# FILLING THE SLOTS: REPEATED MENTIONS GET THEIR ATTRIBUTES BEFORE THEIR INSTANCES IN MMDIT TEXT-TO-IMAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Text-to-image counting benchmarks phrase multiplicity almost exclusively with numerals ("three bicycles"). We show that enumerating the same objects as repeated mentions ("a bicycle, a bicycle and a bicycle") instead lowers counting accuracy by 44.3 points in SD3.5-medium and 59.2 points in FLUX.1-dev, a failure that numeral-only evaluation cannot detect. The penalty is not explained by same-category numerosity: repeated mentions lose instances, and cues that distinguish them, such as contrastive determiners and distinct colors, recover far more instances in FLUX.1-dev than in SD3.5-medium, which instead renders the requested colors on too few objects. Scoring counting and color binding separately shows that both models bind colors once the count is right, so the asymmetry lies in instances. Distinct colors raise the separability of the mentions' attention maps early in both models, but this predicts success only in FLUX.1-dev. In SD3.5-medium, the two ways a mention loses its instance have distinct early signatures: an omitted mention's image-to-text attention decays on its own, whereas fused mentions keep their attention but are never given separate regions. Intervening on joint attention shows that restoring early attention to a mention causally restores its color, through the early image layout, but has no specific effect on its instance. A rendering correction built on this mechanism, EMAR, restores missing color bindings and, across six prompt structures, raises counting accuracy by 7–15 points while lowering omission and, wherever a noun repeats, fusion.

## 1 INTRODUCTION

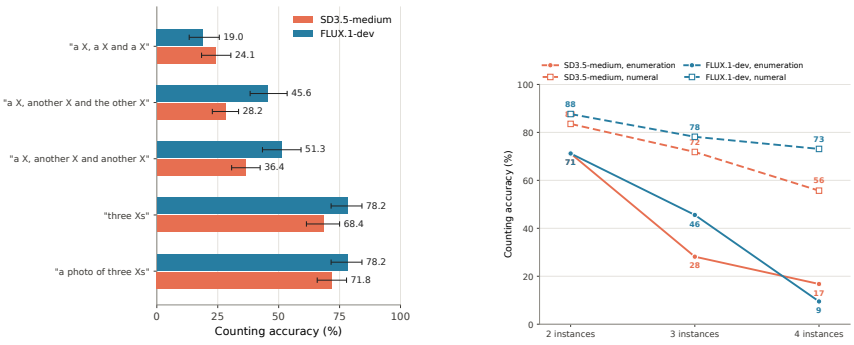
Text-to-image (T2I) models built on Multimodal Diffusion Transformers (MMDiT) (Esser et al., 2024; Labs, 2024) jointly attend over text and image tokens and have substantially improved prompt alignment. Their counting ability is measured by benchmarks such as GenEval (Ghosh et al., 2023) and T2I-CompBench(++) (Huang et al., 2023; 2025), which vary object count, attributes, and relations. These benchmarks phrase multiplicity almost exclusively with numerals or a small fixed template set. Whether measured counting competence survives a change in how multiplicity is *expressed* has not been tested.

Accuracy drops substantially under alternative phrasing. When evaluating prompts with three bicycles, the bare enumeration 'a X, a X, and a X' degrades performance relative to 'three Xs' by 44.3 percentage points for SD3.5-medium and 59.2 percentage points for FLUX.1-dev (Figure 1a). Giving each mention a distinct color ("a red X, a brown X, and a gold X") still costs SD3.5-medium 34.2 points. The gap is present for every enumeration we test and widens with the number of instances (Figure 1b). A numeral-only benchmark would report neither.

**Question 1: Is the gap just the difficulty of producing many objects of one kind?** No. Numeral-matched controls show that both models count "three bicycles" more accurately than three distinct objects, yet the penalty is large whenever every mention repeats one noun, and repetition turns overcounts into omissions. The difficulty lies in realizing several instances from repeated mentions.



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.




(a) Five phrasings, both models.

(b) Repeated-instance scaling.

Figure 1: **Counting accuracy for the same request under different phrasings.** (a) Three bicycles, as three enumerations and two numeral forms. (b) Two to four repeated instances; solid lines enumerate, dashed lines use a numeral. Error bars: 95% prompt-clustered bootstrap CIs.

**Question 2: What makes repeated mentions easier to enumerate?** Cues that tell the mentions apart, contrastive determiners and distinct colors, raise counting far more in FLUX.1-dev than in SD3.5-medium, and only when a noun repeats. Scoring counting and binding separately shows that SD3.5-medium renders the requested colors but on too few objects, and separates genuine instance loss from attribute inheritance.

**Question 3: Does a repeated mention lose its instance and its attribute for the same reason?** No. Intervening on SD3.5-medium's joint attention shows that early attention to a mention causally determines its color, through the early image layout, but not its instance. An omitted mention's attention decays on its own, whereas fused mentions keep their attention but are never separated.

Our contributions are: ❶ an enumeration gap invisible to numeral-only benchmarks, isolated from same-category numerosity by numeral-matched controls; ❷ a cue-use asymmetry between FLUX.1-dev and SD3.5-medium, with failure types that separate lost instances from unbound colors; ❸ a causal account of SD3.5-medium's failures in which early attention controls a mention's attribute but not its instance; and ❹ EMAR, a correction derived from this account that restores missing color bindings in 17 of 19 failing cases and raises counting accuracy across prompt structures.

## 2 Related Work

**MMDiT generation and evaluation.** Diffusion Transformers (Peebles & Xie, 2023) and their multimodal extensions (Esser et al., 2024; Labs, 2024) process text and image tokens in one joint attention (Appendix A.1). MMDiT compositional fidelity is commonly measured by GenEval (Ghosh et al., 2023) and T2I-CompBench(++) (Huang et al., 2023; 2025), which test counting, attributes, and relations using short templates. These benchmarks vary the requested content, but generally do not compare linguistically different realizations of the same multiplicity. Our evaluation instead holds the visual requirement fixed and contrasts numeral expression with explicit noun enumeration.

**Diagnosing and correcting MMDiT alignment.** Recent work rebalances cross-modal attention (Lv et al., 2025), exploits attention sparsity (Kim & Sim, 2025), modifies textual representations (Li et al., 2026), or detects and intervenes on internal omission signals (Baek et al., 2026). Prompt Reinjection attributes alignment loss to progressive prompt forgetting across depth (Yao et al., 2026), and Seg4Diff identifies layers whose joint attention grounds textual concepts (Kim et al., 2025). These methods are developed within numeral-style, fixed-arity prompting. They also carry practical costs: Lv et al. (2025)'s attention rebalancing incurs an unfused path costing 3.5× unmodified generation, and Baek et al. (2026)'s omission intervention requires a pretrained classifier and per-setting intervention hyperparameters. Applied to repeated-mention enumeration, none of the three methods we tested (Lv et al., 2025; Baek et al., 2026; Yao et al., 2026) brings enumeration close to the numeral form (Appendix A.4).

**Multi-object generation, counting, and binding.** Inference-time methods counter subject neglect by strengthening token attention (Chefer et al., 2023), separating subjects or their embeddings (Dahary et al., 2024; Byun et al., 2026), or optimizing attention-map separation (Meral et al., 2024; Kim et al., 2024). Other work targets accurate counts (Binyamin et al., 2025), similar-subject mixing in MMDiTs (Wei et al., 2026), and multi-instance attribute leakage in FLUX (Zhou et al., 2025). Attribute binding has been improved by aligning entity–modifier attention with syntax (Rassin et al., 2023; Trusca et al., 2024). Our goal is diagnostic, and it bears on the premises of these corrections: strengthening early attention to a mention repairs its attribute but not a missing instance, and fused mentions fail by never separating rather than by lacking attention.

# 3 Experimental Framework

## 3.1 Factorial design

| Noun duplication | | Attribute structure | |
|---|---|---|---|
| Level | Structure | Level | Structure |
| $N_1$ | $\text{noun}_1$, $\text{noun}_1$, $\text{noun}_1$ | $A_\varnothing$ | uncolored |
| $N_2$ | $\text{noun}_1$, $\text{noun}_2$, $\text{noun}_1$ | $A_1$ | $\text{color}_1$, $\text{color}_1$, $\text{color}_1$ |
| $N_3$ | $\text{noun}_1$, $\text{noun}_2$, $\text{noun}_3$ | $A_2$ | $\text{color}_1$, $\text{color}_1$, $\text{color}_2$ |
| | | $A_2^-$ | $\text{color}_1$, $\text{color}_1$, uncolored |
| | | $A_3$ | $\text{color}_1$, $\text{color}_2$, $\text{color}_3$ |

Table 1: Factorization of three-slot prompts.

The main suite uses **three slots**, the smallest number at which all duplication regimes exist. The noun-duplication axis $N$ is indexed by the number of distinct nouns: three distinct nouns ($N_3$), one repeated plus one distinct noun ($N_2$), and one noun repeated throughout ($N_1$). The attribute axis is indexed by the number of distinct colors. $A_\varnothing$ is colorless. When every slot is colored, $A_1$ shares one color across all slots, $A_2$ shares one color across two slots and gives the third a different one, and $A_3$ gives every slot a distinct color. $A_2^-$ shares one color across two slots and leaves the third uncolored (Table 1).

Colorless and shared-color prompts use contrastive determiners: $N_1$ is "a X, another X and the other X", and $N_2$ is "a X, a Y and another X". Prompts that give each repeated slot its own color use the indefinite article throughout ("a red X, a brown X and a gold X"). Numeral counterparts are "three Xs" and "one Y and two Xs". The two- and four-instance numeral forms follow the GenEval template ("a photo of two Xs"); at three instances this template and the bare numeral are statistically indistinguishable (SD3.5-medium $68.4\%$ vs. $71.8\%$, difference CI $[-5.7, 12.7]$; FLUX.1-dev $78.2\%$ vs. $78.2\%$, CI $[-8.5, 9.2]$). Nouns are 79 COCO object categories and colors are 20 basic color terms (Appendix A.3).

**Models and evaluation.** We evaluate SD3.5-medium and FLUX.1-dev in their standard configurations on 79 prompt configurations $\times$ 4 seeds per cell. Qwen2.5-VL-7B-Instruct counts each requested base noun and color–noun pair. *Counting accuracy*, the primary endpoint, requires every base-noun count to match; *combined accuracy* also requires every color–noun count to match, with a *leakage-robust* variant ($\geq$). Undercounts are *fusion* if every requested color–noun pair is present and *omission* otherwise. Details are in Appendix A.2.

# 4 Results

## 4.1 Does enumeration matter beyond numerosity?

Without colors, counting accuracy falls as nouns repeat, mostly at the first repetition (Appendix Figure 6a): from $57.6\%$ at $N_3$ to $35.1\%$ and $28.2\%$ at $N_2$ and $N_1$ in SD3.5-medium, and from $61.7\%$ to $40.2\%$ and $45.6\%$ in FLUX.1-dev. Appendix Table 4 pairs each enumeration with a numeral prompt for the same objects. **Pure repetition** is worse than the numeral form in both models and at every instance count, by 12.3 to 63.6 points, although both models count "three bicycles" more

accurately (68.4% / 78.2% for SD3.5-medium / FLUX.1-dev) than three different objects (57.6% / 61.7%). With one distinct noun (**mixed prompts**), FLUX.1-dev still loses 13.0 and 10.8 points, while SD3.5-medium shows no penalty, so its $N_3 \to N_2$ drop reflects mixed-category multiplicity.

**Repetition loses instances.** The noun axis also changes how counting fails (Figure 2). With distinct nouns, most miscounts are overcounts (32.3% / 23.4% of colorless $N_3$ images, against 10.1% / 14.9% omissions). With one noun repeated they are omissions (62.0% / 45.9% at $N_1$), a rise of 51.9 [44.6, 58.9] and 31.0 [20.3, 41.1] points, while overcounts fall by 22.5 [14.2, 30.4] and 14.9 [9.2, 20.9].

What makes enumeration hard is therefore the repetition of a mention, not the objects it requests: the model must assign each of several identical mentions its own instance, and nothing in the mentions says which is which.

### 4.2 What rescues repeated-mention enumeration?

If the difficulty is telling identical mentions apart, cues that distinguish them should help. Appendix Table 5 reports two such cues at $N_1$, where the penalty is largest, and Appendix Figures 6b–6c show the attribute axis at each duplication level.

