AUDITING CONTROL-LOSS ATTRIBUTION IN FEW-STEP DIFFUSION: CONFIGURATION EFFECTS AND MEASUREMENT LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image generation must balance sampling efficiency with faithful realization of requested objects, attributes, and spatial relations. Few-step diffusion reduces sampling cost, but interpreting its prompt-following performance requires separating changes in model weights and inference settings from the behavior of automatic evaluators. We investigate this attribution problem through paired model comparisons, a reference-model experiment crossing two step counts with two guidance settings, and fixed-image evaluator checks. In the 1,200-image reference-model experiment, reducing steps and reducing guidance each lower automatic prompt-validity rates, the fraction of images judged to satisfy all prompt constraints, at both tested levels of the other factor. The added many-step, low-guidance condition shows that this guidance response also occurs at the tested many-step setting. The fixed-image audit identifies label–explanation contradictions and inconsistent relation answers in 16 of 19 cases with definite answers when the subject and reference object are queried in both orders. These findings connect two complementary sources of evidence: paired experiments locate configuration-sensitive scores, and fixed-image checks expose how the evaluator interprets prompt constraints. They motivate an evaluation structure for inference calibration, selecting inference-time settings with pretrained weights fixed: specify supported configuration changes, check task-specific measurement behavior, and test configuration selection on held-out prompts under matched search budgets.

1 INTRODUCTION

Text-to-image generation turns natural-language requests into images, requiring not only recognizable objects but also the correct attributes and spatial arrangement (Ghosh et al., 2023; Huang et al., 2025). Iterative use adds a complementary objective: producing these images with low sampling cost. Few-step diffusion reduces the number of generation updates through approaches such as distillation and consistency learning (Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024). Faithfulness concerns whether the requested composition is realized. For example, an image can contain both a chair and a book while reversing the requested relation between them. When accelerated models differ in prompt-following scores, identifying the source of that difference is essential for deciding what to change.

Interpreting such performance differences requires answering two questions: what changed in the generator, and what did the evaluator measure? Distillation trains a student generator using information from a reference teacher, but deploying the student can change both learned weights and inference-time computation (Salimans & Ho, 2022; Meng et al., 2023). Teacher and student operating points, meaning inference-time configurations, may differ in step counts, guidance settings, and sampling-schedule rules (Meng et al., 2023; Sabour et al., 2024). Classifier-free guidance (CFG), for example, controls conditioning strength by combining predictions with and without the text condition (Ho & Salimans, 2022). A teacher/student endpoint comparison measures final-image scores at the two chosen configurations, combining these changes in one difference. Separately, a prompt-

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.

level score depends on how the evaluator recognizes objects, binds attributes, and resolves relations (Ghosh et al., 2023; Hu et al., 2023). An attribution based on that score therefore requires both a specified configuration change and evidence that the evaluator distinguishes the relevant constraints reliably (Cho et al., 2024; Thrush et al., 2022; Kamath et al., 2023). Our research question is how controlled configuration comparisons and task-specific evaluator checks can jointly clarify the interpretation of this difference.

To separate the configuration changes, we use a four-configuration path: A runs a many-step teacher; B runs that teacher with few steps; B′ also switches its guidance to a student-like setting; and C runs the few-step student pipeline. We then add D, the many-step teacher at the low-guidance setting, to obtain a teacher-only factorial, a design containing every combination of the two selected step counts and two guidance levels. This extra configuration estimates the guidance response at both step counts and exposes the interaction, or dependence of one factor's effect on the other. Pairing prompts and random seeds holds the textual requests and replicate identities consistent across conditions.

The resulting teacher experiment contains 100 prompts, three random replicates per prompt, and four configurations, totaling 1,200 images. It reveals two consistent patterns in the automatic judgments: step reduction lowers success rates at both guidance levels, and guidance reduction lowers success rates at both step counts (Figure 1). The added many-step, low-guidance condition locates the guidance-associated decrease at both tested step counts, enabling a direct comparison of conditional effects.

To interpret these score changes, we examine the underlying judgments: the saved responses reveal contradictions between the evaluator's structured labels and semantic explanations. A fixed-image follow-up asks atomic questions, each concerning one object, attribute, or relation constraint. Among 19 relation cases with definite answers in both subject–object orders, 16 fail reversal consistency: reversing the queried object order should reverse the reported spatial direction (Table 4). A separate object-localization follow-up distinguishes category matching from the geometric placement of detected boxes. Together, these fixed-image checks separate label consistency, relation consistency, and localization coverage, making the measurement behavior behind aggregate scores explicit.

Our contributions are: (1) a comparison framework that distinguishes deployment-system differences from teacher-conditional step and guidance responses; (2) a paired four-corner experiment that measures both conditional responses and their interaction on automatic prompt-validity scores; and (3) a fixed-image measurement audit that exposes inconsistencies relevant to interpreting those scores. These findings motivate an evaluation structure for inference calibration, the selection of inference-time parameters with pretrained weights fixed: define the supported actions, check the evaluation endpoint, and assess selection on held-out prompts excluded from tuning.

