Where to Look Is Not How to Fix: Pre-Denoising Diagnostics and Modality-Dependent Control in Diffusion Composition

Fangzheng Wu
Department of Computer Science
Tulane University
New Orleans, LA 70118
fwu6@tulane.edu

Brian Summa
Department of Computer Science
Tulane University
New Orleans, LA 70118
bsumma@tulane.edu

Abstract

Understanding compositional failures in text-to-image diffusion requires identifying both where stress is detectable and how intervention changes the output. We study these questions through a controlled anchor–stress protocol that jointly evaluates text-encoder diagnostics and denoiser interventions. We introduce a text-only Compositional Stress Index (CSI), which separates common from rare compositions across SD1.5, SDXL, and the SD3 text path and provides an upstream diagnostic coordinate. A matched six-prompt localization study links intervention location to distinct outcomes: residual-minimizing embedding adapters improve representation fit, while downstream cross-attention intervention increases color hit rate (CHR) by 0.0272. Across SD1.5 and SDXL denoiser blocks, the largest positive signed diagnostic-accessibility mean occurs at the deep encoder, whereas selective boost has its largest positive mean CHR response at decoder blocks. Selective subtraction and broad ablation reveal further modality- and architecture-dependent responses, including a substantial CHR decrease when SDXL decoder cross-attention is broadly ablated. We find a diagnosis-control dissociation under our controlled attribute-object composition setting: compositional defects are diagnosable before denoising, but the representation coordinate that exposes risk is not necessarily the coordinate or modality that improves generation.

1 Introduction

Text-to-image diffusion models [12, 26, 29] have achieved remarkable generation quality but remain brittle on semantically rare attribute–object compositions [14, 36]: a model that reliably generates "a red car" may fail on "a blue pizza" or "a glass cat," producing the wrong color, material, or texture. Two threads of work address this. Inference-time methods—Attend-and-Excite [4], SynGen [28], StructureDiffusion [8], Divide-and-Bind [19], and attention-map-controlled synthesis [35]—modify denoiser cross-attention so that attribute tokens better bind to the correct object regions, typically aggregating corrections across all layers. A second line of work studies the upstream text-conditioning pathway, asking whether compositional failures originate in representation geometry [3, 17] or in misalignment between the embedding objective and the denoiser's binding readout [22, 37]. Implicit in both threads is an assumption: the pipeline location where a compositional defect is most *detectable* is also the location where it is most effectively *correctable*. We test this relationship by comparing diagnostic readouts and intervention responses at distinct pipeline sites.

Our central finding is a site–modality dissociation: compositional stress has an upstream diagnostic signature, while intervention responses depend jointly on pipeline location and the type of edit. In the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

tested UNets, the deep encoder has the largest positive signed diagnostic-accessibility mean, whereas decoder blocks have the largest positive mean response to selective boost. This motivates treating diagnosis and intervention as distinct empirical questions: *where to look is not how to fix*.

To make this concrete, consider modern text-to-image diffusion pipelines [7, 26, 27, 29, 30]. A text encoder first produces conditioning embeddings, which are then consumed by a denoising backbone to guide generation. In the UNet-based Stable Diffusion family, including SD1.5 and SDXL [26, 29], these conditioning embeddings are injected through cross-attention layers distributed across encoder, middle, and decoder blocks [11, 29, 33]. This architecture provides multiple internal sites at which a compositional failure may be readable, and multiple sites at which an intervention may affect the output. We use this UNet hierarchy for our block-level localization experiments, while also testing the representation-side diagnostic on an additional non-UNet text-to-image pipeline, namely Stable Diffusion 3 [7]. Our question is therefore: at which sites can compositional failures be *detected*, and at which can they be *fixed*—and does detection at a site imply fixability there?

Table 1 maps this dissociation across three pipeline sites; rows correspond to sites, columns distinguish diagnostic readout from intervention outcomes. We present the evidence in two parts, corresponding to two regions of the pipeline.

Upstream: diagnosis and repair at the text encoder (Table 1, row 1). If a text encoder processes "a blue pizza" compositionally, the full-prompt representation should be roughly recoverable from the separate representations of "blue," "pizza," and the surrounding context. A large deviation from this component-based reconstruction indicates that the full prompt is not factorizing into its parts. We call this deviation a *non-factorization residual* and aggregate it into a Compositional Stress Index (CSI), where higher CSI indicates stronger departure from compositional factorization. Within a deliberately controlled anchor–stress setting that pairs common and rare attribute–object compositions while holding the object and context fixed, CSI separates the two groups perfectly (area under the receiver operating characteristic curve, ROC-AUC$=1.0$) across SD1.5, SDXL, and the SD3 text path, and provides a diagnostic coordinate for the subsequent intervention analyses. We use the text-only CSI component for the pre-denoising interpretation (Appendix B).

This diagnostic coordinate lets us examine how representation changes relate to output changes. We train low-rank adapters to move stress embeddings toward anchor-like structure and compare their responses with denoiser-side cross-attention intervention. The adapters improve representation fit. On a matched six-prompt localization subset, downstream intervention increases CHR by 0.0272, while embedding intervention decreases it. The comparison connects intervention location to distinct representation and image-level outcomes under the tested objectives (§5).

Downstream: per-block modality decomposition inside the denoiser (Table 1, rows 2–3). This representation–output comparison motivates the downstream question: once text conditioning is read out by the denoiser, which internal sites are diagnostic, and which are causally controllable? For UNet-based SD1.5 and SDXL, we compare block groups spanning the deep encoder, middle block, and decoder, using three intervention modalities: selective boosting, selective zeroing, and broad cross-attention ablation. We quantify denoiser-side readout by diagnostic accessibility (DA), a block-level measure of how well cross-attention patterns separate anchor from stress prompts during baseline generation. The results reveal a second dissociation, now *within* the denoiser: D2 has the largest positive signed DA mean, while decoder blocks have the largest positive mean CHR response to selective boosting. Broad ablation produces a different, architecture-dependent profile, with a substantial CHR decrease at SDXL U0 and at early SD1.5 encoder blocks. These profiles characterize how intervention responses vary with both *where* an edit is applied and *how* it modifies cross-attention (§6).

Contributions. (1) Under the tested intervention families, we uncover a diagnosis–control dissociation in text-to-image diffusion: the site where compositional stress is most readable is not necessarily the site where intervention is most effective. (2) We further decompose this dissociation under the tested intervention families, showing that selective token edits and broad cross-attention ablations have different site profiles across deep encoder and decoder blocks. (3) We introduce text-only CSI as a pre-denoising stress coordinate and evaluate it across SD1.5, SDXL, and the SD3 text path under the controlled anchor–stress protocol. (4) Under the tested intervention families, we connect representation and output responses through a matched localization study: residual-minimizing adapters

Table 1: Diagnostic readouts and observed intervention responses across pipeline sites. Entries summarize measured profiles rather than significance decisions; effect sizes, uncertainty, and statistical tests are reported in Sections 5–6. DA peaks refer to signed means.

Pipeline site	Diagnostic readout	Targeted edits	Broad ablation
Text encoder	Residual signal	Residual reduction	—
Deep encoder	Positive DA peak	Small CHR changes	Architecture-dependent
Decoder	Smaller signed DA	Peak mean boost CE	SDXL U0: decrease

improve embedding fit, while downstream intervention increases CHR on the selected six-prompt subset.

2 Related Work

The cross-attention-steering methods introduced in §1 [4, 8, 19, 28, 35] share a design choice: they aggregate attention signal across all layers and resolutions, treating cross-attention as a homogeneous control surface; logit-level masking and entity-region constraints (e.g., MaskDiffusion-style methods, box-guided attention control) target specific resolutions or layers but similarly do not decompose across stages. WiCLP [37] pivots upstream, showing embedding-level repair *can* work with supervision aligned to the denoiser's readout and implicating objective misalignment rather than the embedding space itself. Our findings should not be interpreted as evidence that all upstream interventions fail. WiCLP, for instance, demonstrates that embedding-space interventions can improve compositional generation when optimized with denoising-aligned objectives. Our result is narrower: reducing the CSI-diagnosed residual via residual minimization is not sufficient to guarantee semantic improvement. Marioriyad et al. [22] link entity-missing to attention overlap. A complementary line of work intervenes at training or sampling time rather than at inference attention, including contrastive curriculum fine-tuning [10], sampling-time corrective steering [6], and PAC-Bayesian priors over attention distributions [15]. Our contribution is orthogonal to all of these: we diagnose *where* in the pipeline compositional risk appears and characterize *which modality* works at each site, producing a layer-selection principle that could inform any of these approaches—training-time methods might focus supervision on the deep-encoder representation we identify, while sampling-time correctors gain most by operating on decoder-side attention.