**Determiners and distinct colors.** Replacing "a X, a X and a X" with "a X, another X and another X" raises counting accuracy by 32.3 points for FLUX.1-dev but by 12.3 for SD3.5-medium; with the determiners held fixed, three distinct colors add 55.1 and 10.1 points. The gain comes from distinctness, not from color as such: along the $N_1$ line (orange) of Appendix Figures 6b–6c, one shared color ($A_1$) does not help, and each additional distinct color raises accuracy (Appendix Table 5, lower rows). In FLUX.1-dev, three objects of one category with distinct colors are enumerated more accurately than three different objects (74.1% vs. 56.0%); in SD3.5-medium the line stays well below $N_3$. $N_2$ follows the same ordering (Appendix A.5).

**Colors bind once the count is right.** At $N_1 \times A_3$, 89.8% (SD3.5-medium) and 97.0% (FLUX.1-dev) of count-correct images also carry every requested color–noun pair, so leakage-robust combined accuracy follows counting (30.7% and 71.8%). The cue-use asymmetry at $N_1$ is an asymmetry in instances, not in binding.

**When nouns differ, colors cost binding, not counts.** At $N_3$ color structure has no detectable effect on counting in either model (Appendix A.5); colors help only where they are the one property that separates otherwise identical mentions. They do affect binding. Among count-correct $N_3$ images, the share with every color bound falls from 93.0% at $A_1$ to 78.2% at $A_3$ in SD3.5-medium ($-14.8$ [$-23.1$, $-6.2$]) and stays level in FLUX.1-dev (92.0% $\to$ 93.8%, $+1.8$ [$-4.9$, 8.9]), which binds 15.6 [7.9, 23.9] points more at $A_3$. SD3.5-medium's binding of several distinct colors thus weakens even when the nouns keep the mentions apart.

Both cues thus help far more in FLUX.1-dev than in SD3.5-medium. We call this the *cue-use asymmetry*.

### 4.3 What forms do the failures take?

The cue-use asymmetry of Section 4.2 could reflect either a genuine difference in benefit or a difference in how the same benefit surfaces; Figure 2 distinguishes these by outcome type.

**Where SD3.5-medium's color evidence goes.** Because fusion needs two distinct colors on the repeated noun, the quantity comparable along the $N_1$ rows of Figure 2 is the total undercount, omission plus fusion (gray and orange). In FLUX.1-dev, distinct colors remove undercounts (44.3% $\to$ 25.0% $\to$ 23.1% from $A_1$ to $A_3$) and fusion stays minor (10.1%, 16.1%). In SD3.5-medium, undercounts barely change (67.1% $\to$ 61.7% $\to$ 60.1%); what changes is their form: fusion rises from 16.5% to 48.4% while omission falls from 45.3% to 11.7%. Each additional distinct color is realized in SD3.5-medium mainly as one more color on too few objects, and in FLUX.1-dev as one more object (Appendix Figure 19). This excess fusion is specific to full repetition: at $N_2 \times A_3$ the two models fuse equally often (22.8% vs. 21.2%, difference $-1.6$ [$-8.5$, 4.7]) and differ only in omission ($-9.5$ [$-17.1$, $-2.2$] for FLUX.1-dev), whereas at $N_1 \times A_3$ SD3.5-medium fuses 32.3

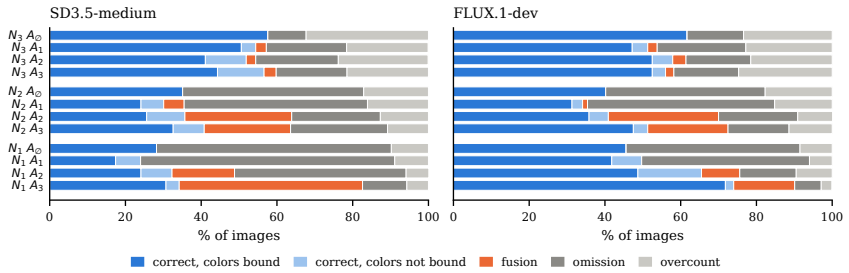



Figure 2: **Outcomes over the $N \times A$ grid.** Fusion requires two distinct colors on the repeated noun (Appendix A.3); elsewhere it reflects disagreement between the evaluator's noun and color–noun counts.

[25.3, 38.9] points more. The colors are rendered, but they do not become separate instances; Section 5.5 traces this to pairs of mentions whose attention is never separated.

**A repeated mention without a cue.** Leaving one repeated slot uncolored ($A_2^-$) gives the lowest combined accuracy in the grid, for two reasons (Appendix A.6). At $N_2$ the bare mention loses instances: removing one color word ("a red bicycle, a red airplane and a brown bicycle" $\rightarrow$ "…and a bicycle") costs $27.8$ counting points in FLUX.1-dev and $12.4$ in SD3.5-medium. At $N_1$ ("a red cat, a cat and another red cat") counting stays near its colorless level and the loss is attribute inheritance: among count-correct images the bare instance is scored in a requested color in $66\%$ / $56\%$ of cases, about twice the evaluator's baseline rate, and leakage-robust scoring recovers most of the combined accuracy ($6.3\% \rightarrow 26.9\%$ / $10.4\% \rightarrow 31.6\%$).

A repeated mention's instance and its attribute thus decouple in both directions: in fusion the attribute appears without its instance, and in inheritance the instance appears without its own attribute. This suggests that instance formation and attribute binding are controlled separately, so an intervention on how the model processes a mention should move one without the other; Section 5 tests this.

### 4.4 An early internal correlate, and its limit

Before intervening, we ask whether the asymmetry is visible in how image tokens attend to the repeated mentions. *Slot separability* $S$ is $1 - \overline{\cos}$ over the pairwise cosine similarities of the three mentions' image-to-text attention maps (Appendix Figure 9; all blocks in Appendix A.8). Distinct colors raise $S$ at the first captured step in both models, before any object is rendered, and the three occurrences are already more distinct in the frozen text encoders (Appendix A.7). Only in FLUX.1-dev does the gain persist and predict the outcome within $A_3$ (block 18, steps 4–9 of 28, $d = 1.35$–$1.51$); in SD3.5-medium it vanishes after the first captured step. As an aggregate over three mentions, $S$ cannot show which mention is lost, and no whole-generation summary separates SD3.5-medium's fusions or omissions from correct generations (Appendices A.9, A.16, and A.17). Section 5 therefore measures single mentions and intervenes on them.

## 5 What does attention to a repeated mention control?

Section 4.3 showed that a repeated mention's attribute and instance can be realized separately, and Section 4.4 that aggregate attention mirrors the cue-use asymmetry without explaining SD3.5-medium's failures. We therefore intervene in SD3.5-medium to ask what attention to a single mention controls. We use single-omission cases for two reasons: the lost mention is identified by its missing color, so its attention can be measured and manipulated; and an omission has lost both its attribute and its instance, so restoring attention shows which of the two returns, whereas a fusion already carries the attribute. If attention controls the attribute but not the instance, this also explains why distinct colors turn SD3.5-medium's omissions into fusions. We show that attention to an omitted mention collapses early (Section 5.1), test what restoring it changes (Section 5.2), trace the pathway (Section 5.3), state the mechanism (Section 5.4), and extend it to fusion (Section 5.5).

### 5.1 OBSERVATION: ATTENTION TO THE OMITTED MENTION COLLAPSES EARLY AND DISSIPATES

For each mention $k$, let $T_k$ be the positions of its color and noun tokens in the CLIP text stream. An image token *attends selectively* to $k$ if its image-to-text attention to $T_k$ is substantial and clearly larger than to the other two mentions (Appendix A.10). The fraction of image tokens attending selectively to $k$ is its *attention coverage*, which we track over the first denoising steps for all 23 single-omission cases (joint blocks 15 and 19).

**Parity, then collapse.** At step 0 the omitted mention's coverage does not differ from that of the same slot and color in correct generations of the same prompt ($+0.004$, 95% CI $[-0.021, +0.029]$). Between steps 0 and 5 it falls by $0.057$ more than this matched reference ($[-0.081, -0.034]$), while the coverage of the two rendered mentions is maintained (Appendix Figure 13a). The omitted mention is not starved of attention from the start; it fails to keep it.

**Dissipation, not capture.** If omission were competition between mentions, the image tokens attending to the omitted mention would be taken over by the rendered ones. They are not (Appendix Figure 13b): only $1\%$ of them attend to another mention by step 6, and they account for $2\%$ of the rendered mentions' gain. Instead, $72\%$ stop attending to object tokens altogether (correct mentions: $46\%$).

**Association with the outcome.** The early coverage predicts which slot will be lost: using steps $\leq 3$, slot-level AUROC is $0.83$ on the original failures and $0.81$ on a position-balanced replication (648 images, colors rotated through all slots), in which slot position has no effect. Low color chroma is an independent risk factor (odds ratio per SD $0.57$ [$0.42$, $0.77$] in the balanced replication). The same early collapse precedes failure when the three nouns are distinct (AUROC $0.79$–$0.83$ from step 2; 14 failing slots), so it is not specific to repeated tokens, although distinct nouns make failure rarer (fully correct $91.7\%$ vs. $82.7\%$, $p = 0.001$).

The collapse could cause the loss or merely reflect an image already drifting away from the object; the interventions below distinguish the two.

### 5.2 INTERVENTION: RESTORING ATTENTION TO THE OMITTED MENTION

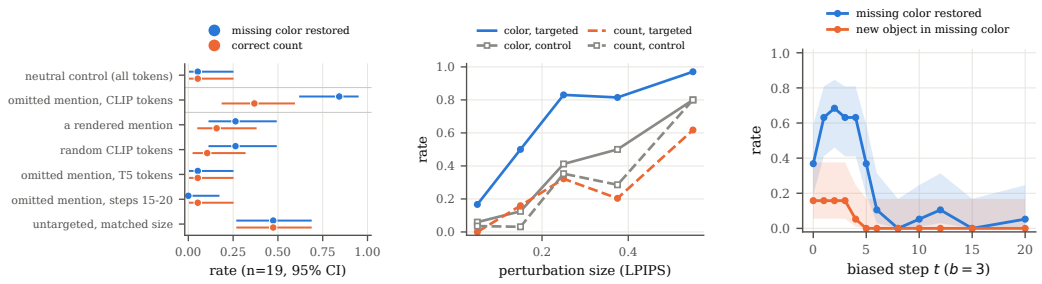
The *targeted attention bias* adds a constant $b$ to the pre-softmax logits from every image token $i$ to the omitted mention's CLIP tokens $T_m$:

$$\ell_{ij} \leftarrow \ell_{ij} + b \cdot \mathbb{1}[j \in T], \qquad t \in \{1, \ldots, 5\}, \text{ joint blocks 9–21} \tag{1}$$

with $T = T_m$ and $b = 2$; control conditions bias other tokens, act late, or perturb the image by a matched amount (Appendix A.10). **Color: a specific effect.** The bias restores the missing color in $16/19$ cases against $1/19$ under a neutral control ($p < 10^{-3}$), and the effect is specific by target (other targets at most $5/19$), by size (odds ratio $5.8$ at matched LPIPS), and by time (none at steps 15–20; Figure 3). **Instance: no specific effect.** The count gain ($7/19$ vs. $1/19$) is matched by an untargeted perturbation of the same size ($9/19$; odds ratio $1.2$ at matched LPIPS, $p = 0.59$), and suppressing attention to rendered mentions neither removes them nor restores the omitted color (Appendix A.10). Early attention thus causes a mention's color to be realized, whereas the instance, settled by about step 3, responds to how strongly the early trajectory is perturbed rather than to which mention is attended.

### 5.3 MEDIATION: THE EFFECT ACTS THROUGH THE EARLY IMAGE LAYOUT

Attention can affect the output only through the evolving image, so we ask whether the targeted bias acts through the coarse layout of the predicted clean image $\hat{x}_0$ (conditions in Appendix A.10). **Blocking the pathway.** Applying the bias while *fixing the coarse layout* (the low-frequency component of $\hat{x}_0$ held at its unmodified value through step 5) still raises attention to the omitted mention, yet color restoration falls from $16/19$ to $6/19$ ($p = 0.002$) and count correction from $7/19$ to $1/19$ ($p = 0.031$); layout fixing alone changes nothing. **Substituting the pathway.** Conversely, a *direct color edit* of $\hat{x}_0$ at the omitted object's location, without touching attention, fades within two steps and restores color no more often than a wrong-color edit ($5/19$ vs. $6/19$), whereas the smaller change made by the attention bias persists (Appendix Figure 14). The early layout is thus a neces-




(a) Bias target ($b = 2$, steps 1–5). (b) Matched perturbation size. (c) Timing of the bias.

