BINDING CRYSTALLIZATION: DIAGNOSING AND GUIDING ATTRIBUTE ASSIGNMENT IN TEXT-TO-IMAGE DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models can generate high-quality images from natural language descriptions, yet often fail to assign attributes to the intended objects. Existing training-free methods mitigate these errors by guiding attention during sampling, but the dynamics governing when and how attribute assignments can be corrected remain insufficiently understood. Through counterfactual prompt switches and step-local interventions, we identify *binding crystallization*: across four prompt categories and three diffusion backbones, attribute assignments become resistant to reversal before spatial refinement is complete. Late denoisers still distinguish opposite assignments, but the sampler writes less of that difference into the latent, and comparable latent writes exert weaker downstream effects on ownership. Intermediate-trajectory probes further show that object presence and attribute ownership are distinct control problems. Together, these findings prescribe *when*, *where*, and *what* to guide. **CrystalBind** is the resulting training-free framework: it sets its intervention window from observed binding transitions, refreshes spatial targets from evolving noun attention maps, and jointly controls presence and ownership through noun excitation, target attraction, and competitor suppression. On T2I-CompBench with SD v1.5, CrystalBind improves the average binding score by 21.3 points over the unguided model and by 3.4 points over CONFORM under a matched budget of 26 guidance updates, while reducing wall-clock time by 28% relative to CONFORM. The same configuration improves SD v2.1 by 14.3 points without retuning. These results show how binding dynamics inform effective and efficient compositional guidance.

1 INTRODUCTION

Text-to-image diffusion models have made high-quality image synthesis accessible through natural language, yet reliably composing requested concepts remains challenging (Rombach et al., 2022; Podell et al., 2024). Their usefulness depends on both visual fidelity and faithful interpretation of concept relations. For example, "a purple car and an orange house" requires both objects and the correct color assignment; exchanging the colors yields a plausible image that violates the prompt. Attribute-binding errors remain common on compositional benchmarks (Huang et al., 2023; 2025).

Recent training-free approaches improve composition through conditional prediction composition or attention-based guidance (Liu et al., 2022; Chefer et al., 2023; Feng et al., 2023; Rassin et al., 2024; Meral et al., 2024; Li et al., 2023b). Final-image gains do not reveal how ownership develops along the trajectory: when an assignment ceases to be correctable, whether emerging noun regions provide usable spatial targets, or whether stronger object activation corrects ownership or only preserves presence. These questions determine guidance timing, spatial targets, and objectives.

We study binding as a state-dependent decision along the reverse trajectory: cross-attention changes the noise prediction, the sampler writes it into the next latent, and that latent shapes subsequent attention. This feedback can make weak early evidence self-reinforcing before visual refinement completes, leaving late denoisers able to distinguish but not realize an alternative assignment.

Counterfactual prompt switches and step-local interventions test this account. Across color, shape, texture, and complex prompts on SD v1.5, SD v2.1, and PixArt-α, assignments resist reversal early,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a phenomenon we call *binding crystallization*. On SD v1.5, coarse noun regions emerge before commitment and refine afterward. Late denoisers still distinguish opposite assignments, but the sampler writes less of their difference into the latent, and comparable realized displacements exert weaker downstream effects on ownership. Trajectory probes and component interventions further separate presence from correct binding: noun activation primarily tracks presence, whereas relative attribute–noun association tracks ownership. Guidance should therefore act while assignments remain responsive, follow evolving regions, and control both presence and ownership.

