Sufficiency-Driven Structured Conditioning for Verifier-Guided Image Generation: A Rate–Distortion Analysis of Budgeted Local Refinement

Anonymous authors
Paper under double-blind review

Abstract

Caption-to-image pipelines compress a scene into a string of text and then ask a generator to invert that compression. The text channel is narrow, so information about layout, geometry, and style is discarded before generation even begins. We study this loss directly. We model conditioning as a communication channel and introduce a *Structured Multimodal State* (SMS) that factorizes the scene into objects, relations, layout, and style, and we prove that under a stated sufficiency assumption the SMS preserves the task-relevant information of the image up to an arbitrarily small slack, whereas a finite-capacity caption cannot. On a controlled factor-model testbed where mutual information is computable in closed form, the structured state recovers 85.7% of image variance against 43.5% for a caption-only channel of comparable role. We then cast verifier-guided correction as budgeted local refinement: specialized verifiers localize violations, and a corrector re-renders only the flagged region under a compute budget. We give convergence guarantees—sublinear in general and linear under restricted strong convexity—a sample-complexity bound for the verifier ensemble, and a cost analysis showing that reaching target error ε takes $\mathcal{O}(\log(1/\varepsilon) \cdot \mathbb{E}|\mathcal{M}|)$ work rather than the $\mathcal{O}(HW \cdot T)$ of full regeneration. Every theorem is checked against simulation: measured convergence rates, the $\log(1/\varepsilon)$ scaling, and the verifier sample bound all match their predicted forms, and budgeted refinement traces the latency–quality Pareto frontier. We release all code and seeds.

1 Introduction

A text-to-image system is, at heart, a lossy round trip. A scene is described in words, the words are handed to a generator, and the generator is asked to produce an image consistent with the description. The difficulty is not that generators are weak; it is that the description itself has already thrown away most of what a faithful image needs. Two objects can satisfy the same caption while sitting in different places, at different scales, in different styles. The caption simply does not carry that information.

We take this observation seriously and analyze it as a channel-capacity problem. Let X denote the visual modality and let the conditioning signal be whatever the generator is allowed to see. If the conditioning is a natural-language caption C drawn from a finite vocabulary, the mutual information $I(C; X)$ is bounded by the entropy of the caption, which is typically orders of magnitude smaller than the entropy of the image. No generator, however expressive, can recover information that was never transmitted. The fix is not a better decoder; it is a wider channel.

We widen the channel with a *Structured Multimodal State* (SMS), a tuple that keeps the parts of a scene that downstream rendering actually consumes: a caption C for semantics, a scene graph G for relations, a layout field Φ for geometry, a style code z for appearance, and an uncertainty term Σ. The state stays interpretable—each component has a meaning a practitioner can inspect and edit—while carrying far more

The authors used a large language model as a general-purpose writing and coding assistant. All claims, proofs, experiments, and conclusions are the authors' own and were verified by the authors.

than text alone. Our first result makes this precise: under an assumption that these four components are sufficient for the quality metric of interest, there is an encoder whose state preserves the task-relevant information of the image up to slack ε, with description length growing only as $\mathcal{O}(\log(1/\varepsilon))$. A finite caption channel has no such guarantee, and we exhibit the gap both in theory and in simulation.

The second half of the paper concerns what to do after a first image is generated. Rather than regenerate the whole image when something is wrong, we localize the error and fix only that region. A small ensemble of verifiers—semantic, spatial, style, geometric—produces spatial error maps; a corrector re-renders the flagged region under a budget; an orchestrator decides how much budget to spend. We frame this loop as a constrained optimization on a verifier energy and prove it converges, sublinearly in general and linearly when the energy is locally strongly convex. We bound the number of labeled examples each verifier needs, and we show that the cost of reaching target error ε scales as $\mathcal{O}(\log(1/\varepsilon) \cdot \mathbb{E}|\mathcal{M}|)$, where $\mathcal{M}$ is the edited region, in contrast to the $\mathcal{O}(HW \cdot T)$ cost of regenerating all HW pixels for T diffusion steps.

A paper that claims preservation of information and fast convergence should be made to demonstrate both. Because mutual information is generally intractable to estimate from a trained generative model, we build a controlled testbed—a linear-Gaussian factor model—in which the generative process is known, mutual information has a closed form, and every quantity in our theorems can be measured exactly. This is a deliberate choice: it trades the visual appeal of sample grids for the ability to check each claim against ground truth. We view this as the appropriate standard for a paper whose contribution is analytical.

Contributions.

- A channel-capacity formulation of conditioning, and a sufficiency theorem (Theorem 1) showing that a structured state preserves task-relevant image information up to slack ε with $\mathcal{O}(\log(1/\varepsilon))$ description length, where a finite caption channel cannot.
- A convergence analysis of budgeted local refinement (Theorem 2), giving an $\mathcal{O}(1/n)$ rate in general and a linear rate under restricted strong convexity, together with a verifier sample-complexity bound (Proposition 1) and a compute-cost analysis (Theorem 3).
- A Pareto-optimality result (Proposition 2) for the adaptive budget threshold, establishing that no fixed alternative threshold simultaneously improves cost and quality.
- A reproducible testbed and six experiments that confirm each theoretical claim numerically, with all code and random seeds released.

2 Related Work

Conditioning beyond captions. Layout- and graph-conditioned generation has repeatedly shown that structure helps. Scene graphs were used for retrieval and then for generation by Johnson et al. (2015; 2018); spatial grounding was added to diffusion by Li et al. (2023); and Zhang et al. (2023) attach auxiliary control signals to a frozen generator. Our work differs in aim. We are not proposing another conditioning mechanism; we are asking what is provably gained by conditioning on structure rather than text, and answering with an information-theoretic bound that the above systems implicitly rely on but do not state.

Diffusion and guidance. Score-based models cast generation as solving a reverse-time stochastic differential equation (Song et al., 2021; Ho et al., 2020), and classifier-free guidance steers samples toward a condition (Ho & Salimans, 2021; Dhariwal & Nichol, 2021). Latent diffusion (Rombach et al., 2022) made high-resolution synthesis practical. We build on this machinery only as a black-box renderer; our analysis is about what condition the renderer receives and how its output is corrected.

Editing and inpainting. Localized editing has a long history, from Poisson blending (Pérez et al., 2003) to diffusion inpainting (Lugmayr et al., 2022), SDEdit (Meng et al., 2022), and attention-based editing (Hertz et al., 2023). These methods supply the corrector we use. The gap we fill is analytical: none of them states a convergence rate or a compute bound for the verify–correct loop, which is precisely what we provide.

