ESG: EARLY SPATIAL GUIDANCE FOR COMPOSITIONAL TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

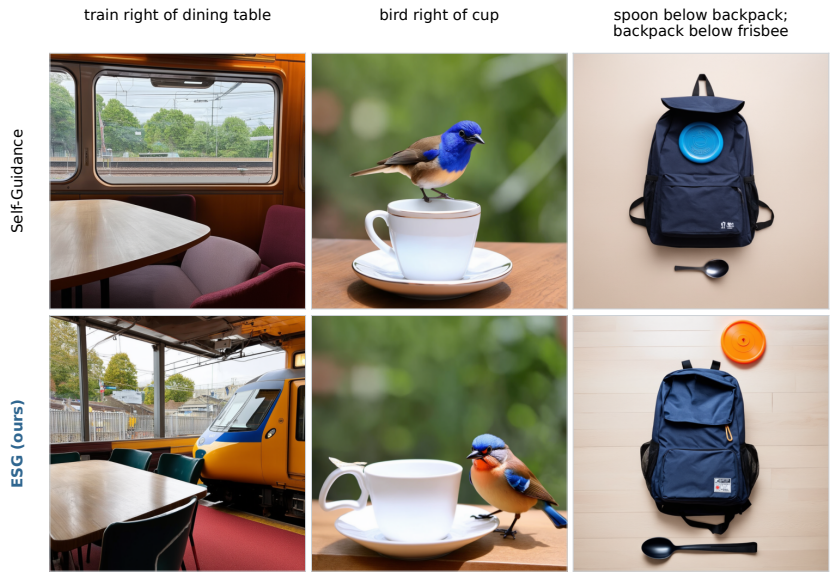


Figure 1: **Spatial composition with ESG.** Paired SD3 generations: GenEval (left), CompBench (middle), and a three-object chain (right). Both methods use medium-budget configurations with identical prompts and initial noise. Full-cohort scores appear in Tables 1 and 3.

ABSTRACT

Text-to-image models often generate the requested objects but fail to arrange them according to the prompt. We present Early Spatial Guidance (ESG), a training-free method for improving spatial composition in rectified flow transformers. ESG reads object locations from native joint attention and updates the latent state to satisfy relative spatial constraints during early sampling. The objective leaves absolute object positions unspecified and handles multiple objects by combining pairwise relations. Normalized updates control intervention strength, while attention-layer calibration adapts the spatial readout to each model. Gradients require only the transformer prefix leading to the selected readout. We analyze the sensitivity of this readout, the interaction of relation gradients, and how interventions propagate through the remaining sampling trajectory. Experiments on SD3 and SD3.5 show improved spatial alignment on public benchmarks and generalization to unseen object pairs. Comparisons with spatial-guidance methods and time-matched native sampling demonstrate a favorable accuracy–cost tradeoff. On controlled multi-object tasks, ESG outperforms the evaluated absolute-centroid objective at matched update budgets, with the clearest gains under repeated guidance. These findings show that early attention-based guidance can coordinate multiple spatial relations at competitive inference cost, without retraining the generator or requiring user-specified layouts.

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.

1 INTRODUCTION

An image containing a cup and a plate need not place the cup to the left of the plate. Text-to-image models can render recognizable objects while missing their relations; object-level benchmarks distinguish spatial correctness from object presence (Ghosh et al., 2023; Huang et al., 2023).

Attention-guided methods use internal object-location estimates to steer sampling (Epstein et al., 2023; Han et al., 2025; Rastegar et al., 2025). The challenge is useful control without expensive iterative refinement: attention centers are imperfect geometric proxies, and prompts specify relations rather than the absolute coordinates and separations imposed by some objectives.

Early Spatial Guidance (ESG) combines a relative objective with low-cost latent intervention (Figure 2). It reads object centers from native joint attention, averages their requested directional differences, and differentiates through the selected transformer prefix. The update norm is a fixed fraction of the latent norm. A fresh full prediction then advances the original sampler; weights and text conditioning stay fixed.

Our analysis characterizes native-centroid sensitivity and the interaction of relation gradients. Experiments compare complete controllers at measured costs, test whether extra native steps suffice, and isolate the effects of objective, timing and readout selection. Public benchmarks, unseen pairs and multi-object scenes evaluate spatial behavior on SD3 and SD3.5. Our contributions are:

- A training-free spatial controller that combines relative object geometry with normalized early latent updates, using only the prefix required by the selected attention readout.
- A method-specific analysis of native-readout stability and relation composition, supported by controlled tests of objective, timing, and readout selection.
- A paired evaluation of spatial accuracy and computation against full controllers, budget-constrained controllers, and time-matched native sampling.

2 EARLY SPATIAL GUIDANCE

Let z_k be the latent at sampling step k, t_k the model timestep, and c the encoded text conditioning. The input supplies entity phrases and directional relations $\mathcal{E}$: a deterministic parser handles supported two-object prompts, and controlled multi-object scenes supply relation graphs. Unsupported prompts retain native sampling and remain in the evaluation. ESG requires latent access and a differentiable entity-spatial readout; the joint-attention implementation shown here uses multimodal diffusion transformers (MMDiT) (Esser et al., 2024).

2.1 OBJECT CENTERS FROM NATIVE ATTENTION

At a selected block, let Q^I denote image queries and $\mathsf{K} = [\mathsf{K}^I; \mathsf{K}^T]$ image and text keys after the pretrained model's own normalizations. For head h and image query i, the native attention probability is

$$A_{ij}^h = \frac{\exp(\langle \mathsf{Q}_{hi}^I, \mathsf{K}_{hj} \rangle / \sqrt{d_h})}{\sum_{j' \in I \cup T} \exp(\langle \mathsf{Q}_{hi}^I, \mathsf{K}_{hj'} \rangle / \sqrt{d_h})}. \tag{1}$$

Here d_h is the head dimension. Retaining image keys in the denominator distinguishes this readout from renormalization over text tokens alone. For entity a with token group $\mathcal{T}_a$, image-grid cell u containing queries $\mathcal{P}_u$, and H heads, define

$$a_a(u) = \frac{1}{H|\mathcal{P}_u|} \sum_{h=1}^{H} \sum_{i \in \mathcal{P}_u} \sum_{j \in \mathcal{T}_a} A_{ij}^h, \qquad \mu_a = \frac{\sum_u a_a(u) x_u}{\sum_u a_a(u)}. \tag{2}$$

The coordinate $x_u \in [0,1]^2$ is the cell center; image y increases downward. Repeated mentions of an entity share one token group and one center. Attention reconstruction uses FP32 and does not replace the pretrained model's forward attention operator.

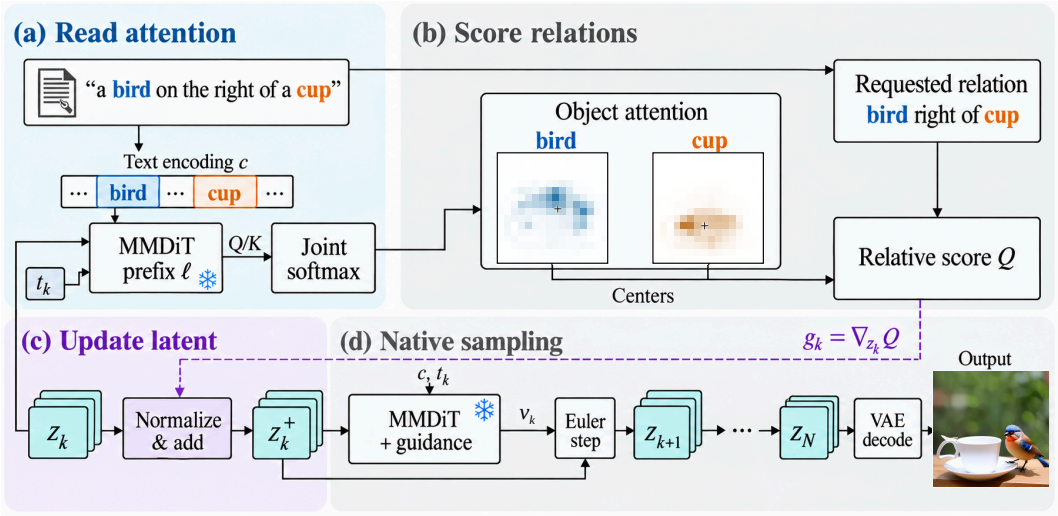


Figure 2: **ESG in an MMDiT sampling step.** (a) At state z_k, text encoding c and timestep t_k, the frozen prefix ends at block ℓ's native queries/keys (Q/K). Joint softmax uses image and text keys. (b) Entity-map centers μ_a (crosses) define the score Q (Eqs. 2–3). (c) Its gradient g_k gives z_k^+ (Eq. 4). (d) A fresh full velocity v_k and z_k^+ enter Euler (Algorithm 1, lines 10–11). The ellipsis repeats sampling, with ESG only at $k \in \mathcal{K}$, until z_N is decoded. Black arrows: forward data; purple dashes: latent gradient; snowflakes: frozen weights. The measured 16×16 maps are from SD3 block 12 before the first update at $k = 2$; colors are scaled separately per entity. The unretouched output is from the same run after eight interventions and all 28 sampling steps.

2.2 Relative objective and normalized update

For relation $e = (a, b, r)$, let d_r select its horizontal or vertical axis and $s_r = -1$ for left/above and $+1$ for right/below. ESG maximizes

$$q_e = s_r d_r^\top (\mu_a - \mu_b), \qquad Q(z_k, c) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} q_e. \tag{3}$$

A larger q_e favors the requested ordering. A common center translation or positive rescaling of an entity map leaves Q unchanged. These are readout properties, not generator invariances. The timestep and readout block are held fixed when evaluating and differentiating Q.

At an intervention step $k \in \mathcal{K}$, we compute $g_k = \nabla_{z_k} Q$ and set

$$R_k = \rho \, \|z_k\|_2 , \qquad \delta_k = R_k \frac{g_k}{\|g_k\|_2}, \qquad z_k^+ = z_k + \delta_k. \tag{4}$$

A zero gradient gives zero displacement. The relative radius ρ controls the update independently of the numerical scale of Q. Given the original sampler step S_k, the next state is

$$z_{k+1} = S_k(z_k + \mathbf{1}_{k \in \mathcal{K}} \delta_k; c). \tag{5}$$

In our Euler implementation, $S_k(z; c) = z + h_k v(z, t_k, c)$, where $h_k = \sigma_{k+1} - \sigma_k < 0$ is the scheduler's noise-level increment, not the difference of scaled model timesteps. The full velocity v is recomputed at the updated state using the original prediction rule (including CFG or CDG where applicable). ESG's spatial objective uses only the conditional readout; it does not itself combine conditional and reference predictions.

Why a prefix is sufficient. For a readout $Q = q \circ H_\ell$, the chain rule gives $\nabla_z Q = DH_\ell(z)^\top \nabla q(H_\ell(z))$. Later blocks are not ancestors of this objective and can be omitted on the gradient call. This is exact for the selected readout, not an approximation to an image-quality gradient. With M interventions, the added cost is $M(C_{\text{prefix,fwd}} + C_{\text{prefix,bwd}})$, including readout operations; all N normal predictions remain complete. Appendix A.1 specifies batching, finite-precision norm correction and the interface required for other architectures.

Algorithm 1 Early Spatial Guidance (ESG)

Input: latent z_0, text conditioning c, entity-token groups $\{\mathcal{T}_a\}$, relations $\mathcal{E}$, block ℓ, indices $\mathcal{K}$, radius ρ, timesteps $\{t_k\}$ and noise levels $\{\sigma_k\}_{k=0}^{N}$.
Fixed: pretrained models and original prediction rule v (including base guidance).

for $k = 0, \dots, N-1$ **do**
  $z_k^+ \leftarrow z_k$
  if $k \in \mathcal{K}$ and the input relations are supported **then**
    Capture conditional $\mathsf{Q}^I, \mathsf{K}^I, \mathsf{K}^T$ at block ℓ from z_k, t_k, c.
    Reconstruct native joint attention; compute maps and centers (Eq. 2).
    Form Q from all requested edges using Eq. 3.
    $g_k \leftarrow \nabla_{z_k} Q$ (differentiate the readout and prefix; weights stay fixed).
    if $\|g_k\|_2 > 0$ **then** $z_k^+ \leftarrow z_k + \rho \|z_k\|_2 g_k / \|g_k\|_2$.
  end if
  $v_k \leftarrow v(z_k^+, t_k, c)$ (fresh full prediction with the original guidance).
  $z_{k+1} \leftarrow z_k^+ + (\sigma_{k+1} - \sigma_k) v_k$ (Euler step).
end for; **return** $\mathrm{Decode}(z_N)$.

3 Analysis of the Spatial Intervention

We analyze native-centroid sensitivity, relation composition and downstream response. Local calculations fix step k and omit its subscript: $z = z_k$, $g = g_k$, $R = R_k$ and $\delta = \delta_k$.

Native-readout sensitivity. Differentiating an attention-weighted center with respect to its map gives

$$\frac{\partial \mu_a}{\partial a_a(u)} = \frac{x_u - \mu_a}{M_a}, \qquad M_a = \sum_u a_a(u). \tag{6}$$

The apparent inverse-mass factor does not imply the same sensitivity to native logits. Let s_{ij} be the joint-attention logits, $b_i = \sum_{j \in \mathcal{T}_a} \mathrm{softmax}(s_i)_j$, and ω_i the fixed pooling weights (with heads included in i). For one coordinate, $\mu = \sum_i \omega_i b_i x_i / M$ and $M = \sum_i \omega_i b_i$.

Proposition 1 (Native-centroid sensitivity)**.** *For finite logits, a nonempty entity token group, fixed positive pooling weights and coordinates in an interval of width W,*

$$\|\nabla_s \mu\|_1 \le W, \qquad |\mu(s+v) - \mu(s)| \le W \tanh \|v\|_\infty, \qquad |D^2 \mu(s)[v,v]| \le \tfrac{9}{2} W \|v\|_\infty^2. \tag{7}$$

The proof follows the native softmax and quotient derivatives, causing the mass factors to cancel after summation. These bounds are independent of entity mass and token count in real arithmetic. They do not establish object presence or numerical stability at underflow. Bounds on the latent-to-logit derivatives are additionally required to control latent curvature (Appendix A.2).

What the relation objective leaves free. For a left-to-right pair, ESG maximizes $q = \mu_{b,x} - \mu_{a,x}$. Our interior-anchor control instead minimizes the two-coordinate L1 distance to $(1/4, 1/2)$ and $(3/4, 1/2)$. Its exact decomposition (Eq. 15, Appendix A) shows additional preferences for pair centering, fixed separation and orthogonal alignment. Changing only this objective therefore tests distinct geometric preferences. Not all absolute losses differ: single-axis boundary targets give the same normalized direction as ESG.

Combining relations. Let $g_e = \nabla q_e$ and $g = |\mathcal{E}|^{-1} \sum_e g_e \neq 0$. If ∇q_e is L_e-Lipschitz along the update segment, then

$$q_e(z + \delta) - q_e(z) \ge R \frac{g_e^\top g}{\|g\|_2} - \tfrac{1}{2} L_e R^2. \tag{8}$$

Aligned relation gradients reinforce each other; maximizing their average need not improve every edge. More specifically, with $q = B\mu$ for the signed relation matrix B and equal weights w, $Q = (B^\top w)^\top \mu$. In a horizontal chain, the middle center cancels directly: $(\mu_{b,x} - \mu_{a,x}) + (\mu_{c,x} - \mu_{b,x}) = \mu_{c,x} - \mu_{a,x}$. It can still move through the shared latent. The graph analysis also gives an exact score threshold for all readout relations in a coordinate-box relaxation; this threshold is nonempty for our corner and grid graphs, but not for chains (Appendix A.2). We therefore evaluate both mean relation accuracy and all-relations success.