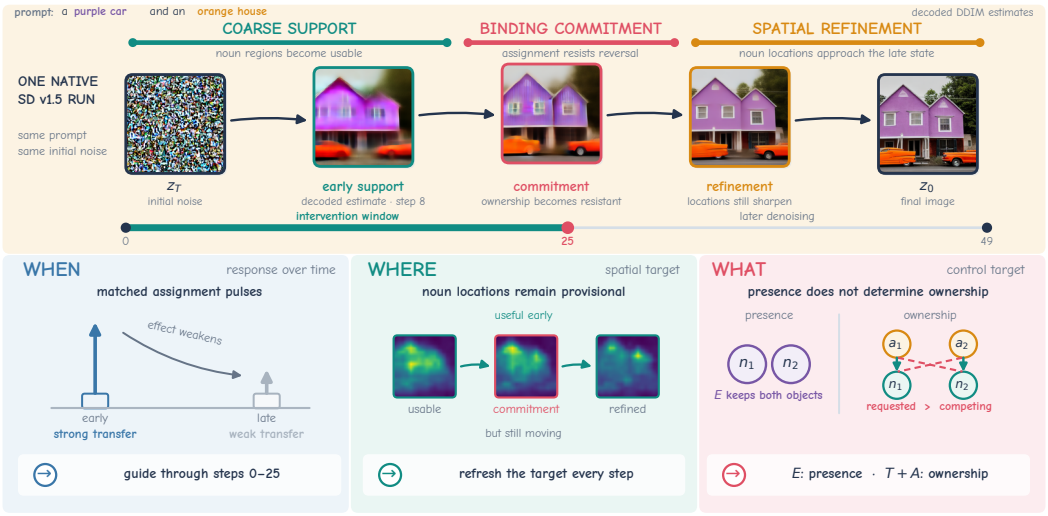


Figure 1: **From denoising dynamics to CrystalBind.** Coarse noun-location support emerges before ownership becomes hard to reverse, yet noun maps continue to move afterward. This measured order determines the intervention window (*when*), live noun targets (*where*), and the separation of presence from assignment (*what*).

CrystalBind implements these requirements through three dynamics-guided choices. The upper tail of prompt-level transition times sets a shared intervention window that covers later-changing prompts without extending into the less responsive late stage. Noun attention maps are refreshed as spatial targets at every guided step, following usable but evolving regions. Noun excitation preserves object presence, while target attraction and competitor suppression direct attribute ownership.

We evaluate CrystalBind on the color, shape, and texture splits of T2I-CompBench over three random seeds. On SD v1.5, it improves average binding by 21.3 points over unguided generation and outperforms CONFORM by 3.4 points under a matched 26-update budget while running 28% faster. The same configuration improves SD v2.1 by 14.3 points without retuning. Timing, target-refresh, and component controls validate the corresponding design choices.

Our contributions are threefold:

1. We identify binding crystallization across three diffusion backbones and characterize how commitment overlaps noun-region emergence and refinement on SD v1.5.
2. We show that late assignment sensitivity does not ensure reassignment, trace this gap through read–write–propagate dynamics, and distinguish object presence from attribute ownership.
3. We introduce CrystalBind, translating these findings into a calibrated intervention window, evolving spatial targets, and a joint presence-and-ownership objective that improves binding at lower runtime under a matched update budget and transfers from SD v1.5 to SD v2.1.

2 Related Work

Training-free compositional guidance. Training-free methods improve composition through conditional prediction composition or attention-based guidance. Composable Diffusion combines denoising predictions associated with different conditions (Liu et al., 2022). Attention-based methods instead act on spatial associations within the evolving generation (Feng et al., 2023; Rassin et al., 2024; Epstein et al., 2023; Agarwal et al., 2023; Wang et al., 2024). Among the closest

attention-guidance methods, A&E increases subject activation (Chefer et al., 2023), D&B adds attribute–object alignment (Li et al., 2023b), and CONFORM contrasts paired and competing attention maps (Meral et al., 2024). CrystalBind differs by grounding its intervention horizon in measured reassignment dynamics, using evolving noun maps as live noun targets, and explicitly separating object presence from attribute ownership. We compare these three closest baselines under both a matched early-update budget and their released schedules.

Layout-guided generation. External-layout methods supply boxes, regions, or planned anchors (Xie et al., 2023; Lian et al., 2024; Yang et al., 2024; Li et al., 2023a; Bar-Tal et al., 2023; Shirakawa & Uchida, 2024). CrystalBind instead uses evolving noun maps as internal targets, allowing guidance before those maps reach their late-denoising state.

Denoising dynamics and phase transitions. Prior work describes coarse-to-fine denoising (Choi et al., 2022; Karras et al., 2022), phase-like generative dynamics (Biroli et al., 2024; Sclocchi et al., 2025; Ambrogioni, 2023; Raya & Ambrogioni, 2023), and critical windows for the emergence of individual features (Li & Chen, 2024); related studies examine inference-time compute and cross-attention heads (Ma et al., 2025; Park et al., 2025). In contrast, we study a relational state: when the ownership of an attribute by one object rather than another loses reversibility. Our counterfactual interventions further distinguish whether the denoiser still reads the competing assignment, whether that difference is written into the latent, and whether the write propagates into final ownership.