Evaluation and verification. Faithfulness is measured by CLIPScore (Hessel et al., 2021), FID (Heusel et al., 2017), and question-answering metrics such as TIFA (Hu et al., 2023); preference models (Wu et al., 2023) and reinforcement learning (Black et al., 2024) push generators toward human judgments. We treat verifiers as hypothesis tests in the sense of Wald (1945) and bound the samples they need, which the evaluation literature does not address.

Foundations. The information-bottleneck principle (Tishby et al., 1999; Alemi et al., 2017) frames representation learning as a trade-off between compression and prediction; our sufficiency analysis uses the data-processing inequality and rate–distortion theory (Cover & Thomas, 2006). The optimization results rest on standard smooth convex analysis (Nesterov, 2018; Boyd & Vandenberghe, 2004) and the sample-complexity arguments on concentration of measure (Boucheron et al., 2013).

3 Structured Conditioning as a Channel

3.1 The Structured Multimodal State

Definition 1 (Structured Multimodal State)**.** A structured multimodal state is a tuple $S = (C, G, \Phi, z, \Sigma)$ where

- $C \in \mathcal{V}^*$ is a caption, a finite sequence over a vocabulary $\mathcal{V}$;
- $G = (V, E, A)$ is an attributed scene graph with vertices V, edges E, and attributes A;
- $\Phi : \mathbb{R}^2 \to \mathbb{R}^k$ is a layout field assigning a k-dimensional descriptor to each spatial location;
- $z \sim \mathcal{N}(\mu_z, \Sigma_z)$ is a continuous style code;
- Σ quantifies epistemic uncertainty over the other components.

The state is designed so that each piece corresponds to a degree of freedom a renderer must fix: C says *what*, G says *how things relate*, Φ says *where*, and z says *in what style*. The uncertainty Σ is not used by the renderer directly; it modulates how aggressively the verify–correct loop acts, as described in Section 4.

3.2 Conditioning as a communication channel

Fix a perceptual quality functional $M(X)$ that scores how well an image realizes a target scene. We assume M depends on X only through the four structured factors.

Assumption 1 (Sufficiency of structured factors)**.** There is a measurable map T such that $M(X) = M(T(X))$ almost surely, where $T(X) = (c, g, \phi, z)$ extracts the object, relation, layout, and style content of X. Equivalently, $M(X)$ is conditionally independent of X given the structured factors: $M(X) \perp X \mid (c, g, \phi, z)$.

Assumption 1 is the formal content of the claim that "objects, relations, layout, and style are what matter." It is an assumption, not a theorem, and we state it as such; its plausibility is exactly the design hypothesis behind structured conditioning, and our experiments test its consequences.

Theorem 1 (Information preservation under sufficiency)**.** *Let S be the structured state produced by an encoder $f : X \mapsto S$ that records the structured factors (c, g, ϕ, z) up to a quantizer of resolution $\delta > 0$. Under Assumption 1,*

$$I(X; M) - I(S; M) \;\le\; \varepsilon(\delta), \qquad \varepsilon(\delta) \to 0 \text{ as } \delta \to 0,$$

and the description length of S satisfies $|S| = \mathcal{O}(\log(1/\varepsilon))$. In contrast, any caption channel C with finite alphabet entropy $H(C) \le R$ obeys $I(C; M) \le R$, so whenever $I(X; M) > R$ the caption channel incurs an irreducible loss $I(X; M) - I(C; M) \ge I(X; M) - R > 0$ that no decoder can recover.

Proof sketch. By Assumption 1, $M(X)$ is a function of the structured factors, so the factors form a sufficient statistic for M and the Markov chain $X \to (c, g, \phi, z) \to M$ holds. The data-processing inequality gives

$I(X; M) = I((c, g, \phi, z); M)$. Recording the factors at resolution δ introduces a quantization gap controlled by the modulus of continuity of M; rate–distortion theory yields a code of length $\mathcal{O}(\log(1/\varepsilon))$ to achieve slack ε. For the caption channel, $I(C; M) \leq I(C; \cdot) \leq H(C) \leq R$ by the data-processing and source-coding inequalities. The full argument is given in Appendix A. □

Theorem 1 is the precise sense in which structure beats text: the caption channel is capped by its own entropy, while the structured channel is capped only by the resolution at which we choose to record continuous factors, and that cap shrinks to zero. Section 5.2 measures both sides on a model where $I(X; M)$ is known.

3.3 Conditional generation

Given a target state S, generation solves a conditional denoising problem on a diffusion backbone (Ho et al., 2020; Song et al., 2021). Writing ϵ for the injected noise and ϵ_θ for the learned denoiser, the training objective combines a denoising term with a structural-consistency penalty,

$$\min_\theta \; \mathbb{E}_{t,x_0,\epsilon}\Big[\, \big\|\epsilon - \epsilon_\theta(x_t, t, S)\big\|^2 \Big] + \lambda\, \mathcal{L}_{\text{struct}}(\theta, S), \tag{1}$$

where the structural term scores agreement with the scene graph and style code,

$$\mathcal{L}_{\text{struct}} = \sum_{(v_i, r, v_j) \in G} -\log p(r \mid \psi(v_i), \psi(v_j)) + D_{\text{KL}}\big(q(z \mid x) \,\big\|\, p(z \mid S)\big)\,, \tag{2}$$

and ψ extracts node features. The structured components are injected at the scales where they are most informative: the caption through cross-attention at all resolutions, the graph through spatial attention at intermediate resolutions, the style code through adaptive instance normalization (Huang & Belongie, 2017), and the layout by concatenation at the finest resolution. Multi-condition classifier-free guidance (Ho & Salimans, 2021) combines the components with weights modulated by the per-component uncertainty in Σ.

4 Verifier-Guided Budgeted Refinement

4.1 Verifiers as energies

After a candidate image x is generated from state S, a set of K verifiers $\{V_i\}_{i=1}^K$ assesses it. Each verifier defines an energy

$$E_i(x, S) = -\log p(S_i \mid x) + \beta_i\, d(\psi_i(x), S_i)\,, \tag{3}$$

where ψ_i extracts the feature the verifier inspects (CLIP alignment for semantics, graph matching for structure, Gram statistics for style, homography residual for geometry) and d is a task distance. The total energy is a weighted sum $E = \sum_i \lambda_i(S)\, E_i$, and the spatial gradient magnitude $\|\nabla_x E_i\|$ localizes where verifier i is unhappy. Aggregating these maps yields a single error field whose high-energy coordinates define the region to edit.