2 Related Work

Acceleration and inference configuration. Acceleration research changes either the learned predictor or its sampling procedure: progressive distillation reduces steps through teacher–student training, whereas denoising diffusion implicit models (DDIM) and DPM-Solver, a numerical solver for diffusion probabilistic models, change sampling computation (Salimans & Ho, 2022; Song et al., 2021; Lu et al., 2022). Guided distillation explicitly absorbs a conditional–unconditional prediction combination into the student (Meng et al., 2023); schedule optimization and guidance-interval selection instead adjust inference with a fixed denoiser (Sabour et al., 2024; Kynkäänniemi et al., 2024). These approaches optimize generation, while our teacher-only experiment measures step and guidance responses at both levels of the other factor under one evaluation protocol. This design distinguishes conditional configuration responses from the broader differences between pretrained teacher and student pipelines.

Compositional evaluation and measurement reliability. GenEval, an object-centric text-to-image benchmark, and T2I-CompBench, a compositional text-to-image benchmark, evaluate object-, attribute-, and relation-level requirements beyond aggregate similarity (Ghosh et al., 2023; Huang et al., 2025). Question-based faithfulness evaluation decomposes prompts into individual checks, while dependency-aware scene graphs account for prerequisites such as object presence before attribute judgments (Hu et al., 2023; Cho et al., 2024). Complementing generator evaluation,

Winoground and What'sUp test evaluators' sensitivity to compositional and spatial distinctions (Thrush et al., 2022; Kamath et al., 2023). Our audit connects these two evaluation targets within the few-step comparison: it checks saved labels against explanations, tests relation consistency under reversed object order, and examines localization coverage on fixed images. The contribution is protocol-specific evidence for interpreting configuration contrasts, rather than a new compositional benchmark or a general estimate of evaluator accuracy. Appendix E provides the extended discussion, including inference optimization and internal-representation probes.

3 Comparison Targets and Measurement

To specify what each contrast can identify, we distinguish generator settings from evaluator outputs and define the comparison path, teacher factorial, and paired statistical analysis.

3.1 Separate the Generator from Its Evaluator

Let $G(w, n, o, c; p, z)$ denote an image generated with denoising weights w, step count n, operating configuration o, other pipeline components c, prompt p, and random input z. The operating configuration includes the guidance formulation and the rule used to construct the sampling schedule. The realized timestep grid may change when n changes under that rule; it is not an independently held-fixed grid. The components c include text encoders, tokenizers, and the decoder. Making these arguments explicit prevents a checkpoint replacement from being silently treated as a denoising-weight intervention. This separation is consistent with work distinguishing diffusion design choices and diagnosing mismatches between training noise schedules and sampling implementations (Karras et al., 2022; Lin et al., 2024a).

Having specified the generator, we distinguish its intended semantic outcome $Y^*(G, p)$ from an evaluator output $\widehat{Y}_J(G, p)$ under a fixed protocol J. For the binary outcome considered here, $Y^* = 1$ means that all required prompt constraints are satisfied, whereas $\widehat{Y}_J = 1$ means that the evaluator judges them satisfied. The existing binary prompt-validity tables report the latter. The global judge, an automated evaluator returning one verdict for the complete prompt, is Qwen2.5-VL-7B-Instruct, a vision–language model (VLM) that processes an image and a textual evaluation request (Bai et al., 2025). The reported success rates are not estimates of evaluator accuracy, and their sampling intervals do not include systematic judgment error. Attribute binding means assigning each requested attribute to the correct object, such as a red chair rather than a red backpack; relation following means realizing the requested spatial arrangement between specified objects. The term "binding prompts" or "relation prompts" identifies a prompt group, not a separately identified binding or relation capability. Contrastive Language–Image Pre-training (CLIP) supplies a secondary text–image embedding alignment score (Radford et al., 2021), not an independent certificate of perceptual quality or constraint satisfaction. Composition-sensitive tests of contrastive models and question-based alternatives to embedding similarity motivate this distinction (Yuksekgonul et al., 2022; Lin et al., 2024b).

3.2 A Comparison Path Is Not a Complete Factorial

For compactness, write μ_X for the mean evaluator output in configuration X. We write $Y - X$ for the paired contrast $\mu_Y - \mu_X$, with a positive value indicating a higher automatic success rate in Y. The initial design is given in Table 1. The labels "teacher" and "student" name the roles in this comparison; they do not establish a direct training genealogy for every checkpoint pair. In particular, a family-level deployment comparison must not be interpreted as a weight-only intervention without component and execution matching. Here identification means that a contrast can be attributed to the specified intervention, rather than to other changes made alongside it.

The endpoint difference satisfies

$$\mu_C - \mu_A = (\mu_B - \mu_A) + (\mu_{B'} - \mu_B) + (\mu_C - \mu_{B'}). \tag{1}$$

The intermediate terms cancel algebraically; this telescoping identity is not evidence that the three terms are independent causes. An intervention contract is a record specifying the deliberately changed factor, the factors intended to remain fixed, and execution checks for those commitments.

Table 1: Comparison configurations. A/B/B′/C form a path. A/B/B′/D form the teacher-only step-by-guidance factorial. "Student-like" names a tested setting, not a proven cross-model equivalence.

Condition	Model role	Steps	Guidance setting
A	Teacher	Many	Teacher reference
B	Teacher	Few	Teacher reference
B′	Teacher	Few	Student-like
C	Student pipeline	Few	Student setting
D	Teacher	Many	Student-like

Under a verified intervention contract, $B - A$ measures step reduction at the reference teacher operating configuration, and $B' - B$ measures a guidance change at few steps. Their scope is conditional on the chosen levels, pipeline, and evaluation protocol. The final term is a student-pipeline contrast unless all other relevant changes have been excluded. Even a valid positive average weight contrast would not exclude harmed examples or the loss of particular internal channels, meaning layer-, head-, or timestep-specific computational pathways involved in conditioning.