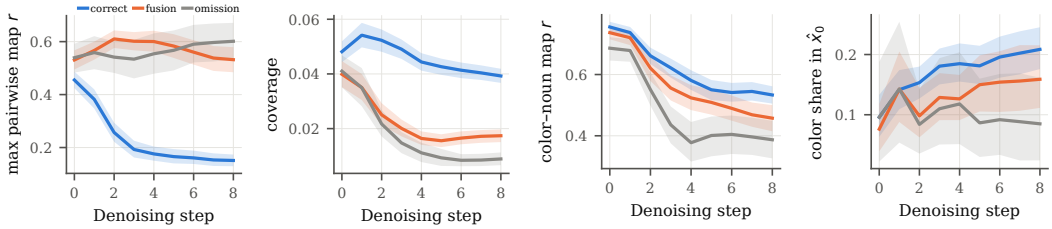
Figure 3: **Intervention on early attention**. (a) Outcomes by bias target. (b) All bias conditions against the image change they cause (LPIPS); targeted conditions bias the omitted mention's CLIP tokens, controls bias other tokens or act late. (c) Bias applied at a single step $t$.

sary mediator but not a sufficient handle: a change that attention does not support is washed out, whereas a change produced through attention is retained.

### 5.4 The causal mechanism

In the first denoising steps, image tokens attend to each mention's color and noun tokens in the CLIP stream. This attention writes the mention's attribute into the coarse image layout, and the changed layout in turn keeps image tokens attending to the mention. A mention whose attention lapses during steps 1–5 loses its coverage by dissipation, and its color is not bound to any object. Sustaining attention to the mention causally restores the attribute, and it does so through the layout. For the *instance*, whether a third object exists appears to be settled by about step 3 and is not specifically controlled by attention to any mention, since only the overall size of the perturbation moves it. Three claims are established directly by intervention: early attention to a mention causes its color to be realized, this attention has no specific effect on whether the instance exists, and its effect on color is mediated by the coarse layout.

### 5.5 Fusion: two mentions that never separate




(a) Map overlap. (b) Coverage. (c) Color binding. (d) Rendered color.

Figure 4: **Fusion vs. omission.** (a) Largest pairwise map overlap. (b–d) The merged pair's weaker mention.

Under distinct colors, SD3.5-medium more often fuses than omits: at $N_1 \times A_3$, $48.4\%$ of generations are fusions and $11.7\%$ omissions (Section 4.3). In a fusion every requested color is rendered, but on too few objects. We characterize fusion on the position-balanced replication of Section 5.1, where attention was captured at each of the first nine denoising steps. Because a fusion involves two mentions sharing one object, we measure pairs of mentions. The *map overlap* of two mentions is the Pearson correlation of their image-to-text attention maps over image tokens (color and noun CLIP tokens, joint blocks 15 and 19). In each generation, the *merged pair* is the pair with the highest overlap at step 5, and its *weaker mention* is the member with lower attention coverage. In omissions, the merged pair contains the lost mention in $94\%$ of cases (chance $67\%$), and its weaker

mention is the lost one in $81\%$; the pair measure thus locates the affected mention (definitions in Appendix A.12).

**Separation fails first.** In correct generations the three mentions' maps separate within the first steps: the largest overlap falls from $0.46$ at step 0 to $0.26$ at step 2 and $0.17$ at step 5 (Figure 4a). In fusions and omissions alike, one pair never separates ($0.61$ and $0.54$ at step 2; $0.58$ and $0.57$ at step 5). This distinguishes fusion from correct generations with AUROC $0.84$ at step 2 and $0.90$ at step 5, and omission with $0.79$ and $0.84$ (Appendix Table 13). The third mention keeps the coverage of a correct generation ($0.069$ vs. $0.063$ at step 5), so the failure is confined to the pair. As in correct generations, the most-overlapping pair is usually adjacent in the prompt ($89\%$ of fusions); in fusion it is, in addition, strongly selected by color similarity (Appendix Table 14). Fusion and omission therefore begin the same way: two mentions that are not given separate regions in the first denoising steps.

**Shared, not dissipated.** The two failures diverge in what happens to the weaker mention of the pair. In omission its coverage dissipates, as in Section 5.1, to $0.009$ at step 5. In fusion it keeps nearly twice as much at step 8, and its color map stays aligned with its own noun map (Figure 4b–c). This retained attention is written into the image. Within the merged pair's region of $\hat{x}_0$, the share of pixels in the weaker mention's color holds steady in fusion from step 1 to step 8, grows in correct generations, and fades in omission (fusion vs. omission at step 8, $p = 0.007$; Figure 4d). A fusion is thus a pair of mentions that share one region while both keep their attention and bind their colors, which the model renders as one object carrying both colors.

**Consistency with the causal mechanism.** Attention to a mention controls its color but not its instance (Section 5.2). Fusion is the state this predicts when two mentions share a region: both colors are realized because both mentions keep their attention, and the instance is missing because the pair never received a second region. The same account predicts how attention reinforcement acts on the two failures. Sustaining attention to every mention (Section 6) turns omissions into fusions (11 of 33) nearly as often as into correct counts (13), and leaves most fusions as they are (95 of 141; Appendix Table 15). Where omission is a mention whose attention decays on its own, fusion is two mentions whose attention persists but is never separated.

## 6 A CORRECTION AS A TEST OF THE MECHANISM

The mechanism of Section 5 was established by biasing attention to a mention already known to be lost. A correction cannot know which mention that is and must act on every generation, so its outcomes test the mechanism out of sample. Each finding of Section 5 yields one prediction. **(P1)** Early attention to a mention causes its color to be realized, and raising attention to a rendered mention does not remove its color (Section 5.2), so reinforcing all mentions should restore a missing color without knowing which one is missing. **(P2)** The instance responds to the size of the early perturbation, not to which mention is attended, so count gains should not exceed those of an untargeted perturbation of matched size, and **(P3)** the correction should add instances to generations that were already correct, producing overcounts. **(P4)** Color and instance are settled within steps 1–5, so the correction should do nothing when applied late, and **(P5)** the effect runs through the coarse layout (Section 5.3), so it should vanish when the layout is held fixed. Meeting P1 while violating P2 or P3 would mean that attention reaches the instance after all; meeting P1 without P4 or P5 would imply a pathway other than the layout.

**Method.** Following P1–P5, the correction acts on attention from image tokens to the mentions' CLIP tokens, only in steps 1–5, and at moderate strength (dose response in Appendix Figure 12). **E**arly **M**ention **A**ttention **R**einforcement (EMAR) first assesses, for each mention, whether its early attention footprint indicates a risk of omission, and then applies Eq. 1 with $T$ the CLIP color and noun tokens of the mentions so identified, $b = 3$, at steps 1–5 and joint blocks 9–21, leaving mentions whose attention is already healthy untouched. It adds no parameters, training, or forward passes.

**Results.** Appendix Figure 15a compares each case before and after EMAR (exact counts out of the 19 single-omission cases; paired bootstrap CIs and exact McNemar tests). EMAR restores the missing color in 17 of 19 omissions, up from 1 ($+84$ points, 95% CI $[+58, +100]$, $p < 10^{-3}$),

and binds all three requested color–object pairs in 8, up from none ($+42$ points, CI $[+21, +63]$, $p = 0.008$). In 7 cases it adds a correct new instance with the rest of the image intact ($+37$ points, CI $[+16, +58]$, $p = 0.016$), and the count becomes correct in 8, up from 1 ($+37$ points, CI $[+11, +63]$, $p = 0.039$). It never removes a color that was already rendered, and it alters the layout in 2 of 19 cases ($-11$ points, CI $[-26, 0]$, $p = 0.50$). EMAR changes images by LPIPS 0.39 on average.

**Which gains follow from the mechanism.** Appendix Figure 15a also compares EMAR with an untargeted perturbation that changes the image as much (the neutral control gives the same values as the unmodified generation). The color gain survives this comparison ($+42$ points, $p = 0.02$), and correct new instances tend in the same direction ($+26$ points, $p = 0.06$). The count gain does not ($-5$ points): perturbing the early trajectory by the same amount corrects the count as often. This is the split the mechanism predicts, in which attention to a mention controls its attribute but not its instance (Appendix Figure 15b). The mechanism also predicts when the correction must fail, and the intervention experiments of Section 5 confirm both conditions for the same bias: applied at steps 15–20 it restores nothing ($0/19$), and with the coarse layout fixed its color restoration drops from $16/19$ to $6/19$.

**Beyond the diagnostic sample.** We applied EMAR to five further $N \times A$ structures and a two-instance prompt, each on close to the main study's sample, and scored every outcome of Appendix A.2 (Appendix Table 12; further structures in Appendix A.11). EMAR raises counting accuracy ($+7.4$ to $+15.0$ points) and leakage-robust combined accuracy ($+5.4$ to $+15.0$) in all six structures, lowers omission in all six ($-5.4$ to $-16.8$) and fusion in every structure with a repeated noun ($-4.2$ to $-12.0$), and keeps binding among count-correct images within about two points of the baseline except at $N_1 \times A_2$ ($-6.0$). At the headline structure $N_1 \times A_3$ (paired images), counting rises from $34.6\%$ to $49.3\%$ and leakage-robust combined accuracy from $30.8\%$ to $44.5\%$.

## 7 Discussion & Limitations

**Discussion.** Numeral-only counting benchmarks overestimate how reliably MMDiT models produce multiple instances when multiplicity is expressed through repeated mentions. Evaluation suites should include such enumerations and score failures by type, because lost instances, colors rendered on too few objects, and inherited colors are different failures. Our results also suggest where corrections should act: strengthening attention to a mention can restore its attribute but not its instance, and fusion calls for separating mentions early rather than adding attention to them.

**Limitations.** Our causal analysis is confined to SD3.5-medium; FLUX.1-dev is characterized only observationally, and the FLUX.1-dev side of the cue-use asymmetry lacks a matching causal account. A preliminary check suggests a different mechanism there (Appendix A.13), which we leave to future work. The suite deliberately restricts nouns and colors to a small, controlled vocabulary over three slots in order to isolate the basic phenomenon; this design does not address how enumeration interacts with more naturalistic combinations of attributes, modifiers, or open-vocabulary nouns, and the main grid relies on a single VLM judge as the counting and binding oracle.

## 8 Conclusion

Expressing multiplicity through repeated mentions exposes a counting failure that numeral-only benchmarks miss: it arises where mentions are lexically identical, consists of lost instances rather than unbound colors, and is relieved by cues that distinguish the mentions, far more in FLUX.1-dev than in SD3.5-medium. In SD3.5-medium, the two ways a repeated mention loses its instance have distinct signatures in the first denoising steps: in omission a mention's attention decays on its own, and in fusion two mentions keep their attention but are never separated. Intervening on joint attention shows that early attention to a mention causally determines, through the early image layout, whether its attribute is realized, but not whether its instance exists. A correction derived from this mechanism, **EMAR**, restores missing color bindings and raises counting and combined accuracy across prompt structures; consistent with the mechanism, its color gains are specific to the attended mention and its count gains are not.

### AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing and to generate synthetic datasets (detailed further in the paper). We have not used generative AI tools for ideation, literature review, or peer review, and all other disclosure categories are not applicable to this work. We have thoroughly reviewed all AI-assisted work—verifying generated prose for factual accuracy and validating synthetic data for correctness—and take full responsibility for all final content, claims, and artifacts in this work.

### REPRODUCIBILITY STATEMENT

Prompt configurations, seeds, sampling settings (Section 3), the evaluator prompt (Appendix A.3), and per-image evaluator outputs are included in the supplementary material.

## REFERENCES

Kanghyun Baek, Jaihyun Lew, Chaehun Shin, Jungbeom Lee, and Sungroh Yoon. Diagnosing and correcting concept omission in multimodal diffusion transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uvSU4dNYv3.

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

Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13242–13251. IEEE, 2025.

Dongnam Byun, Jungwon Park, Jungmin Ko, Changin Choi, and Wonjong Rhee. Dos: Directional object separation in text embeddings for multi-object image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2490–2497, 2026.

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.

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pp. 432–448. Springer, 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.

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.

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

Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. *arXiv preprint arXiv:2509.18096*, 2025.

Kwanyoung Kim and Byeongsu Sim. Pladis: Pushing the limits of attention in diffusion models at inference time by leveraging sparsity. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16238–16248. IEEE, 2025.

Seungwoo Kim, Junyoung Park, and Jiwon Kim. Divide and bind: Disentangling cross-attention maps for multi-subject generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Black Forest Labs. Flux.1-dev. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024. Software/Model.