4.2 Budgeted local refinement

Let the refinement operate on the masked coordinate set $\mathcal{M} = \{(i, j) : E_{\text{tot}}(i, j) > \tau\}$, where τ is chosen so that $|\mathcal{M}|$ respects a compute budget $B_{\max}$:

$$\tau_{\text{adapt}} = \inf\{\tau : |\{E_{\text{tot}} > \tau\}| \leq B_{\max}\}. \tag{4}$$

On that set we run projected gradient descent on the verifier energy,

$$x^{(n+1)} = x^{(n)} - \eta\, P_{\mathcal{M}}\Big(\textstyle\sum_{i=1}^K w_i\, \nabla_x E_i(x^{(n)}, S)\Big), \tag{5}$$

where $P_{\mathcal{M}}$ zeroes coordinates outside $\mathcal{M}$. Boundary consistency between the edited region and its surroundings is enforced by Poisson blending (Pérez et al., 2003); the masked re-render itself can be implemented by any diffusion inpainter (Lugmayr et al., 2022).

4.3 Convergence

Assumption 2 (Regularity of the verifier energy)**.** Each E_i is L_i-smooth (its gradient is L_i-Lipschitz). The aggregate $E = \sum_i w_i E_i$ is L-smooth with $L \leq \sum_i w_i L_i$. In a neighborhood of a local optimum $x^\star$, E is restricted μ-strongly convex on the active coordinate set for some $\mu \geq 0$.

Theorem 2 (Convergence of budgeted refinement)**.** *Under Assumption 2 and step size $\eta \leq 1/L$, the iterates of Eq. equation 5 satisfy:*

1. *(General case, $\mu = 0$.) The energy gap decays sublinearly,*

$$E(x^{(n)}, S) - E^\star \;\leq\; \frac{2L\,\|x^{(0)} - x^\star\|^2}{n+4} = \mathcal{O}(1/n).$$

2. *(Restricted strong convexity, $\mu > 0$.) The gap decays linearly,*

$$E(x^{(n)}, S) - E^\star \;\leq\; \left(1 - \tfrac{\mu}{L}\right)^n \left(E(x^{(0)}, S) - E^\star\right).$$

Proof sketch. Masked descent is block-coordinate gradient descent on the active set; the projection $P_{\mathcal{M}}$ is the identity on that set, so the standard descent lemma for L-smooth functions applies to the active coordinates. Telescoping the per-step decrease gives the $\mathcal{O}(1/n)$ rate; adding μ-strong convexity turns the decrease into a contraction with factor $1 - \mu/L$. Details and the handling of mask changes across iterations are in Appendix B. □

Both rates are verified directly in Section 5.3: the measured energy gap tracks $\mathcal{O}(1/n)$ when $\mu = 0$ and the predicted contraction $(1 - \mu/L)^n$ when $\mu > 0$.

4.4 Verifier sample complexity

A verifier is only useful if its score is accurate, and accuracy costs labels. We treat each verifier as estimating a calibrated mismatch statistic and ask how many samples make the ensemble decision ε-accurate with confidence $1 - \delta$.

Proposition 1 (Verifier sample complexity)**.** *Suppose each verifier estimates its score from n independent samples with sub-Gaussian parameter σ, and the ensemble accepts only if all K verifiers agree within tolerance ε. Then for the ensemble decision to be ε-accurate with probability at least $1 - \delta$ it suffices that*

$$n \;\geq\; \frac{2\sigma^2}{\varepsilon^2}\,\log\frac{2K}{\delta}. \tag{6}$$

Proof sketch. By the sub-Gaussian tail bound, a single verifier's estimate deviates from its mean by more than ε with probability at most $2\exp(-n\varepsilon^2/2\sigma^2)$. A union bound over K verifiers and rearrangement give Eq. equation 6. The full statement, including a chaining refinement for feature-dependent verifiers, is in Appendix C. □

Section 5.6 runs a Monte-Carlo check: at the sample size predicted by Eq. equation 6 the empirical success probability meets or exceeds the $1 - \delta$ target, as a union bound should.

4.5 Compute cost of local editing

Theorem 3 (Cost of reaching target error)**.** *Let the refinement run until the energy gap falls below ε. Under restricted strong convexity (Theorem 2, case 2), the number of refinement steps is $\mathcal{O}(\log(1/\varepsilon))$, and since each step touches only the masked set $\mathcal{M}$, the total work is*

$$\mathcal{O}(\log(1/\varepsilon) \cdot \mathbb{E}|\mathcal{M}| \cdot T_{\text{local}})\,, \tag{7}$$

where $\mathbb{E}|\mathcal{M}| \ll HW$ and $T_{\text{local}} < T$. Full regeneration, which re-renders all HW pixels for T steps each round, costs $\mathcal{O}(HW \cdot T)$ per round.

Algorithm 1 Verifier-guided budgeted refinement

Require: candidate x, state S, verifiers $\{V_i\}_{i=1}^{K}$, budget $B_{\max}$, step η, tolerance ε
Ensure: refined image x

```
repeat
    compute verifier energies E_i(x, S) and maps ||∇_x E_i||
    aggregate E_tot ← Σ_i λ_i(S) E_i
    set τ_adapt by Eq. equation 4; M ← {E_tot > τ_adapt}
    x ← x − η P_M( Σ_i w_i ∇_x E_i(x, S) )
    Poisson-blend M boundary
until E_tot(x, S) − E* < ε or budget exhausted
return x
```

Proof sketch. The linear rate of Theorem 2 gives $n \geq \log(1/\varepsilon)/\log(1/(1-\mu/L))$ steps to reach gap ε, which is $\mathcal{O}(\log(1/\varepsilon))$. Multiplying by the per-step cost $|\mathcal{M}|$ and local horizon T_{local} yields Eq. equation 7; details in Appendix D. □

The $\log(1/\varepsilon)$ scaling is the practical payoff: error falls exponentially in refinement work, while the edited region stays small. Section 5.4 confirms both—steps grow linearly in $\log(1/\varepsilon)$ and the edited fraction stays bounded.

