.5-Large (Esser et al., 2024), Sana-Sprint (Chen et al., 2025) with Sana (Xie et al., 2024), Qwen-Image-Lightning with Qwen-Image (Wu et al., 2025), LCM-SDXL (Luo et al., 2023) with SDXL (Rombach et al., 2022), and Z-Image-Turbo with Z-Image (Cai et al., 2025). All evaluations use $N{=}256$ samples per prompt. We report expected best-of-k on GenEval2, DPG-Bench, and GenAI-Bench, and pass@k on GenEval. Every student leads its teacher at $k{=}1$, but the teacher overtakes within budget in 16 of the 20 comparisons. Appendix B breaks these crossovers down by benchmark subscore and provides prompt-level coverage diagnostics.

The pass@k crossover shows that the distilled student's early advantage does not persist as the sampling budget grows. We next ask whether this contraction is accompanied by reduced diversity among generations for the same prompt. We illustrate this with SD3.5-Large and its Turbo student. Fig. 4 compares 100 samples from each model on three GenEval prompts that both solve perfectly, using CLIP ViT-L/14 embeddings (Radford et al., 2021). Despite identical correctness, the teacher's mean pairwise CLIP distance is 2.5–3.4 times larger, and the PCA plots show the Turbo samples clustering much more tightly.

LCM-SDXL(Luo et al., 2023) is the only student that remains ahead of its teacher across all four benchmarks. Its consistency objective learns to traverse the teacher's denoising trajectory in fewer steps, and the sustained gains show that this trajectory-shortening strategy can preserve coverage rather than inevitably induce mode collapse. Sana-Sprint, however, combines continuous-time

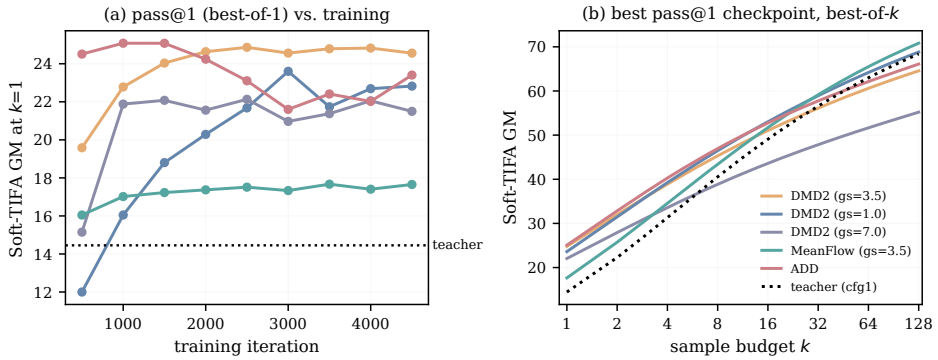

Fig. 5: **Distillation objectives shape coverage across training and sampling budgets.** *(a)* GenEval2 GM at k=1 across training checkpoints. *(b)* Best-of-k curves for the checkpoint from each run with the strongest k=1 performance.

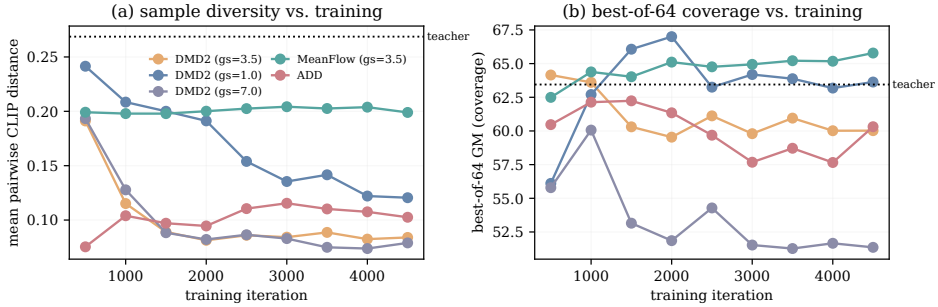

Fig. 6: **Coverage degradation is accompanied by diversity collapse.** Along the same training trajectories, *(a)* reports mean pairwise CLIP distance within each prompt and *(b)* reports GenEval2 GM best-of-64. DMD2 contracts in representation space as its coverage declines, whereas MeanFlow preserves diversity while improving coverage.

consistency distillation (sCM) with latent adversarial distillation (LADD) (Chen et al., 2025). Its teacher–student crossovers therefore reflect the behavior of a hybrid training objective and cannot be attributed to consistency distillation alone. Section 5.2 next examines why coverage behavior differs across training objectives in a controlled FLUX setting.

5.2 What in Distillation Causes the Collapse? A Controlled Loss Ablation

To understand why some students collapse more severely than others, we group distillation methods by their training objectives. Distribution-matching methods such as DMD2 optimize the final-sample distribution and often include adversarial or reverse-KL-like terms that are known to be mode-seeking (Wang et al., 2025). Consistency- and trajectory-based methods such as MeanFlow instead compress the teacher's sampling path by regressing its average velocity field. Although Section 5 shows that few-step distillation generally reduces large-k coverage, those comparisons also change the teacher, architecture, and training data. LCM-SDXL is a notable exception in Fig. 3, but it belongs to a different model family. We therefore use a controlled FLUX.1-dev study to isolate the role of the distillation objective.