On the representation and evaluation side, CLIP embeddings admit sparse linear decompositions [3] and compose single-object attributes but fail under multi-object binding [17]. Prompt-to-Prompt [11] showed cross-attention encodes spatial and semantic structure; FreeU [31] demonstrated that UNet encoder–decoder information flow is a control axis. More broadly, probing and causal intervention research [1, 2, 9, 24, 34] has shown that probe-friendly layers need not be intervention-friendly. We sharpen this gap for diffusion cross-attention by identifying both *which layer* and *which modality* works at each layer, and we quantify a pre-denoising non-factorization signal not captured by these prior analyses. Output-side benchmarks [14, 18, 32] measure *whether* composition fails; our work diagnoses *where* the failure originates and *how* to intervene at each site. Extended related work covering compositional evaluation suites, object-centric representations, and energy-based compositional generation appears in Appendix A.

3 Setup: Sites, Diagnostics, and Interventions

We study controlled compositional stress in text-to-image diffusion. Each example consists of an *anchor* prompt paired with a single *stress* prompt that shares the same object and context but differs in the target attribute. The anchor uses a common attribute–object composition, whereas the stress prompt uses a rarer or less typical composition. Given such pairs, we ask two questions: where in the model is the stress signal diagnostically readable, and where can an intervention causally improve the generated output?

Our experiments instantiate this setup with 432 prompts organized into 216 paired groups, each consisting of one anchor prompt and one stress prompt. Attribute–object combinations are selected from T2I-CompBench [14], CLEVR [16], and ARO [36] lexicons to span common and rare compositional regimes (examples in Appendix N). For example, within one group the anchor might be "a studio photo of a red pizza" and the stress prompt "a studio photo of a blue pizza"—same object, same

context, different attribute. Examples such as "a red car" and "a blue pizza" that appear elsewhere in the paper illustrate the rarity gradient across the dataset, not within-pair variation.

Primary models for generation and denoiser-side intervention are SD1.5 and SDXL, while SD3 is used only for text-path diagnostic extension. For UNet block-group analysis, we partition cross-attention modules by their position in the UNet architecture. Following standard notation, D_k denotes the k-th down-sampling (encoder) stage, M the middle bottleneck, and U_k the k-th up-sampling (decoder) stage. SDXL has five active cross-attention groups: $\{D_1, D_2, M, U_0, U_1\}$; SD1.5 has seven: $\{D_0, D_1, D_2, M, U_1, U_2, U_3\}$ (module counts per group in Appendix K). Throughout, we refer to D_2 as the *deep encoder* and U_0 (SDXL) / U_1 (SD1.5) as the *decoder*. With these sites defined, we next introduce two diagnostic readouts—one at the text encoder, one inside the denoiser—and two families of intervention: embedding-level repair and denoiser-side cross-attention edits.

Text-encoder diagnostic: CSI. We operationalize text-side compositionality as a component-reconstruction test. If the text encoder preserves a factorized representation of attribute, object, and context, then a simple composition rule should explain the full-prompt embedding; large residuals indicate prompt-specific interactions beyond this factorized account. For each prompt p_{full} (e.g., "a studio photo of a blue pizza"), we construct component inputs: p_{attr} (the target attribute, e.g., "blue"), p_{obj} (the object, e.g., "pizza"), and p_{ctx} (the surrounding context template). Let $h(\cdot)$ denote the representation extracted from the text encoder, and let $\hat{h}(\cdot)$ be a deliberately low-capacity composition model fit to predict $h(p_{\text{full}})$ from the component representations. We consider additive, low-rank linear, and low-rank bilinear variants, fit on anchor prompts and evaluated on both anchor and stress prompts. We define the non-factorization residual as

$$R(p_{\text{full}}) = \frac{\left\| h(p_{\text{full}}) - \hat{h}(p_{\text{attr}}, p_{\text{obj}}, p_{\text{ctx}}) \right\|_2}{\|h(p_{\text{full}})\|_2 + \varepsilon}. \tag{1}$$

where ε prevents division by zero. In plain terms: If the encoder preserves component-wise structure for "blue," "pizza," and the context template, then a simple composition model should be able to reconstruct the full-prompt embedding, yielding a small R. If the full prompt instead induces prompt-specific interactions that are not predictable from those components, the residual increases. CSI aggregates this normalized residual with span-geometry features via robust standardization on anchor statistics; we report CSI_{text}, CSI_{span}, and combined CSI. Only CSI_{text} is strictly pre-denoising; combined CSI includes internal attention-derived span features. Encoder extraction, calibration scope, and span-feature definitions are given in Appendix B.

Denoiser diagnostic: DA. The second diagnostic readout operates inside the denoiser. For each UNet block group b, diagnostic accessibility $\text{DA}(b)$ measures how well block-level cross-attention patterns distinguish anchor from stress prompts during baseline generation. We define an attribute-binding measure ABM_b as the fraction of attribute-token attention mass falling within top-quantile object-attention regions, then compute:

$$\text{DA}(b) = \mathbb{E}_{(p_a, p_s)}\big[\text{ABM}_b(p_a) - \text{ABM}_b(p_s)\big], \tag{2}$$

where (p_a, p_s) denotes a matched anchor–stress pair. Positive DA means anchor prompts show tighter attribute-to-object binding at block b than stress prompts. We rank sites by this signed mean difference; negative DA indicates the reverse contrast and is not evidence of absent diagnostic information. Significance against zero does not establish a significant difference between sites. Formal definitions and validation of ABM are given in Appendix H.

To separate *where* an intervention is applied from *what kind* of intervention it is, we organize denoiser-side edits along two axes: *selectivity*, whether the edit targets only the attribute token or the entire cross-attention map, and *sign*, whether the edit strengthens or weakens the conditioning channel. This gives a 2×2 design; the broad-additive cell is omitted because uniformly scaling all attention weights is canceled by post-softmax normalization, leaving three testable denoiser-side families. *Selective additive* (BOOST) amplifies attention to the target attribute token at a chosen block, testing whether stronger local attribute readout improves binding. *Selective subtractive* (ATTR_ZERO) removes the same attribute-token attention, testing the opposite direction as a matched-selectivity control. *Broad subtractive* (ZERO) uniformly flattens all cross-attention at a chosen block, testing whether the block's entire text-conditioning contribution is necessary or harmful. Together with two embedding-level repair families—pooled and token low-rank adapters trained to move stress representations toward

anchor-like structure—this yields five intervention families. Implementation details and statistical testing protocols are given in Appendix D.

These three denoiser modalities span a natural hypothesis space about the defect's structure. If the failure reflects weak attribute readout, BOOST should help by amplifying the target-attribute channel. If the same channel is causally involved but harmful or mislocalized, ATTR_ZERO tests the opposite direction by removing it. If the failure is distributed across the block's entire text-conditioning pathway, ZERO tests whether removing that contribution improves generation. Testing all three at each block allows us to diagnose not just *where* the defect is, but *what kind* of defect it is. All intervention hyperparameters are fixed across block comparisons (strength, mask quantile, timestep schedule, and CFG settings in Appendix D).