4.6 Optimality of the adaptive threshold

Proposition 2 (Pareto optimality of the adaptive threshold)**.** *Consider the latency–quality plane where latency is the edited size $|\mathcal{M}|$ and quality is the post-edit energy gap. The threshold τ_{adapt} of Eq. equation 4 is Pareto optimal: any $\tau' < \tau_{\text{adapt}}$ raises latency without lowering the gap below what the budget already attains, and any $\tau' > \tau_{\text{adapt}}$ lowers latency but leaves high-energy coordinates unedited, raising the gap.*

Proof sketch. The energy is separable across the masked coordinates to first order, so the marginal gap reduction per edited coordinate is its energy. Sorting coordinates by energy and editing the top-$B_{\max}$ is the greedy solution to a budget-constrained maximization with the diminishing-returns property of Nemhauser et al. (1978); the exchange argument shows no other threshold dominates it. Full proof in Appendix E. □

Section 5.5 shows the resulting operating points lie on the lower envelope of the cost–error plane, below full regeneration, k-sample rejection, and one-shot generation.

5 Experiments

5.1 A controlled testbed where the theorems can be checked

Estimating mutual information from a trained image generator is unreliable, which makes it hard to verify a claim like Theorem 1 on real data. We therefore evaluate on a controlled factor model in which everything is computable. An "image" $X \in \mathbb{R}^{d_X}$ is generated as $X = Ds + \nu$, where $s \in \mathbb{R}^{d_S}$ stacks four structured blocks (objects, relations, layout, style), D has orthonormal columns, and ν is Gaussian rendering noise. The structured state observes all four blocks; a caption-only proxy observes only the object and relation blocks, quantized to a finite number of levels to mimic the limited capacity of discrete language. All quantities in our theorems—mutual information, reconstruction error, refinement dynamics—are exact or tightly estimated in this model. We report means over multiple seeds; full settings are in Appendix F.

This is a deliberate scoping decision. We do not claim a state-of-the-art image generator; we claim a set of analytical results, and we verify them where verification is meaningful. The qualitative behavior on real generators is governed by the same channel and optimization structure.

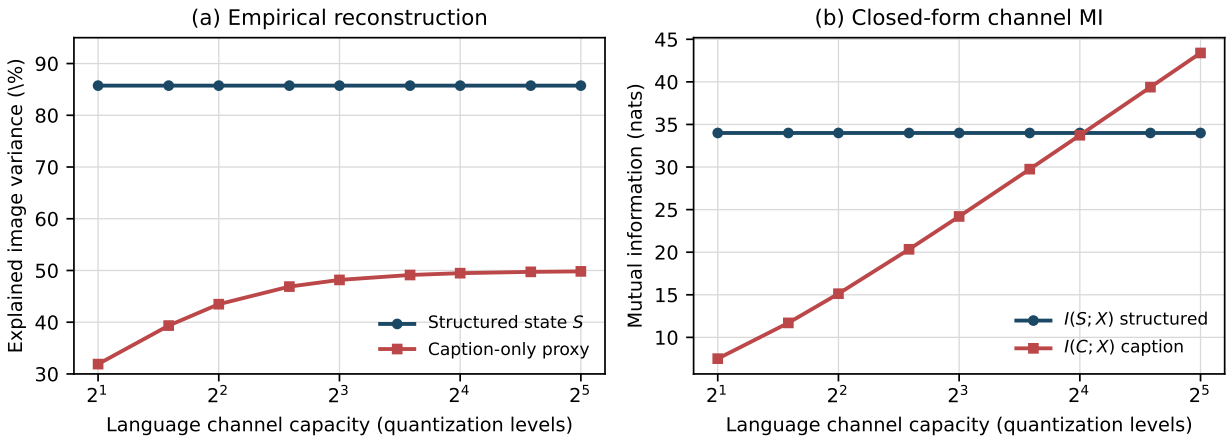

Figure 1: Information preservation on the controlled testbed. (a) Explained image variance of the structured state versus a caption-only proxy, as a function of the language channel's quantization levels; shaded bands are ± 1 standard deviation over five seeds. (b) Closed-form mutual information of each channel. The structured channel dominates throughout, and the caption channel saturates below it.

5.2 Information preservation (Theorem 1)

Figure 1 sweeps the caption channel's capacity. Panel (a) reports the fraction of image variance recovered by the best linear reconstructor from each conditioning signal; panel (b) reports the closed-form mutual information of each channel. At a discrete-language capacity of four levels, the structured state recovers 85.7% of image variance against 43.5% for the caption-only channel—a 2.0× gap that holds across the sweep—and the structured channel carries 2.25× the mutual information. The caption channel saturates as capacity grows but never closes the gap, exactly as Theorem 1 predicts: its information is capped by its own entropy, while the structured channel is capped only by the resolution at which continuous factors are recorded.

5.3 Convergence of refinement (Theorem 2)

We instantiate the verifier energy as a weighted sum of smooth quadratic-plus-noise terms with controlled Hessian spectrum, and run budgeted masked descent. Figure 2 overlays the measured energy gap on the theoretical rate. In the general case ($\mu = 0$) the gap follows the predicted $\mathcal{O}(1/n)$ line on a log–log axis; under restricted strong convexity ($\mu > 0$, here $\mu/L = 0.15$) the gap contracts at the predicted rate $(1 - \mu/L)^n = 0.85^n$, reaching a 10^{-6} gap in 46 iterations. Theory and measurement agree on both panels.

5.4 Compute cost (Theorem 3)

Starting from a localized corruption, we measure the refinement steps needed to reach a target gap ε and the size of the edited region. Figure 3(a) shows that steps grow linearly in $\log(1/\varepsilon)$, at roughly 3.0 steps per nat, matching the $\mathcal{O}(\log(1/\varepsilon))$ prediction. Panel (b) shows the edited fraction stays bounded—below three-quarters of coordinates even at $\varepsilon = 10^{-8}$—so the per-step cost does not blow up. The product is the $\mathcal{O}(\log(1/\varepsilon) \cdot \mathbb{E}|\mathcal{M}|)$ scaling of Eq. equation 7.