From readout control to generated images. Let Φ denote the remaining sampling and decoding computation, including later interventions, and F a smooth terminal spatial surrogate. With $b = D\Phi(z)^\top \nabla F(\Phi(z))$, the first-order change is

$$\Delta F \approx R \, \|b\|_2 \cos \angle(b, g). \tag{9}$$

Local readout ascent is useful only insofar as it aligns with downstream geometry. Appendix A.2 makes this condition explicit through a readout–surrogate mismatch bound. We do not certify those neighborhood assumptions on real models. Instead, matched timing and layer controls test the relevant choices empirically; official discrete success remains an experimental outcome, not a theorem.

4 Experiments

4.1 Evaluation protocol

Models, inputs, and metrics. We evaluate SD3-Medium and SD3.5-Large (Esser et al., 2024; Stability AI, 2024) at 1024×1024 with FP16 model weights and 28 Euler steps unless otherwise stated. Within each comparison, methods share prompts and initial noise. The main SD3 evaluation uses 100 GenEval position prompts with four seeds (400 images per configuration) and all 300 T2I-CompBench spatial prompts with two seeds (600 images). The controlled composition set contains 100 chain, corner, and grid scenes with four seeds on SD3 and two on SD3.5. SD3.5 public and unseen-pair cohorts each contain 200 images per configuration. Unseen pairs exclude unordered object-pair overlap with the public pool.

Position is official GenEval spatial success; Joint additionally requires exactly one instance of each object. CompBench Spatial is its official continuous score multiplied by 100, not a success percentage. Multi-object evaluation reports average edge correctness (Mean relation), all-edge correctness (All relations), and exact-count/all-edge Joint success. We retain unsupported prompts in all denominators and report object presence, counts, continuous geometry and image-quality diagnostics in the appendix.

Baselines and budgets. We compare CFG, Prompt Reinjection, Rectified-CFG++, Self-Guidance, STORM and InfSplign on the same SD3 backbone. The main table fixes CFG scale 7. Baseline adapters retain the released objectives and disclose their architecture-specific changes in Appendix B. Full spatial controllers retain their complete schedules. Low and medium configurations satisfy development-selected $1.5\times$ and $2\times$ native-generation ceilings. These are common ceilings, not assertions of exactly equal elapsed time. ESG and Self-Guidance share a 64-case pair/multi-object development set; STORM and InfSplign share a separate 64-case pair-only set. Both exclude formal SD3 object combinations. Actual runtimes, search protocols and objective-only controls complement these comparisons.

ESG settings and inference cost. SD3 uses block 12: one update at index 2 with $\rho = .005$, four low-budget updates with $\rho = .005$, or eight medium-budget updates with $\rho = .01$. SD3.5 uses a calibrated block-24 readout; the scale-3.5 configuration uses one update for public/unseen pairs and four for multi-object scenes. Text conditioning is cached equally for all methods. Table 1 reports sampling time; budget selection and Table 2 use generation time including cache lookup and decoding, but not model loading or initial text encoding. Both clocks are measured on RTX 3090 GPUs and are never mixed on one axis.

Uncertainty. Paired cluster bootstraps resample unordered object sets with repeated seeds kept together. Complete-method comparisons use a 53-comparison Bonferroni family; objective and native-step controls retain their original separate families (Appendix C). Boldface denotes point estimates, not significance.

4.2 Spatial accuracy at measured computational cost

A single ESG update already improves both public spatial metrics (Table 1). Medium-budget ESG obtains the highest Position and Spatial point estimates among the evaluated configurations. It exceeds full Self-Guidance, STORM and InfSplign on these metrics while using less sampling time.

Table 1: **Spatial accuracy and inference cost on identical SD3 inputs.** All 14 configurations use 400 GenEval position and 600 CompBench spatial images. Position/Joint are percentages; Spatial is the official continuous score $\times 100$. T_G, T_C are sampling seconds per image. Full controllers retain their complete schedules; low/medium groups use common generation-time ceilings. Bold marks the best accuracy in each group. All methods use the same pretrained backbone and CFG scale 7 except the development-selected Rectified-CFG++ guidance rule.

Method	Position ↑	Joint ↑	Spatial ↑	T_G ↓	T_C ↓
Standard guidance and one-update ESG					
CFG	26.50	20.00	30.76	9.51	9.46
Prompt Reinjection	31.50	22.75	32.41	9.80	9.73
Rectified-CFG++ (14 steps)	20.75	12.50	26.93	9.52	9.49
Rectified-CFG++ (28 steps)	28.00	19.75	28.12	19.06	19.00
ESG (one update)	**37.00**	**24.50**	**35.71**	10.37	10.06
Full spatial-controller schedules					
Self-Guidance	43.50	29.75	36.83	85.66	64.89
STORM	49.50	23.50	31.26	223.16	181.26
InfSplign	**50.75**	**37.50**	**39.04**	41.62	41.78
Low budget: $1.5\times$ *generation-time ceiling*					
Self-Guidance	34.25	23.50	34.09	12.73	11.74
ESG	**47.25**	**34.00**	**36.98**	12.75	11.86
Medium budget: $2\times$ *generation-time ceiling*					
Self-Guidance	39.00	27.50	36.60	15.88	14.08
STORM	51.00	**38.75**	37.92	17.55	17.09
InfSplign	47.75	37.00	40.08	16.40	16.34
ESG	**56.50**	38.25	**41.90**	16.05	14.15

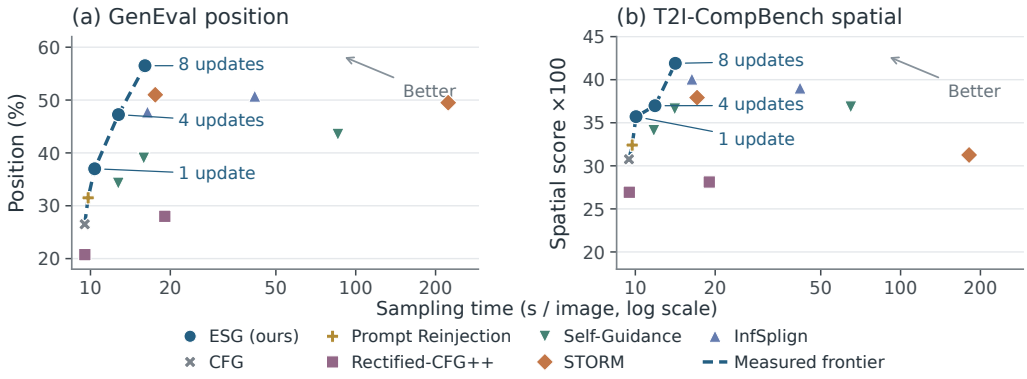


Figure 3: **Accuracy–cost tradeoffs for Table 1.** All 14 configurations are shown; colors and markers identify methods. Both panels use sampling time on a log scale. Dashed lines join the non-dominated measured points, not predictions for untested budgets. GenEval Position is a percentage; CompBench Spatial is a continuous score $\times 100$.

This comparison concerns the complete adapted controllers; it does not isolate which implementation component produces the speed difference.

Within the common budgets, ESG improves Position over Self-Guidance by 13.0 points at the low budget and 17.5 at the medium budget, with positive simultaneous intervals. Their measured generation times are almost equal at each operating point. Medium-budget ESG also has higher Position and Spatial point estimates than constrained STORM and InfSplign, although these simultaneous intervals include zero. STORM leads slightly on Joint success. Figure 3 summarizes the measured tradeoffs; no claim of a globally optimal frontier is made. CLIP and aesthetic diagnostics in Appendix C show that stronger spatial control can accompany lower aesthetic scores.

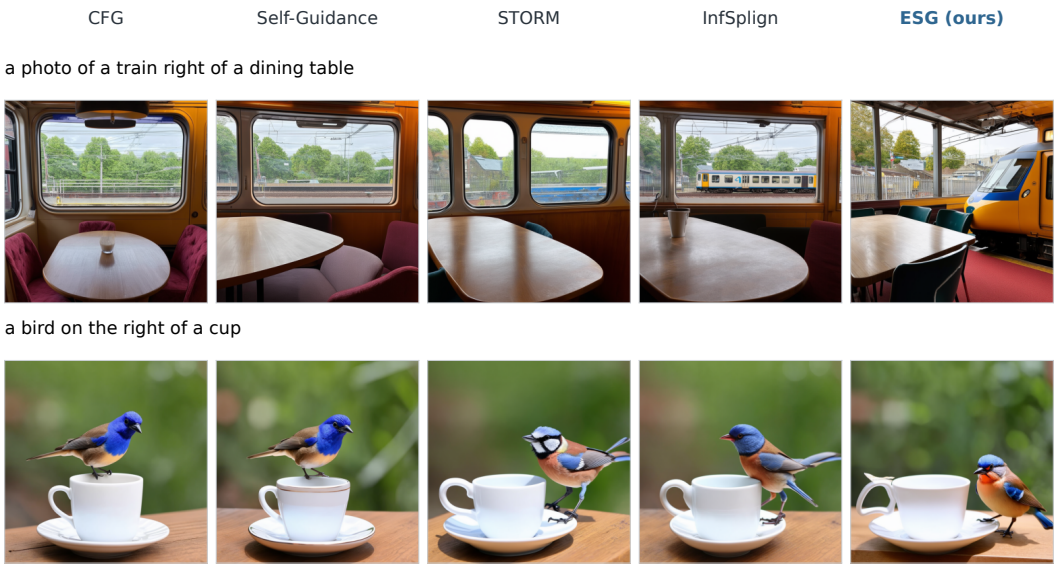


Figure 4: **Paired SD3 generations on public spatial prompts.** The four spatial controllers use medium-budget configurations with the same prompts and initial noise. Original image frames are unretouched. Examples are selected for clarity and relation coverage, not randomly sampled; Table 1 reports full-cohort performance.

Table 2: **Time-matched native-step controls: 200 paired images per arm.** Position/Joint are percentages. T is generation seconds per image from isolated development timing, not concurrent evaluation batches. ESG retains 28 sampling steps; M is its number of interventions. Native step counts are selected without quality scores.

Model	Method	Steps	M	Position ↑	Joint ↑	T ↓
SD3	Native	28	0	26.00	21.00	9.52
SD3	Native (time-matched)	48	0	26.50	22.50	16.11
SD3	ESG	28	8	56.00	38.50	16.04
SD3.5	Native	28	0	26.50	21.00	31.96
SD3.5	Native (time-matched)	40	0	25.00	20.50	45.57
SD3.5	ESG	28	1	39.00	32.00	44.78

Does spending the same time on more steps suffice? Additional native steps are selected by runtime alone on eight disjoint development prompts with two reverse-balanced repeats: 48 steps for SD3 and 40 for SD3.5. Mean generation times differ from ESG by less than 2% (5% tolerance). Evaluation uses 200 paired images per arm and model, reusing existing 28-step and ESG images. This SD3 subset differs from the 400-image main cohort. Increasing native steps does not reproduce ESG's improvement (Table 2): ESG gains 29.5 Position points on SD3 and 14.0 on SD3.5. Joint gains are also positive, and all four simultaneous intervals exclude zero (Appendix I).

4.3 Multiple relations and model transfer

Relative constraints improve more than individual pairs. At the medium budget, ESG improves mean relation accuracy over Self-Guidance by 9.06 points and increases the fraction of fully correct layouts (Table 3). Gains appear in chains, corners and grids; the topology breakdown uses the same images, not a new independent test (Appendix C.1). Selected paired images appear in Appendix D.

The same intervention rule improves SD3.5 after readout-layer calibration (Table 4). At scale 3.5, ESG improves all three spatial endpoints. At scale 7, TexTailor leads on unseen pairs, while ESG leads on public Position and multi-object accuracy. Calibration uses distinct seeds on public prompts, not a prompt-disjoint set; unseen pairs provide the composition holdout. The deeper readout incurs greater measured overhead.

Table 3: **SD3 multi-object composition (400 images per configuration).** Mean relation, All relations and Joint are percentages. Time includes cache lookup, sampling and decoding. Comparisons are paired within each budget; bold indicates the higher accuracy.

Method	Mean relation ↑	All relations ↑	Joint ↑	Time ↓
Low budget: $1.5\times$ *generation-time ceiling*				
Self-Guidance	23.31	3.50	1.50	13.35
ESG	**28.94**	**5.25**	**2.50**	13.53
Medium budget: $2\times$ *generation-time ceiling*				
Self-Guidance	27.56	4.50	3.25	16.55
ESG	**36.62**	**11.25**	**7.25**	16.85

Table 4: **SD3.5 transfer (200 paired images per configuration and cohort).** Public/Unseen: Position; Multi: Mean relation (%). Time is public sampling seconds per image. ESG uses one public/unseen update and four multi-object updates. † Prompt Reinjection development overlaps 32 unseen-pair prompts. Bold denotes the best accuracy per group.

Method	Public ↑	Unseen pairs ↑	Multi ↑	Time ↓
CFG scale 3.5				
CFG	26.50	26.00	18.88	31.72
Prompt Reinjection†	30.50	29.50	25.12	31.76
Rectified-CFG++ ($w = 4.5$)	24.00	25.50	17.75	65.68
ESG	**39.00**	**41.50**	**36.12**	44.47
CFG scale 7				
CFG	29.00	30.00	19.75	31.72
TexTailor	34.00	**40.00**	24.75	31.75
ESG	**38.00**	33.00	**29.00**	44.88

4.4 Controlled tests of the mechanism

Does the relative objective matter? An objective-only comparison holds the main SD3 CFG-7 configuration fixed: layer, radius, eight-update schedule, prompts and initial noise. It replaces only the relative objective with the interior-anchor L1 loss of Eq. 15. Table 5 supports the relative objective's contribution to Position and multi-object correctness; the CompBench interval includes zero. The gain is therefore not explained solely by early intervention or prefix computation. Conversely, the objectives perform similarly with a single pairwise update in the complementary CDG setting. Repeated controls on both models yield larger compositional differences (Appendix E), consistent with the distinct preferences in Eq. 15. These controls isolate one particular absolute objective, not every possible alternative.

Timing, radius and readout. On a separate 200-case SD3 control with CDG, a single early update reaches 47.5% Position versus 36.5% for an index-7 update at the same radius. Smaller per-step radii reduce the benefit of four updates. On paired SD3.5 cases, the selected block-24 readout reaches 41.0% versus 28.0% for block 12. Finite perturbations also show that readout gains become less linear as radius increases. These controlled observations support early intervention and readout calibration in the evaluated settings; they do not prove universal early optimality. Appendix F reports the complete controls.

Compatibility with the base guidance. ESG also improves a CDG trajectory without changing its spatial-update rule. The twelve-update SD3/CDG configuration reaches 78.25% Position. This is a separate operating point, not part of the CFG budget ranking: both its schedule and radius differ from the main eight-update setting. The compatible interface and the empirical benefit should not be confused with independence of performance from the base guidance.

Table 5: **Isolating the objective under the main ESG configuration.** Differences are ESG minus the absolute objective. Position and relation scores are percentages; CompBench Spatial is a continuous score $\times 100$. Intervals are simultaneous over these four primary contrasts. All non-objective settings are identical.