We evaluate control with object-local color hit rate (CHR): the fraction of pixels inside a fixed target-object mask whose color matches the desired attribute. The object mask is derived from baseline final-timestep attention and held fixed across conditions, which fixes the scoring region but does not independently verify object identity or localization; mask stability across interventions is verified in Appendix I. In the full per-block runs, CHR is finite for 52 stress prompts per seed; the other 164 stress prompts have no target color recorded by the scorer. This is a scoring-eligibility criterion, not a verified taxonomy of prompt semantics. We use chromatic binding because it changes attribute semantics while largely preserving object shape and localization, providing the cleanest endpoint for site-localization comparisons. Color prototypes and thresholds are reported in Appendix I. For the audited per-block results, we first average paired differences across seeds, then test across prompts (CE) or matched `group_id` units (DA). We use two-sided sign-flip tests (5,000 Monte Carlo draws, plus-one correction; exact enumeration for at most 16 units) and 2,000 percentile-bootstrap draws, with analysis seed 0. BH correction is applied separately to DA and CE across all active blocks within each model/run. Full CE analyses use 52 prompts with five seeds for boost/attr-zero and three for broad zeroing; DA uses 216 groups. Group IDs do not by themselves establish independence of shared anchor texts.

4 Result 1: Text Encoders Provide a Pre-Denoising Stress Axis

We first ask whether compositional stress is already visible before denoising, using only text-encoder representations. The goal is not to establish CSI as a general-purpose detector of arbitrary compositional failure, but to obtain a calibrated axis that orders prompts by stress severity before testing where intervention can change the generated output. On the controlled anchor–stress protocol, CSI cleanly separates common-composition anchors from rare-composition stress prompts across SD1.5, SDXL, and the SD3 text path (ROC-AUC$=1.0$, pair-ranking$=1.0$; Figure 2). This ceiling-level separation should be interpreted with the protocol in mind: stress prompts are deliberately constructed to be semantically rare, and a frequency-only baseline already reaches ROC-AUC$=0.79$. Thus, the important claim is not that CSI is a universal binary detector, but that text-side non-factorization provides a stable pre-denoising stress coordinate.

While ROC-AUC summarizes ranking performance, Figure 1 visualizes the underlying prompt-level distributions. Across all three models, stress prompts exhibit substantially larger CSI values than anchors, consistent with the near-perfect separation reported above.

The stronger evidence is that CSI remains informative *within* the stress set and aligns with output-side failures. Across stress prompts, CSI correlates with graded stress severity (decile-stress Spearman $\rho=0.87$) and with BLIP-derived attribute failure rate (Spearman $\rho=0.68$); output-side BLIP-VQA labels ($n=4320$) show that CSI predicts attribute failure with ROC-AUC$=0.89$ and binding failure with ROC-AUC$=0.85$. Controls support this interpretation: group-preserving permutation gives null ROC-AUC $= 0.49 \pm 0.01$ (all empirical $p=0.001$), text-side non-factorization is the dominant contributor, span-only geometry is substantially weaker, and robustness to lexical confounds and family ablations is reported in Appendix B. The pattern is also consistent across encoder families: SD1.5 uses a single CLIP text encoder, SDXL uses dual encoders (CLIP-L + OpenCLIP-G), and the SD3 text-path extension uses a T5-XXL sequence mean. Despite these architectural differences, the non-factorization gap appears in all three settings, so we use CSI as a model-agnostic pre-denoising stratification axis for the intervention analyses in §5–§6.

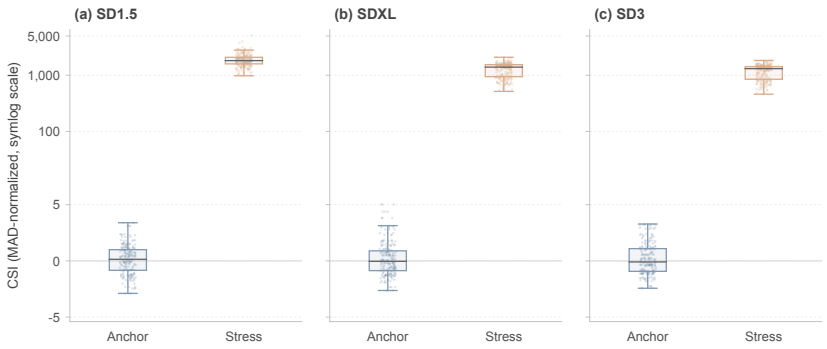

Figure 1: Prompt-level CSI distributions for anchor and stress prompts across SD1.5, SDXL, and SD3. Values are displayed on a MAD-normalized symlog scale. Although absolute magnitudes are not directly comparable across models, stress prompts consistently occupy a much higher CSI regime than anchor prompts. This visualization complements ROC-AUC by showing the magnitude of the anchor–stress separation.

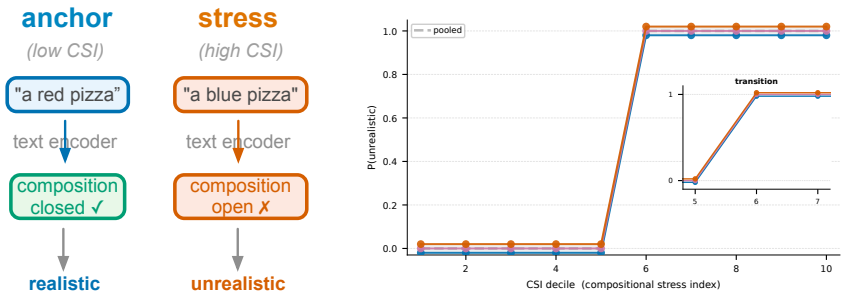

Figure 2: Cross-model CSI decile curves. Probability of stress membership increases monotonically with CSI decile; SD1.5, SDXL, and SD3 collapse to the same transition pattern. ROC-AUC=1.0 in all three models.

5 Result 2: Upstream Representation Repair Does Not Imply Semantic Repair

The upstream analysis established that compositional stress is detectable at the text encoder. We next ask whether repairing that diagnostic coordinate is sufficient to improve generated output. A pooled adapter reduces the representation residual ($\Delta_{\text{repr}} = -0.100$, q=0.0024, 20 prompts), and a token adapter gives $\Delta_{\text{repr}} = -0.184$ (q=0.0008, 19 prompts). Each has seven CHR-evaluable prompts, with observed Δ_{CHR}=0 (q=1). On a separate CSI-stratified localization subset with six CHR-evaluable prompts and five seeds each, downstream intervention increases CHR by +0.0272 (95% CI $[0.0092, 0.0469]$; exact prompt-level p=0.03125, q=0.046875, d_z=1.04), whereas token-level embedding repair decreases CHR by −0.0680 (95% CI $[-0.1071, -0.0327]$; p=0.03125, q=0.03125). The two arms use the same six prompts and seeds and have identical baseline CHR values. These results concern the measured CHR endpoint and do not establish general semantic repair.

A separate rank sweep (4–128) shows no significant positive CHR effect for the pooled adapters and negative CHR effects for the token adapters under within-run correction (Appendix C). Residual improvement is not monotonic in rank, and these data do not rule out capacity or objective choice as explanations. In the localization downstream arm, early ABM increases (Δ=0.000450, q=0.0012), whereas the mid-ABM change does not reach significance (Δ=0.000257, q=0.0708). This pattern motivates, but does not identify, distinct readout mechanisms.

We interpret this as a representation–generation boundary: the upstream coordinate that diagnoses compositional stress is not automatically a control coordinate for the denoiser's binding behavior. This conclusion is specific to residual-minimizing objectives; it does not rule out embedding-level repair with binding-aligned supervision such as WiCLP [37]. The next question is whether the

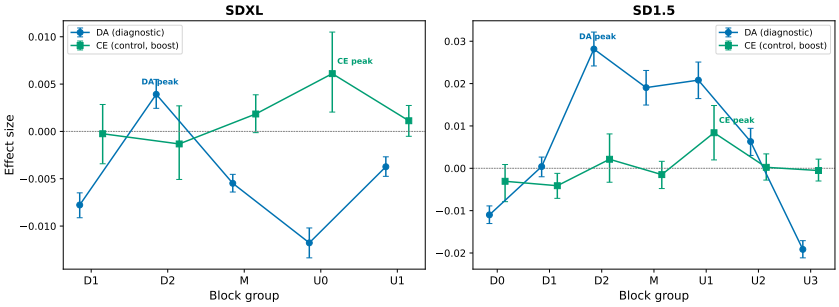

Figure 3: DA (blue) and CE under boost (green) across cross-attention block groups. The plotted DA and CE profiles are historical; updated phase1 means place the largest positive DA at D2 and boost CE at U0/U1. The largest positive signed DA mean occurs at D2. The retained plot uses historical runs and intervals; phase1 prompt-level CE results are reported in Table 6.

