LOST IN ATTENTION: UNDERSTANDING AND CORRECTING COMPOSITIONAL FAILURES IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models remain unreliable at simple compositional requests: objects go missing, counts are wrong, attributes attach to the wrong object and spatial relations are ignored. We trace these failures to how cross-attention allocates its budget: the start token absorbs most of it, the concept named first is favoured, and a concept whose share falls early never appears. The information the text transmits to the image factorises exactly into an *allocation* and a *localization* factor, so stating where an object should be cannot make it appear. We propose Mordant, a training-free guidance method that states compositional requirements as constraints on the frozen model's attention, with multipliers found by dual ascent along the sampling trajectory and a relation margin derived from the frame. On Stable Diffusion 1.4, Mordant raises GenEval from 0.45 to 0.70, T2I-CompBench colour binding from 0.37 to 0.69 and 2D spatial relations from 0.13 to 0.39, and VISOR object accuracy from 33.4% to 79.9%; with its constants unchanged it also raises GenEval on SD 2.1 (0.50 to 0.72), SDXL (0.58 to 0.71), SD3-medium (0.70 to 0.83) and SD3.5-medium (0.70 to 0.83).

1 INTRODUCTION

Text-to-image diffusion models (Ho et al., 2020; Rombach et al., 2022) render photographic detail with ease, yet remain unreliable at simple compositional requests (Feng et al., 2023; Chefer et al., 2023): *a horse and a train* produces a train and no horse (Fig. 1). Benchmarks measure these failures as object presence, numeracy, attribute binding and spatial relations (Ghosh et al., 2023; Huang et al., 2023; Gokhale et al., 2022), and training-free methods steer the frozen model at inference time to repair one or two of them at a time (Chefer et al., 2023; Chen et al., 2024; Han et al., 2025; Marioriyad et al., 2025).

We start from a measurement: the dominant failure is not that an object is drawn in the wrong place or colour but that it is *not drawn at all*. Cross-attention (Hertz et al., 2023) explains why. Attention at every pixel is a budget the prompt's tokens compete for, and the competition is lopsided: the start token absorbs 84 to 89% of the attention at an object's own pixels while contributing only 38 to 48% of what is rendered there, the object named first is drawn far more often than the one named second, and a word's share at the first coarse layer already predicts whether its object will appear. We show that the information the text transmits to the image factorises, token by token, into an *allocation* factor (the token's share of the budget) and a *localization* factor (how far its map is from uniform); spatial guidance acts only on the second, so when allocation has collapsed no statement about where an object should be can make it appear.

Multi-term guidance methods (Agarwal et al., 2023; Meral et al., 2024; Han et al., 2025) also share a less visible limitation: they are sums of fixed-weight penalties, and in our measurements the sign of a term's effect flips as other terms are strengthened. We propose Mordant, which states what the picture must satisfy as *conditions* on the frozen model's attention and lets the sampling trajectory find how hard each must be pushed. The conditions are likelihood-ratio tests between a concept's map and the uniform map or another word's map; the objective regulates both factors, with an effective-area term, a bounded repulsion and an allocation floor; the multipliers are found by dual ascent (Bertsekas, 1982), starting from zero, so a condition the unguided model already meets never

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.

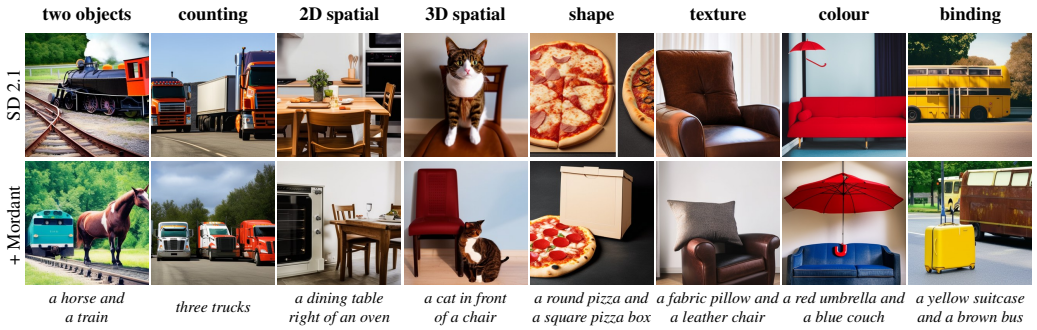


Figure 1: **One approach, eight compositional skills.** SD 2.1 (top) and SD 2.1 with Mordant (bottom) from the same starting noise; the prompt is under each column.

acts; and a forward re-allocation moves part of the start token's attention to each object at the pixels it already occupies. Everything is read from attention the model already computes, with no training, layout, language-model parse or verifier.

On Stable Diffusion 1.4, Mordant raises GenEval from 0.45 to 0.70, T2I-CompBench colour binding from 0.37 to 0.69 and 2D spatial from 0.13 to 0.39, and VISOR object accuracy from 33.4% to 79.9% with the pair placed correctly in 99.97% of the images that contain both; with the same constants it improves SD 2.1, SDXL (Podell et al., 2024), SD3-medium (Esser et al., 2024) and SD3.5-medium (Stability AI, 2024).

Contributions.

- An analysis of compositional failure as an *allocation* failure, with an exact decomposition of the text-to-image information into allocation and localization (Sec. 4).
- Mordant, a constrained program on the frozen model's attention with likelihood-ratio conditions, multipliers found by dual ascent, a derived relation margin and a forward re-allocation of the start token's attention (Sec. 5).
- State-of-the-art results among training-free methods on GenEval, T2I-CompBench and VISOR, on SD 1.4 and four further backbones, with ablations of every component (Sec. 6).

2 Related Work

Guidance on attention. Cross-attention maps localise the prompt's words (Hertz et al., 2023; Tang et al., 2023), and training-free guidance steers sampling with the gradient of a loss on them (Bansal et al., 2024; Epstein et al., 2023). For presence and binding, Attend-and-Excite (Chefer et al., 2023) raises a neglected subject's peak attention, Divide-and-Bind (Li et al., 2023b), A-STAR (Agarwal et al., 2023), SynGen (Rassin et al., 2023), CONFORM (Meral et al., 2024) and Attention Overlap (Marioriyad et al., 2025) separate or contrast the words' maps, and InitNO (Guo et al., 2024) optimises the initial noise. For relations, Layout Guidance (Chen et al., 2024) confines maps to given boxes, STORM (Han et al., 2025) transports them toward target regions, PSG (Rezaei et al., 2025) maximises the probability of the stated order, InfSplign (Rastegar et al., 2025) adds a centroid hinge to presence terms, Noise Refinement (Izadi et al., 2025) sums equally weighted losses with a language-model parse, and Learn-to-Steer (Yiflach et al., 2026) learns the loss per backbone. Every one of these states where a map should lie or how peaked it should be, which acts on localization; the start token's hold on the budget is left where it is, and the terms are combined with fixed weights. Mordant regulates allocation directly and finds its multipliers along the trajectory (the detailed related work and a term-by-term comparison are in Sec. A.1).

Layouts, fine-tuning and sample search. Layout-conditioned models (Li et al., 2023a; Lian et al., 2024), fine-tuning on spatial data (Chatterjee et al., 2024a; Zhang et al., 2025), text-side token merging (Hu et al., 2024) and search over samples with a verifier (Ma et al., 2025; Singhal et al., 2025) need extra input, training or several samples per image; Mordant needs none of these.

3 BACKGROUND

Diffusion models. A diffusion model (Ho et al., 2020; Song et al., 2021) learns to reverse a process that gradually corrupts data with Gaussian noise. With a noise schedule $\bar{\alpha}_t \in (0, 1]$ decreasing in t, a clean sample z_0 is noised as $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and a network $\epsilon_\theta(z_t, t, y)$ is trained to predict the noise from the noisy sample, the time step and a condition y. The prediction is, up to scale, the score of the noisy distribution, $\nabla_{z_t} \log p_t(z_t \mid y) = -\epsilon_\theta(z_t, t, y)/\sqrt{1-\bar{\alpha}_t}$, and sampling starts from pure noise z_T and applies a sampler step (DDPM, DDIM or PNDM (Ho et al., 2020; Song et al., 2021)) that uses ϵ_θ to move from z_t to z_{t-1}. Stable Diffusion (Rombach et al., 2022) runs this process in the latent space of an autoencoder with a U-Net denoiser conditioned on the 77 token embeddings of a CLIP text encoder (Radford et al., 2021): a start token, the prompt's words and padding. Classifier-free guidance (Ho & Salimans, 2021) sharpens the conditioning by extrapolating from the unconditional prediction, $\hat{\epsilon} = \epsilon_\theta(z_t, t, \varnothing) + s\,\big(\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \varnothing)\big)$, with a scale s of about 7.5.

Cross-attention. The text reaches the image only through cross-attention layers inside the U-Net. At a layer with n spatial positions, the image features give queries $Q \in \mathbb{R}^{n \times d}$ and the token embeddings give keys K and values V, and the layer outputs

$$\mathrm{Attn}(Q, K, V) = A\,V, \qquad A = \mathrm{softmax}\big(QK^\top/\sqrt{d}\big), \tag{1}$$

where the softmax runs over the tokens, so each row of A is a distribution over the 77 tokens at one position. The column of A belonging to a word, reshaped to the layer's grid, is that word's *attention map*; it shows where in the image the word is being rendered (Hertz et al., 2023; Tang et al., 2023). Joint-attention transformers such as SD3 (Esser et al., 2024) have no separate cross-attention, but the image-to-text block of their joint attention plays the same role.

Inference-time guidance. Because the prediction is a score, it can be steered at sampling time without retraining. Classifier guidance (Dhariwal & Nichol, 2021) adds the gradient of a classifier's log-likelihood to the score, and training-free guidance replaces the classifier with any differentiable loss $\mathcal{L}(z_t)$ computed from the model's own quantities (Bansal et al., 2024; Yu et al., 2023; Epstein et al., 2023). Its gradient is applied either as a step on the latent, $z_t \leftarrow z_t - \eta\,\nabla_{z_t}\mathcal{L}$ (Chefer et al., 2023; Han et al., 2025), or added to the noise prediction, $\hat{\epsilon} \leftarrow \hat{\epsilon} + w\,\nabla_{z_t}\mathcal{L}$ (Epstein et al., 2023; Rastegar et al., 2025), which tilts the sampling distribution toward low loss. Mordant uses the second form; what differs is the loss, which we state as a constrained program in Sec. 5.

4 LOST IN ATTENTION: WHERE COMPOSITIONAL FAILURES COME FROM

We work with Stable Diffusion 1.4 and the cross-attention of Eq. (1). At a cross-attention layer l with n_l spatial positions, the head-averaged attention $A_l(u, t)$ gives, for every position u, a distribution over the text tokens t: $\sum_t A_l(u, t) = 1$. For a concept c of the prompt (an object noun, an attribute) with token positions $\tau(c)$ we write

$$a_c(u) = \sum_{t \in \tau(c)} A_l(u, t), \qquad E_c = \frac{1}{n_l}\sum_u a_c(u), \qquad p_c(u) = \frac{a_c(u)}{n_l\, E_c}. \tag{2}$$

E_c is the concept's *share* of the layer's attention budget, the expected fraction of the image assigned to it, and p_c is its *map*, a distribution over the image plane.

4.1 ALLOCATION AND LOCALIZATION

Read the attention as a channel from positions to tokens: draw a position U uniformly and a token $T \sim A_l(U, \cdot)$. The token marginal of this joint distribution is the share, $P(T{=}t) = E_t$, and the position given the token is the map, $P(U{=}u \mid T{=}t) = p_t(u)$. The information the text transmits to the image through the layer is therefore

$$I(T;U) \;=\; H(U) - H(U \mid T) \;=\; \sum_t E_t\,\big(\log n_l - H(p_t)\big) \;=\; \sum_t \underbrace{E_t}_{\text{allocation}}\ \underbrace{D_{\mathrm{KL}}\big(p_t \,\big\|\, \mathcal{U}\big)}_{\text{localization}}, \tag{3}$$

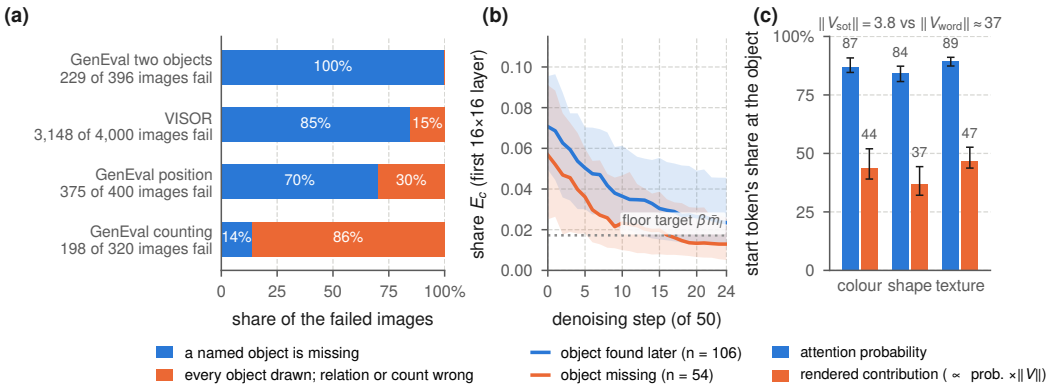


Figure 2: **Compositional failure is an allocation failure** (unguided SD 1.4). **(a)** Among failed images, the share with a named object missing against the share with every object drawn but the relation or count wrong. **(b)** An object's attention share E_c (Eq. (2)) at the first 16×16 decoder layer over the denoising steps, split by whether the object is drawn; dotted: where the allocation floor switches on. **(c)** The start token's share of attention probability at an object's cells against its share of what is rendered.

where $\mathcal{U}$ is the uniform map. Each token contributes the product of two factors: how much of the budget it holds, and how far its map is from carrying no spatial information. The factorisation makes a structural point that the rest of the paper builds on. Every spatial statement, whether a centroid gap, a transport plan, an overlap or a box, is a statement about p_t and moves only the second factor. If the first factor has collapsed, the token transmits nothing wherever its map lies, so geometric guidance alone cannot guarantee that an object appears. The measurements below show that the first factor is the one that fails; Fig. 2 collects them.

4.2 Diffusion models play favorites

The start token is a sink. At the pixels an object occupies, the start token <sot> receives 84 to 89% of the attention probability (SD 1.4, 16×16 layers, 180 prompt-noun pairs from T2I-CompBench's colour, shape and texture sets). Its value vector is an order of magnitude shorter than a word's ($\|V_{\texttt{<sot>}}\| = 3.8$ against about 36), so it contributes only 38 to 48% of the mixture actually rendered there. The prompt's words compete for the remaining 11 to 16% of the probability, which carries most of the content.

Allocation predicts presence, early. A concept's share E_c at the first 16×16 decoder layer separates the objects that a detector later finds (mean share 0.043) from the ones it misses (0.018) with an AUC of 0.75 on unguided VISOR images. The spread of the same maps predicts presence less well (AUC 0.60 to 0.65): the size of a map says less about whether its object will render than the share of the budget behind it.

The object named first is also drawn more often than the one named second (69.1% against 61.7% on a class-balanced VISOR set), the share of an object that will go missing decays during sampling, and once both objects are drawn the relation is almost never wrong; these measurements are in Sec. A.7.

5 Mordant: compositional guidance as a constrained program

Mordant states what the picture must satisfy as conditions on the frozen model's own cross-attention, and lets the sampling trajectory find how hard each condition has to be pushed. Every condition compares a concept's map with the uniform map, which carries no spatial information, or with another word of the same prompt; the objective regulates the two factors of Eq. (3), allocation and localization. The prompt supplies the object nouns with their counts, the attributes of each noun and the relation between two objects when one is stated. No layout, language model or verifier is involved. Fig. 3 shows, on two prompts, the attention maps the program reads and how they change under it.

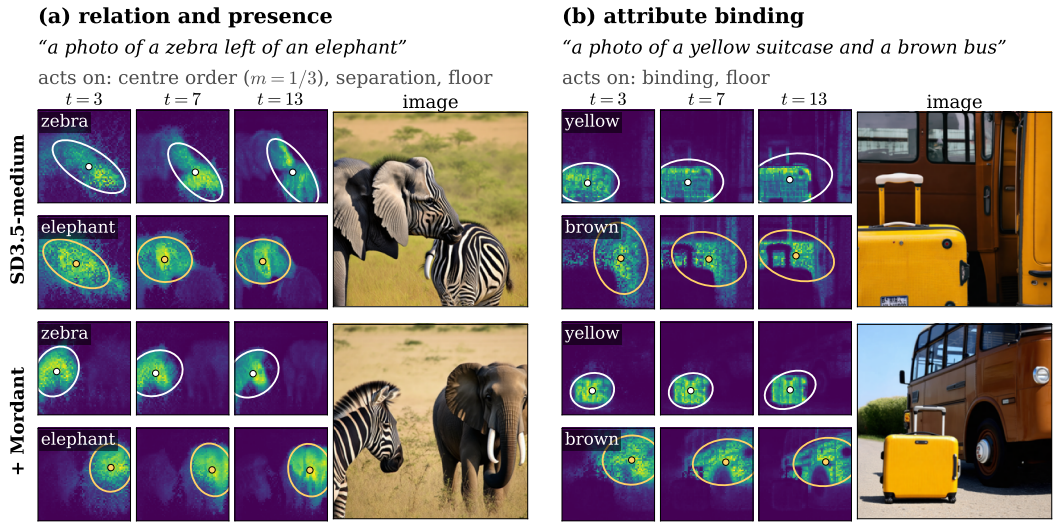


Figure 3: **What Mordant acts on** (SD3.5-medium). Each row is one word's attention map at steps $t = 3$, 7 and 13 of 28, with its centre and spread marked, beside the final image, for the unguided model and Mordant from the same starting noise. Mordant moves the zebra's map to the left of the elephant's (a) and keeps *yellow* on the suitcase and *brown* on the bus (b).

5.1 Readings

At each hooked layer l and guided step, every concept c is read through Eq. (2): its share E_c, its map p_c, and the map's mean μ_c and covariance Σ_c in the coordinates of the unit square. A noun with count k is represented by k instance nodes, obtained by a soft k-means split of its map; the objective acts on nodes and the conditions on nouns. $\mathcal{U}$ is the uniform map on the layer's n cells, $H(p, q) = -\sum_u p(u) \log q(u)$ is the cross-entropy, and $H(p, \mathcal{U}) = \log n$ for every p.

5.2 Loss function

With objects i, j, instance nodes k, attributes a owned by noun $o(a)$ and the set R of object pairs the prompt relates,

$$\min\ F = \underbrace{\kappa \sum_k \mathrm{EP}(p_k)}_{\text{effective area}} + \underbrace{\kappa \sum_{\{i,j\} \notin R} \big[m - \|\mu_i - \mu_j\|\big]_+^2}_{\text{apart}} + \underbrace{\lambda_f \sum_i \big[\log(\beta\, \bar{m}_l) - \log E_i\big]_+}_{\text{allocation floor}} \tag{4}$$

$$\text{s.t.}\quad H(p_i, p_j) \geq \log n \qquad (i \neq j) \qquad \text{(separation)} \tag{5}$$

$$d \cdot (\mu_S - \mu_O) \geq m \qquad \text{(relation)} \tag{6}$$

$$H\big(p_a, p_{o(a)}\big) \leq \min\Big(\min_{j \neq o(a)} H(p_a, p_j),\ \log n\Big) \qquad \text{(binding)} \tag{7}$$

with the margin $m = \frac{1}{3}$ (Thm. 2). Each term is the smallest statement we found for one graded event.

Effective area. The first term measures how much of the image a concept's attention covers. Treat the concept's map p as a distribution over the grid, with centre μ_p and covariance Σ_p; the Gaussian with the same centre and covariance covers an ellipse whose area is proportional to $\sqrt{\det \Sigma_p}$, and $\mathrm{EP}(p) = 2\sqrt{\det(\Sigma_p + s^2 I)}$ is that area (it is the Gaussian's entropy power, $e^H/(\pi e)$ in two dimensions). The term $s^2 I$, with s the width of one grid cell of the layer ($\frac{1}{16}$ of the frame on a 16×16 layer), keeps the area from falling below one cell: a map concentrated on a single cell would otherwise have zero area and an unbounded gradient, and a layer cannot resolve anything finer than its own cell. At a fixed share a compact claim wins its pixels, while a haze loses them everywhere. Unlike the trace of the covariance, the usual concentration term, it charges area and not elongation (Prop. 3), so an oblong or triangular object is not pulled toward a disc.