Binglei Li, Mengping Yang, Zhiyu Tan, Junping Zhang, and Hao Li. TexTailor: Inference-time textual guidance tailoring for multimodal diffusion transformers. In *Computer Vision – ECCV 2026*, pp. 575–593. Springer, 2026.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K Wong. Rethinking cross-modal interaction in multimodal diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5934–5943. IEEE, 2025.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. CONFORM: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9005–9014, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Maria Mihaela Trusca, Wolf Nuyts, Jonathan Thomm, Robert Honig, Thomas Hofmann, Tinne Tuytelaars, and Marie-Francine Moens. Object-attribute binding in text-to-image generation: Evaluation and control. *arXiv preprint arXiv:2404.13766*, 2024.

Tianyi Wei, Dongdong Chen, Yifan Zhou, and Xingang Pan. Enhancing mmdit-based text-to-image models for similar subject generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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 *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=P3b1028tNC`.

Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16712–16722, 2025.

# A APPENDIX

## A.1 MMDIT JOINT ATTENTION

In a dual-stream MMDiT block, text and image hidden states $H_t$ and $H_i$ are projected by modality-specific weights into $Q_m$, $K_m$, $V_m$ ($m \in \{t, i\}$), and the concatenated sequences attend jointly:

$$Q = [Q_t; Q_i], \quad K = [K_t; K_i], \quad V = [V_t; V_i], \quad A = \mathrm{softmax}\Big(QK^\top/\sqrt{d}\Big), \quad O = AV. \quad (2)$$

This joint attention $A$ decomposes into four quadrants by query/key modality: T2T-SA (text-query, text-key), I2I-SA (image-query, image-key), T2I-CA (text-query, image-key), and I2T-CA (image-query, text-key) where SA and CA stand for self-attention and cross-attention, respectively.

## A.2 MODELS AND EVALUATION

We conduct our experiments using the standard configuration for each model. SD3.5-medium: $1024^2$ pixels, 30 steps, CFG 5.0, shift 3.0, DPM++ 2M sampler. FLUX.1-dev: $1024^2$ pixels, 28 steps, guidance 3.5, maximum text length 512. Both use the same prompt configurations and seeds. Each cell has 79 prompt configurations $\times$ 4 seeds (316 images per model), and every grid value uses this matched sample. CIs bootstrap over prompt configurations with seeds kept as clusters.

Qwen2.5-VL-7B-Instruct (Bai et al., 2025) counts each requested base noun and color–noun conjunction. *Counting accuracy*, the primary endpoint, requires every base-noun count to match its target; *combined accuracy* additionally requires every color–noun count to match, with a *leakage-robust* variant ($\geq$) that only requires each conjunction at least as often as requested, since exact matching penalizes color leakage onto uncolored instances. Undercounted images are labeled *fusion* if every requested conjunction is still present and *omission* otherwise; remaining miscounts are *overcount*. Appendix A.3 gives the full scoring rules and worked examples.

## A.3 EVALUATION DETAILS

This appendix describes the scoring of the main suite (Sections 4.1–5.5). Sections 5–6 use a separate outcome protocol with two judges and instance masks (Appendix A.10).

**Targets.** Each prompt is parsed deterministically into a set of *queries* with target counts $c(\cdot)$: one query per distinct base noun $n$, with $c(n)$ the number of slots carrying that noun, and one query per distinct color–noun conjunction $x$ (e.g. "red bicycle"), with $c(x)$ the number of slots carrying that color and noun. Uncolored slots contribute only to their base noun. Table 2 gives examples.

**Questions.** Each query $q$ is asked separately. Qwen2.5-VL-7B-Instruct (bf16, greedy decoding, at most 8 new tokens, images capped at $448^2$ pixels) receives: *"Look carefully at this image. How many distinct, separate instances of "q" are visible in the image? Count only clearly distinguishable individual instances (do not count reflections, drawings-within-the-image, or partial/cut-off duplicates as separate instances unless clearly a distinct individual). Respond with ONLY a single integer, nothing else."* The first integer in the reply is the detected count $d(q)$.

**Counting accuracy.** An image is *count-correct* if $d(n) = c(n)$ for every base noun $n$. Color queries play no role. Counting accuracy is the share of count-correct images and is the primary endpoint throughout.

**Attribute binding.** Binding is scored only on count-correct images and only through the color–noun queries.

- *Combined accuracy (strict)*: count-correct and $d(x) = c(x)$ for every conjunction $x$.
- *Leakage-robust combined accuracy* ($\geq$): count-correct and $d(x) \geq c(x)$ for every $x$. Each requested color–object pair must be present at least as often as requested, but an extra instance in a requested color is not penalized. This matters when a slot is uncolored: in "a red cat, a cat and another red cat" ($c(\text{cat}) = 3$, $c(\text{red cat}) = 2$), an image with three red

cats fails strictly ($d(\text{red cat}) = 3$) although every stated request is met; the uncolored cat has inherited its neighbors' color.

- *Excess-color rate*: among count-correct images, the share with $d(x) > c(x)$ for at least one $x$. Under $A_3$ no slot is uncolored and no instance can legitimately inherit a requested color, so the rate there (20–38%) measures evaluator noise and serves as the baseline for $A_2^-$ (Table 8).

For colorless prompts there are no conjunction queries, and both combined accuracies equal counting accuracy.

**Outcome classes.** Each image receives exactly one label, assigned in this order:

1. *Undercount*: $d(n) < c(n)$ for some base noun. The undercount is a *fusion* if the prompt requests at least two distinct conjunctions and $d(x) \geq c(x)$ for every conjunction; otherwise it is an *omission*.
2. *Overcount*: otherwise, if $d(n) > c(n)$ for some base noun.
3. *Correct count*: otherwise.

A fusion therefore has fewer objects than requested, yet every requested color is found on some object: the colors of the missing instance are carried by the objects that were drawn, for example a two-toned object judged to be both colors. An omission has lost at least one requested color–noun pair together with the instance. The distinction requires at least two independently checkable conjunctions. With none (colorless) or one ("red bicycle" under $A_1$, "red cat" under $A_2^-$ at $N_1$), an undercount cannot be classified and is labeled omission, so fusion is $0\%$ by construction in those conditions (Figure 2).

| Condition | Prompt | Queries: target $c(\cdot)$ |
|---|---|---|
| $N_1 \times A_\varnothing$ | "a bicycle, another bicycle and the other bicycle" | bicycle: 3 |
| $N_1 \times A_1$ | "a red bicycle, another red bicycle and the other red bicycle" | bicycle: 3; red bicycle: 3 |
| $N_1 \times A_2$ | "a red bicycle, another red bicycle and a brown bicycle" | bicycle: 3; red bicycle: 2; brown bicycle: 1 |
| $N_1 \times A_3$ | "a red bicycle, a brown bicycle and a gold bicycle" | bicycle: 3; red, brown, gold bicycle: 1 each |
| $N_1 \times A_2^-$ | "a red cat, a cat and another red cat" | cat: 3; red cat: 2 |
| $N_2 \times A_2$ | "a red bicycle, a red airplane and a brown bicycle" | bicycle: 2; airplane: 1; red bicycle: 1; red airplane: 1; brown bicycle: 1 |
| Numeral | "three bicycles" | bicycle: 3 |

Table 2: **Evaluator queries for example prompts.** Outcomes for the $N_1 \times A_3$ prompt: detected bicycle $= 3$ is a correct count; bicycle $< 3$ with red, brown, and gold bicycle each $\geq 1$ is a fusion; bicycle $< 3$ with any of them $= 0$ is an omission; bicycle $> 3$ is an overcount.

**Vocabulary.** Nouns are 79 COCO object categories and colors are 20 basic color terms (Table 3).

Table 3: COCO Categories (79 Nouns) and Colors (20)

| **Nouns (79 COCO Categories)** | | | | | |
|---|---|---|---|---|---|
| airplane | apple | backpack | banana | baseball bat | baseball glove |
| bear | bed | bench | bicycle | bird | boat |
| book | bottle | bowl | broccoli | bus | cake |
| car | carrot | cat | cell phone | chair | clock |
| couch | cow | cup | dining table | dog | donut |
| elephant | fire hydrant | fork | frisbee | giraffe | hair drier |
| handbag | horse | hot dog | keyboard | kite | knife |
| laptop | microwave | motorcycle | mouse | orange | oven |
| parking meter | pizza | potted plant | refrigerator | remote | sandwich |
| scissors | sheep | sink | skateboard | skis | snowboard |
| spoon | sports ball | stop sign | suitcase | surfboard | teddy bear |
| tennis racket | tie | toaster | toilet | toothbrush | traffic light |
| train | truck | tv | umbrella | vase | wine glass |
| zebra | | | | | |
| **Colors (20)** | | | | | |
| beige | black | blue | brown | cyan | gold |
| gray | green | magenta | maroon | navy | orange |
| pink | purple | red | silver | teal | turquoise |
| white | yellow | | | | |

### A.4 Existing alignment methods on repeated-mention enumeration

We applied three inference-time methods from Section 2 that intervene on MMDiT attention or text conditioning to the $N_1$ prompts of the main suite, with the same prompt configurations, seeds, and evaluator: TACA (Lv et al., 2025), which rescales the temperature of cross-modal attention logits; OSI, the omission correction of Baek et al. (2026), which steers the keys of omission-sensitive attention heads; and Prompt Reinjection (Yao et al., 2026), which reinjects early-block text features into deeper blocks. Figure 5 shows counting accuracy along the attribute axis. None of the methods closes the enumeration gap. For colorless enumeration ($A_\varnothing$), the best method reaches $38.9\%$ (OSI) in SD3.5-medium and $53.2\%$ (TACA) in FLUX.1-dev, against $68.4\%$ and $78.2\%$ for the unmodified models under the numeral form "three Xs". The methods shift the attribute axis up or down without changing its shape. In FLUX.1-dev distinct colors remain the main source of improvement under every method ($A_\varnothing \to A_3$: $+28.5$ baseline, $+29.4$ TACA, $+25.4$ OSI, $+35.7$ Prompt Reinjection), and in SD3.5-medium the axis stays nearly flat ($+5.4$ to $+11.7$). The cue-use asymmetry of Section 4.2 thus persists under all three corrections. These are point estimates on 316 images per cell, without CIs.

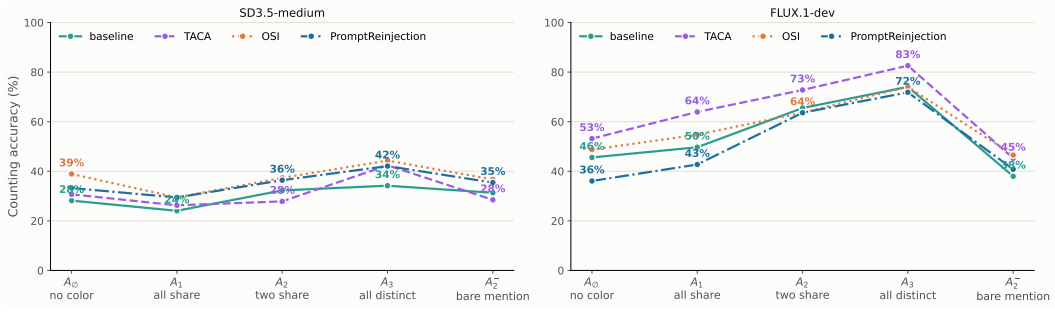
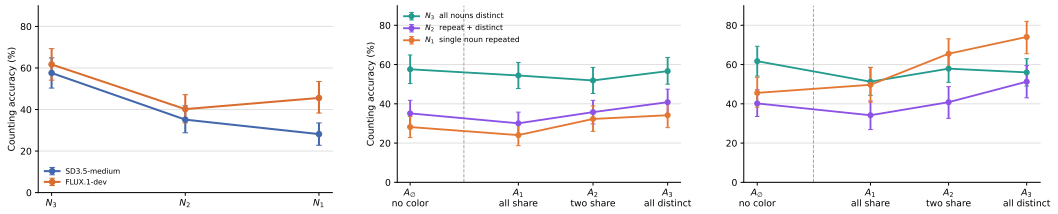



Figure 5: **Counting accuracy of existing alignment methods at** $N_1$ along the attribute axis (79 prompt configurations $\times$ 4 seeds per point). Baseline: unmodified model (Table 6); TACA (Lv et al., 2025); OSI (Baek et al., 2026); Prompt Reinjection (Yao et al., 2026).

### A.5 FULL COUNTING GRID AND ADDITIONAL CUE CONTRASTS

Figure 6 and Tables 4 and 5 give the counting results of Sections 4.1–4.2.




(a) Noun duplication, colorless. (b) Attribute structure, SD3.5-medium. (c) Attribute structure, FLUX.1-dev.