Metric	Pairs	Absolute	ESG	Difference	Simultaneous interval
Position	400	50.75	56.50	+5.75	[+0.25, +11.25]
CompBench spatial	600	40.43	41.90	+1.47	[-1.03, +4.04]
Mean relation	400	27.62	36.62	+9.00	[+4.94, +13.03]
All relations	400	5.75	11.25	+5.50	[+1.25, +10.00]

5 Related Work

Attention-based spatial control. Self-Guidance extracts location, shape and appearance from internal representations (Epstein et al., 2023); Attend-and-Excite strengthens subject activations (Chefer et al., 2023). STORM optimizes spatial transport and already emphasizes early intervention (Han et al., 2025). InfSplign combines relative-position, presence and balance objectives (Rastegar et al., 2025). ESG does not claim these ingredients individually as new. It studies a shared-center directional objective together with exact selected-prefix gradients and normalized early latent updates, and separates their contributions with matched controls. Learn-to-Steer learns a spatial loss over attention maps (Yiflach et al., 2026); ReaDiT and CoMPaSS introduce learned spatial prediction or spatial supervision (Mahajan et al., 2026; Zhang et al., 2025). Stitch constructs intermediate layouts and composes objects (Bader et al., 2025). ESG requires neither auxiliary spatial training nor user-specified coordinates.

Prediction and conditioning guidance. CFG and CDG form predictions from alternative conditions (Ho & Salimans, 2022; Han et al., 2026); APG, CFG-Ctrl and Rectified-CFG++ modify prediction guidance (Sadat et al., 2025; Wang et al., 2026a; Saini et al., 2025). Prompt Reinjection and TexTailor modify transformer conditioning (Yao et al., 2026; Li et al., 2026). ESG instead changes the latent before the original sampling step. Spatial benchmarks (Ghosh et al., 2023; Huang et al., 2023; Wang et al., 2026b) and evidence of seed sensitivity (Li et al., 2025) motivate our paired, object-level evaluation.

6 Conclusion

ESG uses native attention geometry to allocate a small number of spatially directed latent updates. Relative objectives, selected-prefix computation and early intervention together produce favorable measured accuracy–cost tradeoffs. Time-matched native controls rule out extra sampling time alone as the explanation, while objective-only and mechanism controls clarify which choices contribute to the observed benefit.

Limitations. The evaluation presented here covers two related pretrained transformer models and directional spatial tasks. Future work will broaden evaluation to generator families beyond Stable Diffusion, additional relation types and sampling procedures. Readout-layer calibration is model-dependent. ESG does not guarantee object counts or appearance preservation, and stronger spatial control can reduce aesthetic scores. Full-controller comparisons depend on the disclosed adaptations and development protocols. The theory characterizes readout geometry and conditional downstream response; it does not certify image-level success.

Reproducibility Statement

The appendix provides complete theoretical proofs, model and controller settings, baseline adaptations, evaluation definitions, paired uncertainty estimates and additional qualitative results. The inference algorithm is specified in Algorithm 1; the source code and configurations specify the finite-precision implementation and evaluation procedures.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval during early investigation; language polishing, revision feedback including checks of mathematical derivations and reporting consistency, and editing of the mechanism illustration's diagram elements after author drafting; and implementation of experimental code from author-specified experiment matrices and configurations, with human review. The authors led the research conception, experiment design and manuscript development, and are responsible for checking the references, code, results and final claims. Measured attention maps are plotted from recorded arrays; generated-image examples are unretouched outputs of the evaluated models.

REFERENCES

Jessica Bader, Mateusz Pach, Maria A. Bravo, Serge Belongie, and Zeynep Akata. Stitch: Training-Free Position Control in Multimodal Diffusion Transformers. *arXiv preprint arXiv:2509.26644*, 2025. URL https://arxiv.org/abs/2509.26644.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurate computation of the log-sum-exp and softmax functions. arXiv:1909.03469, 2019. URL https://arxiv.org/abs/1909.03469.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Transactions on Graphics*, 42(4):1–10, 2023. doi: 10.1145/3592116. URL https://doi.org/10.1145/3592116.

Samuel N. Cohen and Eliana Fausti. Hyperbolic contractivity and the hilbert metric on probability measures. arXiv:2309.02413, 2023. URL https://arxiv.org/abs/2309.02413.

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion Self-Guidance for Controllable Image Generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16222–16239. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3469b211b829b39d2b0cfd3b880a869c-Abstract-Conference.html.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. arXiv:1704.00805, 2017. URL https://arxiv.org/abs/1704.00805.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Shilong Han, Yuming Zhang, and Hongxia Wang. Guiding Diffusion Models with Semantically Degraded Conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43653–43663, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Han_Guiding_Diffusion_Models_with_Semantically_Degraded_Conditions_CVPR_2026_paper.html.

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)*, pp. 18401–18410, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Han_Spatial_Transport_Optimization_by_Repositioning_Attention_Map_for_Training-Free_Text-to-Image_CVPR_2025_paper.html.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

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*, volume 36, pp. 78723–78747. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract.html.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127:203–244, 2011. URL https://arxiv.org/abs/0811.1067.

Binglei Li, Mengping Yang, Zhiyu Tan, Junping Zhang, and Hao Li. TexTailor: Inference-Time Textual Guidance Tailoring for Multimodal Diffusion Transformers. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2601.02211. To appear.

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Enhancing Compositional Text-to-Image Generation with Reliable Random Seeds. In *International Conference on Learning Representations*, volume 2025, pp. 45444–45465, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/708e0d691a22212e1e373dc8779cbe53-Abstract-Conference.html.

Jay Mahajan, Chang Liu, Rauf Makharov, Viraj Shah, Alexander Schwing, and Svetlana Lazebnik. ReaDiT Guidance: Control for Image and Video Generation using Diffusion Transformer Features. *arXiv preprint arXiv:2609.04649*, 2026. URL https://arxiv.org/abs/2609.04649.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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. URL https://arxiv.org/abs/2512.17851.

Seyedmorteza Sadat, Otmar Hilliges, and Romann Weber. Eliminating Oversaturation and Artifacts of High Guidance Scales in Diffusion Models. In *International Conference on Learning Representations*, volume 2025, pp. 84127–84153, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d1450d6c10c6b6cf1b80964357f5fa08-Abstract-Conference.html.

Shreshth Saini, Shashank Gupta, and Alan C. Bovik. Rectified-CFG++ for Flow Based Models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.07631.

Stability AI. Stable Diffusion 3.5 Large, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-large. Model card.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. CFG-Ctrl: Control-Based Classifier-Free Diffusion Guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11437–11447, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_CFG-Ctrl_Control-Based_Classifier-Free_Diffusion_Guidance_CVPR_2026_paper.html.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in Its Place: Benchmarking Spatial Intelligence of Text-to-Image Models. In *International Conference on Learning Representations*, volume 2026, pp. 44846–44873, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4a5c76c63f83ea45fb136d62db6c7104-Abstract-Conference.html.

Yuxuan Yao, Yuxuan Chen, Hui Li, Kaihui Cheng, Qipeng Guo, Yuwei Sun, Zilong Dong, Jingdong Wang, and Siyu Zhu. Prompt Reinjection: Alleviating Prompt Forgetting in Multimodal Diffusion Transformers for Text-to-Image Generation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06886.

Sapir Esther Yiflach, Yuval Atzmon, and Gal Chechik. Data-Driven Loss Functions for Inference-Time Optimization in Text-to-Image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3525–3535, 2026. URL https://learn-to-steer-paper.github.io/.

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)*, pp. 15253–15265, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_CoMPaSS_Enhancing_Spatial_Understanding_in_Text-to-Image_Diffusion_Models_ICCV_2025_paper.html.

A THEORETICAL DETAILS

We use the main text's fixed-step convention, $z = z_k$, $g = g_k$, $R = R_k$ and $\delta = \delta_k$. The scalar Q is the relative spatial score, distinct from the native query projection Q^I. Q_ℓ below makes the otherwise fixed readout block explicit.

Centroid derivative and invariance. For $N_e = \sum_u a_e(u)x_u$ and $M_e = \sum_u a_e(u) > 0$, differentiating N_e/M_e gives Eq. 6. The weighted sum of derivatives vanishes since $\sum_u a_e(u)(x_u - \mu_e) = 0$. Multiplication of a map by a positive constant cancels in its centroid; a common translation cancels in each pairwise difference. These identities describe the spatial statistic, not an invariance of the whole image generator.

Proposition 2 (Bounded first-order control)**.** *For nonzero $g = \nabla Q(z)$, the unique maximizer of $g^\top \delta$ over $\|\delta\|_2 \le R$ is $\delta_* = Rg/\|g\|_2$. If ∇Q is L-Lipschitz on the update segment, then*

$$Q(z + \delta_*) - Q(z) \ge R\|g\|_2 - \tfrac{1}{2}LR^2. \tag{10}$$

Proof. Cauchy–Schwarz gives $g^\top \delta \le \|g\|_2 \|\delta\|_2 \le R\|g\|_2$, with unique equality at $\delta_* = Rg/\|g\|_2$ for $g \neq 0$. Moreover,

$$Q(z+\delta) - Q(z) - g^\top \delta = \int_0^1 [\nabla Q(z + u\delta) - g]^\top \delta \, du \tag{11}$$

$$\ge -\int_0^1 Lu\|\delta\|_2^2 \, du = -\tfrac{1}{2}L\|\delta\|_2^2. \tag{12}$$

Substitution proves Eq. 10. Rounding error e changes the first-order gain by at most $\|g\|_2 \|e\|_2$; finite-precision implementation measures the realized displacement. A positive affine transformation of Q preserves the normalized direction in exact arithmetic.

Relation-gradient interactions. Differentiation commutes with the finite sum in Eq. 3. Taking the directional derivative of each q_e along $Rg/\|g\|_2$ yields Eq. 8. A positive average derivative does not require all $g_e^\top g$ to be positive. For consistent geometric constraints, shared-coordinate terms can even cancel in the aggregate objective: for a horizontal chain, $(\mu_{b,x} - \mu_{a,x}) + (\mu_{c,x} - \mu_{b,x}) = \mu_{c,x} - \mu_{a,x}$. More generally, define $b_v = |\mathcal{E}|^{-1}(\sum_{e:a=v} s_r d_r - \sum_{e:b=v} s_r d_r)$. The objective and its latent gradient satisfy

$$Q = \sum_v b_v^\top \mu_v, \qquad \nabla_z Q = \sum_v (D_z \mu_v)^\top b_v. \tag{13}$$

This separates the graph's directional coefficients from the generator's spatial Jacobians. The middle centroid of a collinear chain has zero direct coefficient; it can still move through the shared latent state. Orthogonal edges at a corner contribute to different coordinates. Thus the same analytic objective induces different control directions for different graph topologies. We report both edge-average and all-edge outcomes rather than treating a sum of margins as a guarantee for every relation.

Propagation along the sampler. For $S_k(z) = z + h_k v(z, t_k, c)$, with $h_k = \sigma_{k+1} - \sigma_k$, the first-order perturbation propagates through $DS_k = I + h_k D_z v(z, t_k, c)$. Composing the sampler Jacobians, the Jacobians of any subsequent ESG updates, and the decoder Jacobian gives $D\Phi$. If $H = F \circ \Phi$ has a locally Lipschitz gradient with constant L_H, then

$$H(z + \delta_*) - H(z) \ge R\|\nabla H(z)\|_2 \cos\angle(\nabla H(z), g) - \tfrac{1}{2}L_H R^2. \tag{14}$$

This is a local statement about a smooth surrogate. Official image-level success is discrete. Readout AUROC measures prediction, and does not establish the gradient alignment needed by this expression or prove that early intervention is universally optimal.

Finite-perturbation measurements. Four latent states and 56 signed perturbations agree with the predicted local sign in all measured cases. Mean observed-to-linear positive readout change is 0.976, 0.820, and 0.511 at relative radii 0.0005, 0.005, and 0.02. These are mechanism diagnostics, distinct from the large image-level evaluations.

Interior-anchor decomposition. For a left-to-right pair, write $q = \mu_{b,x} - \mu_{a,x}$, $m = (\mu_a + \mu_b)/2$, and $r = \mu_{b,y} - \mu_{a,y}$. The evaluated control satisfies

$$\begin{aligned} L_{\text{abs}} &= \|\mu_a - (\tfrac{1}{4}, \tfrac{1}{2})\|_1 + \|\mu_b - (\tfrac{3}{4}, \tfrac{1}{2})\|_1 \\ &= \max\{2|m_x - \tfrac{1}{2}|, |q - \tfrac{1}{2}|\} + \max\{2|m_y - \tfrac{1}{2}|, |r|\}. \end{aligned} \tag{15}$$

For scalars u, v, $|u - v| + |u + v| = 2\max(|u|, |v|)$. Write $\mu_{a,x} = m_x - q/2$ and $\mu_{b,x} = m_x + q/2$. Their deviations from $1/4$ and $3/4$ are $(m_x - 1/2) \mp (q - 1/2)/2$. Applying the identity to this pair and to the orthogonal-coordinate deviations proves Eq. 15. Away from the ties between the two terms of each maximum, the anchor gradient follows the larger error in that coordinate: common translation or target separation on the relation axis, and common translation or alignment on the orthogonal axis. It need not follow the requested ordering margin.

For the same state and radius, let $g_{\text{abs}} = -\nabla_z L_{\text{abs}} \neq 0$ and $g = \nabla_z Q \neq 0$. The ratio between their first-order increases of Q is

$$\frac{g^\top (R g_{\text{abs}} / \|g_{\text{abs}}\|_2)}{g^\top (R g / \|g\|_2)} = \cos \angle(g, g_{\text{abs}}). \tag{16}$$

The Jacobian from object coordinates to the latent can couple these components, so coordinate-space orthogonality does not imply orthogonal latent gradients. The expression identifies the local efficiency measured against the relation objective. Repeated updates visit different regions of the piecewise anchor loss, giving a mechanism by which initially similar directions can diverge. The matched controls measure the terminal effect; the local calculation does not assume that image-level success is a differentiable loss.

Sampler interface and implementation scope. Write $G_k(z) = z + \delta_k(z)$ at an intervention step. ESG replaces S_k by $S_k \circ G_k$. For a one-step sampler with a locally Lipschitz Jacobian and a fixed perturbation, Taylor expansion gives

$$S_k(z + \delta; c) - S_k(z; c) = DS_k(z; c)\delta + O(\|\delta\|_2^2). \tag{17}$$

Thus the update is defined by the conditional spatial readout, whereas its downstream effect depends on the chosen sampler. A direct conditional prediction, $v = v_c$, is an admissible Euler field without any unconditional branch. Our SD3 adapters capture the conditional member of a two-branch forward pass to preserve the numerical path of the reported experiments. This batching choice is implementation-specific. Extending the readout requires native attention normalization, positional encoding, token ordering, and latent representation. The reported image-generation evidence covers SD3 and SD3.5.

Prediction rules in the evaluated configurations. The principal comparisons use CFG, $v = v_u + \gamma(v_c - v_u)$; the complementary configuration uses CDG, $v = v_d + \gamma(v_c - v_d)$. Here γ is the guidance scale, while u and d denote unconditional and degraded conditioning. Both compose with the same spatial-update rule. ESG optimizes the conditional attention readout. At a fixed latent, timestep, and prompt, changing the base guidance leaves that conditional readout unchanged. The subsequent velocity does change. Writing v_u for the unconditional prediction and v_d for the degraded-condition prediction gives

$$v_\gamma^{\text{CDG}} - v_\gamma^{\text{CFG}} = (\gamma - 1)(v_u - v_d), \qquad DS_k^{\mathcal{B}} = I + h_k D_z v_\gamma^{\mathcal{B}}, \quad \mathcal{B} \in \{\text{CFG}, \text{CDG}\}. \tag{18}$$

The one-step response to the same intervention is $DS_k^{\mathcal{B}}\delta + O(\|\delta\|_2^2)$; the remaining response also includes subsequent sampler, ESG-update and decoder Jacobians. Thus the base guidance changes both the states at which ESG is applied and how its perturbation is propagated. This explains why compatibility does not imply identical gains and motivates reporting CFG and CDG operating families separately. The measured CDG advantage is an empirical result, not a universal ordering implied by these Jacobians.