Apart. Two objects that the prompt does not relate still have to be two objects. The hinge $[m - \|\mu_i - \mu_j\|]_+^2$ repels them with a force $2(m - r)$ at distance $r < m$ and exactly zero beyond; it is a preference, not a condition, because nothing in the prompt states it.

Allocation floor. The last term holds the allocation factor of Eq. (3) away from zero. $\bar{m}_l$ is the median share of an object that a detector finds, measured once per layer on unguided images, and the hinge is taken in log units because shares differ by two orders of magnitude between layers. It is inactive while a concept holds its share and switches on only as that share decays.

Separation and binding are likelihood-ratio tests between maps, which keep a gradient where an overlap statistic has none (Sec. A.8).

5.3 Optimising the loss inside the sampler

With the conditions written $g_c \le 0$ and multipliers $\lambda_c \ge 0$, the loss at a guided step is the augmented Lagrangian

$$\mathcal{L} = F + \sum_c \psi(g_c, \lambda_c), \qquad \psi(g, \lambda) = \frac{[\lambda + \rho g]_+^2 - \lambda^2}{2\rho}, \tag{8}$$

summed over the hooked layers. Its gradient is added to the classifier-free-guided noise prediction, $\hat{\epsilon}_t = \epsilon_\theta(z_t, t, y) + w\,\nabla_{z_t}\mathcal{L}$, during the first 25 of 50 steps, and after each step the multipliers take a dual-ascent step, $\lambda_c \leftarrow [\lambda_c + \rho\,\bar{g}_c]_+$, from $\lambda_c = 0$, with $\bar{g}_c$ the condition's mean over layers. A condition that already holds on the unguided trajectory never acts; a violated one acts immediately through the quadratic part of ψ, and its multiplier grows only while it stays violated. The multipliers are therefore not hand-set weights, and rescaling a condition changes nothing but its multiplier (Prop. 1).

Forward re-allocation. The floor acts through the gradient, one step at a time; the start token's hold on the budget can also be relaxed inside the forward pass. At every cross-attention layer and head h, let $S_i = \{u : \tilde{a}_i(u) > \frac{1}{2}\}$ be the cells where object i already dominates its own map ($\tilde{a}_i$ is a_i rescaled to its range, detached). There we move part of the start token's probability to the object's tokens,

$$A'_h(u, \texttt{<sot>}) = A_h(u, \texttt{<sot>}) - \Delta_h(u), \qquad A'_h(u, t) = A_h(u, t) + \Delta_h(u)/|\tau(i)| \quad (t \in \tau(i)), \tag{9}$$

with $\Delta_h(u) = \min(\delta, [A_h(u, \texttt{<sot>}) - \epsilon_0]_+)$, $\delta = 0.5$ and $\epsilon_0 = 0.05$. Rows remain distributions and the unconditional branch is untouched. Since S_i is where the model already places the object, the edit raises E_i without moving p_i's support; and since the start token's value vector is an order of magnitude shorter than a word's (Sec. 4.2), a small transfer of probability is a large transfer of what is rendered. Algorithm 1 summarises the procedure.

6 Experiments

6.1 Experimental setup

Unless a table says otherwise (Sec. 6.2), every image in this section is generated by Stable Diffusion 1.4 (Rombach et al., 2022) at 512×512 with 50 DDPM steps, classifier-free guidance 7.5 and an fp16 U-Net (every loss in fp32); Mordant acts on the first 25 steps, and image k of a prompt is generated from seed $42+k$ for every method, so every comparison is paired (Sec. E). *GenEval* (Ghosh et al., 2023): 553 prompts, 4 images each (2,212), six tasks graded by Mask2Former and a CLIP colour classifier. *T2I-CompBench* (Huang et al., 2023): the 300 evaluation prompts of each of the eight categories, 10 images each (24,000), graded by BLIP-VQA (colour, shape, texture), UniDet with the benchmark's exact label rule (2D spatial), UniDet R50 with a depth map (3D spatial) and UniDet R50 (numeracy), CLIPScore (non-spatial) and the 3-in-1 metric (complex). *VISOR* (Gokhale et al., 2022): a 1,000-prompt subset of the 25,280-prompt benchmark whose class and relation frequencies match the full set's ($\chi^2 = 5.0$ on 79 degrees of freedom, against 306 for the 1,000-prompt subset in common use; Sec. F), graded by VISOR's OWL-ViT detector. Every constant of Mordant was chosen on held-out prompts that share nothing with these sets (Sec. G). Competitors are run from their authors' released code with default settings on the same backbone, seeds and scorers (Sec. K).

Table 1: **GenEval** by base model: 553 prompts, 4 images each, official scorer; Attr.: colour attribution. On the other backbones only the guidance weight is rescaled (Sec. E), and the MMDiTs run without the forward re-allocation.

Method	Venue	Extra	Single	Two	Count	Colours	Position	Attr.	Overall
SD 1.4†	–	–	**0.99**	0.42	0.38	0.78	0.06	0.08	0.45
+ Attend&Excite†	*TOG'23*	–	0.98	0.62	0.33	0.75	0.11	0.19	0.50
+ Divide&Bind†	*BMVC'23*	–	0.98	0.62	0.32	0.73	0.10	0.18	0.49
+ InitNO†	*CVPR'24*	–	0.97	0.71	0.32	0.75	0.11	0.25	0.52
+ CONFORM†	*CVPR'24*	–	**0.99**	0.70	0.33	0.72	0.11	**0.40**	0.54
+ Attention Overlap†	*TMLR'25*	–	**0.99**	0.74	0.33	0.72	0.15	0.23	0.53
+ Self-Cross†	*CVPR'25*	–	**0.99**	0.74	0.33	0.72	0.13	0.26	0.53
+ STORM†	*CVPR'25*	–	**0.99**	0.41	0.33	0.72	0.15	0.05	0.44
+ Noise Refinement†	*arXiv'25*	–	0.98	0.71	0.33	0.80	0.31	0.31	0.57
+ CoMPaSS (Zhang et al., 2025)	*ICCV'25*	fine-tuning	**0.99**	0.68	0.34	0.73	0.46	0.17	0.56
+ Mordant	–	–	0.98	**0.82**	**0.48**	**0.83**	**0.71**	**0.40**	**0.70**
Improvement			−0.01	+0.40	+0.10	+0.05	+0.65	+0.32	+0.25
SD 2.1†	–	–	**0.99**	0.48	0.47	0.84	0.07	0.15	0.50
+ STORM†	*CVPR'25*	–	**0.99**	0.48	0.42	0.83	0.16	0.09	0.49
+ SPRIGHT†	*ECCV'24*	fine-tuning	**0.99**	0.63	**0.48**	0.78	0.11	0.14	0.52
+ Learn-to-Steer (Yiflach et al., 2026)	*WACV'26*	learned loss	–	–	–	–	0.54	–	–
+ Mordant	–	–	0.98	**0.85**	0.45	**0.86**	**0.78**	**0.42**	**0.72**
Improvement			−0.01	+0.37	−0.02	+0.02	+0.71	+0.27	+0.22
SDXL†	–	–	0.98	0.79	**0.51**	**0.88**	0.11	0.23	0.58
+ Mordant	–	–	**0.99**	**0.88**	0.41	0.87	**0.68**	**0.42**	**0.71**
Improvement			+0.01	+0.09	−0.10	−0.01	+0.57	+0.19	+0.13
SD3-medium†	–	–	**1.00**	0.86	0.61	**0.87**	0.28	0.57	0.70
+ Mordant	–	–	**1.00**	**0.95**	**0.70**	**0.87**	**0.79**	**0.66**	**0.83**
Improvement			0.00	+0.09	+0.09	0.00	+0.51	+0.09	+0.13
SD3.5-medium†	–	–	0.99	0.88	**0.71**	0.82	0.26	0.56	0.70
+ Mordant	–	–	**1.00**	**0.93**	0.70	**0.84**	**0.78**	**0.70**	**0.83**
Improvement			+0.01	+0.05	−0.01	+0.02	+0.52	+0.14	+0.13

† Reproduced using the official released code; other rows as published.

Table 2: **T2I-CompBench** on SD 1.4: 300 prompts per category, 10 images per prompt, the benchmark's official scorers. Extra: what a method needs beyond the frozen model. All categories: Tab. 12.

			Attribute binding			Spatial		
Method	Venue	Extra	Colour	Shape	Texture	2D	3D	Num.
SD 1.4 (Huang et al., 2023; 2025)	–	–	0.38	0.36	0.42	0.12	0.30	0.45
+ Attention Overlap†	*TMLR'25*	–	0.57	0.47	0.57	0.18	–	–
+ REVISION (Chatterjee et al., 2024b)	*ECCV'24*	layout	–	–	–	0.33	–	–
+ CoMPaSS (Zhang et al., 2025)	*ICCV'25*	fine-tuning	0.49	0.43	0.53	0.34	–	–
+ Mordant	–	–	**0.69**	**0.55**	**0.68**	**0.39**	**0.37**	**0.56**
Improvement			+0.31	+0.19	+0.26	+0.27	+0.07	+0.11

† Reproduced using the official released code.

6.2 Comparison with the state of the art

GenEval (Tab. 1). Mordant reaches 0.70 overall from 0.45 unguided, above every training-free method we reproduced (Noise Refinement 0.57, CONFORM 0.54, Self-Cross 0.53, Attention Overlap 0.53, InitNO 0.52, Attend&Excite 0.50, Divide&Bind 0.49, STORM 0.44), and above the published fine-tuned CoMPaSS (0.56). By task it draws both objects in 0.82 of the two-object prompts (0.42 unguided), places the pair in 0.71 of the position prompts (0.06), binds the two colours in 0.40 (0.08), counts correctly in 0.48 (0.38) and keeps the single-colour and single-object tasks (0.83, 0.98). The remaining failures are still mostly missing objects (Tab. 15).

T2I-CompBench (Tab. 2). Mordant raises every category over SD 1.4, by 0.19 to 0.31 on attribute binding and 0.27 on 2D spatial; non-spatial prompts, which name no object pair, are unchanged

Table 3: **VISOR** on the balanced 1,000-prompt subset, 4 images per prompt, OWL-ViT. OA: both objects found; cond: relation correct given both; uncond = OA×cond; V_N: prompts with at least N of 4 images correct. Other detectors: Tab. 10; published full-benchmark rows: Tab. 9.

Method	Venue	Extra	OA	uncond	cond	V_1	V_2	V_3	V_4
SD 1.4†	–	–	33.40	21.30	63.77	50.80	23.40	8.70	2.30
+ Mordant	–	–	**79.88**	**79.85**	**99.97**	**96.10**	**88.40**	**77.40**	**57.50**
Improvement			+46.48	+58.55	+36.20	+45.30	+65.00	+68.70	+55.20
SDXL†	–	–	70.22	45.15	64.29	76.60	55.70	33.20	15.10
+ Mordant	–	–	**82.20**	**81.85**	**99.57**	**96.50**	**90.60**	**81.00**	**59.30**
Improvement			+11.98	+36.70	+35.28	+19.90	+34.90	+47.80	+44.20

† Reproduced using the official released code.

(Tab. 12). It is the highest row in every column, above the published fine-tuned CoMPaSS (0.49 / 0.43 / 0.53 on the attribute sets); STORM, whose released code we could not bring to its published attribute scores, is compared in Sec. K.1. On 2D spatial its 0.39 is above layout-conditioned (REVISION, 0.33) and fine-tuned (CoMPaSS, 0.34) methods.

VISOR (Tab. 3). On the balanced subset, Mordant draws both objects in 79.9% of the images (33.4% unguided) and, given both, places them on the asked side in 99.97% (63.8% unguided); every image of a prompt is correct for 57.5% of the prompts (V_4), against 2.3% unguided. The second-named object is now drawn almost as often as the first (88.1% against 91.0%; unguided 61.7% against 69.1%), and other detectors give the same picture (Tab. 10). The same constants carry to SDXL (guidance weight rescaled only): object accuracy rises from 70.2% to 82.2%, the relation given both objects from 64.3% to 99.6%, and V_4 from 15.1% to 59.3%.

Other backbones. Tab. 1 runs Mordant on the other backbones with no constant re-tuned; only the guidance weight is rescaled, by a ratio measured once per backbone (Sec. E). GenEval rises from 0.50 to 0.72 on SD 2.1, 0.58 to 0.71 on SDXL, and 0.70 to 0.83 on both SD3-medium and SD3.5-medium, carried mostly by position, colour attribution and two objects; counting improves only on SD 1.4 and SD3-medium (Sec. 7).

6.3 Ablation study

Tab. 4 removes one component of Mordant at a time, swaps the relation statistic, and replaces the multipliers by zero (fixed penalties) or by hand-tuned fixed weights; Sec. M shows the same removals on the attention maps of single prompts. Presence and the relation are read on held-out VISOR prompts that share no object pair with the evaluation subset (Sec. G), and attribute binding on T2I-CompBench colour at the full protocol.

Each component carries one graded event. Binding carries the attributes: without it, colour falls from 0.69 to 0.62 (Figs. 12 and 13). The apart term carries the second object that the attribute score needs, since the score is a product over the prompt's noun phrases: without it the second object forms inside or beside the first (Fig. 11), and colour falls to 0.65. The allocation floor carries presence (-3.6 points of object accuracy). Holding every multiplier at zero, so that the same conditions act as fixed penalties, is the largest change (-9.5), and the relation is no longer held: $\text{VISOR}_{\text{cond}}$ falls from 100.0 to 89.6. Hand-tuned fixed weights recover part of the gap (79.8), and at the full protocol stay within 0.02 of Mordant on colour, texture, 2D and 3D but are below it on shape, numeracy, GenEval and VISOR (Sec. J.2). The surprisal as the relation statistic costs 5.1 points on VISOR's sign grader (Sec. A.2). The same ablations read with two other detectors are in Sec. J.

6.4 Qualitative results and failure cases

Fig. 4 sets Mordant beside the training-free methods we ran on SD 2.1, on prompts where the unguided model fails: the competitors miss an object, the count or the arrangement, while Mordant draws the named objects in the stated arrangement and colours; Sec. L adds comparisons on SD 1.4 and uncurated samples.

Table 4: **Ablations**: ✓ the component is used, ✗ it is not. Relation statistic: centre order (ctr) or surprisal; multipliers: found by dual ascent, held at zero (fixed penalties) or hand-tuned fixed weights. VISOR: object accuracy on held-out prompts (800 images, OWL-ViT). Colour: T2I-CompBench colour at the full protocol.

	Sep.	Bind.	Apart	Floor	Re-alloc.	Relation	Multipliers	VISOR OA	Colour
SD 1.4 (unguided)	–	–	–	–	–	–	–	29.10	0.37
Mordant	✓	✓	✓	✓	✓	ctr	dual	**83.10**	0.69
no binding	✓	✗	✓	✓	✓	ctr	dual	–	0.62
no apart term	✓	✓	✗	✓	✓	ctr	dual	–	0.65
no allocation floor	✓	✓	✓	✗	✓	ctr	dual	79.50	–
surprisal as relation statistic	✓	✓	✓	✓	✓	surpr.	dual	78.00	–
multipliers held at zero	✓	✓	✓	✓	✓	ctr	zero	73.60	–
hand-tuned fixed weights	✓	✓	✓	✓	✓	ctr	fixed	79.80	**0.70**

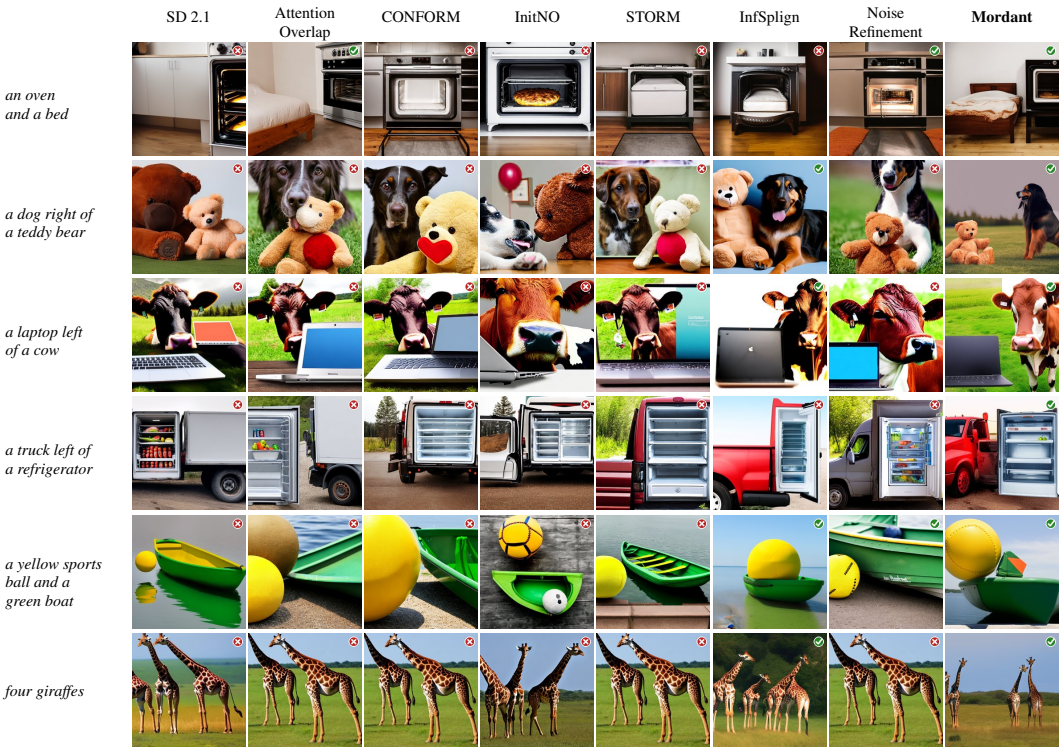


Figure 4: **GenEval on SD 2.1, method by method**: each row is one prompt, every method from the same starting noise, with GenEval's verdict in the corner.

7 Conclusion

Compositional failures of text-to-image diffusion models are, first of all, failures of allocation: a concept that loses its share of the attention budget early is not drawn in the wrong place but not drawn at all. The information the text transmits to the image factorises into an allocation and a localization factor, and spatial guidance, which acts only on localization, cannot by itself make a missing object appear. Mordant states allocation, localization, relations and binding as conditions on the model's own attention, and finds their multipliers along the sampling trajectory instead of fixing weights by hand, so that a condition the model already meets never acts. With one set of constants it raises GenEval, T2I-CompBench and VISOR on SD 1.4 and transfers to four further backbones with only the guidance weight rescaled. It regulates what the frozen model already represents: it moves colour more than texture or shape, it states no count, and an object the backbone renders badly is placed and kept rather than improved; these limitations are discussed in Sec. A.9. More broadly, reliable prompt following requires controlling not only where a concept is generated, but how strongly it participates in generation.

AI USE STATEMENT

We used large language models as assistants to help implement and debug the experiments, to edit the text for readability, to search for and summarise related work, to format references and to prepare figures. We did not use them to generate or clean data, for translation, or for data analysis. The authors designed the study and the method, reviewed all AI-assisted output, checked every reported number against the run that produced it, and take full responsibility for the content of the paper.

ETHICS STATEMENT

Mordant changes how closely an existing text-to-image model follows a prompt; it adds no capability to generate content the backbone cannot already produce, and it inherits the backbone's biases and its potential for misuse, including the generation of misleading images. All experiments use public models (Stable Diffusion 1.4, 2.1, XL, 3 and 3.5) and public benchmarks (VISOR, T2I-CompBench, GenEval) whose prompts describe everyday objects; no personal data were collected and no human subjects were involved. Improving presence and binding makes a model's output more faithful to its prompt, which also makes harmful prompts more faithfully rendered; deployments should keep the safety filters of the backbone, which our method does not alter. The analysis of Sec. 4 documents an order bias inherited from the text encoder (the object named first is drawn more often); we report it as a property of the models studied, and Mordant reduces its effect on the objects named in a prompt.

REPRODUCIBILITY STATEMENT

Sec. 5 and algorithm 1 specify the method completely, and Sec. D lists every setting with its origin: derived, selected on held-out prompts, or fixed. The held-out tuning prompts are disjoint from every evaluation prompt and their construction is described in Sec. G; the balanced VISOR subset is constructed deterministically (Sec. F). Every reported number is produced by the benchmarks' own evaluation code at the versions described in Sec. E, including the exact label-matching rule of T2I-CompBench's 2D scorer and the R50 UniDet backbone of its 3D and numeracy scorers; image k of every prompt is generated with seed $42 + k$ for every method. Baselines were run from their authors' released code with their default settings (Sec. K), and published numbers are marked as such. We report the measured run-to-run spread of every benchmark (Sec. H) and state which differences lie within it. All proofs are in Sec. B. Code, prompt sets and the scripts that produce every table will be released; an anonymised copy is provided as supplementary material.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. CLIP under the microscope: A fine-grained analysis of multi-object representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Aishwarya Agarwal, Srikrishna Karanam, K. J. Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-STAR: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. *OpenAI technical report*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2. https://huggingface.co/black-forest-labs/FLUX.2-dev, 2025.