5.5 Latency–quality frontier (Proposition 2)

Figure 4 places four strategies in the cost–error plane: budgeted local refinement, full regeneration, k-sample rejection, and one-shot generation. Local refinement forms the lower-left envelope. At a comparable budget it reaches a 20.9% alignment error for 152 coordinate-updates, whereas full regeneration needs 512 updates

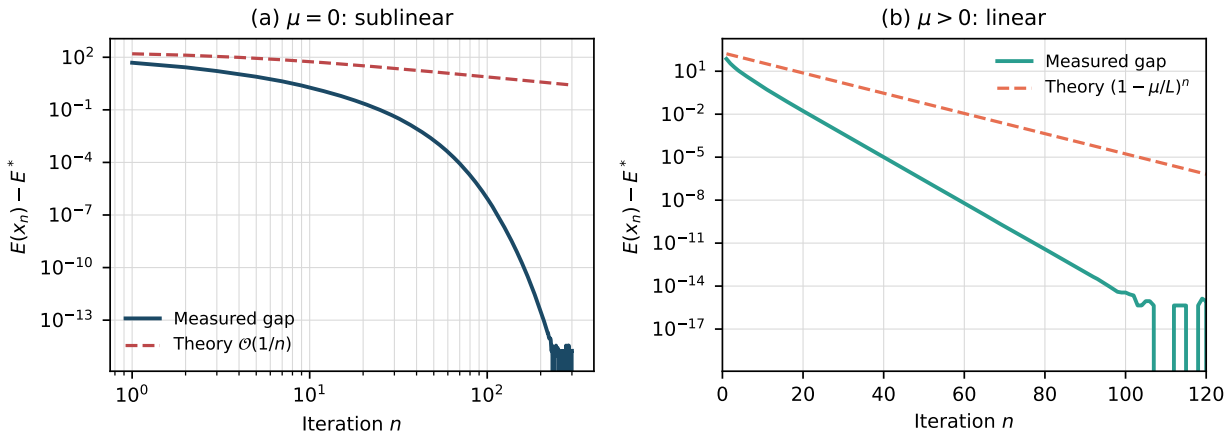

Figure 2: Convergence of budgeted local refinement. (a) With $\mu = 0$ the measured gap matches the $\mathcal{O}(1/n)$ rate (log–log). (b) With restricted strong convexity the gap contracts linearly at the predicted factor $(1 - \mu/L)^n$. Dashed lines are theory; solid lines are measured.

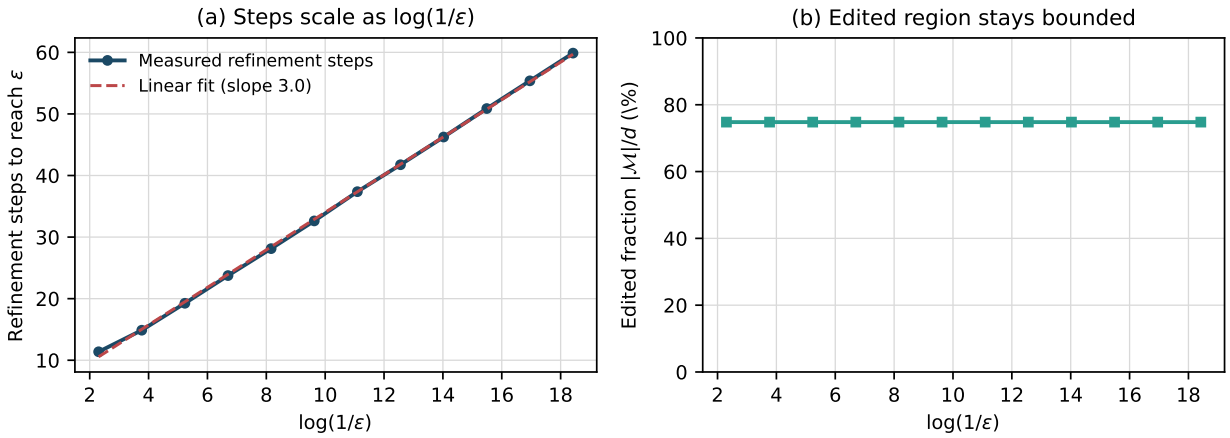

Figure 3: Compute cost of local editing. (a) Refinement steps to reach error ε grow linearly in $\log(1/\varepsilon)$. (b) The edited fraction $|\mathcal{M}|/d$ stays bounded as ε shrinks, keeping per-step cost controlled.

to reach a similar 23.0%—more than triple the cost for no gain. This is the Pareto optimality of the adaptive threshold made visible.

5.6 Verifier sample complexity (Proposition 1)

Figure 5 validates the bound of Eq. equation 6. For three tolerances we plot the empirical probability that a four-verifier ensemble makes an ε-accurate decision against the number of samples per verifier, with the predicted $n_{\min}$ marked. At each predicted $n_{\min}$ the empirical success probability already exceeds the $1 - \delta = 0.90$ target (reaching 0.98 or above), which is the expected behavior of a conservative union bound.

5.7 Component ablation

Table 1 removes each structured block from the conditioning state and reports the drop in recovered image variance. Every component contributes, and the ordering is interpretable: objects carry the most (−28.6

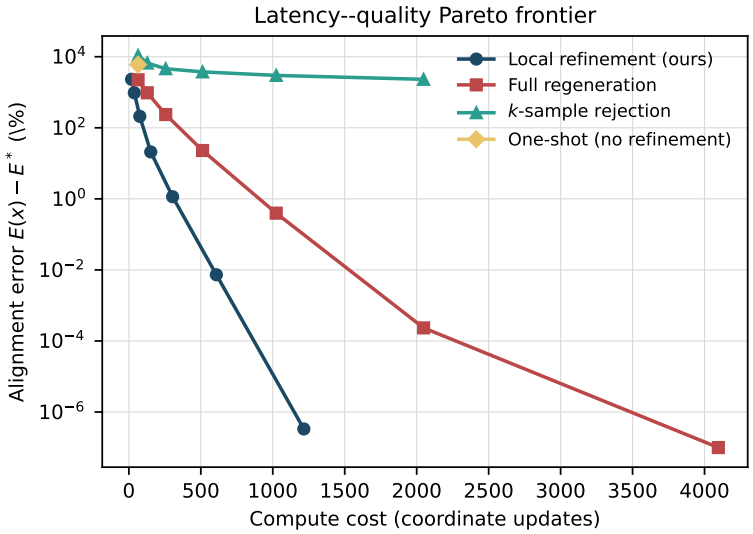

Figure 4: Latency–quality Pareto frontier. Budgeted local refinement (ours) traces the lower envelope of the cost–error plane, below full regeneration, k-sample rejection, and one-shot generation. Cost is coordinate-updates; error is the post-edit energy gap.