Relative and absolute targets. For centroids in the unit image, the horizontal boundary-target loss obeys $|\mu_{a,x} - 0| + |\mu_{b,x} - 1| = 1 - (\mu_{b,x} - \mu_{a,x})$. Its gradient direction is therefore identical to maximizing the corresponding relative margin. A two-coordinate interior-anchor loss, by contrast, also constrains common translation and the orthogonal axis. Equation 15 characterizes the interior-anchor comparator actually evaluated here. The boundary identity and interior decomposition together specify when absolute and relative formulations coincide and when they impose different geometric preferences.

A.1 EXACT COMPUTATION, SAMPLER STATE, AND COST

These statements formalize the implementation and its scope. The chain rule and smoothness bounds are standard tools; their role here is to identify the exact computed objective, the interface required for transfer, and the limits of inference from local control to image-level success.

Proposition 3 (Exact selected-readout gradient)**.** *Fix the latent, conditioning and timestep. Let $H_\ell = F_\ell \circ \cdots \circ F_1$ contain every operation producing the selected native attention readout, and $Q_\ell = q \circ H_\ell$. If these maps are differentiable at the evaluated state, differentiating the prefix gives the same mathematical gradient as attaching q to that intermediate state in a complete model forward pass:*

$$\nabla_z Q_\ell = DH_\ell(z)^\top \nabla q(H_\ell(z)). \tag{19}$$

Proof. The chain rule gives the expression directly. Later blocks and the output head are not ancestors of this loss, and contribute no derivative factor. A readout inside a block includes all operations producing its queries and keys, including normalization and positional encoding, but need not include its attention output. Equality requires the same native prefix and arithmetic path; changed precision, batching or kernels require numerical checks. This proposition does not equate an attention objective with an image metric or an all-layer loss. Each ordinary full prediction is still evaluated at the updated latent.

Proposition 4 (Sampler-state composition and propagation)**.** *Let x be the full sampler state, $z = Px$ its latent component, and $J\delta$ an injected displacement. Define $G_k(x) = x + J\delta_k(Px; c)$ and use $S_k \circ G_k$. If DS_k is L_S-Lipschitz along this displacement, then*

$$\|S_k(x + J\delta) - S_k(x) - DS_k(x)J\delta\|_2 \le \tfrac{1}{2} L_S \|J\delta\|_2^2 . \tag{20}$$

For two trajectories with the same initial state, if each S_k is L_k-Lipschitz on the relevant states and $\|J\delta_k\|_2 \le R_k$, their distance satisfies

$$d_{k+1} \le L_k(d_k + R_k), \qquad d_N \le \sum_{j=0}^{N-1} R_j \prod_{k=j}^{N-1} L_k. \tag{21}$$

Proof. Integrating $[DS_k(x + tJ\delta) - DS_k(x)]J\delta$ over $t \in [0, 1]$ gives the remainder bound. The Lipschitz inequality and triangle inequality give the recurrence, which unrolls by induction from $d_0 = 0$. No unconditional or degraded-condition prediction appears in the intervention definition. For one-step samplers, $P = J = I$; multistep solvers need an explicit history-state policy. These statements give interface compatibility and a displacement bound, not equal improvement across models, a terminal-accuracy guarantee, or universal superiority of early updates.

Operation accounting. For N complete predictions and M selected-readout interventions, suppose comparable per-call costs C_v, C_p and C_b represent the full prediction, prefix forward and prefix backward. With shared fixed cost C_{fixed},

$$T_{\text{ESG}} = C_{\text{fixed}} + NC_v + M(C_p + C_b), \tag{22}$$

$$\frac{T_{\text{ESG}} - T_{\text{native}}}{T_{\text{native}} - C_{\text{fixed}}} = \frac{M}{N}\frac{C_p + C_b}{C_v}. \tag{23}$$

This counts the actual computation graph rather than treating every backward call as a full-model evaluation. At fixed M and prefix depth, inverse-N relative overhead is a prediction only when the call costs remain comparable. CPU offload, memory reuse, host load and kernel dispatch can violate this approximation. We report measured sampling seconds and paired overhead separately; the completed experiments do not establish a step-count scaling law.

Concrete MMDiT execution. Frozen text encoders provide token embeddings and pooled conditioning. The latent is patch-embedded with the pretrained positional encoding; timestep and pooled-text embeddings modulate the transformer blocks. The readout call executes the preceding blocks and the selected block's native query/key projections and normalizations. It captures conditional image queries and both image and text keys, then stops before that block's attention output. A separate differentiable reconstruction computes joint softmax, entity-token sums, head averages, spatial

pooling and centers. Backpropagation follows this readout and the prefix to the input latent, not through later blocks or the VAE.

Our adapters retain a two-member forward batch and select its conditional member, preserving the reported numerical path. The spatial score does not use the reference prediction. After the latent update, a fresh complete forward evaluates both branches required by the original prediction rule; the scheduler receives that combined velocity and the updated latent. In the FlowMatch Euler implementation, the increment is the difference of scheduler sigmas; model timesteps are scaled inputs and are not used directly as integration increments. The decoder also applies the pretrained latent scaling and shift. Algorithm 1 abstracts these fixed decoding operations into Decode.

This diagram describes an MMDiT implementation, not an identical operator in every generator. Porting the intervention requires a differentiable entity-spatial readout with the target model's attention normalization, token order, positional encoding and latent representation. A cross-attention-only model, for example, has a different key denominator. A multistep sampler additionally needs an explicit history-state policy. These interface requirements do not establish equal effectiveness across architectures.

Finite-precision displacement. The implementation differentiates with respect to a floating-point copy of the latent, casts through the model's FP16 forward computation, and reconstructs attention in FP32. Numerical loss scaling is canceled after differentiation. The intended update is $\rho\|z\|_2 g/\|g\|_2$. Casting the updated latent to FP16 changes its realized norm, so a scalar correction targets a one-percent norm tolerance (with a two-percent acceptance check). This correction does not introduce an additional objective gradient. Nonfinite maps or gradients trigger an error rather than a silently substituted score. The real-arithmetic analysis and its error bounds must therefore be distinguished from the numerical implementation.

Relation to monotone spatial losses. For a single relation, consider a spatial loss $L = f(m - q)$, where m is fixed and $f'(m - q) > 0$. At a state with $\nabla q \neq 0$,

$$\frac{-\nabla L}{\|\nabla L\|_2} = \frac{\nabla q}{\|\nabla q\|_2}. \tag{24}$$

Thus a monotone transformation of the same scalar margin is not a distinct normalized one-edge control direction. The spatial component of InfSplign (Rastegar et al., 2025) also uses a transformed centroid separation; its complete controller additionally includes presence and balance terms, multiple readouts, and a different update rule. With multiple relations, summing nonlinear edge losses can change their relative gradient weights. ESG instead uses equal weights and shared entity centers. The objective-only interior-anchor control and the complete InfSplign comparison therefore address different questions.

A.2 Native-readout stability and compositional response

This analysis concerns the frozen ESG readout and equal-weight relation objective. It uses standard softmax calculus, graph linear algebra, and smoothness inequalities, specialized to the actual computation rather than introducing a new optimizer. Its role is to characterize local control and its limits; it does not establish universal superiority over other methods. All statements below concern real arithmetic and fixed conditioning, timestep, model weights, token groups, and pooling weights.

Notation and implementation scope. Flatten each entity's horizontal and vertical coordinates into $\mu(z) \in \mathbb{R}^{2n}$. Let B be the signed relation matrix: every row has $-1, +1$ at the two coordinates on the requested axis, oriented so that $q = B\mu$ is positive for the desired ordering. With m edges, $w_e = 1/m$, $Q = w^\top q$, $g = \nabla_z Q$, and $\delta = Rg/\|g\|_2$ for $g \neq 0$. The radius $R = \rho\|z\|_2$ is fixed during an individual update. The zero-gradient branch applies no update. Coordinates lie in $[a, b]$, with $W = b - a > 0$; the 16×16 pooled grid uses cell centers, so $a = 1/32$, $b = 31/32$, and $W = 15/16$. The coordinate box is a relaxation, not a claim that the generator reaches every point in it.

A.2.1 THE NATIVE CENTROID HAS MASS-INDEPENDENT LOGIT BOUNDS

For a fixed entity token set E, let i index heads and image queries after accounting for their fixed pooling weights $\omega_i > 0$. Set

$$p_{ij} = \mathrm{softmax}(s_i)_j, \quad b_i = \sum_{j \in E} p_{ij}, \quad M = \sum_i \omega_i b_i, \quad \pi_i = \omega_i b_i / M, \quad \mu = \sum_i \pi_i x_i, \tag{25}$$

where $x_i \in [a, b]$ is one spatial coordinate. The softmax denominator includes *all native image and text keys*. Finite logits and a nonempty entity set imply $M > 0$ in real arithmetic.

Proposition 5 (Native-centroid sensitivity and curvature)**.** *Under these fixed-weight assumptions,*

$$\frac{\partial \mu}{\partial s_{ij}} = \frac{\omega_i}{M}(x_i - \mu)p_{ij}(\mathbf{1}_{j \in E} - b_i), \qquad \|\nabla_s \mu\|_1 \le W, \tag{26}$$

$$|\mu(s + v) - \mu(s)| \le W \tanh \|v\|_\infty, \qquad |D^2\mu(s)[v, v]| \le \tfrac{9}{2} W \|v\|_\infty^2. \tag{27}$$

The constants do not contain inverse entity mass or the number of tokens, heads, or queries. The curvature constant is a sufficient bound, not claimed optimal.

Proof. The softmax derivative gives $\partial b_i / \partial s_{ij} = p_{ij}(\mathbf{1}_{j \in E} - b_i)$; the quotient rule gives the displayed Jacobian. Summing its absolute values over keys inside and outside E yields

$$\|\nabla_s \mu\|_1 = 2 \sum_i \pi_i (1 - b_i)|x_i - \mu| \le 2\mathbb{E}_\pi |X - \mathbb{E}_\pi X| \le W. \tag{28}$$

The last inequality follows from the chord bound on $|x - \mu|$ on $[a, b]$: $\mathbb{E}|X - \mu| \le 2(\mu - a)(b - \mu)/W \le W/2$.

For the finite bound, set $\epsilon = \|v\|_\infty$, $b_i' = b_i(s + v)$, and $r_i = b_i'/b_i$. The entity exponential sum and the full partition function each change by a factor in $[e^{-\epsilon}, e^{\epsilon}]$, so $r_i \in [u, v_+] = [e^{-2\epsilon}, e^{2\epsilon}]$. With $t = \mathbb{E}_\pi r$, $\pi_i' = \pi_i r_i / t$. The endpoint chord bound gives

$$\mathrm{TV}(\pi', \pi) = \frac{\mathbb{E}_\pi |r - t|}{2t} \le \frac{(v_+ - t)(t - u)}{t(v_+ - u)} \le \frac{\sqrt{v_+} - \sqrt{u}}{\sqrt{v_+} + \sqrt{u}} = \tanh \epsilon. \tag{29}$$

The middle expression is maximized at $t = \sqrt{u v_+}$, and a bounded-coordinate mean changes by at most W TV. The zero-perturbation case is immediate. This is the finite-discrete likelihood-ratio/total-variation argument underlying the Hilbert-metric bound of Cohen & Fausti (2023), applied to the native centroid.

For the second derivative, follow $s(t) = s + tv$, and write

$$h_i(t) = \log b_i(s(t)) = \mathrm{LSE}_{j \in E}\, s_{ij}(t) - \mathrm{LSE}_j\, s_{ij}(t), \qquad \pi(t) = \mathrm{softmax}(\log \omega + h(t)). \tag{30}$$

This is a proof reparameterization, not a text-only replacement for the native softmax. Log-sum-exp calculus (Gao & Pavel, 2017) gives $h_i' = \mathbb{E}_{p_i^E} v_i - \mathbb{E}_{p_i} v_i \in [-2\epsilon, 2\epsilon]$ and $h_i'' = \mathrm{Var}_{p_i^E} v_i - \mathrm{Var}_{p_i} v_i \in [-\epsilon^2, \epsilon^2]$. Differentiating the expectation twice gives

$$\mu'' = \mathrm{Cov}_\pi(X, h'') + \mathbb{E}_\pi[(X - \mu)(h' - \mathbb{E}_\pi h')^2]. \tag{31}$$

The covariance term is at most $W\epsilon^2/2$ in magnitude by the bounded-range variance inequality. The second is at most $W \mathrm{Var}_\pi(h') \le 4W\epsilon^2$. Adding the bounds proves the claim. □

Finite latent updates. Suppose every logit satisfies $\|\nabla_z s_{ij}\|_2 \le \Lambda$ and $\|\nabla_z^2 s_{ij}\|_{2 \to 2} \le \Gamma$ throughout the relevant convex neighborhood. The chain rule and Proposition 5 give

$$\|\nabla_z^2 \mu\|_{2 \to 2} \le \tfrac{9}{2} W \Lambda^2 + W\Gamma, \qquad L_Q \le 9W\Lambda^2 + 2W\Gamma. \tag{32}$$

The second bound applies to a signed difference of two centroids and to their equal-weight relation average; no extra factor of m is required when the logit bounds hold uniformly. Thus $Q(z + \delta) - Q(z) \ge R\|g\|_2 - L_Q R^2/2$. These expressions expose the latent-to-logit assumptions hidden in an abstract smoothness constant. *We have not certified Λ or Γ on real-model neighborhoods.*

What stability does not imply. The cancellation of inverse mass is a property of native-logit derivatives in real arithmetic. It neither establishes object presence nor compares ESG's terminal performance with a baseline. Mathematically equivalent softmax formulas can differ numerically (Blanchard et al., 2019); underflow in a probability-normalization implementation can invalidate a finite-precision derivative even when the real-arithmetic bound holds.

A.2.2 Graph coefficients distinguish average and all-edge control

Let $c_G = B^\top w$ and $J = D_z \mu$. Then

$$Q = c_G^\top \mu, \qquad g = J^\top c_G, \qquad G = BJ, \qquad K = GG^\top. \tag{33}$$

This is standard incidence-matrix algebra (Jiang et al., 2011). It identifies directions in relation weighting that the linear objective cannot distinguish. For example, in a four-object collinear chain, $Q = (x_4 - x_1)/3$: the middle edge can be reversed even at the box optimum. This counterexample is not a claim about the evaluated four-object cohort, which uses grids.

Relation-gradient response. With $g \neq 0$, define $\beta_e = (Kw)_e / \sqrt{w^\top K w}$. If ∇q_e is L_e-Lipschitz along the actual update segment, Taylor's integral remainder gives

$$q_e(z + \delta) - q_e(z) \geq R\beta_e - \tfrac{1}{2} L_e R^2. \tag{34}$$

Indeed, $\nabla q_e^\top \delta = R(GG^\top w)_e / \left\| G^\top w \right\|_2$. Every edge improves when all $\beta_e > 0$ and $R < 2\beta_e / L_e$ for edges with $L_e > 0$; zero-curvature edges depend only on their linear term. Positive semidefiniteness of K ensures average first-order ascent, not positivity of every component of Kw. For example, $g_1 = (1, 0)$, $g_2 = (-2, 1)$ and equal weights give a negative response on edge 1. ESG remains equally weighted; this analysis introduces no adaptive reweighting.