Luiz F. O. Chamon, Mohammad Reza Karimi, and Anna Korba. Constrained sampling with primal-dual Langevin Monte Carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2411.00568.

Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, and Yezhou Yang. Getting it right: Improving spatial consistency in text-to-image models. In *European Conference on Computer Vision (ECCV)*, 2024a.

Agneet Chatterjee, Yiran Luo, Tejas Gokhale, Yezhou Yang, and Chitta Baral. REVISION: Rendering tools enable spatial fidelity in vision-language models. In *European Conference on Computer Vision (ECCV)*, 2024b.

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), 2023.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

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

DeepFloyd Lab at StabilityAI. DeepFloyd IF. https://github.com/deep-floyd/IF, 2023.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *International Conference on Machine Learning (ICML)*, 2023.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.04312.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2023.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. InitNO: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Woojung Han, Yeonkyung Lee, Chanyoung Kim, Kwanghyun Park, and Seong Jae Hwang. Spatial transport optimization by repositioning attention map for training-free text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 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 (TPAMI)*, 47(5):3563–3579, 2025.

Amir Mohammad Izadi, Seyed Mohammad Hadi Hosseini, Soroush Vafaie Tabar, Ali Abdollahi, Armin Saghafian, and Mahdieh Soleymani Baghshah. Fine-grained alignment and noise refinement for compositional text-to-image generation. *arXiv preprint arXiv:2503.06506*, 2025.

Vivek Jayaram, Ira Kemelmacher-Shlizerman, Steven M. Seitz, and John Thickstun. Constrained diffusion implicit models. *arXiv preprint arXiv:2411.00359*, 2024.

Shervin Khalafi, Dongsheng Ding, and Alejandro Ribeiro. Constrained diffusion models via dual training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2408.15094.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022.

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

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & Bind your attention for improved generative semantic nursing. In *British Machine Vision Conference (BMVC)*, 2023b.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.

Guan-Horng Liu, Tianrong Chen, Evangelos A. Theodorou, and Molei Tao. Mirror diffusion models for constrained and watermarked generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2310.01236.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2303.05499.

Aaron Lou and Stefano Ermon. Reflected diffusion models. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2304.04740.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2501.09732.

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! *Transactions on Machine Learning Research (TMLR)*, 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 (CVPR)*, 2024.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Qwen Team. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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 (NeurIPS)*, 2023.

Sarah Rastegar, Violeta Chatalbasheva, Sieger Falkena, Anuj Singh, Yanbo Wang, Tejas Gokhale, Hamid Palangi, and Hadi Jamali-Rad. InfSplign: Inference-time spatial alignment of text-to-image diffusion models. *arXiv preprint arXiv:2512.17851*, 2025.

Parham Rezaei, Arash Marioriyad, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Why settle for mid: A probabilistic viewpoint to spatial relationship alignment in text-to-image models. *Transactions on Machine Learning Research (TMLR)*, 2025.

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

Yifei Shen, Xinyang Jiang, Yifan Yang, Yezhen Wang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2501.06848.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. In *International Conference on Machine Learning (ICML)*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Stability AI. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, 2024.

Kota Sueyoshi and Takashi Matsubara. Predicated Diffusion: Predicate logic-based attention guidance for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting Stable Diffusion using cross attention. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Tongyi-MAI Team. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. Model: https://huggingface.co/Tongyi-MAI/Z-Image-Turbo.

Binxu Wang, Jingxuan Fan, and Xu Pan. Circuit mechanisms for spatial relation generation in diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2601.06338.

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 (TPAMI)*, 48(7):7692–7706, 2026. arXiv:2411.18301.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation rectification in diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2403.20249.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, Yao Lu, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.

Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi. Guidance with spherical Gaussian constraint for conditional diffusion. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.03201.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sapir Esther Yiflach, Yuval Atzmon, and Gal Chechik. Data-driven loss functions for inference-time optimization in text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3525–3535, 2026. arXiv:2509.02295.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Gaoyang Zhang, Bingtao Fu, Qingnan Fan, Qi Zhang, Runxing Liu, Hong Gu, Huaqi Zhang, and Xinguo Liu. CoMPaSS: Enhancing spatial understanding in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Huixuan Zhang and Xiaojun Wan. R-Bind: Unified enhancement of attribute and relation binding in text-to-image diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6856–6870, 2025.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Chenyi Zhuang, Ying Hu, and Pan Gao. Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

A Extended discussion and analysis

A.1 Related work in detail

Spatial guidance on attention. Cross-attention maps localise the prompt's words (Hertz et al., 2023; Tang et al., 2023), and training-free guidance pushes the latent with the gradient of a loss on them (Bansal et al., 2024; Epstein et al., 2023). Layout Guidance (Chen et al., 2024) confines a token's map to a given box by a backward loss and a forward re-weighting of the attention, and reports that the forward form alone does not raise object accuracy. Prompt-only methods read the target from the relation word: STORM (Han et al., 2025) transports each object's map toward a target region under a cost with a non-overlap term and refines the latent for several iterations at chosen steps; PSG (Rezaei et al., 2025) maximises the squared probability of the stated order under the two maps' marginals projected on the relation axis, or searches over initial noise; InfSplign (Rastegar et al., 2025), a preprint that we cite but do not place in our tables, adds a margin hinge on the centroid gap to variance-based presence and balance terms with fixed weights; Noise Refinement (Izadi et al., 2025) sums equally weighted losses for missing, mixing, binding and a sigmoid of the centroid difference, with a language-model parse and a verifier that triggers a second generation; Learn-to-Steer (Yiflach et al., 2026) replaces the hand-written loss by a relation classifier trained per backbone. Others change the model or its input: RRNet (Wu et al., 2024) trains a network per relation to edit the end-of-text embedding, SPRIGHT (Chatterjee et al., 2024a) and CoMPaSS (Zhang et al., 2025) fine-tune on spatially descriptive data, REVISION (Chatterjee et al., 2024b) renders a reference scene, and EnMMDiT (Wei et al., 2026) optimises the latent of a joint-attention transformer against subject mixing. A mechanistic study of small diffusion transformers finds that, under a contextual text encoder, the relation is carried by the noun tokens rather than by the relation word and is implemented by a few early cross-attention heads (Wang et al., 2026), which is where our relation condition reads it. Our relation statistic is PSG's marginal probability, logged instead of squared, and our centroid form is InfSplign's hinge; what changes is their role: the relation is a condition with a multiplier that starts at zero for a grader that reads a sign and an objective term for one that reads boxes, and its margin is derived (Thm. 2). STORM's non-overlap cost and PSG's distance loss push two maps apart without a stopping point, where our repulsion is a hinge that vanishes beyond the derived gap; InfSplign's presence term is the trace of a map's covariance, where ours is the entropy power, which charges area and not shape (Prop. 3), with a floor on the share, a statement about allocation rather than spread.

Presence and attribute binding. Attend-and-Excite (Chefer et al., 2023) raises the neglected subject's peak attention; Divide-and-Bind (Li et al., 2023b) maximises a map's total variation and binds attributes with a Jensen-Shannon term; A-STAR (Agarwal et al., 2023) penalises the overlap of two subjects' maps and their decay across steps; SynGen (Rassin et al., 2023) pulls a modifier's map to its noun's and away from other words under a dependency parse, and CONFORM (Meral et al., 2024) states the same contrastively; InitNO (Guo et al., 2024) optimises the initial noise on a peak score and a self-attention conflict score; Predicated Diffusion (Sueyoshi & Matsubara, 2024) derives fuzzy-logic losses from propositions, of which Attend-and-Excite is a special case; Attention Overlap (Marioriyad et al., 2025) shows that the overlap of two entities' maps predicts a missing entity better than a map's intensity or spread, and minimises it; R-Bind (Zhang & Wan, 2025) adds an entity-relation-entity term under a language-model parse; ToMe (Hu et al., 2024) and Magnet (Zhuang et al., 2024) act on the text side, merging a noun with its attributes into one token or shifting its embedding along binding directions; Make It Count (Binyamin et al., 2025) fixes the number of instances with a learned layout. Every one of these states where a map should lie or how peaked it should be, which in the decomposition of Sec. 4.1 acts on localization, and the start token's hold on the budget, noted by Layout Guidance and excluded from the statistics of Attend-and-Excite and InitNO, is left where it is. Mordant regulates allocation directly, with a share floor and a forward re-allocation that moves the start token's probability to the object at the cells the object already dominates, without a layout, and states separation and binding as likelihood-ratio tests against the uniform map, which keep a gradient where an overlap statistic has none.

Search and selection. Inference-time compute can also be spent on candidates: best-of-N over initial noise scored by a verifier (Ma et al., 2025), reward-based noise optimisation (Eyring et al., 2024), PSG's search variant, and Feynman-Kac steering (Singhal et al., 2025), which resamples

Table 5: **What each training-free method regulates.** *Alloc.*: the share of the attention budget a concept holds against the start token and the prompt's other words; every other presence statement concerns where attention lies. *Weights*: how the terms are combined. Extra input: what a method needs beyond the prompt and the frozen backbone.

Method	Venue	Extra input	Alloc.	Presence acts on	Relation	Binding	Count	Weights
Attend-and-Excite	*TOG'23*	–	✗	peak	–	–	–	single
A-STAR	*ICCV'23*	–	✗	overlap, decay	–	–	–	fixed
Divide-and-Bind	*BMVC'23*	–	✗	total variation	–	JSD	–	fixed
SynGen	*NeurIPS'23*	parser	✗	–	–	sym. KL	–	fixed
CONFORM	*CVPR'24*	token groups	✗	contrast	–	contrast	–	single
InitNO	*CVPR'24*	–	✗	peak (noise)	–	–	–	fixed
Predicated Diffusion	*CVPR'24*	parser	✗	existence	–	implication	–	fixed
Attention Overlap	*TMLR'25*	–	✗	overlap	–	–	–	single
ToMe, Magnet	*NeurIPS'24*	parser	✗	–	–	text embedding	–	–
PSG	*TMLR'25*	–	✗	–	order prob.	–	–	single
STORM	*CVPR'25*	–	✗	–	transport	transport	–	fixed
InfSplign	*arXiv'25*	–	✗	spread	centroid hinge	–	–	fixed
Noise Refinement	*arXiv'25*	LLM, verifier	✗	peak, overlap	centroid	overlap	–	equal
R-Bind	*EMNLP'25*	LLM parse	✗	focus	ent.-rel.-ent.	ent.-attr.	–	fixed
Learn-to-Steer	*WACV'26*	classifier	✗	–	learned	–	–	learned
Mordant (ours)	–	–	✓	share, entropy power	prob. / centroid	likelihood ratio	instances	multipliers

particles under a reward potential and reports GenEval on the backbones we use. These need several trajectories per image and a scorer; Mordant runs one trajectory and reads only the model's attention.

Constrained formulations. Training-free guidance is usually a fixed-weight penalty sum (Bansal et al., 2024; Yu et al., 2023; Ye et al., 2024; Shen et al., 2024), at times with the step projected back toward the data manifold (He et al., 2024; Yang et al., 2024), and energies are composed at sampling time as products of experts or with Monte Carlo correctors (Du et al., 2023; Wu et al., 2023; Skreta et al., 2025). Constraints proper appear on the sample's support, by reflection or a mirror map (Lou & Ermon, 2023; Liu et al., 2023); on linear measurements, by projection (Jayaram et al., 2024); at training time, where a Lagrangian dual reshapes the learned distribution (Khalafi et al., 2024); and in Langevin sampling under expectation constraints with primal-dual dynamics (Chamon et al., 2024). We know of no compositional guidance method that states its requirements as constraints with multipliers found along the trajectory. Mordant does, with the augmented Lagrangian (Bertsekas, 1982; Nocedal & Wright, 2006) on attention statistics: a condition the model already meets never acts, and rescaling a condition changes only its multiplier (Prop. 1), whereas a penalty sum's optimum moves with every weight. Tab. 5 summarises the comparison.

Where the bias comes from. The start token's sink is the image-side counterpart of attention sinks in language models (Xiao et al., 2024); PSG observed that the second-named object is drawn less often, SPRIGHT that a missing word's attention is spread over the image, and Attention Overlap that the overlap of two entities' maps predicts a missing entity better than a map's intensity or spread; CLIP's text encoder encodes the first-mentioned object most strongly (Abbasi et al., 2025). Our measurements quantify the sink (84 to 89% of the probability, 38 to 48% of what renders), show the order bias on a class-balanced set and the share's decay during sampling, and place both in the allocation-localization identity (Sec. 4).

A.2 The relation follows what the grader reads

Eq. (6) asks the subject's centre to lie at least m past the object's along the relation's direction d, a unit vector read from the relation word ($d = (-1, 0)$ for *left of*, $(0, -1)$ for *above*, with y pointing down); for *left of* it is $x_O - x_S \geq m$. *Near* asks $\|\mu_S - \mu_O\| \leq m$, and depth relations (*in front of*, *behind*) ask for the vertical order, the monocular cue that the nearer object sits lower in the frame, together with nearness, since an occluder overlaps what it hides.

The statement is the same on every benchmark; the statistic of the map that stands in for the grader's reading is not. A grader that reads the *sign of a centre offset* (VISOR) is served by Eq. (6) and two conditions on the soft edges of the above-mean maps $\tilde{p} = [p - \bar{p}]_+ / \sum [p - \bar{p}]_+$: along d the subject's far edge lies before the object's near edge, and across d the two extents overlap, with $\mathrm{hi} = \nu_2/\nu_1$ and $\mathrm{lo} = (\nu_1 - \nu_2)/(1 - \nu_1)$ for a marginal with moments ν_1, ν_2. A grader that reads

Algorithm 1: Mordant sampling (one image)

Input: prompt y with objects, counts, attributes and relation; frozen denoiser ϵ_θ; T=50 steps, guided steps $t < 25$

$z_T \sim \mathcal{N}(0, I)$; $\lambda_c \leftarrow 0$ for every condition c;
for $t = T, \dots, 1$ **do**
  run $\epsilon_\theta(z_t, t, y)$ with the re-allocation of Eq. (9), capturing the cross-attention maps;
  $\hat{\epsilon}_t \leftarrow$ classifier-free-guided prediction;
  if *t is a guided step* **then**
    read shares, maps and moments (Eq. (2)); evaluate F and the conditions g_c on the hooked layers;
    $\hat{\epsilon}_t \leftarrow \hat{\epsilon}_t + w\,\nabla_{z_t}\mathcal{L}$ with $\mathcal{L}$ from Eq. (8);
    $\lambda_c \leftarrow [\lambda_c + \rho\,\bar{g}_c]_+$ for every condition c;
  $z_{t-1} \leftarrow$ scheduler step$(z_t, \hat{\epsilon}_t)$;
return decode(z_0)

boxes (T2I-CompBench's overlap rule, GenEval's size-discounted offset) is served by the surprisal of the stated order under the maps' marginals, added to the objective,

$$-\log P\big(X_S + m < X_O\big), \qquad P\big(X_S + m < X_O\big) = \sum_{x,x'} p_S^{\mathrm{x}}(x)\, p_O^{\mathrm{x}}(x')\, \mathbb{1}[x + m < x'], \quad (10)$$

computed exactly on the grid, with a proximity kernel in place of the order for *near* and the vertical order for depth. A hinge on centres stops once the centres are ordered; the surprisal keeps pulling until the two marginals no longer overlap at the margin, which is what a box grader pays for and what a sign grader does not (measured in Sec. 6.3).

What is set by hand. The margin $m = \frac{1}{3}$ is derived (Thm. 2), and the separation and binding thresholds are the uniform map. The dual step $\rho = 1$, the guidance weight $w = 300$ (200 when the relation enters the objective as Eq. (10)) and the objective's weight $\kappa = 5$ were selected on prompts disjoint from every evaluation prompt (Sec. G); the floor's β and λ_f and the re-allocation δ are listed with every other setting in Sec. D.

A.3 Why the loss takes this form

Three properties motivate the form of the loss: its conditions are scale-free where penalties are not, the relation margin is a third, and the entropy power, unlike the trace, prefers no shape. Proofs are in Sec. B.

Proposition 1 (conditions are scale-free, penalties are not)**.** *Fixed points of the dual ascent $\lambda_c \leftarrow [\lambda_c + \rho g_c]_+$ at a stationary point of $\mathcal{L}$ satisfy the KKT conditions of the program. Rescaling a condition, $g_c \to a_c g_c$ with $a_c > 0$, leaves the feasible set, the KKT points and the ascent's fixed points unchanged (the multiplier becomes λ_c / a_c), whereas the minimiser of a fixed-weight penalty sum $F + \sum_c \gamma_c g_c$ moves with every a_c.*

A fixed-weight recipe's optimum is a property of its weights, a constrained program's a property of its statements. This is the formal counterpart of a measurement: in the fixed-weight version of our method, the sign of a term's effect flipped as other terms were strengthened (Sec. J.2).

Theorem 2 (the margin is a third)**.** *For k objects ordered along an axis of the unit interval, with margin m required between neighbours and between each end object and the frame, the system $x_1 \geq m$, $x_{i+1} \geq x_i + m$, $x_k \leq 1 - m$ is feasible if and only if $m \leq 1/(k+1)$. At $m = 1/(k+1)$ its solution is unique, $x_i = i/(k+1)$, and it maximises the smallest of the $k+1$ gaps; for two objects $m = \frac{1}{3}$.*

Proposition 3 (the trace is a roundness prior)**.** *For $\Sigma \succ 0$ in two dimensions with $\sqrt{\det \Sigma}$ fixed, $\operatorname{tr} \Sigma \geq 2\sqrt{\det \Sigma}$ with equality iff Σ is isotropic; on isotropic maps the two agree in value and gradient.*

Minimising the trace therefore drives every map toward a disc at whatever area it has reached, the shape a round word names and not the one a triangular word names; the entropy power has no preferred shape.

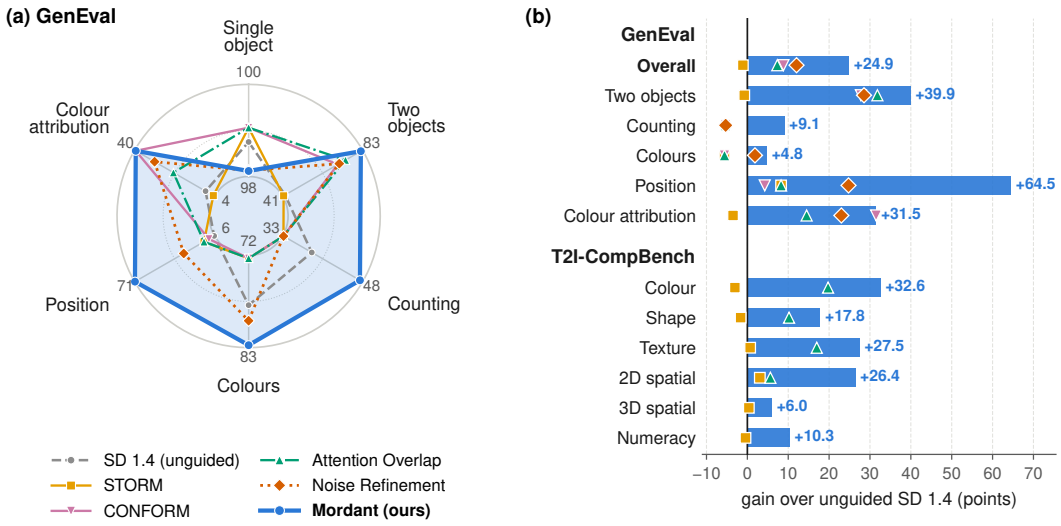


Figure 5: **Mordant against the methods we ran** on SD 1.4, all in our harness. **(a)** The six GenEval tasks, each axis scaled separately between the whole-percent bounds printed at the inner ring and the rim; **(b)** gain over unguided SD 1.4 in points, Mordant as bars and the other methods as markers (T2I-CompBench only for the methods run on it).