Figure 6: **Counting accuracy along the two axes of the suite.** (a) Colorless enumeration across $N_3/N_2/N_1$. (b–c) Attribute structure at full coverage, with the colorless baseline leftmost. Error bars: 95% prompt-clustered bootstrap CIs.

| | | SD3.5-medium | | | FLUX.1-dev | | |
|---|---|---|---|---|---|---|---|
| Structure | Inst. | Enum. | Num. | Δ [95% CI] | Enum. | Num. | Δ [95% CI] |
| Pure repetition | 2 | 71.2 | 83.5 | +12.3 [6.3, 18.7] | 71.2 | 87.7 | +16.5 [9.8, 23.4] |
| Pure repetition | 3 | 28.2 | 68.4 | +40.2 [33.2, 46.8] | 45.6 | 78.2 | +32.6 [24.4, 40.8] |
| Pure repetition | 4 | 16.8 | 55.7 | +38.9 [32.0, 45.9] | 9.5 | 73.1 | +63.6 [56.0, 70.6] |
| Mixed (X, Y, X) | 3 | 35.8 | 37.0 | +1.3 [−7.3, 10.1] | 46.2 | 59.2 | +13.0 [5.1, 20.9] |
| Mixed (X, Y, X, X) | 4 | 20.3 | 17.4 | −2.8 [−8.9, 2.8] | 27.5 | 38.3 | +10.8 [4.1, 17.1] |

Table 4: **Numeral-matched enumeration penalty** (colorless counting accuracy, %; Δ = numeral − enumeration; paired prompt-clustered bootstrap).

| Cue | Contrast at $N_1$ | SD3.5-medium | FLUX.1-dev |
|---|---|---|---|
| Determiners | a, a, a → a, another, another | +12.3 [6.3, 18.0] | +32.3 [25.6, 39.2] |
| Distinct colors | a, a, a → $A_3$ | +10.1 [3.8, 16.5] | +55.1 [46.5, 63.0] |
| Shared-color control | $A_1 \to A_2$ | +8.2 [1.9, 14.6] | +15.8 [9.5, 22.5] |
| | $A_2 \to A_3$ | +1.9 [−4.1, 8.2] | +8.5 [2.5, 15.2] |

Table 5: **Mention-distinguishing cues**: change in counting accuracy at $N_1$ (points [95% CI]). Determiners are those of "a X, a X and a X"; the shared-color control varies only the number of distinct colors.

**Colorless baseline.** Relative to the colorless baseline $A_\varnothing$, which uses contrastive determiners, three distinct colors add 28.5 [19.9, 37.0] points at $N_1$ for FLUX.1-dev (45.6% → 74.1%) and 6.0 [0.6, 11.4] for SD3.5-medium (28.2% → 34.2%). Because $A_3$ uses "a" throughout, this contrast understates the color effect; the main text therefore reports the determiner-matched contrast.

**Cues at $N_2$.** The shared-color control shows the same ordering as at $N_1$. From $A_1$ to $A_2$ and then to $A_3$, FLUX.1-dev gains +6.6 [−0.9, 13.9] and +10.4 [3.5, 17.4] points, and SD3.5-medium +5.7 [0.0, 11.4] and +5.1 [−1.9, 11.7]; only FLUX.1-dev's second step is significant. No determiner-matched bare enumeration exists at $N_2$. Against the colorless baseline, distinct colors add +11.1 [3.2, 19.0] points for FLUX.1-dev and +5.7 [−2.8, 13.9] (not significant) for SD3.5-medium. SD3.5-medium's $N_2$ difficulty is shared with the numeral form (Section 4.1), so its $N_2$ cue effects are not interpreted.

**Two instances.** With the determiner held fixed ("a red X and a brown X" vs. "a red X and a red X"), distinct colors add +45.9 [38.3, 53.2] points for FLUX.1-dev and +17.7 [11.7, 24.1] for SD3.5-medium.

| | $A_\varnothing$ | $A_1$ | $A_2$ | $A_3$ | $A_2^-$ |
|---|---|---|---|---|---|
| *SD3.5-medium* | | | | | |
| $N_3$ | 57.6 | 54.4 | 51.9 | 56.6 | 52.8 |
| $N_2$ | 35.1 | 30.1 | 35.8 | 40.8 | 23.4 |
| $N_1$ | 28.2 | 24.1 | 32.3 | 34.2 | 31.3 |
| *FLUX.1-dev* | | | | | |
| $N_3$ | 61.7 | 51.3 | 57.9 | 56.0 | 56.0 |
| $N_2$ | 40.2 | 34.2 | 40.8 | 51.3 | 13.0 |
| $N_1$ | 45.6 | 49.7 | 65.5 | 74.1 | 38.0 |

Table 6: **Counting accuracy (%), full $N \times A$ grid.** Combined accuracy is in Table 7.

**Cues at $N_3$.** Against the colorless baseline, distinct colors change counting by $-0.9$ [$-9.8$, 7.3] points for SD3.5-medium and $-5.7$ [$-13.0$, 2.2] for FLUX.1-dev, and neither step of the shared-color control is significant ($A_1 \to A_2$: $-2.5$ [$-9.2$, 4.7] and $+6.6$ [0.0, 13.6]; $A_2 \to A_3$: $+4.7$ [$-2.2$, 11.7] and $-1.9$ [$-7.9$, 4.4]). A further condition outside the main grid colors only one of three distinct nouns ("a sports ball, a red fork and a tennis racket"); it is also indistinguishable from colorless ($-0.3$ [$-8.5$, 8.2] and $+3.8$ [$-4.4$, 11.7]).

### A.6 Combined accuracy, leakage-robust scoring, and the bare-mention condition

| | $A_\varnothing$ | $A_1$ | $A_2$ | $A_3$ | $A_2^-$ |
|---|---|---|---|---|---|
| *SD3.5-medium, combined accuracy: strict / $\geq$ (%)* | | | | | |
| $N_3$ | 57.6 | 48.7 / 50.6 | 40.8 / 41.1 | 43.7 / 44.3 | 48.1 / 48.4 |
| $N_2$ | 35.1 | 22.5 / 24.1 | 17.1 / 25.6 | 24.4 / 32.6 | 7.6 / 20.3 |
| $N_1$ | 28.2 | 17.4 / 17.4 | 13.9 / 24.1 | 17.7 / 30.7 | 6.3 / 26.9 |
| *FLUX.1-dev, combined accuracy: strict / $\geq$ (%)* | | | | | |
| $N_3$ | 61.7 | 45.9 / 47.2 | 51.9 / 52.5 | 52.2 / 52.5 | 51.6 / 52.2 |
| $N_2$ | 40.2 | 30.7 / 31.3 | 26.6 / 35.8 | 36.4 / 47.5 | 3.8 / 11.7 |
| $N_1$ | 45.6 | 41.8 / 41.8 | 31.6 / 48.7 | 50.6 / 71.8 | 10.4 / 31.6 |

Table 7: **Combined accuracy, strict and leakage-robust ($\geq$)** (matched sample). For colorless prompts both equal counting accuracy.

**Bare-mention condition.** Table 8 and Figure 7 separate instance loss from attribute inheritance under $A_2^-$. Combined accuracy there is the lowest in the grid ($N_2$: 7.6% / 3.8%; $N_1$: 6.3% / 10.4% for SD3.5-medium / FLUX.1-dev). At $N_2$ counting falls significantly below the colorless baseline in both models, to 23.4% and 13.0%; the one-word minimal pair of Section 4.3 costs 12.4 points (CI $[-18.4, -6.3]$) for SD3.5-medium and 27.8 (CI $[-35.8, -19.6]$) for FLUX.1-dev. At $N_1$ counting is indistinguishable from the colorless baseline in SD3.5-medium and at most modestly lower in FLUX.1-dev, and leakage-robust combined accuracy recovers to 26.9% / 31.6%.

**The partial-sharing dip is metric-induced.** Strict combined accuracy dips at $A_2$, between $A_1$ and $A_3$. Under leakage-robust scoring the attribute axis is monotone at $N_1$ in both models (SD3.5-medium $17.4 \to 24.1 \to 30.7$; FLUX.1-dev $41.8 \to 48.7 \to 71.8$), and counting accuracy shows no dip either (Table 6). At $N_2$ the dip likewise disappears under leakage-robust scoring; it persists only for SD3.5-medium at $N_3$, where the nouns are already distinct. The same definitional coupling largely determines the counting–binding correlation $\phi$ across attribute conditions (Appendix A.14).

### A.7 Text-side occurrence differentiation

Occurrence differentiation ($1 - \overline{\cos}$ over the pairwise cosine similarities of the three occurrence hidden states) is higher under $A_3$ than under $A_\varnothing$ at each of the first four joint blocks in both models (SD3.5-medium $d = 1.00$–$1.30$; FLUX.1-dev $d = 1.50$–$3.09$; Figure 8). Differences in the

| | Model | $\Delta$ vs. $A_\varnothing$ [95% CI] | Excess-color (base.) |
|---|---|---|---|
| $N_2$ | SD3.5-medium | $-11.7$ [$-19.3$, $-4.4$] | 54.1 (20.2) |
| | FLUX.1-dev | $-27.2$ [$-35.1$, $-19.0$] | 61.0 (21.6) |
| $N_1$ | SD3.5-medium | $+3.1$ [$-2.9$, 9.9] | 65.7 (38.0) |
| | FLUX.1-dev | $-7.6$ [$-15.4$, 0.2] | 55.8 (28.6) |

Table 8: **The bare-mention condition** $A_2^-$ (%). $N_2$: "a red bicycle, a red airplane and a bicycle"; $N_1$: "a red cat, a cat and another red cat". $\Delta$: counting accuracy. Excess-color rate with the $A_3$ baseline in parentheses.

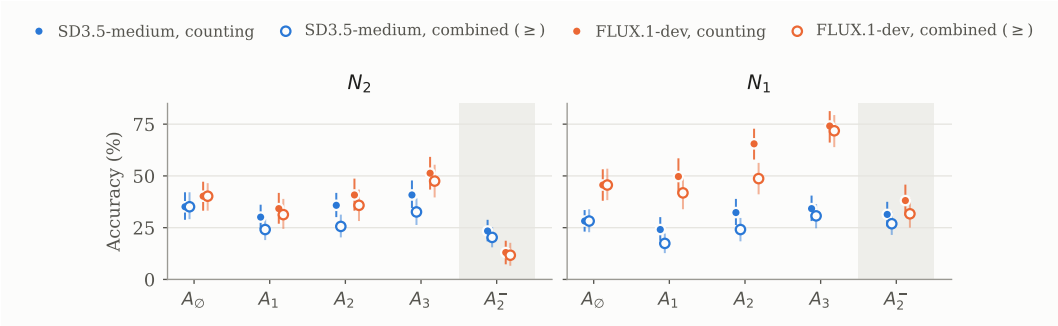



Figure 7: **Counting accuracy (filled) and leakage-robust combined accuracy (hollow) along the attribute axis**, with the bare-mention condition $A_2^-$ shaded. Error bars: 95% prompt-clustered bootstrap CIs.

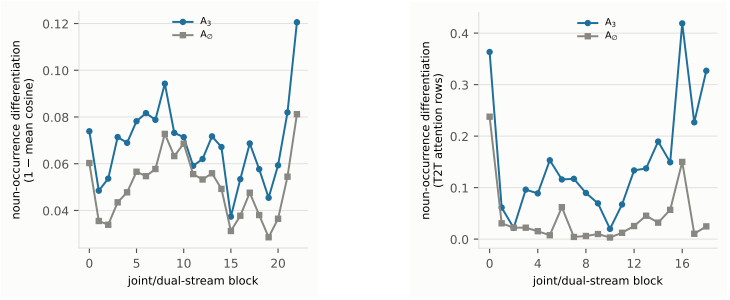
same statistic already exist in the frozen encoder outputs, and their direction depends on the branch (Table 9). SD3.5-medium's CLIP-L output is more differentiated under $A_3$ ($d = +2.40$). The T5 output of both models is *less* differentiated ($d = -0.87$ and $-1.27$), consistent with the distinct determiners of the colorless prompts. Text-side differentiation therefore reflects contextualization by the surrounding words. It shows that both models register the attribute distinction, but it does not distinguish the models and is not used for any claim about them.

| Model | Representation | $A_\varnothing$ | $A_3$ | $d$ | $p$ |
|---|---|---|---|---|---|
| SD3.5-medium | CLIP-L output | 0.162 | 0.353 | $+2.40$ | $1.4 \times 10^{-4}$ |
| | CLIP-G output | 0.525 | 0.545 | $+0.31$ | 0.48 |
| | T5-XXL output | 0.358 | 0.322 | $-0.87$ | 0.056 |
| | Model-side projection | 0.179 | 0.200 | $+0.69$ | 0.12 |
| FLUX.1-dev | T5-XXL output | 0.352 | 0.309 | $-1.27$ | 0.013 |
| | Model-side projection | 0.279 | 0.287 | $+0.30$ | 0.53 |

Table 9: **Occurrence differentiation** ($1 - \overline{\cos}$) **before the joint transformer**, computed at each occurrence's own noun tokens within each encoder branch, before SD3.5-medium's zero-pad and concatenation. SD3.5-medium $n = 11/11$; FLUX.1-dev $n = 8/11$; Welch $t$-tests.