Proposition 6 (Exact all-edge threshold in the coordinate box)**.** *Write an oriented edge as $q_e = x_v - x_u$, with $x \in [a, b]^{2n}$, and let $0 \leq \tau_e < W$. Define*

$$Q_{\max} = \tfrac{W}{2} \|c_G\|_1, \qquad \kappa_e = \min\{(-c_{G,u})_+, (c_{G,v})_+\}, \qquad D_* = \min_e \kappa_e (W - \tau_e). \tag{35}$$

Then

$$\max_{x : \exists e,\ q_e \leq \tau_e} Q(x) = Q_{\max} - D_*. \tag{36}$$

Consequently, $Q(x) > Q_{\max} - D_*$ guarantees $q_e(x) > \tau_e$ for every readout relation. It is an exact uniform score threshold for the box relaxation, not a necessary condition for any particular scene to be correct.

Proof. Since each row of B sums to zero, $\sum_i c_{G,i} = 0$. Independent coordinate maximization gives $Q_{\max} = a \sum_i c_{G,i} + W \sum_i (c_{G,i})_+ = W \|c_G\|_1 / 2$. For a fixed violating edge, other coordinates can stay at maximizing endpoints. If $c_{G,u} < 0 < c_{G,v}$, write $x_u = a + r$, $x_v = b - t$. The violation requires $r + t \geq W - \tau_e$, and its objective loss is $(-c_{G,u})r + c_{G,v} t$. Because $0 < W - \tau_e \leq W$, assigning the whole required displacement to the cheaper coordinate is feasible and yields loss $\kappa_e (W - \tau_e)$. If the strict sign condition fails, some unconstrained maximizing endpoint assignment already has $x_v - x_u \leq 0 \leq \tau_e$, so the minimum loss is zero. Taking the maximum over the finite union of violating-edge sets proves the result. □

Implications for the evaluated relation graphs. Applying the graph formula to all 100 scenes in each SD3/SD3.5 composition cohort gives Table 6. For three-object chains, the middle coefficient cancels and $D_* = 0$, so the strict threshold is empty. Corners and grids admit nonempty thresholds because both endpoints of each requested axis have opposing coefficients. At the exact three-object-chain box optimum, both margins are nonnegative but one may be zero; unlike the four-object-chain counterexample, a reversed edge at that exact optimum is impossible. Combining the preceding results, the sufficient condition

$$Q(z) + R \|g\|_2 - \tfrac{1}{2} L_Q R^2 > Q_{\max} - D_* \tag{37}$$

certifies all readout margins after one exact-arithmetic update. The image evaluation does not certify the required neighborhood constants; we therefore do not report any generated image as certified by this inequality. Detector correctness, object count, and aesthetics remain separate empirical endpoints.

Table 6: **Certificates for readout coordinates, not generated-image scores.** Counts are independent scenes per model. For a common $0 \le \tau < W$, the threshold is $Q > Q_{\max} - \min_e \kappa_e (W - \tau)$.

Topology	Scenes	$Q_{\max}/W$	$\min_e \kappa_e$
Three-object chain	40	$1/2$	0
Three-object corner	40	1	$1/2$
Four-object grid	20	1	$1/4$

A.2.3 Conditional transfer to a terminal spatial surrogate

Let Φ be the fixed remaining sampling/decoding computation, including any later ESG updates, and let $H_e = F_e \circ \Phi$ for a smooth terminal geometry surrogate F_e. Assume on the closed ball $B(z, r)$ that

$$H_e(x) = \alpha_e q_e(x) + b_e + \xi_e(x), \qquad \alpha_e > 0, \quad |\xi_e(x)| \le \varepsilon_e, \tag{38}$$

where α_e, b_e are constant on the ball and $\nabla \xi_e$ is $L_{\xi,e}$-Lipschitz. Define

$$\eta_e = \inf_{0<h\le r} \left(\frac{\varepsilon_e}{h} + \frac{L_{\xi,e} h}{2} \right). \tag{39}$$

For any unit direction u, the centered difference has magnitude at most ε_e/h, and its discrepancy from $u^\top \nabla \xi_e(z)$ is at most $(2h)^{-1} \int_{-h}^{h} L_{\xi,e} |t| \, dt = L_{\xi,e} h/2$. Taking the supremum over directions and infimum over h proves $\|\nabla \xi_e(z)\|_2 \le \eta_e$. If the unconstrained minimizer lies in the ball, $\eta_e = \sqrt{2\varepsilon_e L_{\xi,e}}$; the infimum also handles zero-error and zero-curvature cases.

For $R \le r$, if ∇H_e is $L_{H,e}$-Lipschitz on the update segment, this gives

$$H_e(z + \delta) - H_e(z) \ge R(\alpha_e \beta_e - \eta_e) - \tfrac{1}{2} L_{H,e} R^2. \tag{40}$$

The proof substitutes $\nabla H_e = \alpha_e \nabla q_e + \nabla \xi_e$ into the directional derivative and bounds the remainder. The three terms distinguish relation-gradient agreement, mismatch between the readout and downstream geometry, and finite-step curvature. None of the required uniform real-model neighborhood bounds is certified by our readout AUROC or finite-perturbation diagnostics. Official discrete success additionally requires stable object identity and detection, which this surrogate result does not supply.

Why predictive accuracy and early timing are insufficient alone. The smooth example $H(z) = z$ and $Q_k(z) = z + (2/k)\sin(kz)$ has arbitrarily small uniform value error, but $Q_k'(\pi/k) = -1$ and $H'(\pi/k) = 1$. Predictive closeness alone cannot guarantee a useful control direction. Likewise, with remaining scalar sampler $\Phi(z) = \lambda z$, $\lambda > 0$, a displacement R before this map changes the terminal coordinate by λR, versus R after it. Early intervention can be better or worse depending on λ. Our timing and layer controls test particular implementations; they are not universal early-optimality proofs.

A.2.4 Arithmetic error and normalization limits

If a computed gradient h satisfies $\|h - g\|_2 \le \epsilon \|g\|_2$ for $g \ne 0$ and $0 \le \epsilon < 1$, then

$$\frac{g^\top h}{\|g\|_2 \|h\|_2} \ge \sqrt{1 - \epsilon^2}. \tag{41}$$

Indeed $g^\top h > 0$, and projection onto $h/\|h\|_2$ gives $\|g - h\|_2^2 \ge \|g\|_2^2 - (g^\top h/\|h\|_2)^2$. For an actual displacement $\widehat{\delta} = Rh/\|h\|_2 + e$ with $\|e\|_2 \le \tau$, smoothness along its segment yields

$$Q(z + \widehat{\delta}) - Q(z) \ge R\|g\|_2 \sqrt{1 - \epsilon^2} - \tau \|g\|_2 - \tfrac{1}{2} L_Q (R + \tau)^2. \tag{42}$$

This standard geometric refinement separates gradient error from vector rounding. Norm matching alone does not bound either error parameter, and neither has been certified across real-model neighborhoods.

For the state-dependent radius, write $T(z) = z + \rho \|z\|_2 u(z)$, $u = g/\|g\|_2$, and $H = \nabla^2 Q$. At $z \neq 0$, $g \neq 0$,

$$DT = I + \rho \left[u \frac{z^\top}{\|z\|_2} + \frac{\|z\|_2}{\|g\|_2} (I - uu^\top) H \right]. \tag{43}$$

A convex neighborhood with $\|z\|_2 \leq B_0$, $\|g\|_2 \geq m_0 > 0$, and $\|H\| \leq L$ therefore gives $\mathrm{Lip}(T) \leq 1 + \rho + \rho B_0 L / m_0$. Bounded readout curvature alone is insufficient: $Q(z) = \frac{1}{2}\|z - (1,0)\|_2^2$ has Hessian I, yet its normalized update approaches opposite nonzero displacements from opposite sides of $(1,0)$ and is zero at that point.

B Implementation and Baseline Protocols

SD3-Medium (Esser et al., 2024) and SD3.5-Large (Stability AI, 2024) use 28 Euler steps at 1024×1024 resolution and FP16 model weights, except for the explicitly labeled Rectified-CFG++ predictor–corrector configurations. Native attention projections are read at block 12 for SD3 and block 24 for SD3.5. Entity maps sum over their T5 spans, average over attention heads, and pool the image grid to 16×16. Cell coordinates are $((u_x + 1/2)/16, (u_y + 1/2)/16)$. The single-update configuration uses zero-based index 2 and radius 0.005. Table 7 lists the selected SD3 controller settings.

Table 7: **Selected SD3 spatial-controller configurations.** Indices are zero-based within 28 Euler steps. ESG strengths are relative latent radii ρ; Self-Guidance strengths multiply the unnormalized location-energy gradient in the velocity prediction and are not directly comparable to ρ.

Configuration	Intervention indices	Strength
ESG (one update)	$\{2\}$	0.005
ESG (low budget)	$\{2, 4, 6, 8\}$	0.005
ESG (medium budget)	$\{2, 4, 5, 7, 8, 10, 11, 13\}$	0.01
Self-Guidance (full)	$\{2, 3, \ldots, 25\}$	5,000
Self-Guidance (low budget)	$\{2\}$	10,000
Self-Guidance (medium budget)	$\{2, 4\}$	10,000

Input interpretation. An input-only parser identifies entity phrases and relations. Two-object templates require a unique, non-overlapping token span for each entity. In controlled scene graphs, repeated mentions of an entity are aggregated into one token group; groups remain disjoint. The generation prompt is unchanged. Unsupported prompts retain native sampling and are included in the score. CompBench has 215 supported directional prompts and 85 proximity or other spatial prompts in the 300-prompt list. Controlled multi-object cases provide the relation graph as part of their input specification; no claim of unrestricted free-text scene-graph parsing is made. The 100 controlled scenes contain 40 three-object chains with two collinear edges, 40 three-object corners with two orthogonal edges sharing one object, and 20 four-object grids with four edges. Each scene uses four seeds on SD3 and two on SD3.5. All requested edges are evaluated; an image contributes its mean edge correctness and its all-edge indicator.

Computation-budget selection. Unscored timing probes determine feasible update counts under the two native-generation ceilings. Each candidate is evaluated on 64 development cases. The selected ESG settings use four early updates at radius 0.005 and eight early updates at radius 0.01. Self-Guidance uses one and two early updates at strength 10,000. All methods are frozen before their formal comparison. Budget selection for ESG and Self-Guidance uses the same 16 pair prompts and 16 multi-object scenes, each with two seeds (64 cases), and six candidates per method and budget: two schedules and three strengths. Their selection score equally weights pair and multi-object relation accuracy. STORM and InfSplign use a separate shared set of 32 two-object prompts with two seeds (64 cases), with three strengths each for their full and constrained schedules. The latter selection uses pairwise relation accuracy. Neither development set shares prompts or object combinations with the formal SD3 cohorts. These are matched development protocols within each controller study, rather than an identical development task mixture across all methods. The selected medium-budget configurations use four STORM updates and six InfSplign updates. Timing is measured on RTX

3090 GPUs. Text conditioning is cached equally, and generation time includes lookup, sampling, and decoding. The budget-specific table uses this total-generation clock for every method. Update counts alone do not measure equal computation: a gradient through one selected transformer prefix differs from gradients through six or all attention blocks. We therefore report intervention schedules and readout depths alongside measured runtime, rather than treating backward passes as identical function evaluations. The main 14-configuration table and full-range frontier uniformly use sampling-only time; these two clocks are not combined on one axis.

Baseline identity. Self-Guidance is the released location-energy method adapted to SD3 attention and flow prediction, with deterministic targets derived from the relation graph. The two-coordinate L1 centroid objective is retained separately as an ablation. Rectified-CFG++ retains its author predictor–corrector rule and is evaluated at 14 and 28 steps with development-selected guidance. Prompt Reinjection uses the MMDiT conditioning intervention; its SD3.5 configuration is separately selected. Full STORM includes iterative refinement; full InfSplign includes its compound spatial, presence, and balance terms. Both are evaluated as disclosed SD3 adaptations of the released methods, using their author implementations of the objective functions.

Self-Guidance adaptation and targets. We use the centroid-location energy released with the authors' SDXL demonstration[1] (Epstein et al., 2023). The adapter averages this energy over all 24 SD3 joint-attention layers, retaining the full image–text attention denominator. Its gradient is added to the guided velocity prediction before the Euler step. To supply targets from the same relation input as ESG, each axis is treated as a directed acyclic ordering graph. An entity's rank is the length of its longest incoming path. For constrained entities, a rank r_e on an axis with maximum rank $r_{\max} > 0$ maps to $0.25 + 0.5 r_e / r_{\max}$; unconstrained coordinates are 0.5. Thus a left/right pair receives axis coordinates 0.25 and 0.75, with both orthogonal coordinates fixed at 0.5. One target is shared by every relation involving the same entity. This specifies the released location energy on a common MMDiT backbone, rather than reproducing the original paper's generator.

The full Self-Guidance configuration selects between effective strengths 1,000 and 5,000 on a separate 32-prompt, two-seed development set, using official position accuracy; a tie selects 5,000. These prompts do not overlap the formal prompts. The selected configuration retains all 24 intervention steps.

STORM and InfSplign adaptations. Both methods read six SD3 joint-attention blocks, with zero-based indices $4, 8, 12, 16, 20, 23$. Entity spans are averaged into object maps, retaining the native image–text attention denominator. STORM pools the readout to its 16×16 map and keeps the released transport, smoothing, clipping, and refinement rules (Han et al., 2025). The 50-step author schedule is mapped to 28 steps: intervention occupies indices 1–13, with refinement at 3, 6, 8, and 11 and a 30-iteration limit per refinement. Object order and axes are canonicalized for left/right and above/below relations. A finite derivative at the center of the source's Gaussian construction preserves its forward values while avoiding an undefined intermediate derivative. The selected strength is 80 in both profiles. Full STORM executes a mean of 126.82 and 102.34 updates on GenEval position and CompBench, respectively; its budget configuration stops after four accepted updates.

InfSplign averages the released compound objective over the six blocks, retaining spatial, presence, and balance terms with equal weights, $\alpha = 0.75$, and margin 0.25 (Rastegar et al., 2025). Its gradient modifies the guided flow prediction. Full InfSplign intervenes at all 28 steps; its budget configuration uses six updates. The selected strength is 4,000 in both profiles. These strengths multiply different energies and are not comparable to ESG's relative displacement radius. In both adapters, numerical loss scaling is canceled in the gradient; retries and executed backward passes are included in measured cost. The two-object prompts do not provide separate attribute-binding annotations. Native forward-trajectory and finite-gradient checks precede the formal comparison.

Layer selection and transfer. SD3 uses the fixed block-12 readout from method development. To select the readout after transfer, SD3.5 calibration scores blocks 6, 12, 18, 24, and 30 using native-readout prediction of final position success. Eligibility requires AUROC at least 0.75 and lower 95% bound at least 0.65; among eligible layers within 0.02 of the best AUROC, the shallowest is

[1] `https://dave.ml/selfguidance/`.

chosen. This selects block 24. The calibration replay processes 200 existing development trajectories through the first two sampling steps and reads five candidate prefixes at index 2. The three recorded RTX 3090 jobs total 20.4 GPU-minutes, including model loading; the replay generates no new endpoint images. This one-time cost excludes the previously generated endpoints and their evaluation and is separate from per-image sampling times. Calibration and evaluation use distinct seeds on the same public prompt set. A separate paired layer comparison holds prompts and seeds fixed. Unseen-pair tests exclude unordered object-pair overlap with the public pool. The Prompt Reinjection development set overlaps 32 SD3.5 unseen-pair prompts; that row is not an untouched prompt holdout.

C Statistics and Additional Evaluation

Table 8: **End-to-end timing at common computation ceilings.** Every row uses the same 400 position and 600 CompBench images. T_G, T_C include conditioning lookup, sampling, and decoding; these measured totals are distinct from the sampling-only columns in the main table. Bold marks the best spatial scores within each budget group.

