PLAN BEFORE YOU DENOISE: SEMANTIC AVAILABILITY IN IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A text-to-image model can render a visual answer far more reliably than it can execute the instruction that determines it. On controlled Boolean-program prompts (SYNBOOL), FLUX renders the correct colour on only 27.1% of unresolved instructions at 15 operations (collapsing from 60.8% at 3 ops to chance levels $\leq$ 27.1% for $\geq$ 7 ops), but reaches 80.7% when the answer is supplied. We formalize this execution gap through three complementary computational constraints: (1) *Content*: an exact decoder-relative velocity-risk law proves that when semantic readouts cannot compute the instruction, the population-optimal velocity collapses to uninformative drift, forfeiting conditional guidance despite statistically complete input. (2) *Adaptive compute*: on a stylized random-oracle chain, achieving nontrivial advantage with fewer than L adaptive rounds requires $\Omega(M)$ total queries, establishing that parallel width cannot substitute for serial depth. (3) *Timing*: a finite-Euler controllability bound proves that late semantic disclosure monotonically contracts the remaining Gramian, escalating the minimum steering energy and rendering standard linear guidance ineffective. These constraints illuminate complementary failure modes for one-shot planning (single-pass degradation under spatial crowding) and iterative painting (compounding routing errors). Combining them into a constraint-guided hybrid strategy, PLAN+PAINT, pays serial span in an upstream autoregressive planner and spatial work across incremental rendering passes. PLAN+PAINT boosts exact placement on dense 3×3 relational grids to 0.623–0.773 (vs. 0.401–0.525 for Plan and 0.159–0.529 for Paint), reaches 0.745 on GenEval spatial prompts (vs. 0.395 unresolved), and achieves 0.770 on T2I-ReasonBench entity prompts (vs. 0.345 unresolved, matching an answer-supplied oracle of 0.775).

1 INTRODUCTION

Text-to-image diffusion and flow-matching models render objects, materials, and styles with remarkable photorealistic fidelity. Yet they remain notoriously unreliable at *following* instructions rather than merely depicting visual concepts. Compositional benchmarks document persistent failures on attribute binding, counting, and spatial relations (Huang et al., 2023; Ghosh et al., 2023; Sun et al., 2025). These failures survive better synthetic captions (Betker et al., 2023), scaled rectified-flow backbones (Peebles & Xie, 2023; Esser et al., 2024), and test-time attention steering (Chefer et al., 2023; Feng et al., 2023b). While visual capacity and cross-attention binding each present distinct challenges, we argue that a fundamental class of these failures shares a structural cause: an instruction prompt is often not an illustrative caption of a scene, but an algorithmic specification whose visual consequence must be worked out before anything can be drawn. Whereas depiction maps recognized lexical tokens directly to spatial distributions, instruction execution requires evaluating discrete dependencies over input clauses. In autoregressive models, sequential token generation provides a natural scratchpad for intermediate derivation; in contrast, standard continuous diffusion samplers iteratively refine spatial latents under static cross-attention conditioning, lacking an explicit mechanism to register and propagate discrete symbolic state transitions across denoising time.

Language models faced the same challenge. The field's answer was to allocate inference compute to intermediate symbolic tokens before committing to an output (Wei et al., 2022; Snell et al., 2024; Li et al., 2024). A diffusion sampler appears to offer its own version of that dial: the number of denoising steps. However, sampler updates perform numerical refinement of continuous visual

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.

latents under a static conditioning signal; scaling denoising steps improves perceptual convergence rather than discrete symbolic reasoning (Ma et al., 2025). Consequently, an active line of work externalizes reasoning, employing language or multimodal planners to generate layouts, recaptions, or structured correction loops rendered by a downstream diffusion backbone (Yang et al., 2024a; Lian et al., 2024; Wu et al., 2024). Notably, Gonuguntla (2026) go beyond end-to-end evaluations by using plan corruptions and repairs to disentangle planning errors from execution failures. What remains lacking, however, is an analytical framework that connects conditional generative dynamics directly to the computational availability of instructions: *how do semantic readout constraints in the conditioning interface translate into flow-matching velocity errors, and how does the timing of semantic disclosure dictate the minimal corrective effort required during sampling?*

The execution gap. We begin with the empirical reality. On controlled tasks with 15 Boolean operations (SYNBOOL), FLUX renders the correct colour on only 27.1% of unresolved instructions. When the ground-truth answer is explicitly appended to the prompt, accuracy jumps to 80.7%, reaching 96.7% when given only the concise visual target. While unresolved generation reaches 60.8% on simple 3-operation tasks, it collapses steeply to chance levels (24.8–27.1%) once depth reaches 7 operations and beyond, remaining trapped near chance across a $12.5\times$ sampler-budget sweep (8 to 100 Euler steps) and across different model architectures (Qwen-Image). Furthermore, the renderer faithfully depicts whichever visual target it is given, following a deliberately false answer on 86.8% of tasks. The primary bottleneck is thus instruction execution rather than rendering capability. Substantial answer decodability does not guarantee successful visual execution: on shallow strata, probe accuracy reaches 0.927 while unresolved rendering reaches only 0.646; with increasing serial depth, both quantities deteriorate (Section 3). An upstream chain-of-thought planner closes this gap, lifting end-to-end accuracy to 58.2–68.8%, and the system error admits a clean stage-wise decomposition across planner and renderer components.

Semantic availability and computational accessibility. The concept that resolves this puzzle is not whether information is present in the prompt string, but whether it is *available* at the generative interface. An instruction can be losslessly encoded while the interaction required to act on it remains unresolved. Grounded in computational notions of usable information (Xu et al., 2020), availability demands that the interaction has become directly usable by the generative vector field, not merely recoverable in principle (Franzese et al., 2024). In practice, our probing experiments reveal an empirical bottleneck at the conditioning interface: on shallow Boolean subcircuits a Transformer probe achieves 0.927 accuracy, declining to 0.481 at deep serial dependencies (overall 0.563); meanwhile, unresolved image rendering collapses from 0.646 on shallow circuits to chance (0.271 on 15 ops) as serial depth scales. The conditioning interface's difficulty in resolving deep answers thus serves as an observed bottleneck for generative output. Formally, semantic availability is **decoder-relative**: with respect to a specified semantic readout class $\mathcal{H}$, an interaction is fully available when the best $\mathcal{H}$-restricted velocity field attains the answer-conditioned population optimum.

Constraints on semantic computation. Availability must be actively computed, encountering three complementary structural constraints:

1. **Content.** The conditioning signal must carry a resolved decision in an accessible subspace. Grounded in usable information (Xu et al., 2020), an exact decoder-relative velocity risk law in our sign-product flow model proves that if the downstream readout cannot compute the required interaction ($\mathcal{A}^{(2)}_{\mathcal{H}} = 0$), an optimal restricted velocity collapses to uninformative drift even when the prompt is statistically fully informative (Theorem 1, Corollary 1).

2. **Adaptive compute.** Resolving serial dependencies requires adaptive rounds. On a stylized random-oracle reference chain (Brown-Cohen et al., 2026), we prove that with fewer than L sequential query rounds, achieving nontrivial advantage requires $\Omega(M)$ total queries, leaving the terminal answer close to chance under sublinear budgets $Q = o(M)$ (Theorem 2). This isolates *work* (total operations) from *span* (longest sequential dependency chain), demonstrating that parallel width cannot generically replace sequential compute.

3. **Timing.** The resolved condition must arrive early on the sampler's finite Euler grid. Delaying condition disclosure monotonically contracts the remaining controllability Gramian, strictly raising the minimum correction energy required to steer the trajectory (Theorem 3). In practice, standard linear guidance interpolation incurs severe control inefficiency when delayed, expending $16\times$ more energy without recovering early-disclosure accuracy.

These constraints motivate a constraint-guided allocation strategy across pipeline stages: serial dependencies can be resolved sequentially by an autoregressive language model before passing a compiled condition to the renderer, whereas spatial work is distributed across canvas coordinates.

Contributions.

- **Analytical velocity risk and timing energy bounds in conditional flow matching.** In an analytically tractable conditional flow-matching model, we provide a mechanism model deriving the exact population velocity risk under restricted semantic readout constraints (Xu et al., 2020) (content), establish an adaptive-round lower bound on a stylized random-oracle reference chain (Brown-Cohen et al., 2026) (adaptive compute), and prove that late semantic disclosure monotonically contracts the controllability Gramian, escalating the minimum correction energy (timing; Section 4).
- **Controlled empirical diagnosis of the execution gap.** Across depth, sampler-budget, and architectural controls on SYNBOOL, we diagnose the failure of multi-step unresolved generation (collapsing to 24.8–27.1% and position bias) and evaluate the qualitative predictions of timing and disclosure interventions (Sections 2–3).
- **Constraint-guided inference allocation.** Guided by these constraints, we analyze two complementary allocation paradigms—paying serial span in an upstream planner versus distributing spatial work across visual edit rounds—and formulate their hybrid PLAN+PAINT. On dense 3×3 relational grids, PLAN+PAINT achieves 0.623–0.773 exact placement (vs. 0.401–0.525 for one-shot planning and 0.159–0.529 for painting), transferring to 0.745 on GenEval and 0.770 on T2I-ReasonBench entity prompts (Section 5).

2 EXPERIMENTAL SETUP

SynBool benchmark. Our controlled benchmark, **SynBool** (synthetic Boolean-program prompts), compiles nested Boolean DAG programs into natural-language instructions. Each instruction lists four candidate colours $C = (c_0, c_1, c_2, c_3)$ in fixed order. Two subcircuits evaluate binary bits $(b_1, b_2) \in \{0,1\}^2$, selecting target colour index $Y = 2b_1 + b_2 \in \{0,1,2,3\}$. Each task specifies a single geometric shape on a blank white canvas, enabling deterministic pixel-level scoring. The benchmark evaluates five prompt variants: *Unresolved* (program only), *Answer-supplied* (program plus explicit correct answer), *Answer-plus-distractor* (answer embedded in length-matched non-causal text), *Minimal* (target colour only, e.g., "A solid blue circle on a plain white background"), and *Wrong-answer* (program plus a deliberately incorrect colour). An operation ladder of 3, 7, 15, and 31 operations varies program depth at constant visual complexity (1,000 tasks evaluated per ladder setting).

Renderers. Our primary renderer is FLUX.2-Klein-base-4B (Black Forest Labs, 2026) at 512×512 resolution, classifier-free guidance 4.0, and 28 Euler steps. Sampling sweeps at 8, 28, and 100 steps evaluate the effect of denoising compute. Architectural generality is evaluated on Qwen-Image (Wu et al., 2025), featuring a different visual backbone and text encoder.

Scoring. Each rendered image is mapped to a predicted candidate index $\widehat{Y} \in \{0,1,2,3\}$ using foreground HSV filtering against the candidate palette. We report four-choice accuracy as the primary metric across all controlled conditions (full per-cell numerical results in Appendix D.1).

Planner pipeline. Planning operates as a three-stage pipeline: an upstream *planner* executes the instruction via chain-of-thought autoregression (using the generator's own Qwen3 text encoder); a *compiler* translates the resulting decision into either an explicit full prompt or a concise minimal visual instruction; and the *renderer* generates the image. The compiled prompt is encoded once and cached across all sampling steps (*cached conditioning*; justified formally in Section 4.1).

3 THE EXECUTION GAP AND WHAT CLOSES IT

Controlled experiments on SYNBOOL establish the execution gap, probe its representation interface, and demonstrate how an upstream planner closes it.

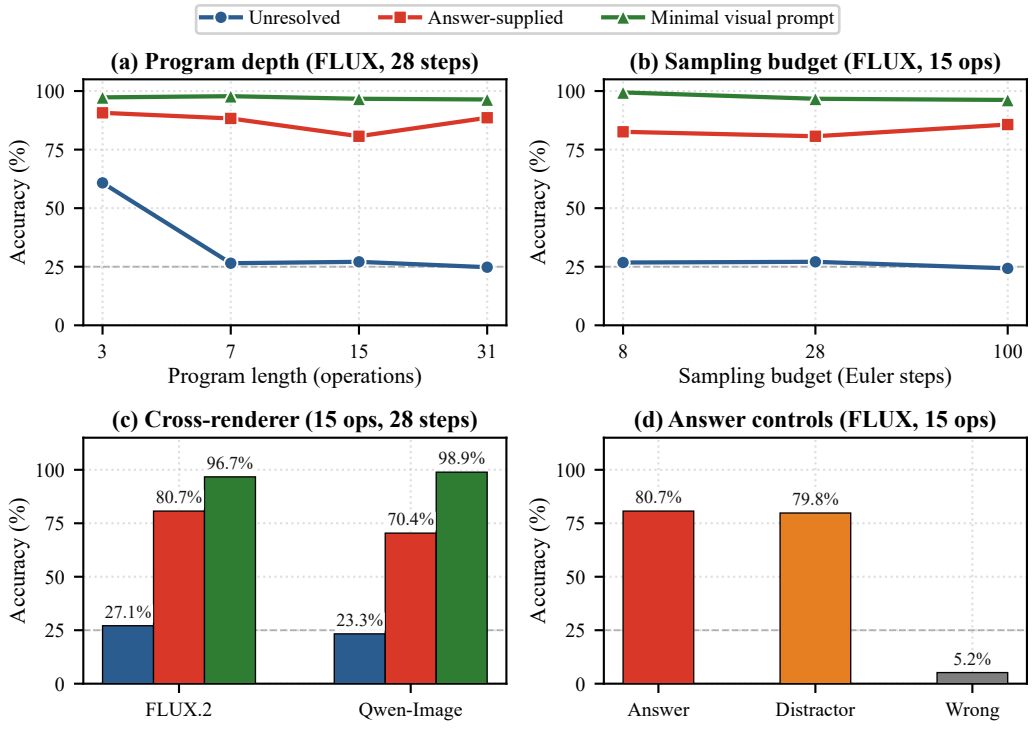


Figure 1: Four-choice colour accuracy (%) on SYNBOOL. Top legend applies to panels (a)–(c); panel (d) evaluates answer controls. (a) Unresolved accuracy reaches 60.8% at 3 operations, but collapses steeply to chance levels (24.8–27.1%, dashed grey line at 25.0%) for ≥ 7 operations, while answer-supplied and minimal conditions remain high. (b) A 12.5× sampling-budget sweep (8–100 Euler steps) yields no systematic improvement. (c) The execution gap generalises across model architectures (FLUX and Qwen-Image). (d) Supplying the correct answer rescues accuracy (80.7%), distractors preserve performance (79.8%), while a deliberately wrong answer reduces ground-truth accuracy to 5.2%, with the injected false colour faithfully rendered on 86.8% of tasks.

3.1 FINDING 1: RENDERERS DEPICT ANSWERS THEY CANNOT INFER

Figure 1 shows the gap. At 15 operations, FLUX renders the correct colour on only 27.1% of unresolved instructions, but reaches 80.7% when the answer is supplied. The visual concepts are thus easily drawn when identified. The minimal condition reaches 96.7%, demonstrating that concise compilation further reduces prompt interference. The answer-plus-distractor arm achieves 79.8%, showing that prompt length alone does not cause the failure: once the terminal decision is exposed, the renderer executes it reliably.

3.2 FINDING 2: THE RENDERER FOLLOWS WHICHEVER DECISION IS EXPOSED

If exposing the decision is what governs generation, the renderer should follow whichever decision is exposed. On a 1,000-task evaluation, supplying the correct answer yields 80.7% accuracy (Figure 1d), whereas supplying a deliberately incorrect answer drops accuracy to 5.2% against ground truth, with the image faithfully depicting the false colour on 86.8% of tasks. Supplying a correct derivation trace yields 78.1%, and randomly permuting its intermediate derivation lines yields 78.7%. The decisive input is the terminal visual decision, not the reasoning narrative.

3.3 FINDING 3: MORE DENOISING STEPS DO NOT RESOLVE THE INSTRUCTION

Figure 1 shows that while simple 3-operation tasks achieve 60.8% unresolved accuracy, performance drops sharply to chance levels (24.8–27.1%) once depth reaches 7 operations and beyond. We observe no systematic improvement over the tested 8–100 Euler-step range (12.5× budget sweep).