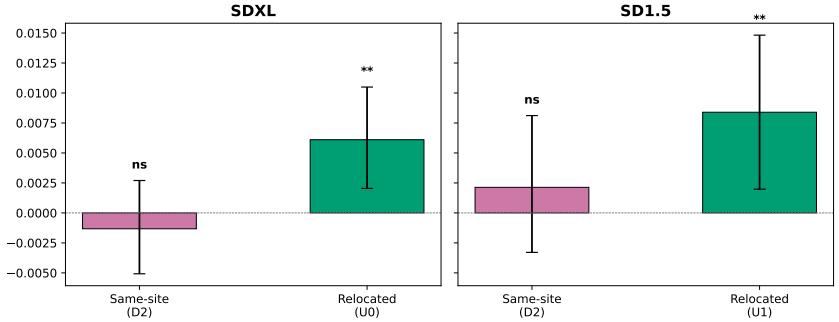

Figure 4: Same-site (D2, purple) vs. relocated (decoder, green) under boost. Updated prompt-level decoder-minus-D2 tests give p=0.0028 (SDXL) and p=0.1882 (SD1.5). Plotted intervals are historical; updated inference is reported in Appendix F.

binding channel can be reached more directly inside the denoiser, and if so, at which blocks and with which modality. §6 addresses this with a per-block sweep of cross-attention interventions.

6 Result 3: Diagnostic and Control Sites Dissociate Inside the Denoiser

Having found no positive CHR improvement from the tested residual-minimizing adapters, we next intervene inside the denoiser. D2 has the largest positive signed DA mean in both models across the audited baseline datasets (Appendix E). Under boost baselines, SDXL has $\mathrm{DA}(\mathrm{D2}) = +0.0039$ (q=0.0174), and SD1.5 has $+0.0282$ (q=0.00028). These tests establish a positive anchor–stress contrast at D2; they do not test whether D2 is significantly more diagnostic than every other block.

Selective boost at D2 has no detected CHR effect in either model (SDXL p=0.4093; SD1.5 p=0.4819). Decoder-minus-D2 is positive and nominally significant in SDXL ($+0.0074$, p=0.0028), but only directionally consistent in SD1.5 ($+0.0063$, p=0.1882). Decoder-minus-baseline is $+0.0061$ in SDXL (95% CI $[0.0016, 0.0111]$, p=0.0168, q=0.0840) and $+0.0084$ in SD1.5 ($[0.0008, 0.0156]$, p=0.0384, q=0.1344). Neither survives within-model correction across blocks. Selective ATTR-ZERO shows no significant positive CHR effect; SD1.5 D0 has a small negative effect (-0.0050, q=0.0014). Failure to detect an effect at D2 is not an equivalence result. The separately configured timestep and fractional-attenuation sweeps are descriptive controls (Appendices L and O).

Broad ablation gives a different, architecture-dependent site profile. SDXL D2 has no detected change ($+0.0054$, q=0.8273), and SD1.5 D2 has a positive mean ($+0.0577$, p=0.0356) that does not survive FDR (q=0.0623). SDXL U0 shows a substantial decrease (-0.1338, q=0.0010), while SDXL D1 improves ($+0.0200$, q=0.0330). SD1.5 D0/D1 show substantial decreases (q=0.0007). Thus, the data support modality- and architecture-dependent responses, but do not establish that D2 ablation is beneficial or harmless in both models.

Together, the modality decomposition separates the largest positive signed DA mean at D2 from the largest positive boost CE mean at the decoder (Table 2). The decoder-minus-D2 contrast is supported in SDXL but not significant in SD1.5; decoder-minus-baseline does not survive per-block



Figure 5: Same-site (D2) vs. relocated (decoder) boost. D2 boost produces no visible change or degradation; decoder boost increases CHR in these examples. "Gold guitar": CHR 0.125→0.114→**0.244**. "Blue apple": 0.239→0.224→**0.372**. "Red horse": 0.148→0.160→**0.329**.

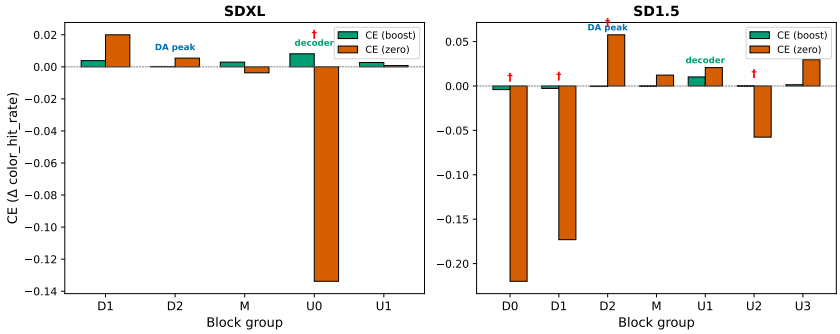

Figure 6: CE under boost (green) and zeroing (orange) at each block group. The retained plot shows historical estimates; updated prompt-level inference is in Appendix G. Broad zeroing substantially reduces CHR at SDXL U0; absence of a significant D2 effect does not establish tolerance.

FDR in either model. Broad ablation has an architecture-dependent profile, including a large negative response at SDXL U0. These findings support evaluating diagnostic readouts and intervention operators separately, with the strength of each comparison stated explicitly.

7 Discussion

Taken together, the upstream and downstream analyses reveal a specific property of diffusion pipelines: as the conditioning representation flows from text encoder to denoiser cross-attention, the features that are most informative about *which prompts will fail* and the features that are most responsive to *interventions that correct failure* live in different places. Text-encoder residuals provide a diagnostic coordinate; D2 has the largest positive signed DA mean; and decoder blocks have the largest positive boost CE means. Together these measurements locate distinct diagnostic and intervention-response profiles along the pipeline. The site-level mismatch extends to the prompt level as well: CSI-stratified analysis shows no consistent positive correlation between pre-denoising stress severity and decoder

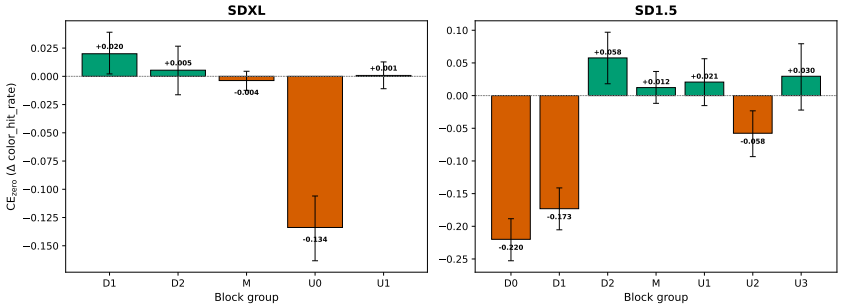

Figure 7: Zeroing fragility curve. Colors indicate the sign of the estimated CHR change, not corrected significance or functional necessity. SDXL has a large negative response at U0; SD1.5 at D0/D1. Historical plot annotations require the updated inference in Table 8.

Table 2: CE under three modalities at D2 and decoder (both models). Bold: FDR $q<0.05$. Boost p-values are uncorrected; FDR across block groups does not reach significance (see text).

Model	Block	Selective additive (boost)	Selective subtractive (`attr_zero`)	Broad subtractive (zero)
SDXL	D2	-0.0013 (ns)	-0.0011 (ns)	$+0.0054$ (ns)
	U0	$+0.0061$ (p=0.0168)	-0.0002 (ns)	$\mathbf{-0.1338}$ (q=0.0010)
SD1.5	D2	$+0.0021$ (ns)	-0.0029 (ns)	$+0.0577$ (q=0.0623)
	U1	$+0.0084$ (p=0.0384)	-0.0031 (ns)	$+0.0207$ (ns)

boost response (SDXL partial $\rho = -0.17$, p=0.007; SD1.5 $\rho = +0.11$, ns), so diagnostic severity does not predict which prompts respond to intervention. Diagnosis and intervention are not aligned by default along the pipeline; whether they can be realigned depends on choices about objective (embedding repair, §5) and modality (per-block decomposition, §6), not merely site.