3.3 Teacher Interaction and Statistical Scope

The teacher experiment crosses two step counts with two guidance settings (Table 1): A and B use reference guidance at many and few steps, respectively, while D and B′ use low guidance at those same step counts. Each configuration uses the same prompts and seed replicates. D completes the teacher factorial; C remains a separate student-pipeline comparison. Adding D permits an interaction on the success-probability difference scale, also called the risk-difference scale:

$$I = (\mu_{B'} - \mu_B) - (\mu_D - \mu_A). \tag{2}$$

It compares the guidance effect at few and many steps. It also measures the difference between the two allocations along A→B→B′ and A→D→B′. The total change is identical along both paths, but the assigned conditional effects can differ. A confidence interval containing zero does not establish negligible interaction; that would require an equivalence margin, a prespecified maximum effect considered practically negligible, and sufficient precision. The interaction is defined on the evaluator probability scale, not on every possible transformation of that scale.

The teacher audit pairs full prompt text and actual seed across conditions. It uses prompt-cluster bootstrap intervals: prompts are resampled with replacement, retaining all paired conditions and seed replicates together within each sampled prompt. Each prompt group has 50 clusters and 150 prompt–seed examples. Reported intervals are pointwise 95% intervals for individual contrasts, not simultaneous guarantees across all contrasts. Intervals for the separate A/B/B′/C pipeline comparison resampled individual prompt–seed examples and are separately labeled in the appendix. Sharing a seed is a pairing convention; it does not guarantee identical latent trajectories, the sequences of intermediate image representations updated during sampling, across different pipelines.

4 Controlled Configuration Comparisons

We first examine how automatic prompt-validity scores respond to step count and guidance within a fixed teacher pipeline. The four-corner experiment is the primary controlled comparison; teacher/student pipeline contrasts and teacher dose sweeps then provide context about cross-pipeline differences and configuration sensitivity. Throughout, the reported outcomes are evaluator judgments, whose semantic interpretation is examined in Section 5.

4.1 The Teacher-Only Four-Corner Experiment

To test whether the measured guidance response is specific to few-step sampling and whether its size depends on step count, we cross both factors within one teacher pipeline. We use Stable Diffusion 3.5 Medium (SD3.5-medium), based on the transformer architecture of Esser et al. (2024), with 28 or 4 steps and guidance values 4.5 or 0 passed to the application programming interface (API),

Table 2: Teacher four-corner experiment: automatic global prompt-validity counts, not verified semantic accuracy. Each group has 150 examples per condition.

Condition	Steps	API guidance	Binding prompts	Relation prompts
A	28	4.5	148/150	121/150
B	4	4.5	93/150	76/150
B′	4	0	38/150	23/150
D	28	0	106/150	69/150

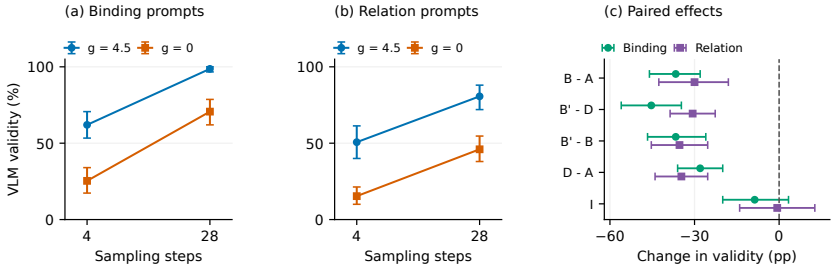


Figure 1: SD3.5 teacher responses under the recorded VLM protocol. (a,b) Automatic prompt-validity rates, with g denoting API guidance. Lines connect only the two tested step counts. (c) Paired effects in percentage points (pp): B−A and B′−D reduce steps at reference and low guidance; B′−B and D−A reduce guidance at few and many steps; $I = (B' - B) - (D - A)$ is their interaction. Error bars are pointwise 95% prompt-cluster bootstrap intervals from 10,000 resamples (50 prompts, three paired seeds per group). Rate and contrast intervals are computed separately. These are evaluator-label responses, not verified semantic accuracy; interaction intervals crossing zero do not establish additivity.

holding the recorded remaining settings fixed.[1] Here API guidance 0 disables the additional CFG branch under the tested implementation; text conditioning remains present. It should not be called unconditional generation. The run comprises 100 interleaved prompts, three seed bases, and four conditions at 1024×1024 resolution. Each seed base is an offset added to the prompt index to obtain a generation seed. The prompts are a study-specific template set with 50 attribute-binding and 50 relation requests, rather than the official test split of an external benchmark. Their object, color, and spatial-relation constraints follow the categories studied by the object-centric benchmark GenEval and the compositional text-to-image benchmark T2I-CompBench (Ghosh et al., 2023; Huang et al., 2025); these references motivate the categories, not the provenance of our generated prompts. The recorded maximum text sequence length is 256 tokens.

The execution validator reports 1,200 of 1,200 expected rows and 300 paired examples. The manifest is the per-image record of prompts, seeds, conditions, and generated-file identifiers. Separate inspection of it and the saved judgments reproduces the pairing keys, structured-response parsing, and request hashes, which are checksums identifying the recorded evaluation requests. These checks establish consistency of the supplied records, not independent verification of every image byte or pipeline component. Table 2 reports the observed global judgments. The measurement limitations established in Section 5 apply to all entries.