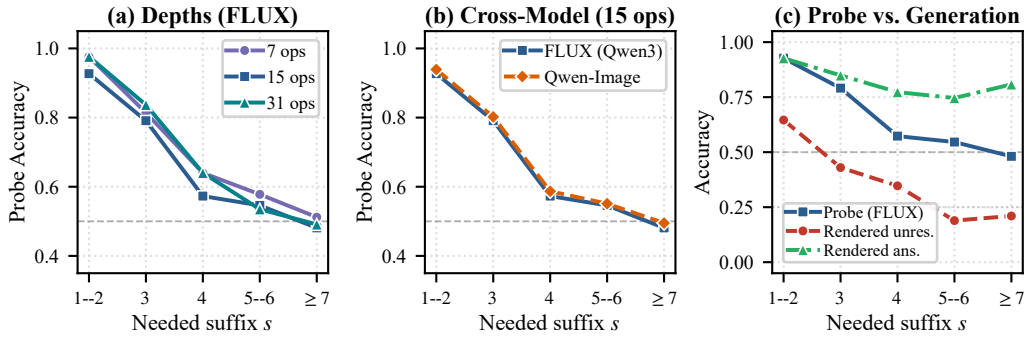


Figure 2: Decodability of answer Y and rendered image accuracy across needed-suffix strata s. (a) Probes collapse at serial depth $s \geq 4$ across program lengths. (b) FLUX and Qwen-Image exhibit similar depth-dependent degradation. (c) Probe decodability drops from 0.927 to 0.481 (overall 0.563), while unresolved rendering decays from 0.646 to chance (0.189–0.210, overall 0.271) and answer-supplied stays high (0.807).

Switching the backbone from FLUX to Qwen-Image reproduces the same failure at 15 operations (23.3%). On multi-step programs, the unresolved prediction marginal is dominated by a positional bias over the four options (38.8%, 37.1%, 13.9%, 10.2%). Across the evaluated models and sampling budgets, this failure reflects a limitation of the conditioning interface: the renderer has not received a resolved instruction, and scaling continuous denoising compute alone does not recover the missing decision.

3.4 Finding 4: Probe decodability collapses with serial depth

We probe frozen text representations across layers using linear, MLP, and four-layer Transformer sequence probes (Appendix C.4). Answer-supplied controls reach 1.000 accuracy early. On unresolved tasks stratified by needed-suffix length s (the circuit depth required to resolve the target), probe accuracy collapses from 0.927 at $s = 1$–2 to 0.481 at $s \geq 7$ (overall 0.563; Figure 2; Table 2). Unresolved rendered accuracy falls steeply from 0.646 to chance (0.189–0.210) for $s \geq 5$ (overall 0.271). Even at $s = 1$–2 where probe decodability is 0.927, unguided rendering lags at 0.646 (vs. 0.927 answer-supplied), confirming that decodability does not imply visual execution without explicit prompting.

3.5 A planner supplies the missing execution

On 15 operations, an autoregressive chain-of-thought planner solves 721/1,000 tasks ($p = 0.721$). Rendered with the explicit compiler, end-to-end success reaches 58.2% (582/1,000), and reaches 68.8% (688/1,000) under the minimal compiler, compared to 27.1% unresolved. Planner accuracy degrades smoothly with program depth (dropping from 91.3% at 3 ops to 31.2% at 31 ops as reasoning tokens scale from 291 to 1,468), whereas unresolved rendering exhibits a sharp early collapse and then remains near chance (direct answering without chain-of-thought likewise stays flat at chance $\approx 25\%$).

3.6 Decomposing pipeline accuracy

The composed pipeline separates cleanly into two stages. Let $p = \mathbb{P}(\widehat{Y}_{\text{plan}} = Y)$, $r_+ = \mathbb{P}(\widehat{Y}_{\text{image}} = Y \mid \widehat{Y}_{\text{plan}} = Y)$, and $r_- = \mathbb{P}(\widehat{Y}_{\text{image}} = Y \mid \widehat{Y}_{\text{plan}} \neq Y)$. End-to-end accuracy is

$$\text{Acc}_{\text{image}} = p\, r_+ + (1 - p) r_-. \tag{1}$$

On SynBool, incorrect plans are virtually never rectified by the renderer ($r_- = 0$ for the minimal compiler; $r_- = 0.025$ for the explicit compiler). At 15 ops on FLUX, substituting $p = 0.721$ and $r_+ = 0.798$ yields $0.721 \times 0.798 + 0.279 \times 0.025 = 0.582$. With the minimal compiler, r_+ rises to 0.954 ($r_- = 0$), lifting accuracy to $0.721 \times 0.954 = 0.688$. Across 15 of 16 depth–compiler–renderer cells, r_+ matches the empirical ceiling of the answer-supplied arm within 5 percentage points.

4 THEORETICAL ANALYSIS

Section 3 leaves one fundamental question open: why does a generative backbone reading an unresolved prompt across 28 denoising steps fail to compute the instruction? To answer this, we examine the smallest setting where the instruction must be computed while visual rendering is held fixed, formalizing the three constraints.

Setting. Let $A \in \{-1,1\}^m$ be independent uniform attributes and $Y = f(A) \in \{-1,1\}$ a balanced Boolean label. The clean image $D \in \mathbb{R}^d$ encodes Y via the signs of its first k coordinates:

$$p(D = x \mid Y = y) = \phi_d(x)\,(1 + y g_k(x))\,, \qquad g_k(x) = \prod_{i=1}^{k} \operatorname{sign}(x_i), \tag{2}$$

so $g_k(D) = Y$ almost surely and marginally $D \sim \mathcal{N}(0, I_d)$. Generation follows the affine flow-matching path (Lipman et al., 2023; Liu et al., 2023)

$$X_\sigma = (1-\sigma)D + \sigma G, \qquad G \sim \mathcal{N}(0, I_d), \quad \sigma \in [0,1], \tag{3}$$

from noise $\sigma = 1$ to clean $\sigma = 0$. Let $v_\sigma^Y(x) = \mathbb{E}[\dot{X}_\sigma \mid X_\sigma = x, Y]$, $v_\sigma^0(x)$ be the unconditional drift, and $v_\sigma^F(x) = \mathbb{E}[\dot{X}_\sigma \mid X_\sigma = x, F]$ for conditioning representation F computed before sampling. Write $\eta(F) = \mathbb{E}[Y \mid F] \in [-1,1]$.

4.1 CONTENT CONSTRAINT: DECODER-RELATIVE SEMANTIC AVAILABILITY

If a prompt contains all variables A that determine $Y = f(A)$, then statistically the answer is fully present: $A_{\text{stat}}(F_{\text{raw}}) := \mathbb{E}[\mathbb{E}(Y \mid A)^2] = 1$. An unconstrained Bayes decoder could evaluate $f(A)$ with zero error. Hence, the execution gap cannot be blamed on missing statistical information alone. Grounded in computational notions of information (Xu et al., 2020), we formalize that the gap arises from constraints on the *semantic decoding capacity* of the conditional velocity field.

Definition 1. *Let $\mathcal{H} \subset L^2(P_A)$ be a fixed finite-dimensional linear semantic readout subspace containing the constant function 1, representing admissible readouts over a prescribed feature set (e.g., degree-r interactions $\mathcal{H}_r = \operatorname{span}\{\prod_{i\in S} A_i : |S| \le r\}$). Define the L^2-**accessibility** of Y with respect to $\mathcal{H}$ as*

$$\mathcal{A}_{\mathcal{H}}^{(2)} := 1 - \inf_{h \in \mathcal{H}} \mathbb{E}[(Y - h(A))^2] = \|\Pi_{\mathcal{H}} f\|^2_{L^2(P_A)} \in [0,1], \tag{4}$$

where $\Pi_{\mathcal{H}}$ is orthogonal projection. The admissible velocity field class is

$$\mathcal{V}_{\mathcal{H}} := \{v : \; A \mapsto v_\ell(x, A) \in \mathcal{H} \text{ for each coordinate } \ell \in [d] \text{ and a.e. } x\}\,, \tag{5}$$

*with unrestricted measurable spatial dependence on $x \in \mathbb{R}^d$. The **restricted velocity risk** is*

$$R_{\mathcal{H},\sigma} := \inf_{v \in \mathcal{V}_{\mathcal{H}}} \mathbb{E}\|v(X_\sigma, A) - v_\sigma^Y(X_\sigma)\|^2. \tag{6}$$

The concept of L^2-accessibility $\mathcal{A}_{\mathcal{H}}^{(2)}$ directly inherits from the framework of usable information under computational constraints (Xu et al., 2020), corresponding to the predictive coefficient of determination under hypothesis class $\mathcal{H}$. When $\mathcal{H} = L^2(\sigma(F))$ is unconstrained, $\Pi_{\mathcal{H}} f = \mathbb{E}[Y \mid F]$ and $\mathcal{A}_{\mathcal{H}}^{(2)} = A_{\text{stat}}(F)$ recovers statistical availability. Our contribution is establishing the direct analytical link between restricted semantic readouts and the population velocity risk of conditional flow matching (Theorem 1).

Theorem 1 (Decoder-relative velocity risk). *Under Eqs. (2)–(6), let $\mathcal{H} \subset L^2(P_A)$ be a fixed finite-dimensional semantic readout subspace containing 1, and let $\mathcal{V}_{\mathcal{H}}$ denote velocity fields whose dependence on A lies in $\mathcal{H}$. Then, for every $\sigma \in (0,1)$,*

$$R_{\mathcal{H},\sigma} \ge J_{k,\sigma}\big(1 - \mathcal{A}_{\mathcal{H}}^{(2)}\big), \tag{7}$$

where $J_{k,\sigma} := \mathbb{E}\|v_\sigma^0(X_\sigma) - v_\sigma^Y(X_\sigma)\|^2$. The bound is sharp at both endpoints: if $\mathcal{A}_{\mathcal{H}}^{(2)} = 0$, an optimal restricted field is the unconditional drift v_σ^0 and $R_{\mathcal{H},\sigma} = J_{k,\sigma}$; if $\mathcal{A}_{\mathcal{H}}^{(2)} = 1$ (equivalently $f \in \mathcal{H}$), then $R_{\mathcal{H},\sigma} = 0$.

Within this model, at the extreme $\mathcal{A}_{\mathcal{H}}^{(2)} = 0$, an optimal restricted field collapses to the unconditional drift v_σ^0, strictly guaranteeing that the generated sample $\widehat{X}$ is independent of Y (Corollary 1). For intermediate $0 < \mathcal{A}_{\mathcal{H}}^{(2)} < 1$, it quantifies the continuous population velocity penalty incurred by the readout's residual semantic deficit.

Corollary 1 (Separation under full statistical information). *Let $F_{\mathrm{raw}} = A$, with raw degree-r readout $\mathcal{H}_r^{\mathrm{raw}} = \mathrm{span}\{\chi_S(A) : |S| \le r\}$ $(1 \le r < m)$. For parity $f(A) = \prod_{i=1}^m A_i$, Walsh–Fourier expansion gives $\mathcal{A}^{(2)}_{\mathcal{H}_r^{\mathrm{raw}}} = \sum_{|S|\le r} \widehat{f}(S)^2 = 0$. Thus:*

1. ***Unresolved input**: $A_{\mathrm{stat}}(F_{\mathrm{raw}}) = 1$, yet $R_{\mathcal{H}_r^{\mathrm{raw}},\sigma} = J_{k,\sigma}$ (the velocity field collapses to uninformative drift);*
2. ***Planned input** $F_{\mathrm{plan}} = (A, f(A))$: $A_{\mathrm{stat}}(F_{\mathrm{plan}}) = 1$, and a linear readout $\mathcal{H}_1^{\mathrm{plan}} = \mathrm{span}\{1, A_1, \ldots, A_m, f(A)\}$ achieves $R_{\mathcal{H}_1^{\mathrm{plan}},\sigma} = 0$.*

Connection to sampler compute. This velocity collapse explains Finding 3.3 (where scaling Euler updates from 8 to 100 yielded no gain): when restricted readouts collapse to uninformative drift ($R_{\mathcal{H},\sigma} = J_{k,\sigma}$), the field contains zero conditional signal along the target mode. Additional sampling steps merely compute a more accurate numerical trajectory under uninformative drift, rather than synthesizing missing semantic guidance.

4.2 Adaptive-compute constraint: who must compute it

Theorem 1 characterizes the loss under an unresolved condition, but does not explain why the backbone cannot resolve it. A natural question is why the backbone cannot resolve this internally, given its substantial nominal compute: dozens of transformer layers and 28 forward passes. Resolving this question requires distinguishing three distinct computational constraints.

Definition 2. *An explicitly specified instruction DAG has* work N *(atomic operations) and* program span L *(longest dependency path). A component has* width B *(atomic operations completed in one round). An algorithm has* R adaptive rounds *if queries in round $r+1$ may depend on answers from round r.*

Work and span are fundamentally different resources. Placing five independent objects incurs work five; placing each relative to its predecessor requires span four. Any schedule respecting a DAG requires $R \ge \max\{\lceil N/B \rceil, L\}$, while a Brent–Graham greedy schedule achieves $R \le \lceil N/B \rceil + L - 1$ (Appendix B.2). To establish an intrinsic separation that cannot be shortcut by parallel width or algebraic re-association, we analyze a stylized random-oracle reference chain (Brown-Cohen et al., 2026).

Let $\theta_0 \in [M]$ be public. For $i = 1, \ldots, L-1$, draw independent random functions $O_i : [M] \to [M]$ and set $\theta_i = O_i(\theta_{i-1})$. Draw terminal oracle $O_L : [M] \to \{-1, +1\}$ and define $Y = O_L(\theta_{L-1})$.

Theorem 2 (Adaptive-round lower bound). *Any querying algorithm with $R < L$ adaptive rounds and at most Q total oracle queries satisfies*

$$\Pr(\widehat{Y} = Y) \;\le\; \tfrac{1}{2} + \frac{Q}{2M}. \tag{8}$$

Proposition 1 (Reduction to decision decoding). *Any generation algorithm $\Gamma(F, G)$ producing an image with $g_k(\Gamma(F, G)) = Y$ with probability p induces an answer decoder $\widehat{Y} := g_k(\Gamma(F, G))$ achieving the identical success rate p under the same oracle query budget.*

Proposition 1 (proof in Appendix B.2) connects decision round bounds to visual generation in the stylized random-oracle setting (Brown-Cohen et al., 2026), establishing that serial span cannot be generically bypassed by parallel width. Together with Theorem 1, this formalizes an *objective-level failure mechanism*: when the conditioning interface cannot resolve deep interactions, the population-optimal velocity collapses to uninformative drift, preventing the sampling trajectory from exploiting the unread decision. Upfront planning resolves this objective-level bottleneck by paying the serial span L in autoregressive reasoning tokens, caching $f(A)$ so that subsequent velocity evaluations require only a trivial linear readout $\mathcal{H}_1$.

4.3 Timing constraint: when it must arrive

Theorems 1 and 2 establish what must exist and who must compute it, but leave open *when* it must reach the sampler.

Definition 3. *A sampler runs M_s Euler updates on grid $\sigma_0 > \cdots > \sigma_{M_s} = 0$ with steps $h_j > 0$. The condition is* disclosed at index J *($0 \le J < M_s$) if updates before J operate under the population-optimal field in the restricted decoder class, which equals the uninformative drift v_σ^0 when $\mathcal{A}_{\mathcal{H}}^{(2)} = 0$, and those from J onward use suffix controls $u_j(G, Y)$ under the resolved condition ($\mathcal{A}_{\mathcal{H}}^{(2)} = 1$). Let u_j be the velocity correction at step j, with drift multiplier $\omega_j := 1 + h_j \frac{1-2\sigma_j}{q_{\sigma_j}}$ for marginal variance factor $q_\sigma := (1-\sigma)^2 + \sigma^2$, magnification $P := \prod_{j=0}^{M_s-1} \omega_j$, and suffix factors $\pi_j := \prod_{\ell>j} \omega_\ell$. The* correction energy *is $\mathcal{E}_J = \mathbb{E} \sum_{j \ge J} h_j \|u_j\|^2$, and the* remaining controllability Gramian *is $W_J = \sum_{j \ge J} h_j \pi_j^2$.*

Theorem 3 (Disclosure and correction energy). *Start from independent Gaussian noise and disclose at update J. For target accuracy $a > 1/2$, the infimum correction energy with which any measurable suffix control reaches accuracy a is*

$$\inf \mathcal{E}_J = \frac{P^2 \, \Theta_k(a)}{2 \, W_J}, \qquad \Theta_k(a) := \int_0^{2a-1} F_{\Delta_k}^{-1}(r) \, dr, \tag{9}$$

where $\Delta_k = \min_{i \le k} G_i^2$ is the squared distance to the nearest sign boundary with CDF F_{Δ_k}. Because every term $h_j \pi_j^2 > 0$, W_J strictly contracts in J, strictly increasing the minimum energy.