Table 6: **Controllability of the layout** (balanced VISOR, images with both objects found, OWL-ViT boxes; nats). $I(R;Y)$: information between the asked relation and the one rendered; $H(Y)$: entropy of the rendered relation. Both-found against the product of the two per-object rates tests whether the objects render independently.

	both found	product	$P(Y{=}R \mid \text{both})$	$H(Y)$	$I(R;Y)$
SD 1.4	0.33	0.43	0.27	1.39	0.01
Mordant	0.80	0.80	0.93	1.39	1.07

A.4 Summary against the methods we ran

Fig. 5 summarises the comparison on SD 1.4: among the methods we ran from their authors' released code, Mordant has the highest value on GenEval overall and on every GenEval task except single-object (where all methods are within 0.01 of each other), tying CONFORM on colour attribution, and the largest gain over the unguided model on every T2I-CompBench category shown.

A.5 Analysis: controllability and presence

Tab. 6 measures how much of the layout the prompt controls on the balanced VISOR images in which both objects are found, reading the relation from the detector's boxes. Unguided, the asked relation carries 0.01 nats about the rendered one ($P(Y{=}R \mid \text{both}) = 0.27$, chance 0.25): the layout is set by the prior. Under Mordant the mutual information is 1.07 of the 1.39 nats available and the relation is read correctly in 92.7% of the images with both objects. The dispersion of the detected centres is the same as the unguided model's (standard deviations of 0.15 to 0.19 of the frame in both), so the relation is set without collapsing where the pair sits.

A.6 Cost

Mordant adds one backward pass through the U-Net at each of the 25 guided steps; the dual update is a handful of scalars per image and the forward re-allocation adds nothing measurable. Timed at batch size 1 on the same GPU, unguided SD 1.4 generates an image in 1.30 s at a peak of 3.2 GiB of memory and Mordant in 4.35 s at 6.5 GiB, that is 3.3× the unguided time and 2× its memory (the fixed-weight form of Sec. 6.3 takes 3.85 s), with no training, no second generation and no auxiliary

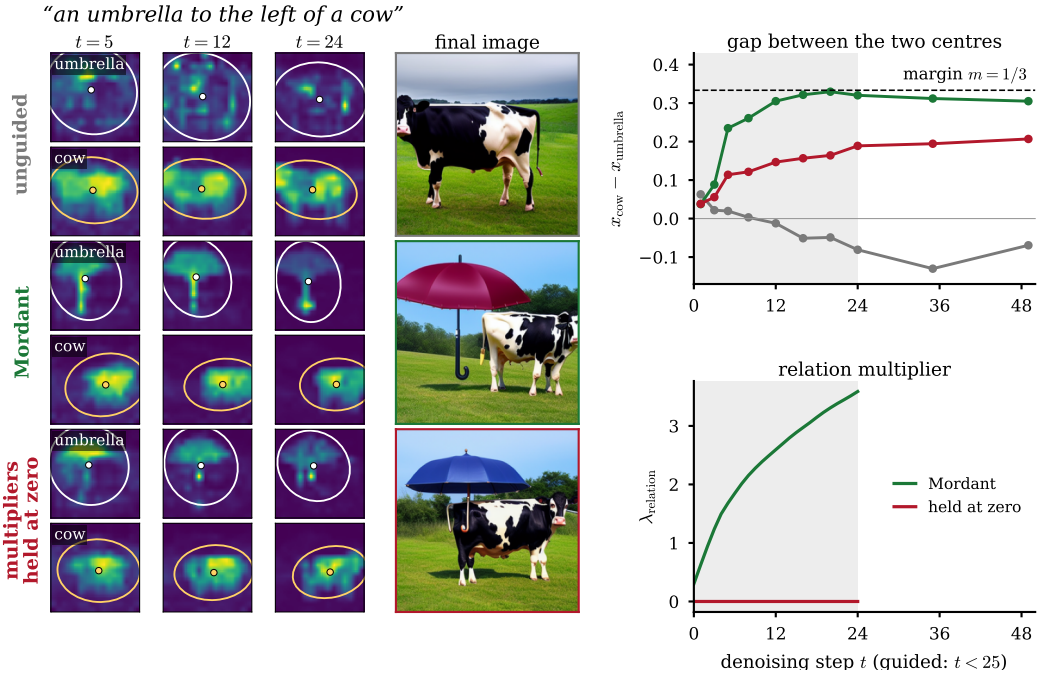


Figure 6: **How the conditions act** (SD 1.4, *an umbrella to the left of a cow*). Left: each word's attention at three steps, with its centre and one-standard-deviation ellipse, and the final image, for three arms from the same noise. Right: the gap between the two centres against the margin $m = 1/3$, and the relation's multiplier; held at zero, the same condition acts as a fixed penalty and gets about half as far.

model. The competitors' released code, timed the same way in the fp32 their pipelines load (Tab. 24), takes 4.1 s per image for Attention Overlap, 5.1 s for Attend&Excite, 6.5 s for Divide&Bind and for CONFORM, 11.5 s for InitNO, 13.8 s for Noise Refinement and 11 to 17 s for STORM, the slowest, whose pipeline takes 2.2 s with its guidance switched off: measured against its own unguided pipeline, STORM's guidance multiplies the time per image by 5.0 to 7.7, Mordant's by 3.3. Methods that call a language model to parse the prompt (Izadi et al., 2025) add that model's cost on top; Mordant reads only the attention maps the model already computes.

How the conditions act. Fig. 6 reads the cross-attention maps along the trajectory for one prompt the unguided model fails, under the frozen model, under Mordant, and under the same conditions with every multiplier held at zero, i.e. as the fixed quadratic penalties a fixed-weight method applies. Unguided, the umbrella's map drifts to the wrong side of the cow's and no umbrella is drawn; under the program the two centres separate and come within 0.02 of the derived margin $m = \frac{1}{3}$ by $t = 16$, while the multiplier of the order condition grows for as long as the gap is short of it; with the multipliers held at zero the same penalty moves the pair about half as far and the umbrella is drawn over the cow. Sec. M shows the same reading for the apart term (a second object that fragments without it), the binding condition (a colour and a material word that stay on their own noun with it and straddle both nouns without it), and the multipliers of every condition over the guided steps.

A.7 Further measurements of the allocation failure

The concept named first is favored. On the T2I-CompBench 2D-spatial prompts whose two objects are both in the grader's vocabulary (198 prompts, 1,980 images), the unguided model draws the first-named object in 81.7% of images and the second-named one in 63.5%. The gap is not a property of the objects: on our balanced VISOR set, in which every class appears equally often in each position (1,000 prompts, 4,000 images), the first-named object is still found more often than the second (69.1% against 61.7%). The text encoder is known to encode the first-mentioned object of a

caption most strongly (Abbasi et al., 2025); the diffusion model inherits the preference as a gap in presence.

The unfavored concept is starved during sampling. Shares are not fixed; they drift as the image commits. For *a green bench to the left of a blue bowl*, the two nouns hold 0.046 and 0.037 of the first coarse layer's budget at the first step, and 0.0095 and 0.0041 by step 24 of 50, when the layout has settled. The model then draws the bench and no bowl. The object that loses its share early is not drawn in the wrong place; it is not drawn.

Placement is not the bottleneck. With spatial guidance, the relation between two objects is almost never wrong once both are drawn: on VISOR our guided model places the pair correctly in 99.97% of the images that contain both (Sec. 6.2), and the failures that remain are images in which one object is missing.

A.8 Separation and binding as likelihood-ratio tests

Since $H(p, q) = H(p) + D_{\mathrm{KL}}(p\|q)$ and $\log n = H(p) + D_{\mathrm{KL}}(p\|\mathcal{U})$, Eq. (5) is $D_{\mathrm{KL}}(p_i\|p_j) \geq D_{\mathrm{KL}}(p_i\|\mathcal{U})$: object j's map explains object i's no better than the uniform map does. Eq. (7) asks an attribute's map to be explained better by its own noun's map than by any other object's or by the uniform map; its gradient, $\log(p_j/p_{o(a)})$ with respect to p_a, moves the attribute toward cells where its noun dominates the rival at any distance, which an overlap statistic cannot do when the maps are disjoint. Neither carries a threshold.

A.9 Limitations in full

Mordant regulates what the frozen model already represents. Re-allocating attention to a concept helps only if the model can render that concept: an object the backbone draws badly unguided is placed and kept, not improved, and the classes that remain hardest on VISOR are those whose rendering, or whose name, is unstable in the backbone. Attribute binding is stated on where an attribute attends, so it moves colour, a property of the whole region, far more than texture or shape, whose rendering depends on how much the attribute participates at its noun's cells and on the value vectors, which the maps do not read. Counting enters only through the instance nodes of a counted noun, which keep instances distinct but state no count; surplus instances, the base model's most common counting error, are not removed, and on SDXL the guided model counts worse than the unguided one (Tab. 1). The multipliers are updated once per sampling step, so the sampler approaches rather than reaches the program's KKT points. Which statistic stands in for the relation depends on what the grader reads (Sec. A.2); we fix it by the grader's documented rule and ablate the swap on VISOR (Sec. 6.3). Each guided step costs one backward pass through the denoiser during the first half of sampling (Sec. A.6). Our main experiments use Stable Diffusion 1.4, and the other backbones are evaluated on GenEval only (Sec. 6.2); the statements are defined on any cross-attention map, including the image-text block of a joint-attention transformer, but the guidance weight has to be rescaled for a backbone with a different prediction target or latent size, and the forward re-allocation is implemented for U-Net cross-attention only.

B Proofs

Proof of Prop. 1. The ascent's fixed point $\lambda = [\lambda + \rho g]_+$ gives $g = 0$ when $\lambda + \rho g \geq 0$ and $\lambda = 0,\ g < 0$ otherwise; in both cases $g \leq 0$, $\lambda \geq 0$ and $\lambda g = 0$. Since $\partial_z \psi(g_c, \lambda_c) = [\lambda_c + \rho g_c]_+ \nabla g_c = \lambda_c \nabla g_c$ at the fixed point, stationarity of $\mathcal{L}$ is $\nabla F + \sum_c \lambda_c \nabla g_c = 0$. Under $g_c \to a_c g_c$, $\{a_c g_c \leq 0\} = \{g_c \leq 0\}$, and (z, λ) is a KKT point for g iff (z, λ') with $\lambda'_c = \lambda_c / a_c$ is one for ag; the fixed-point conditions are the same. A stationary point of $F + \sum_c \gamma_c a_c g_c$ solves $\nabla F + \sum_c \gamma_c a_c \nabla g_c = 0$, a different equation for every $a_c \neq 1$. □

Proof of Thm. 2. The $k + 1$ gaps $x_1,\ x_2 - x_1, \ldots, 1 - x_k$ are each at least m and sum to 1, so $(k + 1)m \leq 1$; $x_i = i/(k + 1)$ is feasible for every $m \leq 1/(k + 1)$. At $m = 1/(k + 1)$ every gap is at least m and they sum to $(k + 1)m$, so all equal m; and for any layout the smallest of $k + 1$ gaps summing to one is at most $1/(k + 1)$, with equality iff all are equal. □

Proof of Prop. 3. With eigenvalues λ_1, λ_2, $\operatorname{tr}\Sigma = \lambda_1 + \lambda_2 \geq 2\sqrt{\lambda_1\lambda_2}$ with equality iff $\lambda_1 = \lambda_2$. On $\Sigma = s^2 I$ both equal $2s^2$, and $\partial_\Sigma \operatorname{tr}\Sigma = I = \sqrt{\det\Sigma}\,\Sigma^{-1} = \partial_\Sigma 2\sqrt{\det\Sigma}$. □

C Implementation details

Backbone and sampling. Stable Diffusion 1.4 at 512×512, 50 DDPM steps, classifier-free guidance 7.5, an fp16 U-Net with every loss computed in fp32; image k of every prompt is generated from seed $42 + k$ for every method, so all comparisons are paired. Guidance acts on the first 25 of the 50 steps.

Layers. The loss reads the nine cross-attention layers of SD 1.4 at 16×16 and coarser, plus the three at 32×32: the two 16×16 layers of the last down block, the 8×8 bottleneck, the three 16×16 layers of the first up block and the three 32×32 layers of the second. Attention is averaged over heads and resized to a fixed 64×64 grid, and every reading of Eq. (2) is taken there; the floor's reference uses the layer's native grid. The objective's weight on layer l is $\kappa \min\left(1, (16/g_l)^8\right)$ with g_l the layer's native grid, which keeps the effective-area term on the layers of grid 16 and coarser, where presence is decided. The forward re-allocation of Eq. (9) acts on all sixteen cross-attention layers.

Tokens. Each object's map sums the attention of all its word pieces; a word that the tokenizer splits into several pieces forms one group, so a multi-piece word (*pentagonal*, *postbox*) does not shift the positions of the words after it. For the attribute sets, the object slot is the head noun of each noun phrase, read with a part-of-speech tagger and the benchmark's attribute vocabulary; when a prompt names the same noun twice, the k-th mention takes the k-th occurrence of its tokens.

Counts. A noun with count k is split into k instance nodes by a soft k-means of its map (temperature 0.01 on squared distances in the unit square); the effective-area term acts on each node, weighted by the node's share of the noun's map.

Relations. Tab. 7 lists the relation words and their directions. The soft edges of Sec. A.2 are computed on the above-mean map $\tilde{p}$, one marginal per axis. The surprisal form of Eq. (10) uses the exact bilinear form on the 64-cell marginal; for *near* the kernel is $\mathbb{1}[\min(|\Delta x|, |\Delta y|) \leq 0.3]\,\mathbb{1}[\max(|\Delta x|, |\Delta y|) > m]$, close on one axis and apart on the other; for depth relations it is the vertical order.

Table 7: **Relation vocabulary.** d is the direction of Eq. (6), with y pointing down.

Kind	Phrases	Statement
Order	left of, on the left of, to the left of	$d = (-1, 0)$
	right of, on the right of, to the right of	$d = (1, 0)$
	above, on top of, on the top of	$d = (0, -1)$
	below, underneath, on the bottom of	$d = (0, 1)$
	top left of, bottom right of, …	$d = (\pm 1, \pm 1)/\sqrt{2}$
Near	next to, near, on side of, beside	$\|\mu_S - \mu_O\| \leq m$
Depth	in front of / behind, hidden by	vertical order (nearer is lower) and nearness

Dual ascent. Every multiplier starts at zero for each image and takes one step per guided step, using the condition's mean over the hooked layers; the multipliers are per image, since a batch mixes prompts.

Floor reference. $\bar{m}_l$ is the median share of an object that a detector finds, measured once on unguided VISOR images for each layer: 0.0431, 0.0138 and 0.0097 on the three 16×16 up-block layers, 0.0124 and 0.0325 on the two 16×16 down-block layers, and 0.0003, 0.0013 and 0.0038 on the three 32×32 layers. A second measurement on a different set of prompts reproduced the six up-block values to within 0.008.

Table 8: **Every setting of Mordant on SD 1.4** and where it comes from. *Held-out*: selected on the tuning prompts of Sec. G, which share no prompt with any evaluation set.

Component	Setting	Value	Origin
Relation	margin m	$\frac{1}{3}$	derived (Thm. 2)
Conditions	separation, binding thresholds	uniform map	derived (likelihood-ratio tests)
Solver	dual step ρ	1	held-out; sharp optimum
	guidance weight w	300; 200 when the relation is in the objective	held-out
Objective	weight κ (effective area, apart)	5	held-out
	area floor s	one grid cell	fixed
Allocation	floor β, weight λ_f	0.4 of $\bar{m}_l$, 0.25	$\bar{m}_l$ measured per layer; β, λ_f from the fixed-weight form
	re-allocation δ, ϵ_0	0.5, 0.05	fixed
Schedule	guided steps; layers	first 25 of 50; nine (Sec. C)	fixed
Sampling	steps / scheduler / CFG / precision	50 / DDPM / 7.5 / fp16 U-Net, fp32 losses	backbone default

D SETTINGS

E EVALUATION PROTOCOL

Sampling. Every image we generate, ours and every competitor's, comes from Stable Diffusion 1.4 at 512×512 with 50 denoising steps and classifier-free guidance 7.5, except the backbone study of Tabs. 1 and 16 (Sec. E). Our rows (the unguided model, Mordant and its ablations) use the DDPM scheduler with an fp16 U-Net and every loss in fp32; Mordant's gradient acts on the first 25 of the 50 steps. Competitor rows use the scheduler and precision that each released pipeline loads by default, fp32 with the PNDM scheduler for the SD 1.4 pipelines of the Attend-and-Excite family (Sec. K), which is why every competitor is read against its own switched-off control rather than against our fp16 base row. Image k of a prompt is drawn from a generator seeded with $42 + k$, for every method and every benchmark, so each comparison is paired image by image. Images per prompt: 4 on VISOR (seeds 42 to 45), 4 on GenEval (42 to 45) and 10 on T2I-CompBench (42 to 51).

Prompt sets. *GenEval*: the benchmark's 553 prompts (80 single-object, 99 two-object, 80 counting, 94 colours, 100 position, 100 colour attribution), 2,212 images. *T2I-CompBench*: the 300 evaluation prompts of each of the eight categories (colour, shape, texture, 2D spatial, 3D spatial, numeracy, non-spatial, complex), 24,000 images; the colour file is the benchmark's 2023 release, in which 5 prompts are duplicated, so a colour run holds 2,950 images of 295 distinct prompts. *VISOR*: the balanced 1,000-prompt subset of Sec. F, 4,000 images. The T2I 2D grader compares a prompt noun to UniDet's label strings verbatim; 102 of its 300 prompts contain a noun that no label matches (*wallet*, or *cow*, whose label is the string "cattle, cow"), so those prompts score zero for every method and the attainable score is capped near 0.66. We keep the upstream rule for comparability with published numbers.

Scorers. Every number is produced by the benchmarks' own evaluation code at the following versions. *GenEval* (Ghosh et al., 2023): its Mask2Former Swin-S detector trained on COCO (Cheng et al., 2022) at the benchmark's confidence threshold, and its zero-shot CLIP ViT-L/14 colour classifier (Radford et al., 2021) over the ten colour words on the detected object's crop; a task's score is the mean over its images of a per-image boolean, and the overall score is the unweighted mean of the six task scores. *T2I-CompBench* (Huang et al., 2023): BLIP-VQA (Li et al., 2022) for colour, shape and texture, the product over the prompt's noun phrases of the probability of "yes"; UniDet (Zhou et al., 2022) for 2D spatial relations, with the benchmark's spaCy noun extraction, its exact match between a prompt noun and a detector label and its scoring rule (a pair placed wrongly or a single detected object scores zero); UniDet with the R50 backbone and a DPT depth map (Ranftl et al., 2021) for 3D spatial relations, read at the benchmark's own crop; UniDet R50 for numeracy, where half of each object's credit is presence and half the exact count; CLIPScore (Hessel et al., 2021)

with ViT-B/32 for non-spatial prompts; and the benchmark's 3-in-1 combination of BLIP-VQA, the 2D scorer and CLIPScore for complex prompts. *VISOR* (Gokhale et al., 2022): OWL-ViT (Minderer et al., 2022), queried with the prompt's two class names, and VISOR's rule on the sign of the centroid offset; object accuracy (OA) is the fraction of images with both objects detected, $\text{VISOR}_{\text{cond}}$ the fraction of those with the relation correct, $\text{VISOR}_{\text{uncond}}$ their product, and VISOR_N the fraction of prompts with at least N of their images correct. For the detector-robustness ablation (Tab. 10) we re-detect every VISOR image with GenEval's closed-set Mask2Former, which sees exactly the 80 COCO classes VISOR names, and with the open-vocabulary Grounding DINO (Liu et al., 2024) queried with the two class names, and apply VISOR's rule to their boxes, so that no conclusion depends on one detector.

Object words and token groups. Two properties of the tokens a guidance method is handed are part of the protocol, because they decide which map the loss reads. First, the object words distributed with the T2I-CompBench prompt files are not always the head of the noun phrase the grader asks about: in 120 texture and 59 shape prompts the listed object is the attribute word itself (*rubber*, *oval*), in others a modifier of the head (*desk* for *a metallic desk lamp*), and in a few the same word twice. We read each noun phrase's head with a part-of-speech tagger and the benchmark's own attribute vocabulary, which corrects 28 colour, 103 shape and 147 texture records of 300, and leaves GenEval and VISOR unchanged. Second, a word that the tokenizer splits into several pieces (*pentagonal*, *postbox*, *fire hydrant*) forms one token group, so that the words after it keep their positions; 129 of the 300 shape prompts contain such a word. Both apply to every row of Mordant and, for the fixed-weight form of Sec. 6.3, to the attribute, 3D and numeracy sets; on VISOR, GenEval and 2D spatial, whose object words already equal their phrase heads and where a multi-piece word occurs in under 3% of the prompts, the fixed-weight rows were generated without them. The unguided model reads no token and is unaffected, and competitors use their authors' own token pickers (Sec. K). Measured on a fixed-weight recipe at 10 images per prompt, the two together are worth about $+0.06$ on texture, $+0.03$ on shape and $+0.01$ on colour.