Controlled design. We distill the same FLUX.1-dev teacher into the same four-step student architecture on GPIC 1M (Chandrasegaran et al., 2026) at 512×512 resolution, varying the distillation loss:

- **DMD2** (Yin et al., 2024) combines distribution matching with an adversarial term against real data, a recipe adopted by several recent large text-to-image models (Wu et al., 2025; Zhao et al., 2026a; Cai et al., 2025);
- **ADD** (Sauer et al., 2024) uses an adversarial objective against real data without the DMD distribution-matching term;
- **MeanFlow** (Geng et al., 2025) regresses the teacher's average velocity field without an adversarial or distribution-matching term.

We evaluate each training checkpoint with 128 GenEval2 samples per prompt. All students use an embedded guidance scale of 3.5; for DMD2, we additionally evaluate scales 1.0 and 7.0. The FLUX.1-dev teacher uses a 50-step Euler sampler without CFG.

Result 1: distribution-matching and adversarial objectives exhibit coverage collapse. Fig. 5 reports $k{=}1$ quality across training and the best-of-k curves of each run's strongest $k{=}1$ checkpoint. Fig. 6(b) tracks best-of-64 GM across training. DMD2 follows the same non-monotonic trajectory at each guidance scale: coverage improves early and then declines. ADD also peaks early and remains below the teacher, whereas MeanFlow improves throughout training and finishes above it. Higher guidance further amplifies the late-training decline within DMD2, but the contrast between DMD2 and MeanFlow remains under their shared standard guidance.

Separating objective and guidance effects. We use two complementary comparisons to distinguish the role of the distillation objective from that of guidance. First, at the same embedded guidance scale of 3.5, DMD2 and MeanFlow exhibit different training dynamics: DMD2 loses coverage and diversity during later training, whereas MeanFlow preserves diversity while improving coverage. These contrasting trajectories suggest that the distillation objective affects coverage and diversity even at a fixed guidance scale. Second, holding the DMD2 objective fixed, we vary guidance over $\{1.0, 3.5, 7.0\}$. Even at guidance 1.0, best-of-64 performance declines after its peak alongside decreasing CLIP diversity, while stronger guidance amplifies the late-training degradation. The guidance ablation further suggests that elevated guidance amplifies, but does not fully explain, the observed collapse.

Result 2: pass@k degradation is accompanied by reduced diversity. Fig. 6 measures within-prompt diversity along the same training trajectories. MeanFlow maintains nearly constant CLIP diversity while its coverage improves. DMD2 instead contracts sharply in representation space, and its later checkpoints remain substantially below MeanFlow and the teacher. The decline in DMD2's pass@k performance therefore coincides with a loss of generative diversity rather than only a change in the benchmark score. Although ADD's CLIP diversity remains relatively stable across training checkpoints, it stays substantially below the teacher's, indicating a persistent diversity deficit rather than progressive contraction during the observed training period.

Mechanism and scope. The results follow the training objective. Distribution-matching and adversarial objectives place pressure on the student to concentrate probability mass in high-density regions of the target distribution. MeanFlow instead learns a shortened teacher trajectory without directly reshaping its output distribution. In this controlled setting, the divergence-based objectives are associated with reduced diversity and coverage collapse, whereas the consistency-style objective preserves both. Final-checkpoint curves and per-prompt success profiles are provided in Appendix C.

6 Distribution Coverage in Video Forcing Models

Video forcing methods. With the insight from text-to-image diffusion distillation, we extend our evaluation to few-step video diffusion models. Video forcing methods accelerate autoregressive video generation by distilling a multi-step video diffusion process into a shorter sampler, while preserving the causal rollout structure needed for long video generation. We evaluate two representative families: self-forcing (Huang et al., 2025) and causal forcing (Zhao et al., 2026b), compared with their teacher Wan model (Wan et al., 2025). Self-forcing is a pure distribution-matching distillation approach. Causal forcing includes two kinds of few-step approaches, allowing us to separate a distribution-matching variant from a consistency-distillation variant.

Benchmark selection. We evaluate VBench-20 using a fixed set of 20 prompts for each of 16 VBench dimensions and $n = 64$ independently sampled videos per dimension prompt. For each dimension d, we apply Eq. 2 to the normalized per-video scores $s_{p,i,d}$ and average over its prompt set $\mathcal{P}_d$. Thus each point is the exact expected maximum over uniformly chosen size-k subsets of the 64 videos; $k = 1$ uses the mean of all 64 scores. We retain the source evaluation's zero scores for missing color records (7–16 of 1,280 records per method). We report every integer $k \in [1, 64]$ and three aggregates: quality averages seven dimension curves with weight 0.5 for dynamic degree and 1 otherwise; semantic equally averages nine dimension curves; overall equally averages quality and semantic. Selection occurs separately within each dimension.