Empirical comparison on a 3,200-task grid. On 3,200 paired-noise tasks, unresolved accuracy is 27.8% (891/3,200). Early disclosure at $J = 0$ reaches 87.4% (2,796/3,200) at unit gain with mean energy 1956.6. Delaying disclosure gives 69.7% at $J = 7$ (energy 1205.8), 39.6% at $J = 14$ (energy 653.0), and 28.3% at $J = 21$ (energy 566.0). Theorem 3 establishes that minimum theoretical energy scales as $\inf \mathcal{E}_J \propto 1/W_J$, guaranteeing feasibility for large budgets when $W_J > 0$. In practice, standard sampling uses heuristic linear guidance $v_j = v_I + \lambda(v_E - v_I)$, which exhibits severe inefficiency when applied late: increasing λ from 1 to 4 at $J = 21$ expends $16\times$ more energy (9103.4 vs. 566.0) while accuracy stays near chance (28.7% vs. 28.3%). Thus, late disclosure escalates theoretical minimum energy and renders standard guidance largely ineffective.

These three constraints motivate a clear guideline: resolve discrete dependencies explicitly, compile and expose the result early on the sampling trajectory, and cache the representation.

5 Strategies: Pay the Span in the Planner, the Work in the Renderer

These constraints outline what a compositional pipeline must resolve. They do not dictate how execution should be partitioned across components. We compare two distinct strategies and their hybrid combination (formalized in Appendix C.6, Figure 3): (1) **Plan Before You Draw** *pays the span up front*, resolving the entire dependency chain in one planner call and compiling visual instructions exposed at $J = 0$ for single-pass rendering; (2) **Agentic Painting** *spreads the work over rounds*, inspecting the canvas and executing each relational clause as an incremental edit pass; and (3) **Plan + Paint** *pays the span in the planner and the work in the renderer*, resolving the chain up front in serial reasoning tokens and executing sequential object placement incrementally.

This resource view gives insight into how each strategy alone fails. Plan clears the span reliably, but single-pass rendering degrades under increasing spatial load, frequently dropping objects, consistent with a finite single-pass spatial capacity. Agentic Painting spreads spatial work, but pays span in noisy perceptual hops that compound multiplicatively. PLAN+PAINT mitigates both failure modes.

Relational grid datasets. We evaluate on two relational benchmarks, each evaluating 1,000 tasks: SYNGRID-2×2 (4 cells, 2–4 objects) and SYNGRID-3×3 (9 cells, 2–5 objects). Object 1 is anchored at a named cell, and subsequent objects are placed relative to their immediate predecessor (above, below, left, right). Scoring requires every object in its designated cell.

5.1 Plan Before You Draw pays the span

One planner call clears the chain, sufficient on four cells: Plan reaches 0.809–0.864 against 0.081–0.360 unresolved (Table 1). On dense SYNGRID-3 × 3, Plan drops to 0.401–0.525: while plan

Table 1: Exact placement accuracy on relational grids with FLUX (1,000 paired tasks per benchmark). All arms share programs and seeds. On dense 3×3 grids, PLAN+PAINT outperforms Plan (spatial degradation) and Paint (compounding sequential errors).

SYNGRID-2×2 (4 cells, low spatial crowding)					SYNGRID-3×3 (9 cells, high spatial crowding)				
Obj.	**Unres.**	**Plan**	**Paint**	**P+P**	**Obj.**	**Unres.**	**Plan**	**Paint**	**P+P**
2	0.322	0.864	0.763	**0.995**	2	0.073	0.525	0.529	**0.773**
3	0.081	0.823	0.822	**0.992**	3	0.008	0.427	0.298	**0.727**
4	0.360	0.809	0.790	**0.979**	4	0.007	0.422	0.175	**0.657**
					5	0.002	0.401	0.159	**0.623**

accuracy remains high (0.825–0.975), the one-shot renderer struggles under spatial crowding, dropping objects.

5.2 AGENTIC PAINTING SPREADS THE WORK

Rendering one object per step removes spatial overload, reaching 0.763 on 2×2 and 0.529 on 3×3 at two objects (Table 1). However, on 3×3 grids, accuracy declines steeply to 0.298 (3 objects), 0.175 (4 objects), and 0.159 (5 objects) as perceptual routing errors compound across sequential hops.

5.3 PLAN + PAINT PAYS BOTH

PLAN+PAINT resolves the full sequence with the LLM planner, then renders incrementally. On 2×2 grids it reaches 0.979–0.995. On dense 3×3 grids it reaches 0.623–0.773, surpassing Plan by 15–30 percentage points and Paint by 24–48 percentage points. An oracle arm compiling true intermediate states reaches 0.725–0.900, confirming that incremental execution unlocks the renderer's capacity. These results illustrate a fundamental trade-off: one-shot planning pays serial span up front but risks cross-attention saturation under spatial crowding, whereas iterative painting isolates placement but compounds routing errors across sequential hops. Decoupling discrete reasoning from incremental visual rendering achieves optimal resource alignment across span and work.

5.4 NATURAL PROMPTS: GENEVAL AND T2I-REASONBENCH

The same principles transfer to natural prompts. On 400 GenEval two-object spatial prompts, unresolved FLUX achieves 0.395 accuracy; Plan reaches 0.648, Paint reaches 0.535–0.555, and PLAN+PAINT reaches 0.745 (Appendix Figure 9). On 200 T2I-ReasonBench entity prompts, unresolved scores 0.345, Paint 0.385, Plan 0.547, and PLAN+PAINT reaches 0.770, virtually matching the answer-supplied oracle of 0.775 (evaluated with Gemini 3.1 Pro under benchmark rubric; Appendix Figure 10; Appendix D.7).

6 CONCLUSION

A text-to-image model may render visual scenes with high fidelity while its unresolved generation on multi-step instructions collapses to chance (24.8–27.1%), dominated by superficial prompt-position heuristics. We formalize this execution gap through three complementary constraints: the condition must carry a resolved decision (**content**); serial dependencies require sequential rounds (**adaptive compute**); and the result must reach the sampler before its controllability Gramian contracts (**timing**). Our analysis provides an objective-level mechanism model: within a restricted semantic readout subspace, the population-optimal velocity collapses to uninformative drift despite statistically complete input. Across evaluated pretrained models and sampling budgets, continuous denoising does not resolve multi-step instructions, whereas exposing visual consequences substantially improves generation: **resolve the instruction, compile its visual consequence, cache the condition, and expose it from the first update.** Autoregressive language models resolve serial dependencies through sequential token generation, while diffusion models distribute spatial work across the canvas; PLAN+PAINT combines upstream serial resolution with incremental spatial execution.

AI USE STATEMENT

In this work, generative AI tools were used exclusively for language polishing, grammar correction, and improving stylistic clarity of the text. Generative AI was not used for ideation, theoretical derivations, or experimental analysis. The authors have reviewed, verified, and take full responsibility for all content presented in this manuscript.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our theoretical and empirical findings, complete mathematical proofs and derivations are provided in Appendices B, B.2, and B.3. Experimental protocols, model configurations, prompt templates, and evaluation rubrics are detailed in Appendices C and C.2. The code used for benchmark generation, controlled interventions, evaluation, and inference-allocation experiments will be released upon publication.

REFERENCES

Iskander Azangulov, Peter Potaptchik, Qinyu Li, Eddie Aamari, George Deligiannidis, and Judith Rousseau. Adaptive diffusion guidance via stochastic optimal control, 2025. URL https://arxiv.org/abs/2505.19367.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers, 2022. URL https://arxiv.org/abs/2211.01324.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. OpenAI technical report, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Black Forest Labs. FLUX.2 [klein] 4B Base: Model card, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B.

Jeremiah Blocki, Seunghoon Lee, and Samson Zhou. On the security of proofs of sequential work in a post-quantum world. In *2nd Conference on Information-Theoretic Cryptography (ITC 2021)*, pp. 22:1–22:27, 2021. doi: 10.4230/LIPIcs.ITC.2021.22. URL https://doi.org/10.4230/LIPIcs.ITC.2021.22.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023. URL https://arxiv.org/abs/2211.09800.

Jonah Brown-Cohen, David Lindner, and Rohin Shah. Quantifying the necessity of chain of thought through opaque serial depth, 2026. URL https://arxiv.org/abs/2603.09786.

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*, 42(4):1–10, 2023. URL https://arxiv.org/abs/2301.13826.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.11215.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 7654–7664, 2022. doi: 10.18653/v1/2022.acl-long.527. URL https://aclanthology.org/2022.acl-long.527/.

Kai-Min Chung, Serge Fehr, Yu-Hsuan Huang, and Tai-Ning Liao. On the compressed-oracle technique, and post-quantum security of proofs of sequential work. In *Advances in Cryptology – EUROCRYPT 2021*, pp. 598–629, 2021. doi: 10.1007/978-3-030-77886-6_21. URL https://doi.org/10.1007/978-3-030-77886-6_21.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas M"uller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.03206.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. GoT: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a. URL https://arxiv.org/abs/2305.15408.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2212.05032.

Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. MINDE: Mutual information neural diffusion estimation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.09031.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023. URL https://arxiv.org/abs/2310.11513.

Ashritha Gonuguntla. The plan, not the decoder: Diagnosing and repairing compositional failure in reasoning-augmented text-to-image generation, 2026. URL https://arxiv.org/abs/2608.21713.

Georg A. Gottwald, Shuigen Liu, Youssef Marzouk, Sebastian Reich, and Xin T. Tong. Localized diffusion models, 2025. URL https://arxiv.org/abs/2505.04417.

Xinran Gu, Kaifeng Lyu, Jiazheng Li, and Jingzhao Zhang. Data mixing can induce phase transitions in knowledge acquisition. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.18091.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.01626.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.11239.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023. URL https://arxiv.org/abs/2307.06350.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level CoT. *arXiv preprint arXiv:2505.00703*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.00364.

Alexander Kozachinskiy, Tomasz Steifer, and Przemysław Wałęga. Parity, sensitivity, and transformers, 2026. URL https://arxiv.org/abs/2602.05896v3. Version 3, 5 September 2026.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2404.07724.

Joseph Lehec. Representation formula for the entropy and functional inequalities. *Annales de l'Institut Henri Poincare, Probabilites et Statistiques*, 49(3):885–899, 2013. doi: 10.1214/11-AIHP464. URL https://arxiv.org/abs/1006.3028.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.12875.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2305.13655.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2206.01714.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Yuxi Liu. Perfect diffusion is TC0—bad diffusion is turing-complete, 2025. URL https://arxiv.org/abs/2507.12469.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps, 2025. URL https://arxiv.org/abs/2501.09732.

Song Mei and Yuchen Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models, 2023. URL https://arxiv.org/abs/2309.11420.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023. doi: 10.1162/tacl_a_00562.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021. URL https://arxiv.org/abs/2112.00114.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://arxiv.org/abs/2212.09748.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.01952.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, 2022. URL https://arxiv.org/abs/2204.06125.

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. URL https://arxiv.org/abs/2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.11487.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://arxiv.org/abs/2010.02502.

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. URL https://arxiv.org/abs/2011.13456.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024. doi: 10.1162/tacl_a_00663.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2I-ReasonBench: Benchmarking reasoning-informed text-to-image generation, 2025. URL https://arxiv.org/abs/2508.17472.

Binxu Wang and John J. Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications, 2024. URL https://arxiv.org/abs/2412.09726.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. GenArtist: Multimodal LLM as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zixuan Wang, Xingyu Dang, Jason D. Lee, and Kaifeng Lyu. The power of power law: Asymmetry enables compositional reasoning, 2026. URL https://arxiv.org/abs/2604.22951.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

Chenfei Wu, Jiahao Zhu, Wenyi Wang, Zhiheng Yang, Chen Li, Yi Huang, Junyang Lin, et al. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting LLM-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6327–6336, 2024. URL https://arxiv.org/abs/2311.16090.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10689.

Zhiyang Xun and Eric Price. Query lower bounds for diffusion sampling, 2026. URL https://arxiv.org/abs/2604.10857.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *International Conference on Machine Learning*, 2024a. URL https://arxiv.org/abs/2401.11708.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2Img: Iterative self-refinement with GPT-4V for automatic image design and generation. In *European Conference on Computer Vision (ECCV)*, 2024b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.10601.

TABLE OF CONTENTS

1 Introduction **1**

2 Experimental setup **3**

3 The execution gap and what closes it **3**

3.1 Finding 1: Renderers depict answers they cannot infer 4

3.2 Finding 2: The renderer follows whichever decision is exposed 4

3.3 Finding 3: More denoising steps do not resolve the instruction 4

3.4 Finding 4: Probe decodability collapses with serial depth 5

3.5 A planner supplies the missing execution 5

3.6 Decomposing pipeline accuracy 5

4 Theoretical analysis **6**

4.1 Content constraint: decoder-relative semantic availability 6

4.2 Adaptive-compute constraint: who must compute it 7

4.3 Timing constraint: when it must arrive 7

5 Strategies: pay the span in the planner, the work in the renderer **8**

5.1 Plan Before You Draw pays the span 8

5.2 Agentic Painting spreads the work 9

5.3 Plan + Paint pays both 9

5.4 Natural prompts: GenEval and T2I-ReasonBench 9

6 Conclusion **9**

A Related work **17**

A.1 Instruction following and compositionality in text-to-image models 17

A.2 External planners and reasoning-augmented generation 17

A.3 Inference-time compute and sequential depth 18

A.4 When conditioning acts: guidance and the sampler's time axis 18

A.5 Theoretical accounts of what these models can compute 18

B Theoretical analysis and proofs **19**

B.1 Content constraint: Theorem 1 19

B.2 Adaptive-compute constraint: Theorem 2 24

B.3 Timing constraint: Theorem 3 26

C Comprehensive experimental protocols **28**

C.1 FLUX configuration 28

C.2 Flow-control tasks, interventions, and scoring 28

C.3 SYNBOOL tasks, scorer, and planner 28
C.4 Representation probing 29
C.5 Relational grids 29
C.6 Multi-round agentic painting 29
C.7 Natural-image validation on GenEval 29
C.8 Natural-image validation on T2I-ReasonBench 29

D Complete experimental results **30**

D.1 Accuracy across all SYNBOOL cells 30
D.2 Probing semantic availability in conditioning representations 30
D.3 Disclosure–gain interventions 30
D.4 Single-update responses 30
D.5 Relational-grid qualitative examples 32
D.6 Natural-image validation on GenEval 32
D.7 Natural-image validation on T2I-ReasonBench 32

A Related work

This appendix situates the paper in five established literatures. The central message is this: the empirical phenomenon we study is widely encountered; the remedy we analyze is widely deployed; but the theoretical quantity that explains both has not been isolated: namely, *when a task interaction becomes semantically available to the generative velocity field*.

A.1 Instruction following and compositionality in text-to-image models

Diffusion and flow models produce photorealistic images from natural-language text (Ho et al., 2020; Song et al., 2021b; Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022; Podell et al., 2024; Peebles & Xie, 2023; Esser et al., 2024). However, prompt adherence consistently lags visual quality. Compositional benchmarks isolate this shortfall across attribute binding, counting, and spatial relations (Huang et al., 2023; Ghosh et al., 2023). More recent benchmarks target prompts whose visual content must be inferred or deduced rather than read off directly (Sun et al., 2025). That reasoning-informed regime is precisely the setting this paper formalizes.

Existing remedies act at three primary levels:

1. *Supervision and synthetic data:* improved recaptioning and synthetic text-image pairings (Betker et al., 2023);
2. *Architecture and training objectives:* scaled DiT architectures and rectified-flow objectives (Esser et al., 2024; Peebles & Xie, 2023);
3. *Inference-time manipulation:* cross-attention steering, energy-based compositions, and structured guidance (Chefer et al., 2023; Feng et al., 2023b; Liu et al., 2022; Hertz et al., 2023).

All three lines improve how faithfully a *stated* visual specification is rendered. Our measurements separate that question from a prior one: whether the specification has been executed at all. On the same task population, FLUX renders an explicitly stated colour at 80.7–96.7% accuracy, but executes the unresolved instruction that selects that colour at only 27.1%. The compositional failures documented in the literature are consistent with both explanations. Holding the visual target fixed while varying only how its answer is specified empirically distinguishes them.

Instruction following has also been studied as a supervised objective for instruction-based image editing (Brooks et al., 2023). There, the instruction is executed by a model trained directly on instruction-image pairs. That line changes the training distribution. We hold a pretrained renderer fixed and investigate what reaches its conditioning interface.

A.2 External planners and reasoning-augmented generation

A growing family of systems places a language or multimodal planner upstream of the generative backbone. RPG performs global multi-category planning and regional generation (Yang et al., 2024a). LMD prompts an LLM to emit a bounding-box layout that a downstream grounded controller renders (Lian et al., 2024; Li et al., 2023). SLD, Idea2Img, and GenArtist close the loop by iteratively detecting and correcting errors across successive passes (Wu et al., 2024; Yang et al., 2024b; Wang et al., 2024). Unified multimodal models have also emerged that interleave textual reasoning with visual token generation (Fang et al., 2025; Jiang et al., 2025). Crucially, Gonuguntla (2026) go beyond end-to-end benchmark gains by introducing systematic plan ablations, plan corruptions, and geometric repairs to disentangle planning errors from generative decoding limits. These systems establish that external planning and iterative editing can be highly effective.