Method	Position ↑	Joint ↑	Spatial ↑	T_G ↓	T_C ↓
Low budget: $1.5\times$ *generation-time ceiling*					
Self-Guidance	34.25	23.50	34.09	13.36	12.35
ESG	**47.25**	**34.00**	**36.98**	13.35	12.46
Medium budget: $2\times$ *generation-time ceiling*					
Self-Guidance	39.00	27.50	36.60	16.50	14.66
STORM	51.00	**38.75**	37.92	18.03	17.57
InfSplign	47.75	37.00	40.08	16.90	16.78
ESG	**56.50**	38.25	**41.90**	16.66	14.75

Table 9: **Paired ESG–Self-Guidance differences.** Values are percentage points for binary metrics and score points for CompBench ($\times 100$). Simultaneous intervals use the unified 53-comparison Bonferroni family and 1,000,000 cluster-bootstrap samples.

ESG minus Self-Guidance	Gain	95% CI	Simultaneous CI
Spatial / low	2.89	[1.05, 4.85]	[-0.15, 6.31]
Spatial / medium	5.30	[2.97, 7.71]	[1.44, 9.45]
Position / low	13.00	[8.75, 17.42]	[6.00, 20.54]
Position / medium	17.50	[12.25, 22.77]	[8.59, 26.52]
Mean relation / low	5.62	[2.81, 8.50]	[0.94, 10.56]
All relations / low	1.75	[-0.75, 4.25]	[-2.50, 5.75]
Mean relation / medium	9.06	[5.62, 12.44]	[3.25, 14.81]
All relations / medium	6.75	[3.50, 10.25]	[1.50, 12.75]

The simultaneous intervals support six of eight positive ESG–Self-Guidance differences, including position and mean relation accuracy at both budgets. They include zero for low-budget CompBench and low-budget all-relations success. Medium-budget ESG exceeds STORM and InfSplign in both public spatial point estimates; the corresponding simultaneous intervals include zero. In the one-update comparison with full InfSplign, the position interval favors InfSplign.

Table 10: **Paired comparisons with STORM and InfSplign.** Differences are ESG minus the named comparator. Medium-budget ESG is compared with medium-budget baselines; one-update ESG is compared with full controllers. Intervals use the same 53-comparison family as the preceding table.

Comparison	Metric	Gain	95% CI	Simultaneous CI
ESG medium – STORM (medium budget)	Position	5.50	[-0.50, 11.50]	[-4.55, 15.59]
ESG medium – InfSplign (medium budget)	Position	8.75	[3.00, 14.50]	[-1.00, 18.56]
ESG one update – STORM (full)	Position	-12.50	[-20.00, -4.75]	[-25.07, 0.75]
ESG one update – InfSplign (full)	Position	-13.75	[-18.94, -8.66]	[-22.73, -5.30]
ESG medium – STORM (medium budget)	Spatial	3.98	[1.44, 6.56]	[-0.28, 8.38]
ESG medium – InfSplign (medium budget)	Spatial	1.81	[-0.69, 4.30]	[-2.45, 6.02]
ESG one update – STORM (full)	Spatial	4.44	[1.27, 7.64]	[-0.88, 9.86]
ESG one update – InfSplign (full)	Spatial	-3.34	[-5.97, -0.82]	[-7.91, 0.80]

The intervals quantify paired differences for both full-controller and constrained-budget settings, complementing the accuracy–cost comparison in the main text.

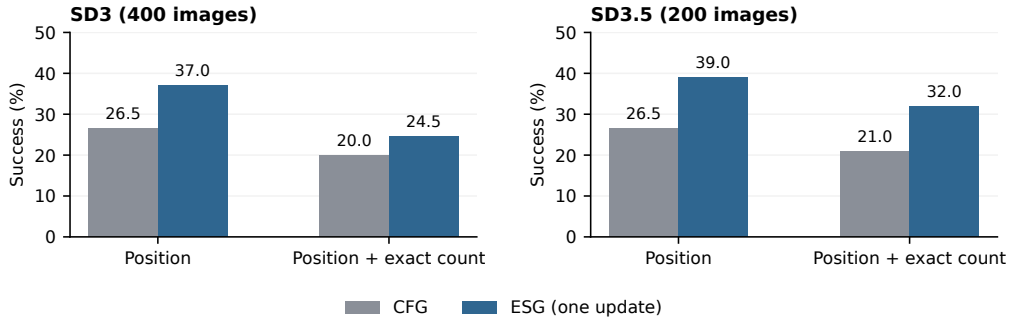


Figure 5: **Spatial intervention transfers across the two evaluated models.** Bars compare matched prompts and initial noise within each model. SD3 uses CFG scale 7 and 400 images; SD3.5 uses scale 3.5 and 200 images. Joint success additionally checks exact object counts.

Official and continuous metrics. GenEval uses its released object detector, matching rules, and position thresholds. Joint success additionally requires exact one-of-each-object counts. Continuous geometry uses detected-box centers normalized to the image and a signed separation along the requested axis; undefined pairs retain explicit coverage counts. CompBench uses the released UniDet spatial scoring function on the full 300-prompt list. Scores in this paper are spatial-subtask results, not the full GenEval or CompBench aggregate.

Vision-language diagnostics. The Qwen3-VL-8B-Instruct evaluator (Bai et al., 2025) receives fixed requested, opposite, and equivalent-relation questions, together with existence and count questions. Relation margins subtract opposite-question affirmative probabilities from requested-question probabilities. These probabilities are diagnostic scores, not calibrated correctness probabilities; multiple question forms are not independent evaluators. Detector and VLM disagreements are retained.

Table 11: **Image-quality diagnostics on all images in the budget cohorts.** C is image–text CLIP cosine similarity $\times 100$ and A is aesthetic-predictor score. Subscripts identify GenEval position (G), CompBench (C), and multi-object scenes (M).

Method	$C_G \uparrow$	$A_G \uparrow$	$C_C \uparrow$	$A_C \uparrow$	$C_M \uparrow$	$A_M \uparrow$
Self-Guidance (low budget)	33.66	5.47	32.77	5.75	34.09	5.16
ESG (low budget)	33.76	5.49	32.78	5.71	34.38	5.10
Self-Guidance (medium budget)	33.77	5.49	32.75	5.76	34.16	5.11
ESG (medium budget)	33.78	5.40	32.73	5.65	34.64	4.94

CLIP (Radford et al., 2021) and aesthetic scores include failures as well as successes. The medium-budget ESG aesthetic means are lower than Self-Guidance by approximately 0.092, 0.110, and 0.163 on the three cohorts. These measurements quantify the accompanying change in image-quality scores. Regional crop similarity is a separate paired diagnostic: mean CLIP cosine is 0.901 on 275 valid object pairs out of 400 possible slots in the 200-image response study.

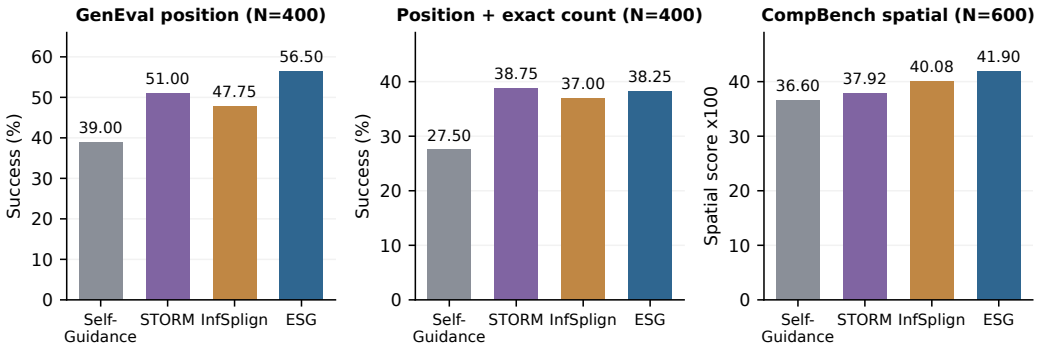


Figure 6: **Four spatial controllers under the same medium computation ceiling.** All methods use identical prompts and initial noise. ESG has the highest position and CompBench point estimates; STORM has the highest joint position/count point estimate. Bars show means, and paired uncertainty is reported in Appendix C.

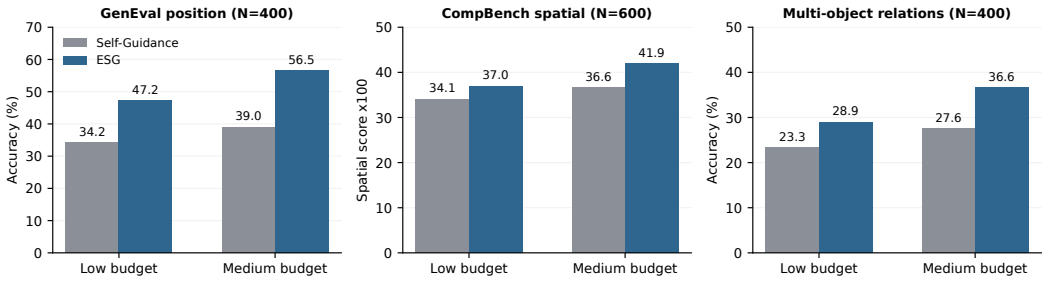


Figure 7: **Scaling the intervention budget.** Low and medium budgets use the same input cohorts. Public spatial metrics and controlled multi-object mean relation accuracy show how the two selected operating points change the outcome. Paired intervals are reported above.

C.1 Results by scene topology

The controlled cohort contains three scene types specified when the prompts were constructed: chains, corners, and grids. Table 12 reports every type at both computation levels. At the medium budget, ESG improves both mean relation accuracy and all-relations success in all three types (Figure 8). This decomposition locates the aggregate benefit across distinct relation structures; it is descriptive and uses the same images as Table 3, rather than an additional independent test.

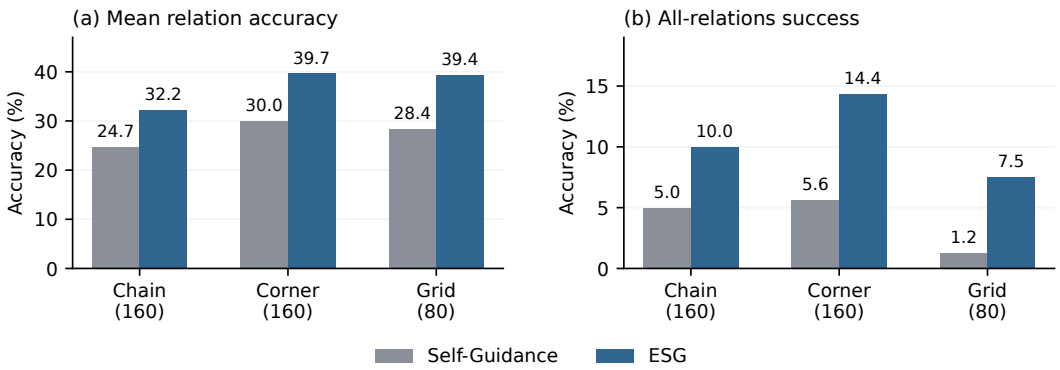


Figure 8: **Medium-budget gains across relation structures.** Numbers below each scene type give images per method. Both panels use the same paired 400-image SD3 cohort, partitioned into its three predefined scene types. Bars are point estimates; the confirmatory paired intervals for aggregate outcomes are reported above.

Table 12: **All scene types at both computation budgets.** N is the number of images per configuration in that type. Scores are percentages. Every row is a subset of the same 400-image paired cohort; no extra samples or parameter selection are introduced.

Scene type	Configuration	N	Mean relation ↑	All relations ↑	Joint ↑
Chain	Self-Guidance / low	160	21.88	3.12	1.88
Chain	ESG / low	160	25.31	4.38	2.50
Chain	Self-Guidance / medium	160	24.69	5.00	3.75
Chain	ESG / medium	160	32.19	10.00	7.50
Corner	Self-Guidance / low	160	25.62	5.00	1.25
Corner	ESG / low	160	33.12	8.12	3.75
Corner	Self-Guidance / medium	160	30.00	5.62	3.75
Corner	ESG / medium	160	39.69	14.37	9.38
Grid	Self-Guidance / low	80	21.56	1.25	1.25
Grid	ESG / low	80	27.81	1.25	0.00
Grid	Self-Guidance / medium	80	28.44	1.25	1.25
Grid	ESG / medium	80	39.38	7.50	2.50

D Qualitative Results

Selected paired examples show visible changes in object arrangement. Methods share the same prompt, seed, and initial latent. Original frames are resized without cropping or retouching; aggregate results are reported in the tables. Figure 4 presents the public-benchmark examples in the main text.

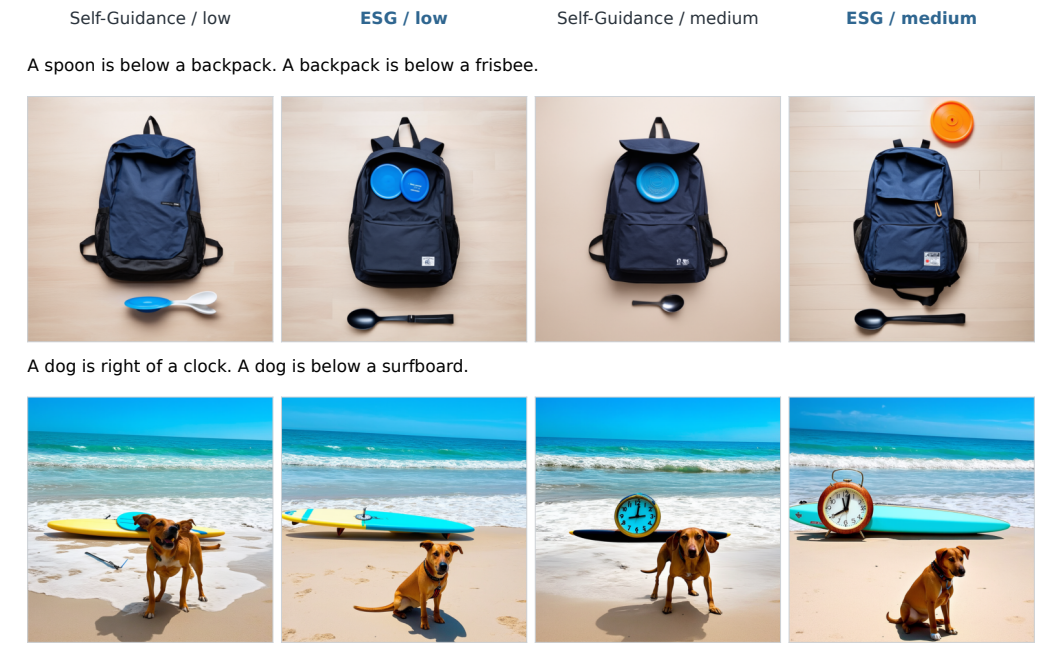


Figure 9: **Relation composition on SD3.** A vertical chain and an orthogonal relation pair are evaluated at two computation levels. Low and medium settings use common generation-time ceilings for Self-Guidance and ESG.

