PROMPT REINJECTION FOR SPATIAL REASONING IN MMDITS: SIGNAL AMOUNT, DEPTH ALLOCATION, AND CONTENT-SPECIFIC REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal diffusion transformers (MMDiTs) forget prompt semantics as depth grows, and spatial relations suffer most. Prompt Reinjection (PR) counters this by repairing every text token at every layer, but neither the number of repaired positions nor the choice of layers has been examined; we therefore treat PR as a probe of its own redundancy. Repair saturates early: eight real positions match full-scale repair, and the injection weight, not the token count, controls the response; past a bounded window, amplification collapses performance on SD3. Most repaired positions are padding, which joint attention does not mask and which carries no signal. Depth adds a secondary trend at the default weight: with span-matched controls, dense repair over the early half matches full PR while late-only repair is no better than baseline, and several allocations perform comparably. As one instantiation, an early-full/late-sparse schedule (Layer-Adaptive Sparse Prompt Reinjection, LASP) reaches 40.1% on GenEval Position, matches full PR with a third of the token-layer pairs, and remains robust where full PR collapses; we make no optimality claim. The injected content matters—noise, a permutation, or a constant removes the gain—and the pattern transfers to T2I-CompBench and directionally to FLUX.1-Schnell.

1 INTRODUCTION

Text-to-image models generate images from natural-language prompts, and spatial instructions remain their most error-prone dimension in GenEval Ghosh et al. (2023), where position accuracy is the lowest across all instruction types. Multimodal diffusion transformers (MMDiTs), such as SD3 and FLUX, depart from cross-attention DiTs Peebles & Xie (2023); Chen et al. (2023) by jointly processing text and visual tokens through bidirectional attention, so that text representations evolve with depth rather than remaining fixed. Because supervision is applied only to visual tokens, fine-grained text semantics are progressively lost as depth increases, a phenomenon termed prompt forgetting Yao et al. (2026): objects are missed and spatial relations misplaced.

Prompt Reinjection (PR) Yao et al. (2026) counters this forgetting by re-injecting shallow-layer text features into deeper layers. Whether the degradation it addresses is token-local or global, and where it concentrates, remains unclear: if token-local, each affected token must be restored individually; if global, any sufficiently large subset suffices. We separate the locus of degradation from the reach of the repair signal: probes show that degradation is category-selective, with spatial-relation tokens degrading fastest (Appendix B), whereas the repair signal propagates through joint attention beyond the repaired positions. PR takes the most conservative route, repairing every token at every layer with a fixed injection weight. Whether this full-scale repair is necessary remains unexamined in this setting.

We therefore ablate three control dimensions—token count, layer span, and injection weight—each corresponding to an untested assumption in the original PR design. Our experiments focus on spatial reasoning. Of the instruction-following dimensions that GenEval evaluates, spatial relations are hardest for MMDiTs: on SD3-medium the base model scores 29.9% on Position against 82.8% on Colors (10-seed, Appendix D), and spatial-relation tokens show the sharpest decodability collapse with depth (Appendix B). Spatial reasoning is thus a sensitive diagnostic target for prompt forget-

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.

ting. PR suits that role because it is training-free and its three parameters (token count, layer span, injection weight) map one-to-one onto the dimensions we ablate.

We ask when repair helps and when it hurts: about eight positions already suffice, repair becomes harmful past a weight threshold, and, at the default weight, the gain concentrates in the early-to-middle layers. LASP (Layer-Adaptive Sparse Prompt Reinjection) follows from this characterization: the dense early stage acts as a degeneration gate that preserves clean text features for sparse late repair. If dense early repair is what carries the gain, an early-full/late-sparse allocation should suffice at a fraction of the signal, and we confirm that prediction empirically. The characterization itself does not depend on the LASP recipe.

PR repairs far more tokens than needed. Eight positions match full-scale repair, and injection weight matters more than token count; the collapse past the safe window is sharp, and its weight threshold is non-monotonic in token count. The two ends fail for different reasons: too few repaired tokens leave the stream inconsistent, whereas too much injected signal overwrites it. Depth adds a secondary trend: at the default weight, the effective repair region is the early-to-middle block, and the early-full/late-sparse schedule built on it remains robust at amplified weights where full repair collapses. A dominant signal-amount axis and a secondary depth trend thus carry implications for MMDiT design—for text-side interventions generally, not only PR. Core findings hold on T2I-CompBench, and the sparse-repair benefit reproduces directionally on FLUX.1-Schnell.

1. **Signal-amount axis.** Using PR as a probe (Appendix B), we show that repairing a handful of real tokens significantly improves overall spatial reasoning (p=0.005), whereas repairing content-free padding positions does not. The benefit saturates at a small number of positions; injection weight dominates, and the safe operating window is bounded by a collapse whose weight threshold is non-monotonic in token count. We further find that conditional-branch-only injection matches full PR at the default weight yet is the most fragile at amplified weights (Section 5.4), and that H(L3) attention entropy carries no discriminating signal at fixed token count.
2. **Secondary depth trend.** The pre-declared Holm-corrected test favored late-layer repair, but its controls are not span-matched; with equal 11-layer windows, dense early repair (L3–L13) matches full PR and late-only repair is no better than baseline (Table 1). At the default weight, the effective repair region is the early-to-middle block.
3. **Content-specific repair.** Replacing the injected shallow features with norm-matched noise, a random feature permutation, or a content-free constant removes the gain (each significantly below the real-feature condition, $p \leq 0.007$), so the repair requires the real text features rather than any perturbation.
4. **A practical instantiation, not an optimal one.** LASP (Figure 1) applies full PR in early layers and sparse top-K PR in late layers. It matches full PR at the default weight with a third of the pairs, is more robust to over-injection than full PR, and is the only such schedule we validate across benchmarks and architectures; we make no optimality claim.

2 Related Work

2.1 MMDiTs and Prompt Reinjection

Diffusion Transformers (DiTs) Peebles & Xie (2023) kept the U-Net's cross-attention conditioning Ho et al. (2020); Rombach et al. (2022), in which text features are computed once and stay fixed. MMDiTs such as SD3 Esser et al. (2024) and FLUX Black Forest Labs (2024) take a different route: text and visual tokens are concatenated into one sequence and processed through bidirectional attention, so text co-evolves with visual features. That interaction is also less stable: with no supervision on the text side, text features drift from their semantics as depth grows—prompt forgetting Yao et al. (2026)—surfacing as missed objects and misplaced spatial relations.

Yao et al. traced the drift to a supervisory asymmetry, since the diffusion objective supervises only visual tokens, and proposed Prompt Reinjection, a training-free intervention that re-injects shallow-layer text features into deeper blocks. In concurrent work, Spingarn et al. Spingarn et al. (2026) report that contextless, position-tagged word embeddings are enough for generation, though they propose no repair mechanism. PR repairs every token at every layer at a fixed weight, and neither its

token set, layer span, nor weight has been tested. We therefore treat PR as a probe and ablate those three parameters, asking which parts of the full-scale recipe are necessary and when amplification turns destructive.

2.2 Token Attribution and Training-Free Interventions

Two attribution lines study what individual tokens carry. Circuit analysis traces how cross-attention heads read spatial-relation and object-attribute information, and how this depends on the text encoder Wang et al. (2026); DAAM instead attributes generation to word tokens in U-Net cross-attention Tang et al. (2023). Attribution of this kind feeds attention-based interventions. Attend-and-Excite amplifies under-attended objects Chefer et al. (2023), and Prompt-to-Prompt reweights token influence by editing cross-attention maps Hertz et al. (2023); both target U-Nets. Layout-constraint methods such as BoxDiff Xie et al. (2023) operate at inference, while GLIGEN Li et al. (2023) and ControlNet Zhang et al. (2023) require extra training.

All of these methods operate inside cross-attention blocks, where information stays local to token categories. In an MMDiT's joint attention that locality is gone: for the spatial dimension studied here, repair needs only a small subset of the sequence's leading block, whatever the category, whereas compositional attribute binding requires denser coverage (Appendix D). The repair is content-specific as well: arbitrary substitutes remove the gain (Section 5.3). TACA Lv et al. (2025) is the MMDiT-specific exception, but it temperature-scales cross-modal attention and needs LoRA fine-tuning, whereas PR is training-free and changes no weights. Our findings describe this injection family; whether they carry over to other text-side interventions remains open. They also refine the attribution picture: because the repair signal propagates beyond the injected positions, the repaired tokens need not be semantically selected, and token identity matters only when the budget is extremely small; the injected content and the injection weight dominate instead.

3 Preliminaries

3.1 Prompt Reinjection

Prompt Reinjection (PR) Yao et al. (2026) is a training-free inference-time intervention: given a prompt with T tokens, it extracts shallow-layer text features (layer 2) and re-injects them into deeper layers via distribution anchoring, orthogonal Procrustes alignment, and residual injection with weight w. Formally, let $h_l^{\text{txt}} \in \mathbb{R}^{T\times d}$ denote the text features at layer l; for each target layer, PR first token-wise standardizes origin and target features,

$$\tilde{h}_l^{\text{txt}} = (h_l^{\text{txt}} - \mu_l)/\sigma_l, \tag{1}$$

(per-token mean μ_l and std σ_l over the hidden dimension), computes the orthogonal Procrustes rotation aligning the origin to the target,

$$R_l = UV^T, \quad \text{where} \quad (\tilde{h}_{L_{\text{ori}}}^{\text{txt}})^T \tilde{h}_l^{\text{txt}} = U\Sigma V^T, \tag{2}$$

and applies residual injection with weight w, followed by a layer-norm stabilization and inverse standardization:

$$h_l^{\text{txt}} \leftarrow \text{LN}\Big(\tilde{h}_l^{\text{txt}} + w\,\tilde{h}_{L_{\text{ori}}}^{\text{txt}} R_l\Big)\cdot\sigma_l + \mu_l, \quad l \in [3, 24]. \tag{3}$$

Because the token count is smaller than the hidden dimension (T=333 $<$ d=4096), the cross-covariance in Eq. 2 is rank-deficient and R_l is not uniquely identified; we use the SVD-based nearest orthogonal solution. We adopt the origin layer (L_{ori}=2) and target span ($[3, 24]$) from Yao et al. (2026). In its default configuration, PR repairs all 333 positions of the text sequence across all 22 layers at w=0.025; the sequence concatenates the CLIP-L/CLIP-G embeddings (77 positions) with the T5 embeddings (256 positions) Raffel et al. (2020), and the GenEval prompts studied here contain 16–24 effective tokens in total, so the large majority of repaired positions are padding. On SD3 (CFG=7.0) PR is applied to both passes; FLUX.1-Schnell uses no CFG, so a single pass is repaired. PR modifies the full text sequence, including the CLIP token embeddings, but not the CLIP pooled projection. The rotation R_l is computed once at the first denoising step and held fixed across the trajectory, while origin features are recaptured at every step (Appendix A).