That literature leaves open an analytical account of the generative interface itself, which this paper addresses:

1. *Why the decoder cannot perform the computation itself.* We provide an objective-level obstruction rather than an empirical observation. When the required interaction is inaccessible ($\mathcal{A}^{(2)}_{\mathcal{H}} = 0$), the population-optimal conditional velocity field collapses to uninformative drift ($v^{\star}_{\mathcal{H},\sigma} = v^{0}_{\sigma}$). Under standard sampling of this learned field, increasing sampler steps does not introduce answer dependence, so continuous denoising compute alone cannot recover the missing decision.

2. *What property a plan must have.* Our answer is *semantic availability*. It is a formal property of the representation the velocity field conditions on, not of the plan's surface text. This explains why the renderer obediently follows a supplied visual decision even when it is false (Finding 2), why a concise minimal visual instruction (96.7%) outperforms an answer embedded amid program text (80.7%), and why the error cleanly admits a stage-wise decomposition into planning and rendering components.

A.3 Inference-time compute and sequential depth

In language modelling, spending test-time compute on intermediate symbolic results before committing to an answer is a foundational technique (Nye et al., 2021; Wei et al., 2022; Wang et al., 2023; Yao et al., 2023; Snell et al., 2024). Its power has a well-understood theoretical basis: chain-of-thought allows a constant-depth transformer to simulate serial polynomial-time computation that it cannot express in a single forward pass (Feng et al., 2023a; Li et al., 2024).

A diffusion sampler appears to offer its own version of test-time compute via its denoising step count. However, sampler steps refine a continuous latent under a largely static text condition. Work on inference-time scaling for diffusion accordingly looks for gains in noise search, tree search, and verifier-guided reward optimization rather than in longer denoising trajectories (Ma et al., 2025). Our step sweep is the corresponding negative control on the semantic axis: between 8 and 100 Euler steps, unresolved instruction accuracy stays flat at 24.3–27.1%, while answer-supplied accuracy remains at 80.7–85.7%. Denoising compute cannot substitute for the serial discrete computation the instruction requires. In this sense, our adaptive-compute analysis is the flow-matching counterpart of the chain-of-thought depth arguments in NLP.

A.4 When conditioning acts: guidance and the sampler's time axis

That conditioning behaves differently at different noise levels is well documented. eDiff-I observes that generation relies heavily on the text prompt at early noise levels and largely ignores it late, motivating stage-specialized expert denoisers (Balaji et al., 2022). Kynkäänniemi et al. (2024) show that classifier-free guidance is harmful at high noise, unnecessary at low noise, and beneficial primarily in an intermediate interval. Stochastic optimal-control formulations adapt the guidance strength dynamically along the trajectory (Azangulov et al., 2025). Sampler theory and discrete schedules establish the numerical grids on which these dynamics operate (Song et al., 2021a; Karras et al., 2022).

Our timing analysis addresses a fundamentally different question. Prior works schedule the *strength* of a guidance signal whose semantic content is fixed throughout, optimizing perceptual quality and Fréchet Inception Distance. We vary *when the semantic content becomes available at all* along the trajectory, measuring symbolic task accuracy. The two can be told apart experimentally. Theorem 3 proves that delaying disclosure monotonically removes terms from the controllability Gramian W_J, raising the infimum correction energy for *any* admissible control ($K_J(a) \propto 1/W_J$). While late correction remains theoretically feasible under a sufficiently large control budget, our empirical sweeps show that heuristic linear guidance updates $v_I + \lambda(v_E - v_I)$ become severely inefficient late in the sampling process: at $J = 21$, scaling the gain from 1 to 4 increases energy by $16\times$ (from 566.0 to 9103.4) while accuracy remains near chance (28.7% vs. 28.3%; Figures 5 and 6). The guidance-interval finding and ours are entirely compatible: guidance gain is best concentrated in the middle of the trajectory, while early disclosure minimizes the required steering energy.

A.5 Theoretical accounts of what these models can compute

Parity is the canonical hard function for bounded-depth attention (Hahn, 2020; Chiang & Cholak, 2022). An extensive expressivity literature characterizes what constant-depth and log-precision Transformers can express (Merrill & Sabharwal, 2023; Strobl et al., 2024; Kozachinskiy et al., 2026). Related work studies when compositional structures are learnable (Gu et al., 2025; Wang et al., 2026). A separate complexity literature investigates sequential query rounds for opaque black-box compositions where active inputs are revealed adaptively (Chung et al., 2021; Blocki et al., 2021; Brown-Cohen et al., 2026). For diffusion models, recent theoretical analyses explore sampling complexity (Chen et al., 2023), score approximation under graphical models (Wang &

Vastola, 2024; Mei & Wu, 2023; Gottwald et al., 2025), computational limits of the sampler (Liu, 2025; Xun & Price, 2026), and information-theoretic/entropy views of the forward-reverse process (Franzese et al., 2024; Lehec, 2013).

Our theoretical analysis operates directly on the population objective of conditional flow matching under restricted semantic decoding, importing computational constraints from the theory of usable information (Xu et al., 2020). Rather than bounding network parameter expressivity, Theorem 1 shows that the excess velocity risk is lower-bounded by $J_{k,\sigma}(1 - \mathcal{A}^{(2)}_{\mathcal{H}})$. Even when an answer is statistically fully determined, if the downstream decoder cannot access the required interaction ($\mathcal{A}^{(2)}_{\mathcal{H}} = 0$), the optimal restricted velocity field collapses to uninformative drift (Corollary 1). Drawing on the complexity of opaque serial depth (Brown-Cohen et al., 2026), Theorem 2 uses a stylized random-oracle reference model to isolate serial span L from parallel width B, which Proposition 1 reduces to visual output accuracy. Finally, Theorem 3 establishes the finite-Euler controllability Gramian bound on the implemented sampling grid.

B Theoretical analysis and proofs

This appendix provides the complete formal setting, lemmas, and proofs for the three theoretical constraints established in Section 4.

B.1 Content constraint: Theorem 1

Formal setting and restricted decoder space. Let $A = (A_1, \dots, A_m) \in \{-1, 1\}^m$ be independent uniform Rademacher attributes with product probability measure P_A, and let $Y = f(A) \in \{-1, 1\}$ be a balanced Boolean target label satisfying $\mathbb{E}_{P_A}[Y] = 0$. Fix dimension $d \geq k \geq 1$. Define the sign-product function

$$g_k(x) = \prod_{i=1}^{k} \operatorname{sign}(x_i), \qquad x \in \mathbb{R}^d. \tag{10}$$

Conditioned on $Y = y \in \{-1, 1\}$, the clean data density is defined by tilting the standard Gaussian density $\phi_d(x)$:

$$p(D = x \mid Y = y) = \phi_d(x)\,(1 + y g_k(x))\,. \tag{11}$$

Because $\int \phi_d(x) g_k(x)\, dx = 0$, this is a valid probability density on $\mathbb{R}^d$. Moreover, $g_k(D) = Y$ almost surely, and the unconditional data marginal is standard Gaussian: $D \sim \mathcal{N}(0, I_d)$.

Generation follows the affine conditional flow-matching path (Lipman et al., 2023; Liu et al., 2023):

$$X_\sigma = (1 - \sigma)D + \sigma G, \qquad G \sim \mathcal{N}(0, I_d), \quad \sigma \in [0, 1], \tag{12}$$

with time velocity $\dot{X}_\sigma = G - D$. Define the scalar parameters:

$$q_\sigma = (1 - \sigma)^2 + \sigma^2, \qquad b_\sigma = \frac{\sigma}{1 - \sigma}. \tag{13}$$

Under the reference Gaussian pair (D, G), the conditional expectation of $g_k(D)$ given $X_\sigma = x$ is

$$H_\sigma(x) := \mathbb{E}[g_k(D) \mid X_\sigma = x] = \prod_{i=1}^{k} \left[2\Phi\left(\frac{(1 - \sigma)x_i}{\sigma\sqrt{q_\sigma}} \right) - 1 \right]. \tag{14}$$

Notice that $|H_\sigma(x)| < 1$ for all $x \in \mathbb{R}^d$ whenever $\sigma \in (0, 1)$. We define the spatial drift vector $w(x) := -b_\sigma \nabla H_\sigma(x)$, the spatial weight function $\ell_\sigma(x) := \frac{b_\sigma^2 \|\nabla H_\sigma(x)\|^2}{1 - H_\sigma(x)^2}$, and the unconditional baseline velocity risk:

$$J_{k,\sigma} := \mathbb{E}_{X_\sigma \sim \mu_\sigma}\left[\|v^0_\sigma(X_\sigma) - v^Y_\sigma(X_\sigma)\|^2 \right] = \mathbb{E}_{X \sim \mathcal{N}(0, q_\sigma I_d)}[\ell_\sigma(X)], \tag{15}$$

where $\mu_\sigma = \mathcal{N}(0, q_\sigma I_d)$ denotes the unconditional marginal distribution of X_σ.

To formalize the computational capacity of the downstream generative backbone, let $\mathcal{H} \subset L^2(P_A)$ be a fixed finite-dimensional linear semantic readout subspace containing the constant function 1 ($1 \in \mathcal{H}$), representing admissible readouts over a prescribed feature set with basis $\Phi(A) = (\phi_1(A), \dots, \phi_p(A))^\top$ with $\phi_1(A) \equiv 1$ (e.g., degree-r polynomial interactions or linear readouts

of prompt tokens). Let $\Pi_{\mathcal{H}} : L^2(P_A) \to \mathcal{H}$ denote the orthogonal projection operator. The L^2-*accessibility* of target $Y = f(A)$ with respect to $\mathcal{H}$ is:

$$\mathcal{A}_{\mathcal{H}}^{(2)} := \|\Pi_{\mathcal{H}} f\|^2_{L^2(P_A)} = \langle f, \Pi_{\mathcal{H}} f \rangle_{L^2(P_A)} \in [0, 1]. \tag{16}$$

The family of admissible velocity fields under restricted semantic decoding $\mathcal{H}$ is defined as:

$$\mathcal{V}_{\mathcal{H}} := \left\{ v(x, A) = \sum_{j=1}^{p} c_j(x)\phi_j(A) : \phi_j \in \mathcal{H},\ c_j : \mathbb{R}^d \to \mathbb{R}^d \text{ measurable} \right\}. \tag{17}$$

Because $1 \in \mathcal{H}$, $\mathcal{V}_{\mathcal{H}}$ naturally contains all purely spatial drifts $c(x) \cdot 1$. The excess restricted velocity risk is the infimum population error over $\mathcal{V}_{\mathcal{H}}$:

$$R_{\mathcal{H},\sigma} := \inf_{v \in \mathcal{V}_{\mathcal{H}}} \mathbb{E}\left[\|v(X_\sigma, A) - v_\sigma^Y(X_\sigma)\|^2\right]. \tag{18}$$

Lemma 1 (Bayes velocity under attribute tilt). *For any $\sigma \in (0, 1)$:*

1. *The conditional density of X_σ given $Y = y$ is $p_\sigma^y(x) = \phi_{q_\sigma I_d}(x)(1 + yH_\sigma(x))$.*
2. *The unconstrained optimal Bayes velocity field given Y is:*
$$v_\sigma^y(x) = v_\sigma^0(x) + w(x)\frac{y}{1 + yH_\sigma(x)}, \qquad v_\sigma^0(x) := \frac{2\sigma - 1}{q_\sigma}x. \tag{19}$$
3. *The conditional distribution of attributes A given $X_\sigma = x$ has Radon–Nikodym derivative with respect to P_A:*
$$\frac{dP_{A|X_\sigma = x}}{dP_A}(a) = 1 + f(a)H_\sigma(x). \tag{20}$$

Proof. Under the unconditioned reference measure, (D, X_σ) is jointly Gaussian. Given $X_\sigma = x$, the coordinates of D are conditionally independent Gaussians with mean $\frac{1-\sigma}{q_\sigma}x_i$ and variance $\frac{\sigma^2}{q_\sigma}$. For a scalar Gaussian $Z \sim \mathcal{N}(\mu, \tau^2)$, $\mathbb{E}[\text{sign}(Z)] = 2\Phi(\mu/\tau) - 1$. Applying this coordinate-wise yields $\mathbb{E}[g_k(D) \mid X_\sigma = x] = H_\sigma(x)$. Tilting the unconditioned Gaussian marginal $\phi_{q_\sigma I_d}(x)$ by $1 + yg_k(D)$ produces $p_\sigma^y(x) = \phi_{q_\sigma I_d}(x)(1 + yH_\sigma(x))$.

For the optimal Bayes velocity, inverting $X_\sigma = (1 - \sigma)D + \sigma G$ gives $D = \frac{X_\sigma - \sigma G}{1-\sigma}$, so $\dot{X}_\sigma = G - D = \frac{G - X_\sigma}{1-\sigma}$. Taking the conditional expectation given $X_\sigma = x, Y = y$ and applying Tweedie's formula under the Gaussian noise model, $\mathbb{E}[G \mid X_\sigma = x, Y = y] = -\sigma\nabla \log p_\sigma^y(x)$. Therefore:

$$\begin{aligned} v_\sigma^y(x) = \mathbb{E}[\dot{X}_\sigma \mid X_\sigma = x, Y = y] &= \frac{-\sigma\nabla \log p_\sigma^y(x) - x}{1 - \sigma} \\ &= \frac{-\sigma\left(-\frac{x}{q_\sigma} + \frac{y\nabla H_\sigma(x)}{1 + yH_\sigma(x)}\right) - x}{1 - \sigma} = \frac{2\sigma - 1}{q_\sigma}x - \frac{\sigma}{1 - \sigma}\frac{y\nabla H_\sigma(x)}{1 + yH_\sigma(x)} \\ &= v_\sigma^0(x) + w(x)\frac{y}{1 + yH_\sigma(x)}, \end{aligned} \tag{21}$$

where $w(x) := -b_\sigma \nabla H_\sigma(x)$ and $b_\sigma := \frac{\sigma}{1-\sigma}$. Finally, the joint density of (X_σ, A) is $p_\sigma(x \mid A = a)P_A(a) = \phi_{q_\sigma I_d}(x)(1 + f(a)H_\sigma(x))P_A(a)$. Integrating over a with $\mathbb{E}_{P_A}[f(A)] = 0$ yields the marginal density $p_\sigma(x) = \phi_{q_\sigma I_d}(x)$. Dividing $p_\sigma(x, a)$ by $p_\sigma(x)$ yields (20). □

Theorem 4 (Restatement of Theorem 1). *Under the sign-product conditional flow model in Eqs. (2)–(6), let $\mathcal{H} \subset L^2(P_A)$ be a fixed finite-dimensional linear semantic readout subspace with $1 \in \mathcal{H}$ and orthonormal basis $\Phi(A) = (\phi_1(A), \ldots, \phi_p(A))^\top$ in $L^2(P_A)$. Define the projection vector $a := \mathbb{E}_{P_A}[f(A)\Phi(A)] \in \mathbb{R}^p$ (satisfying $\|a\|^2 = \mathcal{A}_{\mathcal{H}}^{(2)}$) and the symmetric coupling matrix $T := \mathbb{E}_{P_A}[f(A)\Phi(A)\Phi(A)^\top] \in \mathbb{R}^{p \times p}$. Then the excess restricted velocity risk is given by the exact formula:*

$$R_{\mathcal{H},\sigma} = \mathbb{E}_{X \sim \mathcal{N}(0, q_\sigma I_d)}\left[\ell_\sigma(X)\left\{1 - \left(1 - H_\sigma(X)^2\right)a^\top\left(I_p - H_\sigma(X)^2 T^2\right)^{-1}a\right\}\right]. \tag{22}$$

Consequently, the universal lower bound holds:

$$J_{k,\sigma}\left(1 - \mathcal{A}_{\mathcal{H}}^{(2)}\right) \le R_{\mathcal{H},\sigma} \le J_{k,\sigma}. \tag{23}$$

Moreover, the bound is sharp at both endpoints:

1. *If $\mathcal{A}_{\mathcal{H}}^{(2)} = 0$ ($f \perp \mathcal{H}$), an optimal restricted velocity field is the unguided drift $v^\star(x, A) = v_\sigma^0(x)$, achieving the maximum risk $R_{\mathcal{H},\sigma} = J_{k,\sigma}$.*

2. *If $f \in \mathcal{H}$, then $v_\sigma^Y \in \mathcal{V}_{\mathcal{H}}$ and $R_{\mathcal{H},\sigma} = 0$.*

Proof. We divide the proof into four steps.