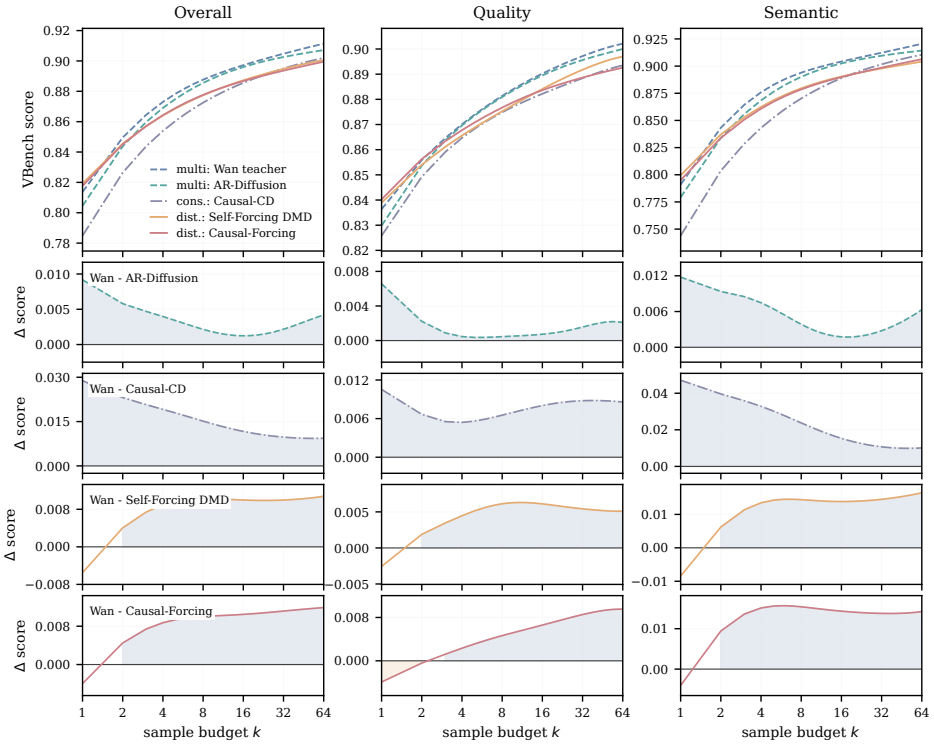

Fig. 7: **VBench-20 best-of-k curves for video forcing models.** Each curve uses the exact subset-averaged estimator in Eq. 2. Solid curves denote distribution-matching methods (Self-Forcing DMD and Causal-Forcing), dash-dot curves denote consistency distillation (Causal-CD), and dashed curves denote multi-step baselines. The four lower rows show $\Delta_m(k) = s_{\text{Wan}}(k) - s_m(k)$ for AR-Diffusion, Causal-CD, Self-Forcing DMD, and Causal-Forcing, respectively. Positive differences favor Wan.

Analysis. Both the quality and semantic best-of-k curves in Fig. 7 show signs of mode collapse in the distribution-matching models. Self-Forcing DMD and Causal-Forcing initially outperform Wan, but their advantage quickly reverses as the sampling budget grows, consistent with our controlled image study in Section 5.2.

Multi-step AR-Diffusion and trajectory-based consistency distillation show a different pattern. Their quality scores remain below the Wan teacher across the evaluated budgets, but their semantic scores tend to catch up as more samples are drawn. This suggests that semantic diversity may be better preserved in these approaches, even when visual quality lags behind. These results motivate further work on distillation methods that improve visual quality while preserving semantic diversity.

7 Conclusion

We use pass@k to probe whether few-step diffusion students truly inherit their teachers' distribution coverage. After validating the diagnostic on CFG, where the coverage–quality tradeoff is independently known, we show that the answer depends on the distillation objective: distribution-matching and adversarial losses concentrate the student's output distribution, improving single-draw performance at the cost of large-budget coverage, whereas consistency and trajectory-based losses better preserve the teacher's coverage. The same divide holds in few-step causal video generation, suggesting it reflects a fundamental property of the training objective rather than a domain-specific artifact. These results challenge the common assumption that distilled models are drop-in replacements for their teachers, and highlight that the choice of distillation loss has consequences for distribution coverage that single-draw metrics do not reveal.

Limitations. Pass@k is deliberately benchmark-relative: it characterizes access to successful generations under a specified evaluation criterion rather than every form of perceptual or semantic diversity. Its large sample pools also limit how many models and training configurations can be evaluated, particularly proprietary systems. We hope our findings motivate greater attention to diversity-preserving distillation methods that retain both sampling efficiency and distribution coverage.

Reproducibility Statement

We provide detailed experimental configurations including model architectures, training hyperparameters, and evaluation protocols in the main text and appendix. Our method builds on publicly available codebases and pretrained models. Code is available at `https://anonymous.4open.science/r/diffusion-pass-k-39C5/`.

AI use statement

Generative AI tools such as OpenAI Codex, were used to assist with drafting and language editing, manuscript organization, literature discovery and summarization, citation checking and formatting, and critical review of the manuscript, including feedback on the presentation and interpretation of previously obtained experimental results. The authors reviewed all AI-assisted text and suggestions and checked cited references against the original sources. The authors take responsibility for the final content of the manuscript, including text and artifacts produced with the aid of generative AI.

References

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Keshigeyan Chandrasegaran, Kyle Sargent, Suchir Agarwal, Michael Jang, Michael Poli, Juan Carlos Niebles, Justin Johnson, Jiajun Wu, and Li Fei-Fei. Gpic: A giant permissive image corpus for visual generation. *arXiv preprint arXiv:2605.30341*, 2026.

Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16185–16195, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75460–75482, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

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. *Advances in neural information processing systems*, 30, 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*, 2020.

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.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Genai-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024.

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.

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

Advaith Ravishankar, Serena Liu, Mingyang Wang, Todd Zhou, Jeffrey Zhou, Arnav Sharma, Ziling Hu, Léopold Das, Abdulaziz Sobirov, Faizaan Siddique, Freddy Yu, Seungjoo Baek, Yan Luo, and Mengyu Wang. Setting-matched and semantics-scaled benchmarking of one-step generative models against multistep diffusion and flow models. *arXiv preprint arXiv:2603.14186*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

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

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yifei Wang, Weimin Bai, Colin Zhang, Debing Zhang, Weijian Luo, and He Sun. Uni-Instruct: One-step diffusion model through unified diffusion divergence instruction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 57097–57132, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation Fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026a.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026b.