Signal amount. For a configuration repairing k_l tokens in layer l at weight w, the per-layer and total signal are

$$s_l = k_l\, w, \qquad C = \sum_{l=3}^{24} k_l, \qquad S = \sum_{l=3}^{24} s_l = w\, C, \tag{4}$$

where s_l is the per-layer signal, C the number of repaired token-layer pairs, and S the total signal; full PR injects $7326\, w$ (C=7326), LASP (split $L9$) $2451\, w$ (C=2451). Eq. 4 is a descriptive summary of the linear regime, not a generative predictor: within that regime the score rises with S, but different (k, w) pairs reaching the same S can score differently, and beyond the regime S does not predict collapse (Appendix C). Sections 4.1–5.2 characterize how the GenEval score depends on k and w: the injection weight w is the dominant lever, while the token count k is redundant over an intermediate band; the safe operating window is bounded by a collapse at high w whose threshold $w^*(k)$ is non-monotonic in k.

3.2 Experimental Setup

All experiments use SD3-medium Esser et al. (2024) and FLUX.1-Schnell Black Forest Labs (2024), evaluated on GenEval Position Ghosh et al. (2023) (100 prompts, 1000 images per condition, official Mask2Former pipeline Cheng et al. (2022)) and T2I-CompBench spatial subsets Huang et al. (2023). SD3-medium uses 28 rectified-flow steps with classifier-free guidance Ho & Salimans (2022) at scale 7.0 and 1024×1024; FLUX.1-Schnell uses 4 distilled steps with no CFG (full configuration in Appendix E). Unless noted, results are 10-seed on SD3-medium, with seed(prompt, sample) $= 42 +$ prompt_idx$\times1000 +$ sample_idx, sample_idx $\in \{0..9\}$; paired tests share prompts and seeds. All reported p-values are paired Wilcoxon signed-rank tests over prompts, confidence intervals are paired bootstrap over prompts, and full test statistics are reported in Table 17; we mark significance in the text only where a claim depends on it. Full PR and uniform 8-token (freq_top8) serve as baselines. Two families of confirmatory tests are pre-specified and Holm-corrected (the four depth comparisons and the six LASP weight-sweep comparisons); all other contrasts are exploratory and reported uncorrected. Our reproduced base (29.9%) exceeds the 21.5% reported in Yao et al. (2026), likely reflecting checkpoint-version differences; all comparisons are therefore within-setup. GenEval Position can be decomposed into its detection and relation stages (Appendix C); all primary claims use the official Mask2Former score.

4 Deconstructing PR Redundancy: A Dominant Axis and a Secondary Trend

Our ablations reveal a dominant form of redundancy and a secondary trend. The first is a **token-redundancy/weight axis**: w dominates, token count is redundant from about eight positions up to the effective pool size, and the safe window is bounded by a collapse whose threshold is non-monotonic in token count. The second is a **depth trend at the default weight**: dense repair is most effective over the early-to-middle block, while later layers can be sparse. We characterize each in turn.

4.1 Signal Amount and Repair Location

If degradation were localized to particular tokens, repairing only those should recover most of PR's gain. SD3 conditions on a single 333-position sequence (CLIP 77 + T5 256), which PR repairs in full. We ask how *many* positions are needed and whether their *content* matters; the encoder-specific breakdown is deferred to Appendix A.

We first restrict repair to a single linguistic category—prepositions, nouns, or shape nouns (Table 5, Appendix A). No category recovers full PR: repairing prepositions matches baseline and nouns give a small gain. A shape-noun-gated variant (`pr_shape2`) reaches 33.4%: it repairs the last shape noun when one is present (44/100 prompts) and otherwise falls back to unrestricted repair. On the matched subset it reaches 33.6%, against 32.7% for the baseline and 36.4% for full PR, so one semantic position carries some signal but is far from sufficient. Category-gated PR (Appendix A) confirms that no single category suffices, so spatial information is not localized to a category.

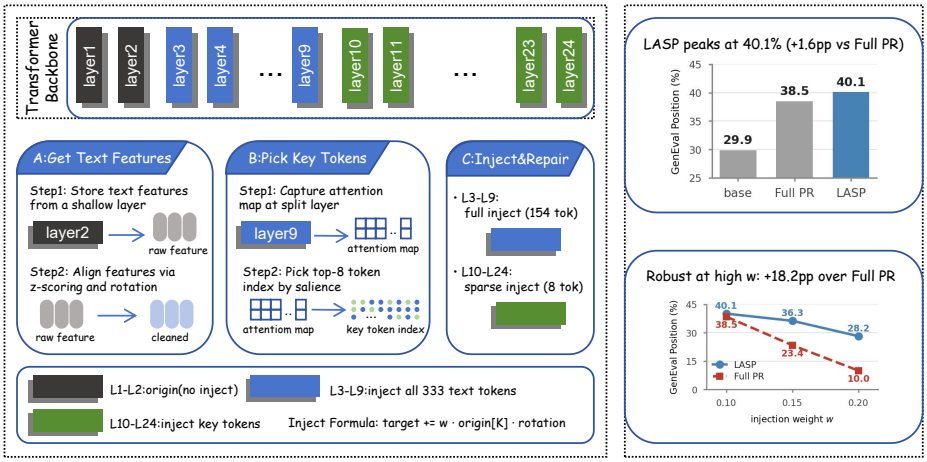

Figure 1: An early-full/late-sparse schedule (LASP), shown as one instantiation of the depth axis. Left: full repair on early layers (blue) and sparse 8-token repair on late layers (green). Right: its weight response relative to full PR.

We next vary how eight positions are chosen (Table 8, Appendix A). Five criteria—random, frequency-matched, part-of-speech-based, and attention-salience—cluster within 1.6pp, and the k-ablation is flat for $k \geq 8$ except a dip at $k=2$. Within this pool the choice of positions is thus largely immaterial once k is large enough; at very small k, however, identity still matters—the single shape-noun position lifts its matched subset to 33.6% whereas four frequency-based positions stay at or below baseline (Appendix C).

The *content* of the repaired positions matters more than their sequence position. Eight padding positions reach only 28.8%, significantly below real-token positions ($p=0.0003$) and no better than baseline: repairing content-free positions adds perturbation without supplying repair signal. This mirrors the padding-budget observation of Appendix C: padding carries little attention, so repairing it adds volume without signal. The CLIP-vs-T5 split is analyzed in Appendix A.

4.2 The Depth Trend: LASP and Early-Block Repair

PR applies repair uniformly across all 22 layers, but it is not obvious that every layer needs the same treatment. To test this, we evaluate LASP, a split-point heuristic (Figure 1): early layers receive full PR as a degeneration gate preserving clean text features; late layers receive sparse 8-token PR. With the split at L9, full PR is applied to L3–L9 and sparse PR to L10–L24 (1-indexed, with the code's block index one lower).

At the split layer L_{split}, LASP selects the top-K text tokens by their total visual-to-text attention mass $\mathcal{K}_{\text{sel}}$ (eq. 5, Appendix B), then applies full PR on layers $l \leq L_{\text{split}}$ and sparse PR restricted to $\mathcal{K}_{\text{sel}}$ on layers $l > L_{\text{split}}$ (eq. 6, Appendix B). Full PR is the special case $L_{\text{split}} \geq 24$; the selection is recomputed per denoising step, while R_l stays fixed. We sweep the split point across L7, L9, L11.

Table 1 reports the depth controls: LASP with split at L9 (full PR on L3–L9, sparse PR on L10–L24) reaches 34.4%, comparable to uniform 8-token repair and approaching full PR while repairing only a third of the token-layer pairs.

To quantify where repair matters (Table 1), we compare layer configurations with a single pre-declared test: a binomial GEE clustered by prompt, Holm-corrected over the four depth comparisons.

The pre-declared test and its confound. Pure late-layer PR (L9–L24, full tokens) reaches 33.3%, a significant improvement over baseline (Holm $p=0.011$) that is statistically indistinguishable from full PR. Early-layer PR (L3–L8) reaches only 31.4%, significantly below full PR (statistics in Ta-

Table 1: Depth controls (SD3-medium, **10-seed**, ±1 SD). Span-matched controls show that dense early repair matches full PR while late-only repair is no better than baseline: 11 layers each (L3–L13 vs. L14–L24) gives 35.4% vs. 30.8%, and 6 layers each (L3–L8 vs. L19–L24) gives 31.4% vs. 29.0%. The earlier asymmetric contrast (L3–L8 vs. L9–L24) confounds position with span. LASP approaches full PR at a third of the pairs, and a uniform allocation at matched volume performs comparably at the default weight (Table 17).

Condition	GenEval Position (%)	Token-layer pairs
Full PR	34.6 ± 4.18	7326
Uniform 8-token PR (freq_top8)	33.4 ± 2.97	176
LASP, split at L7	32.0 ± 2.45	1801
LASP, split at L9	34.4 ± 4.39	2451
LASP, split at L11	33.7 ± 3.38	3101
Early-layer only (L3–L8, full)	31.4 ± 2.54	1998
Late-layer only (L9–L24, full)	33.3 ± 3.47	5328
Late-layer only, matched (L19–L24)	29.0 ± 2.53	1998
Early half, full (L3–L13)	35.4 ± 2.29	3663
Late half, full (L14–L24)	30.8 ± 2.75	3663
Uniform 111-token (volume-matched)	35.2 ± 3.19	2442

ble 17). These windows are not span-matched (6 vs. 16 layers, 1998 vs. 5328 repaired pairs), so the late-layer lead may reflect more repaired signal rather than its position; we retain the pre-declared result as confounded and resolve it with span-matched controls (next paragraph).

Span-matched controls reverse the direction. With 11 layers per side, full repair over the early half (L3–L13) reaches **35.4%**, matching full PR and improving on baseline by $+5.5$pp, whereas full repair over the late half (L14–L24) reaches only 30.8%, no better than the unrepaired baseline. The early-vs-late gap is $+4.6$pp (p=0.0008). A second matched window at 6 layers per side (L3–L8 vs. L19–L24, 1998 pairs each) reproduces the direction (31.4% vs. 29.0%), with late-only repair again no better than baseline. The effective repair region is the early-to-middle block, not the late layers—a reversal that shows how span-confounded layer ablations can invert the direction of an effect.