Step 1: Pointwise directional decomposition and conditioned least squares. The target velocity discrepancy $v_\sigma^Y(x) - v_\sigma^0(x) = w(x)\frac{f(A)}{1+f(A)H_\sigma(x)}$ is collinear with the spatial gradient vector $w(x)$. Any candidate velocity field $v \in \mathcal{V}_{\mathcal{H}}$ decomposes uniquely as

$$v(x, A) = v_\sigma^0(x) + w(x)\psi(x, A) + v^\perp(x, A), \tag{24}$$

where $\psi(x, \cdot) \in \mathcal{H}$ for almost every x, and $w(x)^\top v^\perp(x, A) = 0$. By orthogonality in $\mathbb{R}^d$,

$$\|v(x, A) - v_\sigma^Y(x)\|^2 = \|w(x)\|^2 \left(\psi(x, A) - \frac{f(A)}{1 + f(A)H_\sigma(x)}\right)^2 + \|v^\perp(x, A)\|^2. \tag{25}$$

Hence the pointwise minimizer requires $v^\perp \equiv 0$. Parameterizing $\psi(x, A) = c(x)^\top \Phi(A)$ with $c(x) \in \mathbb{R}^p$, the objective reduces to solving a conditioned least-squares problem at each fixed $x \in \mathbb{R}^d$.

Step 2: Conditioned Gram matrix and normal equations. Under the conditional measure $P_{A|X_\sigma=x}$ from Lemma 1, the Gram matrix of the basis $\Phi(A)$ is

$$\mathbb{E}\left[\Phi(A)\Phi(A)^\top \mid X_\sigma = x\right] = \mathbb{E}_{P_A}\left[\Phi(A)\Phi(A)^\top\big(1 + f(A)H_\sigma(x)\big)\right] = I_p + H_\sigma(x)T. \tag{26}$$

For any unit vector $u \in \mathbb{R}^p$, $|u^\top T u| = |\mathbb{E}_{P_A}[f(A)(u^\top\Phi(A))^2]| \le \mathbb{E}_{P_A}[(u^\top\Phi(A))^2] = 1$, so $\|T\|_{\mathrm{op}} \le 1$ and $T^2 \preceq I_p$. Since $|H_\sigma(x)| < 1$ for all $\sigma \in (0, 1)$, $I_p + H_\sigma(x)T \succ 0$ is strictly positive definite and invertible.

Next, consider the cross-moment with the target scalar $\zeta(x, A) = \frac{f(A)}{1+f(A)H_\sigma(x)}$:

$$\begin{aligned}\mathbb{E}\left[\Phi(A)\zeta(x, A) \mid X_\sigma = x\right] &= \mathbb{E}_{P_A}\left[\Phi(A)\left(\frac{f(A)}{1 + f(A)H_\sigma(x)}\right)\big(1 + f(A)H_\sigma(x)\big)\right] \\ &= \mathbb{E}_{P_A}\left[f(A)\Phi(A)\right] = a.\end{aligned} \tag{27}$$

The denominator of $\zeta(x, A)$ cancels the Radon–Nikodym density ratio exactly. The normal equations for $c(x)$ are thus:

$$(I_p + H_\sigma(x)T)c^\star(x) = a \implies c^\star(x) = (I_p + H_\sigma(x)T)^{-1}a. \tag{28}$$

Step 3: Pointwise residual and coordinate-reflection symmetrization. The conditional second moment of the target scalar is:

$$\begin{aligned}\mathbb{E}\left[\zeta(x, A)^2 \mid X_\sigma = x\right] &= \mathbb{E}_{P_A}\left[\frac{f(A)^2}{1 + f(A)H_\sigma(x)}\right] \\ &= \frac{1}{2}\left[\frac{1}{1 + H_\sigma(x)} + \frac{1}{1 - H_\sigma(x)}\right] = \frac{1}{1 - H_\sigma(x)^2},\end{aligned} \tag{29}$$

where we used $f(A)^2 = 1$ and $\Pr(f(A) = \pm 1) = 1/2$. Subtracting the variance explained by the projection, $a^\top c^\star(x) = a^\top(I_p + H_\sigma(x)T)^{-1}a$, and scaling by $\|w(x)\|^2 = b_\sigma^2\|\nabla H_\sigma(x)\|^2$, the pointwise conditional risk given $X_\sigma = x$ is

$$\mathbb{E}\left[\|v^\star(x, A) - v_\sigma^Y(x)\|^2 \mid X_\sigma = x\right] = \ell_\sigma(x)\left[1 - \big(1 - H_\sigma(x)^2\big)a^\top(I_p + H_\sigma(x)T)^{-1}a\right]. \tag{30}$$

We now integrate over $x \sim \mu_\sigma = \mathcal{N}(0, q_\sigma I_d)$. Consider the single-coordinate reflection $\tau : \mathbb{R}^d \to \mathbb{R}^d$ given by $\tau(x_1, x_2, \ldots, x_d) = (-x_1, x_2, \ldots, x_d)$. The spherical Gaussian measure μ_σ is invariant under τ. For the sign-product function $H_\sigma(x) = \prod_{i=1}^k[2\Phi(\frac{1-\sigma}{\sigma\sqrt{q_\sigma}}x_i) - 1]$, the first factor changes sign while all other factors remain unchanged. Thus, for any $k \ge 1$:

$$H_\sigma(\tau(x)) = -H_\sigma(x), \qquad H_\sigma(\tau(x))^2 = H_\sigma(x)^2, \qquad \ell_\sigma(\tau(x)) = \ell_\sigma(x). \tag{31}$$

Pairing x and $\tau(x)$ in the expectation symmetrizes the matrix inverse:

$$\frac{1}{2}\left[(I_p + H_\sigma(x)T)^{-1} + (I_p - H_\sigma(x)T)^{-1}\right] = (I_p - H_\sigma(x)^2T^2)^{-1}, \tag{32}$$

which follows from $(I+HT)^{-1}+(I-HT)^{-1}=(I-HT)(I-H^2T^2)^{-1}+(I+HT)(I-H^2T^2)^{-1}=2(I-H^2T^2)^{-1}$. Integrating over μ_σ gives the exact risk formula (22).

Step 4: Operator inequality, universal lower bound, and sharp endpoints. Because T is symmetric with $\|T\|_{\mathrm{op}}\le 1$, we have $0\preceq T^2\preceq I_p$. Since $H_\sigma(X)^2\in[0,1)$,

$$I_p-H_\sigma(X)^2T^2\succeq\big(1-H_\sigma(X)^2\big)I_p\succ 0. \tag{33}$$

Inverting both sides in the Loewner order yields $(I_p-H_\sigma(X)^2T^2)^{-1}\preceq\frac{1}{1-H_\sigma(X)^2}I_p$. Multiplying by $1-H_\sigma(X)^2>0$ gives:

$$\big(1-H_\sigma(X)^2\big)(I_p-H_\sigma(X)^2T^2)^{-1}\preceq I_p. \tag{34}$$

Therefore, the quadratic form satisfies $\big(1-H_\sigma(X)^2\big)a^\top(I_p-H_\sigma(X)^2T^2)^{-1}a\le\|a\|^2=\mathcal{A}_{\mathcal{H}}^{(2)}$. Substituting this upper bound into (22) yields:

$$R_{\mathcal{H},\sigma}\ge\mathbb{E}_{\mu_\sigma}\left[\ell_\sigma(X)\big(1-\mathcal{A}_{\mathcal{H}}^{(2)}\big)\right]=J_{k,\sigma}\big(1-\mathcal{A}_{\mathcal{H}}^{(2)}\big). \tag{35}$$

For the endpoints:

1. If $\mathcal{A}_{\mathcal{H}}^{(2)}=0$, then $a=0$, which implies $c^\star(x)=0$, so $v^\star(x,A)=v_\sigma^0(x)$ and $R_{\mathcal{H},\sigma}=\mathbb{E}_{\mu_\sigma}[\ell_\sigma(X)]=J_{k,\sigma}$.

2. If $f\in\mathcal{H}$, note that for any Boolean $Y=f(A)\in\{-1,1\}$, $Y^2=1$, so
$$\frac{Y}{1+YH_\sigma(x)}=\frac{Y(1-YH_\sigma(x))}{1-Y^2H_\sigma(x)^2}=\frac{Y-H_\sigma(x)}{1-H_\sigma(x)^2}=\left[\frac{1}{1-H_\sigma(x)^2}\right]f(A)-\left[\frac{H_\sigma(x)}{1-H_\sigma(x)^2}\right]. \tag{36}$$
Substituting this identity into $v_\sigma^Y(x)$ gives:
$$v_\sigma^Y(x)=\left[v_\sigma^0(x)-\frac{H_\sigma(x)}{1-H_\sigma(x)^2}w(x)\right]+\left[\frac{1}{1-H_\sigma(x)^2}w(x)\right]f(A). \tag{37}$$
Because $1\in\mathcal{H}$, the first term (purely spatial measurable drift) is in $\mathcal{V}_{\mathcal{H}}$ (paired with the constant basis function $1\in\mathcal{H}$). The second term is in $\mathcal{V}_{\mathcal{H}}$ because $f\in\mathcal{H}$. Thus $v_\sigma^Y\in\mathcal{V}_{\mathcal{H}}$, attaining zero residual risk $R_{\mathcal{H},\sigma}=0$.

□

Corollary 2 (Restatement of Corollary 1). *Let $A\in\{-1,1\}^m$ be uniform Rademacher attributes and let $Y=f(A)=\prod_{i=1}^m A_i$ be the m-parity function. Define the degree-r interaction subspace ($1\le r<m$) under raw representation as:*

$$\mathcal{H}_r^{\mathrm{raw}}:=\mathrm{span}\left\{\prod_{i\in S}A_i:|S|\le r\right\}. \tag{38}$$

Then:

1. *Under raw prompt representation $F_{\mathrm{raw}}=A$, the answer is statistically fully determined ($A_{\mathrm{stat}}(F_{\mathrm{raw}})=1$). Yet because $\langle f,\prod_{i\in S}A_i\rangle=0$ for all $|S|\le r<m$, $\Pi_{\mathcal{H}_r^{\mathrm{raw}}}f=0$, giving $\mathcal{A}_{\mathcal{H}_r^{\mathrm{raw}}}^{(2)}=0$ and maximum velocity risk:*
$$R_{\mathcal{H}_r^{\mathrm{raw}},\sigma}=J_{k,\sigma}. \tag{39}$$

2. *Under planned representation $F_{\mathrm{plan}}=(A,f(A))$, statistical information is identical with $A_{\mathrm{stat}}(F_{\mathrm{plan}})=1$. A linear readout $\mathcal{H}_1^{\mathrm{plan}}=\mathrm{span}\{1,A_1,\dots,A_m,f(A)\}$ directly accesses f, yielding $\mathcal{A}_{\mathcal{H}_1^{\mathrm{plan}}}^{(2)}=1$ and:*
$$R_{\mathcal{H}_1^{\mathrm{plan}},\sigma}=0. \tag{40}$$

Corollary 3 (Zero answer dependence under uninformative drift). *Suppose $\mathcal{A}_{\mathcal{H}}^{(2)}=0$. Then the population-optimal restricted velocity field is $v^\star(x,A)=v_\sigma^0(x)=\frac{2\sigma-1}{q_\sigma}x$, which has zero functional dependence on A. Assume the sampler accesses the prompt attributes A solely through the evaluated velocity field, with its initial latent $X_0=G\sim\mathcal{N}(0,I_d)$ and any auxiliary sampling randomness independent of A. Then any discrete or continuous sampling trajectory produces a*

terminal latent X_{M_s} whose distribution is completely independent of A and target $Y = f(A)$. Consequently, the generated decision accuracy cannot exceed the unconditioned baseline accuracy $\frac{1}{2}$.

Proof. When $\mathcal{A}^{(2)}_{\mathcal{H}} = 0$, Theorem 4 establishes that the minimizer of the flow-matching objective over $\mathcal{V}_{\mathcal{H}}$ is $v^\star(x, A) = v^0_\sigma(x)$. The continuous flow ODE in σ is $\frac{dX_\sigma}{d\sigma} = v^0_\sigma(X_\sigma)$, which in generative sampling is integrated backward from noise $\sigma = 1$ to clean data $\sigma = 0$ via Euler updates $X_{j+1} = X_j - h_j v^0_{\sigma_j}(X_j)$ (or in time $t = 1 - \sigma$, $\frac{dX_t}{dt} = -v^0_{1-t}(X_t)$). More generally, any discrete numerical integration rule takes the recursive form

$$X_{j+1} = \Psi_j\big(X_0, X_1, \ldots, X_j, v^0_{\sigma_j}, \xi_j\big), \tag{41}$$

where ξ_j denotes auxiliary sampling noise independent of A. Because $X_0 = G$ is independent of A and the transition map Ψ_j depends on the prompt only through v^0 (which contains zero dependence on A), an immediate induction on j establishes that the entire trajectory $(X_0, X_1, \ldots, X_{M_s})$ is statistically independent of the prompt attributes A. Because $Y = f(A)$ is conditionally centered and completely determined by A, the terminal state X_{M_s} is statistically independent of Y, yielding $\Pr(g_k(X_{M_s}) = Y) = \frac{1}{2}$. □

Remark: Flow-matching population objective vs. general multi-step iteration. It is crucial to distinguish the population-level obstruction of flow matching from general multi-step iterative computing. An arbitrary multi-step procedure could in principle compute nonlinear interactions across steps by using intermediate latents as computational memory (e.g., in a 2-step scalar iteration, setting $X_{1/2} = A_1|G|$ and $X_0 = A_2 X_{1/2} = A_1 A_2 |G|$ successfully computes the parity of two bits). However, standard diffusion and flow-matching backbones do not execute arbitrary state machines; they train a neural velocity network $v_\theta(x, \sigma, F)$ against the instantaneous regression objective $\mathbb{E}\|v_\theta - \dot{X}_\sigma\|^2$. When the conditioning interface restricts instantaneous semantic decoding ($\mathcal{A}^{(2)}_{\mathcal{H}} = 0$), the population-optimal field at every step averages over the unresolved attributes to the unguided drift $v^0_\sigma(x)$. Thus, the backbone learns an uninformative velocity field that is completely decoupled from the answer. Integrating this learned field across multiple steps merely executes an unguided generative trajectory, confirming that iterative sampling along the learned field cannot recover from the conditioning bottleneck.

Theoretical implications: Fourier spectrum and multi-hop error compounding. The restricted semantic decoding framework of Theorem 1 yields two direct theoretical implications:

1. *Walsh–Fourier interaction spectrum beyond parity.* For general Boolean targets $f(A)$, the accessibility under degree-r decoders is characterized by the low-degree Fourier weight:
$$\mathcal{A}^{(2)}_{\mathcal{H}_r} = \sum_{|S| \le r} \widehat{f}(S)^2. \tag{42}$$
While parity places all Fourier weight at degree m (yielding $\mathcal{A}^{(2)}_{\mathcal{H}_r} = 0$ for all $r < m$), even smoother Boolean functions exhibit significant lower bounds. For example, for a majority-of-seven rule under degree-1 linear decoding, $\mathcal{A}^{(2)}_{\mathcal{H}_1} = \frac{175}{256} \approx 0.6836$, leaving an ineliminable lower bound $\ge \frac{81}{256} J_{k,\sigma} \approx 0.3164 J_{k,\sigma}$.
2. *Noisy intermediate states and compounding agentic error.* In multi-turn generation (such as Strategy 2: Agentic Painting in Section 5), intermediate visual states are resolved across sequential perceptual hops. Let transmission through an imperfect perceptual hop be modeled by a binary symmetric channel (BSC) with crossover error $e \in (0, 1/2]$. With decoder class $\mathcal{H} = L^2(\sigma(\widetilde{Y}))$, the conditional expectation is $\Pi_{\mathcal{H}} Y = \mathbb{E}[Y \mid \widetilde{Y}] = (1 - 2e)\widetilde{Y}$, so $\mathcal{A}^{(2)}_{\mathcal{H}} = (1 - 2e)^2$. By Theorem 1, the excess velocity risk satisfies:
$$R_{\mathcal{H},\sigma} \ge J_{k,\sigma}\big(1 - (1 - 2e)^2\big) = 4e(1 - e) J_{k,\sigma}. \tag{43}$$
When $n - 1$ independent noisy hops are composed sequentially, the compound crossover error rate becomes $e_n = \frac{1 - (1-2e)^{n-1}}{2}$. Consequently, $\mathcal{A}^{(2)}_{\mathcal{H}} \to 0$ and $R_{\mathcal{H},\sigma} \to J_{k,\sigma}$ exponentially fast as the chain length n grows, providing the formal basis for why purely iterative perceptual hopping without upstream planning degrades multiplicatively.