Two mechanistic observations sharpen the dissociation. At D2, neither tested token-level intervention shows an FDR-significant CHR effect and per-timestep analysis shows DA–CE anti-correlation—D2 is most diagnostic in early denoising precisely when boost is most destructive. We interpret this carefully: our tested attribute-token operators (additive boost, matched-selectivity zeroing, fractional attenuation) do not reach the defect, but the underlying structure is not uniquely identified, since the pattern is equally consistent with a distributed defect across D2's cross-attention output and with a localized defect in a representation subspace our operators miss. At the decoder, U0's negative DA (-0.0118, FDR q=0.00033 in the SDXL boost baseline) indicates a reverse anchor–stress ABM contrast; the decoder does not lack signal but faithfully executes whatever binding it receives, including incorrect upstream binding. Overriding existing binding with a competing signal is one possible interpretation, not a mechanism established by these comparisons. These findings motivate evaluating diagnostic readouts and intervention operators jointly while selecting intervention sites using measured output responses.

Beyond our own protocol, the modality–site dissociation has implications for existing inference-time methods. Attend-and-Excite [4] and SynGen [28] apply gradient-based corrections across all cross-attention layers simultaneously; our per-block decomposition suggests their effectiveness may concentrate at decoder blocks, because decoder blocks have the largest positive boost CE means in our sweeps. If the encoder-side contribution is zero or negative, all-layer methods may succeed *despite* their encoder-side component rather than because of it—a testable prediction: restricting A&E/SynGen updates to decoder-only cross-attention should preserve or improve performance while reducing compute, and restricting to encoder-only should degrade performance to baseline or worse.

More broadly, the diagnosis–control mismatch is not specific to diffusion models. Any system with a separated encoder and decoder—including encoder-decoder language models, retrieval-augmented generation pipelines, and vision-language models—faces the same structural question: does diagnosing a defect in the encoder tell you how to fix it in the decoder? Our results suggest the answer is not automatic. The features that make a defect *visible* (high mutual information with the anchor/stress label) need not be the features that make it *fixable* (responsive to targeted intervention). Intervention design must be informed by the decoder's readout structure, not merely by the encoder's diagnostic signal.

7.1 Scope and Interpretation

The central conclusion of this work is not that a particular intervention site universally solves compositional failures. Rather, our evidence establishes that diagnostic accessibility, representation modification, and semantic improvement should not be assumed to coincide. Accordingly, we do not claim (i) that all upstream interventions fail, (ii) that decoder-side intervention provides a universal repair mechanism, or (iii) that the observed diagnosis-control dissociation has been established for relational or role-binding compositions. Our results are supported for the evaluated attribute-object setting and the tested intervention modalities.

8 Limitations

We distinguish established from suggestive findings. **Established**: (i) CSI separates anchor from stress cross-model (ROC-AUC$=1.0$, permutation null$=0.49$, within-stress $\rho=0.68$), though partly reflecting the controlled protocol; (ii) D2 has the largest positive signed DA mean in the six baseline datasets and differs from zero after FDR; these are not six independent replications; (iii) selective subtraction shows no significant positive CHR effect across the reported blocks; (iv) embedding repair reduces residuals without semantic improvement. **Suggestive**: (v) decoder boost as a preferred selective site (baseline contrasts nominally significant in both models but not after per-block FDR; decoder-minus-D2 significant only in SDXL); (vi) D2 defect structure (consistent with but not uniquely identified by our operators); (vii) generality beyond chromatic binding.

Key caveats: boost CE is small ($+0.006$–0.008) and does not survive within-model FDR; CE is color-only (52 valid prompts per seed); SD3 is text-path only; two UNet architectures, no DiT; fixed intervention hyperparameters; CSI-based routing and a universal decoder repair interface unvalidated; broad-additive cell untested; WiCLP-aligned embedding repair not tested; object mask from baseline attention, held fixed across conditions (empirically verified stable: median $|\Delta|=0.2\%$, $p95=1.5\%$, $r\geq 0.92$). We characterize a modality-dependent structure, not a universal law. External-mask and background-color sensitivity checks further limit the interpretation to a small chromatic response rather than established object-local semantic repair (Appendix J).

9 Conclusion

We characterize compositional diagnosis and intervention through complementary measurements of text representations, cross-attention, and generated color agreement. Text-only CSI provides an upstream stress coordinate across SD1.5, SDXL, and the SD3 text path under the controlled anchor–stress protocol. Within the two tested UNets, the largest positive signed DA mean occurs at D2, while the largest positive selective-boost CE mean occurs at decoder blocks. A matched six-prompt localization study further links intervention location to output response, with downstream intervention increasing CHR as residual-minimizing adapters improve representation fit. Broad ablation reveals distinct architecture-dependent sensitivity profiles. Together, these findings provide an empirical site–modality map and motivate choosing intervention sites through output-response evaluation alongside diagnostic analysis. Extending this framework to broader compositional benchmarks, binding-aligned objectives, and DiT backbones offers concrete directions for future work. Code, CSI computation scripts, prompts, and seeds will be released on GitHub upon publication.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[3] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). *NeurIPS*, 37:84298–84328, 2024.

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

[5] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, et al. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *ICML*, pages 8489–8510, 2023.

[6] Debottam Dutta, Jianchong Chen, Rajalaxmi Rajagopalan, Yu-Lin Wei, and Romit Roy Choudhury. Steer away from mode collisions: Improving composition in diffusion models, 2026.

[7] 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.

[8] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[9] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *NeurIPS*, 34:9574–9586, 2021.

[10] Evans Xu Han, Linghao Jin, Xiaofeng Liu, and Paul Pu Liang. Progressive compositionality in text-to-image generative models. *arXiv preprint arXiv:2410.16719*, 2024.

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

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[13] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

[14] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*, 36: 78723–78747, 2023.

[15] Eric Hanchen Jiang, Yasi Zhang, Zhi Zhang, Yixin Wan, Andrew Lizarraga, Shufan Li, and Ying Nian Wu. Unlocking the potential of text-to-image diffusion with pac-bayesian theory. *arXiv preprint arXiv:2411.17472*, 2024.

[16] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017.

[17] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? Probing compositionality in large image models. In *Findings of EACL*, pages 1487–1500, 2024.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[19] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864*, 2023.

[20] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

[21] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023.

[22] Arash Marioriyad, Mohammadali Banayeeanzade, Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Attention overlap is responsible for the entity missing problem in text-to-image diffusion models! *arXiv preprint arXiv:2410.20972*, 2024.

[23] Arash Marioriyad, Parham Rezaei, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Diffusion beats autoregressive: An evaluation of compositional generation in text-to-image models. *arXiv preprint arXiv:2410.22775*, 2024.

[24] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *NeurIPS*, 35:17359–17372, 2022.

[25] Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36:50173–50195, 2023.

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[28] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *NeurIPS*, 36:3536–3559, 2023.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[31] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free lunch in diffusion U-Net. In *CVPR*, pages 4733–4743, 2024.

[32] Tristan Thrush, Ryan Jiang, Max Bartolo, et al. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pages 5238–5248, 2022.

[33] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023.

[34] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, et al. Investigating gender bias in language models using causal mediation analysis. *NeurIPS*, 33:12388–12401, 2020.

[35] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. In *AAAI*, volume 38, pages 5544–5552, 2024.

[36] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *ICLR*, 2023.

[37] Arman Zarei, Keivan Rezaei, Samyadeep Basu, Mehrdad Saberi, Mazda Moayeri, Priyatham Kattakinda, and Soheil Feizi. Improving compositional attribute binding in text-to-image generative models via enhanced text embeddings. *arXiv preprint arXiv:2406.07844*, 2024.