Volume-matched control. A uniform allocation at the same signal volume performs comparably, and we cannot detect an ordering among the allocation schedules at this sample size (Tables 1 and 17): we therefore treat depth allocation as a secondary, largely interchangeable axis, and select LASP for its cost and cross-architecture validation rather than for superiority.

Both the split (L9) and the weight used below (w=0.10) are the best of coarse one-dimensional sweeps (split L7/L9/L11 and $w \in \{0.025, \ldots, 0.20\}$), selected post hoc rather than pre-registered. Dense early repair is what makes sparse late repair affordable.

The next section refines the weight axis and addresses an alternative mechanistic explanation.

5 Analysis and Cross-Validation

5.1 The Inverted-U of Injection Weight

The injection weight w is the dominant lever of PR's gain, yet standard PR fixes it at 0.025; how does the score respond to amplification, and where does it become destructive?

We sweep $w \in \{0.01, 0.025, 0.05, 0.10, 0.15, 0.20\}$ with the token set fixed at 8, using frequency-based selection (Figure 2a and Table 3). The resulting curve is an inverted U: performance rises to a broad plateau over moderate weights and drops sharply to 28.6% at $w = 0.20$, below the 29.9% unrepaired baseline. The plateau is broad, spanning w=0.05 to w=0.15, and the 8-token curve stays above full PR's default-weight score throughout it. Token count shifts the collapse boundary but not

this qualitative response (Table 2). The injection weight therefore acts as the dominant lever, and the token count compensates only within the linear regime without moving the boundary.

Amplification up to w=0.10 raises the score; w=0.20 distorts semantics and degrades object detection. The 8-token optimum at w=0.10 (35.8%) is numerically above full PR's default-weight 34.6%. The safe window is architecture-dependent (FLUX tolerates higher weights, Section 5.4).

The collapse boundary is non-monotonic in token count. Table 2 reports the collapse boundary of the $k \times w$ grid on SD3, with $w^*(k)$ defined as the first grid weight at which the score falls below the unrepaired baseline: $w^*(4)$=0.05 (29.1% at w=0.05 and 28.5% at w=0.10, both below the 29.9% baseline), $w^*(8)$=$w^*(16)$=$w^*(32)$=0.20, and w^*_{full}=0.15. Both extremes collapse earlier than the intermediate band—too few tokens create partial inconsistency, too many tokens spread across late layers overwrite representations (Section 5.2)—and only the band degrades least at w=0.20; Section 6.2 traces both failure modes. The full regime analysis is in Appendix C.

Table 2: Collapse boundaries in the $k \times w$ plane (GenEval Position, **10-seed**, 1000 images per cell). Bold: the effective band (the k=32 column injects padding positions).

w	k=4	k=8	k=16	k=32	**full (333)**
0.15	23.7	**34.8**	**33.4**	**34.3**	23.4
0.20	21.7	**28.6**	**27.4**	**24.9**	10.0

5.2 The Bounded Regime Across Repair Schedules

We now ask whether this regime is specific to one schedule or generic across them (Table 3).

Table 3: Weight sweep on GenEval Position (**10-seed**, ±1 SD). Full PR crashes at $w \geq 0.15$; LASP peaks at 40.1% (w=0.10).

w	**Uniform 8-tok**	**LASP**	**Full PR**
0.025	33.4 ± 3.0	34.4 ± 4.4	34.6 ± 4.2
0.05	35.3 ± 3.2	38.1 ± 4.0	38.5 ± 4.0
0.10	**35.8** ± 2.5	**40.1** ± 3.8	**38.5** ± 3.3
0.15	34.8 ± 3.4	36.3 ± 2.6	23.4 ± 3.6
0.20	28.6 ± 2.4	28.2 ± 2.8	10.0 ± 3.5

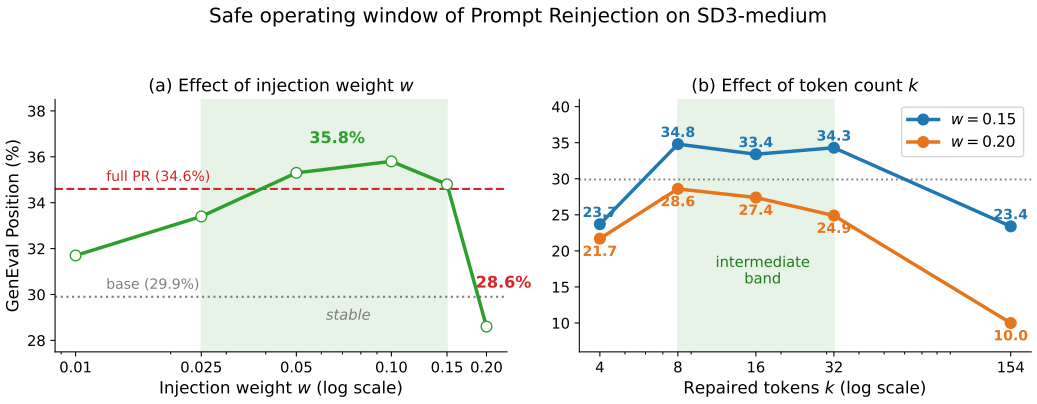

Figure 2: Safe operating window of Prompt Reinjection on SD3-medium, GenEval Position, 10-seed. (a) Inverted-U over injection weight w for 8-token repair. (b) Collapse boundary over token count k at w=0.15 and w=0.20; shaded: the effective band.

Table 3 shows three patterns. First, moderate amplification improves every configuration. Second, collapse is a phase transition whose threshold is *non-monotonic* in token count (Table 2): the effective band survives w=0.15 while full PR and k=4 fall below baseline, and at w=0.20 every

configuration is below baseline, the band degrading least. The k=32 column admits padding, deepening the collapse without changing its threshold (Appendix C), whereas padding injection alone amplifies but does not cause the collapse.

Third, the schedules separate under amplification (Table 3): full PR collapses to 23.4% at w=0.15 and 10.0% at w=0.20, whereas the sparse schedules degrade far less (uniform 8-token: 34.8%/28.6%; LASP: 36.3%/28.2%)—sparsity, not depth, is what survives high weight. We feature LASP for its pair efficiency and cross-architecture validation: it matches full PR at the default weight with a third of the pairs, peaks at 40.1% at w=0.10 (Holm p=0.020 vs. base), and stays above full PR at amplified weights, using 66% fewer pairs. Amplification to w=0.10 buys large spatial gains at little quality cost—HPSv2 at most ∼3% below base and CLIP slightly above—whereas past the boundary both the spatial and the image-quality metrics fall (Appendix D); the collapse is therefore a general over-injection failure, not a spatial-only one.

5.3 Does the Injected Content Matter?

The sections above vary *where* repair is applied while leaving the injected features fixed. We now do the converse: at the same eight `freq_top8` positions and w=0.025, we vary only *what* is injected, replacing the captured shallow features with norm-matched Gaussian noise, the same features under a fixed random permutation of the hidden dimensions, or a constant token-invariant vector, injecting into both CFG passes exactly as the default recipe does. If the benefit came from a generic perturbation of the text stream rather than from the content of the injected features, all three substitutes should match the real features.

They do not (Table 4). The real-feature condition reaches 33.4%, significantly above the unrepaired baseline, whereas norm-matched noise, a fixed permutation of the hidden dimensions, and a constant vector all fall significantly below it ($p \leq 0.007$). The gain therefore requires the real shallow features: it is content-specific, rather than a generic perturbation or a distributional anchor of the text stream.

Table 4: Origin-content control at a fixed set of eight positions and weight w=0.025 (10-seed GenEval Position, 1000 images per condition). Only the injected content is varied.

Injected content	**GenEval Position (%)**
Real shallow features (PR)	**33.4**
Random permutation	30.0
Norm-matched noise	28.6
Content-free constant	27.9
None (baseline)	29.9

5.4 Cross-Benchmark and Cross-Architecture Validation

On T2I-CompBench spatial subsets (10-seed UniDet), LASP approaches full PR (Appendix E).

We replicate the core findings on FLUX.1-Schnell with four conditions (Table 19): the default-weight contrasts are directional, with smaller margins and weaker significance, whereas amplified repair is clearly significant, by 5.3 points at w=0.10 (p=0.0016). On FLUX the sparse stage uses static freq_top8 selection (Appendix E). The sparse sweep improves monotonically to a plateau at w=0.30–0.40, and a full-PR sweep also plateaus with no collapse (Appendix E), so FLUX's immunity to over-injection is independent of token count. A shorter schedule does not explain FLUX's tolerance: SD3 collapses at the same weight at 12 steps (Appendix E). The collapse is therefore SD3-specific, confounded with the concurrent absence of classifier-free guidance and the T5-only encoder; FLUX.1-dev already declines under amplified w Yao et al. (2026). Sparse repair does not degrade non-spatial attributes (Appendix D).

SD3's classifier-free guidance batches the conditional and unconditional embeddings into a single forward pass; isolating each branch (Table 14) shows that the repair signal flows through the con-

ditional text path—conditional-only injection matches full PR at the default weight but is the most fragile under amplification—while null-only injection is worse than no repair.

6 DISCUSSION

6.1 SYNTHESIS AND IMPLICATIONS

Our three findings—weight dominance, the early-to-middle depth trend, and content-specificity—are diagnostic, and they support a practical instantiation: as few as 8 positions suffice (42× reduction), and LASP reaches **40.1%**, matches full PR at equal weight (Table 17), and remains robust where full PR collapses (36.3% vs. 23.4% at w=0.15) with 66% fewer token-layer pairs—a reduction in signal volume, not wall-clock speedup (Appendix E). We do not claim LASP is optimal: a volume-matched uniform allocation performs comparably at slightly lower cost (Table 1), and we feature LASP because it is the only schedule we validate across benchmarks and architectures that matches full PR at the default weight at a third of the pairs. The repair is content-specific: replacing the injected shallow features with noise, a permutation, or a constant removes the gain (Section 5.3). Core findings hold on T2I-CompBench and transfer to FLUX.1-Schnell, while finer-grained ablations are SD3-only. Over-injection is not specific to the spatial composite: at the recommended w=0.10 the spatial gain is large and the quality cost is small (HPSv2 at most ∼3% below base; Appendix D), whereas at w=0.20 both the Position score and the quality metrics fall. Our contribution is diagnostic: PR remains the intervention, and we reveal when full-scale repair is redundant and when amplification makes it harmful, reframing it from a black-box tool into an interpretable probe of MMDiT text degradation. One design implication follows: because joint attention is not masked at padding positions, interventions over the full padded sequence spend most of their budget on content-free positions, so masking padding or introducing dedicated register tokens Darcet et al. (2024) may recover that budget. These constraints are not specific to PR: any intervention that writes text-side residuals into a joint-attention stream inherits the same padding budget and is bounded by the same two failure modes (Section 6.2), so our characterization describes a family of interventions rather than a single recipe.