B.2 ADAPTIVE-COMPUTE CONSTRAINT: THEOREM 2

Proposition 2 (DAG schedule lower bound). *For a specified operation DAG with N vertices, program span L, and execution width B, any valid schedule requires*

$$R \geq \max\left\{ \left\lceil \frac{N}{B} \right\rceil, L \right\}. \tag{44}$$

Proof. In any round r, at most B operations can be evaluated, so $R \cdot B \geq N$, giving $R \geq \lceil N/B \rceil$. Along any directed dependency path of length L, the i-th node cannot begin until the $(i-1)$-th node completes; thus, at most one node from this path can be evaluated in any round, requiring $R \geq L$. □

Proposition 3 (Brent–Graham schedule upper bound). *A valid schedule exists satisfying $R \leq \lceil N/B \rceil + L - 1$.*

Proof. Let $N_i \geq 1$ be the number of nodes at topological level $i \in \{1, \ldots, L\}$, so that $\sum_{i=1}^{L} N_i = N$. A greedy level-by-level schedule executes level i in $\lceil N_i/B \rceil$ rounds. Using the integer identity $\lceil n/B \rceil = 1 + \lfloor (n-1)/B \rfloor$ for any integer $n \geq 1$:

$$R = \sum_{i=1}^{L} \left\lceil \frac{N_i}{B} \right\rceil = L + \sum_{i=1}^{L} \left\lfloor \frac{N_i - 1}{B} \right\rfloor \leq L + \left\lfloor \frac{\sum_{i=1}^{L}(N_i - 1)}{B} \right\rfloor = L + \left\lfloor \frac{N-L}{B} \right\rfloor \leq L + \left\lceil \frac{N}{B} \right\rceil - 1, \tag{45}$$

where the final inequality holds because $\frac{N-L}{B} \leq \frac{N-1}{B} < \lceil \frac{N}{B} \rceil$ for $L \geq 1$ and the floor function maps to integers. □

Random-oracle chain and Proof of Theorem 2. Let $M \geq 2$. Let $\theta_0 \in [M]$ be fixed and public. For $i = 1, \ldots, L-1$, draw independent random functions $O_i : [M] \to [M]$ uniformly from all M^M functions. Draw $O_L : [M] \to \{-1, 1\}$ uniformly from $\{-1, 1\}^M$. The target label is $Y = O_L(\theta_{L-1})$, where $\theta_i = O_i(\theta_{i-1})$.

Theorem 5 (Restatement of Theorem 2). *Any querying algorithm with $R < L$ adaptive rounds and at most Q total queries across all oracles satisfies*

$$\Pr(\widehat{Y} = Y) \leq \frac{1}{2} + \frac{Q}{2M}. \tag{46}$$

Proof. We construct a coupling between the true oracle system and a round-dependent ideal simulation.

1. True system. Sample independent random variables $\theta_1, \ldots, \theta_{L-1} \sim \mathrm{Unif}([M])$ and $Y \sim \mathrm{Unif}(\{-1, 1\})$, and set $\theta_L = Y$. Independently draw random fallback tables $B_1, \ldots, B_L$, where each $B_i : [M] \to [M]$ $(i < L)$ is drawn uniformly from M^M and $B_L : [M] \to \{-1, 1\}$ uniformly from $\{-1, 1\}^M$. The true oracles are defined by:

$$O_i(x) = \begin{cases} \theta_i & \text{if } x = \theta_{i-1}, \\ B_i(x) & \text{otherwise,} \end{cases} \qquad (i = 1, \ldots, L). \tag{47}$$

Because the marginal distribution of $O_i(x)$ is uniform and independent across all $x \in [M]$, this matches the true random-oracle chain identically.

2. Round-dependent ideal simulator. In each adaptive round $r \in \{1, \ldots, R\}$, the ideal simulator answers any query x to oracle index i by:

$$\widetilde{O}_i^{(r)}(x) = \begin{cases} \theta_i & \text{if } i \leq r \text{ and } x = \theta_{i-1}, \\ B_i(x) & \text{otherwise.} \end{cases} \tag{48}$$

Crucially, for all future levels $i > r$, the round-r simulator ignores θ_i entirely and returns only $B_i(x)$.

3. Hit probability per query. Suppose that in round r, the querying algorithm submits an evaluation query x to oracle index i with $i > r$. The query address x is a measurable function of the algorithm's internal coin flips and the simulator responses received in earlier rounds $1, \ldots, r-1$. By construction of $\widetilde{O}^{(\ell)}$ for $\ell < r$, all prior responses depend solely on the fallback tables $B_1, \ldots, B_L$ and the prefix

chain $(\theta_1, \ldots, \theta_{r-1})$. However, because $i > r$, we have $i - 1 \geq r$. The target active address θ_{i-1} is drawn uniformly from $[M]$ independently of $(B_1, \ldots, B_L, \theta_1, \ldots, \theta_{r-1})$ and the algorithm's coin flips. Consequently, unconditionally:

$$\Pr(x = \theta_{i-1}) = \frac{1}{M}. \tag{49}$$

4. Bad event and union bound. Define the bad event $\mathcal{B}$ as the event that during the execution of the algorithm under the ideal simulator, the algorithm queries some oracle index $i > r$ at address $x = \theta_{i-1}$ in some round $r \in \{1, \ldots, R\}$. Since the total number of queries issued across all R rounds is at most Q, applying the union bound over all queries gives:

$$\Pr(\mathcal{B}) \leq \frac{Q}{M}. \tag{50}$$

5. Coupling of executions. On the complementary event $\mathcal{B}^c$, no query in any round r with $i > r$ evaluates address $x = \theta_{i-1}$. For any such query, both the true oracle $O_i(x)$ and the simulator $\widetilde{O}_i^{(r)}(x)$ evaluate to $B_i(x)$. For all queries with $i \leq r$, both $O_i(x)$ and $\widetilde{O}_i^{(r)}(x)$ agree by definition. Therefore, on $\mathcal{B}^c$, the answers returned by the ideal simulator are identically equal to the answers that would be returned by the true oracles $(O_1, \ldots, O_L)$. By induction over rounds, the algorithm's full query transcript and its terminal prediction $\widehat{Y}$ in the true system coincide identically with those under the ideal simulation on $\mathcal{B}^c$.

6. Statistical independence and conclusion. Because $R < L$, the condition $i \leq r \leq R < L$ implies that the ideal simulator never queries or evaluates the terminal entry $\theta_L = Y$ at any point in any round. Hence, in the ideal simulation, the generated transcript, the algorithm's final output $\widehat{Y}$, and the occurrence of the bad event $\mathcal{B}$ are jointly statistically independent of $Y \sim \mathrm{Unif}(\{-1, 1\})$. It follows that $\Pr(\widehat{Y} = Y, \mathcal{B}^c) = \Pr(\widehat{Y} = Y \mid \mathcal{B}^c) \Pr(\mathcal{B}^c) = \frac{1}{2} \Pr(\mathcal{B}^c)$. We conclude:

$$\begin{aligned} \Pr(\widehat{Y} = Y) &\leq \Pr(\widehat{Y} = Y, \mathcal{B}^c) + \Pr(\mathcal{B}) \\ &= \frac{1}{2} \Pr(\mathcal{B}^c) + \Pr(\mathcal{B}) = \frac{1}{2}(1 - \Pr(\mathcal{B})) + \Pr(\mathcal{B}) = \frac{1}{2} + \frac{\Pr(\mathcal{B})}{2} \leq \frac{1}{2} + \frac{Q}{2M}. \end{aligned} \tag{51}$$

□

Proposition 4 (Restatement of Proposition 1). *Consider any visual generation algorithm $\widehat{D} = \Gamma^{O_1, \ldots, O_L}(\theta_0, G)$ querying the oracle chain $(O_1, \ldots, O_L)$ using at most R adaptive rounds and at most Q total queries across all pipeline components (including intermediate text representations $F = \Phi^{O_1, \ldots, O_L}(\theta_0)$ and denoiser passes). If $\widehat{D}$ achieves generation accuracy $p := \Pr(g_k(\widehat{D}) = Y)$, then the induced decision decoder $\widehat{Y} := g_k(\widehat{D})$ achieves identical success rate $\Pr(\widehat{Y} = Y) = p$ within the same (R, Q) query budget.*

Proof. Let $\Gamma^{O_1, \ldots, O_L}$ be any visual generation pipeline taking public address θ_0 and initial Gaussian noise $G \sim \mathcal{N}(0, I_d)$, querying the oracle chain within R adaptive rounds and Q total queries to produce a generated image $\widehat{D} = \Gamma^{O_1, \ldots, O_L}(\theta_0, G) \in \mathbb{R}^d$.

Define the Boolean decision decoder $\widehat{Y} := g_k(\widehat{D}) = \prod_{i=1}^{k} \mathrm{sign}(\widehat{D}_i) \in \{-1, 1\}$. Because $g_k : \mathbb{R}^d \to \{-1, 1\}$ is a deterministic spatial post-processing functional evaluated solely on $\widehat{D}$, computing $\widehat{Y}$ requires zero additional oracle queries and zero additional adaptive rounds. Hence, $\widehat{Y}$ is a valid query algorithm operating under the identical budget of R rounds and Q queries. By definition, $\Pr(\widehat{Y} = Y) = \Pr(g_k(\widehat{D}) = Y) = p$.

When $R < L$, Theorem 5 immediately implies that any such querying algorithm must satisfy $\Pr(\widehat{Y} = Y) \leq \frac{1}{2} + \frac{Q}{2M}$. Consequently:

$$p = \Pr\big(g_k(\widehat{D}) = Y\big) \leq \frac{1}{2} + \frac{Q}{2M}. \tag{52}$$

This establishes that expanding the task domain from discrete Boolean classification to continuous image generation cannot circumvent the serial computational lower bound of the underlying decision problem. □

Continuous score-risk corollary. To show how discrete uncertainty translates into score risk, consider a single-coordinate continuous renderer where clean data $X_0 = (\gamma Y, V)$ with $V \sim \mathcal{N}(0, \tau^2 I_{d-1})$. At diffusion time t, noisy coordinate $X_{t,1} = \mu_t Y + \bar{\sigma}_t Z$ with independent Gaussian noise $Z \sim \mathcal{N}(0,1)$ and signal-to-noise ratio $\varrho_t = \mu_t/\bar{\sigma}_t$. The Bayes posterior variance of Y from $X_{t,1}$ alone is

$$c(\varrho_t) := \mathbb{E}_{Y,Z}\left[\operatorname{sech}^2\left(\varrho_t(\varrho_t Y + Z)\right)\right]. \tag{53}$$

We formalize the information flow for score estimation under an incomplete oracle search: the query algorithm first executes $R < L$ rounds using public address θ_0 and independent algorithmic randomness to produce query transcript T (prior to, and independently of, the realization of diffusion noise Z). The score estimator $\widehat{s}_{t,1}(X_t, T)$ operates on (X_t, T) without direct access to label Y. Under this specification, on the non-collision event $\mathcal{B}^c$, the transcript T provides no information about Y, so conditioning on T does not reduce the posterior uncertainty of Y given $X_{t,1}$. Therefore,

$$\mathbb{E}\left[(\widehat{s}_{t,1} - s^\star_{t,1})^2 \mid \mathcal{B}^c\right] \geq \beta_t^2 c(\varrho_t), \qquad \beta_t := \frac{\mu_t}{\bar{\sigma}_t^2}. \tag{54}$$

Because $\Pr(\mathcal{B}^c) \geq 1 - \frac{Q}{M}$ and $0 \leq c(\varrho_t) \leq 1$, taking total expectations yields:

$$\mathbb{E}\left[(\widehat{s}_{t,1} - s^\star_{t,1})^2\right] \geq \Pr(\mathcal{B}^c)\beta_t^2 c(\varrho_t) \geq \beta_t^2 \left[c(\varrho_t) - \frac{Q}{M}\right]_+. \tag{55}$$

B.3 Timing constraint: Theorem 3

Setting and Controllability Gramian. Consider the discrete Euler noise grid $1 = \sigma_0 > \sigma_1 > \cdots > \sigma_{M_s} = 0$ with $M_s \geq 2$ and step sizes $h_j = \sigma_j - \sigma_{j+1} > 0$. Before disclosure of the target condition ($j < J$), the generation process integrates the population-optimal velocity field under inaccessible conditioning $\mathcal{A}^{(2)}_{\mathcal{H}} = 0$, which by Theorem 1 is the unguided drift $v^0_\sigma(x) = \frac{2\sigma - 1}{q_\sigma} x$. The unguided Euler step is $X_{j+1} = \omega_j X_j$, where

$$\omega_j := 1 + h_j \frac{1 - 2\sigma_j}{q_{\sigma_j}}. \tag{56}$$

At update step J and subsequent steps ($j \geq J$), the target condition Y is disclosed. Suffix controls $u_j = u_j(G, Y)$ may depend jointly on the initial noise sample G and target Y, producing the controlled update $X_{j+1} = \omega_j X_j + h_j u_j(G, Y)$.

Define the total reference magnification $P := \prod_{j=0}^{M_s-1} \omega_j$, suffix factors $\pi_j := \prod_{\ell > j} \omega_\ell$ (with $\pi_{M_s-1} = 1$), and remaining controllability Gramian $W_J := \sum_{j \geq J} h_j \pi_j^2$.

Lemma 2 (Prefix positivity). *If $h_j \leq \sigma_j$ for all j, then $\omega_j > 0$ for all j. Consequently, $P > 0$ and $\pi_j > 0$ for all j.*

Proof. For $\sigma_j \leq 1/2$, $1 - 2\sigma_j \geq 0$, so $\omega_j \geq 1 > 0$. For $\sigma_j > 1/2$, $2\sigma_j - 1 > 0$. Using $h_j \leq \sigma_j$:

$$\omega_j = 1 - h_j \frac{2\sigma_j - 1}{q_{\sigma_j}} \geq 1 - \sigma_j \frac{2\sigma_j - 1}{(1-\sigma_j)^2 + \sigma_j^2} = \frac{(1-\sigma_j)^2 + \sigma_j^2 - 2\sigma_j^2 + \sigma_j}{q_{\sigma_j}} = \frac{1 - \sigma_j}{q_{\sigma_j}} > 0. \tag{57}$$

At the boundary $\sigma_0 = 1$, $\omega_0 = 1 - h_0 = \sigma_1 > 0$ because the grid has an interior point. □

Theorem 6 (Restatement of Theorem 3). *Let $X_0 = G \sim \mathcal{N}(0, I_d)$ and let the condition be disclosed at update J. For target accuracy $a \in (1/2, 1]$, the infimum correction energy $\mathcal{E}_J = \mathbb{E}\sum_{j \geq J} h_j \|u_j(G, Y)\|^2$ is*

$$\inf \mathcal{E}_J = \frac{P^2\, \Theta_k(a)}{2\, W_J}, \qquad \Theta_k(a) := \int_0^{2a-1} F^{-1}_{\Delta_k}(r)\, dr, \tag{58}$$

where $\Delta_k = \min_{i \leq k} G_i^2$. Because $h_j \pi_j^2 > 0$ for all j, W_J strictly decreases in J, strictly increasing the minimum energy.

Proof. Unrolling the linear recursion from step J to the terminal step M_s:

$$X_{M_s} = PG + \sum_{j=J}^{M_s-1} h_j \pi_j u_j(G, Y) = PG + \delta(G, Y), \qquad \delta(G, Y) := \sum_{j=J}^{M_s-1} h_j \pi_j u_j(G, Y). \tag{59}$$

Applying the Cauchy–Schwarz inequality:

$$\|\delta(G,Y)\|^2 = \left\| \sum_{j=J}^{M_s-1} \sqrt{h_j}\pi_j \cdot \sqrt{h_j} u_j(G,Y) \right\|^2$$
$$\leq \left(\sum_{j=J}^{M_s-1} h_j \pi_j^2 \right) \left(\sum_{j=J}^{M_s-1} h_j \|u_j(G,Y)\|^2 \right) = W_J \sum_{j=J}^{M_s-1} h_j \|u_j(G,Y)\|^2. \quad (60)$$

Equality holds if and only if $u_j(G,Y) = \frac{\pi_j}{W_J}\delta(G,Y)$ for all $j \geq J$. Taking expectations, any control sequence achieving net displacement $\delta(G,Y)$ requires energy:

$$\mathcal{E}_J \geq \frac{\mathbb{E}\|\delta(G,Y)\|^2}{W_J}. \quad (61)$$

Without control ($\delta = 0$), the terminal sign product is $g_k(PG) = \prod_{i=1}^k \operatorname{sign}(PG_i) = g_k(G)$ (since $P > 0$ by Lemma 2), which is independent of Y, yielding baseline accuracy $\Pr(g_k(PG) = Y) = 1/2$. Let $E = \{g_k(G) \neq Y\}$ denote the error event, which has probability $\Pr(E) = 1/2$. To achieve final target accuracy $a \in (1/2, 1]$, the perturbation $\delta(G,Y)$ must flip the sign product $g_k(PG + \delta(G,Y))$ to equal Y on a sub-event $E_a \subseteq E$ of measure $\Pr(E_a) = a - 1/2$.