Both step reduction and guidance reduction decrease automatic success labels at the tested levels (Figure 1; Table 3). In particular, D−A shows a guidance-associated decrease at 28 steps as well as at four steps. The added corner therefore rules out the narrow description that the observed label decrease occurs only in the few-step condition. It does not establish the semantic magnitude of that decrease or rank the two interventions independently of their chosen doses.

The binding results illustrate why the added corner matters for attribution. Along A→B→B′, reducing steps and then guidance assigns approximately 36.67 percentage points of the observed decline to each change. Along A→D→B′, reducing guidance first assigns 28.00 points to guidance and 45.33 points to steps. Both paths connect the same endpoints and recover the same total decline,

[1]Exact model artifact: `https://huggingface.co/stabilityai/stable-diffusion-3.5-medium`. The architecture paper and this artifact's model card serve different provenance roles.

Table 3: Teacher contrasts in percentage points, with prompt-cluster 95% bootstrap intervals. These are contrasts of automatic labels. Interaction intervals do not establish additivity.

Contrast	Binding prompts	Relation prompts
B−A	-36.67 $[-46.00, -28.00]$	-30.00 $[-42.67, -18.00]$
B′−D	-45.33 $[-56.00, -34.67]$	-30.67 $[-38.67, -22.67]$
B′−B	-36.67 $[-46.67, -26.00]$	-35.33 $[-45.33, -25.33]$
D−A	-28.00 $[-36.00, -20.00]$	-34.67 $[-44.00, -25.33]$
I	-8.67 $[-20.00, +3.33]$	-0.67 $[-14.00, +12.67]$

apart from rounding, but give different descriptive allocations to the two factors. This example does not establish a statistically reliable ranking of their importance; it shows why reporting only one path can obscure the dependence of attribution on the reference configuration. The four-corner design makes that dependence explicit and estimable, rather than treating the intermediate contrasts as universal contributions. The binding interaction interval allows a negative interaction of 20 percentage points; the relation interval also allows nontrivial departures from zero. Neither supports a claim that steps and guidance act independently. The experiment contains no student condition, so it cannot settle whether student weights harm or repair semantic control. Its contribution is the missing teacher comparison and a more precise definition of the effects that would need reliable measurement.

4.2 Teacher–Student Pipeline Comparisons

Beyond the within-teacher experiment, the SD3.5 and FLUX pipeline comparisons report positive C−B′ differences under the global VLM protocol (Appendix A). FLUX here denotes Black Forest Labs' FLUX.1 text-to-image model family (Labs, 2024); the comparison uses its dev and schnell variants in the reference and accelerated-model roles, respectively.[2] A compensation interpretation requires distinguishing a pipeline-level gain from a weight-specific effect: these experiments compare loaded pipelines, not an independently verified intervention on denoising weights alone. For FLUX in particular, a shared numerical guidance setting does not establish a shared guidance operation across the selected variants. For SD3.5, checkpoint-specific settings and component identity also require an execution-level account. These results characterize deployment-system differences rather than isolated weight effects. Even before considering evaluator error, neither the positive average contrasts nor the operating-point intervals rule out sample-specific damage, local channel loss, or guidance effects elsewhere in the tested models' parameter ranges.

4.3 Teacher Dose Sweeps: Sensitivity, Not Policy Validation

To examine responses between the coarse comparison endpoints, the teacher dose sweeps vary one inference parameter over a grid with 100 prompts and three seed bases per setting. Under the sweep evaluation protocol, the SD3.5 four-step guidance sweep changes the mean of the two prompt-group rates from 0.153 at API guidance 0 to 0.640 at guidance 3.0, then 0.500 at guidance 4.5. The FLUX four-step sweep changes from 0.350 at guidance 0 to 0.490 at guidance 2.0, then 0.210 at guidance 4.5. At fixed FLUX guidance 3.5, increasing steps from 4 to 28 changes the score from 0.360 to 0.883.

These are exploratory response curves, and the FLUX guidance curve is not flat. They do not establish out-of-sample prediction: the same prompts are used to select and evaluate the best swept setting. They also operate on teachers, so their responses do not establish the benefit of changing deployed-student settings. The 28-step comparison changes inference cost, and equal step counts need not imply equal compute when CFG uses additional evaluations. The SD3.5 shared endpoints also disagree between the A/B/B′/C pipeline-comparison and sweep reports by more than rounding can explain. Differences in text settings or loading behavior are candidate explanations, not verified explanations of the discrepancy. Until execution and image identities reconcile those endpoints, we do not pool the runs or use their cross-run student-versus-best-teacher intervals as decisive evidence. In addition, nominal guidance values must be grouped by their effective implementation: API val-

[2]Artifact model cards: `https://huggingface.co/black-forest-labs/FLUX.1-dev` and `https://huggingface.co/black-forest-labs/FLUX.1-schnell`.

Table 4: Atomic measurement checks. Known-answer controls test a narrow synthetic domain; natural-image reversal checks test consistency, not visual accuracy.

Check	Result
Known-control presence answers correct	24/24
Known-control color answers correct	22/24
Known-control relation answers correct	16/16
Natural relation cases with two determinate directions	19/24
Inconsistent reversal among those cases	16/19
Natural relation cases with determinate relation verdict	3/24

ues that all disable extra CFG are not distinct linear guidance doses. Thus the sweeps motivate calibration but do not yet validate a calibration policy.