6.2 WHY THE BOUNDARY EXISTS, AND HOW TO USE IT

The window is two-sided for two different reasons. PR adds a fixed shallow-feature direction to the residual stream: the per-position magnitude is set by w, the number of perturbed positions by k_l, and the layer's own representation does not scale with the injected term. Too few repaired positions leave the repaired and untouched tokens inconsistent, which strong injection amplifies (*under-repair mismatch*); too many positions, or too large a weight, let the injected term overwrite the representation (*over-write*). This also explains why content and region matter: only the real shallow features carry task-relevant semantics, and the early-to-middle block is where they have drifted least (Appendix B). In practice, repair a handful of real positions in the leading block at a moderate weight, leave padding positions untouched, and re-tune the weight per architecture, since the collapse threshold is SD3-specific.

6.3 LIMITATIONS AND OPEN QUESTIONS

Our characterization is within-setup on dual-stream MMDiTs and spans the PR injection family (Procrustes alignment with z-score anchoring). Other injection families remain to be tested. With 100 prompts our tests have limited power; we report paired differences with confidence intervals and avoid equivalence claims. The depth evidence is mixed by design: the pre-declared GEE test favored late-layer repair (Table 17), but its windows are not span-matched, and a matched 11-vs-11-layer control reverses the direction (early half 35.4% vs. late half 30.8%). We therefore report depth as a directional early-to-middle trend rather than an established late-layer effect. A uniform allocation at matched volume performs comparably (Table 1); we therefore select LASP for cost and cross-architecture validation, not superiority, and make no optimality claim for any configuration. Attention entropy is uninformative at fixed token count (Appendix F). Open questions remain: whether FLUX.1-dev reproduces SD3's inverted-U, whether the vertical-relation gains reflect relation-specific drift, and whether the findings extend beyond GenEval's templated prompts.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools (large language model assistants) only for tasks with recommended disclosure: assistance with manuscript drafting and revision, including editing for readability and suggestions on the presentation of contributions and narrative positioning; generating non-core code snippets for experimental scripts and analysis tools under the authors' direction; and providing language and style suggestions for the abstract, introduction, and discussion sections. No task requiring mandatory disclosure applies to this work.

We have not used generative AI tools for generating synthetic datasets, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, implementing experimental methods, designing experimental methodology, interpreting experimental results, or creating scientific figures. All scientific judgments, method design, statistical analyses, and final text are authored and verified by the human authors.

We have reviewed all AI-assisted work. All AI-suggested text was edited and verified by the authors. AI-generated code was reviewed, integrated, and tested by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details.

REFERENCES

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

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

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

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

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

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.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K. Wong. Taca: Rethinking cross-modal interaction in multimodal diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):1–67, 2020.

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.

Nurit Spingarn, Noa Cohen, Tamar Rott Shaham, and Tomer Michaeli. Text-to-image models need less from text encoders than you think. *arXiv preprint arXiv:2606.03715*, 2026.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. URL https://aclanthology.org/2023.acl-long.310.

Binxu Wang, Jingxuan Fan, and Xu Pan. Circuit mechanisms for spatial relation generation in diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yuxuan Yao, Yuxuan Chen, Hui Li, Kaihui Cheng, Qipeng Guo, Yuwei Sun, Zilong Dong, Jingdong Wang, and Siyu Zhu. Prompt reinjection: Alleviating prompt forgetting in multimodal diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2026.

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)*, 2023.

A TOKEN REDUNDANCY: QUANTITY OVER IDENTITY

Section 4.1 characterizes repair quantity and location: within the sequence's leading block, repair suffices from about eight injected positions up to the effective pool size, and within the set of real tokens the choice of positions only modulates per-position signal density (identity matters only at very small k), whereas restricting repair to the T5 block is ineffective at the default weight. Three controls support this: token substitution (Table 5), five position-selection criteria (Table 8), and k-ablation (Table 7); the collapse boundaries across position counts and the padding mechanism are analyzed in Appendix C (Table 2).

Table 5: Token-substitution results on GenEval Position (10-seed). Repair is restricted to the positions of one category (prepositions, nouns, or shape nouns); the shape-noun variant falls back to unrestricted repair when no shape noun is present.

Condition	**GenEval Position (%)**
Base (no repair)	29.9
Full PR (all tokens repaired)	34.6
Repair prepositions (pr_prep)	29.9
Repair nouns (pr_noun)	30.5
Shape-noun-gated PR (pr_shape2)†	**33.4**

† `pr_shape2` repairs the last shape-noun position when one is present (44/100 prompts) and otherwise performs unrestricted full repair; on the 44-prompt matched subset it scores 33.6% against 32.7% (base) and 36.4% (full PR), so one semantic position carries some signal but is far from sufficient.

Table 5 shows that no single linguistic category suffices, and even the single shape-noun position carries only part of the gain (33.6% on its matched subset), consistent with spatial information being distributed rather than localized to a category.

Category-gated PR (CG), which repairs prepositions at full strength, attenuates nouns by a factor γ, and leaves all other tokens untouched, also fails to recover full PR at any setting (Table 6).

Table 6: Category-gated PR on GenEval Position (**5-seed**). No γ exceeds full PR (34.6%).

γ **(noun attenuation)**	**GenEval Position (%)**
0.10	31.0
0.20	30.8
0.30	29.6
0.50	30.2
0.70	29.4

Table 7: Entropy-guided k-ablation on GenEval Position (**10-seed**, ± 1 SD). The injected positions lie in the sequence's leading block.

k	**GenEval Position (%)**	k	**GenEval Position (%)**
0 (base)	29.9 ± 1.70	6	32.7 ± 3.3
1	29.8 ± 2.2	8	**33.4** ± 3.4
2	29.5 ± 2.8	12	32.5 ± 3.1
4	31.5 ± 2.5	16	32.9 ± 3.8

Performance is approximately flat for $k \geq 8$ (32.5–33.4% for $k = 8, 12, 16$), with no further gain once the selection pool is covered. The 333-position default therefore repairs far more positions than needed. See Section 4.1 for discussion.

Table 8: Strategy ablation (GenEval Position, **10-seed**, ± 1 SD). Eight positions in the sequence's leading block, chosen by five criteria; narrow spread.

Selection strategy	GenEval Position (%)
Base (no repair)	29.9 ± 1.70
Random	31.8 ± 3.09
Preposition-top8	32.7 ± 4.82
Noun-top8	33.0 ± 2.57
Frequency-based (freq_top8)	**33.4** ± 2.97
Entropy-guided	**33.4** ± 3.44
Full PR (all 333 positions)	**34.6** ± 4.18

The five selection strategies cluster within 1.6pp of each other (best-vs-random Wilcoxon p=0.265, n=100 paired prompts), confirming that token count, not selection strategy, drives repair effectiveness.

LN-only variant. We additionally tested `pr_shape2_lnonly`, which uses the same shape-noun gating as `pr_shape2` (including its full-repair fallback) but skips the orthogonal Procrustes alignment, using layer-normalization-only anchoring. This variant isolates the contribution of the Procrustes rotation.

Table 9: Anchor (distribution anchoring) vs. Rotation (orthogonal Procrustes) ablation on GenEval Position (10-seed). All variants use the same shape-noun gating; only the alignment pipeline differs.

Anchor	Rotation	GenEval Position (%)
×	×	29.9
✓	×	32.4
✓	✓	**33.4**

Table 9 decomposes the alignment pipeline. Distribution anchoring alone (LN-only) recovers most of the gain (32.4%, +2.5pp over base); adding orthogonal Procrustes rotation contributes an additional +1.0pp (33.4%). Both components are beneficial, but anchoring dominates in this setting. Yao et al. Yao et al. (2026) report that rotation provides the larger improvement when cross-layer geometric mismatch is significant; in our setting, anchoring contributes more. The bounded window holds under both alignment variants at the tested weight.

Encoder Block and Repair Region on SD3

SD3 concatenates the CLIP (77 positions) and T5 (256 positions) embeddings into a single 333-position sequence. Position-selection criteria are consumed directly against this sequence, so criteria computed on the T5 tokenization fall within the leading (CLIP) block; the controls of Section 4.1 therefore probe that block. We separate the two blocks explicitly (Table 10; 10-seed). Eight positions in the leading CLIP block reach 33.4%, matching full PR (34.6%), whereas eight real positions in the T5 block reach 30.9% at the default weight w=0.025, and recover to 33.1% at w=0.05 and 34.9% at w=0.10, matching full PR. The CLIP prefix is thus the more efficient repair region per unit injection weight on SD3, while the T5 block requires heavier amplification; injection weight can therefore compensate for a less efficient region. This block structure is specific to SD3's dual text encoder, as FLUX.1-Schnell uses a single T5 encoder and has no analogous split (Appendix E).

Implementation Details

PR is formalized in Section 3 (Eqs. 1–3). The per-layer rotation R_l is computed once at step 0 and held fixed across denoising steps, while origin features are recaptured per step; the origin layer (2) and this asymmetric design follow Yao et al. Yao et al. (2026), who report layer 2 as optimal across SD3/FLUX. Our focus is on token/layer/weight redundancy rather than origin-layer optimization.

Table 10: Encoder-block repair on SD3-medium (**10-seed**, 100 prompts). The eight-position conditions inject the same number of token-layer pairs (176); only the block and content differ.

Condition	w	**GenEval Position (%)**
8 real positions, CLIP block	0.025	33.4
8 padding positions, CLIP block	0.025	28.8
8 real positions, T5 block	0.025	30.9
8 real positions, T5 block	0.05	33.1
8 real positions, T5 block	0.10	34.9
8 padding positions, T5 block	0.025	30.4
Full PR	0.025	34.6

B THE DEPTH TREND AND REPRESENTATIONAL EVIDENCE

Section 4.2 establishes the depth trend; this appendix collects the layer controls and the representational probe behind it.