Mingyuan Zhou, Yi Gu, Huangjie Zheng, Liangchen Song, Guande He, Yizhe Zhang, Wenze Hu, and Yinfei Yang. Score distillation of flow matching models. *arXiv preprint arXiv:2509.25127*, 2025.

A Additional CFG Results

A.1 Full Guidance Sweep

Fig. 8 extends the selected comparison in Fig. 2 to all evaluated model–benchmark pairs and to CFG$\in \{1, 3.5, 7\}$. Stronger guidance usually improves performance at small sampling budgets, but the advantage narrows and often reverses as k increases. The clearest reversal occurs on GenAI-Bench (Fig. 8d). Moderate guidance is not uniformly harmful: for FLUX.1-dev on GenEval2 (Fig. 8f), CFG$= 3.5$ remains ahead throughout the measured range. This suggests that well-tuned guidance can improve one-sample performance without the same loss in large-budget gains observed under stronger guidance.

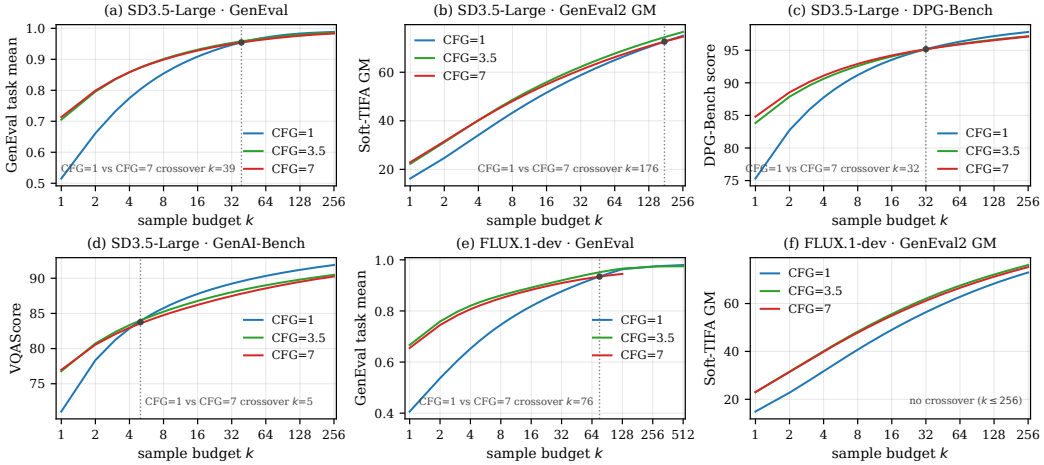

Fig. 8: **Full CFG sweep.** Pass@k and best-of-k curves for SD3.5-Large (a–d) and FLUX.1-dev (e–f) under CFG$\in \{1, 3.5, 7\}$. Dotted lines mark where CFG$= 1$ first matches or overtakes CFG$= 7$.

A.2 Per-Prompt Success Rates

Fig. 9 shows the per-prompt success rates underlying the curves. A sample counts as successful when it is correct on GenEval, reaches Soft-TIFA GM≥ 0.3 on GenEval2, reaches 0.9 on DPG-Bench, or reaches VQAScore≥ 0.9 on GenAI-Bench. The GenEval2 cutoff is deliberately lenient and is shared by every GenEval2 per-prompt diagnostic in the paper (Figs. 2 and 17) as well as by the pass@k re-analysis of Appendix D: a low cutoff asks whether the model *ever* produces a mostly-correct sample for a prompt, which is the coverage notion that pass@k targets, whereas a cutoff near 1 increasingly reflects single-sample polish—precisely the quantity that guidance improves—rather than coverage. No quantitative claim depends on this binarization: best-of-k curves are computed on the raw continuous scores (Eq. 2), and Appendix D reports how the crossovers shift under stricter cutoffs ($\tau = 0.7$ and 0.9). Higher guidance shifts probability mass out of the intermediate-success prompts and toward the extremes. This makes already tractable prompts more reliable, but leaves fewer intermediate prompts that can benefit from additional samples. Indeed, in five of the six settings in Fig. 9, the CFG$= 1$ baseline leaves fewer prompts in the zero-success bin than CFG$= 7$, including SD3.5-Large on GenEval2 and DPG-Bench and FLUX.1-dev on GenEval and GenEval2; stronger CFG therefore does not improve generation uniformly across prompts, but concentrates probability on prompts the model already solves and contracts the effective sampling distribution.

B Additional Distillation Results

This section provides the complete crossover statistics and per-family diagnostics behind the distillation results in the main paper.

B.1 Crossovers Across Families

Fig. 10 breaks the aggregate crossovers in Fig. 3 into benchmark-specific subscores. Each cell is the smallest integer k at which the teacher matches or exceeds the student; k=1 means that the teacher already leads, and NC denotes no crossover within N=256. Crossovers are widespread