5 Measurement Audit and Limits of Attribution

The configuration contrasts raise a measurement question: do the recorded judgments consistently represent the prompt constraints they are intended to assess? We examine stored global responses, test individual constraints and subject–object reversals on fixed images, and assess the coverage of an independent localization-based check. Together, these tests distinguish record integrity, answer consistency, and the evidence needed for semantic attribution.

5.1 Global Labels Can Contradict Their Own Explanations

Inspection of the teacher experiment's stored responses confirms that all 1,200 parse successfully and reproduce the saved Boolean (true/false) values, but reveals a dependence among their error fields. Specifically, every response satisfies `object_error = not prompt_valid`. All 526 judged failures consequently include an object-error label, with no attribute-only or relation-only failure signature. This dependence is present in the raw answers, not introduced by the summarizer. It prevents treating the error fields as independent evidence of distinct failure mechanisms.

Inspection also reveals direct contradictions in the primary outcome. For a prompt requesting a chair to the left of a book, one answer describes the chair as being on the right and explicitly mismatching the request, but returns global success and no relation error. Another response accepts a prompt while explaining that its requested spatial relation is reversed. These examples establish internal inconsistency without requiring us to decide which image description is correct. They do not provide an evaluator false-positive rate: neither the explanation nor the Boolean field is independent ground truth. We therefore preserve the original labels rather than selectively flipping them to agree with the narrative.

5.2 Atomic Questions Do Not Resolve Natural-Image Relations

To inspect individual constraints behind the global verdicts, the follow-up fixes 48 generated images depicting natural objects from four prompts, three seeds, and the four teacher conditions; we call these natural images to distinguish them from synthetic geometric controls. It adds eight known-answer geometric controls and separates object presence, object-specific color, and ordered relation questions. An ordered relation identifies a subject and a reference object: "chair left of book" must become "book right of chair" when their order is reversed. An individual answer is determinate when it returns a specific category or direction rather than uncertainty. The derived constraint verdict is satisfied or violated when supported by consistent determinate answers, unknown when relevant answers are uncertain or inconsistent, and unassessable when prerequisite object presence is not confirmed. The pipeline aggregates atomic answers rather than allowing the judge to freely assign a global success label. All 304 saved question responses are parseable and reproducible under the recorded summary logic. This is an execution result, not proof that the observations are correct.

Simple control performance does not transfer automatically to natural scenes. Among 19 natural relation cases with determinate answers in both query directions, 16 fail the prescribed reversal relation. The follow-up contains only two relation prompts, so this fraction must not be presented

as an error rate for all prompts or all VLMs. Across the 24 relation images, the derived relation verdict is unknown for 15, unassessable for six, and satisfied for three. An overall prompt verdict can still be a failure because another constraint is violated; a low overall unknown rate therefore does not imply adequate relation coverage. The known-control color errors expose another dependency: answers assign a color to an absent target, despite a separate presence answer identifying its absence. Consequently, decomposing a question removes one source of aggregation freedom but does not guarantee correct grounding of each atomic observation. The follow-up does not recover a reliable semantic endpoint for the four-corner experiment.

5.3 Localization Adds Evidence but Leaves Coverage Gaps

A separate localization probe evaluates the same 48 natural images and eight controls using Grounding DINO, an object detector queried with category names (Liu et al., 2024).[3] A candidate is a returned bounding box for a queried category; the geometric relation proxy compares the centers of two uniquely detected boxes rather than asking for a free-form relation answer. Two queried natural categories each have a unique candidate in 26 of 48 images, but uniqueness is not verified category correctness. Only four of 12 two-dimensional relation cases receive a determinate geometric proxy; the 12 depth-relation cases are outside the two-dimensional rule's scope. There are no complete four-condition paired relation cases. Comparing only the surviving cases would therefore change the sample composition across conditions.

Beyond box coverage, category correctness also matters: known negative controls, queries for categories deliberately absent from a control image, expose ambiguity despite accurate localization of some present objects. In a follow-up over the eight control images, single-category queries name one target class, joint-category queries name all tested classes, and reversed-joint queries reverse that list's order. All three modes produce detections for eight of eight absent-category queries. The detector's token score measures a candidate box's match to a text token, a word or word fragment produced by the query tokenizer. All 42 retained single-category candidates have their maximum token score on the queried target word, so removing non-target token maxima does not fix this observed failure. These findings stop this particular geometric-category adaptation branch. They neither establish that natural-object localization is universally unreliable nor supply a substitute semantic label for the original images.

Taken together, the global-response, atomic-question, and localization checks identify specific measurement failures and unresolved outcomes without estimating their full prevalence in the original benchmark. Although the execution records are internally consistent, the judgments and localization coverage do not yet support the intended semantic attribution. The configuration contrasts therefore remain records of evaluator response rather than established effects on true semantic control.

6 Implications for Inference Calibration

Configuration sensitivity identifies settings worth considering, while the measurement audit specifies what must be checked before using their scores to choose a deployment configuration. We connect these observations through a budget-constrained parameter-selection problem and define the held-out comparisons needed to evaluate its practical benefit. The resulting formulation is an evaluation target for future calibration, rather than a validated deployment recipe.