LASP selection and injection. Let $A^{(l)} \in \mathbb{R}^{N\times T}$ denote the visual-to-text attention matrix at layer l (summed over heads). At the split layer, LASP selects the top-K text tokens by total visual attention mass:

$$\mathcal{K}_{\text{sel}} = \underset{j\in[T]}{\arg\,\text{top-K}} \sum_i A_{i,j}^{(L_{\text{split}})}, \tag{5}$$

then applies PR with a layer-dependent token mask:

$$h_l^{\text{txt}} \leftarrow \text{LN}\Big(\tilde{h}_l^{\text{txt}} + w\left(M_l \odot \tilde{h}_{L_{\text{ori}}}^{\text{txt}}\right)R_l\Big)\cdot\sigma_l + \mu_l, \quad M_l = \begin{cases}\mathbf{1}_T & l \le L_{\text{split}} \\ \mathbf{1}[j\in\mathcal{K}_{\text{sel}}] & l > L_{\text{split}}\end{cases} \tag{6}$$

where M_l is a token mask (all-ones for $l \le L_{\text{split}}$, the indicator of $\mathcal{K}_{\text{sel}}$ for $l > L_{\text{split}}$), $\odot$ broadcasts over the hidden dimension, and all other notation follows Section 3. The top-K selection is restricted to effective (non-padding) tokens; the padding-injection variant of Appendix C is the only exception.

Table 11: Depth control experiments (10-seed). Rows whose label ends in `_L8` (`full_early_L8`, `full_late_L8`, `lasp_L8`) use the code's 0-based block index (code L8 = paper split L9); the span-matched `L13` halves repair 11 layers each. OD and CSA decompose the official Mask2Former score (Appendix C).

Condition	**OD**	**CSA**	**Official**
Base	0.789	0.379	29.9%
full_early_L8 (L3–L8, 333 tok)	0.819	0.383	31.4%
full_late_L8 (L9–L24, 333 tok)	0.854	0.390	**33.3%**
full_early_L13 (L3–L13, 333 tok)	0.847	0.418	**35.4%**
full_late_L13 (L14–L24, 333 tok)	0.829	0.372	30.8%
lasp_L8	0.847	0.406	**34.4%**
Full PR	0.864	0.400	**34.6%**

Table 11 decomposes the GenEval Position score into OD and CSA components for the depth control experiments. The asymmetric L8 controls alone would suggest that late-layer repair recovers most of full PR's gain (33.3% vs. 31.4%), but the span-matched L13 controls reverse this: dense early repair (L3–L13) reaches 35.4%—above full PR's 34.6%—while late-only repair (L14–L24) falls to 30.8%, no better than baseline. The decomposition shows the early half improving both stages (OD $0.789 \to 0.847$, CSA $0.379 \to 0.418$), whereas the late half improves OD only marginally (0.829) with CSA below baseline (0.372). LASP's hybrid design (0.847 OD, 0.406 CSA) combines a dense early stage with sparse late repair, consistent with the early block carrying the gain and sparse late repair preserving it at a third of the pairs.

LAYER-WISE TOKEN DECODABILITY

We train a per-layer multinomial logistic-regression probe (standardized features, C=1) to decode each T5 sub-token's semantic category (spatial-relation, noun, color, numeral) from the text-branch representation of every MMDiT layer. Features are collected in a single denoising step from random latents (no CFG), following the protocol of Yao et al. (2026). Each probe is evaluated with 5-fold cross-validation stratified by prompt (tokens of a prompt never appear in both training and test folds), stricter than the single held-out split of the original protocol; reported accuracies are fold means. The identical token set is used across all conditions, so accuracies are directly comparable. Under 8-token repair the injected positions lie in the sequence's leading block, so all semantic T5 tokens in the evaluation set are *unrepaired*. Figure 3 shows the layer-wise decodability curves.

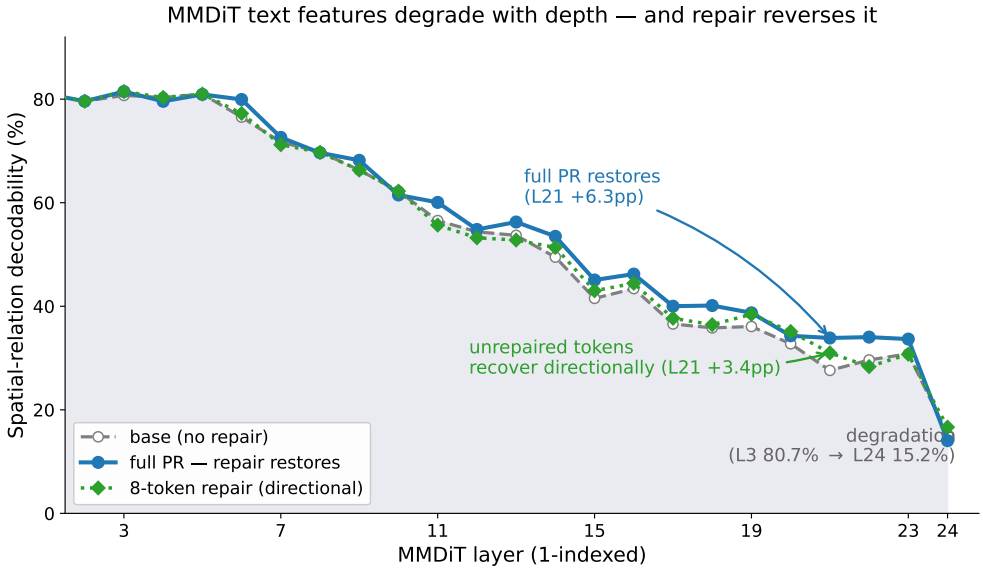

Figure 3: Text-feature degradation and repair in SD3-medium (layer-wise probe). Spatial-relation decodability collapses with depth under base generation (dashed) and is raised by full PR (blue), with sparse 8-token repair (green) showing the same tendency in late layers; the behavioral improvement is significant (p=0.005; Appendix B).

Table 12 reports per-layer accuracy by category (spatial-relation rows are 5-fold CV means; noun/color/numeral are same-set references). All categories both degrade with depth and are recovered by repair; the degradation and recovery are largest for spatial-relation (L9 66.3% → L24 15.2% under base generation), consistent with spatial reasoning being the dimension most sensitive to text drift. Full PR raises spatial decodability in late layers (L21 33.9% vs. base 27.6%); 8-token repair shows the same tendency for unrepaired tokens (L21 31.0%; per-prompt paired tests $p \geq 0.18$). Noun and color accuracies are near-identical across conditions.

C INJECTION WEIGHT AND THE OPERATING WINDOW

The official GenEval Position evaluation can be decomposed into two stages: OD (the fraction of images in which both prompted objects are detected by the official Mask2Former pipeline) and CSA (the fraction of those in which the spatial relation is also correct). The two components multiply to the official score by construction, so the decomposition redistributes—rather than independently re-measures—the benchmark score, and CSA is defined only on the detection-passing subset. We use it as diagnostic scaffolding for hypothesizing failure modes; all primary claims use the undecomposed official score.

At w=0.20, OD collapses (0.672) while CSA among the surviving detections rises (0.426 vs. 0.379 at baseline); at moderate weights (w=0.05–0.10), OD reaches 0.880–0.887, above both the baseline (0.789) and full PR (0.864). The composite drop at w=0.20 is therefore mediated by object detection

Table 12: Layer-wise token decodability by category (all 553 GenEval prompts). Spatial-relation rows are 5-fold cross-validation means (prompt-stratified); noun/color/numeral are same-set references. Spatial-relation shows the largest, sharpest degradation and recovery.

Layer	Base	Full PR	8-tok	Layer	Base	Full PR	8-tok
Spatial-relation (5-fold CV)				*Color*			
L9	66.3	68.2	66.3	L9	63.3	64.4	64.4
L13	53.7	56.3	52.8	L13	52.2	52.2	54.4
L17	36.6	40.0	37.7	L17	52.2	50.0	48.9
L21	27.6	**33.9**	31.0	L21	57.8	55.6	53.3
L24	15.2	14.0	16.7	L24	20.0	20.0	17.8
Noun				*Numeral*			
L9	93.0	93.7	93.0	L9	100.0	100.0	100.0
L13	92.0	91.7	92.2	L13	100.0	100.0	100.0
L17	89.2	91.2	90.7	L17	95.8	95.8	95.8
L21	91.7	88.7	91.5	L21	95.8	95.8	100.0
L24	75.4	73.9	75.4	L24	16.7	20.8	12.5

Table 13: OD and CSA decomposition for the uniform 8-token weight sweep (**10-seed**).

Condition	OD	CSA	Official
Base	0.789	0.379	29.9%
8tok_w0.01	0.821	0.386	31.7%
freq_top8 ($w = 0.025$)	0.850	0.393	33.4%
8tok_w0.05	0.880	0.401	35.3%
8tok_w0.10	**0.887**	0.404	35.8%
8tok_w0.20	0.672	**0.426**	28.6%
Full PR	0.864	0.401	34.6%

rather than by a collapse of spatial reasoning among detected pairs—though the CSA rise must be read with care, since it is measured only on the images that still pass detection at w=0.20.

The dissociation between OD and CSA across the weight sweep is consistent with a finer-grained structure in which object detection and spatial understanding respond to injection strength at different rates: OD peaks earlier and collapses harder, while CSA does not. Because CSA is conditioned on detection, however, the surviving subset at high weight is smaller and plausibly easier, so we treat this pattern as a diagnostic hypothesis rather than as evidence that spatial reasoning tolerates over-injection; the official composite score is the authoritative metric.

The collapse pattern in the main text (Table 3; Table 2 in Section 5.1) is consistent across the weight sweep: the injection weight w is the dominant lever (for $k\geq 8$ the score peaks at w=0.05–0.10 and collapses by w=0.20), and collapse is a phase transition whose weight threshold is non-monotonic in k. The effective band (from eight positions up to the effective pool size) significantly exceeds both extremes at w=0.15 (p<0.001–0.0002) and at w=0.20 (p=0.004–0.016, paired Wilcoxon). Too few tokens fail because strong injection onto a sparse set creates partial inconsistency between repaired and unrepaired tokens (the k=2 dip of Section 4.1 is the same mechanism); too many tokens overwrite layer representations wholesale. Both failure modes are architecture-specific (FLUX.1-Schnell shows neither, Appendix E), so the window bounds PR's operating regime on SD3.

Three regimes and the collapse threshold. The grid decomposes into a **linear** regime ($w\leq 0.05$, $k\leq 32$) where the score rises with $S{=}k\,w$; a **saturated** regime (once k covers the selection pool, marginal gain is zero); and a **collapse** regime ($w \geq w^*(k)$). The per-k collapse threshold $w^*(k)$ is non-monotonic: $w^*(4){=}0.05$, $w^*(8){=}w^*(16){=}w^*(32){=}0.20$, and $w^*_{\text{full}}{=}0.15$ (full PR, 333 positions). Eq. 4 is not a conserved quantity: the matched per-layer-signal pair (k=8, w=0.05, s_l=0.4) scores 35.3% while (k=16, w=0.025, s_l=0.4) scores 32.9%, and S cannot predict the collapse regime at all. Eq. 4 therefore summarizes the linear regime only, and token count and weight are

at best approximate substitutes within it; within the grid, the $w \geq 0.05$ cells use freq_top8 with the corresponding K (so the $k=4$ column injects four frequent function words), while the $w=0.025$ entries of the $k=4$ and $k=16$ columns come from the salience-based k-ablation runs.