For each sample in E, flipping the sign product requires changing the sign of at least one coordinate $i \in \{1, \ldots, k\}$. Changing the sign of coordinate i from $\operatorname{sign}(PG_i)$ requires $|\delta_i(G,Y)| > P|G_i|$ in the opposing direction. To minimize the Euclidean norm $\|\delta(G,Y)\|^2$, it is strictly optimal to perturb only the coordinate with the smallest magnitude:

$$i_\star = \arg\min_{i \leq k} G_i^2, \qquad \Delta_k := \min_{i \leq k} G_i^2 = G_{i_\star}^2. \quad (62)$$

To achieve total mass $a - 1/2$ at minimal expected displacement $\mathbb{E}\|\delta\|^2$, the controller greedily selects the samples in E having the smallest values of Δ_k, namely $E_a = E \cap \{\Delta_k \leq \tau_a\}$, where $\tau_a := F_{\Delta_k}^{-1}(2a-1)$.

To establish that the infimum energy is exactly $\frac{P^2\Theta_k(a)}{2W_J}$, consider for any $\varepsilon > 0$ the approximating perturbation:

$$\delta_\varepsilon(G,Y) := -P(1+\varepsilon)G_{i_\star} e_{i_\star} \mathbf{1}_E \mathbf{1}_{\{\Delta_k \leq \tau_a\}}. \quad (63)$$

For every sample in E_a, coordinate $i_\star$ of the terminal latent becomes:

$$PG_{i_\star} + (\delta_\varepsilon)_{i_\star} = PG_{i_\star} - P(1+\varepsilon)G_{i_\star} = -\varepsilon PG_{i_\star}, \quad (64)$$

which strictly reverses its sign because $\varepsilon > 0$ and $P > 0$. Since all other coordinates remain unchanged, $g_k(PG + \delta_\varepsilon) = -g_k(PG) = Y$ on E_a, achieving exact success probability $\frac{1}{2} + \Pr(E_a) = a$. The expected squared displacement is:

$$\mathbb{E}\|\delta_\varepsilon(G,Y)\|^2 = (1+\varepsilon)^2 P^2 \mathbb{E}\left[\Delta_k \mathbf{1}_E \mathbf{1}_{\{\Delta_k \leq \tau_a\}}\right] = (1+\varepsilon)^2 P^2 \Pr(E) \mathbb{E}\left[\Delta_k \mathbf{1}_{\{\Delta_k \leq \tau_a\}}\right]$$
$$= (1+\varepsilon)^2 \frac{P^2}{2} \int_0^{2a-1} F_{\Delta_k}^{-1}(r)\, dr = (1+\varepsilon)^2 \frac{P^2}{2} \Theta_k(a). \quad (65)$$

Setting $u_{j,\varepsilon}(G,Y) = \frac{\pi_j}{W_J}\delta_\varepsilon(G,Y)$ attains equality in (61), yielding energy $\mathcal{E}_{J,\varepsilon} = (1+\varepsilon)^2 \frac{P^2\Theta_k(a)}{2W_J}$. Taking $\varepsilon \downarrow 0$, $\mathcal{E}_{J,\varepsilon} \to \frac{P^2\Theta_k(a)}{2W_J}$.

Conversely, any control achieving accuracy a must have $\|\delta(G,Y)\|^2 \geq P^2\Delta_k$ almost everywhere on an error subset of measure $a - 1/2$, so $\mathbb{E}\|\delta(G,Y)\|^2 \geq \frac{P^2}{2}\Theta_k(a)$. Dividing by W_J proves that $\frac{P^2\Theta_k(a)}{2W_J}$ is the exact infimum. □

Continuous-time limit and energy divergence. As the Euler mesh size $\max h_j \to 0$, $P \to 1$ and $\pi_j^2 \to 1/q_\sigma$. The Gramian sum converges to the continuous integral:

$$W(s) = \int_0^s \frac{dr}{(1-r)^2 + r^2} = [\arctan(2r-1)]_0^s = \arctan(2s-1) + \frac{\pi}{4}. \quad (66)$$

Because $W(s) = s + \mathcal{O}(s^2)$ as $s \downarrow 0$, the continuous-limit infimum correction energy satisfies

$$\mathcal{E}_{\min}(s) = \frac{\Theta_k(a)}{2W(s)} \sim \frac{\Theta_k(a)}{2s} \quad \text{as } s \downarrow 0, \quad (67)$$

diverging hyperbolically as disclosure is delayed toward the clean endpoint.

Implications for schedule allocation. Theorem 6 establishes that within a fixed discretization grid and linear corrective dynamics, earlier disclosure strictly reduces the required steering energy ($K_J(a) \propto 1/W_J$). Under a finite control energy budget $\mathcal{E} \leq B$, achieving target accuracy a imposes distinct necessary and sufficient disclosure deadlines:

1. *Necessary condition (infeasibility deadline).* For any budget $B < K_J(a) = \frac{P^2\Theta_k(a)}{2W_J}$, target accuracy a is strictly unattainable. Thus, a necessary condition for feasibility is $W_J \geq \frac{P^2\Theta_k(a)}{2B}$. Whenever the set $\{J : W_J \geq \frac{P^2\Theta_k(a)}{2B}\}$ is non-empty, we define the latest admissible disclosure index $J_{\max} := \max\{J : W_J \geq \frac{P^2\Theta_k(a)}{2B}\}$. Delaying disclosure past $J_{\max}$ renders accuracy a strictly infeasible under budget B. (If the set is empty, accuracy a cannot be attained at any update index).
2. *Sufficient condition.* Conversely, for any strictly larger budget $B > K_J(a)$ (or equivalently, whenever $W_J > \frac{P^2\Theta_k(a)}{2B}$), an admissible suffix control achieving accuracy a with energy at most B is guaranteed to exist (by taking $\varepsilon > 0$ sufficiently small in the construction of $\mathcal{E}_{J,\varepsilon}$ above). At the exact boundary $B = K_J(a)$, feasibility depends on whether the infimum is attained in the underlying function space.

As disclosure approaches the clean endpoint ($s \downarrow 0$), $W(s) \sim s$ drives the infimum steering energy to diverge as $\frac{\Theta_k(a)}{2s}$, requiring unbounded corrective amplitudes that exceed any pre-specified finite energy budget.

C Comprehensive Experimental Protocols

C.1 FLUX configuration

We use FLUX.2-Klein-base-4B (Black Forest Labs, 2026) at 512×512 resolution with BF16 weights. Conditioning tensors concatenate hidden layers 9, 18, and 27 of the Qwen3 text encoder. At 512^2, images comprise 1,024 spatial tokens with 128 latent channels ($d = 131{,}072$ latent dimensions). The default sampler uses 28 Euler updates with classifier-free guidance 4.0. Compute sweeps include 8, 16, 28, 50, and 100 Euler steps.

C.2 Flow-control tasks, interventions, and scoring

Flow-control experiments use 3,200 generated two-parity tasks. All intervention conditions share identical initial noise $X_0 \sim \mathcal{N}(0, I_d)$. After disclosure at update $J \in \{0, 7, 14, 21\}$, the applied velocity is

$$v_j = v_I + \lambda(v_E - v_I), \tag{68}$$

where v_I and v_E are velocity fields evaluated under unresolved and resolved prompt representations, and $\lambda \in \{1, 2, 4\}$. A continuous opponent-colour score maps RGB values into target-versus-competitor margin.

C.3 SynBool tasks, scorer, and planner

Each task samples four candidate colours without replacement and one target geometric shape (circle, square, diamond, or triangle). Boolean gates are drawn from $\{\mathrm{AND}, \mathrm{OR}, \mathrm{XOR}, \mathrm{NOT}, \mathrm{NAND}\}$ and compiled into programs of 3, 7, 15, or 31 operations. The five evaluated prompt variants are:

1. **Unresolved:** full instruction and colour table, no terminal answer.
2. **Answer-supplied:** full instruction followed by an explicit declaration of the selected colour.
3. **Answer-plus-distractor:** the correct target colour embedded in non-causal distractor text of comparable length.
4. **Minimal:** only the visual target, e.g., "A solid blue circle on a plain white background."
5. **Wrong-answer:** the unresolved instruction followed by an explicit but incorrect terminal colour.

The primary colour scorer converts pixels to HSV, retaining foreground pixels with saturation $S \geq 0.22$ and value $V \geq 0.18$. Each pixel is assigned to the nearest candidate hue, and majority vote yields predicted index $\widehat{Y}$. Strict scoring additionally requires hue purity ≥ 0.80 and chromatic coverage between 0.03 and 0.85 of the total canvas.

C.4 REPRESENTATION PROBING

We evaluate 5,000 tasks per depth (1,500 at 3 operations). FLUX Qwen3 representations are extracted from layers 0–36 with masked-mean and last-token pooling, and from the concatenation of layers 9, 18, and 27. Qwen-Image uses the final Qwen2.5-VL text representation after discarding template tokens. Three probe families use a 70/10/20 train/val/test split: linear softmax, a two-layer MLP with hidden width 512/ReLU/dropout 0.1, and a four-layer Transformer encoder probe with width 512, 8 heads, and attention pooling. Tasks are stratified by *needed-suffix length* s, the minimal trailing Boolean-gate depth needed to determine the target.

C.5 RELATIONAL GRIDS

SYNGRID-2×2 contains four cells and chains of 2–4 objects; SYNGRID-3×3 contains nine cells and chains of 2–5 objects (1,000 tasks evaluated per benchmark). Object 1 is anchored at a named cell, and every subsequent object is specified relative to its predecessor (above, below, left of, or right of). Exact placement requires every object to occupy its designated cell.

C.6 MULTI-ROUND AGENTIC PAINTING

In Agentic Painting, a VLM (Qwen2.5-VL-7B) receives the current canvas and overall relational instruction, identifies existing objects, resolves the next requested placement, and issues one structured edit action. The image-conditioned renderer returns a new canvas, and the loop repeats. In PLAN+PAINT, the LLM planner first resolves the entire relation sequence; subsequent edit calls execute that fixed plan, eliminating repeated visual re-inference. An oracle-step condition supplies ground-truth grid states to estimate raw edit capacity. Figure 3 illustrates the architectural blueprints of the three evaluated paradigms.

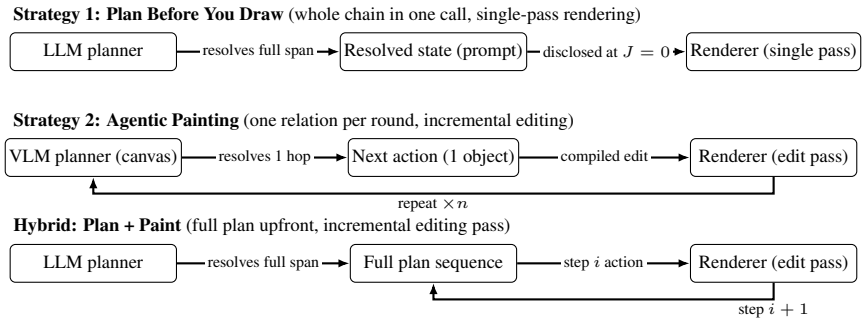


Figure 3: Allocation paradigms. *Plan Before You Draw* resolves the full chain upfront in one LLM call and renders once. *Agentic Painting* uses a visual agent to perceive the current canvas, resolve one relation, and call an editing renderer repeatedly. PLAN+PAINT resolves the full plan upfront and executes the plan through incremental rendering passes.

C.7 NATURAL-IMAGE VALIDATION ON GENEVAL

The GenEval position subset contains 400 two-object prompts. Scoring follows the standard pipeline such that Mask2Former detects instances, and a centroid relation test evaluates spatial adherence.

C.8 NATURAL-IMAGE VALIDATION ON T2I-REASONBENCH

We evaluate 200 prompts from the entity split. Each unresolved prompt describes an entity indirectly; an explicit paraphrase naming the entity serves as an answer-supplied oracle. FLUX.2-Klein

renders every arm. Qwen3-VL-8B-Instruct is used as planner/editor in the corresponding arms. Automated evaluation is conducted with Gemini 3.1 Pro as the multimodal judge following the benchmark rubric.

D COMPLETE EXPERIMENTAL RESULTS

D.1 ACCURACY ACROSS ALL SYNBOOL CELLS

Figure 4 reports four-choice colour accuracy across the full grid of depths, sampling budgets, renderers, and controls; it is the appendix companion to Figure 1 in the main text and uses the same values in every shared cell.

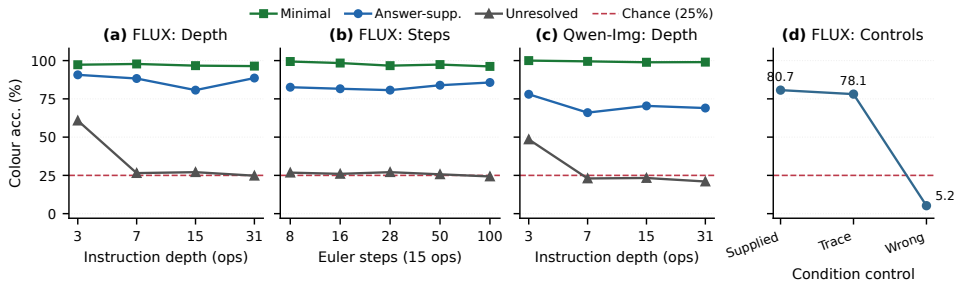


Figure 4: Four-choice colour accuracy (%) across depth, Euler-step budgets, a second renderer, and answer controls (the appendix companion to Figure 1). (a) Unresolved accuracy is elevated on shallow 3-operation programs (60.8%) but collapses to the 25% chance level by 7 operations and beyond, while the answer-supplied and minimal conditions stay high. (b) Scaling the sampling budget from 8 to 100 Euler steps yields no systematic improvement (unresolved accuracy stays near 24.8–27.1%). (c) A second renderer (Qwen-Image) reproduces the near-chance failure at 15 operations. (d) Supplying the terminal answer or a full derivation trace restores high accuracy, whereas a deliberately wrong answer reduces ground-truth accuracy to 5.2%, with the injected false colour faithfully rendered on 86.8% of tasks.

D.2 PROBING SEMANTIC AVAILABILITY IN CONDITIONING REPRESENTATIONS

Table 2 provides the complete numerical values for the probing and generative accuracy measurements visualized in Figure 2 of the main text. While probe decodability declines sharply with needed serial depth, rendered unresolved accuracy drops from 0.646 to chance levels (0.189–0.210), consistent with the generative backbone failing to execute multi-step dependencies without explicit prompting.

At $s \geq 7$, probe accuracy approaches 0.50, while unresolved rendering accuracy drops to chance levels (0.189–0.210, overall 0.271). Crucially, even on shallow strata ($s = 1$–2) where text representations achieve 0.927 decodability, the unguided backbone renders the target with only 0.646 accuracy (rising to 0.927 when the answer is explicitly supplied), confirming that representation decodability does not translate directly to reliable visual execution without explicit prompting.

D.3 DISCLOSURE–GAIN INTERVENTIONS

Figures 5 and 6 present results for the 3,200-task paired-noise intervention experiment.

D.4 SINGLE-UPDATE RESPONSES

Table 3 reports the observed response to single-update condition perturbations at different points along the 28-step trajectory.

Table 2: Decodability of answer Y from text conditioning and rendered image accuracy across needed-suffix strata s (four-layer Transformer sequence probe; corresponding to Figure 2 in the main text). Evaluated across 1,000 SYNBOOL tasks naturally partitioned by circuit topology into needed-suffix strata s. Accuracy falls to 0.189–0.210 for $s \geq 5$; overall unresolved accuracy is $271/1{,}000 = 0.271$, whereas explicit answers yield $807/1{,}000 = 0.807$. The "All" column reports the overall sample micro-average $\sum_s \text{correct}_s/1{,}000$.