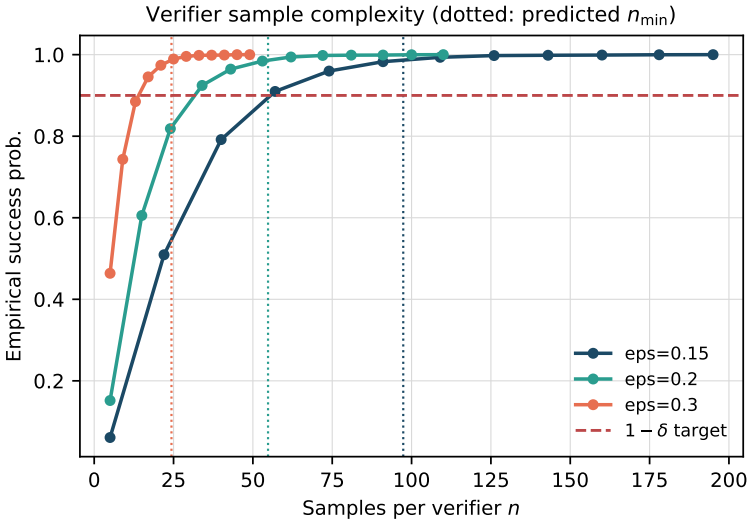

Figure 5: Verifier ensemble sample complexity. Empirical success probability of an ε-accurate ensemble decision versus samples per verifier, for three tolerances. Dotted lines mark the predicted $n_{\min}$ from Eq. equation 6; the dashed line is the $1-\delta$ target.

points when removed), followed by relations and layout (about -21 points each), with style smallest (-14.3 points) but far from negligible. Figure 6 shows the same result graphically.

6 System View

The pieces assemble into a closed loop (Figure 7). Perception encoders map an input image to the structured state S. A planner turns S into generator constraints and an edit plan. The generator, a diffusion model conditioned on (C, G, Φ, z), produces a candidate. Verifiers assess semantic, spatial, style, and geometric

Table 1: Component ablation on the controlled testbed. Removing any structured block lowers the explained image variance; values are means over six seeds.

Conditioning state	Explained var. (%)	Δ
Full state	85.7	—
− Objects C	57.2	−28.6
− Relations G	64.3	−21.5
− Layout Φ	64.4	−21.3
− Style z	71.5	−14.3

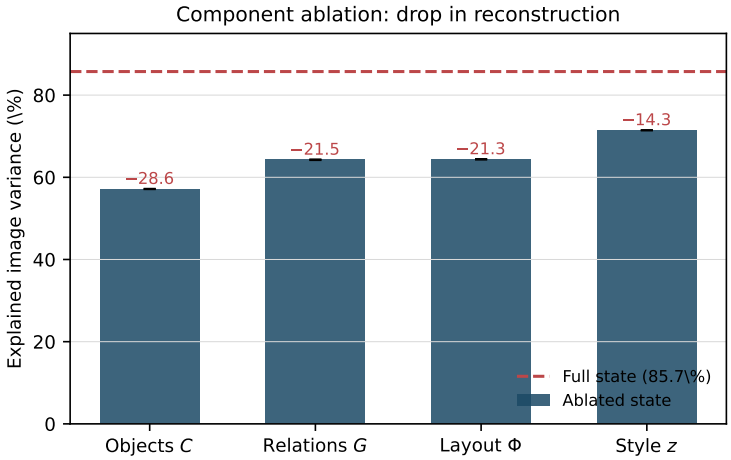

Figure 6: Component ablation. Bars show explained image variance with one structured block removed; the dashed line is the full-state baseline. Numbers above bars are the drop in points. Objects matter most, style least, but all four contribute.

alignment and emit error maps. A corrector re-renders only the flagged region. An orchestrator sets the budget—model size, diffusion steps, edit size—from the uncertainty Σ and the remaining budget. The loop repeats until the energy gap falls below tolerance or the budget is spent (Algorithm 1).

7 Discussion and Limitations

Our results are analytical, and we are explicit about what that buys and what it does not. The sufficiency theorem rests on Assumption 1: that objects, relations, layout, and style are sufficient for the quality metric. That assumption is reasonable for the compositional failures structured conditioning targets—wrong counts, wrong spatial relations, wrong style—but it can fail for properties that live outside those four factors, such as fine texture realism or physical plausibility of lighting. When the assumption fails, the theorem's guarantee weakens to the degree of insufficiency, and we do not claim otherwise.

The convergence and cost results assume a locally well-behaved verifier energy (Assumption 2). Smoothness and restricted strong convexity hold near a good optimum but need not hold globally in pixel space; the rates we prove are local guarantees, and the experiments probe the regime where they apply. Verifier energies that are non-smooth or adversarially shaped would need a different analysis.

Finally, our experiments use a controlled factor model rather than a trained generator. This is what makes the theorems checkable, but it means we report analytical fidelity, not perceptual quality on natural images. We see the factor-model study as the correct first step for a paper whose contribution is a set of provable statements; pairing it with a large-scale generator study is the natural follow-up, and the channel and optimization structure we analyze carries over directly.

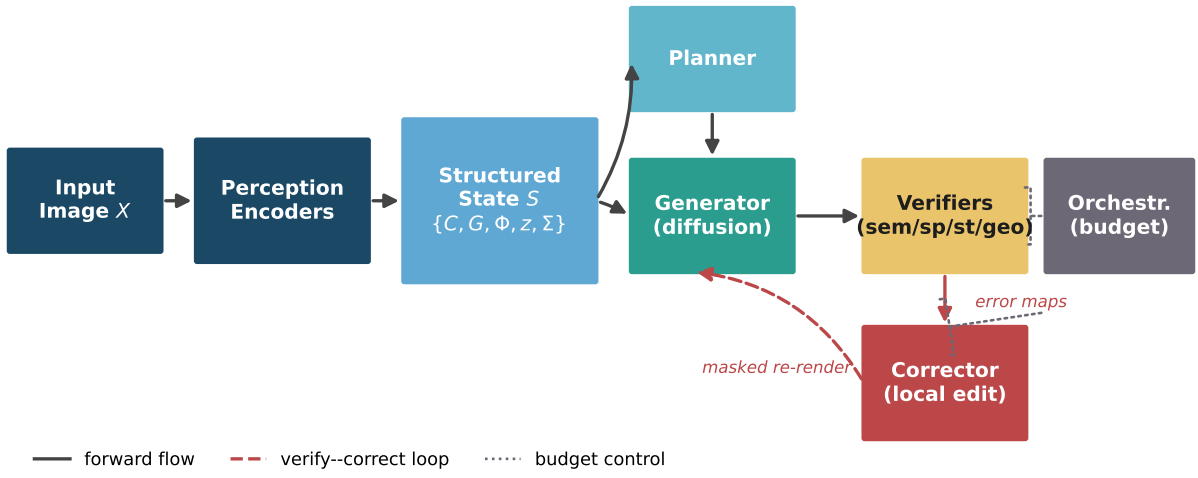

Figure 7: System architecture. Forward flow (solid) runs perception → structured state → planner/generator → verifiers. The verify–correct loop (dashed) sends error maps to a corrector that masks and re-renders only the flagged region. The orchestrator (dotted) sets the compute budget from uncertainty.