Padding-injection mechanism. The $k=32$ grid column injects padding positions to reach 32 tokens (its effective-only variant is identical to $k=16$, since the selector truncates at the effective set; we use the padding variant here), which deepens the $w=0.20$ collapse (24.9% vs. 27.4% for $k=16$) without changing its threshold. MMDiT joint attention does not mask padding positions (SD3's joint-attention block is applied without an attention mask, so padding keys participate), and the 333-position text sequence contains only $\sim$16–24 effective tokens, so roughly 90% of injected positions are padding. Normalizing by position share, attention is concentrated on the effective tokens (2.4–$7.9\times$ their share) while padding receives only 0.5–$0.9\times$ its share (at or below proportional) across the four probed depths (60 prompts of both counting and position types, split-layer single-step forward). Injecting into all positions therefore adds perturbation over a large share of the sequence without adding repair signal, consistent with a general trend in which more injected positions produce a deeper collapse: perfectly monotone at $w=0.20$ ($28.6/27.4/24.9/10.0\%$ for $k=8/16/32/333$) and with a small dip at $w=0.15$ ($34.8/33.4/34.3\%$ for $k=8/16/32$, then 23.4% at full). Padding positions are not ignored by attention, but they carry no repair signal; injecting beyond the effective set adds perturbation without repair benefit.

D Bounded Amplification and Spatial Specificity

CFG branch ablation. SD3's CFG batches the conditional and unconditional text embeddings together; the main text injects both passes. Table 14 isolates each branch.

Section 5.4 reports the branch ablation: conditional-only injection matches full PR at the default weight but is the most fragile under amplification, so the collapse tracks conditional-pass over-injection rather than the null branch.

Table 14: CFG branch ablation on SD3-medium (10-seed, 1000 imgs/condition).

Condition	GenEval Position (%)
Base (no repair)	29.9
Full PR (both passes)	34.6
Conditional pass only	**36.1**
Unconditional pass only	27.0

Table 15: Performance on GenEval Colors, Color+Attribute, and Position subsets (10-seed).

Condition	Colors	Color+Attr	Position
Base (no repair)	82.77%	57.80%	29.9%
8-token PR	86.38%	58.80%	33.4%
Full PR	87.55%	63.40%	34.6%

8-token PR already captures most of full PR's gain on Colors, but lags on Color+Attribute binding, indicating that attribute binding demands more tokens. This asymmetry suggests that different semantic dimensions respond differently to injection strength along the weight axis: simple color attribution saturates early, while compositional attribute binding requires more signal. The token budget is therefore dimension-dependent: the "any 8 tokens suffice" pattern holds for spatial reasoning (Position), while compositional attribute binding needs more.

Image-text alignment and human preference. We also ask whether weight amplification ($w=0.10$, $4\times$ the PR default) trades spatial accuracy for image quality. We measure two complementary metrics: CLIP Radford et al. (2021) image-text similarity (ViT-L/14, global semantic alignment) and HPSv2 Wu et al. (2023) (human preference score, pretrained on human pairwise comparisons; captures fine-grained quality issues like blur/artifact/texture that CLIP misses). Both

are computed on 10-seed GenEval generations (1000 images per condition, Table 16). At the amplified operating point (w=0.10) the spatial composite gains substantially while quality is nearly unchanged: HPSv2 is 0.2883 for LASP and 0.2941 for 8-token repair, against 0.2978 at base (about 1–3% lower), and CLIP is slightly higher (30.76/30.75 vs. 30.00). At the w=0.20 crash condition both metrics fall—HPSv2 to 0.2729 (-8%) and CLIP to 29.34 (-2%)—so over-injection degrades image quality as well, and the failure is not confined to the spatial metric; the detection-stage decomposition above attributes the composite drop to object-recognition failure.

Table 16: Image quality and alignment metrics (10-seed GenEval). CLIP = ViT-L/14 image-text similarity (logit scale; $30 \approx 0.30$ cosine); HPSv2 = human preference score (pretrained model, no real human scoring).

Condition	CLIP	HPSv2	Position
Base	30.00	0.2978	29.9%
Full PR (w=0.025)	30.61	0.3010	34.6%
LASP (w=0.025)	30.25	0.3003	34.4%
8-tok PR (w=0.10)	30.75	0.2941	35.8%
LASP (w=0.10)	30.76	0.2883	**40.1%**
8-tok PR (w=0.20, crash)	29.34	**0.2729**	28.6%

E Cross-Benchmark and Cross-Architecture Validation

Summary statistics

Table 17 collects the pre-declared and exploratory comparisons reported in the main text.

T2I-CompBench

LASP is comparable to full PR on both 2D (0.3379 vs. 0.3375) and 3D (0.4225 vs. 0.4265) benchmarks, so cross-benchmark robustness is not a GenEval artifact. The near-identical scores on T2I-CompBench, achieved with substantially fewer repaired tokens, likewise indicate that a sparse subset of repaired positions suffices, without requiring per-token precision.

FLUX.1-Schnell

MMDiT blocks. SD3-medium: 24 uniform; FLUX.1-Schnell: 19 dual + 38 single = 57.

PR origin layer. L2 (0-indexed 1) for both.

PR target layers. SD3: L3–L24 (22 layers); FLUX: L3–L57 (55 layers).

PR injection weight w. 0.025 (default) for both.

LASP split layer. SD3: L9 (38% early); FLUX: L17 (30% early).

LASP k (sparse tokens). 8 for both.

LASP token selection. SD3: per-step attention salience; FLUX: static freq_top8.

Inference steps. SD3: 28; FLUX: 4 (distilled).

CFG scale. SD3: 7.0; FLUX: none (no classifier-free guidance; single pass; diffusers default embedded guidance 3.5).

Resolution. 1024×1024 for both.

Seeds per prompt. 10 for both.

Section 5.4 reports cross-architecture validation. All FLUX experiments use the Schnell (distilled, 4-step) variant. The generation configuration is as follows (see the side-by-side configuration summary above):

- **Model**: black-forest-labs/FLUX.1-Schnell

Table 17: Per-comparison statistics on GenEval Position (10-seed). *Pre-declared depth family*: binomial GEE clustered by prompt (effect is an odds ratio; unadjusted p shown here, with the family Holm-corrected over the four comparisons and the primary adjusted value reported in Section 4.2). *Exploratory* rows: paired Wilcoxon signed-rank tests over prompts with paired bootstrap confidence intervals (Δ in pp). The pre-registered LASP weight-sweep family (six comparisons) gives Holm p=0.020 for the w=0.10 headline.

Comparison	**Effect / Δ**	**95% CI**	p
Pre-declared depth family (GEE; odds ratio)			
Late-only (L9–L24) vs. base	1.17	—	0.003
Early-layer only (L3–L8) vs. full PR	0.87	—	0.006
Late-only vs. full PR	0.94	—	0.125
Late vs. early (direct)	1.09	—	0.043
Exploratory paired comparisons (Δ in pp)			
Early half (L3–L13) vs. full PR	$+0.8$	$[-1.2, +2.7]$	0.34
Early vs. late half (11/11)	$+4.6$	$[+2.3, +7.0]$	0.0008
Early-layer only vs. late-matched (6/6)	$+2.4$	$[+0.7, +4.1]$	0.0026
Uniform-111 vs. LASP	$+0.8$	$[-0.9, +2.6]$	0.52
Uniform-111 vs. full PR (default weight)	$+0.6$	$[-1.1, +2.3]$	0.48
Early half vs. LASP	$+1.0$	$[-0.9, +2.9]$	0.28
LASP vs. full PR (equal weight, w=0.10)	$+1.6$	$[-1.4, +4.7]$	0.217
LASP vs. full PR (default weight)	-0.2	$[-2.1, +1.6]$	0.83
8-token vs. full PR (default weight)	-1.2	$[-3.2, +0.9]$	0.34
Content control (vs. real features)			
Noise	—	—	0.0004
Permutation	—	—	0.007
Constant	—	—	< 0.0001
FLUX.1-Schnell			
8-token vs. base (default weight)	$+1.4$	—	0.25
LASP vs. full PR (default weight)	-0.4	—	0.27
LASP (w=0.10) vs. base	$+5.3$	—	0.0016

Table 18: T2I-CompBench 2D and 3D spatial scores (10-seed UniDet).

Condition	**Spatial 2D**	**Spatial 3D**
Base (no repair)	0.3158	0.4060
8-token PR (salience-selected)	0.3280	0.4195
LASP (split L9)	**0.3379**	0.4225
Full PR	0.3375	**0.4265**

- **Resolution**: 1024×1024
- **Inference steps**: 4 (schnell default)
- **Classifier-free guidance**: none (Schnell's default; single pass)
- **Memory**: `enable_model_cpu_offload()` (32 GB VRAM)
- **Token selection**: freq_top8 (static; the 8 most frequent English words matched in the T5-tokenized prompt: "the, a, of, and, to, in, is, that"). Unlike SD3 LASP, which dynamically selects top-8 tokens per step via attention salience at the split layer, FLUX LASP uses this static selection throughout generation, as the 4-step distilled schedule offers limited benefit from per-step reselection.
- **LASP split layer**: L17 (at a comparable relative depth; SD3's split at L9/24 $\approx$ 38% early vs. L17/57 $\approx$ 30% in FLUX's absolute block indexing, where 19 dual-stream blocks precede 38 single-stream blocks sequentially)
- **Prompts**: GenEval Position metadata (100 prompts)
- **Evaluation**: Official Mask2Former pipeline

All FLUX generations use 10 distinct seeds per prompt, matching the SD3 protocol.

LASP token salience (SD3). For SD3 LASP, token salience is defined as the mean over attention heads and visual token positions of the softmax-normalized visual-to-text attention weights at the split layer (L9), computed on the conditional CFG pass. The top-8 salient tokens are selected per denoising step and held fixed across all sparse layers (L10–L24) within that step.

LASP cost ratio. SD3 LASP repairs 7 full layers (L3–L9, 333 tokens each) + 15 sparse layers (L10–L24, 8 tokens each) = 2451 token-layer pairs vs. full PR's 7326, i.e., 33.5% of full PR's pairs. FLUX LASP (split L17) repairs 15 full layers (L3–L17, 512 tokens each) + 40 sparse layers (L18–L57, 8 tokens each) = 8000 pairs vs. full PR's 28160, yielding ≈28% cost.