Metric / Setting	$s = 1$–2	$s = 3$	$s = 4$	$s = 5$–6	$s \geq 7$	**All**
FLUX text-conditioning probes						
7 ops	0.976	0.814	0.640	0.578	0.512	0.598
15 ops	0.927	0.791	0.573	0.546	0.481	0.563
31 ops	0.976	0.837	0.640	0.535	0.491	0.580
Qwen-Image text-conditioning probe						
15 ops	0.939	0.802	0.587	0.551	0.495	0.575
Rendered image accuracy (FLUX, 15 ops)						
Unresolved	0.646	0.430	0.347	0.189	0.210	0.271
Answer-supplied	0.927	0.849	0.773	0.746	0.808	0.807

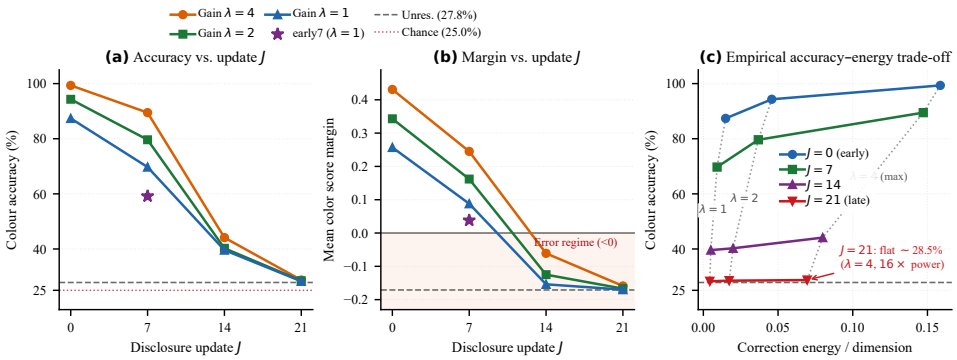


Figure 5: Disclosure timing, guidance gain, score margin, and empirical correction-energy trade-off over 3,200 paired-noise tasks. In panel (c), dotted lines trace iso-gain contours ($\lambda \in \{1, 2, 4\}$); each curve terminates at maximum tested gain $\lambda = 4$, where late interventions ($J = 14, 21$) accumulate strictly less total energy due to shorter remaining horizons. Late high-gain steering does not recover early-disclosure accuracy.

Table 3: Single-update responses on eight response tasks.

update	amplitude	colour	strict	mean margin change
0	0.25	4/8	2/8	+0.0093
0	1.00	4/8	2/8	+0.0315
6	0.25	3/8	1/8	+0.0035
6	1.00	4/8	1/8	+0.0135
13	0.25	3/8	1/8	+0.0006
13	1.00	3/8	1/8	+0.0022
20	0.25	3/8	1/8	+0.0001
20	1.00	3/8	1/8	+0.0006

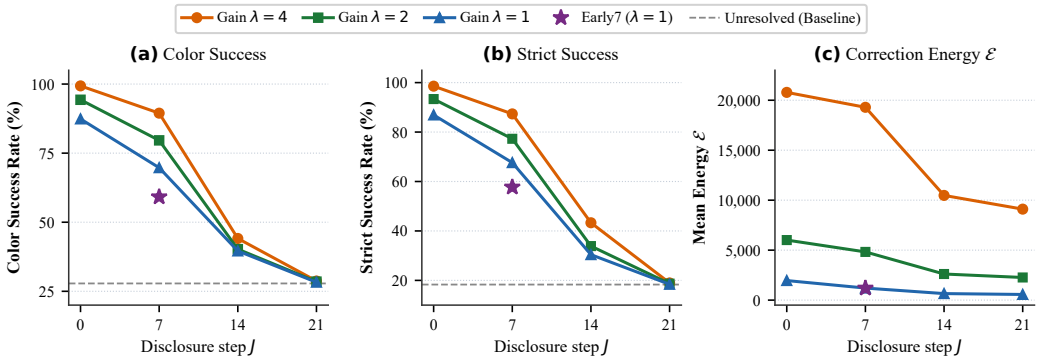


Figure 6: Colour success rate (%), strict success rate (%), and mean correction energy $\mathcal{E}$ across disclosure timing J and guidance gain λ for the 3,200-task paired-noise intervention grid. (a) Early disclosure ($J = 0, 7$) achieves high colour success (up to 99.3% at $\lambda = 4$). Accuracy drops substantially at $J = 14$ and approaches the unresolved baseline at $J = 21$ across the tested gains. (b) Strict success exhibits an even steeper collapse (falling to 18.4–19.0% at $J = 21$). (c) Mean correction energy increases markedly with gain λ for early interventions ($J = 0, 7$), but remains truncated for late interventions ($J = 14, 21$) due to the shortened remaining Euler horizon; late high-gain steering fails to recover early-disclosure accuracy despite high energy expenditure, qualitatively agreeing with the timing analysis.

D.5 Relational-grid qualitative examples

Figures 7 and 8 illustrate step-by-step visual editing sequences on SynGrid-2×2 and SynGrid-3×3.

D.6 Natural-image validation on GenEval

Figure 9 reports overall spatial accuracy across constraint-guided allocation strategies on the 400 GenEval two-object position samples. The unresolved FLUX baseline places the intended spatial relation correctly on only 0.395 of samples. Upfront planning (Plan) substantially lifts overall spatial accuracy to 0.648. Multi-round painting alone (Paint) achieves 0.535–0.555, limited by visual perception noise in intermediate rounds. Combining both paradigms, our hybrid Plan+Paint strategy reaches **0.745**, outperforming planning alone by 9.7 percentage points and elevating spatial accuracy by 35 percentage points over the unresolved baseline.

D.7 Natural-image validation on T2I-ReasonBench

Figures 10 and 11 report automated evaluation scores and qualitative visual cases on 200 entity prompts from T2I-ReasonBench (Sun et al., 2025). Each item pairs an *unresolved* prompt that describes the target entity only indirectly with an *explicit* paraphrase that names it (serving as an answer-supplied oracle).

Paint lifts entity identification by only $+0.040$ over the unresolved baseline (0.385 vs. 0.345), showing that Paint provides only a modest improvement in entity identification, whereas upfront resolution yields substantially larger gains in this evaluation. Upfront planning alone recovers a substantial portion (0.547), while our hybrid Plan+Paint surges to 0.770 (+0.425 over unresolved), effectively matching the 0.775 answer-supplied oracle to within 0.005. This demonstrates that executing the global plan upfront and synthesizing entities incrementally delivers the highest overall fidelity.

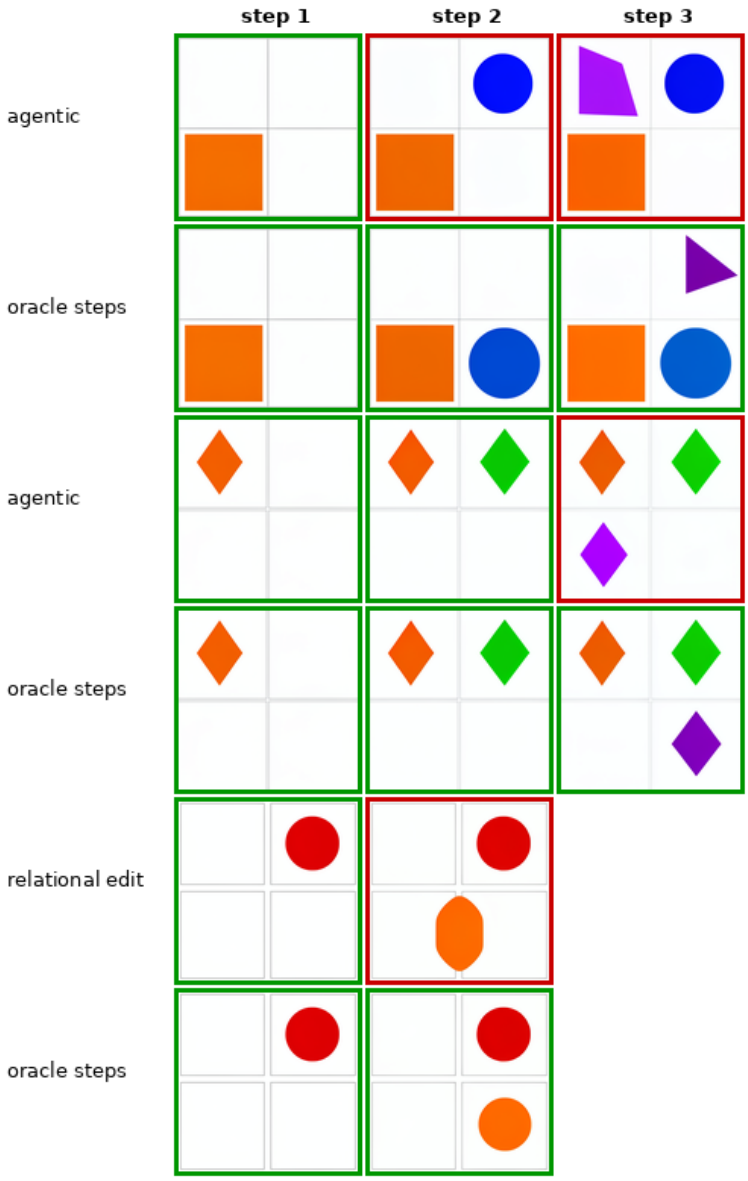


Figure 7: Step-by-step generation sequences on SYNGRID-2 $\times$ 2. Green outlines indicate correct intermediate placement and red outlines indicate errors.

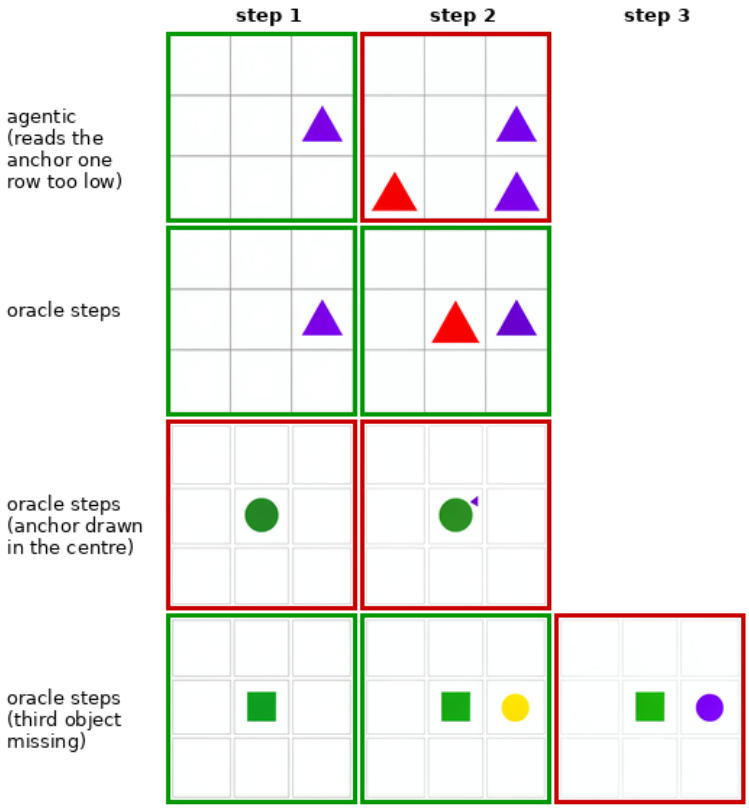


Figure 8: Step-by-step generation sequences on SYNGRID-3 × 3, illustrating spatial crowding and visual-perception errors.

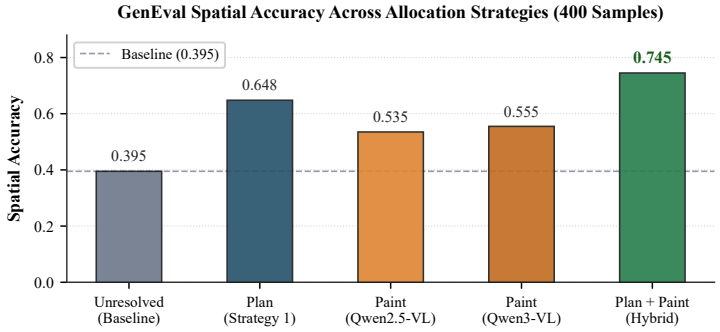


Figure 9: Overall spatial accuracy on GenEval two-object position prompts across constraint-guided allocation strategies (400 samples per arm). Upfront planning (Plan) lifts accuracy from 0.395 to 0.648; multi-round painting alone (Paint) achieves 0.535–0.555; the hybrid PLAN+PAINT strategy reaches **0.745**, outperforming planning alone by 9.7 percentage points and elevating spatial accuracy by 35 percentage points over the unresolved baseline.

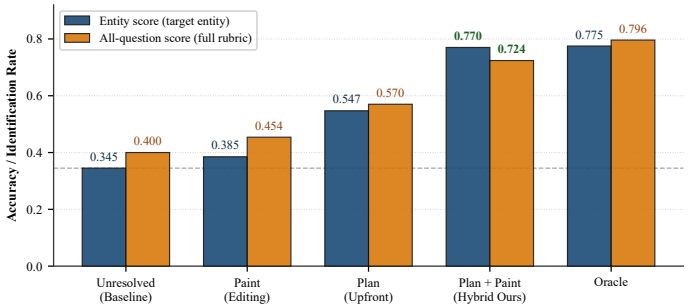


Figure 10: Automated T2I-ReasonBench evaluation (judged by Gemini 3.1 Pro under the benchmark rubric) across constraint-guided allocation strategies on 200 entity-split prompts, reporting entity identification scores alongside all-question rubric scores. Upfront planning (Plan) substantially improves entity identification from 0.345 to 0.547, while Paint alone achieves 0.385. Our hybrid PLAN+PAINT strategy reaches **0.770**, matching the answer-supplied oracle (0.775) within 0.005 (with all-question rubric scores reaching 0.724 vs. 0.796 oracle). Oracle explicitly names the target entity.

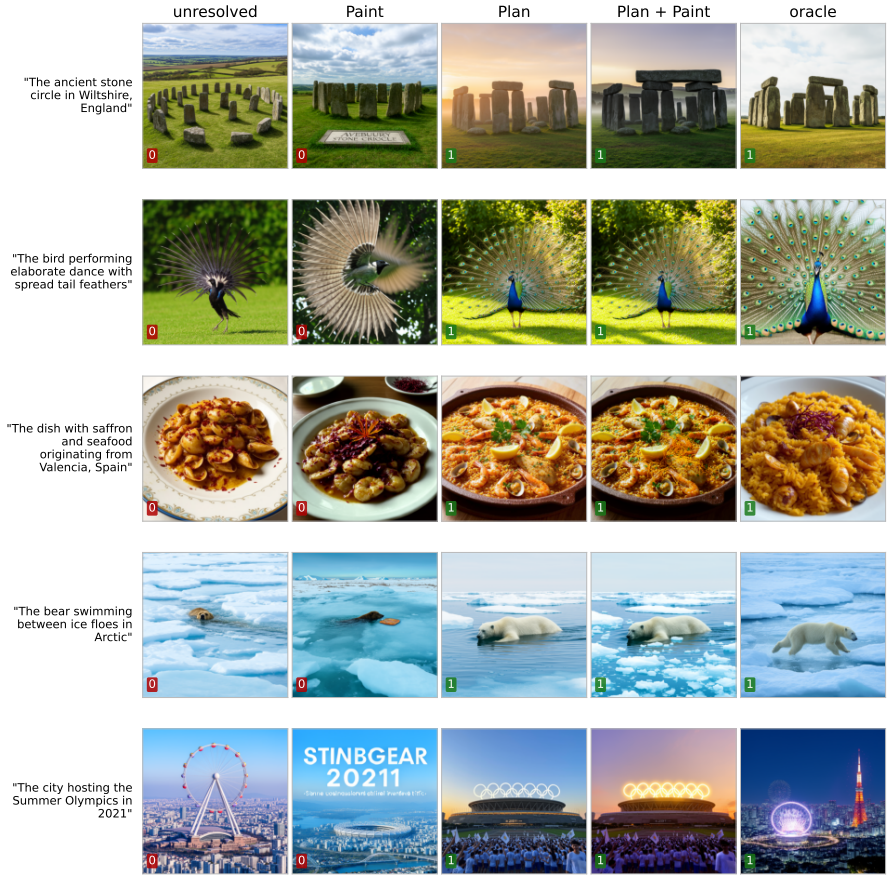


Figure 11: Qualitative cases from T2I-ReasonBench. Each row is one prompt that names its subject only indirectly (left label); the five columns are the arms of Figure 10, and the corner tag is the Gemini 3.1 Pro entity-identification score (green 1, red 0). Unresolved rendering produces a plausible but wrong subject (a generic stone circle, a crow-like bird rather than a peacock, a pasta dish rather than paella, London rather than the Olympic host city); two rounds of Paint edit this wrong canvas without recovering the intended entity, at times adding hallucinated on-image text. Resolving the entity first (Plan and PLAN+PAINT) draws the correct subject and matches the answer-supplied oracle.