The observed teacher sensitivity suggests studying inference calibration: selecting inference-time parameters while keeping pretrained model weights frozen. This is parameter selection for generation, not calibration of an evaluator's confidence probabilities. Optimized sampling schedules and learned time discretization already demonstrate the value of changing inference parameters without retraining the denoiser (Sabour et al., 2024; Tong et al., 2025). Guidance-update modifications offer a complementary intervention: adaptive projected guidance reduces high-guidance oversaturation by modifying the update direction (Sadat et al., 2024). The open question here is whether configuration selection improves reliably measured prompt-constraint satisfaction under a specified deployment budget. However, the beneficiary and action space must be specified before calling a diagnostic actionable. Calibrating a teacher at four steps, modifying a student at its supported settings,

[3]Detector artifact: `https://huggingface.co/IDEA-Research/grounding-dino-tiny`; recorded revision `a2bb814dd30d776dcf7e30523b00659f4f141c71`.

and choosing between teacher and student systems are different problems. An operation available to one model need not exist or have the same meaning in another. We therefore do not issue a general instruction to tune guidance rather than replace weights.

For a specified generator and budget, a candidate calibration rule could select an inference parameter θ using development prompts:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta_{\text{valid}}} \widehat{Q}_{\text{cal}}(\theta) \quad \text{subject to} \quad K(\theta) \leq K_{\max}. \tag{3}$$

Here Θ_{valid} is the set of parameter settings supported by the generator, $\widehat{Q}_{\text{cal}}(\theta)$ is the mean score on development prompts under a fixed evaluation protocol, $K(\theta)$ is measured inference cost, and $K_{\max}$ is the allowed cost limit. The selected setting is $\widehat{\theta}$; the subscript "cal" denotes the development set used for calibration. This expression defines a decision target, not a new algorithm or an achieved experimental result. Any proposed diagnostic-guided restriction of the search space must earn its value against ordinary search under the same total budget, including diagnostic cost.

A deployment evaluation must separate three tests: (1) lock configuration selection using development prompts and evaluate on held-out prompts; (2) compare with default settings and budget-matched search, reporting inference cost as well as the fixed-protocol score; and (3) establish that the evaluation is meaningful for the task, including failures and abstentions, cases for which the evaluator declines to assign a definite outcome, rather than only successfully detected cases. Changing seeds on the same prompts does not alone provide a prompt-level holdout. Let B^* denote the teacher configuration with the highest observed mean of the binding- and relation-prompt scores in a sweep. Selecting B^* on the evaluation set makes it a descriptive best-in-sweep configuration, not an independently validated policy. The existing teacher sweeps satisfy neither a deployed-student test nor this held-out policy comparison, and no completed standard-evaluation integration is reported here. These requirements define the next test of inference calibration with fixed pretrained weights without treating its expected benefit as an achieved result.

7 Limitations

Our findings are scoped to the tested configurations and evaluation protocols. The teacher factorial examines one checkpoint at two step counts and two guidance settings; additional models and intermediate settings would help characterize the generality and shape of these responses. The reported contrasts describe automatic evaluator outputs, while the fixed-image audits identify specific consistency and coverage issues rather than estimate population-wide semantic accuracy. Teacher/student comparisons are interpreted at the pipeline level because weight changes are not fully isolated from other components. The dose sweeps characterize configuration sensitivity; establishing deployment benefit requires configuration selection on development prompts and evaluation on held-out prompts under matched budgets. Internal causal mechanisms remain outside the scope of the present experiments.

8 Conclusion

Paired configuration experiments and fixed-image measurement checks provide complementary evidence for interpreting few-step performance gaps. In the 1,200-image teacher experiment, step reduction lowers automatic prompt-validity rates at both tested guidance levels, and guidance reduction lowers them at both tested step counts. The added many-step, low-guidance condition completes the comparison and permits measurement of step–guidance interaction. The fixed-image audit identifies label–explanation contradictions, relation-reversal inconsistencies, and localization coverage gaps, revealing distinct measurement issues behind prompt-level scores. Together, these findings yield a concrete evaluation structure: measure conditional configuration responses, inspect the corresponding task-specific judgments, and evaluate configuration selection on held-out prompts under matched budgets. This structure provides a basis for testing inference calibration with fixed pretrained weights.

ARTIFICIAL INTELLIGENCE (AI) USE STATEMENT

In this work, we used generative AI tools for grammar correction and rewriting. We have not used generative AI tools for generating synthetic numerical results to substitute for measured experimental outcomes, and AI-assisted recruitment or evaluation of human participants are not applicable to this work. Additionally, we used generative AI tools for literature discovery, manuscript organization and language editing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-

$$\backslash$$

alpha : $Fasttrainingofdiffusiontransformerforphotorealistictext - to - imagesynthesis.In$*International conference on learning representations*, $volume2024, pp.57611 - -57640, 2024.$

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In *ICLR*, 2024.

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.

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.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

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

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5392–5399. IEEE, 2024a.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023a.

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*, 2023b.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, June 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *Advances in neural information processing systems*, 37:117340–117362, 2024.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42947–42975. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sabour24a.html.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 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*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL https://arxiv.org/abs/2303.01469.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022.

Vinh Tong, Trung-Dung Hoang, Anji Liu, Guy Van den Broeck, and Mathias Niepert. Learning to discretize denoising diffusion odes. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 47244–47282, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/750046157471c56235a781f2eff6e226-Paper-Conference.pdf.

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.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023.

A Teacher–Student Comparison Details

Table 5 reports the A/B/B′/C pipeline comparisons, which were collected separately from the teacher-only four-corner experiment. The SD3.5 comparison used the medium base model and the TensorArt medium-turbo pipeline;[4] the FLUX comparison used dev and schnell in the respective reference and student roles. The main prompt subset contained 50 binding and 50 relation prompts with seed bases 1000, 2000, and 3000. The original intervals resampled matched prompt–seed examples rather than prompt clusters, using 5,000 resamples for VLM outputs and bootstrap seed 42. These intervals do not account for systematic evaluation errors or unresolved cross-pipeline differences.