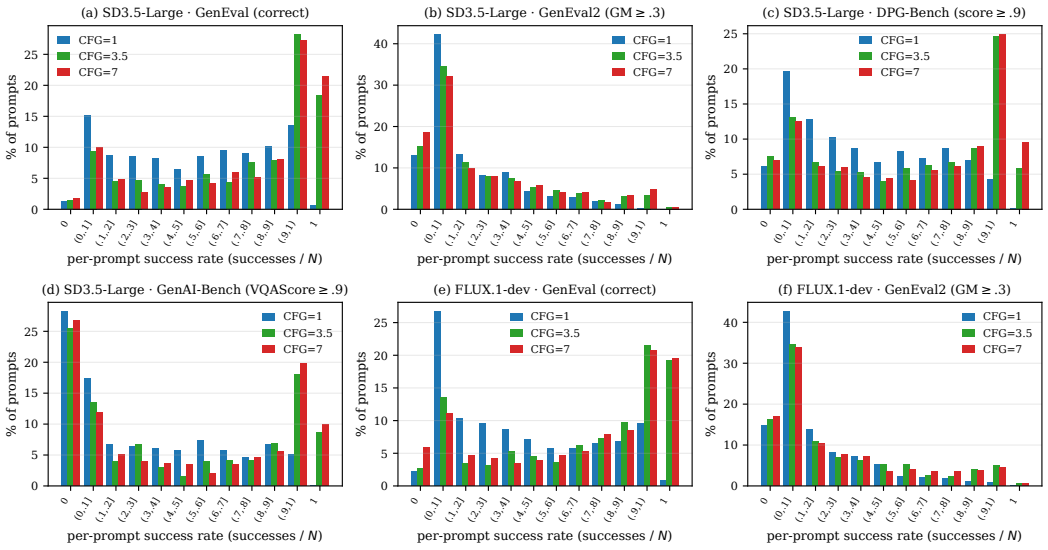

Fig. 9: **Per-prompt success-rate distributions for the CFG sweeps.** Panel order follows Fig. 8.

but family-dependent: they occur across most subscores for SD3.5-Turbo, Qwen-Image-Lightning, and Z-Image-Turbo, whereas LCM-SDXL remains ahead on every aggregate benchmark and is overtaken only on selected subscores, consistent with better coverage preservation under consistency distillation.

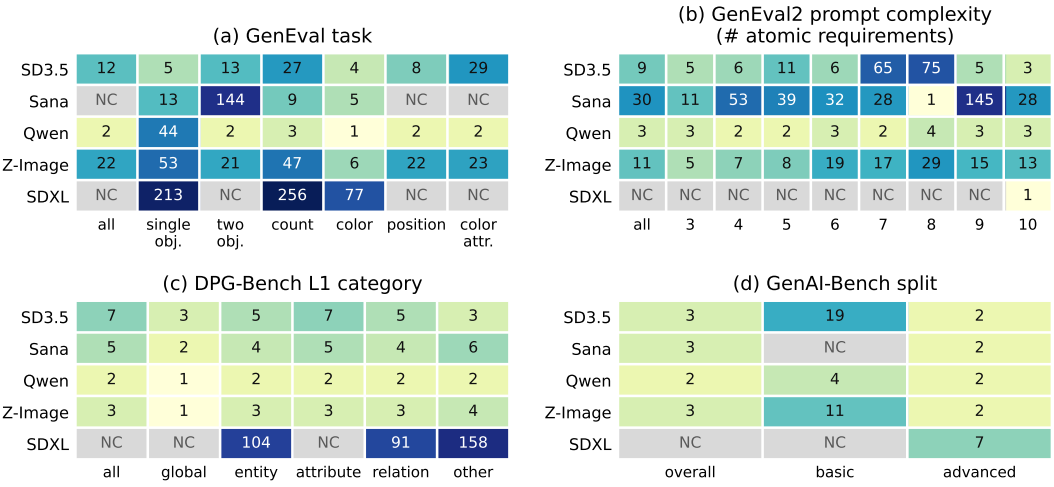

Fig. 10: **Teacher–student crossover budgets by benchmark subscore.** Rows are model families and columns are benchmark-specific tasks, atom counts, categories, or splits. Darker cells indicate later crossovers; gray cells marked NC do not cross within $N{=}256$.

B.2 Prompt-Level Coverage Diagnostics

Success redistribution across prompts. Figs. 11–15 show that the aggregate crossovers reflect a redistribution of success across prompts rather than a uniform change in correctness. For SD3.5-Turbo, Sana-Sprint, Qwen-Image-Lightning, and Z-Image-Turbo, distillation produces more student-only successes at small budgets, but these gains are concentrated on prompts the student solves repeatedly; as k grows, the teacher continues to recover prompts outside this set. The GenEval2 failure distributions show that better sample-level compositional accuracy can therefore coexist with weaker prompt coverage. LCM-SDXL behaves differently: its gains persist across prompts and sampling budgets, consistent with the absence of an aggregate crossover under consistency distillation.

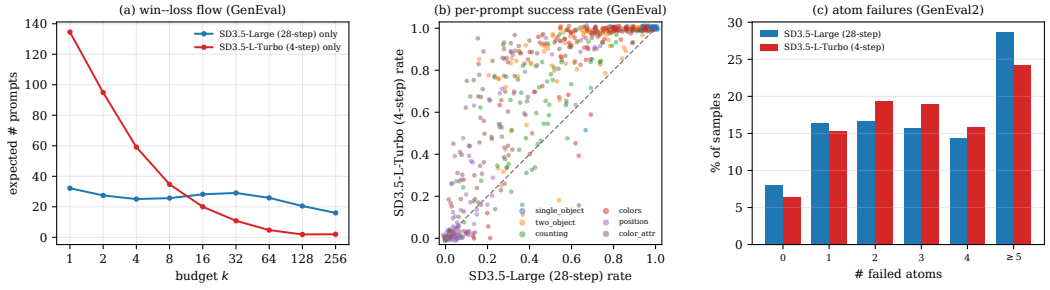

Fig. 11: **Per-prompt diagnostics for SD3.5-Large and Turbo.** Panels show exclusive wins over k, per-prompt success rates, and GenEval2 failed-atom counts.