3 Rethinking Compositional Denoising

3.1 From repeated conditioning to assignment commitment

Let $u = 0, \dots, T$ count reverse steps from noise. At step u, cross-attention contributes to $\epsilon_\theta(z^{(u)}, c)$, the sampler writes this prediction into $z^{(u+1)}$, and the new latent changes the queries used at the next step. Conditioning is thus evaluated repeatedly on a state that already contains its previous effects.

Consider two local modes $y \in \{+, -\}$ corresponding to the requested and swapped assignments. Their posterior odds separate into text preference and evidence already stored in the latent:

$$\log \frac{\rho_+(z, c)}{\rho_-(z, c)} = \log \frac{\pi_+(c)}{\pi_-(c)} + \log \frac{p_u(z \mid +)}{p_u(z \mid -)}. \tag{1}$$

At high noise, the two state densities overlap and the second term is weak. As signal-to-noise ratio rises, the latent becomes more diagnostic of its current branch; in the shared-covariance Gaussian case with latent means $\bar{z}_+$ and $\bar{z}_-$, the expected separation scales as $\bar{\alpha}_u \|\bar{z}_+ - \bar{z}_-\|^2/\sigma_u^2$ (Appendix B). Thus, text-favored assignments gain increasing support from the trajectory.

The update-level mechanism is explicit for deterministic DDIM (Song et al., 2022). For prompts c^+ and c^- with identical words but opposite assignments, their one-step difference at a shared latent is

$$B_u = \kappa_u g \left[\epsilon_\theta(z^{(u)}, c^+) - \epsilon_\theta(z^{(u)}, c^-)\right], \tag{2}$$

where g is the classifier-free guidance scale and κ_u is fixed by the sampler schedule. After both branches resume the same text, the persistent change in an ownership readout r is approximately

$$\delta r_v \approx \nabla r(z^{(v)})^\top \Phi_{v \leftarrow u+1} B_u. \tag{3}$$

Here $\Phi_{v \leftarrow u+1}$ is the product of the Jacobians for updates under the shared conditioning from $u+1$ to v. Equation 3 separates three requirements: the denoiser must *read* the assignment difference, the sampler must *write* it into the latent, and subsequent dynamics must *propagate* that write along an ownership-relevant direction. Late text sensitivity can outlast ownership control.

3.2 Finite-horizon assignment control

A declining prompt-switch response does not identify why reassignment fails. The denoiser may stop distinguishing the two assignments, the sampler may write a smaller latent displacement, or the remaining trajectory may no longer propagate that displacement into ownership. Distinguishing the last mechanism requires measuring ownership change per unit write. For nonzero B_u, let $\widehat{B}_u = B_u/\|B_u\|_2$ and define the local *assignment control authority*

$$\mathcal{C}_{u \to v} = \nabla r(z^{(v)})^\top \Phi_{v \leftarrow u+1} \widehat{B}_u, \qquad \delta r_v \approx \|B_u\|_2 \mathcal{C}_{u \to v}. \tag{4}$$

This signed, state- and time-dependent quantity is the downstream ownership response per unit latent displacement. Together, Eqs. 2–4 factor assignment control into the text-induced denoiser difference (read), its schedule-scaled latent displacement (write), and the downstream response per unit displacement (propagate). Final assignment control requires all three factors: text sensitivity can persist after the write magnitude and downstream response have diminished.

A prompt switch changes the assignment for every remaining step. Its nonlinear rollout asks whether the remaining control can overcome assignment evidence already stored in $z^{(u)}$. Binding crystallizes when the opposite assignment can no longer reverse final ownership. Because both stored evidence and $\mathcal{C}_{u\to v}$ are trajectory-dependent, the boundary varies across prompts and seeds.

These alternatives determine the experimental sequence. Prompt switches test whether accumulated history resists reassignment; short paired pulses separate reading, writing, and propagation; and a time–strength sweep indexes ownership response by the realized latent displacement. A loss of control authority has a distinctive signature: later switches preserve more of the initial assignment even while the late denoiser still distinguishes the alternatives, and comparable assignment-directed writes have weaker persistent effects later.