Baselines. Published numbers are quoted from the papers that report them and marked as such. Where a method has not reported a benchmark, we run its released code with its default settings on the same backbone and seeds and score it with the same scorers (Sec. K), and we report its own switched-off control wherever its code allows one. STORM, CONFORM and Attention Overlap need two entities; on single-entity prompts their code samples the unguided backbone, which is what their GenEval single-object, counting and colours tasks measure.

Other backbones. Tab. 1 runs Mordant on the backbones on which the methods we compare with report results besides SD 1.4: SD 2.1 (STORM (Han et al., 2025), CoMPaSS (Zhang et al., 2025), Learn-to-Steer (Yiflach et al., 2026), RRNet (Wu et al., 2024) and Attention Overlap (Marioriyad et al., 2025)), SDXL (Podell et al., 2024) (ToMe (Hu et al., 2024), Attention Overlap, Noise Refinement (Izadi et al., 2025) and PSG (Rezaei et al., 2025)), SD3-medium and SD3.5-medium (Esser et al., 2024; Stability AI, 2024) (the similar-subject MMDiT method of Wei et al. (2026)). Each backbone is generated in both arms, unguided and guided, with identical sampler settings, and every constant of Tab. 8 is kept except the three below. *Guidance weight*: SD 1.4's $w = 300$ (200 on prompts whose relation is in the objective) multiplied by the ratio, measured once per backbone, at which the guidance perturbs the noise prediction by the same relative amount as on SD 1.4: w is 300 (200) on SD 2.1, whose latent size and prediction target equal SD 1.4's, 900 (600) on SDXL and 2400 on SD3-medium and SD3.5-medium, one weight for every prompt on the last two. *Energy reference grid*: the grid to which the objective's per-layer weighting is referenced is 32×32 on SDXL, its coarsest cross-attention grid at 1024×1024, and the single grid of the joint blocks on the MMDiT backbones, all of whose blocks see the same grid. *Allocation floor*: the per-layer share table $\bar{m}_l$ was measured on SD 1.x, so on every backbone other than SD 1.x and SD 2.1 the floor's target is derived from the prompt instead, as a fixed fraction of the share held by the prompt's content tokens divided among its entities and their attributes, which matches the measured table on a two-object prompt. SD 2.1 is the 512-pixel base checkpoint and SDXL runs at 1024×1024, both with SD 1.4's schedule (50 DDPM steps, guidance on the first 25, CFG 7.5, fp16). On the MMDiT backbones (SD3-medium, SD3.5-medium) the maps are the image-row, text-column block of the joint attention, the forward re-allocation is not applied, since it is implemented for U-Net cross-attention only, and sampling uses 28 flow-matching steps with guidance on the first 14.

Reproducing the unguided model. Generated with the settings above and scored with the official code, unguided SD 1.4 reads within 0.02 of its published T2I-CompBench scores in every category (colour 0.369, shape 0.371, texture 0.402, 2D 0.127, 3D 0.314, numeracy 0.456 against 0.377, 0.358, 0.416, 0.125, 0.303, 0.446) and 0.45 on GenEval against the published 0.43; on the same harness SD 1.5 reads 0.45 against GenEval's own 0.43, with every task within 0.06. On the balanced VISOR subset it reads OA 33.40 (Mask2Former 33.73, Grounding DINO 41.88), 3.5 points above the published value on the full benchmark (29.86).

Compute. Timings are wall-clock seconds per image at batch size 1, all measured on the same GPU. An unguided image takes 1.30 s and 3.2 GiB of memory; Mordant 4.35 s and 6.5 GiB, 3.3× the unguided time; its fixed-weight form 3.85 s and 6.5 GiB (Sec. A.6). The competitors' released pipelines, at batch size 1 in fp32 on the same GPU, take 4.1 to 17 s per image, with STORM the slowest at 11 to 17 s against 2.2 s with its guidance switched off (Tab. 24).

F THE BALANCED VISOR SUBSET

The full VISOR benchmark is exactly balanced: 80 COCO classes, each paired with the 79 others under 4 relations (25,280 prompts), so every class fills 1.25% of the object slots and every relation 25% of the prompts. The 1,000-prompt subset in common use is not a uniform sample of it: its class frequencies deviate from uniform with $\chi^2 = 306$ on 79 degrees of freedom, and the over-represented classes (*baseball bat*, *traffic light*, *hair drier*) are among those the detector finds least often, which lowers every method's object accuracy on it. We evaluate on a balanced 1,000-prompt subset drawn from the full set: 250 prompts per relation, subjects taken round-robin over the 80 classes and partners cycled with a coprime stride, which gives $\chi^2 = 5.0$ on 79 degrees of freedom and 24 to 28 slots per class. Its class and relation marginals match the full benchmark's; the unguided model's object accuracy on it (33.40) is 3.5 points above the full benchmark's published value (29.86), so we compare methods on it only through rows generated and scored in the same harness, and keep the published full-benchmark rows in a separate table (Tab. 9).

G HELD-OUT TUNING PROMPTS

Every choice made after the method's statements were fixed (the dual step ρ, the guidance weight w, the objective's weight κ, the form of the objective, and the per-grader relation statistic) was made on prompts that share nothing with an evaluation prompt. *VISOR*: 200 prompts balanced over the 80 classes ($\chi^2 = 17.6$ on 79 degrees of freedom) and the four relations (50 each), with no object pair shared with the balanced evaluation subset, 4 images each. *T2I-CompBench*: 300 prompts per category drawn from the benchmark's own 700-prompt training split, after removing the prompts that also occur in the evaluation split (0, 0, 0 and 2 for colour, shape, texture and 2D), 2 images each; these prompts served to choose constants, and every T2I-CompBench number we report is at the benchmark's 10 images per prompt. The held-out results for the chosen settings and their neighbours are in Sec. J.1.

H RUN-TO-RUN SPREAD

GPU non-determinism in the fp16 attention changes individual images between two runs of the same configuration and seeds, so every benchmark has a floor below which differences are noise. Measured by repeating identical configurations: VISOR object accuracy varies by 0.5 to 0.9 points on 2,000 images and by 1.3 to 2 points on 528; a GenEval task by about 15 images of 320 to 400 (about 0.01 on the overall score); and T2I-CompBench 2D spatial by about 0.008 at 10 images per prompt. On the 800 held-out VISOR images of Tab. 4, two runs of the full program on different machines differ by 4.2 / 0.8 / 2.5 points under OWL-ViT / Mask2Former / Grounding DINO, which is why each ablation cell is read against the run generated on its own machine. We call a difference real only when it exceeds these spreads, and say so when it does not.

I EXTENDED RESULTS

I.1 VISOR: PUBLISHED FULL-BENCHMARK ROWS

Tab. 9 lists the published VISOR results on SD 1.4. They are computed on the full 25,280-prompt benchmark, whereas every row of Tab. 3 is generated and scored in our harness on the balanced 1,000-prompt subset of Sec. F, on which the unguided model reads 3.5 points of object accuracy higher. The two tables are therefore not comparable row for row; the published rows are a guide to the scale of what each method reports, and the in-harness rows of Tab. 3 are the comparison.

Table 9: **Published VISOR results on SD 1.4**, full 25,280-prompt benchmark, as reported by Han et al. (2025) unless cited otherwise. Not comparable with Tab. 3, which is a different prompt set. Extra: what a method needs beyond the frozen model and the prompt.

Method	Extra	OA	uncond	cond	V_1	V_2	V_3	V_4
SD 1.4 (Han et al., 2025)	–	29.86	18.81	62.98	46.60	20.11	6.89	1.63
+ Composable (Liu et al., 2022)	–	23.27	14.99	64.41	39.44	14.56	4.84	1.12
+ StructureDiff. (Feng et al., 2023)	–	28.65	17.87	62.36	44.70	18.73	6.57	1.46
+ Attend&Excite (Chefer et al., 2023)	–	42.07	25.75	61.21	49.29	19.33	4.56	0.08
+ Divide&Bind (Li et al., 2023b)	–	46.03	31.62	68.70	64.72	37.82	18.64	5.30
+ InitNO (Guo et al., 2024)	–	60.40	35.18	58.24	71.20	42.71	20.09	6.72
+ CONFORM (Meral et al., 2024)	–	60.73	38.48	62.33	73.01	45.82	25.57	9.52
+ STORM (Han et al., 2025)	–	61.01	**57.58**	94.39	**85.93**	**69.71**	49.01	25.70
+ REVISION (Chatterjee et al., 2024b)	layout	53.96	52.71	**97.69**	77.79	61.02	44.90	27.15
+ CoMPaSS (Zhang et al., 2025)	fine-tuning	**65.56**	57.41	87.58	83.23	67.53	**49.99**	**28.91**

I.2 VISOR: DETECTOR ROBUSTNESS, PER RELATION, AND PRESENCE

Tab. 10 is the detector-robustness ablation of Sec. 6.3: the images of Tab. 3 read with the three detectors of Sec. E, VISOR's OWL-ViT, the closed-set Mask2Former and the open-vocabulary Grounding DINO, with VISOR's rule applied to each detector's boxes, and object accuracy split into the two per-object presence rates. Under every detector the relation is correct in more than 99% of the images that contain both objects once the model is guided, so what separates the rows is presence. The product of the two per-object rates tests whether the two objects render independently: for the unguided model the product exceeds the both-found rate under both detectors that report per-object rates, so the two objects compete; guided, the two coincide. Tab. 11 gives object accuracy per relation.

Table 10: **VISOR under three detectors** (the images of Tab. 3, VISOR's rule applied to each detector's boxes). OA: both objects detected; cond: relation correct given both objects; first / second: fraction of images in which the first- and the second-named object is detected; product: first $\times$ second, the both-found rate that independence would give; second-missing: of the images in which exactly one of the two objects is detected, the share missing the second-named one.

Detector	Method	OA	cond	first	second	product	second-missing
OWL-ViT	SD 1.4	33.40	63.77	0.69	0.62	0.43	56.00%
	Mordant	**79.88**	**99.97**	**0.91**	**0.88**	0.80	58.00%
Mask2Former	SD 1.4	33.73	61.97	0.69	0.63	0.43	55.00%
	Mordant	**79.60**	**99.78**	**0.90**	**0.89**	0.80	50.00%
Grounding DINO	SD 1.4	41.88	56.72	–	–	–	–
	Mordant	**91.65**	**99.89**	0.96	0.95	0.92	58.00%

I.3 T2I-COMPBENCH: ALL EIGHT CATEGORIES, CONTROLS AND TOKEN SOURCES

Tab. 12 extends Tab. 2 with the non-spatial and complex categories, STORM's own switched-off control and its three token sources (Sec. K.1). Two facts about the two last categories: non-spatial (CLIPScore) and complex separate almost nothing on this backbone, moving by less than 0.01

Table 11: **VISOR object accuracy per relation**, balanced subset (250 prompts per relation), under the three detectors.

Detector	Method	above	below	left of	right of
OWL-ViT	SD 1.4	34.30	29.60	33.30	36.40
	Mordant	**79.70**	**77.00**	**82.00**	**80.80**
Mask2Former	SD 1.4	34.20	32.80	32.40	35.50
	Mordant	**78.50**	**77.40**	**80.90**	**81.60**
Grounding DINO	SD 1.4	42.80	41.00	39.80	43.90
	Mordant	**91.60**	**90.30**	**92.80**	**91.90**

between the unguided model, STORM and its control, and the published T2I-CompBench++ values for them (0.77 and 0.65 for SD 1.4) are GPT-4V scores on a different scale from the CLIP and 3-in-1 numbers below, so we quote the benchmark's original 0.3079 and 0.3080 instead.

Table 12: **T2I-CompBench on SD 1.4, all categories** (300 prompts per category, 10 images per prompt; scorers as in Sec. E). Rows marked † are generated and scored in our harness; STORM's control is its own pipeline (fp32, PNDM) with the guidance switched off. STORM's token sources: the benchmark's object words with spaCy adjectives (the row of Tab. 2), its own spaCy rule for nouns and adjectives, and the benchmark's object and attribute words with spaCy filling the gaps. [c]: scored with the CLIP component of the complex metric only, before the 3-in-1 scorer was in place, and not comparable with the 3-in-1 column. Unmarked rows are published; Extra as in Tab. 2; SD 1.4's non-spatial and complex are the benchmark's original CLIP and 3-in-1 values.

Method	Extra	Colour	Shape	Texture	2D	3D	Numeracy	Non-spat.	Complex
SD 1.4 (Huang et al., 2023; 2025)	–	0.38	0.36	0.42	0.12	0.30	0.45	**0.31**	0.31
+ STORM†, guidance switched off (fp32, PNDM)	–	–	0.36	0.41	0.13	0.31	0.45	**0.31**	0.29[c]
+ Attention Overlap†	–	0.57	0.47	0.57	0.18	–	–	–	–
+ STORM†: benchmark nouns, spaCy adjectives	–	0.34	0.35	0.41	0.16	0.32	0.45	**0.31**	0.29[c]
+ STORM†: its own spaCy rule	–	–	0.36	0.42	–	–	–	–	–
+ STORM†: benchmark nouns and attributes	–	–	0.35	0.40	–	–	–	–	–
+ STORM (Han et al., 2025)	–	0.65	**0.60**	**0.75**	0.16	–	–	–	–
+ REVISION (Chatterjee et al., 2024b)	layout	–	–	–	0.33	–	–	–	–
+ CoMPaSS (Zhang et al., 2025)	fine-tuning	0.49	0.43	0.53	0.34	–	–	**0.31**	–
+ Mordant	–	**0.69**	0.55	0.68	**0.39**	**0.37**	**0.56**	**0.31**	**0.33**

† Reproduced using the official released code; other rows as published.

SDXL. ToMe (Hu et al., 2024) reports T2I-CompBench attribute binding on SDXL only, against its own SDXL baseline; Tab. 13 places those published rows beside the benchmark's own published SDXL row. We evaluate Mordant on SDXL with GenEval (Tab. 1) and have not run T2I-CompBench on SDXL at the benchmark's protocol, so the table is context for the attribute numbers of Tab. 2, not a comparison; the two published SDXL baselines differ by 0.05 on colour, which is why every comparison we draw is against a baseline generated in the same harness.

Table 13: **SDXL attribute binding** on T2I-CompBench, published rows only: the benchmark's SDXL row (Huang et al., 2025) and ToMe (Hu et al., 2024), whose gains are reported against its own SDXL baseline (0.64 / 0.54 / 0.56). Extra: what a method needs beyond the frozen model and the prompt.

Method	Extra	Colour	Shape	Texture
SDXL (Huang et al., 2025)	–	0.59	0.47	0.53
+ ToMe (Hu et al., 2024)	–	**0.77**	**0.61**	**0.69**

I.4 GenEval: controls and guided fractions

Tab. 14 adds to Tab. 1 STORM's own switched-off control, the Noise Refinement row with its four failed images excluded rather than counted as failures, and the number of images each method actually guided. Half of GenEval's prompts name one entity, and STORM, CONFORM and Attention

Overlap cannot form their loss on them, so their pipelines sample the unguided model there; the single-object, counting and colours scores of those three rows are identical because they are the same fallback images. Read against its own control, STORM gains 0.026 overall (position $+0.115$, two objects $+0.066$) and loses on colour attribution; read against our fp16 base row it would appear to lose 0.011, which is the precision and scheduler difference, not the method (Sec. K).

Table 14: **GenEval on SD 1.4, extended** (553 prompts, 4 images each, GenEval's official scorer). Guided: images generated with the method active, of 2,212; the rest are the method's own plain-SD fallback. Noise Refinement (excl.): the same run with the 4 images its code failed on excluded from the mean instead of counted as failures. Extra as in Tab. 1.

Method	Extra	Guided	Single	Two	Count	Colours	Position	Attr.	Overall
SD 1.4†	–	0	**0.99**	0.42	0.38	0.78	0.06	0.08	0.45
+ Attend&Excite†	–	2116	0.98	0.62	0.33	0.75	0.11	0.19	0.50
+ InitNO†	–	–	0.97	0.71	0.32	0.75	0.11	0.25	0.52
+ CONFORM†	–	–	**0.99**	0.70	0.33	0.72	0.11	**0.40**	0.54
+ Attention Overlap†	–	1184	**0.99**	0.74	0.33	0.72	0.15	0.23	0.53
+ Self-Cross†	–	–	**0.99**	0.74	0.33	0.72	0.13	0.26	0.53
+ Divide&Bind†	–	–	0.98	0.62	0.32	0.73	0.10	0.18	0.49
+ STORM†, switched-off control	–	0	**0.99**	0.35	0.33	0.73	0.03	0.07	0.42
+ STORM†	–	1184	**0.99**	0.41	0.33	0.72	0.15	0.05	0.44
+ Noise Refinement†	–	2208	0.98	0.71	0.33	0.80	0.31	0.31	0.57
+ Noise Refinement† (excl.)	–	2208	0.98	0.71	0.33	0.81	0.31	0.31	0.57
+ Mordant	–	2212	0.98	**0.82**	**0.48**	**0.83**	**0.71**	**0.40**	**0.70**

† Reproduced using the official released code; other rows as published.

Table 15: **GenEval failures by reason**, from the grader's per-image verdicts (SD 1.4, 553 prompts $\times$ 4 images). Position: missing object / wrong side; counting: too few / too many; two objects, colours and colour attribution: missing object / wrong colour. The dominant failure of the unguided model is a missing object on every task but counting; guidance removes most of those and leaves the wrong-side and wrong-colour counts nearly unchanged.

	Position ok (miss / side)	Counting ok (few / many)	Two objects ok (miss)	Colours ok (miss / colour)	Colour attr. ok (miss / colour)	Single ok (miss)
SD 1.4	25 (265 / 110)	123 (149 / 48)	167 (229)	293 (32 / 51)	32 (267 / 101)	316 (4)
Mordant + frame condition	282 (91 / 27)	144 (135 / 41)	322 (74)	313 (37 / 26)	146 (156 / 98)	315 (5)
Mordant	283 (87 / 30)	152 (124 / 44)	325 (71)	311 (39 / 26)	158 (152 / 90)	314 (6)

I.5 Unguided backbones

Tab. 16 places the guided 0.9B model among unguided backbones of up to 32B parameters, each generated in our harness at its own default resolution and step count and scored with the same code, 4 images per prompt on GenEval and 10 on T2I-CompBench. On GenEval every backbone up to FLUX.1-dev, fourteen times the size of SD 1.4, scores below 0.59 on position, the task that measures spatial relations, and only the two most recent backbones (Qwen-Image, FLUX.2-dev) are above 0.45 there. On T2I-CompBench the 2D score of SD 1.4 with Mordant (0.392) is above every unguided backbone in the table, published or measured. The unguided SDXL row is the one of Tab. 1, where the same backbone with Mordant reads 0.708.

I.6 The competitors' own protocols

Three of the compared methods evaluate on a protocol of their own. We run Mordant on each, with the same constants, and score it with the method's own metric as we re-implemented it; the competitor rows are the released code run and scored the same way (Sec. K). *DA-Score* is Noise Refinement's T2I-CompBench metric (Izadi et al., 2025): for each image, the mean over the prompt's question list of a VQA model's probability of "yes", on 25 prompts per category (24 distinct for colour), 10 images each; it rescores the images of Tab. 2 and needs no new generation. *Similar subjects* is the

Table 16: **Unguided backbones against SD 1.4 with Mordant.** GenEval (overall and the two-object, position and colour-attribution tasks) and T2I-CompBench (colour, shape, texture, 2D spatial). Backbones: SD 1.4 (Rombach et al., 2022), SDXL (Podell et al., 2024), IF-XL (DeepFloyd Lab at StabilityAI, 2023), SPRIGHT (Chatterjee et al., 2024a) (an SD 2.1 fine-tune), SD 3 (Esser et al., 2024), SD 3.5 (Stability AI, 2024), SANA 1.5 (Xie et al., 2025), Z-Image-Turbo (Tongyi-MAI Team, 2025), FLUX.1 (Black Forest Labs, 2024), Qwen-Image (Qwen Team, 2025), FLUX.2 (Black Forest Labs, 2025), DALL·E 3 (Betker et al., 2023). Parameter counts are the denoiser's. Rows marked † were generated and scored by us at the backbone's default resolution and step count, in bfloat16 for the transformer models, with 4 images per prompt on GenEval and 10 on T2I-CompBench; unmarked rows are published, with their citation, and a backbone with both a measured and a published part is listed once for each. Extra: SPRIGHT is a fine-tuned SD 2.1. FLUX.2-dev is at 50 steps (0.8493 at 28). SD 2.1 is the 512-pixel base checkpoint.