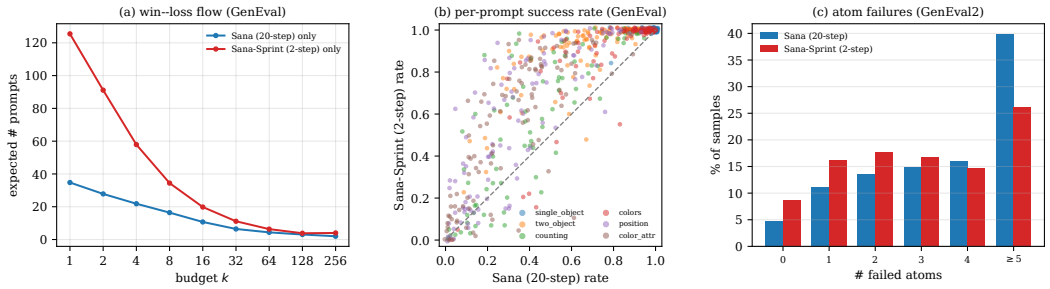

Fig. 12: **Per-prompt diagnostics for Sana and Sana-Sprint.** Panels follow Fig. 11.

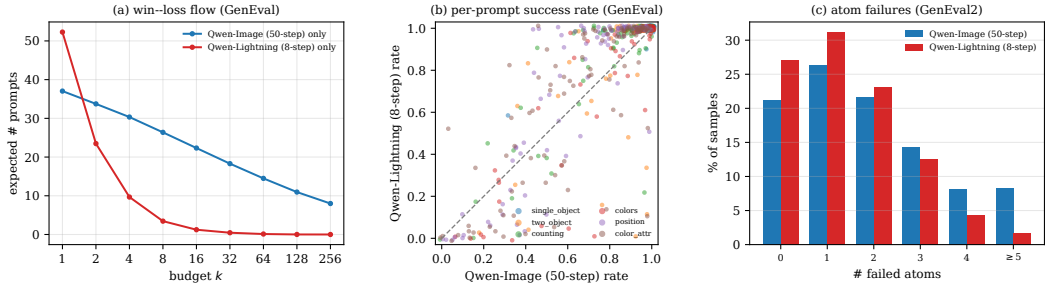

Fig. 13: **Per-prompt diagnostics for Qwen-Image and Lightning.** Panels follow Fig. 11.

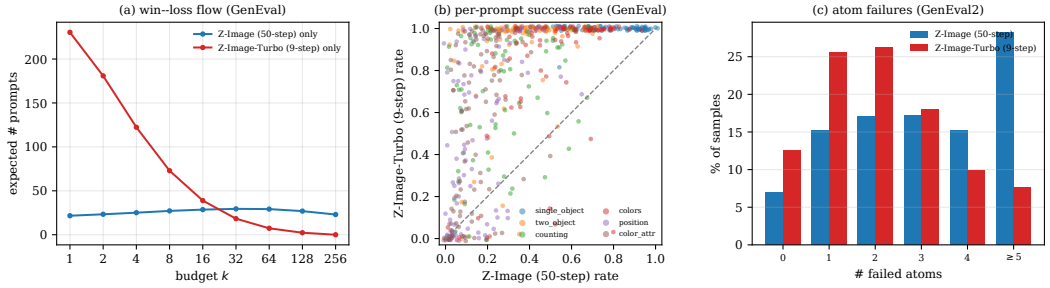

Fig. 14: **Per-prompt diagnostics for Z-Image and Z-Image-Turbo.** Panels follow Fig. 11.

C Additional Controlled FLUX Results

C.1 Coverage at the Final Checkpoint

Fig. 16 compares the best-of-k curves at iteration 4500. Under embedded guidance 3.5, DMD2 begins with strong small-budget performance, but the teacher overtakes as k increases. With embedded guidance 1, DMD2 shows a narrowing advantage as k increases, whereas at guidance 7 it falls below the teacher from k=8 onward and trails by more than 11 points at k=64. ADD follows the

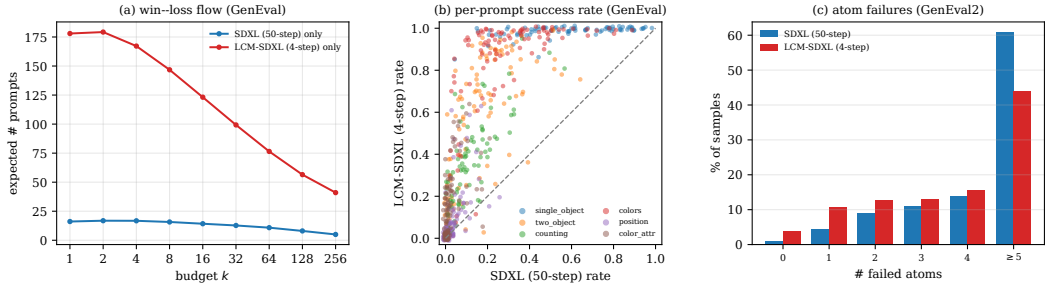

Fig. 15: **Per-prompt diagnostics for SDXL and LCM.** Panels follow Fig. 11; LCM retains its advantage across the measured range.

same pattern as DMD2 at guidance 3.5: it leads through k=16 and falls below the teacher at k=32 and 64. MeanFlow remains above the teacher across the measured range.