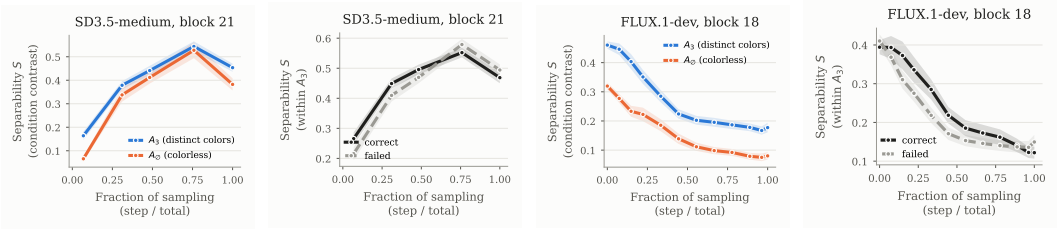
### A.8 Step-resolved separability at every captured block

With Bonferroni correction over all cells (FLUX.1-dev 84, SD3.5-medium 25), the $A_3 > A_\varnothing$ separability gain holds at every captured step at FLUX.1-dev blocks 3, 9, and 12. At block 18 it holds at every step except steps 9 and 12. At block 6 it holds only at steps 0–6 and 27. It is absent at blocks 0 and 15. In SD3.5-medium it holds only at step 2 at blocks 9–21, and at every step at block 3. The within-$A_3$ correct-vs.-failed difference survives correction only at FLUX.1-dev block 18, steps 4–9. At the first captured step the condition gain is 0.460 vs. 0.320 at FLUX.1-dev block 18, step 0 ($p = 5 \times 10^{-11}$), and significant at all five SD3.5-medium blocks at step 2 ($p \leq 2 \times 10^{-4}$). SD3.5-medium's block-3 gain lies at near-zero absolute separability ($S < 0.06$). The FLUX.1-dev outcome contrast at block 18 is 0.515 vs. 0.348 ($d = 1.35$), 0.462 vs. 0.296 ($d = 1.51$), and 0.376




(a) SD3.5-medium. (b) FLUX.1-dev.

Figure 8: **Text-side occurrence differentiation inside the joint transformer**, $A_{\varnothing}$ vs. $A_3$, means over 36 examples.




(a) SD3.5-medium, condition contrast. (b) SD3.5-medium, outcome contrast. (c) FLUX.1-dev, condition contrast. (d) FLUX.1-dev, outcome contrast.

Figure 9: **I2T slot separability at the deepest captured block** (SD3.5-medium block 21, FLUX.1-dev block 18; $N_1$). (a, c) Condition contrast: $A_3$ vs. $A_{\varnothing}$. (b, d) Outcome contrast: correct vs. failed generations within $A_3$. Mean ± SEM; all blocks in Appendix Figure 10.

vs. 0.239 ($d = 1.38$) at steps 4, 6, and 9 (all $p < 3 \times 10^{-4}$). Prompt-level contextualization is constant within $A_3$, so it cannot produce this within-condition association.

### A.9 Failure-type signatures with multiplicity control

Within $N_1 \times A_3$, we compared fusion and omission with correct generations on separability, minimum occurrence attention mass ($M_{\min}$), and own noun–color alignment at every captured block (step-averaged). After Holm correction within each model, one signature survives: lower own noun–color alignment for fusion in FLUX.1-dev at block 18 ($d = -2.02$, $p_{\text{Holm}} = 0.005$; Table 10). No signature survives for SD3.5-medium, and none for omission in either model. With 12 generations per category these contrasts have limited power, so the result is a null result for whole-generation summaries rather than evidence that omission or fusion leaves no internal trace; Sections 5.1 and 5.5 find such traces with per-mention and per-pair measures. Illustrative maps are in Figure 11.

| Model | Contrast | Metric | Block | $d$ | $p$ | $p_{\text{Holm}}$ |
|---|---|---|---|---|---|---|
| FLUX.1-dev | fusion vs. correct | own noun–color alignment | 18 | −2.02 | 0.0001 | **0.005** |
| | omission vs. correct | own noun–color alignment | 18 | −1.23 | 0.009 | 0.35 |
| | fusion vs. correct | separability | 18 | −1.14 | 0.012 | 0.46 |
| SD3.5-medium | fusion vs. correct | $M_{\min}$ | 21 | −1.00 | 0.030 | 0.90 |
| | fusion vs. correct | $M_{\min}$ | 19 | −0.91 | 0.043 | 1.00 |

Table 10: **Strongest failure-type signatures**, $N_1 \times A_3$, 12 items per category (step-averaged; Welch $t$). Holm correction is over all metric × block × contrast tests within a model.

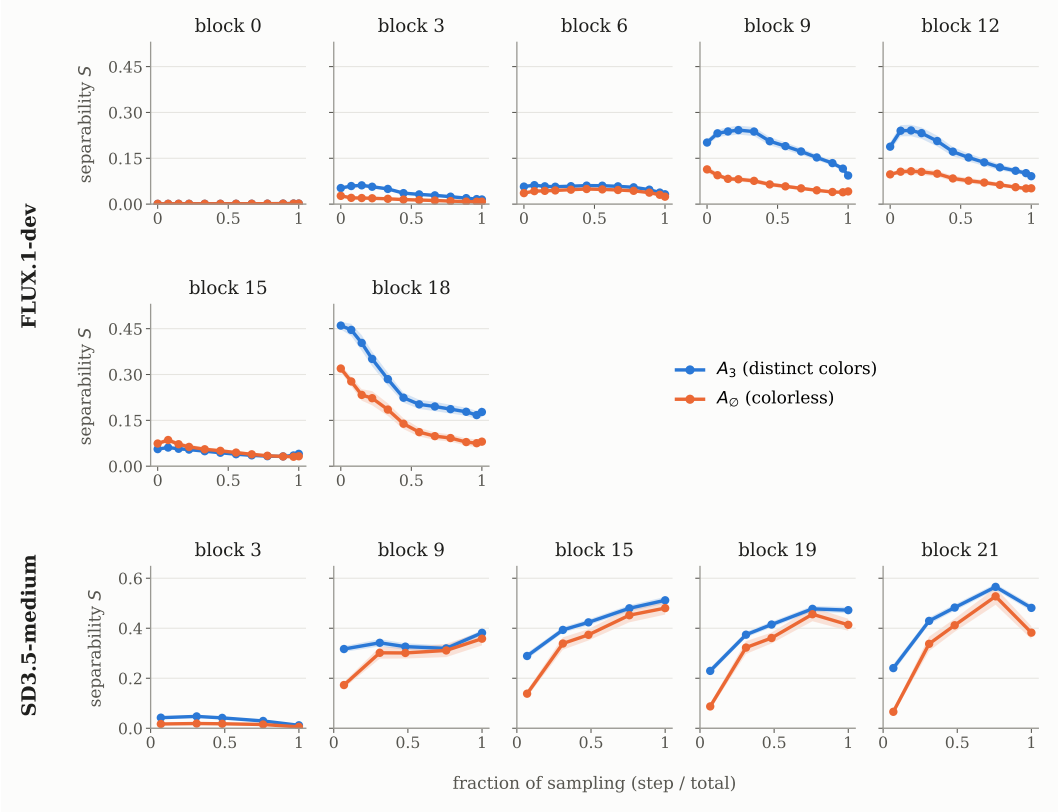



Figure 10: **I2T slot separability at every captured block** ($N_1$; $A_3$ vs. $A_\varnothing$, mean ± SEM over 36 vs. 24 generations). Triangles mark steps where the $A_3 > A_\varnothing$ gain survives Bonferroni correction over all captured cells of that model. The $y$-axis scale differs by panel.

### A.10 Causal analysis and rendering correction: samples, definitions, and outcomes

**Generations.** All generations in Sections 5–6 use SD3.5-medium with the main-study configuration and encode text on the same device and at the same precision as the evaluated images. We verified that they reproduce the evaluated image (mean absolute pixel difference ≈ 1/255). Every manipulated generation is paired with the unmodified one (same prompt, seed, and initial noise), and LPIPS between the two measures the size of the image change.

**Samples.** *Observation*: 23 single-omission cases from $N_1 \times A_3$ (4 seeds × 79 configurations), with correct generations of the same prompts as matched references. *Interventions and correction*: the 19 of these without ambiguity under manual inspection (four were excluded: one judge error and three unclear cases). *Balanced replication*: 648 images from 6 nouns and color triplets with 0–3 achromatic colors, each triplet rotated through all slot orders. *Distinct nouns*: the same colors, orders, and seeds with three distinct nouns; "kite" is excluded because SD3.5-medium renders "black kite" and "red kite" as the bird.

**Attention coverage.** For mention $k$ with CLIP token set $T_k$ (color and noun), let $m_k(i) = \sum_{j \in T_k} A_{ij}$ for image token $i$ (head-averaged image-to-text attention), and $P(i,k) = m_k(i) / \sum_{k'} m_{k'}(i)$. Image token $i$ attends to object tokens if its total object-token attention exceeds the uniform per-token share, and attends selectively to $\arg\max_k P(i,k)$ if, in addition, the margin between its largest and second-largest $P$ exceeds the 75th percentile of a null that permutes each mention's attention across image tokens. The attention coverage of $k$ is the margin-weighted fraction of the 4096 image tokens attending selectively to $k$; Figure 13b follows each such token




(a) FLUX.1-dev, omission (block 18, step 27 of 28): all three noun maps fall on the single rendered bicycle.




(b) FLUX.1-dev, fusion (block 18, step 27 of 28): the noun maps spread over the one rendered bus rather than over separate objects.




(c) SD3.5-medium, omission (block 19, step 22 of 30): “navy” and “blue” fall on the two rendered signs; the omitted “gray” does not localize on either.




(d) SD3.5-medium, fusion (block 19, step 22 of 30): “orange” and “black” both fall on the orange car, “silver” on the other.

Figure 11: **Illustrative I2T attention maps for four $N_1 \times A_3$ failures**, one per model and failure type. Each row overlays the maps of the three noun occurrences and their color tokens on the generated image; maps are normalized separately, so they show where each token attends, not relative magnitude. †: omitted slot. (b) and (c) are closest to their category’s median noun-map separability; (a) and (d) are illustrative.

from step 0 to step 6. Coverage is averaged over joint blocks 15 and 19. Attention to T5 tokens shows no mention-selective coverage, so all analyses use the CLIP stream.

**Outcome labels.** For each generation, Qwen2.5-VL and Qwen2-VL count the base noun and each color–noun pair, and instance masks are assigned to their nearest requested color (CIEDE2000). Color presence uses the consensus of the two VLM judges ($\kappa = 0.75$); the segmentation-based color check serves only to locate instances. The primary outcomes are whether the missing color is restored, whether the count of the repeated noun is correct, and whether a correct new instance appears. Each manipulated image is classified against its unmodified counterpart as: new object in the missing color (a new instance, IoU $< 0.3$ with every original instance, carries the missing color), existing object recolored, color off any object, or unchanged/other. A *correct new instance* additionally requires that both other requested colors and the layout are kept.

**Intervention conditions.** All add $b$ to pre-softmax image-to-text attention logits (Eq. 1) in joint blocks 9–21, conditional branch only, unless noted. *Target* ($b = 2$, steps 1–5): the omitted mention's CLIP tokens (targeted attention bias); a rendered mention's CLIP tokens; random non-object CLIP tokens; the omitted mention's T5 tokens; every text token (neutral control, LPIPS $0.005$); function-word tokens (untargeted perturbation of matched size, LPIPS $0.30$); the omitted mention's CLIP tokens at steps 15–20. *Strength*: $b \in \{0.25, 0.5, 1, 1.5, 2, 3\}$, restoring the color in $4, 10, 14, 17, 16, 19$ of 19 cases (logistic midpoint $b \approx 0.7$); count correction stays at or below $7/19$, and the other two colors are kept in only $13/19$ cases at $b = 3$. *Heads* ($b = 3$, blocks 9/15/19/21): CLIP attention heads were clustered ($k$-means, $k = 4$, bootstrap ARI $0.84$) on features measured in unmanipulated generations (color and noun selectivity, concentration, spatial localization, and the attention deficit of omitted mentions); the bias is restricted to the color-selective cluster, to a random head set of equal size, or applied to all heads. *Timing*: $b = 3$ at a single step $t \in \{0, \ldots, 6, 8, 10, 12, 15, 20\}$. *Suppression* ($b = -2$ or $-3$) in 30 correct generations: the targeted color is lost in $\leq 4/30$ cases for every head set, against $0/30$ for late or neutral suppression. The *rendering correction* EMAR uses $T = \bigcup_k T_k$ with $b = 2$ at steps 1–5.