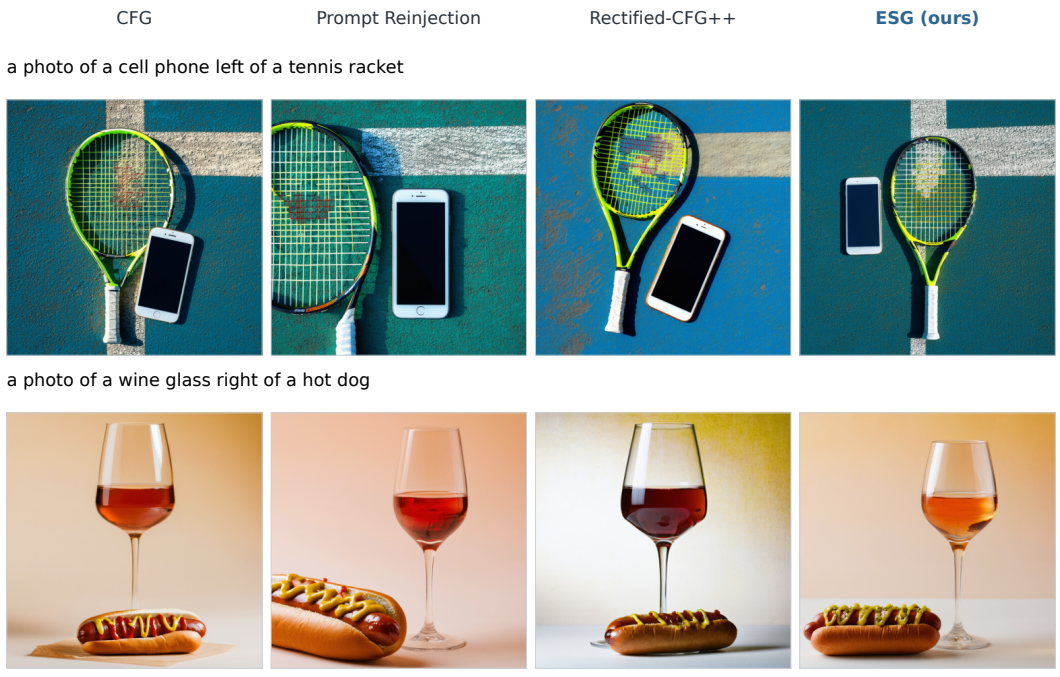


Figure 10: **SD3.5 public spatial prompts.** The phone–racket scene reverses the horizontal arrangement produced by the comparators. The second row compares alternative arrangements of a hot dog and a wine glass. CFG, Prompt Reinjection, and ESG use scale 3.5; Rectified-CFG++ uses 4.5. ESG applies one block-24 update.

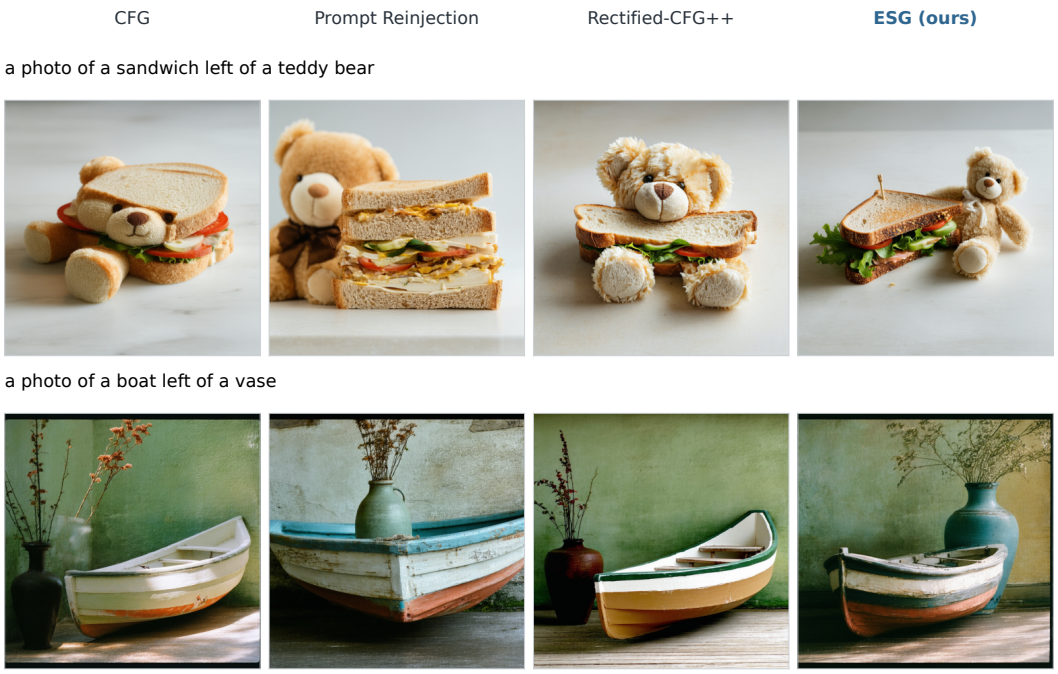


Figure 11: **SD3.5 on unseen object pairs.** ESG separates the sandwich from the teddy bear and places the boat to the left of the vase. CFG, Prompt Reinjection, and ESG use scale 3.5; Rectified-CFG++ uses 4.5. The Prompt Reinjection cohort has the development overlap disclosed in Table 4.

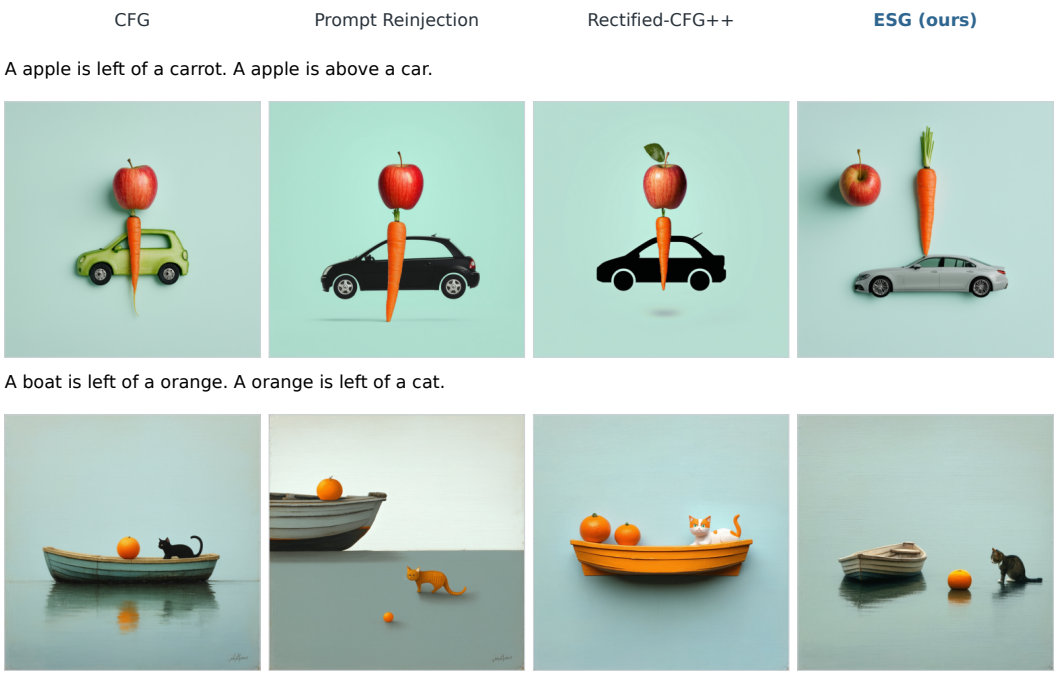


Figure 12: **SD3.5 multi-object compositions.** Two orthogonal relations place the apple left of the carrot and above the car; a horizontal chain places the boat left of the orange and the orange left of the cat. Columns share the prompt and initial noise. CFG, Prompt Reinjection, and ESG use scale 3.5; Rectified-CFG++ uses 4.5.

E MATCHED OBJECTIVE CONTROLS

These experiments compare the relative objective and the two-coordinate interior-anchor L1 objective with the same selected readout, normalized latent rule, radius, and intervention indices. SD3 uses CDG scale 7, degraded block 1, and readout block 12; SD3.5 uses CDG scale 3.5, degraded block 2, and readout block 24. Both use 28 Euler steps and $\rho = 0.005$. Four updates use indices $2, 4, 6, 8$; twelve use $0, \ldots, 11$. The four- and twelve-update configurations use the same prompts, seeds, layers and radii.

Table 13: **Multi-object objective controls (400 SD3/200 SD3.5 images/configuration).** Mean: relation accuracy; All: all-relations success; Joint: all-relations/exact-count success (%). Time: sampling seconds/image. Bold compares ESG and Centroid L1 at matched update counts.

Model	Objective	Updates	Mean ↑	All ↑	Joint ↑	Time ↓
SD3	CDG	0	16.69	0.50	0.50	9.68
SD3	Centroid L1	1	21.19	1.50	**1.00**	10.52
SD3	ESG	1	**22.94**	**1.75**	**1.00**	10.54
SD3	Centroid L1	4	24.88	3.00	2.25	13.08
SD3	ESG	4	**30.63**	**5.25**	**2.50**	13.04
SD3	Centroid L1	12	38.81	11.75	6.50	19.81
SD3	ESG	12	**48.00**	**19.50**	**8.00**	19.75
SD3.5	CDG	0	20.62	3.50	3.00	31.51
SD3.5	Centroid L1	4	33.88	8.00	4.50	77.76
SD3.5	ESG	4	**38.00**	**10.50**	**5.50**	79.18
SD3.5	Centroid L1	12	45.62	21.50	10.00	171.40
SD3.5	ESG	12	**58.88**	**35.50**	**16.00**	170.90

The eight-comparison family combines two backbones, two update counts, and two metrics (mean relation and all-relations success). We retain the study's paired object-set bootstrap with 100 clusters, 10,000 draws, and Bonferroni-adjusted intervals. This mechanism family is separate from the 53-comparison complete-method family. The objectives share image identities, prompts, initial latents and settings; measured times come from separate execution batches. We compare the objectives at identical update schedules and report measured latency, without assuming identical wall time.

Table 14: **Paired relative-minus-interior-anchor gains.** All eight planned matched-objective comparisons are shown, in percentage points. The simultaneous intervals correct over this eight-comparison family.

Model / updates	Metric	Gain	95% CI	Simultaneous CI
SD3 / 4	Mean relation	5.75	[3.12, 8.31]	[1.95, 9.36]
SD3 / 4	All relations	2.25	[0.25, 4.50]	[-0.50, 5.25]
SD3 / 12	Mean relation	9.19	[5.81, 12.75]	[4.44, 14.17]
SD3 / 12	All relations	7.75	[3.00, 12.75]	[1.50, 14.50]
SD3.5 / 4	Mean relation	4.12	[0.12, 8.00]	[-1.50, 9.62]
SD3.5 / 4	All relations	2.50	[-1.50, 6.50]	[-2.50, 8.50]
SD3.5 / 12	Mean relation	13.25	[7.75, 18.75]	[5.66, 20.97]
SD3.5 / 12	All relations	14.00	[6.00, 22.00]	[3.00, 25.50]

What the control isolates. For multiple objects, the L1 control averages the pairwise interior-anchor losses over the same relation edges as ESG. Each edge assigns its own targets at $1/4$ and $3/4$ on the relation axis and $1/2$ on the orthogonal axis; a shared entity participates in every incident edge loss. This differs from the single graph-ranked target per entity used by full Self-Guidance. The L1 comparator and ESG share the readout and normalized update, isolating the choice between these pairwise objectives. Full Self-Guidance additionally uses its released centroid conventions, all-layer energy aggregation, and velocity-space correction. Its complete-method comparison answers a different question and remains in Table 1. The one-update pair results show that much of the initial

benefit is shared by early normalized centroid control. Repeated and multi-object results quantify the additional benefit of relative constraints.

Presence, arrangement, and count diagnostics. Table 15 partitions the same 400 SD3 position images using the official detector and evaluator. ESG reduces both missing joint detections and incorrect arrangements when both objects are detected. Position/count joint success rises from 20.00% for CFG to 38.25% for ESG. Remaining errors include geometric mistakes and incorrect counts after a correct arrangement. These categories describe evaluator outcomes; they do not identify whether an undetected object is visually absent or establish appearance preservation.

Table 15: **Detector-based outcome decomposition (400 images/configuration).** Columns are mutually exclusive and sum to 100%. Missing means at least one requested object lacks a position-threshold detection; wrong relation is evaluated only when both are detected; count failure follows correct position. Self-Guidance and ESG use medium-budget configurations.

Method	Missing	Wrong relation	Count failure	Joint success
CFG	19.25	54.25	6.50	20.00
Self-Guidance	18.75	42.25	11.50	27.50
ESG	12.00	31.50	18.25	38.25

F Timing, Radius, and Readout Controls

The following control uses 200 paired SD3 cases with CDG scale 7 and the block-12 readout. The single early intervention is at index 2; the later control uses index 7 with the same radius. Four full-radius interventions use $\rho = .005$ per update, whereas the smaller-radius control uses .00125 per update. These settings are distinct from the CFG main comparison and use their own paired cases.

Table 16: **Timing and cost ablations on SD3.** All methods use 28 sampling steps and 200 paired cases. Position, exact count, and joint success are percentages. Bold marks the best displayed accuracy point estimate; time and update count describe cost.

Method	Position ↑	Count ↑	Joint ↑	Time (s) ↓	Updates ↓
CDG	37.00	60.50	25.50	9.62	0
Later update (index 7)	36.50	62.00	29.00	10.44	1
Text-normalized readout	46.00	60.00	33.00	10.45	1
ESG (one early update)	47.50	62.00	33.50	10.46	1
Four smaller updates	43.50	61.50	32.00	12.93	4
Four full-radius updates	**52.00**	**63.50**	**36.50**	12.92	4

SD3.5 readout-layer control. On 200 paired public cases (100 prompts, two seeds), blocks 12 and 24 yield 28.0% and 41.0% Position, respectively. Both use 28 Euler steps, CDG scale 3.5 with degraded block 2, and one update at index 2 with $\rho = .005$; prompts and initial noise are identical.

Table 17: SD3.5 public-cohort computation settings (200 paired images per configuration). Position and Joint are percentages; time is sampling seconds per image.

Method	Position ↑	Joint ↑	Sampling (s) ↓
CFG	26.50	21.00	31.72
Rectified-CFG++ (14 steps)	22.50	15.50	32.87
Rectified-CFG++ (28 steps)	24.00	19.00	65.68
ESG	39.00	32.00	44.47

Table 18: Image-quality diagnostics for the 13 SD3 configurations with complete quality evaluation. C is CLIP image–text cosine similarity $\times 100$; A is aesthetic score. G/C denote the 400-image GenEval and 600-image CompBench cohorts. The 14-step Rectified-CFG++ control is not included in this quality panel.

Method	$C_G \uparrow$	$A_G \uparrow$	$C_C \uparrow$	$A_C \uparrow$
Standard guidance and one-update ESG				
CFG	33.66	5.48	32.72	5.76
Prompt Reinjection	34.11	5.25	32.97	5.48
Rectified-CFG++ (28 steps)	33.46	5.46	32.57	5.75
ESG (one update)	33.60	5.51	32.73	5.74
Full spatial-controller schedules				
Self-Guidance (full)	33.63	5.32	32.83	5.63
STORM (full)	31.08	4.61	31.55	4.98
InfSplign (full)	33.69	5.18	32.74	5.51
Low budget: $1.5\times$ *generation-time ceiling*				
Self-Guidance (low budget)	33.66	5.47	32.77	5.75
ESG (low budget)	33.76	5.49	32.78	5.71
Medium budget: $2\times$ *generation-time ceiling*				
Self-Guidance (medium budget)	33.77	5.49	32.75	5.76
STORM (medium budget)	33.56	5.39	32.69	5.65
InfSplign (medium budget)	33.76	5.47	32.82	5.71
ESG (medium budget)	33.78	5.40	32.73	5.65

G MAIN-CONFIGURATION OBJECTIVE CONTROL

This objective-only comparison holds the main SD3 CFG scale 7, eight-update schedule, radius 0.01, layer 12, prompts and initial noise fixed. Both objectives are evaluated on the same 1,400 paired cases: 400 position, 600 CompBench and 400 multi-object cases. Only the objective changes; this is not a full Self-Guidance implementation comparison. The study uses its own four prespecified primary contrasts and paired object-set cluster bootstrap. Repeated seeds remain within clusters. Its intervals are separate from the 53-comparison complete-method family.