Appendix Contents

A Extended Related Work 14
B CSI Definition and Discovery Details 14
C Embedding-Level Repair Details 15
D Intervention Modality Matrix 15
E DA Stability Across Conditions 16
F Same-Site vs Relocated Comparison 16
G Full Per-Block CE Tables 16
G.1 CE under selective additive boost 16
G.2 CE under selective subtractive attr_zero 17
G.3 CE under broad subtractive zeroing 17
H ABM Metric and Validation 17
I Color Hit Rate Details 18
J Rebuttal Sensitivity Checks 18
K Block Inventories 18
L Per-Timestep Phase Analysis 19
M Power Analysis for Decoder Boost 19
N Prompt Examples 19
O Fractional Attenuation Curve 19
P Compute Resources 20

A Extended Related Work

A growing body of work examines compositional generation and evaluation in controlled settings. Okawa et al. [25] show that compositional abilities in diffusion models emerge multiplicatively rather than additively, consistent with our finding that compositional entanglement in text representations resists simple additive decomposition. Marioriyad et al. [23] provide a systematic evaluation comparing diffusion and autoregressive architectures on compositionality, reinforcing that compositional failures are architecture-general. Broader evaluation suites—T2I-CompBench++ [14], Winoground [32], CREPE [21], and SugarCrepe [13]—measure compositional success at the output level across attribute, relation, and object categories; our work is complementary, diagnosing *where* inside the pipeline these failures originate rather than *whether* they occur.

Object-centric and disentangled representation learning [20] provides a conceptual parallel: slot-based models enforce compositional structure by design, whereas text-to-image diffusion models inherit whatever compositional structure CLIP-family encoders provide. Energy-based compositional generation [5] offers an alternative compositional mechanism through explicit concept combination at sampling time. Our per-block analysis connects to these perspectives by showing that D2's cross-attention—the site where compositional defects are most visible—does not support token-level decomposition of the defect, consistent with a representation that lacks the disentangled structure needed for targeted repair.

B CSI Definition and Discovery Details

The non-factorization residual R (Eq. 1) is computed under three low-complexity composition models: additive, low-rank linear, and low-rank bilinear. Composition parameters are fit within training folds only (leave-one-group-out); test groups are never used for fitting.

Encoder extraction. SD1.5 uses the sequence mean of the final CLIP hidden states (768 dimensions). SDXL concatenates the sequence means of the final CLIP-L (768) and OpenCLIP-G (1280) hidden states with the projected pooled OpenCLIP-G output (1280), yielding one 3328-dimensional vector. A single composition model is fit to this concatenated representation; encoder-wise CSI scores are not averaged. SD3 uses only the sequence mean of the final T5-XXL hidden states, not joint CLIP/T5 conditioning. Each final vector is ℓ_2-normalized.

The three composition models are:

- **Additive**: $\hat{h} = h(p_{\text{attr}}) + h(p_{\text{obj}}) + h(p_{\text{ctx}})$.
- **Low-rank linear**: $\hat{h} = W_{\text{attr}}h(p_{\text{attr}}) + W_{\text{obj}}h(p_{\text{obj}}) + W_{\text{ctx}}h(p_{\text{ctx}})$, where W_i are rank-r matrices fit per LOGO fold.
- **Low-rank bilinear**: adds a rank-r bilinear interaction term between attribute and object representations.

For span geometry, define component matrix $B = [h(p_{\text{attr}}), h(p_{\text{obj}}), h(p_{\text{ctx}})] \in \mathbb{R}^{d\times 3}$ and $f = h(p_{\text{full}})$. The projected vector is $\hat{f} = B(B^\top B + \lambda I)^{-1}B^\top f$, yielding:

$$R_{\text{span}} = \frac{\|f - \hat{f}\|_2}{\|f\|_2 + \varepsilon}, \qquad \theta_{\text{span}} = \arccos\left(\frac{\|\hat{f}\|_2}{\|f\|_2 + \varepsilon}\right). \tag{3}$$

CSI aggregates text-residual and span-geometry features:

$$\text{CSI}_{\text{text}} = \tfrac{1}{2}(z(r_{\text{lin}}) + z(r_{\text{bilin}})), \quad \text{CSI}_{\text{span}} = \tfrac{1}{2}(z(R_{\text{span}}) + z(\theta_{\text{span}})), \tag{4}$$

$$\text{CSI} = \tfrac{1}{2}(\text{CSI}_{\text{text}} + \text{CSI}_{\text{span}}), \tag{5}$$

where $z(\cdot)$ is robust standardization using anchor-only statistics per model. No weights are tuned on downstream labels. Specifically, $z(v) = (v - \text{median}_A\, v)/(s_A + 10^{-8})$, where s_A is the unscaled median absolute deviation on anchors; a non-finite or near-zero value ($< 10^{-8}$) falls back to the population standard deviation, and then to 1 if still degenerate.

Calibration and diagnostic scope. The saved discovery scores use all anchors within each model for this standardization, separately from fold-wise composition fitting. They therefore do not establish fully held-out calibration or deployment performance. Pair-level partitions can also share an anchor text across groups and do not establish generalization to unseen anchor texts. The saved span features include internal mid-attention quantities, so only $\mathrm{CSI}_{\mathrm{text}}$ supports a strictly pre-denoising interpretation; combined-CSI associations are retrospective diagnostics and should not be attributed to a text-only score without recomputation.

Discovery is assessed by a Ladder Consistency Criterion (LCC): monotonic decile trend, first-to-last delta CI excluding 0, and consistent direction across models. The stricter Correlation Threshold Sensitivity check ($\rho \geq 0.90$) is reported as sensitivity analysis only.

Several robustness analyses confirm that the CSI signal is not an artifact. Group-preserving label-permutation yields null ROC-AUC $= 0.49 \pm 0.01$ against a true ROC-AUC of 0.89–1.00 (all empirical $p = 0.001$), and CSI ROC-AUC=1.0 holds in all four conflict-type strata (chromatic oddness, material–object, geometric pattern, state–environment), so the signal is not driven by a single attribute class.

We further apply a three-level covariate control protocol. Level 1 uses covariate-adjusted analyses controlling for perplexity, subtoken count, attribute frequency, prompt length, and template family. Level 2 uses near-matched subsets (n_{pairs}=140, canonical tier); near-matched LOGO residual effects remain significant (linear mean diff=0.543, q<0.001; bilinear mean diff=0.681, q<0.001), and strict (n=112) and relaxed (n=178) tiers preserve the same pattern. Level 3 bounds interpretation: these controls weaken lexical/tokenization explanations but do not eliminate all language-side priors.

A predictive-asymmetry breakdown clarifies which CSI families carry the signal. Within-dataset family ablation across 7 partitions yields combined ROC-AUC=0.9999, text-only=0.9999, and span-only=0.6695; cross-model transfer across 30 directions yields text-only mean ROC-AUC=0.9999, combined=0.9986, and span-only=0.5745. Text families dominate; span-only is near chance in transfer.

C Embedding-Level Repair Details

Two low-rank adapter families (rank 32, AdamW, 800 epochs, MSE objective) are evaluated. The original pooled-adapter run has 20 representation-evaluable prompts ($\Delta_{\text{repr}} = -0.0998$, q=0.0024); the original token-adapter run has 19 (-0.1836, q=0.0008). Each has seven CHR-evaluable prompts with observed zero paired change (p=q=1); this small-sample observation does not establish equivalence. Within each run, BH correction covers representation residual and CHR. Separate full-scale rank-sweep runs evaluate 216 representation prompts and 52 CHR prompts: pooled adapters show no significant positive CHR effect, while token adapters have negative CHR effects. Their residual improvements are not monotonic in rank; these runs must not be conflated with the original small runs. No correction across all rank configurations is claimed. WiCLP [37] uses a binding-aligned objective, so our findings remain specific to the tested residual-minimizing objectives.

For E5 localization, CHR is evaluated on six prompts after averaging five seeds; downstream ABM uses 24 prompts. Downstream BH correction covers early ABM, mid ABM, and CHR, giving CHR q=0.046875. The embedding arm uses its separate two-endpoint family (residual and CHR), giving CHR q=0.03125. With six prompts, exact two-sided sign-flip inference enumerates 64 sign patterns. An exploratory paired comparison of CHR gains is $+0.0952$ (95% CI $[0.0434, 0.1509]$, raw p=0.03125); it was not a prespecified superiority test. Identical baseline CHR values do not by themselves verify identical images or masks.