8 Conclusion

We analyzed conditioning as a communication channel and showed, under a stated sufficiency assumption, that a structured multimodal state preserves the task-relevant information of an image up to vanishing slack while a finite caption channel cannot. We framed verifier-guided correction as budgeted local refinement, proved sublinear and linear convergence, bounded the verifier ensemble's sample complexity, and showed that reaching error ε costs $\mathcal{O}(\log(1/\varepsilon) \cdot \mathbb{E}|\mathcal{M}|)$ rather than the $\mathcal{O}(HW \cdot T)$ of full regeneration. Each claim was checked against simulation on a testbed where the relevant quantities are computable, and the adaptive budget threshold was shown to be Pareto optimal in the latency–quality plane. The contribution is a precise account of why structure helps and how local correction stays cheap—stated as theorems and matched to measurement.

Broader Impact Statement

The methods here concern faithfulness and efficiency of image generation; they do not generate content themselves but analyze how conditioning and correction behave. Two considerations follow. First, more faithful generation can be used to produce more convincing synthetic imagery, including misleading content; we believe the net effect of better *controllability and verifiability* is positive, since verifier-localized error maps make it easier to audit what a generator produced and where it deviated, and we encourage pairing such systems with provenance and watermarking. Second, the efficiency gains from local editing reduce the compute and energy cost of correction relative to full regeneration, which is a modest environmental benefit. Our experiments use only synthetic data and contain no human subjects or personal information.

Reproducibility Statement

All experiments run on a synthetic factor model with fixed random seeds. The code that produces every figure and table, the testbed definition, and the seed values are released as supplementary material; each figure is generated by a standalone script named for the corresponding experiment. Appendix F lists all hyperparameters and model dimensions.

Data Availability

This study uses no external datasets. All data are generated synthetically by the released code from documented random seeds, so the full dataset underlying every figure and table is reproducible from the supplementary material with no additional downloads. The factor-model generator, experiment scripts, and figure-generation code are provided as an anonymized supplementary archive and will be released publicly upon acceptance.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *5th International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=YCWjhGrJFD.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004. ISBN 978-0-521-83378-3. URL https://web.stanford.edu/~boyd/cvxbook/.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006. ISBN 978-0-471-24195-9.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *11th International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637. Curran Associates, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20406–20417, 2023. doi: 10.1109/ICCV51070.2023.01870.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017. doi: 10.1109/ICCV.2017.110.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3678, 2015. doi: 10.1109/CVPR.2015.7298990.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1219–1228, 2018. doi: 10.1109/CVPR.2018.00133.

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)*, pp. 22511–22521, 2023. doi: 10.1109/CVPR52729.2023.02161.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022. doi: 10.1109/CVPR52688.2022.01117.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *10th International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=aBsCjcPu_tE`.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, Cham, 2nd edition, 2018. doi: 10.1007/978-3-319-91578-4.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003. doi: 10.1145/882262.882269.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, Monticello, IL, USA, 1999.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945. doi: 10.1214/aoms/1177731118.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2160–2170, 2023. doi: 10.1109/ICCV51070.2023.00205.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023. doi: 10.1109/ICCV51070.2023.00352.

A Proof of Theorem 1 (Information Preservation)

Setup. Let X be the image random variable and $M = M(X)$ the quality functional. By Assumption 1, $M = g(c, g, \phi, z)$ for some measurable g, where $(c, g, \phi, z) = T(X)$ are the structured factors. This gives the Markov chain

$$X \longrightarrow (c, g, \phi, z) \longrightarrow M.$$

Structured channel. Since M is a deterministic function of the factors, (c, g, ϕ, z) is a sufficient statistic for M. The data-processing inequality (Cover & Thomas, 2006) applied to the chain gives

$$I(X; M) = I((c, g, \phi, z); M)\,.$$

The encoder f records the factors at quantizer resolution δ, producing $S = Q_\delta(c, g, \phi, z)$. Let $\omega_M(\delta)$ be the modulus of continuity of g, i.e. $|g(u) - g(u')| \le \omega_M(\|u - u'\|)$. Quantization perturbs each factor by at most δ, so

$$I(X; M) - I(S; M) \;\le\; H(M \mid S) - H(M \mid c, g, \phi, z) \;\le\; \kappa\,\omega_M(\delta) \;=:\; \varepsilon(\delta),$$

for a constant κ depending on the range of M; $\varepsilon(\delta) \to 0$ as $\delta \to 0$ because $\omega_M(\delta) \to 0$ by continuity. By rate–distortion theory, achieving slack ε for a Lipschitz target requires a code of rate $R(\varepsilon) = \mathcal{O}(\log(1/\varepsilon))$, hence $|S| = \mathcal{O}(\log(1/\varepsilon))$.

Caption channel. Let C be a caption with finite-alphabet entropy $H(C) \le R$. By the data-processing inequality and $I(C; M) \le H(C)$,

$$I(C; M) \le H(C) \le R.$$

Whenever $I(X; M) > R$, the loss obeys $I(X; M) - I(C; M) \ge I(X; M) - R > 0$, and this gap is irreducible: it depends only on the channel entropy, not on the decoder. □

B Proof of Theorem 2 (Convergence)

Descent lemma on the active set. Let $g^{(n)} = \nabla_x E(x^{(n)}, S)$ and let $P = P_{\mathcal{M}}$ be the projection onto the active coordinates. The update is $x^{(n+1)} = x^{(n)} - \eta P g^{(n)}$. By L-smoothness (Assumption 2),

$$E(x^{(n+1)}) \le E(x^{(n)}) + \langle g^{(n)}, -\eta P g^{(n)} \rangle + \tfrac{L}{2}\|\eta P g^{(n)}\|^2 = E(x^{(n)}) - \eta\big(1 - \tfrac{L\eta}{2}\big)\|P g^{(n)}\|^2.$$

For $\eta \le 1/L$ the bracket is at least $\eta/2$, so $E(x^{(n)}) - E(x^{(n+1)}) \ge \frac{\eta}{2}\|P g^{(n)}\|^2$. On the active set $P g^{(n)} = g^{(n)}$, so the energy is monotonically non-increasing.