3.3 Locating the Loss of Reassignment Leverage

A final image reveals which assignment won, but not when the alternative lost the ability to redirect the trajectory. To locate that boundary, let $p =$ "a purple car and an orange house" and let p' swap only the assignment. Starting from identical noise, we use p for the first t_{sw} DDIM steps and p' afterward. If ownership remains malleable, the final image follows p'; persistence of p indicates that the early assignment resists reversal. We compare CLIP image–text logits using

$$s(t_{\mathrm{sw}}) = \frac{\exp \ell(p')}{\exp \ell(p) + \exp \ell(p')}. \tag{5}$$

Because p and p' contain the same words and differ only in assignment, their logit difference controls for lexical content and measures relative support. When an image supports both descriptions equally, $\ell(p) \approx \ell(p')$, so Eq. 5 centers the score at 0.5. Aggregation over prompts traces reassignment dynamics; official metrics evaluate CrystalBind's final outputs.

Evaluating every even switch step over $279/261/287/272$ valid color/shape/texture/complex prompts yields 28,574 SD v1.5 images. The mean curves cross 0.5 at steps $10/8/10/4$, with first-to-last decreases of 9.9–15.2 points (Figure 2). SD v2.1 crosses at $8/4/16/6$ and PixArt-α at $16/12/24/8$; every model–category pair loses responsiveness despite different crossing times.

3.4 Category Timing, Prompt Variation, and Spatial Order

A category-mean crossing locates the shared trend but cannot determine a window that covers heterogeneous prompts. We therefore recover prompt-level variation by linearly interpolating each prompt's 0.5 crossing T_i and treating non-crossing curves as censored. Observed medians are near

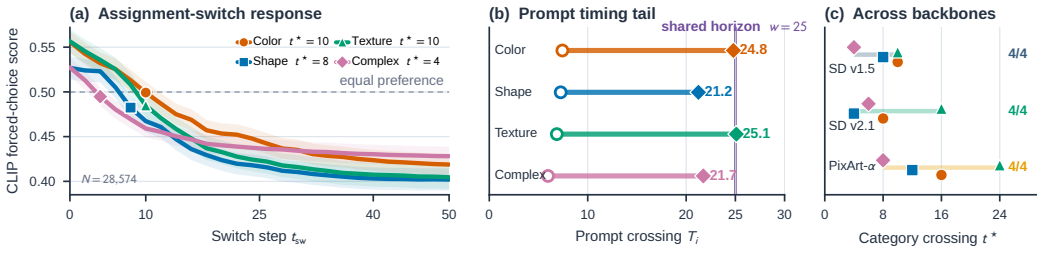


Figure 2: **Binding loses reversibility early, but not at one universal step.** (a) Across four prompt categories, CLIP forced-choice support for the later, swapped assignment decreases with switch time; markers identify the first category-mean crossing of 0.5. (b) Median-to-q_{95} prompt-level ranges expose the long timing tail that sets the shared horizon $w = 25$. (c) All four categories become less responsive to reassignment on each backbone, while their numerical crossings shift with the model.

step 7, whereas $q_{95}(T_i)$ reaches $24.8/21.2/25.1$ for color/shape/texture. The upper tail identifies later-changing prompts that a mean-based endpoint would miss.

The transition time identifies when control is lost, but a spatial controller also needs a usable target before that loss. If noun maps become informative only after commitment, they cannot guide binding in time; if they have already stabilized beforehand, freezing a mature target becomes feasible. The feedback mechanism instead suggests an overlap: coarse noun support emerges early enough to route attributes, while noun locations continue refining through the commitment interval. To distinguish these possibilities, at every vanilla step we compare each parsed noun's normalized 16×16 attention map with the mean of its final five maps. Thresholds of 0.85 and 0.90 mark coarse noun-location support and persistent agreement with the late-denoising reference, respectively.

All categories follow the same order, and within-prompt pairing preserves the positive delay (Figure 3a). In panel (b), trajectories that differ only during two intervening steps retain the largest assignment difference around steps 6–10 and almost none late.

This overlap determines the target design. Before t^*, noun maps already locate candidate objects, yet remain provisional. Waiting for a late map misses the most responsive interval; freezing the first usable map locks in an immature location estimate. CrystalBind instead intervenes during the overlap and refreshes its noun target as the scene evolves; Section 5.2 tests both.