D Intervention Modality Matrix

Table 3 summarizes the selectivity×sign intervention matrix. We test five intervention families across two pipeline stages (embedding and denoiser). At the denoiser level, three of four cells are tested; the "broad additive" cell (uniformly scaling all post-softmax attention entries) is a mathematical identity under renormalization and is therefore untested. A proper broad-additive control would require pre-softmax logit manipulation, which our hook architecture does not support.

Table 3: Selectivity × sign intervention matrix. Three denoiser cells and two embedding cells are tested; broad additive is a no-op under post-softmax renormalization.

	Selective (token-level)	**Broad** (block/pool-wide)
Embedding additive	Token adapter (MSE)	Pooled adapter
Denoiser additive	Boost (attr × obj mask)	Untested (identity)
Denoiser subtractive	Attr_zero (attr × obj mask)	Zero (uniform $1/k$)

E DA Stability Across Conditions

D2 has the largest positive signed DA mean in all six baseline datasets, and its contrast against zero survives within-model, within-run FDR (Table 4). Boost and attr-zero baseline statistics coincide and must not be treated as independent replications. No prompt IDs span multiple group IDs in these files, but shared anchor texts or base identities remain a separate dependence question.

Table 4: DA at D2 across all experimental conditions. D2 is the DA peak in every condition.

Model	Condition	DA(D2)	FDR q
SDXL	Boost baseline	$+0.0039$	0.0174
SDXL	Zero baseline	$+0.0035$	0.0332
SDXL	Attr_zero baseline	$+0.0039$	0.0174
SD1.5	Boost baseline	$+0.0282$	0.00028
SD1.5	Zero baseline	$+0.0296$	0.00028
SD1.5	Attr_zero baseline	$+0.0282$	0.00028

F Same-Site vs Relocated Comparison

Table 5 reports the phase1 paired comparisons after averaging five seeds within each of 52 prompts. Decoder-minus-D2 is nominally significant in SDXL (p=0.0028), but not SD1.5 (p=0.1882). Neither D2 baseline contrast is significant; this does not establish equivalence.

Table 5: Same-site (D2) vs relocated (decoder) under boost. Tests use 52 prompt means after pairing five seeds; displayed significance is nominal, before multiplicity correction.

Model	Comparison	Δ CHR	p	Sig.?
SDXL	Same-site (D2) vs baseline	-0.0013	0.4093	No
	Relocated (U0) vs baseline	$+0.0061$	0.0168	Yes
	Relocated vs same-site	$\mathbf{+0.0074}$	$\mathbf{0.0028}$	**Yes**
SD1.5	Same-site (D2) vs baseline	$+0.0021$	0.4819	No
	Relocated (U1) vs baseline	$+0.0084$	0.0384	Yes
	Relocated vs same-site	$+0.0063$	0.1882	No

G Full Per-Block CE Tables

Tables 6, 7, and 8 report prompt-level CE at all active blocks. Boost uses the phase1 runs throughout; earlier expansion results are not mixed into its rows. Boost and broad-zero tables display raw p, with bold effects only when within-run BH $q < 0.05$; attr-zero displays q. No boost block survives FDR. Attr-zero has no significant positive effect, with a negative effect at SD1.5 D0. Broad zeroing improves SDXL D1 and decreases SDXL U0 and SD1.5 D0/D1 after FDR; SD1.5 D2 and U2 do not survive correction.

G.1 CE under selective additive boost (432 prompts × 5 seeds)

Table 6 reports per-block CE under selective additive boost for SDXL and SD1.5. The largest positive mean responses occur at decoder blocks U0 and U1, respectively, but no block survives FDR correction.

Table 6: CE_{boost} per block. Columns show raw p-values. No block survives FDR correction; decoder peaks (U0, U1) are nominally significant with consistent cross-model direction.

Block	SDXL CE	SDXL p	SD1.5 CE	SD1.5 p
D0	–	–	-0.0031	0.1854
D1	-0.0002	0.8826	-0.0041	0.0182
D2	-0.0013	0.4093	$+0.0021$	0.4819
M	$+0.0018$	0.1242	-0.0015	0.2763
U0	$+0.0061$	0.0168	–	–
U1	$+0.0011$	0.1788	$+0.0084$	0.0384
U2	–	–	$+0.0002$	0.8658
U3	–	–	-0.0005	0.6401

G.2 CE under selective subtractive attr_zero (432 × 5 seeds, both models full-scale)

Table 7 reports per-block CE under selective attribute-token zeroing for both models, together with BH-adjusted q-values. No block shows a significant positive effect; SD1.5 D0 shows a significant negative effect.

Table 7: $\text{CE}_{\text{attr_zero}}$ per block with prompt-level BH q. SD1.5 D0 has a small significant negative effect; no block has a significant positive effect.

Block	SDXL CE	SDXL FDR q	SD1.5 CE	SD1.5 FDR q
D0	–	–	$\mathbf{-0.0050}$	0.0014
D1	$+0.0003$	0.8514	-0.0010	0.4669
D2	-0.0011	0.8514	-0.0029	0.0602
M	-0.0003	0.8514	-0.0022	0.1225
U0	-0.0002	0.8514	–	–
U1	$+0.0007$	0.8514	-0.0031	0.0970
U2	–	–	-0.0018	0.2634
U3	–	–	-0.0031	0.2312

G.3 CE under broad subtractive zeroing (432 × 3 seeds)

Table 8 reports per-block CE under broad subtractive zeroing for both models. After FDR correction, zeroing increases CHR at SDXL D1 and decreases CHR at SDXL U0 and SD1.5 D0/D1.

Table 8: CE_{zero} per block. Bold: FDR $q < 0.05$. Columns show raw p-values; boldface indicates FDR-corrected significance.

Block	SDXL CE	SDXL p	SD1.5 CE	SD1.5 p
D0	–	–	$\mathbf{-0.2199}$	0.0002
D1	$\mathbf{+0.0200}$	0.0132	$\mathbf{-0.1731}$	0.0002
D2	$+0.0054$	0.6619	$+0.0577$	0.0356
M	-0.0037	0.4395	$+0.0123$	0.3511
U0	$\mathbf{-0.1338}$	0.0002	–	–
U1	$+0.0008$	0.9126	$+0.0207$	0.4597
U2	–	–	-0.0576	0.0286
U3	–	–	$+0.0296$	0.4325

H ABM Metric and Validation

ABM is computed per cross-attention head per timestep. For probability matrix $P \in \mathbb{R}^{q \times k}$: (1) object map $\mathbf{o} = \sum_{j \in \text{obj}} P_{:,j}$; (2) attribute map $\mathbf{a} = \sum_{j \in \text{attr}} P_{:,j}$; (3) mask $M_i = \mathbf{1}[\mathbf{o}_i \geq Q_{0.80}(\mathbf{o})]$; (4) ABM $= \sum_i \mathbf{a}_i M_i / \sum_i \mathbf{a}_i$. Averaged across heads, timesteps, and seeds.

To validate ABM as a proxy for compositional binding, we correlate it with the CHR endpoint. Stress-prompt ABM at D2 correlates negatively with stress CHR: SDXL $\rho = -0.235$ (p=0.0001); SD1.5 $\rho = -0.124$ (p=0.046). The correlation is modest ($R^2 \approx 5\%$), indicating ABM is a meaningful but partial proxy. DA (anchor−stress ABM) is weakly positively correlated with stress CHR ($\rho = +0.17$);

this is an arithmetic consequence: anchor ABM is uncorrelated with stress CHR ($\rho \approx 0$), so DA inherits its structure from the stress ABM term.

I Color Hit Rate Details

The object-local color hit rate (CHR) measures target-color agreement within an attention-derived object region. In the audited full per-block runs, 52 of 216 stress prompts have finite CHR across the required seeds; the other 164 have no target color recorded by the scorer. Baseline diagnostics did not attribute these exclusions to invalid masks. Absence of a recognized target color is not a verified semantic category label. Seed-level scores are paired within prompt and then averaged, giving 52 statistical units for CE, rather than $52 \times$ seeds. Earlier expansion runs have only six CHR-evaluable stress prompts and are not used in the phase1 boost table.