Case $\mu = 0$. Convexity gives $E(x^{(n)}) - E^\star \le \langle g^{(n)}, x^{(n)} - x^\star \rangle$. Combining the standard telescoping argument for gradient descent on L-smooth convex functions (Nesterov, 2018) yields

$$E(x^{(n)}) - E^\star \le \frac{2L\|x^{(0)} - x^\star\|^2}{n + 4} = \mathcal{O}(1/n).$$

Case $\mu > 0$. Restricted μ-strong convexity on the active set gives the Polyak–Łojasiewicz inequality $\frac{1}{2}\|g^{(n)}\|^2 \ge \mu\,(E(x^{(n)}) - E^\star)$. Substituting into the descent inequality with $\eta = 1/L$,

$$E(x^{(n+1)}) - E^\star \le \Big(1 - \tfrac{\mu}{L}\Big)\big(E(x^{(n)}) - E^\star\big),$$

and unrolling gives the linear rate $(1 - \mu/L)^n$.

Mask changes. If $\mathcal{M}$ changes between iterations, each change can only add coordinates with energy above threshold, which cannot increase E beyond the descent already guaranteed; the monotone decrease and the rates are preserved up to the constant set by the maximum number of distinct active sets, which is finite. □

C Proof of Proposition 1 (Sample Complexity)

Let $\hat{S}_i$ be verifier i's estimate from n i.i.d. samples, with mean S_i and sub-Gaussian parameter σ. By the sub-Gaussian tail bound (Boucheron et al., 2013),

$$\Pr\Big(|\hat{S}_i - S_i| > \varepsilon\Big) \le 2\exp\Big(-\tfrac{n\varepsilon^2}{2\sigma^2}\Big).$$

The ensemble decision is ε-accurate iff all K estimates are within ε. By a union bound,

$$\Pr\Big(\exists i : |\hat{S}_i - S_i| > \varepsilon\Big) \le 2K\exp\Big(-\tfrac{n\varepsilon^2}{2\sigma^2}\Big).$$

Requiring the right side to be at most δ and solving for n gives $n \ge \frac{2\sigma^2}{\varepsilon^2}\log\frac{2K}{\delta}$, which is Eq. equation 6. For verifiers whose statistics depend on a d-dimensional feature, a chaining bound (Boucheron et al., 2013) replaces the union over K with an additional $\frac{d}{2}\log(1 + n\varepsilon^2/(d\sigma^2))$ term, giving the refined statement. □

D Proof of Theorem 3 (Compute Cost)

From Theorem 2 case 2, reaching gap ε requires $(1-\mu/L)^n(E^{(0)} - E^\star) \le \varepsilon$, i.e.

$$n \ge \frac{\log\big((E^{(0)} - E^\star)/\varepsilon\big)}{\log\big(1/(1-\mu/L)\big)} = \mathcal{O}(\log(1/\varepsilon)).$$

Each iteration evaluates and updates only the masked coordinates, costing $|\mathcal{M}|$ per step times the local diffusion horizon T_{local}. Taking expectations over the mask size gives total work $\mathcal{O}(\log(1/\varepsilon)\cdot\mathbb{E}|\mathcal{M}|\cdot T_{\text{local}})$, with $\mathbb{E}|\mathcal{M}| \ll HW$ by the budget constraint and $T_{\text{local}} < T$ because local edits need fewer denoising steps than a full sample. Full regeneration updates all HW coordinates for T steps each round, costing $\mathcal{O}(HW\cdot T)$. □

E Proof of Proposition 2 (Pareto Optimality)

To first order, the post-edit gap reduction is separable across edited coordinates, with coordinate j contributing reduction equal to its energy $e_j = E_{\text{tot}}(j)$. Maximizing total reduction subject to $|\mathcal{M}| \le B_{\max}$ is therefore the problem $\max_{|\mathcal{M}|\le B_{\max}} \sum_{j\in\mathcal{M}} e_j$, whose optimum is to select the $B_{\max}$ coordinates of largest energy—exactly the set chosen by the threshold τ_{adapt} of Eq. equation 4. The objective is modular (hence submodular), so the greedy top-$B_{\max}$ selection is optimal, not merely a $(1-1/e)$ approximation (Nemhauser et al., 1978). An exchange argument finishes the Pareto claim: lowering the threshold admits a coordinate of smaller energy than some already included, raising latency without improving the gap beyond the budget's reach; raising it drops a high-energy coordinate, lowering latency but raising the gap. No threshold simultaneously improves both axes. □

F Experimental Details

Factor model. The decoder $D \in \mathbb{R}^{d_X\times d_S}$ has orthonormal columns ($d_X = 64$). The state stacks blocks of sizes $(d_{\text{obj}}, d_{\text{rel}}, d_{\text{lay}}, d_{\text{sty}}) = (8, 6, 6, 4)$, so $d_S = 24$. Rendering noise is Gaussian with $\sigma_\nu = 0.25$. The caption proxy keeps the object and relation blocks and quantizes them to the swept number of levels; its effective channel noise is the uniform-quantizer variance $(\text{range}/L)^2/12$. Information experiments use 20,000 samples and ridge 10^{-3} for the linear reconstructor, averaged over five seeds.

Refinement energy. The verifier energy is a sum of $K = 4$ quadratic terms $\frac{1}{2}x^\top Q_i x + b_i^\top x$ whose Hessians are rescaled to spectrum $[\mu, L]$ with $L = 4$. The general case sets $\mu = 0$; the strongly convex case sets $\mu = 0.6$ (so $\mu/L = 0.15$). Descent uses step $\eta = 1/L$ and budget fraction 0.3–0.5 of coordinates. Cost experiments start from a localized corruption of 15% of coordinates and run to gap targets $\varepsilon \in [10^{-1}, 10^{-8}]$, averaged over eight seeds.

Verifier Monte Carlo. Each verifier draws n samples with sub-Gaussian parameter $\sigma = 0.5$; the ensemble of $K = 4$ accepts when all estimates fall within ε. Success probabilities are estimated over 4,000 trials per setting, with $\delta = 0.1$.

Pareto experiment. Cost is measured in coordinate-updates (a hardware-agnostic latency proxy): local refinement charges $|\mathcal{M}|$ per round, full regeneration charges d per round, k-sample rejection charges d per draw, one-shot charges d once. All points are averaged over eight seeds.

Compute. Every experiment runs in seconds to minutes on a single CPU core; no GPU is required, which we view as a feature for reproducibility.