3.5 Why late assignment changes lose downstream influence

Adjacent switch conditions already show that a two-step text difference has its largest persistent effect around steps 6–10 and little effect after step 30. To explain this decline, we run paired two-step pulses. Two branches share their noise and prefix; one receives the opposite assignment for two updates, after which both resume identical text. Repeating both assignment orientations and orienting differences toward the active base assignment cancels static prompt-order preference.

The denoiser response is slightly larger at step 40 than at step 10, showing that assignment sensitivity persists late (Figure 4). Meanwhile, $|\kappa_u|$ falls by more than half, the latent write shrinks, and ownership change per write falls by 88%; final CLIP and BLIP-VQA effects drop by an order of

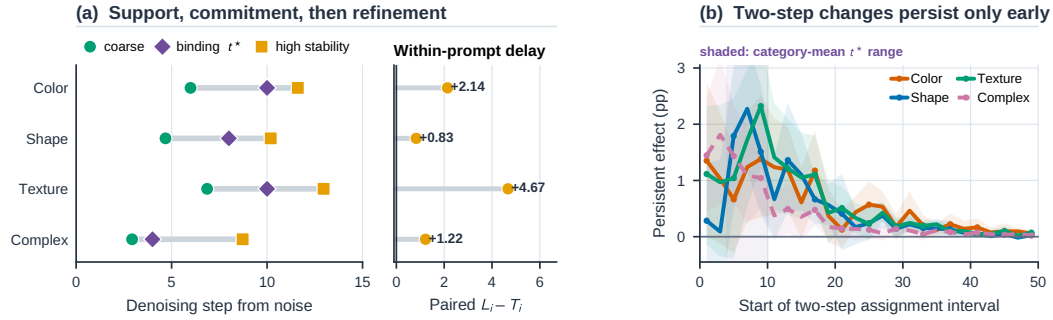


Figure 3: **Ownership commits while noun locations still refine.** (a) Every category follows the same stage order; the aligned right column separately reports the positive within-prompt delay from crystallization to high noun-map stability. (b) A two-step assignment change persists most strongly around the shaded transition range and weakly at late steps. Bands are 95% prompt-level intervals.

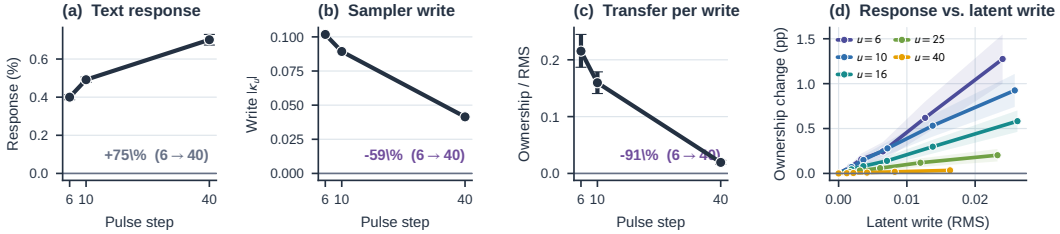


Figure 4: **Why late assignment changes lose downstream influence.** Three-seed paired pulses separate denoiser response, sampler write, and downstream transfer (a–c). Across 275 Color prompts, late perturbations produce weaker ownership changes at comparable realized latent displacements (d; 95% intervals).

magnitude while object presence remains unchanged. These measurements identify a sharp loss of *assignment control authority*: text is still read, but its write barely redirects the trajectory.

The time–strength sweep separates control authority from write magnitude. We apply a two-step assignment-directed perturbation $\epsilon_\gamma = \epsilon_{\text{base}} + \gamma(\epsilon_{\text{opp}} - \epsilon_{\text{base}})$ for $\gamma \in \{0, 0.25, 0.5, 1, 2, 4\}$, then restore the base text. Panel (d) measures ownership after four updates under the shared conditioning against the *actual* latent displacement. At step 10, $\gamma = 1$ writes 0.0072 RMS and changes ownership by 0.279 points; at step 40, $\gamma = 2$ writes more (0.0082 RMS) but changes ownership by only 0.018 points. The local ownership response per unit latent write falls $18.9\times$, validating the predicted decline in control authority at a fixed lag and when response is indexed by the realized latent displacement. Appendices D.10 and D.11 report the complete protocol and response surfaces.