**Further intervention results.** Color restoration rises smoothly with dose, from $1/19$ at $b = 0$ to $19/19$ at $b = 3$ (half restored at $b \approx 0.7$), and is concentrated in a cluster of color-selective heads ($13/19$ vs. $4/19$ for a random head set of equal size, $p = 0.012$). Suppressing attention to the most strongly attended rendered mention neither removes it ($0/23$) nor restores the omitted color more than an untargeted perturbation, and suppressing attention in correct generations rarely removes color either ($\leq 4/30$). In FLUX.1-dev, raising the attention mass of the least-attended occurrence corrected none of nine failures at any strength, so raw attention mass is not the operative variable.

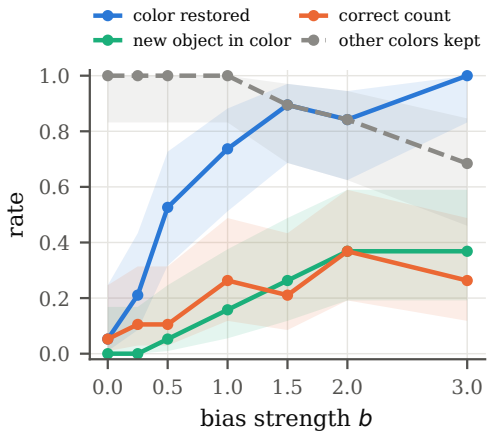
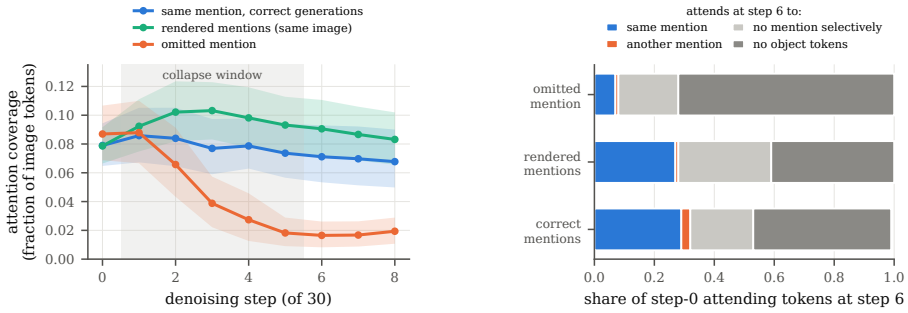



Figure 12: **Effect of bias strength** $b$ for the targeted attention bias (steps 1–5, joint blocks 9–21; 19 single-omission cases; 95% Wilson CIs). Mean LPIPS to the unmodified image is $0.01$, $0.07$, $0.15$, $0.22$, $0.32$, $0.38$, and $0.47$ at $b = 0, 0.25, 0.5, 1, 1.5, 2, 3$.

**Mediation conditions.** Each is run alongside its own unmodified generation. *Coarse layout fixing*: the Gaussian low-pass component ($\sigma = 32$ px) of the predicted clean image $\hat{x}_0$ is replaced by that of the unmodified generation at steps 1–5 (or 1–8), leaving the high-frequency residual free. Layout fixing alone changes nothing ($0/19$ colors, LPIPS $0.001$). Under layout fixing, the bias still raises attention to the omitted mention by the same amount at steps 1–2 and by half to two thirds as much at steps 3–5; fixing the layout through step 8 gives $5/19$ colors and $1/19$ counts. After the bias ends, the elevated attention persists in unconstrained generations ($+0.04$ at steps 6–8) but returns to baseline when the layout was fixed. *Direct color edit*: the mean of $\hat{x}_0$ is moved toward the missing color inside a disk at the omitted object's step-0 location at steps 1–2. The edit is large (missing-color likelihood $+0.37$) but fades to $+0.09$ within two steps and barely raises attention to the omitted mention ($+0.008$ vs. a wrong-color edit, CI $[+0.001, +0.016]$). Its outcomes do not differ from a wrong-color or background edit ($5/19$ vs. $6/19$ and $4/19$, both $p = 1.0$) and are below the attention




(a) Attention coverage over denoising steps.

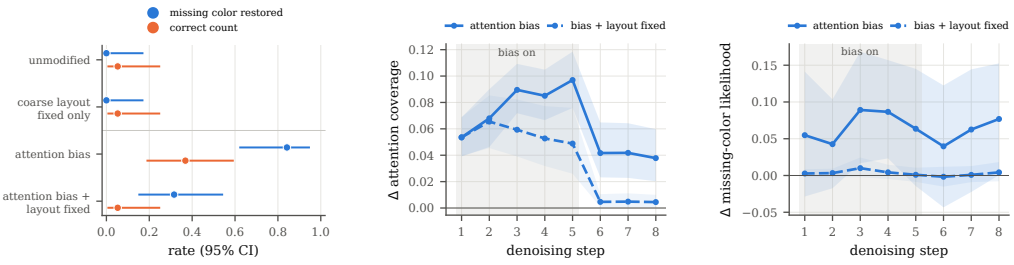
(b) Step-0 attention at step 6.

Figure 13: **Observation (association only)**, single-omission cases. (a) Attention coverage of each mention's CLIP color tokens; the reference is the same mention in correct generations of the same prompt (mean, 95% bootstrap CI). (b) Where the image tokens attending selectively to a mention at step 0 attend at step 6.

bias ($p = 0.003$), whose change to the image is largest at step 3 ($+0.089$ [$+0.014$, $+0.167$]) and still positive at step 8.

| Condition | Color restored | New object in color | Correct count | Correct new instance | LPIPS ↓ |
|---|---|---|---|---|---|
| Neutral control | 1/19 | 0/19 | 1/19 | 0/19 | 0.01 |
| Targeted attention bias | 16/19 | 7/19 | 7/19 | 5/19 | 0.38 |
| toward a rendered mention | 5/19 | 3/19 | 3/19 | 3/19 | 0.29 |
| toward random CLIP tokens | 5/19 | 0/19 | 2/19 | 0/19 | 0.15 |
| toward omitted mention's T5 tokens | 1/19 | 0/19 | 1/19 | 0/19 | 0.01 |
| Untargeted perturbation, matched size | 9/19 | 3/19 | 9/19 | 2/19 | 0.30 |
| Targeted bias, coarse layout fixed (steps 1–5) | 6/19 | 1/19 | 1/19 | 1/19 | 0.09 |
| Targeted bias, coarse layout fixed (steps 1–8) | 5/19 | 0/19 | 1/19 | 0/19 | 0.07 |
| EMAR (rendering correction) | 17/19 | 7/19 | 8/19 | 7/19 | 0.39 |

Table 11: **Intervention, mediation, and correction outcomes** on the 19 single-omission cases. Paired exact McNemar tests are reported in Sections 5.2–5.3.

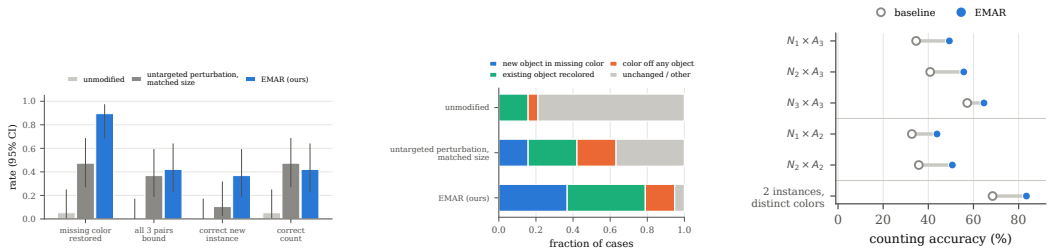



(a) Outcomes.

(b) Attention to the omitted mention.

(c) Missing color in the predicted image.

Figure 14: **Mediation**: the targeted bias with and without fixing the coarse layout of the predicted image $\hat{x}_0$. (a) Outcomes under each condition. (b) Change in attention to the omitted mention and (c) in the missing color's likelihood at the omitted object's location, relative to the unmodified generation; shading marks the bias window (mean, 95% CI).




(a) Outcomes before vs. after in the pilot sets.

(b) How the missing color returns.

(c) Gated correction, six other subsets.

Figure 15: **EMAR before vs. after.** (a–b) The single-omission cases: (a) outcomes without correction, under an untargeted perturbation of matched size, and with EMAR (95% CIs); (b) how the missing color is realized, mainly a new object or a recolored one. (c) Counting accuracy on six further prompt structures, comparing the baseline against EMAR applied only to images the baseline undercounts (baseline kept otherwise). Further prompt structures are evaluated in Appendix A.11.

### A.11 Generalization of the rendering correction

Section 6 evaluates EMAR ($b = 3$) beyond the 19 diagnostic cases on twelve prompt structures: the $A_3$ and $A_2$ cells at every noun level, the $A_1$ cells at $N_1$ and $N_3$, the three $A_\varnothing$ cells, and the two-instance prompt (Table 12). EMAR is applied to the images the unmodified model undercounts and the unmodified generation is kept otherwise, so every outcome is compared on the same images, scored as in Appendix A.3.

Counting accuracy rises in eleven of the twelve structures, from $+5.0$ points ($N_2 \times A_\varnothing$) to $+15.0$ ($N_1 \times A_1$ and the two-instance prompt), and is unchanged at $N_3 \times A_\varnothing$, where undercounts are rare. Omission falls in every structure, and fusion falls wherever it is common (from $-4.2$ to $-12.0$ points). Part of the recovered instances become overcounts, which rise in every structure. Binding among count-correct images stays within about two points of the baseline except at $N_1 \times A_2$ ($-6.0$) and $N_1 \times A_1$ ($+6.9$, not significant). The gain is largest where repeated mentions make undercounting common and smallest at $N_3$, where the nouns already distinguish the mentions, consistent with the mechanism identified in Section 5.

### A.12 Fusion analysis: definitions and tables

All quantities in Section 5.5 use the position-balanced replication (6 nouns, 4 seeds, color triplets with 0–3 achromatic colors rotated through all slot orders), without "kite". Attention maps are head-averaged image-to-text attention to a mention's CLIP color and noun tokens, averaged over joint blocks 15 and 19. The merged pair and its weaker mention are fixed at step 5 and followed over steps 0–8. The rendered-color measure takes the top $15\%$ of image tokens by the merged pair's summed attention map, and within this region counts pixels of the $128^2$ preview of $\hat{x}_0$ whose nearest requested color is the weaker mention's (CIELAB distance below 30); it is reported for chromatic weaker colors. Confidence intervals resample prompts. Table 13 gives the step-resolved separation of the outcomes, Table 14 which pair of mentions merges, and Table 15 how EMAR moves each outcome on the main-study prompts.