Inference overhead. The "33.5% token cost" refers to token-layer pairs (signal volume), not wall-clock time. Measured on RTX 3090 (SD3-medium, 28 steps, 1024×1024, 5 timed iterations after 2 warmup): base = 9.86 ± 0.04 s/image, full PR = 15.05 ± 0.14 s, LASP = 15.10 ± 0.04 s. LASP and full PR have near-identical latency ($\Delta < 0.1$ s) and identical peak VRAM (6.97 GB), because runtime is dominated by per-step tensor transformations—origin-feature extraction, standardization, the R_l products, and layer normalization across 22 layers for both CFG passes—which all repair variants share; the one-time step-0 Procrustes SVD is only a small part, and the residual injection step (8 vs. 333 positions per layer) contributes negligibly. The practical benefit of LASP is thus token efficiency rather than inference speedup, achieving comparable performance with 66% fewer repaired token-layer pairs.

Validation Results

Table 19: Cross-architecture validation on FLUX.1-Schnell vs SD3-medium (**10-seed**).

Condition	FLUX Acc (%)	SD3 Acc (%)
Base (no repair)	26.2 ± 3.37	29.9 ± 1.70
8-token PR (freq_top8)	27.6 ± 3.01	33.4 ± 2.97
LASP	**28.9** ± 3.96	**34.4** ± 4.39
Full PR	**29.3** ± 3.85	**34.6** ± 4.18

The four-condition comparison (Table 19) transfers the core pattern observed on SD3: 8-token repair (freq_top8) improves over the base model (27.6% vs. 26.2%), and LASP closely tracks full PR (28.9% vs. 29.3%). The absolute gains are smaller on FLUX (+1.4% to +3.1%) compared to SD3 (+3.5% to +4.7%), consistent with the T5 encoder's greater robustness to text feature degradation Yao et al. (2026). The qualitative pattern transfers across architectures: sparse repair approaches full PR, and depth-adaptive deployment narrows the gap further, supporting that the benefits of text-feature repair transfer across MMDiT architectures, while the safe operating range of the injection weight is architecture-dependent.

LASP Weight Sweep on FLUX

To test the bounded amplification regime on FLUX, we sweep $w \in \{0.025, 0.05, 0.10, 0.20, 0.30, 0.40\}$ for LASP (split L17). The split layer is placed at a comparable relative depth: SD3's split at L9/24 (≈38% early) vs. L17/57 (≈30% early) in FLUX's absolute block indexing, where the 19 dual blocks precede the 38 single blocks sequentially (not interleaved). Table 20 reports the results.

In contrast to SD3's inverted-U crash at w=0.20, FLUX.1-Schnell improves monotonically with increasing w, saturating at w=0.30 (32.4%) and plateauing thereafter (32.2% at w=0.40). We additionally swept w for *full PR* on FLUX to test whether this tolerance depends on token count (Table 21): full PR also rises to a plateau (30.5% at w=0.05, peaking at 33.5% at w=0.10, then 32.2–32.4% up to w=0.30) with no collapse, so FLUX's immunity to over-injection is independent of the number of repaired tokens. This tolerance is a property of Schnell: on FLUX.1-dev (full inference steps), Yao et al. Yao et al. (2026) report that overall GenEval declines from 0.7002 at w=0.025 (origin layer 2) to 0.6010 at w=0.10. Schnell therefore brackets the FLUX weight response from above, with FLUX.1-dev lying between SD3's inverted-U and Schnell's monotonic plateau; the

Table 20: LASP weight sweep on FLUX.1-Schnell (10-seed GenEval Position, ± 1 SD where available).

Injection weight w	**GenEval Position (%)**
0.025 (PR default)	28.9 ± 3.96
0.05	30.0 ± 3.85
0.10	31.5 ± 3.11
0.20	32.2 ± 2.86
0.30	**32.4**
0.40	32.2

safe operating window is modulated by the text encoder's robustness and the total injected signal. Inference steps are ruled out as the cause of the difference: with a 12-step SD3 schedule the collapse persists at the same weight (base 23.5%; full PR 19.0% at w=0.15 and 6.1% at w=0.20); the remaining confounds (classifier-free guidance and the T5-only encoder) are noted in Section 5.4.

Table 21: Full PR weight sweep on FLUX.1-Schnell (10-seed GenEval Position, 1000 imgs/condition). Like LASP, full PR shows no collapse.

Injection weight w	**GenEval Position (%)**
0.025 (PR default)	29.3 ± 3.85
0.05	30.5 ± 4.08
0.10	**33.5** ± 2.73
0.20	32.2 ± 2.18
0.30	32.4 ± 3.44

The absolute improvement from base to w=0.20 on FLUX (+6.0%, from 26.2% to 32.2%) is comparable to SD3's peak gain (+5.9%, from 29.9% to 35.8% at w=0.10). The two architectures thus achieve similar total benefit from text feature repair, but distribute it across the weight range differently: SD3 concentrates the gain within a narrow safe band, while FLUX accumulates it gradually across a wider range.

Per-relation analysis

Decomposing GenEval Position by spatial relation type reveals that PR/LASP improvements on SD3 concentrate in vertical relations (above/below) with minimal effect on horizontal (left/right) (Tables 22 and 23). On SD3, LASP w=0.10 improves "above" by $+20.8$pp (44.2%→65.0%) and "below" by $+13.6$pp, while "left of" and "right of" change by only $+0.6$pp and $+3.3$pp. The gain is not driven by a single relation type: excluding any one relation (leave-one-out over the 100 prompts), LASP w=0.10 still exceeds base by at least $+6.5$pp. On FLUX.1-Schnell, the "above" improvement replicates (52.3%→60.8%), but "below" does not rise (24.6%→23.9% for LASP, 24.3% for full PR); the vertical-dominant pattern is therefore driven by "above" in FLUX. This differential is consistent with the prompt forgetting mechanism, and it motivates a direct measurement of per-relation feature drift.

Table 22: Per-relation-type GenEval Position accuracy on SD3-medium (10-seed). PR/LASP improvements concentrate in vertical relations.

Condition	**above**	**below**	**left of**	**right of**
Base (no repair)	44.2%	36.8%	22.6%	14.1%
Full PR	53.5%	41.8%	21.1%	18.5%
LASP (w=0.025)	50.8%	42.9%	20.5%	19.6%
LASP (w=0.10)	**65.0%**	**50.4%**	23.2%	17.4%

Table 23: Per-relation-type GenEval Position accuracy on FLUX.1-Schnell (10-seed). The "above" improvement replicates; "below" is flat.

Condition	above	below	left of	right of
Base (no repair)	52.3%	24.6%	15.8%	10.0%
LASP	60.8%	23.9%	18.9%	10.4%
Full PR	61.5%	24.3%	17.9%	11.5%

F Excluding Attention Entropy as a Confound

We show here that H(L3) does not provide a usable mechanism signal: at fixed token count its variation is negligible and it cannot rank strategies.

Table 24: H(L3) entropy for seven PR variants with varying token counts (5-seed).

Condition	H(L3) mean	Official	Token count
Base	4.7327	29.9%	0
pr_prep	4.7323	29.9%	$\sim$3
pr_noun	4.7314	30.5%	$\sim$10
CG $\gamma = 0.20$	4.7321	30.8%	$\sim$5
CG $\gamma = 0.50$	4.7318	30.2%	$\sim$5
pr_shape2	4.7325	33.4%	$1^{\dagger}$
Full PR	4.6772	34.6%	333

†Gated: one position when a shape noun is present, otherwise unrestricted repair (Appendix A).

Across these seven mixed-token conditions, H(L3) correlates only moderately with performance ($r = -0.76$, p=0.047; rank correlation $\rho = -0.50$, n.s.), and the correlation is not stable under token-count control: the partial correlation changes sign and magnitude with plausible codings of the gated shape-noun condition (from $+0.77$ to -0.20). Mixed-token comparisons can therefore suggest an entropy–performance link without establishing one, so we rely on the equi-token control below.

Table 25: H(L3) entropy under equi-token control (all repairing exactly 8 tokens). H(L3) measured on 5-seed; the Official column reports the 10-seed GenEval Position score. Friedman test (non-parametric repeated-measures): $\chi^2 = 266.06$, $p < 0.001$; between-strategy spread $\Delta = 0.005$. Spearman $\rho = -0.87$, $p = 0.054$ ($n = 5$).

Strategy (8 tokens)	H(L3) mean	Official
random_8	4.7256	31.8%
noun_top8	4.7210	33.0%
entropy_k8	4.7204	33.4%
freq_top8	4.7203	33.4%
prep_top8	4.7207	32.7%

When token count is held constant, H(L3) varies by only $\Delta = 0.005$ across strategies—negligible against the 1.6pp performance spread—and it does not rank them (Spearman $\rho = -0.87$, $p = 0.054$, $n = 5$; 80% power at this n would require $|\rho| > 0.92$). A Friedman test detects systematic rank differences ($\chi^2 = 266$, $p < 0.001$), but the effect size is far too small to explain the spread. H(L3) is therefore not usable as a mechanism: at fixed token count it carries no discriminating signal, and its mixed-token correlation is weaker than previously reported and unstable under token-count control.

G QUALITATIVE EXAMPLES

This appendix provides qualitative comparisons on fixed-seed generations. All images use identical random seeds across conditions.

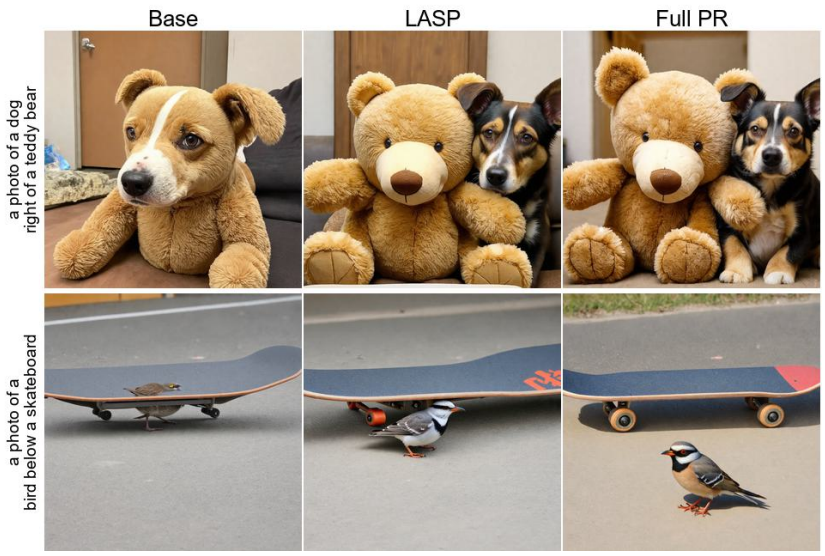

Figure 4: Qualitative comparison on SD3. Base (left), LASP (middle), Full PR (right), w=0.025. Top: base misses the teddy bear; LASP and full PR restore "dog right of teddy bear". Bottom: base occludes the bird under the skateboard; LASP and full PR place it below.