4 From Process Evidence to CrystalBind

4.1 Separating Object Presence from Attribute Ownership

The timing and spatial analyses determine when and where to guide, but the intervention must still distinguish two different errors: an object may be absent, or it may be present with the wrong attribute. These failures require different signals. Noun activation measures whether an object has support, whereas ownership depends on an attribute's relative alignment with its requested noun versus competitors. For K attribute–noun pairs with normalized noun map $m_k^{(u)}$ and attribute map $a_k^{(u)}$, we define the assignment margin

$$b_k^{(u)} = \langle a_k^{(u)}, m_k^{(u)} \rangle - \frac{1}{K-1} \sum_{j \neq k} \langle a_k^{(u)}, m_j^{(u)} \rangle. \tag{6}$$

Strong noun activation does not determine this margin's sign; conversely, a favorable margin cannot preserve an object with negligible noun support. This separation predicts that activation features better identify final presence, whereas assignment features better identify correct binding; interventions on their objectives should affect these outcomes differently.

We test this separation on 2,658 unmodified trajectories from 886 color/shape/texture prompts and three seeds. Activation features summarize noun strength and concentration; assignment features compare every attribute with its requested and competing nouns. Final object presence, correct-versus-swapped binding, and the official benchmark score are measured on the final images, independently of the intermediate attention features.

The trajectory probes recover this specialization, and paired interventions establish its functional consequence (Figure 5). CrystalBind's noun-excitation component restores object presence, while its directional ownership objective improves binding with essentially no further presence change. Together, the activation and assignment signals best predict the official benchmark score (AUC 0.741, versus $0.683/0.704$ separately). Presence is necessary but insufficient for ownership.

These results produce three concrete requirements:

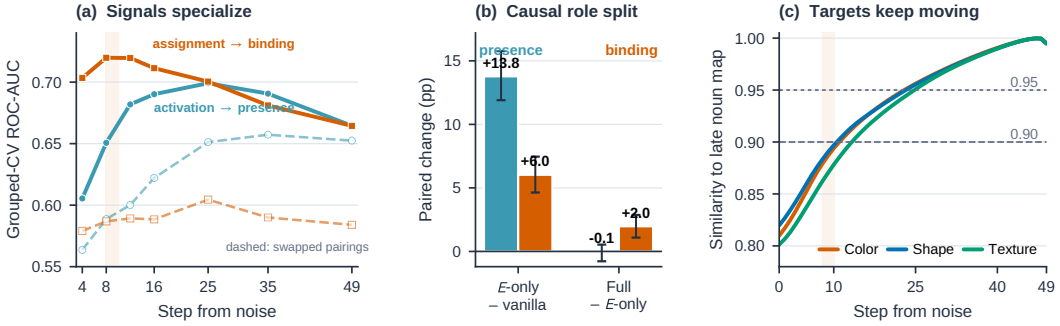


Figure 5: **Presence, ownership, and location evolve differently.** (a) Activation features better predict presence and assignment features better predict correct binding; dashed curves swap the feature sets across outcomes. (b) Within CrystalBind, noun excitation restores presence, while the directional ownership objective improves binding with almost no further presence gain (95% CIs). (c) Noun maps keep moving after crystallization.

- **When:** cover the transition-time *distribution*. The category-level crossing identifies the phenomenon; the prompt-level upper tail covers later-changing prompts without continuing guidance through later steps with little measured effect.
- **Where:** follow the evolving noun-location maps. They are usable before ownership commits but continue to move afterward, so targets should be refreshed rather than delayed or frozen.
- **What:** preserve object presence and direct attribute ownership. A complete objective couples a noun-excitation component with a directional ownership objective.

4.2 Method

We use spaCy to extract attribute–noun pairs (a_k, n_k). At each guided step u, CrystalBind aggregates their cross-attention maps, smooths them, and applies unit-mass normalization

$$\mathcal{N}(x)_i = \frac{x_i}{\sum_j x_j + \varepsilon}. \tag{7}$$

The noun target $\mu_k^{(u)} = \mathrm{stopgrad}[\mathcal{N}(g_\sigma * A_{n_k}^{(u)})]$ is recomputed at every step. Detachment keeps it a reference during the current update; recomputation lets it follow the evolving scene.