			GenEval				T2I-CompBench			
Model	Extra	Params	Overall	Two	Position	Attr.	Colour	Shape	Texture	2D
SD 1.4†	–	0.9B	0.45	0.42	0.06	0.08	0.37	0.37	0.40	0.13
SD 2.1† (base)	–	0.9B	0.51	0.50	0.07	0.15	0.49	0.42	0.49	0.17
SPRIGHT† (SD 2.1)	fine-tuning	0.9B	0.52	0.63	0.11	0.14	0.62	0.48	0.59	–
SDXL†	–	2.6B	0.58	0.79	0.11	0.23	0.59*	0.47*	0.53*	0.21*
IF-XL (Ghosh et al., 2023)	–	4.3B	0.61	0.74	0.13	0.35	–	–	–	–
SD 3-medium†	–	2B	0.70	0.86	0.29	0.59	0.81*	0.59*	0.73*	0.32*
SD 3.5-medium†	–	2.5B	0.71	0.88	0.24	0.60	0.80	0.59	0.73	–
SD 3.5-large†	–	8B	0.71	0.92	0.24	0.58	–	–	–	–
SANA 1.5†	–	4.8B	0.80	0.94	0.59	0.62	0.76	0.55	0.68	–
Z-Image-Turbo†	–	6B	0.75	0.91	0.48	0.57	–	–	–	–
FLUX.1-dev†	–	12B	0.67	0.84	0.23	0.44	–	–	–	0.29*
Qwen-Image†	–	20B	0.74	0.92	0.48	0.60	–	–	–	–
FLUX.2-dev†	–	32B	0.85	0.94	0.72	0.76	–	–	–	–
DALL·E 3 (Huang et al., 2025)	–	–	–	–	–	–	0.78	0.62	0.70	0.29
SD 1.4 + Mordant	–	0.9B	**0.70**	**0.82**	**0.71**	**0.40**	**0.69**	**0.55**	**0.68**	**0.39**

† Reproduced using the official released code; other rows as published. * From T2I-CompBench++ (Huang et al., 2025).

similar-subject success rate of the MMDiT method of Wei et al. (2026): 90 prompts naming 2, 3 or 4 visually similar subjects, 10 images each, an image counting only when Grounding DINO finds every subject (confidence ≥ 0.5 with two subjects, ≥ 0.4 with three or four); Mordant guides every named subject, STORM and Attention Overlap the first two. *COCO-Comp* is Attention Overlap's presence benchmark (Marioriyad et al., 2025): 1,275 two-object COCO prompts, one image each, scored by a VQA model asked whether each object is present (soft: the mean over the two objects; harsh: both present).

Table 17: **The compared methods' own protocols**, SD 1.4, generated and scored in our harness with each protocol's metric. DA-Score on the T2I-CompBench images of Tab. 2; similar subjects: success rate (%); COCO-Comp: presence soft / harsh (%); its unguided row is our SD 1.4 run, while Attention Overlap's own pipeline with its guidance switched off reads 72.00 / 44.20 (a different scheduler and precision). "–": the method was not run on that protocol.

		DA-Score (Noise Refinement)				Similar subjects	COCO-Comp
Method	Extra	Colour	Shape	Texture	2D	SR	soft / harsh
SD 1.4†	–	0.74	0.66	0.75	0.79	13.00	78.90 / 57.90
+ Attention Overlap†	–	–	–	–	–	22.00	94.80 / 89.60
+ Noise Refinement†	–	–	–	–	–	**34.80**	–
+ STORM†	–	0.70	0.67	0.74	0.83	12.20	–
+ Mordant	–	**0.90**	**0.79**	**0.91**	**0.87**	28.80	**97.90 / 95.90**

† Reproduced using the official released code; other rows as published.

J FURTHER ABLATIONS AND SENSITIVITY

J.1 SENSITIVITY OF THE STEP CONSTANTS

Tab. 18 varies the three step constants around the chosen values on the held-out VISOR prompts, on the form that preceded the final one, with the trace in place of the entropy power and the frame condition in place (2.7 / 1.6 points below the final objective under OWL-ViT and Mask2Former and level under Grounding DINO, Tab. 4); the constants were kept when the objective changed. The dual step ρ has a sharp optimum: its neighbours lose one to five points of object accuracy, and at $\rho = 0.3$ the relation accuracy given both objects also falls (97.8% against 100%), because the multipliers grow too slowly to hold the order. The guidance weight w and the objective's weight κ are flat in object accuracy; their values were set on the held-out colour-binding prompts (Sec. G).

Table 18: **Step constants on held-out prompts** (VISOR, 200 held-out prompts: object accuracy under OWL-ViT / Mask2Former / Grounding DINO), on the form with the trace objective and the frame condition.

Setting	VISOR OA	
$\rho = 1$, $w = 300$, $\kappa = 5$ (chosen)	79.90 / 79.40 / 92.50	
$\rho = 0.3$ / $\rho = 3$	78.60 / 76.10 / 89.50	76.30 / 74.80 / 89.60
$w = 200$ / $w = 400$	80.40 / 80.40 / 92.50	**81.00** / 80.40 / 91.90
$\kappa = 2.5$ / $\kappa = 10$	80.10 / **81.00** / 92.30	80.90 / 77.50 / **92.90**

J.2 WHY CONDITIONS AND NOT FIXED WEIGHTS

Before the constrained formulation, Mordant was a sum of fixed-weight penalties (its fixed-weight form, Tab. 4). Tuning it exposed the behaviour that Prop. 1 formalises: a term's effect is a property of the other terms' weights. On VISOR (500 near-balanced prompts, 4 images each), lengthening the guidance window from 25 to 50 steps adds 1.2 points of object accuracy when the forward re-allocation acts on the six hooked layers and 1.0 when it acts on the up-sampling blocks, and *removes* 2.1 points when it acts on every cross-attention layer. Removing the relational gate of the pair barrier adds 3.4 points of VISOR_4 with the narrower re-allocation, removes 1.0 with the widest, and removes 1.5 points of object accuracy once the loss also reads two more 16×16 layers. The re-allocation's best strength moves from 0.7 to 0.5 when those two layers are added. Each term's best weight, and even the sign of its effect, depended on what the others were doing; no fixed setting was a property of the statements.

The direct test is the penalty row of Tab. 4: the final program with every multiplier held at zero, so that each condition acts only through the quadratic part of Eq. (8), $\frac{\rho}{2}[g_c]_+^2$, one weight for every condition and nothing tuned. On the held-out VISOR prompts it loses 9.5 / 13.6 / 9.1 points of object accuracy against the run of the full program generated alongside it, and it loses the relation: $\text{VISOR}_{\text{cond}}$ falls from 100.0 / 100.0 / 99.9 to 89.6 / 88.7 / 86.8, so the order is violated on a tenth of the images that hold both objects, where every other guided row of the table holds it on 99.7 to 100.0% of them. Three readings of the same statements on the same 800 images order themselves: one untuned weight each, 73.6 / 67.8 / 84.1; weights tuned by hand over many rounds, the fixed-weight form, 79.8 / 79.6 / 91.3; multipliers found by dual ascent from zero, 83.1 / 81.4 / 93.3. The multipliers replace the tuning and exceed it, and Fig. 6 shows the mechanism on one prompt: the quadratic penalty of a violated order condition barely moves the pair, and the multiplier that accumulates while the condition stays violated is what does.

Tab. 19 reads the fixed-weight form at the evaluation protocol of Sec. 6, beside Mordant. The constrained program gives up 0.01 of colour and 0.02 of texture and is level on 2D and 3D; it gains 0.069 on shape, 0.017 on numeracy, 0.034 on GenEval and 2.8 points of VISOR object accuracy under OWL-ViT (1.5 and 1.5 under Mask2Former and Grounding DINO), with 4.5 more points of VISOR_4. On the controllability measurement of Tab. 6 the fixed-weight form carries 1.18 nats about the rendered relation ($P(Y{=}R \mid \text{both}) = 0.96$, OWL-ViT boxes), against 1.07 for Mordant.

The same measurements explain why the floor is a bounded term of the objective and not a condition. Stated as a condition, its multiplier grows without bound on an object the model will not draw:

Table 19: **The fixed-weight form at the evaluation protocol**, SD 1.4. VISOR: the balanced subset of Tab. 3, object accuracy under OWL-ViT / Mask2Former / Grounding DINO and VISOR$_4$ under OWL-ViT; GenEval: overall; T2I-CompBench: 10 images per prompt.

	VISOR OA	V_4	GenEval	Colour	Shape	Texture	2D	3D	Numeracy
Fixed-weight form	77.05 / 78.10 / 90.12	53.00	0.67	**0.70**	0.48	**0.69**	**0.39**	**0.37**	0.54
Mordant	**79.88 / 79.60 / 91.65**	**57.50**	**0.70**	0.69	**0.55**	0.68	**0.39**	**0.37**	**0.56**

VISOR-132 object accuracy fell to 43.9 / 50.2 / 70.3 under the three detectors, against 54.1 / 58.4 / 72.7 for the same program with the floor in the objective (an earlier screening round, before the frame and separation conditions). Raising its target to the full rendered median and its weight to the objective's ($\beta = 1$, $\lambda_f = 5$) starves every other statement: 8.7 / 20.3 / 37.3, on the form with the frame condition (Tab. 21). Raising the target alone is level under two detectors and 4.9 points lower under Mask2Former (59.3 / 62.5 / 83.5 against 60.6 / 67.4 / 82.2).

J.3 Removing one statement at a time on VISOR-132

Tab. 20 repeats the leave-one-out of Tab. 4 on the 132-prompt VISOR screening set (4 images each, three detectors), for the program with the trace objective and the frame condition that preceded the final form. Every statement that the final form keeps costs object accuracy under at least two of the three detectors when it is removed; the two binding conditions it no longer carries (*cover* and *no leak*, which asked the attribute to cover its noun and to avoid the rival) were dropped at this screening stage.

Table 20: **Leave-one-out on the VISOR screening set**, for the program with the trace objective and the frame condition (VISOR-132: object accuracy under OWL-ViT / Mask2Former / Grounding DINO).

	VISOR-132 OA
Full program	**64.60** / **67.80** / 81.80
− frame	61.00 / 65.90 / **82.80**
− separation	63.60 / 64.60 / 82.20
− allocation floor	63.80 / 64.00 / 78.60
− edge conditions (relation)	61.90 / 63.60 / 79.40
floor, margin, apart and binding only	60.60 / 63.80 / 80.30

On the form with the frame condition, Tab. 21 removes the allocation terms and the frame on VISOR-132; the final form is that table's row without the frame. The forward re-allocation is the largest single component (12.5 points under Mask2Former); removing the floor and the frame together costs 3.6 points under the two stronger detectors; the frame alone changes sign between detectors (+3.8 / −1.7 / 0.0 when removed), which the held-out sets confirm (Sec. J.4) and which is why the final form drops it; and raising the floor's weight to the objective's collapses object accuracy.

J.4 Effect of the frame condition

An earlier form of Mordant carried a fourth condition, on the frame: that no object be denser on the outermost ring of cells of the picture than a uniform map would be, $\log\left(n\,\bar{p}_\partial(p_i)\right) \le 0$ with $\bar{p}_\partial$ the mean of the map on the ring (the outer $1/64$ of the frame on the 64×64 reading grid), a condition on the object's extent with no margin to set. It prices the border: a formed object leaves only its pedestal on the ring, while an object that runs off the picture, the cropped, zoomed-in failure, puts real mass there. Its multiplier is zero on every trajectory whose objects stay inside the picture, so on most prompts it never acts; where it acts it pulls an object inward. The measurements say it is not needed. On the held-out VISOR prompts (Tab. 4) the form with the condition reads 82.6 / 81.0 / 92.4 against 83.1 / 81.4 / 93.3 without it, inside the spread, and on VISOR-132 (Tab. 21) removing it reads +3.8 / −1.7 / 0.0. At the full protocol the form with the condition reads VISOR 80.67 / 79.62 / 91.08 (balanced 1,000 prompts), GenEval 0.6917, and T2I-CompBench colour 0.6908, shape 0.5271, texture 0.6769, 2D 0.3906, 3D 0.3787 and numeracy 0.5626 at 10 images per prompt; the final form

Table 21: **Leave-one-out on VISOR-132 for the form with the frame condition** (4 images per prompt; OA and $\text{VISOR}_{\text{cond}}$ under three detectors; single runs, whose spread on this set is 1.3 to 2 points). The final form is the highlighted row, the same program without the frame; the removals below the rule are made from the form with it.

	OWL-ViT OA / cond	Mask2Former OA / cond	Grounding DINO OA / cond
SD 1.4 (unguided)	–	26.30 / 58.30	33.70 / 56.20
Mordant, fixed-weight form	55.90 / 99.70	63.30 / 99.70	81.80 / **99.80**
Mordant	**64.40** / **100.00**	65.70 / 99.70	82.20 / 99.30
+ frame condition	60.60 / 99.70	**67.40** / 99.40	82.20 / 99.10
removed from the form with the frame condition			
− forward re-allocation	57.40 / **100.00**	54.90 / **100.00**	79.20 / 99.50
− allocation floor	60.40 / 99.70	63.30 / 99.10	80.90 / 98.80
− floor and frame	59.90 / **100.00**	63.80 / **100.00**	78.60 / **99.80**
floor at the median ($\beta = 1$)	59.30 / **100.00**	62.50 / **100.00**	**83.50** / **99.80**
floor weight $\lambda_f = 5$	28.60 / **100.00**	40.20 / **100.00**	55.50 / 99.00
both ($\beta = 1$, $\lambda_f = 5$)	8.70 / **100.00**	20.30 / 99.10	37.30 / 91.90

at the same protocol reads 79.88 / 79.60 / 91.65 on VISOR, 0.7011 on GenEval and 0.6942 / 0.5494 / 0.6761 / 0.3916 / 0.3735 / 0.5591 on T2I-CompBench. The final form drops the condition: with it, VISOR and GenEval are level, every T2I-CompBench category but shape is within 0.005, and shape is 0.022 lower; without it, the only statement about the frame left in the program is the margin of Thm. 2, derived from the frame's width rather than imposed on the maps.

J.5 The relation statistic depends on the grader

Replacing the centroid and edge conditions by the surprisal of Eq. (10) on VISOR-132 lowers object accuracy from 64.6 / 67.8 / 81.8 to 56.3 / 61.0 / 79.6, all of it presence (the second object is found in 72.5% of images against 80.7%) with the relation as accurate as before (99.3 to 99.8% given both); the fixed-weight form, which uses the same surprisal, reads 55.9 / 63.3 / 81.8. On the final form and the held-out prompts (Tab. 4) the swap reads the same way: the surprisal on VISOR costs 5.1 / 2.8 / 0.3 points of object accuracy (78.0 / 78.6 / 93.0 against 83.1 / 81.4 / 93.3) with $\text{VISOR}_{\text{cond}}$ at 100.0 / 99.8 / 99.9. The surprisal does not stop once the centres are ordered; on a grader that reads only the sign of the centre offset, that persistence is spent against presence, while a grader that pays for non-overlapping boxes rewards it.

J.6 What did not work

We record the directions that were measured and closed, since several are natural first ideas. *Optimising the initial noise* on the program's own loss or on the weakest word's peak attention, before sampling, lowered VISOR-132 object accuracy by 5 to 6 points under OWL-ViT, 7 to 16 under Mask2Former and 3 to 4 under Grounding DINO: the first step's maps are hazes, and their gradient moves the noise in directions unrelated to where objects will form. *A step normalised to a fixed fraction of the noise prediction* in place of w never lets the guidance switch off once the conditions hold, and lost 4 to 9 points of object accuracy under OWL-ViT and up to 17 under Mask2Former; normalising each condition by its initial violation overflowed fp16 on conditions that start near zero. *Latent refinement* in the style of Attend-and-Excite (repeated gradient steps on the latent at selected timesteps) lowered attribute binding monotonically in the step size. *Image-space scores* (a CLIP loss on the decoded prediction, with the attribute swaps as negatives) were inert on colour, texture and shape at every weight, view and schedule, consistent with CLIP's known insensitivity to binding. *A single information objective* in place of the effective area and the separation condition (the mutual information between entity and cell) rewards spreading the union of the maps over the frame, which is a push to the border in disguise.

Table 22: **Ablations under three detectors**: the VISOR cells of Tab. 4 re-detected with Mask2Former and Grounding DINO, VISOR's rule applied to their boxes. The grey line is the difference from Mordant.

	VISOR OA, held-out (OWL-ViT / Mask2Former / Grounding DINO)
SD 1.4 (unguided)	29.10 / 31.30 / 38.90
Mordant, fixed-weight form	79.80 / 79.60 / 91.30
Mordant	**83.10 / 81.40 / 93.30**
− allocation floor	79.50 / 80.50 / 91.30 −3.60 / −0.90 / −2.00
multipliers at zero (penalties)	73.60 / 67.80 / 84.10
$\text{VISOR}_{\text{cond}}$ 89.60 / 88.70 / 86.80	−9.50 / −13.60 / −9.10
surprisal as the relation statistic	78.00 / 78.60 / 93.00 −5.10 / −2.80 / −0.30

K BASELINES RUN FROM RELEASED CODE

Every competitor row marked "generated and scored in our harness" is the authors' own repository, at the commit given in Tab. 23, run through one driver that hands the pipeline a prompt, a seed and token indices and writes each image under the same name our own pipeline uses, so that the same scorers read every row. These rows are reproductions on our protocol: the same backbone, prompt sets, seeds and scorers as our rows, with each method's published default settings. They are not the authors' numbers, which are quoted separately and marked as published wherever they exist; where a reproduction disagrees with the paper we report both (Sec. K.1).

Environment. The repositories pin diffusers 0.12 to 0.16; they run here under the diffusers version that generates everything else, with four names that moved between those versions restored by a compatibility layer (`randn_tensor`, the `pipeline_utils` and `cross_attention` modules, and the optional batch-size argument of `prepare_attention_mask`). One environment means the weights, the VAE and the scheduler each pipeline loads by default (PNDM for SD 1.4) are identical across competitor rows, so a difference between two rows is the method. Every competitor runs at 50 steps, classifier-free guidance 7.5, 512×512, batch size 1 and fp32: four of the repositories load fp32 by default and STORM's guided path hard-codes it, so fp32 is the precision at which STORM's published 2D score, the one number that validates the integration, was produced; CONFORM's notebook asks for fp16 and is run at fp32 for uniformity. Image k of a prompt is drawn from a generator seeded with $42 + k$, as for our rows.

Token indices. Attend-and-Excite's driver asks the user for the token indices to steer, and no competitor publishes an automatic rule, so one rule is used for every method that takes indices: the object words of the benchmark record, located in the tokenised prompt, which is the list the scorer grades, with spaCy nouns as the fallback where a record word cannot be found (38 of 553 GenEval prompts, 50 of 1,000 VISOR, 5 of 300 T2I 2D). Adjectives, where a method binds them (STORM's attribute term, CONFORM's token groups), follow STORM's own rule: spaCy adjectives, each bound to the nearest noun to its right. STORM's spaCy rule for nouns was tried first and rejected because it fails on the benchmarks' templates: *photo* is tagged as a noun in *a photo of a dog right of a teddy bear*, so it returned *photo* as the first object on 99 of 100 GenEval position prompts, and *top* on 90 of 300 T2I prompts of the form *a girl on the top of a frog*. Noise Refinement uses its own token matching, and CoMPaSS needs no indices.

Prompts a method cannot act on. STORM's transport plan, CONFORM's contrastive loss and Attention Overlap's centre-of-mass distance are undefined on one entity, and half of GenEval's prompts name one (single object, counting, colours). Those prompts are not dropped, which would score them zero: they are generated by the method's own pipeline with the guidance off, at the same seed, through the `run_standard_sd` switch of the Attend-and-Excite family or, for CONFORM (no switch) and STORM (its denoising loop computes the transport loss before consulting the switch and raises on one entity), a plain Stable Diffusion pipeline assembled from the method pipeline's own components. Attend-and-Excite, Divide-and-Bind and InitNO take one index and stay guided.

Each run records how many images the method guided (Tab. 14); a GenEval row in which half of the images are plain Stable Diffusion is, on those tasks, a row about the backbone.

Controls. A competitor's pipeline differs from ours in precision and scheduler, so its scores are read against its own pipeline with the guidance switched off wherever the code allows one, not against our fp16 DDPM base row. STORM is the case that shows why: on GenEval its released code reads -0.011 overall against our base row and $+0.026$ against its own control, with position $+0.115$ and two objects $+0.066$; its counting and colours are one-entity tasks on which it never acts, and there the -0.05 against our base row is the pipeline difference, not the method, while single object is identical to four decimals in both of its runs. The shift between the two readings is largest on the detector-scored benchmarks (up to 0.074 on GenEval two objects, 0.028 on T2I numeracy) and negligible under a VQA scorer (0.010).