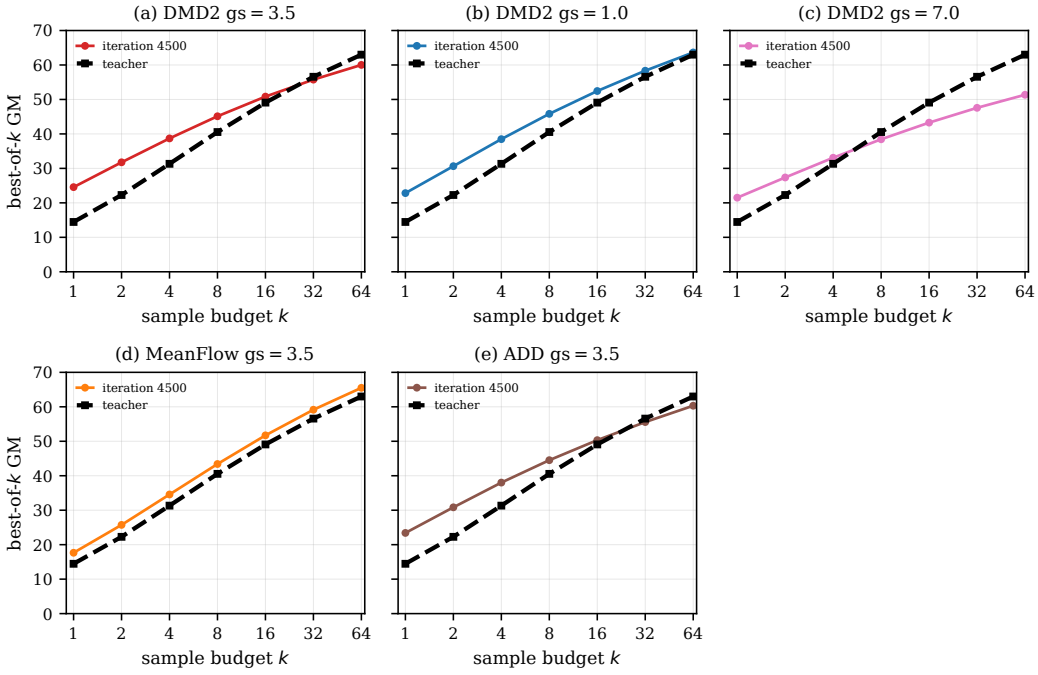

Fig. 16: **Best-of-k curves at iteration 4500.** The top row shows DMD2 with embedded guidance 3.5, 1, and 7; the bottom row shows MeanFlow and ADD with guidance 3.5. The dashed black curve is the shared 50-step FLUX.1-dev teacher.

C.2 Per-Prompt Success Rates

Fig. 17 shows the per-prompt success-rate distributions at the selected checkpoints, with success defined by the same cutoff (Soft-TIFA GM≥0.3) used for all GenEval2 per-prompt diagnostics (Appendix A.2). Relative to the teacher, the divergence-based students shift mass toward both extremes: DMD2 (guidance 3.5, 7) and ADD solve more prompts at high rates but also leave more prompts with zero successes, mirroring the redistribution observed under strong guidance in Fig. 2. MeanFlow instead tracks the teacher's distribution most closely and leaves the fewest prompts unsolved. The high-success mass accounts for the strong small-budget scores of DMD2 and ADD, whereas their enlarged zero-success bin limits the gains available at large k, consistent with the teacher overtaking them in Fig. 16.

D Evaluator False Positives and the Robustness of Pass@k

Unlike deterministic code-execution tests, detector-based evaluators have nonzero false-positive (FP) rates. If a sample of an unsolved prompt passes the evaluator with probability ε, the prompt's

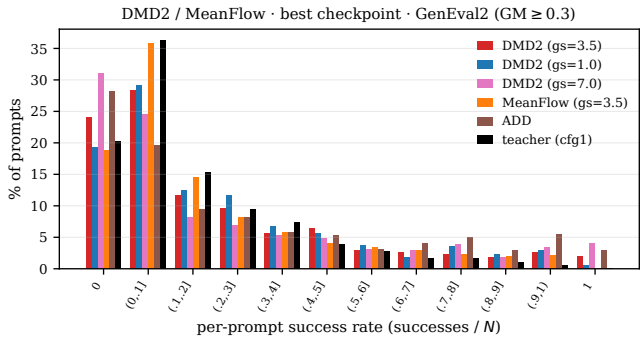

Fig. 17: **Per-prompt success rates in the controlled study.** A sample counts as a success when its GenEval2 geometric-mean score is at least 0.3, the cutoff shared with Figs. 2 and 9.

observed pass@k contribution saturates toward $1-(1-\varepsilon)^k$ rather than 0, so at large budgets FP mass accumulates and compresses differences between models at the extreme upper tail of the curves. Because every model on a benchmark is scored by the same evaluator, this inflation is shared: orderings and crossovers at moderate budgets are far more robust than absolute values at $k \to n$. This appendix quantifies both effects.

Two FP diagnostics. An FP-driven "solved" prompt has a signature: its pass@256 rests on one or two passing samples out of 256. We therefore recompute all comparisons under two variants. First, an *FP-immune statistic* that requires at least two successes, which a single false positive cannot satisfy:

$$\text{pass}_{\geq 2}@k = \mathbb{E}_{\text{instances}}\left[1 - \binom{n-c}{k} \Big/ \binom{n}{k} - c\binom{n-c}{k-1} \Big/ \binom{n}{k}\right]. \tag{6}$$

Second, for continuous evaluators we sweep the success threshold τ (success $=$ score $\geq \tau$), since FP scores concentrate near the threshold.