Let $\hat{a}_k = \mathcal{N}(g_\sigma * A_{a_k}^{(u)})$ and let $e_k^{(u)} = |S_k|^{-1} \sum_{i \in S_k} (g_\sigma * A_{n_k}^{(u)})_i$, where S_k contains the noun map's five highest-response locations. CrystalBind couples a noun-excitation component for object presence with a directional ownership objective comprising target attraction and competitor suppression:

$$\mathcal{L}_T = -\sum_k \langle \hat{a}_k, \mu_k^{(u)} \rangle, \qquad \mathcal{L}_A = \sum_{k \neq j} \langle \hat{a}_k, \mu_j^{(u)} \rangle, \qquad \mathcal{L}_E = \sum_k (1 - e_k^{(u)}), \tag{8}$$

$$\mathcal{L} = \mathcal{L}_T + \mathcal{L}_A + 0.5\mathcal{L}_E, \quad z^{(u)} \leftarrow z^{(u)} - 2\,\frac{\nabla_{z^{(u)}} \mathcal{L}}{\|\nabla_{z^{(u)}} \mathcal{L}\| + 10^{-8}}. \tag{9}$$

For $u = 0, \ldots, 25$, this guidance update precedes the ordinary DDIM update; steps 26–49 use DDIM alone. The endpoint is calibrated once from the measured $q_{95}(T_i)$ values and held fixed across categories and evaluation seeds. The measured tail thus sets one intervention horizon for single-pass generation that covers later-changing prompts; measuring an individual run's exact transition would require counterfactual continuations that test reversibility.

5 Method Evaluation

5.1 Protocol and main comparison

We compare binding, image quality, and efficiency at a matched update budget. We evaluate 300 prompts per T2I-CompBench color, shape, and texture split with official BLIP-VQA (Huang et al., 2023; 2025; Li et al., 2022); CLIP-IQA measures image quality (Wang et al., 2023). Unless noted,

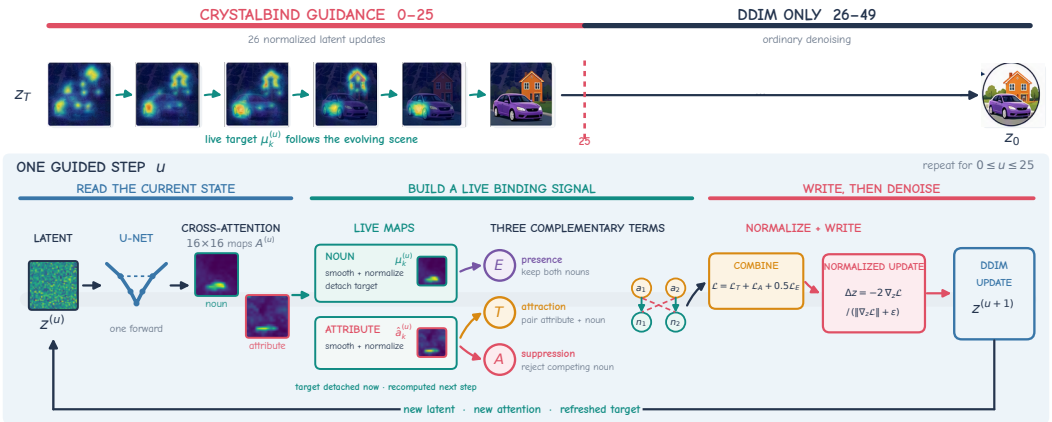


Figure 6: **CrystalBind.** Each guided step uses one gradient-enabled pass to obtain live noun and attribute maps, then applies a normalized latent update before DDIM. Steps 26–49 use DDIM alone.

experiments use 50-step DDIM, classifier-free guidance (CFG) 7.5, and three random seeds. CrystalBind applies one update at steps 0–25.

On SD v1.5, we report the baselines' released schedules and a matched 26-update setting; on SD v2.1, all guided methods use the same 26-update protocol.

Table 1: **Main results: binding, efficiency, and transfer.** Scores are three-seed means; time is averaged over three runs, and IQA averages the three binding categories. Released SD v1.5 rows follow the published schedules; matched rows and all SD v2.1 attribute-binding rows use 26 updates. IR denotes iterative refinement. Complex reports non-MLLM 3-in-1 and MLLM ShareGPT4V-CoT. Parentheses show gains over vanilla; bold and underline mark the best and second-best guided results in each block.