Table 23: **Competitor protocols.** Repository commit, the published defaults each run uses, what the loss is handed, and the fallback on prompts the loss cannot be formed on. Every row: SD 1.4, 512×512, 50 steps, CFG 7.5, fp32, batch size 1, seeds $42 + k$, the pipeline's default scheduler.

Method	Commit	Settings (the authors' defaults)	Loss reads	One entity
Attend-and-Excite (Chefer et al., 2023)	163efdf	attention at 16×16, up to 25 latent updates per step, thresholds 0.05 / 0.5 / 0.8 at steps 0 / 10 / 20, step scale 20 decaying to 10, Gaussian smoothing ($\sigma = 0.5$, kernel 3)	object nouns	guided
Divide-and-Bind (Li et al., 2023b)	644dfda	as above with the total-variation loss and the max-attention indicator (the paper's variant)	object nouns	guided
InitNO (Guo et al., 2024)	fa021c4	initial-noise optimisation on, attention at 16×16, up to 25 latent updates per step	object nouns	guided
CONFORM (Meral et al., 2024)	f18d658	NT-Xent contrastive loss at temperature 0.5, up to 30 updates per step, refinement at steps 0 / 10 / 20 with 20 refinement steps, step scale 20, smoothing	one group per object: noun plus its adjectives	plain SD
STORM (Han et al., 2025)	88dbc3a	attention at 16×16, up to 25 updates per step, thresholds 0.9 / 0.95 / 0.99 / 0.995 / 0.999 at steps 0 / 5 / 10 / 15 / 20, step scale 20 decaying to 10, smoothing ($\sigma = 0.5$, kernel 3)	object nouns and adjectives	plain SD
Attention Overlap (Marioriyad et al., 2025)	357935c	centre-of-mass distance loss, attention at 16×16, up to 25 updates per step, step scale 20 decaying to 10; its three-entity branch hard-codes the token positions of the authors' template, so the first two nouns are steered	first two nouns	plain SD
Noise Refinement (Izadi et al., 2025)	d4cc1b0	two stages; constraints written from the benchmark records' objects and attributes (the authors' GPT-4 parse covers only their own 150 prompts); verifier: the BLIP VQA model their code loads (not the scorer's); the Attend-and-Excite update schedule; failed on the 4 images of *a photo of an orange orange*	its own token matching	guided
CoMPaSS (Zhang et al., 2025)	defdedd	the authors' released SD 1.4 U-Net (full fine-tune) with their positional-encoding processor and inference recipe (PNDM, 50 steps, CFG 7.5, fp16); no guidance	–	–

Cost. Tab. 24 gives the wall-clock time per image of each released pipeline at batch size 1, read from the runs above, beside our own pipeline at the same batch size on the same GPU. The competitors run in fp32 and our rows in fp16, so the fairest reading is each method against its own guidance-free pipeline: Mordant takes 3.3× the time of the unguided model (4.35 against 1.30 s), STORM 5.0 to 7.7× the time of its own pipeline with the guidance switched off (11 to 17 against 2.2 s), and STORM is the slowest method in the table.

The 2D label matcher. Every T2I 2D score in the paper uses the benchmark's verbatim match between a prompt noun and a UniDet label string. STORM's 2D run was first scored with an earlier matcher that split UniDet's compound labels ("cattle, cow") into words and graded the last detected instance, which read 0.1662; the row in Tab. 2 is the same cached detections re-scored under the benchmark's rule (0.1578), which is also the rule under which our unguided model reproduces the published 0.1246 to within 0.003.

K.1 Reproducing STORM

STORM is the one training-free method with published numbers on both VISOR and T2I-CompBench, so its re-run is what says whether a shimmed pipeline measures what the authors measured. Tab. 25

Table 24: **Seconds per image at batch size 1**, every row on the same GPU. Competitors from their released code in fp32 (mean over a run; STORM's range spans its GenEval and T2I runs), except CoMPaSS, which applies no guidance and runs its own fp16 recipe; our rows in fp16 with the losses in fp32.

Method	s / image	Method	s / image
Attention Overlap	4.1	STORM, guidance off	2.2
Attend-and-Excite	5.1	STORM	11 to 17
Divide-and-Bind	6.5	Noise Refinement	13.8
CONFORM	6.5	CoMPaSS (fine-tuned weights)	0.5
InitNO	11.5	SD 1.4, unguided (ours, fp16)	1.30
		Mordant, fixed-weight form	3.85
		Mordant	4.35

gives the four T2I categories STORM reports, at the benchmark's protocol of 300 prompts and 10 images each, from its released code at its own defaults.

Table 25: **STORM's published T2I-CompBench scores against its released code** on SD 1.4 (300 prompts $\times$ 10 images, fp32, the benchmark's scorers). Control: STORM's own pipeline with the guidance switched off; base: our pipeline in fp32. The 2D score reproduces; the attribute scores stay at the control's level.

Category	Published	Released code	Its control	Base (fp32)	Code $-$ control
2D spatial	0.16	0.16	0.13	0.13	$+0.03$
Colour	0.65	0.34	–	0.37	(-0.03 vs. base)
Shape	0.60	0.35	0.36	0.37	-0.01
Texture	0.75	0.41	0.41	0.40	-0.00

The spatial term reproduces. On 2D spatial, the one category whose prompts carry no attribute, the released code reads 0.1578 against the published 0.1613, $+0.032$ over its own switched-off control, and that control (0.1263) agrees with our fp32 unguided pipeline (0.1263) and with our fp16 one (0.1273), so the two pipelines are measuring the same thing. Its GenEval position score moves the same way ($+0.115$ over its control, Tab. 14).

The attribute term does not. Colour, shape and texture land at the control's level (-0.003 to -0.005) and 0.24 to 0.35 below the published values. This is not an inactive term: parsed from the run logs, the attribute loss is non-zero on 99.7% of the logged steps on texture and 82.2% on shape, and the transport term is non-zero on every step. It is not the token source: the same cells with STORM's own spaCy rule for nouns read 0.3553 (shape) and 0.4154 (texture), with the benchmark's object and attribute words and spaCy filling the gaps 0.3541 and 0.3994, and with a hand-checked annotation of every prompt's head nouns and attributes handed to the code at its own index convention (a partial run, paired with the base on the same prompts and seeds) -0.022 ± 0.042 on shape, $+0.015 \pm 0.071$ on texture and -0.067 ± 0.049 on colour, where the published rows would need $+0.24$ and $+0.35$ over the base. It is not the scorer: the same BLIP-VQA path reproduces the benchmark's own SD 1.4 row to within 0.015 on all three categories. And the benchmark's scoring rule does not hide a protocol that would recover it: the attribute score is the product of BLIP's answers over a prompt's noun phrases, and truncating that product to its first K phrases (Tab. 26) lifts the base, the control and the released code alike, so that even at $K = 1$ the code reads 0.694 on texture against the published 0.754.

What the released code does. Reading the code against the paper explains the split. The attribute term is an entropic optimal-transport cost between a noun's map and its adjective's whose ground cost is defined on flattened raster indices rather than on 2D positions, so it is strongly anisotropic (moving an adjective's map eight cells along a row and along a column change the loss by different orders of magnitude); it is added only from the seventh step on and is exactly zero whenever an adjective has no index. Adjectives are spaCy adjectives, each bound to the nearest noun on its right, kept only when a single CLIP token decodes to the word: the multi-token shape words (*triangular*, *spherical*, *pyramidal*, *oblong*, *conical*, *cylindrical*, *pentagonal*, *teardrop*) therefore never bind, and

Table 26: **BLIP-VQA product truncated to the first K noun phrases** (SD 1.4, 10 images per prompt; $K = 8$ is the official score). No truncation brings STORM's released code near its published row (texture 0.7539, shape 0.5983); Mordant is shown for scale.

	Texture				Shape			
	K=1	K=2	K=3	official	K=1	K=2	K=3	official
SD 1.4	0.70	0.46	0.42	0.40	0.78	0.44	0.39	0.37
STORM, switched off	0.70	0.47	0.43	0.41	0.77	0.43	0.37	0.36
STORM, released code	0.69	0.46	0.43	0.41	0.76	0.42	≈0.37	0.35
Mordant	**0.87**	**0.73**	**0.69**	**0.68**	**0.83**	**0.60**	**0.56**	**0.55**

glass, *leather*, *fabric* and *rubber* are tagged as nouns, so on texture 72, 155 and 56 of the 300 prompts reach the attribute term with two, one and no bound adjective. There is no Attend-and-Excite peak term anywhere in the code; the latent takes a clipped gradient step at every step of the first 25, and the refinement iterations run at steps 5, 10, 15 and 20 only while the loss exceeds a threshold and the second noun's centroid lies below the first's, which happens on 56 to 63% of the images of each set. The published attribute row is also difficult to place on the benchmark's own scale: its SD 1.4 texture (0.7539) is above the SD 3 (0.7334) and DALL·E 3 (0.7036) rows of T2I-CompBench++'s table. We therefore quote STORM's attribute scores as published, report the released code's scores beside them, and draw no conclusion from the gap beyond the fact that it does not come from our harness.

L More qualitative results and failure cases

All images in this appendix are SD 1.4 at 512×512 (the SD 2.1 comparison is Fig. 4) under the protocol of Sec. C; image k of a prompt is generated from seed $42 + k$ by every method, so each row or triple compares one initial noise. Competitor images come from the authors' released code with its default settings, run on the same backbone and seeds in our harness (Sec. K). Every image is shown at reduced resolution; none is retouched.

Fig. 7 repeats the same-seed comparison of Fig. 4 on SD 1.4, with the competitors. On every row the unguided model draws one of the named objects and drops the other, or draws two clocks for four. The presence methods (Attend&Excite, InitNO, CONFORM, Attention Overlap) and the spatial one (STORM) mostly reproduce that failure on the same seed, and when they do draw the second object they put it where the prior puts it: the cake under the backpack, the bird on top of the skateboard. Mordant draws both objects, and the relation, the colours and the count then follow, with the style of the unguided image preserved. Its failures are of the kinds the analysis predicts: two objects the prior draws as one are fused rather than separated (last row), a class the backbone rarely renders is replaced by a generic one, an object of a vertical relation is pushed into the frame's edge, and counts above three are overshot (Fig. 9).

M How the conditions act on the attention maps

The tables measure what Mordant changes; this appendix shows *how*, on prompts the unguided model fails. For each prompt we take a seed at which the unguided image fails its grader (image k of a prompt is seed $42 + k$, Sec. C), and generate from that seed three times: the frozen model, the full program, and the program with one statement removed. Every prompt word's cross-attention map is read at the five 16×16 cross-attention layers (the band on which every condition acts, averaged over the five layers; Sec. C) at steps $t = 1, 5, 8, 12, 24$ and 49 of 50; guidance acts at $t < 25$. Each map is normalised to its own range, so the strips show the geometry of a map and not its magnitude, which decays over sampling for every word. The final image is shown at the right of each strip. The maps are the model's own attention on the trajectory that the program shapes; the program reads them and never writes into them. Sampling on a GPU is not bit-reproducible, so the same configuration can differ in detail between two runs; the three strips of a figure always come from one run. The removals are those of Tab. 4: the separation condition Eq. (5), the binding condition Eq. (7), the allocation floor and the apart term of Eq. (4), the forward re-allocation Eq. (9), and the program with every multiplier held at zero, where the same conditions act as fixed quadratic penalties (Eq. (8) with $\lambda_c = 0$).



Figure 7: **SD 1.4 against the training-free methods, same seed, on GenEval.** Each row is one prompt and one initial noise: image k of the prompt, seed $42 + k$, in every column. The rows cover the two-object, position, colour-attribution and counting tasks and were chosen where the unguided model fails the grader by a missing object (too few objects in the counting row) and most of the training-free methods fail on the same seed. Competitor images are generated with the authors' released code and default settings on the same backbone and seeds. In the last row Mordant fails as well: the bat and the fork are fused into one object.

A relation on a sign grader (Fig. 6). For *an umbrella to the left of a cow*, the unguided model draws a cow and no umbrella; the two words' maps share a centre through $t = 8$, after which the umbrella's drifts to the right of the cow's (gap -0.08 of the frame at $t = 24$). Under the full program the umbrella's map has moved left and the cow's right by $t = 5$; the gap between the two centres grows from 0.23 of the frame at $t = 5$ to 0.26 at $t = 8$, 0.31 at $t = 12$ and 0.32 at $t = 16$, just short of the margin $m = \frac{1}{3}$, and the multiplier of the centre-order condition grows for as long as the gap is short of it (0.3 at the first guided step, 2.6 at $t = 12$, 3.6 at $t = 24$). With the multipliers held at zero the same condition acts only as its quadratic penalty: the gap grows from 0.11 at $t = 5$ to 0.19 at $t = 24$, about half the margin, and the umbrella is drawn over the cow rather than beside it. The quadratic penalty of a violated condition is what a fixed-weight method applies; on this prompt it moves the pair half as far, and the multiplier that dual ascent accumulates while the condition stays violated moves it the rest of the way.

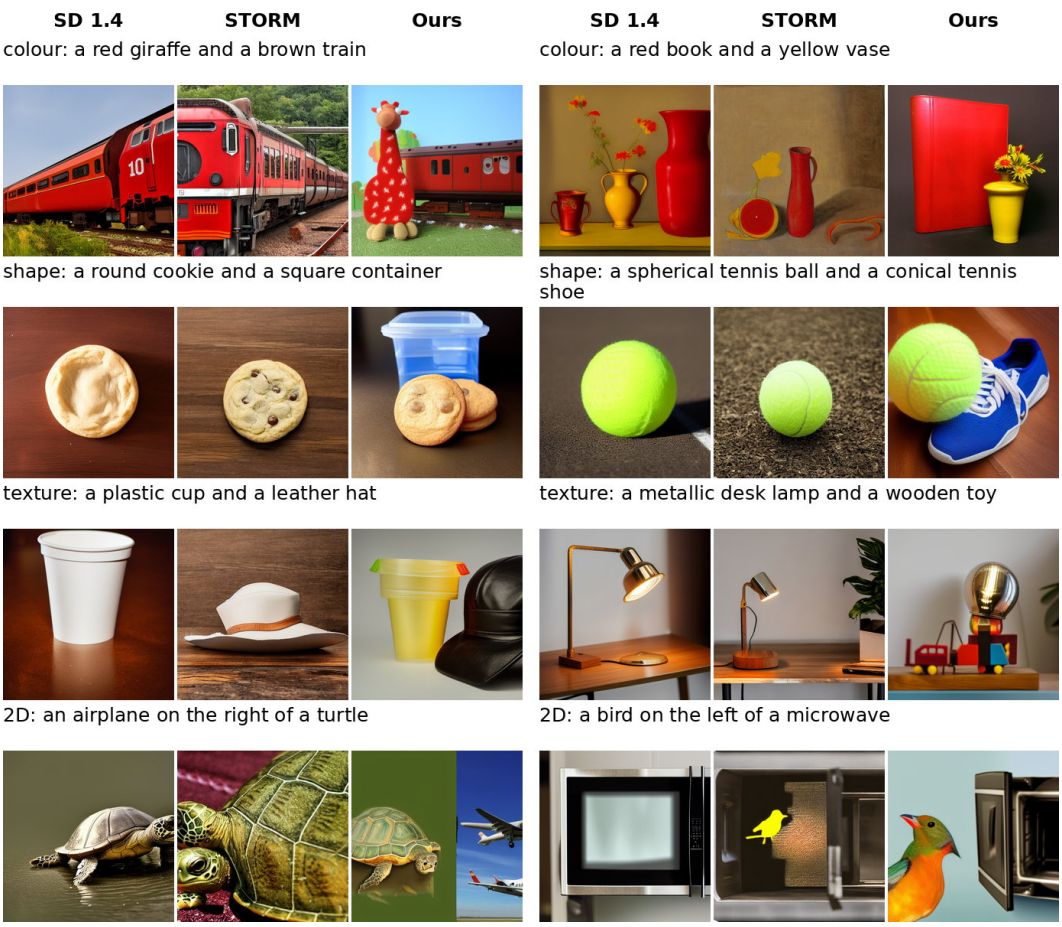


Figure 8: **T2I-CompBench, same seed.** Two prompts from each of the colour, shape, texture and 2D-spatial categories, image 0, each as unguided SD 1.4, STORM (released code, default settings) and Mordant. Prompts were chosen where both the unguided model and STORM fail the category's grader (a BLIP-VQA score below 0.25, or a 2D-spatial score of 0). The unguided column and STORM's pipeline run the U-Net in fp32, Mordant in fp16 (Sec. C); the seeds are the same. As on GenEval, the unguided model and STORM render one of the two objects; the guided images add the second one with its attribute or on the stated side.

Two objects with no relation (Fig. 11). For *a photo of a baseball bat and a bear*, the unguided model draws the bear holding the bat; the bat's map is a thin stripe inside the bear's region and the detector finds no bat. Under the full program the bat's map is a compact vertical bar at a third of the frame's width from $t = 5$, the bear's map sits at two thirds, and the bat stands beside the bear in the picture. The separation condition is satisfied on both trajectories from $t = 5$ (the cross-entropy $H(p_{\text{bat}}, p_{\text{bear}})$ exceeds $\log n$ at $t = 12$ by 1.3 nats with the apart term and by 1.0 without it), so it is not what keeps the two claims apart; the apart term is. Without it the first steps are the same, but from $t = 12$ the bat's map fragments into three vertical stripes and the picture shows three bats: nothing in the remaining program says where a second claim should sit relative to the first, and the model forms it in several places.

Colour attribution (Fig. 12). For *a photo of a black car and a green parking meter*, the unguided model draws a green car and no meter: the meter's map stays diffuse over the whole frame at every step (the mass in its top tenth of cells is 0.20 to 0.23, against 0.10 for a uniform map), and the maps of *black* and *green* both lie on the car. Under the full program the meter's map contracts to a compact region at the right of the frame by $t = 8$ (0.34 of its mass in its top tenth of cells, 0.42 by $t = 24$), *green* sits on it and *black* on the car, and the picture is a black car beside green meters. The binding multipliers act at $t \leq 4$, while the meter's region forms, and again more weakly after $t = 15$ as



Figure 9: **Failure cases of Mordant on GenEval**, one per kind, each with the reason read off the grader's per-image verdict (seeds $42 + k$ of the benchmark's prompts). *Merged pair*: two objects the prior draws as one are fused rather than separated, and neither is detected. *Hard class*: an object the backbone rarely renders (a hair drier) becomes a generic appliance. *Attribute leak*: the first object's colour reaches the second, which the grader then reads as green. *Border crop*: the lower object of a vertical relation is pushed into the bottom edge and is not detected. *Count*: an extra instance is drawn. *Wrong side*: a physically plausible arrangement, a person riding the board, overrides the stated relation.

the sharpening maps briefly re-violate the condition (Fig. 14); in between the condition holds by itself. Without the binding condition the meter is still drawn, because separation and the apart term are unchanged, but at $t = 24$ both colour words carry mass at the meter's column and the meter comes out black and green. At these layers the map-level difference is modest (the adjective's map is explained better by its own noun than by the rival by 0.44 nats with the condition and 0.34 without, at $t = 24$); the effect is visible in the pixels more than in the coarse maps.

Texture binding, where the map itself moves (Fig. 13). For *a fabric scarf and a glass bowl*, the unguided model draws the scarf only. Under the full program the bowl's map forms at the bottom of the frame by $t = 5$ and the scarf's above it, and the map of *fabric* follows the scarf: at $t = 8$ it is explained by the scarf's map better than by the bowl's by 0.71 nats, and by 0.73 at $t = 12$. Without the binding condition the two objects are still drawn, and *glass* still lies on the bowl, which the model binds by itself (0.5 to 0.9 nats on every trajectory); but the map of *fabric* straddles both objects until late in sampling (0.02, -0.02 and 0.08 nats at $t = 5, 8, 12$), which is the case Eq. (7) is written for: an attribute that no other statement places, on a noun that the model does not associate with it.