Table 19: **Objective-only control in the main configuration.** Position, mean-relation and all-relations scores are percentages; CompBench is a continuous score $\times 100$. Differences are relative ESG minus absolute objective. Intervals are simultaneous over the four primary contrasts.

Metric	Pairs	Absolute	ESG	Difference	Simultaneous interval
Position	400	50.75	56.50	+5.75	[+0.25, +11.25]
CompBench spatial	600	40.43	41.90	+1.47	[-1.03, +4.04]
Mean relation	400	27.62	36.62	+9.00	[+4.94, +13.03]
All relations	400	5.75	11.25	+5.50	[+1.25, +10.00]

Three simultaneous intervals exclude zero; the CompBench interval does not. Thus the main-configuration control supports the relative objective's benefit for position and compositional correctness without implying improvement on every endpoint. Tables preserve all metrics and denominators, including quality trade-offs. Rates/probabilities are $\times 100$; margins, count MAE, seconds, CLIP, aesthetics, budget ratios and call counts retain native units. Brackets give valid-image counts; – marks unavailable aggregate bookkeeping. Timing comes from the original batches, not an isolated comparison.

Table 20: Main-configuration objective control: all position diagnostics (400 images per arm).

Metric	Absolute	Relative ESG
Joint count/position	34.00 [400]	38.25 [400]
Position	50.75 [400]	56.50 [400]
Exactly one of each	52.50 [400]	56.00 [400]
Both detected	83.75 [400]	88.00 [400]
VLM: one of each	74.94 [399]	73.18 [399]
Detector/VLM single consensus	46.75 [400]	49.50 [400]
VLM requested relation	75.25 [400]	82.25 [400]
VLM consistent requested relation	60.25 [400]	68.25 [400]
Detector count MAE	0.284 [400]	0.264 [400]
VLM count MAE	0.165 [399]	0.168 [399]
Count evaluator disagreement	47.37 [399]	44.61 [399]
Unresolved entity consensus	45.75 [400]	43.00 [400]
Box relation margin	-0.055 [335]	0.006 [352]
VLM relation margin	0.630 [400]	0.725 [400]
VLM equivalent-relation margin	0.614 [400]	0.701 [400]
VLM single-count probability	85.83 [400]	85.79 [400]
VLM existence probability	95.81 [400]	97.29 [400]
Detector count MAE a	0.223 [400]	0.210 [400]
VLM count MAE a	0.102 [400]	0.115 [400]
Detector count MAE b	0.345 [400]	0.318 [400]
VLM count MAE b	0.228 [399]	0.221 [399]
Sampling time (s)	16.218 [400]	16.054 [400]
CLIP similarity	33.729 [400]	33.783 [400]
Aesthetic score	5.467 [400]	5.400 [400]
Generation time (s)	16.784 [400]	16.658 [400]
Generation / native time	–	1.666 [400]
Intervention calls	–	8.000 [400]

Table 21: Main-configuration objective control: all CompBench diagnostics (600 images per arm).

Metric	Absolute	Relative ESG
CompBench spatial	40.43 [600]	41.90 [600]
Sampling time (s)	14.367 [600]	14.148 [600]
CLIP similarity	32.879 [600]	32.734 [600]
Aesthetic score	5.696 [600]	5.650 [600]
Generation time (s)	14.938 [600]	14.754 [600]
Generation / native time	–	1.480 [600]
Intervention calls	–	5.733 [600]

Table 22: Main-configuration objective control: all multi-object diagnostics (400 images per arm).

Metric	Absolute	Relative ESG
Mean relation	27.62 [400]	36.62 [400]
All relations	5.75 [400]	11.25 [400]
All object counts exact	17.75 [400]	21.25 [400]
Joint relation/count	3.00 [400]	7.25 [400]
All objects detected	43.00 [400]	52.25 [400]
Mean VLM relation margin	0.324 [400]	0.459 [400]
All VLM relations consistent	14.75 [400]	23.25 [400]
Minimum box margin	-0.537 [171]	-0.409 [207]
Sampling time (s)	16.198 [400]	16.229 [400]
CLIP similarity	34.195 [400]	34.645 [400]
Aesthetic score	4.987 [400]	4.943 [400]
Generation time (s)	16.762 [400]	16.853 [400]
Generation / native time	–	1.685 [400]
Intervention calls	–	8.000 [400]

H Additional Paired Illustrations

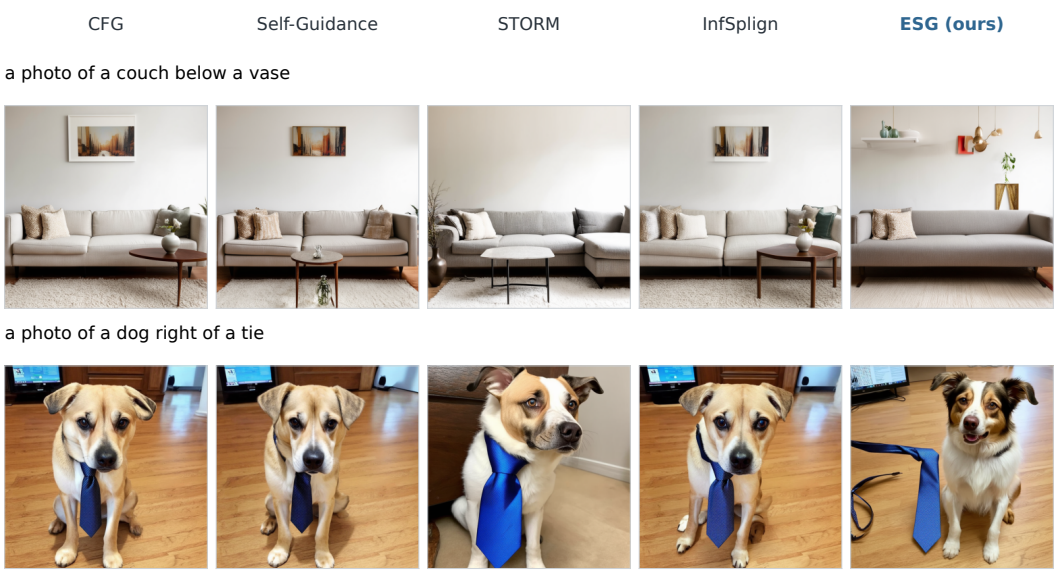


Figure 13: **Paired SD3 comparisons with spatial-guidance methods.** Columns share prompts and initial noise. Self-Guidance, STORM, InfSplign, and ESG use medium-budget configurations; CFG is the native baseline. The first row places the vase above the couch. The second places the dog right of the tie. Images are selected for visual clarity; full-cohort results are in Table 1.

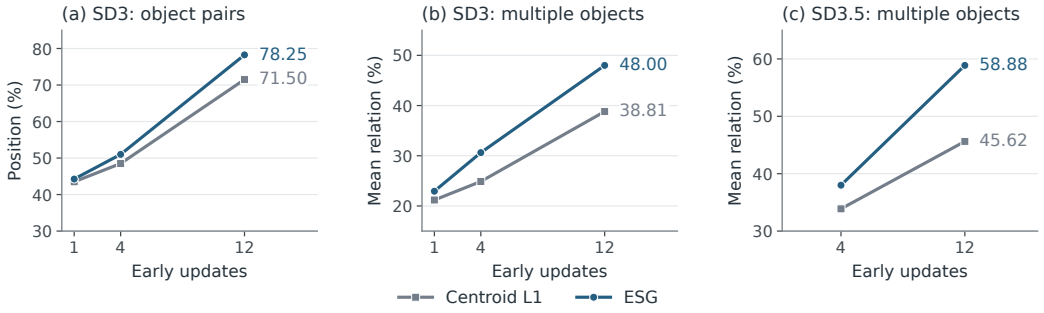


Figure 14: **Relative constraints under repeated guidance.** ESG and Centroid L1 share readouts, radii, schedules, prompts and initial noise. (a) SD3 position: 400 images/configuration. (b,c) Multi-object scenes: 400 SD3 and 200 SD3.5 images/configuration. All panels use CDG; the 1/4/12-update schedules are specified in the text. Lines connect evaluated settings. Multi-object scores and intervals appear in Appendix E.

I TIME-MATCHED NATIVE SAMPLING

Each model uses 100 public position prompts with two seeds, for 200 images per arm. The native 28-step and ESG images are exactly reused from their corresponding cohorts; the native extra-step images share their initial noise. SD3 uses the first two seeds of the four-seed main cohort. These are paired subsets, not independent replications. The ESG configuration is fixed before native-step selection: eight block-12 updates at $\rho = .01$ on SD3, and one block-24 update at $\rho = .005$ on SD3.5.

Timing protocol. Eight disjoint development prompts and two repeats in reversed order select integer native step counts by measured generation time alone. The selected counts are 48 and 40, with native/ESG timing ratios 1.0045 and 1.0178. The tolerance is symmetric five percent. Conditioning is cached equally; generation time includes cache lookup, sampling and VAE decoding, but excludes loading, initial text encoding and evaluation. Timing is measured separately from concurrent formal generation. This isolates an operating-point comparison, not a universal runtime scaling law.

Table 23: ESG minus time-matched native sampling. Intervals use a Bonferroni divisor of six and 100,000 paired object-set bootstrap draws.

Model	Metric	Difference (pp)	Simultaneous interval (pp)
SD3	Joint	+16.00	[+7.50, +24.75]
SD3	Position	+29.50	[+19.70, +39.71]
SD3.5	Joint	+11.50	[+5.39, +18.18]
SD3.5	Position	+14.00	[+7.43, +21.29]

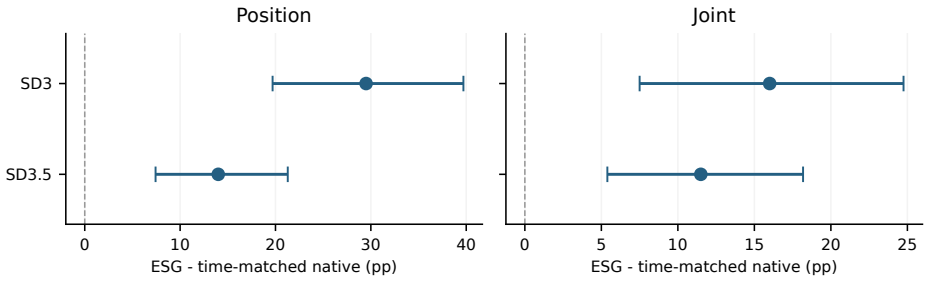


Figure 15: Paired gains over time-matched native sampling. Error bars are the simultaneous intervals reported in the preceding table, not standard errors.

Complete metric definitions. Position, exact-count, joint, presence, consensus, VLM decision and VLM probability scores are displayed $\times 100$. CLIP retains its evaluator's 100-scaled similarity; aesthetics, count mean absolute error (MAE), and signed margins retain native units. Brackets give valid denominators out of 200. Missing outputs are not imputed. VLM requested-relation scores use fixed questions; consistent relation additionally requires agreement across requested/opposite formulations. Count consensus requires agreement between the detector and VLM. These are supplementary diagnostics, not independent ground-truth evaluators or replacements for official Position.

Table 24: SD3 native-step control: all 23 shared evaluation metrics (200 images per arm).

Metric	Native 28	Native matched	ESG
Aesthetic score	5.499 [200]	5.453 [200]	5.440 [200]
CLIP similarity	33.786 [200]	33.816 [200]	33.726 [200]
Detector/VLM single-object consensus	47.50 [200]	53.00 [200]	47.00 [200]
Unresolved entity consensus	44.50 [200]	37.00 [200]	46.50 [200]
Count-evaluator disagreement	43.50 [200]	36.50 [200]	47.74 [199]
Detector count MAE	0.273 [200]	0.240 [200]	0.275 [200]
Detector count MAE (object A)	0.220 [200]	0.205 [200]	0.205 [200]
Detector count MAE (object B)	0.325 [200]	0.275 [200]	0.345 [200]
Exactly one of each object	53.00 [200]	58.50 [200]	55.00 [200]
Joint count/position	21.00 [200]	22.50 [200]	38.50 [200]
VLM single-count probability	86.55 [200]	87.50 [200]	85.26 [200]
VLM existence probability	91.64 [200]	92.20 [200]	96.51 [200]
Position	26.00 [200]	26.50 [200]	56.00 [200]
Both objects detected	81.00 [200]	79.00 [200]	87.50 [200]
Box relation margin	-0.384 [161]	-0.346 [157]	0.001 [175]
VLM: one of each object	75.00 [200]	77.00 [200]	72.36 [199]
VLM consistent relation	28.50 [200]	31.50 [200]	67.00 [200]
VLM count MAE	0.142 [200]	0.133 [200]	0.181 [199]
VLM count MAE (object A)	0.095 [200]	0.080 [200]	0.105 [200]
VLM count MAE (object B)	0.190 [200]	0.185 [200]	0.256 [199]
VLM equivalent-relation margin	0.134 [200]	0.205 [200]	0.693 [200]
VLM relation margin	0.154 [200]	0.211 [200]	0.719 [200]
VLM requested relation	44.00 [200]	47.50 [200]	81.50 [200]

Table 25: SD3.5 native-step control: all 23 shared evaluation metrics (200 images per arm).

Metric	Native 28	Native matched	ESG
Aesthetic score	6.072 [200]	6.064 [200]	6.079 [200]
CLIP similarity	34.112 [200]	34.114 [200]	33.878 [200]
Detector/VLM single-object consensus	58.00 [200]	57.50 [200]	62.00 [200]
Unresolved entity consensus	32.50 [200]	33.50 [200]	30.50 [200]
Count-evaluator disagreement	31.82 [198]	33.50 [200]	32.50 [200]
Detector count MAE	0.230 [200]	0.270 [200]	0.228 [200]
Detector count MAE (object A)	0.165 [200]	0.195 [200]	0.180 [200]
Detector count MAE (object B)	0.295 [200]	0.345 [200]	0.275 [200]
Exactly one of each object	61.50 [200]	61.00 [200]	64.50 [200]
Joint count/position	21.00 [200]	20.50 [200]	32.00 [200]
VLM single-count probability	89.52 [200]	88.71 [200]	89.80 [200]
VLM existence probability	93.31 [200]	93.32 [200]	93.85 [200]
Position	26.50 [200]	25.00 [200]	39.00 [200]
Both objects detected	80.50 [200]	82.00 [200]	83.50 [200]
Box relation margin	-0.409 [161]	-0.412 [164]	-0.171 [166]
VLM: one of each object	81.82 [198]	79.50 [200]	80.00 [200]
VLM consistent relation	29.50 [200]	28.50 [200]	51.00 [200]
VLM count MAE	0.116 [198]	0.135 [200]	0.117 [200]
VLM count MAE (object A)	0.055 [200]	0.100 [200]	0.065 [200]
VLM count MAE (object B)	0.177 [198]	0.170 [200]	0.170 [200]
VLM equivalent-relation margin	0.060 [200]	0.071 [200]	0.446 [200]
VLM relation margin	0.050 [200]	0.083 [200]	0.444 [200]
VLM requested relation	41.50 [200]	41.00 [200]	62.50 [200]