SPATIAL RELATION REPAIR ACROSS ARCHITECTURES

Figures 4, 5, and 6 show qualitative comparisons of base model outputs vs. repaired outputs on prompts where the base model misplaces a critical spatial relation, on SD3 and FLUX.1-Schnell respectively. On SD3, the base model typically misplaces or omits the target spatial relation (e.g., "left of" rendered as "right of"); LASP and Full PR both achieve faithful spatial layouts. This visual progression mirrors the quantitative ordering in Table 1: LASP approaches Full PR at reduced token cost. On FLUX.1-Schnell, the base model exhibits similar spatial failures, confirming that the mechanism transfers across MM-DiT architectures despite the architectural and encoder differences. In both settings, the 8-token variants (SD3 panels in Figure 5) demonstrate that sparse repair is visually indistinguishable from the substantially more expensive full PR.

OVER-INJECTION FAILURE

Figure 7 provides visual evidence for the bounded amplification regime by sweeping LASP across four injection weights (w=0.05, 0.10, 0.15, 0.20) on SD3. At w=0.05 and w=0.10, LASP produces clean spatial layouts with recognizable objects, consistent with its GenEval scores (38.1% and 40.1%, Table 3). At w=0.15, the image begins to degrade: fine-grained textures are replaced by artifacts and object identities become ambiguous (36.3%, still above baseline). At w=0.20, the degradation is pronounced: objects are heavily distorted or become unrecognizable, and the spatial layout breaks down. This visual collapse mirrors the quantitative drop below the unrepaired baseline (28.2% vs. 29.9%, Table 3) and illustrates why the bounded regime imposes a hard practical ceiling on injection weight for SD3.

Over-injection is a continuous phenomenon rather than a sharp threshold: sporadic object-identity or texture failures occur even at the default w=0.025 (a small minority of the 1000 images per condition). Figure 7 shows the most extreme w=0.20 cases, while the aggregate numbers in Table 3 average over all 1000 images per condition.

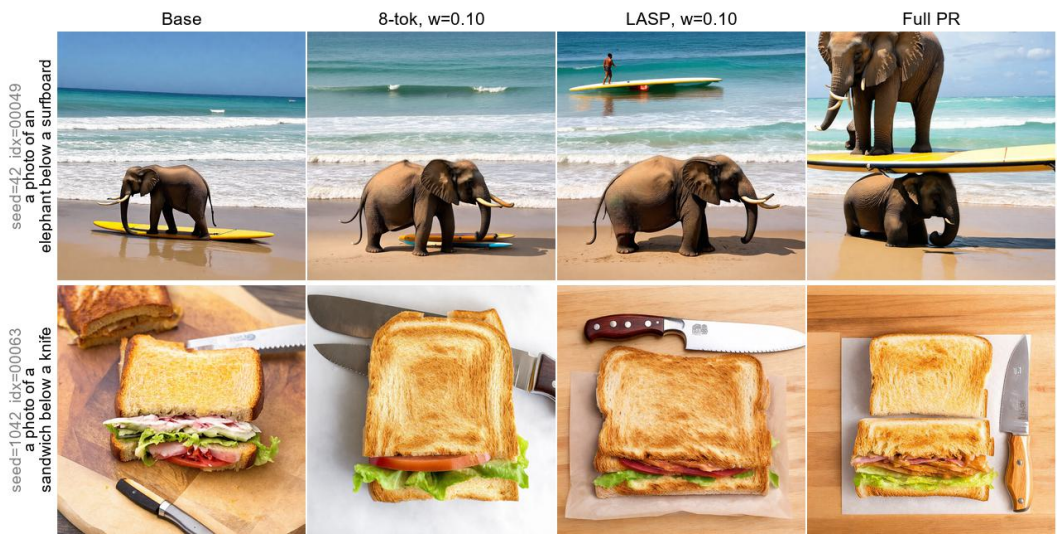

Figure 5: Spatial relation repair on SD3. Two cases where the base model misplaces the target relation (elephant below surfboard; sandwich below knife); LASP places the knife above the sandwich, while full PR leaves it to the right.



Figure 6: Spatial relation repair on FLUX.1-Schnell. Two cases where the base model inverts the target relation (frisbee right of motorcycle; tennis racket right of spoon); LASP and full PR restore the correct layout.

Object Detection Failures

Figures 8 and 9 contrast base-model and repaired outputs on prompts where the base model fails to detect or correctly render an object, complementing the heuristic observations in Appendix C; the weight sweep is shown separately in Figure 7. On SD3, Figure 8 illustrates three cases where the base model fails to detect or correctly render objects (distortion, category confusion, and omission). On FLUX.1-Schnell, Figure 9 illustrates that base-model failures take a different form than SD3's: objects may be missed entirely, rather than distorted.

In both architectures, object fidelity is visibly more fragile under weight variation than spatial layout, consistent with the detection-mediated picture of the collapse (Appendix C).



Figure 7: Over-injection failure on SD3. LASP (split L9) across four injection weights (w=0.05, 0.10, 0.15, 0.20, left to right): clean layouts at moderate weights give way to marked object and spatial breakdown at w=0.20. Per-weight scores in Table 3.

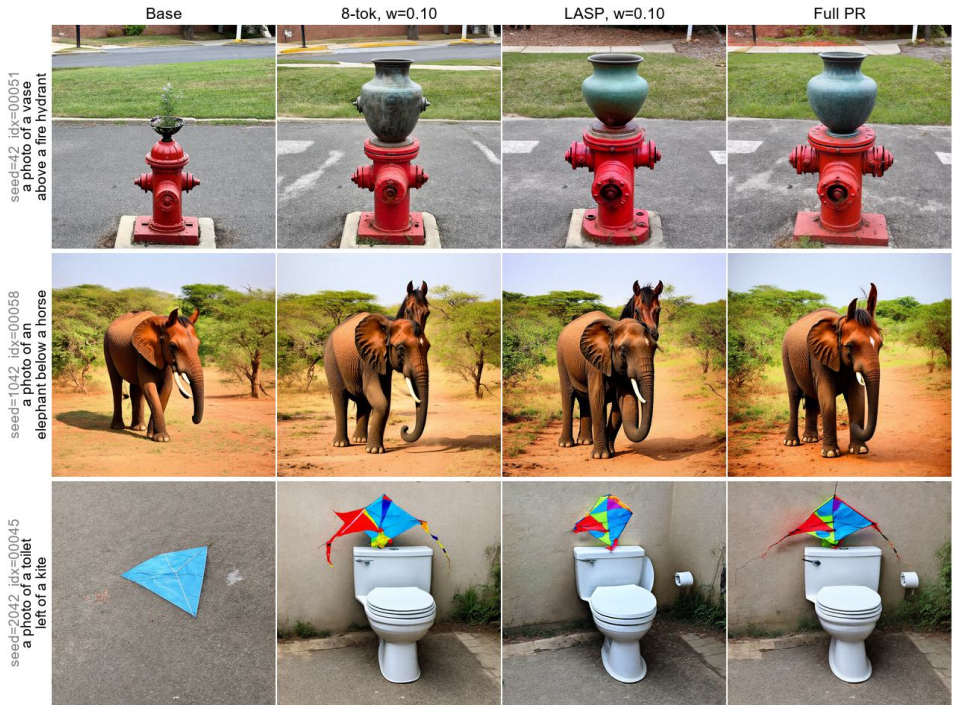

Figure 8: Object detection failures on SD3. Three representative base-model failure cases: distortion (vase above fire hydrant), category fusion (elephant below horse), and omission (toilet left of kite).

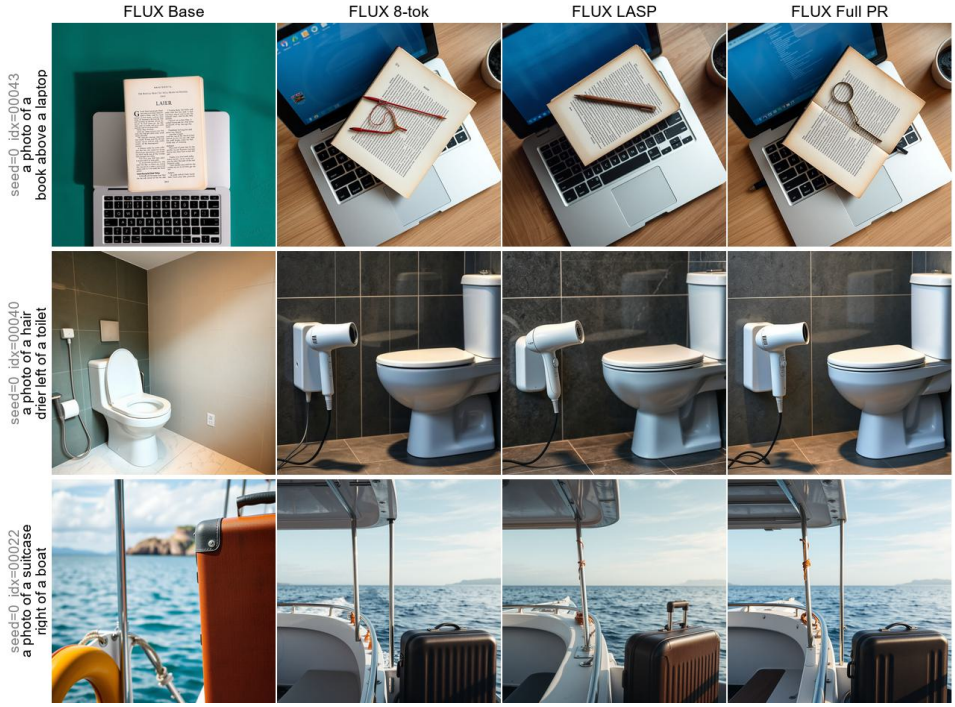

Figure 9: Object detection failures on FLUX.1-Schnell. Three representative base-model failure cases: object merging (book above laptop), omission (hair drier left of toilet), and weak visibility (suitcase right of boat); FLUX exhibits different failure modes than SD3's distortion.