The multipliers over the guided steps (Fig. 14). Fig. 14 reads the multipliers of Sec. 5.3 on the full program's trajectory for two of the prompts, after each step's dual update. On *an umbrella to the left of a cow* the centre-order condition of Eq. (6) stays just short of $\frac{1}{3}$ for the whole window, and its multiplier grows monotonically to 3.6; the edge condition (the subject's far edge before the object's near edge along d) is met by $t = 8$, and its multiplier, which peaked at 0.50 at $t = 3$, has returned to zero by then. One across-d condition binds from $t = 4$ and its multiplier levels off near 1.0; the separation condition never binds on this prompt. On the colour prompt the separation condition binds twice: its multiplier rises to 0.25 at $t = 5$ while the meter's map contracts beside the car, falls back to zero by $t = 10$ once the two maps no longer explain each other, and rises again after $t = 15$ as the maps sharpen onto adjacent objects; the binding multiplier of *green* peaks at 0.14 at $t = 2$, is zero from $t = 5$ and stays below 0.05 afterwards. No multiplier is set by hand: each one is the running sum of its condition's violation, and it returns to zero as soon as the condition holds.

The other prompts. The same strips were made for four further prompts. On two relational prompts the relation carries the picture and no single removal breaks it: for *a photo of a backpack right of a sandwich* (unguided, no backpack) every arm draws the sandwich on the left and the backpack on the right, with the backpack's map at the right edge from $t = 5$, and for *a backpack to the left of a cat* (unguided, a backpack strap on a cat at the right) every arm draws the backpack left of the cat. For *a circular pendant light and a triangular wall hook* the unguided model draws a conical lamp and no hook; the full program draws a round pendant and a separate hook, and removing separation, the floor, the apart term or the re-allocation each merges the hook into the lamp or loses it. The shape words play no part in this: they carry no binding condition (Sec. 5.2), and their maps sit on their nouns in every arm, so the picture's shapes are the model's own. For *a photo of four bowls* the unguided model draws three bowls and so does the full program, whose map holds three claims in a row by $t = 16$; without the forward re-allocation the map holds three blobs by $t = 8$ and four by $t = 24$, and the picture has four bowls. This is the one prompt on which a removal did better, on one seed, and it is consistent with what the re-allocation does: it raises the share of the claims that already exist, which on a counted noun can commit the map to the claims it has before the fourth forms.

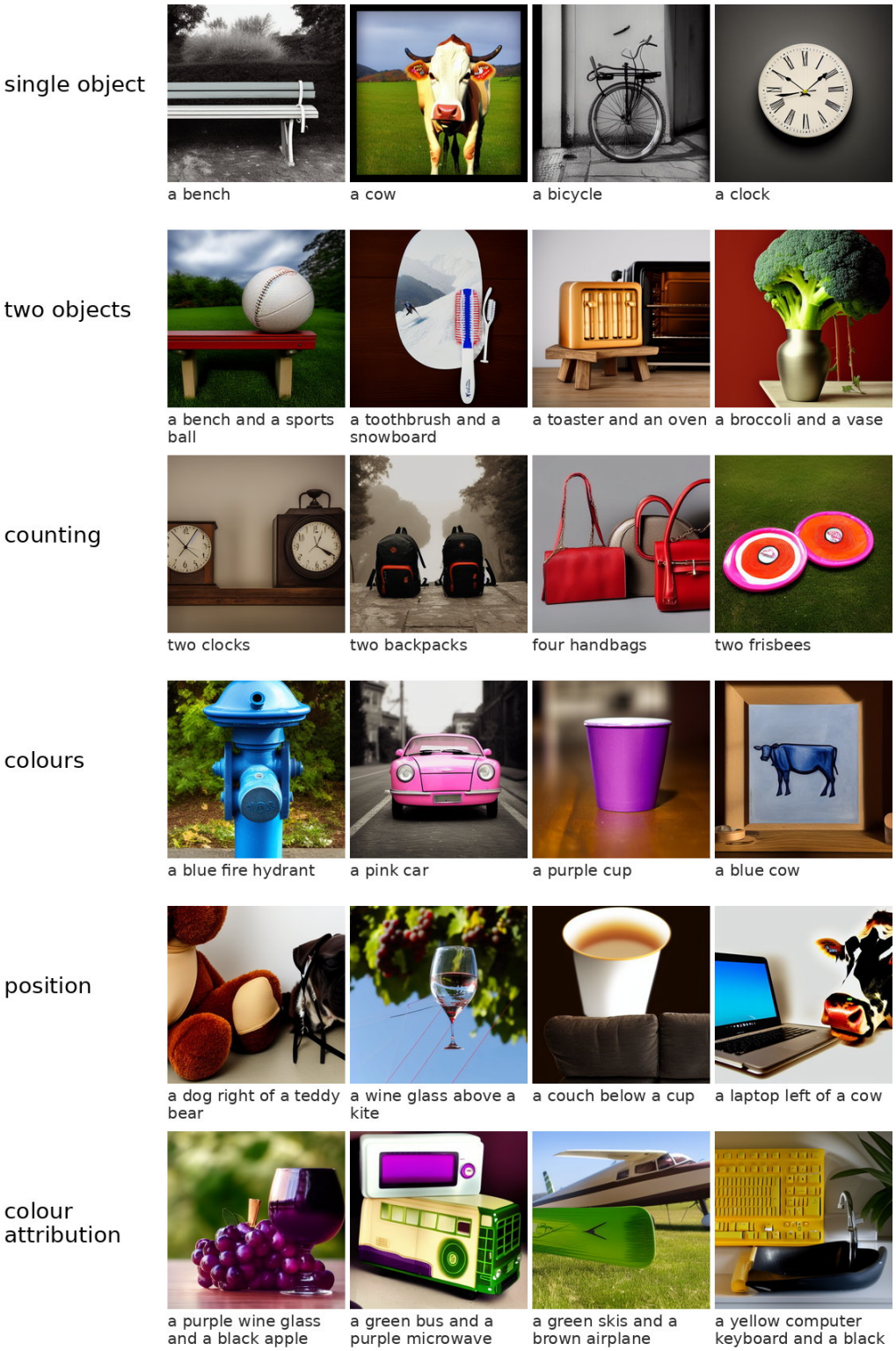


Figure 10: **Uncurated samples of Mordant on GenEval.** The first four prompts of each of the six tasks in the benchmark's own order, image 0, with no selection.

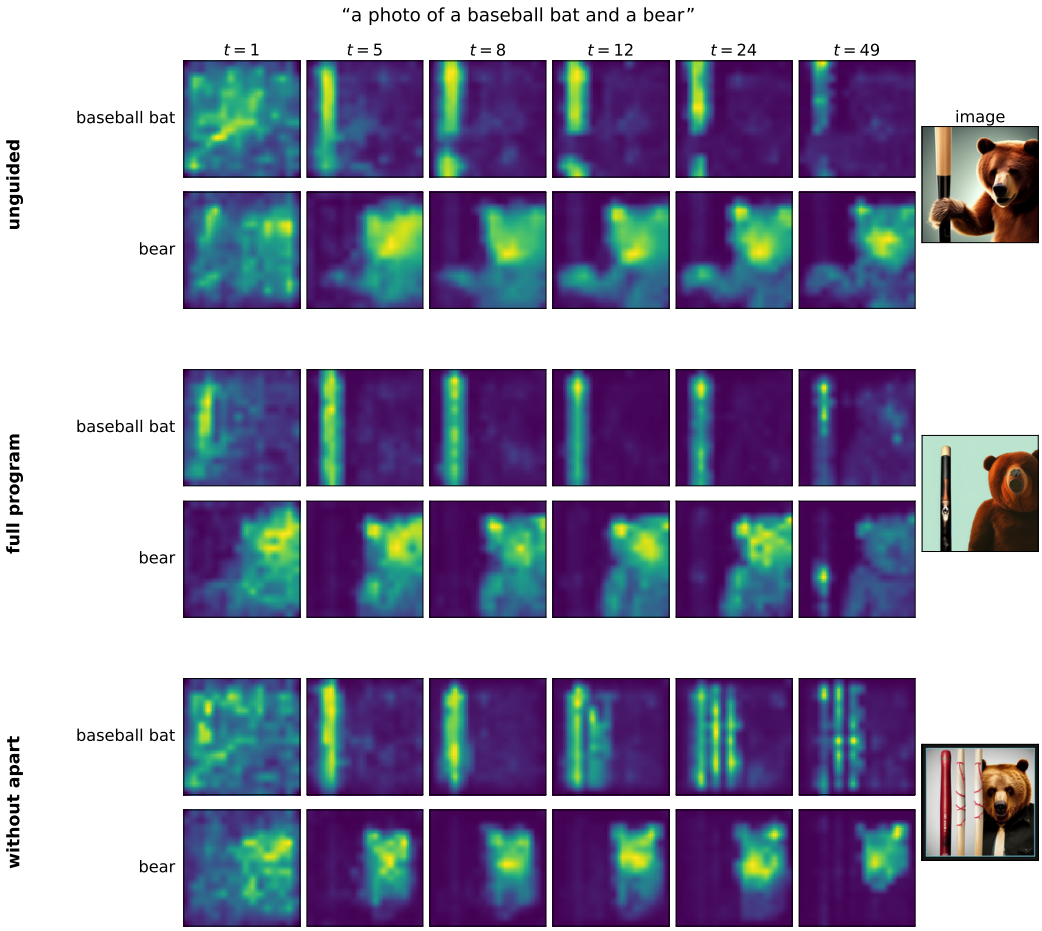


Figure 11: **Two objects and the apart term.** *A photo of a baseball bat and a bear*; layout as in Fig. 6. Top: the frozen model (the bat is drawn in the bear's paws and not detected). Middle: Mordant (the bat's map is a compact bar a third of the frame from the bear's from $t = 5$). Bottom: without the apart term of Eq. (4) the bat's map splits into three stripes from $t = 12$ and the picture has three bats; the separation condition holds on both trajectories, so it is the apart term that fixes where the second claim forms.

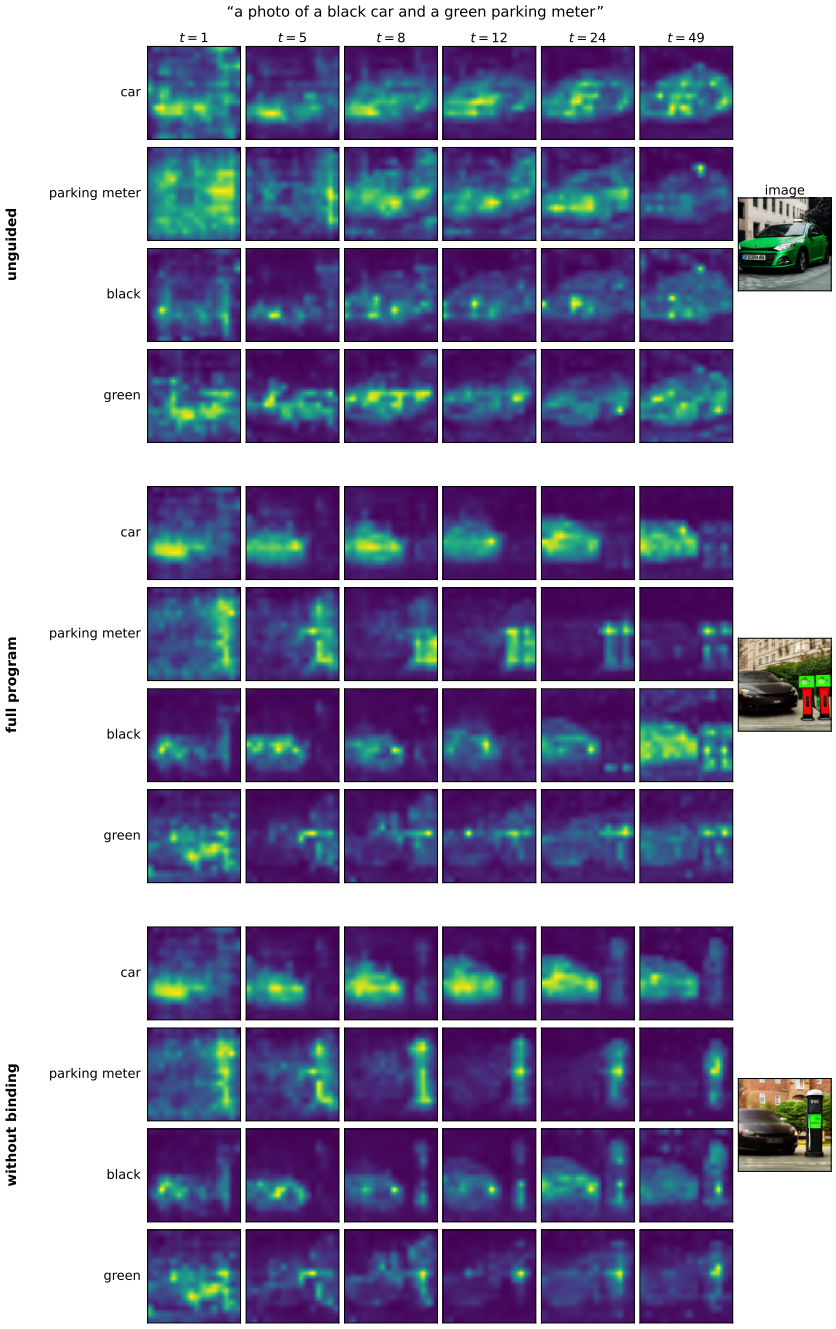


Figure 12: **Colour attribution.** *A photo of a black car and a green parking meter*; rows are the two nouns and the two colour words, layout as in Fig. 6. Top: the frozen model (a green car, no meter; the meter's map never contracts). Middle: Mordant (the meter's map contracts to the right by $t = 8$; *green* sits on it and *black* on the car). Bottom: without the binding condition Eq. (7) the meter is still drawn, but both colour words reach it and it comes out black and green.

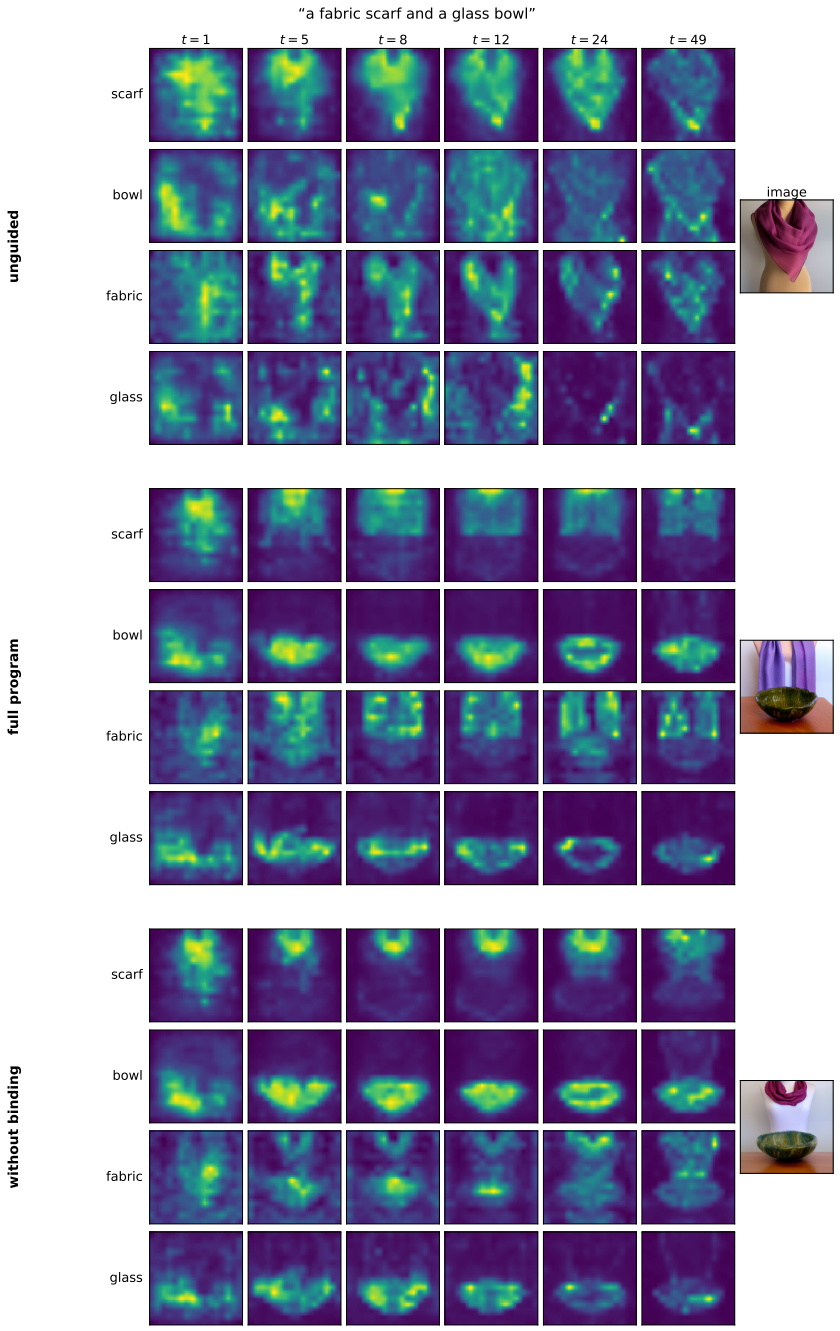


Figure 13: **Texture binding.** *A fabric scarf and a glass bowl*; rows are the two nouns and the two material words, layout as in Fig. 6. Top: the frozen model (a scarf, no bowl). Middle: Mordant (the bowl forms at the bottom by $t = 5$; *fabric* lies on the scarf and *glass* on the bowl). Bottom: without the binding condition the objects are still drawn but the map of *fabric* straddles both until $t = 24$; *glass* is bound by the model on its own.

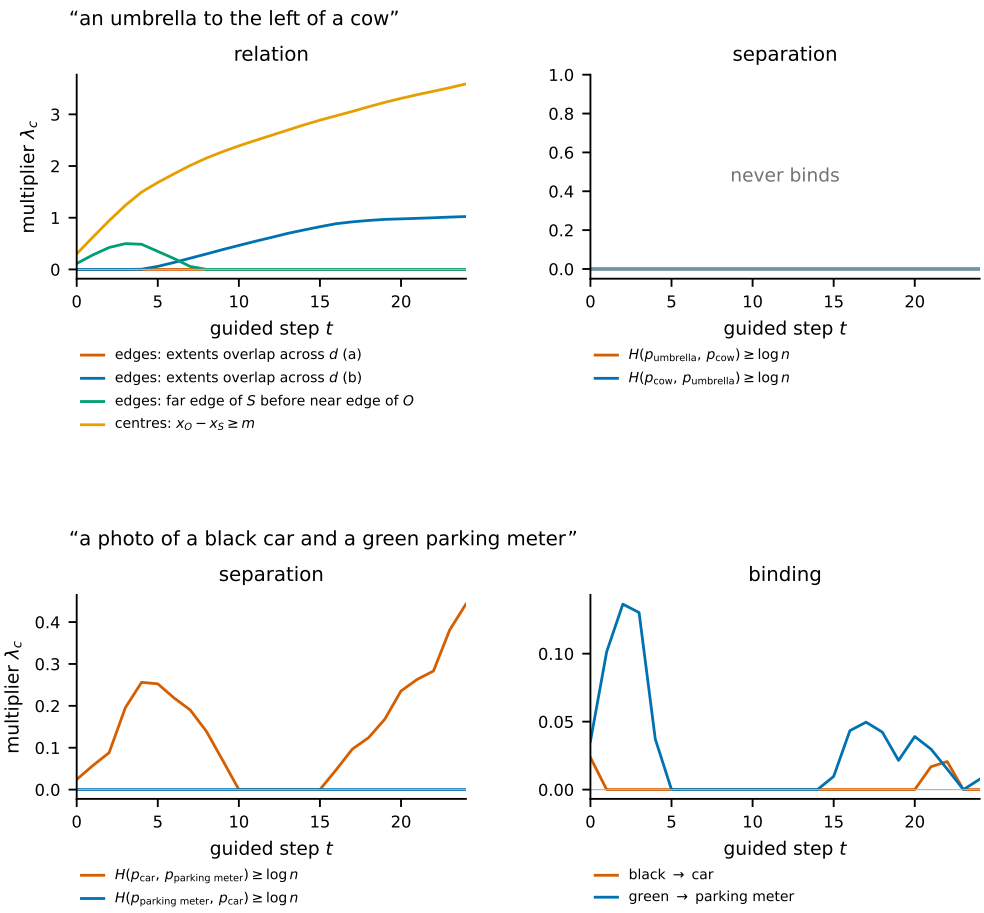


Figure 14: **Multipliers over the guided steps** of the full program, one line per condition. Top: *an umbrella to the left of a cow*, where the centre-order multiplier grows while the gap stays short of $\frac{1}{3}$; bottom: *a black car and a green parking meter*, where separation binds twice and the binding of *green* is settled early.