Table 5: Teacher–student pipeline contrasts under the automatic evaluation protocol. C$-$B′ compares loaded pipelines rather than isolated denoising weights. These runs are separate from the teacher-only factorial.

Family	Contrast	Binding prompts	Relation prompts
SD3.5	B$-$A	-0.347 $[-0.427, -0.267]$	-0.320 $[-0.400, -0.240]$
SD3.5	B′$-$B	-0.373 $[-0.467, -0.280]$	-0.380 $[-0.467, -0.293]$
SD3.5	C$-$B′	$+0.393$ $[+0.300, +0.480]$	$+0.400$ $[+0.313, +0.480]$
FLUX	B$-$A	-0.567 $[-0.647, -0.487]$	-0.480 $[-0.567, -0.393]$
FLUX	B′$-$B	-0.033 $[-0.127, +0.060]$	$+0.013$ $[-0.080, +0.107]$
FLUX	C$-$B′	$+0.587$ $[+0.507, +0.667]$	$+0.513$ $[+0.427, +0.600]$

The corresponding CLIP deltas for B$-$A, B′$-$B, and C$-$B′ are -0.038, -0.052, and $+0.051$ in SD3.5, and -0.056, $+0.008$, and $+0.046$ in FLUX. These broadly aligned directions do not demonstrate preserved perceptual quality alongside selectively lost semantic control. No pure weight-damage or no-damage conclusion follows from this table.

[4] Student artifact: https://huggingface.co/tensorart/stable-diffusion-3.5-medium-turbo. This source identifies the checkpoint and is not evidence of a particular unreported distillation objective.

B Configuration Sweeps and Baseline Screening

Dose settings. The recorded SD3.5 four-step guidance grid is $\{0, 0.5, 1, 1.5, 2, 3, 4.5\}$. The FLUX four-step grid is $\{0, 0.5, 1, 1.5, 2, 3, 3.5, 4.5\}$, and its fixed-guidance 3.5 step grid is $\{4, 8, 12, 16, 28\}$. Each setting uses 100 prompts and three seed bases in the sweep reports. For SD3.5, nominal values at or below one must not be interpreted as distinct CFG mixing coefficients when the implementation bypasses the extra branch. The original best-in-sweep selections and student comparisons reuse evaluation prompts; they are not held-out estimates.

Cross-run comparability. The SD3.5 pipeline-comparison table implies equal-weight averages of the binding- and relation-prompt scores near 0.577 at four steps and guidance 4.5, and 0.200 at guidance 0. The sweep reports 0.500 and 0.153 at those nominal endpoints. These values come from different runs and must not be silently reconciled by choosing the preferred table. The later teacher factorial is a further run with its own audited records. The identity of nominal settings alone is not sufficient to merge these results.

PixArt baseline screening. PixArt-α, an additional text-to-image diffusion-transformer family tested in the pilot (Chen et al., 2024), received teacher automatic-label rates of 0.400 and 0.220 for binding and relation prompts.[5] An easier prompt-set follow-up reported 0.280 and 0.260, below the study's chosen 0.60 threshold for each group. This motivated stopping cross-family interpretation for that branch. The threshold is an operational screening choice, not a theorem about identifiability. A floor effect means that a baseline is so low that further deterioration is difficult to detect; low evaluator scores do not by themselves prove that the teacher's true semantic ability is at such a floor. Neither the initial run nor the follow-up is counted as evidence of an absent operating-point effect.

C Prediction-Space Trajectory Analysis

The SD3.5 trajectory probe compares model predictions along separately generated teacher, low-guidance teacher, and student trajectories. Prediction distance is measured by root mean squared error (RMSE), the square root of the mean squared elementwise difference between prediction tensors. The reported improvement is the distance of B′ to the guided teacher reference minus the distance of C to the same reference, so positive values indicate that C is closer. The recorded improvements are 0.1177, 0.1151, 0.1149, and 0.0857 across four step indices; the corresponding reported fractions of cases in which C is closer are each 0.958. The cases were selected through the pipeline-comparison failure/recovery analysis, rather than sampled as an unselected population test. The rounded aggregate does not establish its sample count; the complete selection manifest is required to establish its denominator and selection criteria.

After the first update, the compared predictions can be evaluated at different latent states. Matching the step index does not establish equal noise level, and model outputs must be interpreted using the actual prediction parameterization rather than assuming epsilon prediction, in which the model estimates the noise added to the image representation. Thus the observation does not identify a correction learned at a common state. It also does not show that greater proximity causes a semantic improvement. Similarly, an unchanged or increased aggregate text contribution cannot exclude token-specific routing or binding failures. A mechanism claim would require a common-state, common-noise-level comparison and a targeted intervention with appropriate controls and a valid semantic endpoint. Replacing a final latent with a successful donor latent is not such evidence: it can reproduce the donor by construction. No causal channel-localization or guidance-correction result is asserted here. The original channel-removal proposal remains unresolved, rather than being rejected by a positive average deployment contrast.

[5] The run records identify `https://huggingface.co/PixArt-alpha/PixArt-XL-2-1024-MS` and `https://huggingface.co/PixArt-alpha/PixArt-LCM-XL-2-1024-MS`. Here LCM denotes latent consistency model (Luo et al., 2023b). The model cards identify artifacts, not independently verified revisions for those runs.