11 color prototypes (RGB, 0–1 scale): red (0.87,0.21,0.21), orange (0.95,0.56,0.20), yellow (0.93,0.86,0.22), green (0.23,0.70,0.31), blue (0.22,0.40,0.90), purple (0.58,0.35,0.86), pink (0.95,0.60,0.76), brown (0.55,0.36,0.24), black (0.08,0.08,0.08), white (0.95,0.95,0.95), gray (0.55,0.55,0.55). Distance threshold: 0.35. Object mask quantile: 0.80. Minimum mask fraction: 0.002. Scoring eligibility follows the recorded target colors and finite endpoint values; a separate prompt taxonomy must not be inferred from missing CHR alone.

J Rebuttal Sensitivity Checks

A separate held-out SDXL evaluation used 220 prompts, five seeds, and 20 active cross-attention modules per site (3.01% mean-JS mismatch). U0–baseline CHR was $+0.003775$ (95% CI $[+0.002328, +0.005240]$), and U0–D2 was $+0.008149$ ($[+0.006190, +0.010142]$). These estimates use prompt-level inference and are separate from the original experiments. With baseline-only GroundingDINO+SAM masks shared across all arms, the U0–baseline interval crossed zero, while U0–D2 remained positive. A continuous target-color score (negative mean CIEDE2000) showed comparable U0–baseline changes inside the mask ($+0.154668$) and in the background ($+0.156614$); the inside-minus-background contrast was -0.001946 (95% CI $[-0.055536, +0.054245]$). Thus increased CHR alone does not establish localized semantic repair.

K Block Inventories

SDXL and SD1.5 have different numbers of cross-attention modules per block group (Table 9). Counting cross-attention (`attn2`) only, SDXL concentrates modules in D2 (20) and U0 (30); the retained SD1.5 inventory counts all patched attention modules (2–6 per group), including self-attention, rather than verified cross-attention-only modules. D0 has no cross-attention in SDXL; U0 has none in SD1.5. Post-hoc module normalization alone does not rule out intervention-exposure differences between sites.

Table 9: Module inventories per block group. SDXL counts cross-attention only; SD1.5 retains the legacy all-patched attention inventory. The two columns use different counting conventions and must not be compared as intervention exposure.

Block	SDXL	SD1.5
D0	0	4
D1	4	4
D2	20	4
M	10	2
U0	30	0
U1	6	6
U2	0	6
U3	–	6

L Per-Timestep Phase Analysis

To rule out that all-timestep averaging masks a controllable window at D2, we run boost restricted to early ($\leq 33\%$), mid (33–66%), or late ($> 66\%$) denoising phases separately (Table 10). D2's DA is strongest in the early phase ($+0.016$), yet boost CE is most negative there (-0.010). The mid-phase D2 mean is slightly positive ($+0.001$); this descriptive table does not establish a significant benefit. U0 boost CE is also strongest in the early phase ($+0.008$), consistent with compositional decisions being made early.

Table 10: Per-phase DA and CE at D2 (SDXL, 100 prompts $\times$ 3 seeds).

Phase	D2 DA	D2 CE (boost)	U0 CE (boost)
Early	$+0.016$	-0.010	$+0.008$
Mid	$+0.013$	$+0.001$	$+0.002$
Late	$+0.008$	-0.000	$+0.001$

M Power Analysis for Decoder Boost

We address whether the FDR non-significance of the decoder-vs-baseline boost reflects effect absence or sample-size limitation. For each model we compute the per-prompt effect size d_z (mean paired difference divided by its standard deviation) on the color-stress subset and the power of a paired test at the current sample size.

Effect size on the full color-stress subset (n=52 stress prompts $\times$ 5 seeds, aggregated to per-prompt means) is d_z=0.345 for SDXL (U0) and d_z=0.297 for SD1.5 (U1). These are conventionally classified as small-to-moderate. Power at the current sample size is 0.685 (SDXL) and 0.557 (SD1.5). Reaching 80% power would require 68 and 91 color stress prompts respectively—i.e., 16 to 39 additional prompts beyond our current 52. BH thresholds depend on the ranks of all tested p-values and do not impose a fixed factor-of-m penalty on every comparison. These descriptive power estimates do not establish why a result fails FDR.

The six-prompt E5 localization subset has prompt-level d_z=1.04 for downstream CHR. This selected subset uses a separate protocol; its effect size does not establish that higher CSI predicts greater response or that full-set averaging merely dilutes a common effect.

These power calculations are conditional on the observed effect sizes and do not establish that the underlying effect is nonzero or that limited power explains FDR non-significance. The paired same-site versus relocated comparison is reported separately from the per-block comparisons.

Table 11: Effect size and power for decoder boost at the full color-stress subset.

Model	Block	d_z (per-prompt)	n_{prompt}	Power	n for 80% power
SDXL	U0	0.345	52	0.685	68
SD1.5	U1	0.297	52	0.557	91

N Prompt Examples

Table 12 shows representative anchor–stress pairs from the 432-prompt dataset. Each group shares the same object and context; only the target attribute changes. The stress attribute is selected from a lower-frequency pool to induce compositional rarity. Conflict types span four categories: chromatic oddness (e.g., blue pear), material–object (glass cat, metallic banana), geometric pattern (striped refrigerator), and state–environment (dry horse).

O Fractional Attenuation Curve

To rule out the possibility that full attribute zeroing (α=0) is simply too aggressive, we test fractional attenuation at $\alpha \in \{0.00, 0.25, 0.50, 0.75\}$ on SDXL (432 prompts $\times$ 3 seeds), where α scales the attribute-token attention weight and α=1.0 is the unmodified baseline. Table 13 reports CE at D2 and U0 (decoder) at each level.

Table 12: Representative anchor–stress pairs.

Group	Anchor	Stress
B001	a studio photo of a black cat	a studio photo of a glass cat
B005	a studio photo of a yellow banana	a studio photo of a metallic banana
B028	a studio photo of a green pear	a studio photo of a blue pear
B045	a studio photo of a brown guitar	a studio photo of a gold guitar

All effect sizes remain below ± 0.002, well within noise (compare decoder boost CE $= +0.006$, nominally significant at p=0.0168). D2 shows a weakly monotonic trend (ρ=1.0, but n=4; random chance yields this ordering with probability $1/24$), with direction consistent with broad ablation's positive effect (both reduce attribute signal at D2), but the magnitude is two orders of magnitude smaller than broad ablation (SD1.5 D2 zeroing CE $= +0.058$). U0 is completely flat (ρ=0.0): the decoder responds to additive amplification (boost) but not to attenuation of existing attribute attention at any level, consistent with the "competing signal" mechanism described in §7.

Table 13: Fractional attenuation CE at D2 and U0 (SDXL, n=156 color-prompt pairs per level). No attenuation level produces practically meaningful effect sizes.

α	D2 CE	U0 CE
0.00	-0.0006	$+0.0002$
0.25	$+0.0003$	-0.0006
0.50	$+0.0008$	$+0.0005$
0.75	$+0.0011$	$+0.0002$
Reference: boost U0 $= +0.006$ (p=0.0168)		
Reference: zeroing D2 $= +0.058$ (SD1.5, q=0.0623)		

P Compute Resources

All experiments were run on NVIDIA A100 (80GB) GPUs via SLURM. Table 14 breaks down approximate GPU-hours by experiment phase. The total ($\sim$159 GPU-hours) includes all three analysis components (CSI pipeline, embedding repair, and per-block modality decomposition) across two primary models and one extension model.

Table 14: Compute budget by experiment phase.

Experiment	GPU-hours (approx.)
CSI pipeline (cross-model factorization, SD1.5+SDXL+SD3)	24
Embedding repair (pooled + token adapters, rank sweep)	12
Output-side BLIP scoring	4
Per-block boost sweep (432 $\times$ 5 seeds $\times$ 2 models)	43
Per-block zeroing sweep (432 $\times$ 3 seeds $\times$ 2 models)	20
Attr_zero full-scale (432 $\times$ 5 seeds $\times$ 2 models)	36
Fractional attenuation (432 $\times$ 3 seeds $\times$ 4 levels)	8
Per-timestep phase analysis	4
Qualitative image generation	1
CLIP auxiliary subset	7
Total	$\sim$159