### A.13 A preliminary check on FLUX.1-dev

The causal mechanism of Section 5 is established only for SD3.5-medium. As a preliminary check of whether it transfers to FLUX.1-dev, we measured, on 9 single-omission cases sampled from FLUX.1-dev's own generations, (i) the same per-mention attention coverage as Figure 13, restricted to FLUX.1-dev's dual-stream blocks 0–18, and (ii) the total (unfiltered) image-to-text attention mass – the fraction of each image token's attention spent on all text positions, regardless of mention – compared against the same statistic in SD3.5-medium on a matched 9-case set (same prompts and seeds).

| Prompt | | Count | Comb. ($\geq$) | Bind \| count | Fusion | Omission | Overcount |
|---|---|---|---|---|---|---|---|
| $N_3 \times A_3$ | baseline | 57.2 | 45.1 | 78.8 | 3.4 | 18.5 | 20.9 |
| | EMAR | **64.6** | **50.5** | 78.1 | 2.0 | **4.7** | **28.6** |
| $N_2 \times A_3$ | baseline | 40.8 | 32.6 | 79.8 | 22.8 | 25.6 | 10.8 |
| | EMAR | **55.7** | **45.3** | 81.2 | **16.1** | **8.9** | **19.3** |
| $N_1 \times A_3$ | baseline | 34.6 | 30.8 | 89.1 | 48.3 | 11.3 | 5.8 |
| | EMAR | **49.3** | **44.5** | 90.3 | **36.3** | **4.1** | **10.3** |
| $N_3 \times A_2$ † | baseline | 53.8 | 43.8 | 81.4 | 2.5 | 22.5 | 21.2 |
| | EMAR | **65.0** | **52.5** | 80.8 | 0.0 | **7.5** | **27.5** |
| $N_2 \times A_2$ | baseline | 35.8 | 25.6 | 71.7 | 28.2 | 23.4 | 12.7 |
| | EMAR | **50.6** | **37.3** | 73.8 | **21.8** | **9.2** | **18.4** |
| $N_1 \times A_2$ | baseline | 32.7 | 24.6 | 75.3 | 16.2 | 44.2 | 6.9 |
| | EMAR | **43.8** | **30.4** | **69.3** | **11.9** | **34.6** | **9.6** |
| $N_3 \times A_1$ † | baseline | 57.5 | 56.2 | 97.8 | 2.5 | 23.8 | 16.2 |
| | EMAR | **65.0** | **63.8** | 98.1 | 1.2 | **11.2** | 22.5 |
| $N_1 \times A_1$ † | baseline | 23.8 | 17.5 | 73.7 | 0.0 | 66.2 | 10.0 |
| | EMAR | **38.8** | **31.2** | 80.6 | 0.0 | **42.5** | **18.8** |
| $N_3 \times A_\varnothing$ † | baseline | 67.5 | 67.5 | – | – | 3.8 | 28.8 |
| | EMAR | 67.5 | 67.5 | – | – | 1.2 | 31.2 |
| $N_2 \times A_\varnothing$ † | baseline | 36.2 | 36.2 | – | – | 48.8 | 15.0 |
| | EMAR | 41.2 | 41.2 | – | – | **35.0** | **23.8** |
| $N_1 \times A_\varnothing$ † | baseline | 23.8 | 23.8 | – | – | 58.8 | 17.5 |
| | EMAR | **33.8** | **33.8** | – | – | **42.5** | **23.8** |
| two instances | baseline | 68.5 | 64.6 | 94.4 | 20.8 | 7.3 | 3.5 |
| | EMAR | **83.5** | **79.6** | **95.4** | **9.2** | **1.9** | **5.4** |

Table 12: **EMAR across prompt structures** (SD3.5-medium, %). EMAR is applied where the unmodified model undercounts, keeping the unmodified generation otherwise. Outcomes follow Appendix A.3; –: undefined without requested colors. †20 prompt configurations $\times$ 4 seeds. Bold: paired 95% CI excludes zero.

| Signal | Contrast | $t$=0 | 1 | 2 | 3 | 5 | 8 |
|---|---|---|---|---|---|---|---|
| pairwise map overlap | fusion vs. correct | 0.60 [0.54, 0.65] | 0.70 [0.65, 0.75] | 0.84 [0.80, 0.88] | 0.88 [0.85, 0.92] | 0.90 [0.87, 0.93] | 0.89 [0.86, 0.92] |
| pairwise map overlap | omission vs. correct | 0.63 [0.55, 0.71] | 0.69 [0.62, 0.76] | 0.79 [0.72, 0.85] | 0.82 [0.75, 0.88] | 0.84 [0.77, 0.90] | 0.88 [0.82, 0.93] |
| weaker-mention coverage | fusion vs. omission | 0.48 [0.40, 0.56] | 0.50 [0.41, 0.59] | 0.52 [0.44, 0.60] | 0.56 [0.48, 0.64] | 0.66 [0.57, 0.74] | 0.69 [0.61, 0.76] |
| weaker-mention color–noun $r$ | fusion vs. omission | 0.61 [0.53, 0.70] | 0.61 [0.52, 0.68] | 0.61 [0.52, 0.69] | 0.64 [0.56, 0.72] | 0.61 [0.53, 0.69] | 0.58 [0.50, 0.66] |

Table 13: **AUROC by denoising step** $t$ (95% CI).

| Outcome | color distance ($\Delta E_{00}/10$) | achromatic colors in pair | adjacent in prompt |
|---|---|---|---|
| fusion | $-0.38 \pm 0.07$ ($p < 0.001$) | $+0.33 \pm 0.24$ ($p = 0.173$) | $+1.59 \pm 0.26$ ($p < 0.001$) |
| omission | $-0.18 \pm 0.08$ ($p = 0.025$) | $+0.60 \pm 0.36$ ($p = 0.100$) | $+0.94 \pm 0.34$ ($p = 0.005$) |
| correct | $-0.11 \pm 0.04$ ($p = 0.006$) | $+0.37 \pm 0.16$ ($p = 0.022$) | $+1.18 \pm 0.17$ ($p < 0.001$) |

Table 14: **Which pair of mentions merges** (conditional logit, ±SE).

Neither observation supports the SD3.5-medium mechanism as it stands: FLUX.1-dev shows no selective early collapse that distinguishes the omitted mention from rendered ones (panel b), and its overall text-attention mass is already lower and declines further than SD3.5-medium's throughout sampling (panel a), consistent with a general shift of attention toward image self-attention rather than a mention-specific failure signal. This is evidence against transplanting the SD3.5-medium mechanism onto FLUX.1-dev unmodified, not a positive account of what causes FLUX.1-dev's omissions; we return to this in the Limitation paragraph of Section 7.

### A.14 Counting–binding dependence ($\phi$)

The phi coefficient between the binary counting and binding outcomes,

$$\phi = \frac{P(\mathrm{C} \cap \mathrm{B}) - P(\mathrm{C})P(\mathrm{B})}{\sqrt{P(\mathrm{C})(1 - P(\mathrm{C}))P(\mathrm{B})(1 - P(\mathrm{B}))}}, \tag{3}$$

follows the same ordering in both models (Figure 17): high under $A_1$ (0.70 / 0.81), lower under $A_2$ and $A_3$ (0.32 / 0.40 and 0.23 / 0.39), and near zero under $A_2^-$ ($-0.06$ / 0.06). Most of this ordering follows from the metric definitions. Under $A_1$ the color–noun target equals the base-noun target, so the two outcomes are scored against nearly the same event. Under $A_2^-$ binding fails through inheritance largely independently of counting (Section 4.3). We therefore treat $\phi$ as a property of the evaluation, not of either model.

### A.15 Robustness details

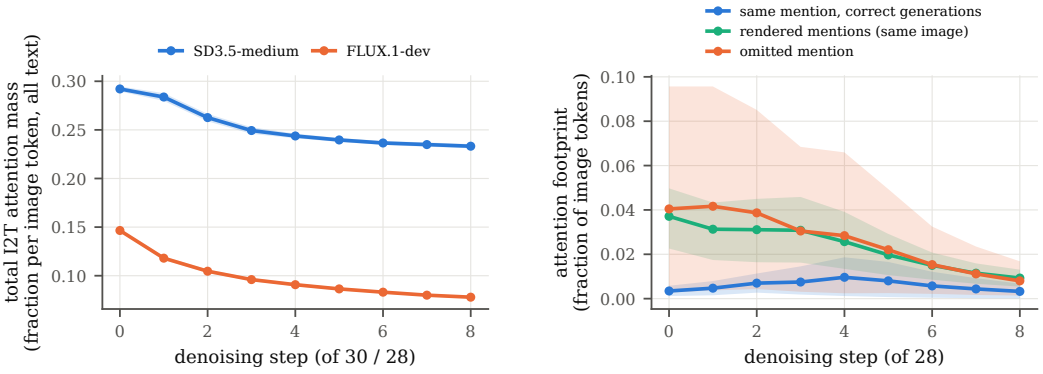
*Four slots.* Colorless counting with four distinct nouns (43.4% FLUX.1-dev, 39.2% SD3.5-medium) falls when all four nouns are repeated (9.5%, 16.8%; differences $-33.9$, CI $[-41.1, -26.6]$, and $-22.5$, CI $[-30.1, -14.9]$). *Slot order at $N_2$.* Colorless counting for the three orders (repeat at first and last / first two / last two positions): SD3.5-medium 35.1 / 32.6 / 35.8%; FLUX.1-dev 40.2 / 38.0 / 46.2%. The grid uses the first order. Table 4 uses the order that matches the numeral form.

### A.16 T2T self-attention and the copying alternative

One alternative account is that repeated occurrences copy one another through text-to-text self-attention. We tested it on the failed-versus-correct contrast within $N_1 \times A_3$. In FLUX.1-dev (12 omission + 12 correct, blocks 0–18), cross-noun attention is significantly *lower* for omission at several blocks (e.g. block 1, $p < 0.001$), and own-adjective attention is *higher* from block 7 onward (e.g. block 14, $p = 0.008$). Both effects run opposite to what copying predicts. In SD3.5-medium (12 fusion + 12 correct), cross-noun and own-adjective attention differ from correct items only at blocks 0–3. Only an attention-sink ratio is directionally consistent with copying, intermittently, in both models.

| Prompt | Baseline | With EMAR ($b = 3$) correct | fusion | omission | overcount |
|---|---|---|---|---|---|
| $N_1 \times A_3$ | correct | 59 | 23 | 2 | 16 |
| | fusion | 21 | 67 | 4 | 7 |
| | omission | 39 | 33 | 18 | 9 |
| | overcount | 35 | 29 | 0 | 35 |
| two instances | correct | 76 | 11 | 1 | 12 |
| | fusion | 48 | 37 | 6 | 9 |
| | omission | 68 | 21 | 11 | 0 |
| | overcount | 56 | 0 | 0 | 44 |

Table 15: **Outcomes without (rows) and with (columns) EMAR** (% of row).




(a) Total I2T attention mass, FLUX.1-dev vs. SD3.5-medium.

(b) FLUX.1-dev's own per-mention coverage.

Figure 16: **A preliminary check of the SD3.5-medium mechanism on FLUX.1-dev.** (a) Unlike the per-mention coverage, this quantity is not filtered by mention selectivity; SD3.5-medium's is higher throughout and the two never cross. (b) In contrast to SD3.5-medium (Figure 13a), the omitted mention's coverage does not collapse selectively: it declines together with the rendered mentions' coverage rather than diverging from them.

### A.17 Value-composition separability

We also tested the separability $S_V$ of the attention-weighted value composition $O_i = \sum_{t \in \text{occ}_i} I2T * V_t$ at each occurrence's peak image-query row, on the pilot items (3 correct + 3 failed colorless $N_1$ per model). In SD3.5-medium $S_V$ shows at most a weak correct-greater-than-failed tendency that co-emerges with $S$. In FLUX.1-dev it shows no reliable separation.

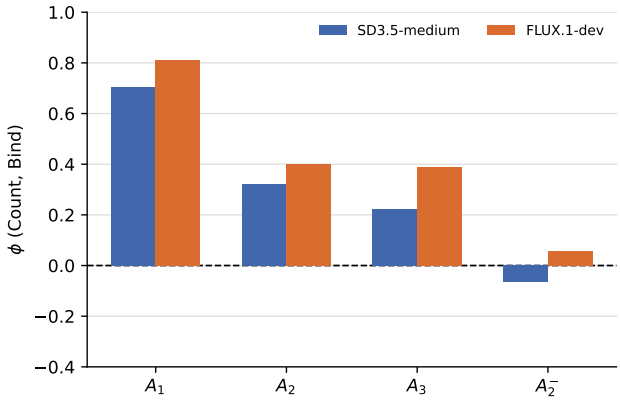



Figure 17: **Counting–binding dependence.** Mean $\phi$ (Eq. 3) across $N_1$–$N_3$; dashed line, independence.

## A.18 Qualitative examples

### A.18.1 EMAR




Figure 18: **Examples of the rendering correction and of the causal conditions** (missing color in bold). Columns are the six single-omission cases in which EMAR renders all three color–object pairs with the correct count and the rest of the image kept. Rows below EMAR show the same cases under the targeted attention bias, the targeted bias with the coarse layout fixed, and a direct color edit of the predicted image; fixing the layout removes the bias's effect in five of the six cases, and the direct edit restores none. Labels give the judged outcome.

### A.18.2 BASELINE

Figures 19 show matched-noise samples for the first four $N_1 \times A_3$ prompt configurations, with the evaluator's outcome label under each image.




(a) “a red bicycle, a brown bicycle and a gold bicycle”




(b) “an orange car, a black car and a silver car”




(c) “a yellow motorcycle, a white motorcycle and a maroon motorcycle”

Figure 19: **Matched-noise samples for $N_1 \times A_3$ prompts (1/2).** Rows are models; each column uses matched initial noise, and all four seeds are shown. Prompt configurations 0–3 are the first four in dataset order, not selected by appearance. Each tile gives the evaluator’s detected count of the repeated noun and the outcome label.