D Evaluation Protocol and Data Integrity

The teacher-only experiment's manifest and judgment tables, stored as comma-separated values (CSV), contain 1,200 matching unique keys. The checked seed bases are 1000, 2000, and 3000, with actual seed equal to seed base plus prompt index. The raw-response audit reproduces the point estimates and prompt-cluster intervals reported in Section 4. It checks record consistency, not natural-image ground truth.

The atomic follow-up contains 48 natural-image cases and 304 question responses including the known controls. The localization follow-up contains 56 image cases in total; the subsequent category-query study uses only the eight geometric controls. They must not be pooled as independent estimates of the same population error rate. The shape-domain negative controls are informative about that configuration but cannot establish natural-category accuracy or inaccuracy.

The reported evidence comprises the pipeline comparisons, teacher factorial, configuration sweeps, and fixed-image audits described above. Standard GenEval integration and held-out calibration are outside this evaluation set; the tables do not include results from those protocols.

E Extended Related Work

Few-step generation. Progressive distillation trains a sequence of students with fewer sampling steps (Salimans & Ho, 2022), while consistency models and latent consistency models use consistency-based learning to support short sampling trajectories (Song et al., 2023; Luo et al., 2023b). Distribution-matching and adversarial objectives provide additional approaches to matching image distributions at low sampling cost (Yin et al., 2024; Sauer et al., 2024), and Hyper-SD studies efficient generation across step regimes (Ren et al., 2024). Guided distillation explicitly learns a single prediction matching a conditional–unconditional combination before reducing sampling steps (Meng et al., 2023). Thus, guidance behavior must be considered when interpreting a teacher/student comparison, rather than assumed unchanged by acceleration. These contributions concern how accelerated generators are trained. Our comparison instead treats pretrained artifacts as fixed systems and examines the effects of their inference configurations and evaluation protocols. This distinguishes a deployment audit from a controlled comparison of training objectives.

Inference-time configuration. DDIM changes the sampling process while retaining the diffusion training objective, while DPM-Solver and the unified predictor–corrector framework UniPC improve numerical sampling efficiency (Song et al., 2021; Lu et al., 2022; Zhao et al., 2023). These approaches distinguish the quality of a short sampling trajectory from the effect of replacing pretrained weights. The modular diffusion design analysis of Karras et al. (2022) and the schedule–implementation mismatches studied by Lin et al. (2024a) further motivate explicitly recording sampling choices. Align Your Steps optimizes sampling schedules, while Learning to Discretize Denoising Diffusion ODEs (LD3) learns the time grid for diffusion ordinary differential equations (ODEs) with the pretrained denoiser fixed (Sabour et al., 2024; Tong et al., 2025). Restricting guidance to a noise interval can also improve sample and distribution quality (Kynkäänniemi et al., 2024). Adaptive projected guidance instead changes the guidance update to mitigate oversaturation (Sadat et al., 2024); neither result establishes the cause of the semantic-score responses in our tested pipelines. These results establish inference configuration as a substantive optimization target. Our teacher experiment instead measures conditional step and guidance responses under a fixed evaluation protocol; it does not introduce a new schedule optimizer or establish semantic gains from one.

Compositional evaluation. GenEval, an object-centric text-to-image evaluation benchmark, and T2I-CompBench, a benchmark for compositional text-to-image generation, motivate evaluation beyond aggregate similarity (Ghosh et al., 2023; Huang et al., 2025). GenEval emphasizes object-centric properties, whereas T2I-CompBench spans attribute binding, relationships, and more complex compositions. Text-to-Image Faithfulness evaluation with question Answering (TIFA) decomposes prompt faithfulness into visual questions (Hu et al., 2023). Davidsonian Scene Graph (DSG) organizes atomic questions into dependency graphs to address coverage and inconsistent answers, including attributes assigned to absent objects (Cho et al., 2024). VQAScore uses visual question answering (VQA) to score image–text alignment through the probability of an affirmative answer (Lin et al., 2024b). Complementary evaluator benchmarks isolate compositional distinctions:

Winoground pairs captions with identical words in different orders, the Attribution, Relation, and Order (ARO) benchmark tests composition-sensitive matching, and What'sUp varies spatial relations while holding object identity fixed (Thrush et al., 2022; Yuksekgonul et al., 2022; Kamath et al., 2023). These studies motivate task-specific evaluator checks; their reported failure rates are not transferred to our judge or image distribution. Our audit builds on this established measurement concern rather than claiming that evaluator inconsistency is new. Its focus is the interpretation of few-step configuration contrasts: saved global responses, subject–object reversal checks, and localization coverage test whether this particular protocol supports the intended attribution.

Intervention and internal representations. Attend-and-Excite and Prompt-to-Prompt intervene on attention to influence generated content (Chefer et al., 2023; Hertz et al., 2022). Diffusion Attentive Attribution Maps (DAAM) associate prompt words with image regions, while Diffusion Hyperfeatures combines representations across denoising layers and timesteps (Tang et al., 2023; Luo et al., 2023a). These studies distinguish interventions that alter generation from analyses that characterize internal representations. Our trajectory probe belongs to the latter category: it records model predictions without modifying the sampling process and compares them with a guided reference. Similarity to that reference is compatible with the prediction-matching objective studied in guided distillation (Meng et al., 2023); it does not by itself identify the training objective of the tested checkpoint or a causal mechanism. Appendix C states the conditions needed to interpret those observations mechanistically.