Tab. 1: **Teacher–student crossovers and endpoint gaps survive FP-immune re-analysis.** For each comparison: the budget at which the multi-step model first overtakes ($k_\times$) and the k=256 gap Δ_{256} (multi-step minus few-step), under the standard statistic and the FP-immune $\text{pass}_{\geq 2}@k$ of Eq. 6. Success thresholds: VQAScore ≥ 0.8 (GenAI-Bench), Soft-TIFA GM ≥ 0.3 (GenEval2); GenEval is binary. Requiring two successes shifts each crossover by at most one budget doubling and preserves (GenEval: enlarges) every gap; SDXL/LCM is the coverage counter-example reported in the main text and remains one.

		standard		require ≥ 2	
benchmark	comparison	$k_\times$	Δ_{256}	$k_\times$	Δ_{256}
GenAI-Bench	SD3.5-L / Turbo	8	+.063	16	+.059
	Sana / Sprint	8	+.063	16	+.063
	Qwen-Image / Lightning	4	+.083	8	+.072
	Z-Image / Turbo	8	+.108	16	+.101
	SDXL / LCM	256	+.004	—	−.009
	CFG 1 vs. 7 (SD3.5-L)	16	+.042	32	+.042
GenEval	Qwen-Image / Lightning	2	+.014	4	+.019
	Z-Image / Turbo	32	+.039	64	+.044
GenEval2	Z-Image / Turbo	8	+.156	16	+.191

Robustness of the reported comparisons. Table 1 summarizes the re-analysis over all teacher–student pairs with retained per-sample scores, plus the CFG sweep. Every crossover reported in the main text survives the FP-immune statistic, shifting by at most one budget doubling, and every endpoint gap is preserved; on GenEval the gaps grow, because the few-step students' tails rest on 1–2-hit prompts slightly more than their teachers'. Threshold sensitivity behaves as the FP model predicts: at $\tau = 0.7$ all crossovers occur earlier, and at the strictest cut ($\tau = 0.9$) the upper tail compresses, delaying the SD3.5-L and Sana crossovers toward the edge of the measured budget while Qwen-Image ($k_\times$=8) and Z-Image ($k_\times$=64) persist. The robust finding is therefore the ordering

itself: the few-step student leads at small budgets and its teacher overtakes at larger ones. The exact crossover budget depends on how strictly success is defined.

Direct calibration of ε: a null-pair test. We measure each evaluator's FP rate directly by scoring images against *mismatched* prompts: image directory i keeps prompt i's evaluation program, but contains images generated for prompt $j = (i + \delta) \bmod P$ (a fixed-point-free assignment; δ coprime with P). A pass under this pairing is counted as a false positive. This measures FPs on unrelated content; FPs on near-miss images are what the two diagnostics above target. We score 4 mismatched pairs per prompt using one model's images: 800×4 pairs for GenEval2 Soft-TIFA (Z-Image) and 527×4 for GenAI-Bench VQAScore (Qwen-Image). Because the null pairing makes image content irrelevant to the prompt, ε is a property of the evaluator alone and does not depend on which model generated the images.

Tab. 2: **Null-pair calibration.** Measured FP rate of each evaluator and the most it can add to each model's pass@256.

evaluator	model	unsolved $1-\hat{p}$	FP inflation	measured Δ_{256}
Soft-TIFA GM ≥ 0.3	Z-Image	.011	.001 [.005]	$+.156$
$(1/3200$ null passes)	Turbo	.168	.014 [.077]	
VQAScore ≥ 0.8	SD3.5-L	.082	0 [.036]	$+.063$
$(0/2108$ null passes)	Turbo	.144	0 [.063]	
	Sana	.089	0 [.039]	$+.063$
	Sprint	.152	0 [.067]	
	Qwen-Image	.061	0 [.027]	$+.083$
	Lightning	.144	0 [.063]	
	Z-Image	.070	0 [.031]	$+.108$
	Turbo	.178	0 [.078]	

The point-estimate FP rate is $\hat{\varepsilon} = 3.1 \times 10^{-4}$ for Soft-TIFA $(1/3200)$ and 0 for VQAScore $(0/2108)$; the corresponding pass@256 floors for a completely unsolved prompt are $f = 1 - (1 - \varepsilon)^{256} = .077$ and 0, respectively (.316 and .305 with the one-sided 95% Clopper–Pearson upper bound; for $0/2108$ this is the rule-of-three bound 1.4×10^{-3}). FP inflation in Table 2 is the most that FPs can add to a model's pass@256 $\hat{p}$: $(1 - \hat{p})f/(1 - f)$, where $1 - \hat{p}$ is the fraction of prompts the model leaves unsolved; bracketed values use the upper-bound f. Rows list the teachers and students of the coverage-loss families of Table 1; the gap is given on the teacher's row.

At the measured FP rates, false positives cannot account for the teacher–student gaps. A false positive can only make an unsolved prompt look solved, and teachers leave few prompts unsolved (the $1 - \hat{p}$ column). So even if every false positive went to the teacher and none to the student, the teacher's pass@256 would rise by at most its FP inflation. With the point estimates this is below the measured gap for every family, and with the 95% upper bounds in brackets it is too. Moreover, both models share the same evaluator, and the student has more unsolved prompts, so it gains more false solves than the teacher.