(a) SD v1.5 attribute binding, efficiency, and image quality

Method	Updates	Time ↓	Color ↑	Shape ↑	Texture ↑	Avg. ↑	IQA ↑
Vanilla	0	1.58	0.378	0.362	0.399	0.380	0.779
Released schedules (IR retained where applicable)							
A&E	25+IR	**4.64**	0.581	<u>0.478</u>	0.588	0.549	**0.821**
D&B	25+IR	<u>5.96</u>	<u>0.601</u>	0.461	<u>0.615</u>	<u>0.559</u>	<u>0.799</u>
CONFORM	90	7.70	**0.684**	**0.492**	**0.661**	**0.612**	0.797
26-update schedules (steps 0–25)							
A&E	26	4.48	<u>0.574</u>	<u>0.451</u>	0.543	0.523	<u>0.810</u>
D&B	26	4.55	0.557	0.425	0.561	0.514	0.792
CONFORM	26	<u>4.42</u>	**0.638**	0.439	<u>0.599</u>	<u>0.559</u>	0.804
CrystalBind	26	**3.17**	**0.638** (+0.261)	**0.497** (+0.135)	**0.642** (+0.243)	**0.592** (+0.213)	**0.812**

(b) SD v2.1 attribute binding, efficiency, and image quality

Method	Updates	Time ↓	Color ↑	Shape ↑	Texture ↑	Avg. ↑	IQA ↑
Vanilla	0	2.47	0.518	0.420	0.506	0.481	0.814
A&E	26	12.74	0.644	<u>0.501</u>	0.592	0.579	0.822
D&B	26	14.83	0.524	0.424	0.513	0.487	0.813
CONFORM	26	<u>12.48</u>	**0.684**	<u>0.501</u>	<u>0.657</u>	<u>0.614</u>	<u>0.823</u>
CrystalBind	26	**4.20**	<u>0.673</u> (+0.155)	**0.533** (+0.113)	**0.667** (+0.161)	**0.624** (+0.143)	**0.826**

(c) Complex-prompt composition

Metric	Backbone	Vanilla	A&E	D&B	CONFORM	CrystalBind
3-in-1	SD v1.5	0.309	0.323	0.320	**0.343**	<u>0.330</u> (+0.021)
ShareGPT4V-CoT	SD v1.5	76.711	76.889	<u>77.022</u>	76.778	**77.133** (+0.422)
3-in-1	SD v2.1	0.327	0.339	0.327	**0.349**	<u>0.344</u> (+0.017)
ShareGPT4V-CoT	SD v2.1	77.689	<u>77.911</u>	77.511	77.578	**78.178** (+0.489)

The released schedules provide a higher-compute comparison: CONFORM reaches 0.612 with 90 updates and 7.70 seconds, while A&E and D&B under their published schedules remain below CrystalBind. At 26 updates, CrystalBind improves vanilla by $26.1/13.5/24.3$ color/shape/texture points, or 21.3 points on average (Table 1). It exceeds CONFORM, A&E, and D&B by $3.4/7.0/7.8$ average points; its largest CONFORM margins are on shape ($+5.8$) and texture ($+4.3$), where ownership is often limiting. It is also 28–30% faster and attains the highest CLIP-IQA. Within the matched block, CrystalBind therefore combines the highest average binding and CLIP-IQA with the lowest runtime.

5.2 Validating Dynamics-Guided Design Choices

We connect diagnosis to design by varying one choice at a time: guidance timing, target refresh, and separate control of object presence and attribute ownership.

Moving the same ten updates later steadily lowers the mean (Figure 7a). Extending the early prefix from 16 to 25 adds $1.3/2.4/2.0$ color/shape/texture points before gains flatten; guidance through step 49 slightly reduces all three scores. The measured transition tail therefore identifies an effective endpoint, retaining the responsive prefix while avoiding late updates with weak ownership control.

The spatial-order analysis tests three target strategies: freeze an early map, wait for stability, or refresh it. Live noun targets outperform maps frozen at steps 8 and 13. Even the best delayed schedule, starting after the largest category-mean 0.90 noun-map stability step, trails by 2.64 points: coarse noun-location support is already usable, so delay discards high-leverage steps. These controls show that useful targets appear early and continue to move.

The presence–ownership analysis predicts that noun excitation and directional terms solve different errors. Adding directional ownership to noun excitation improves color/shape/texture by

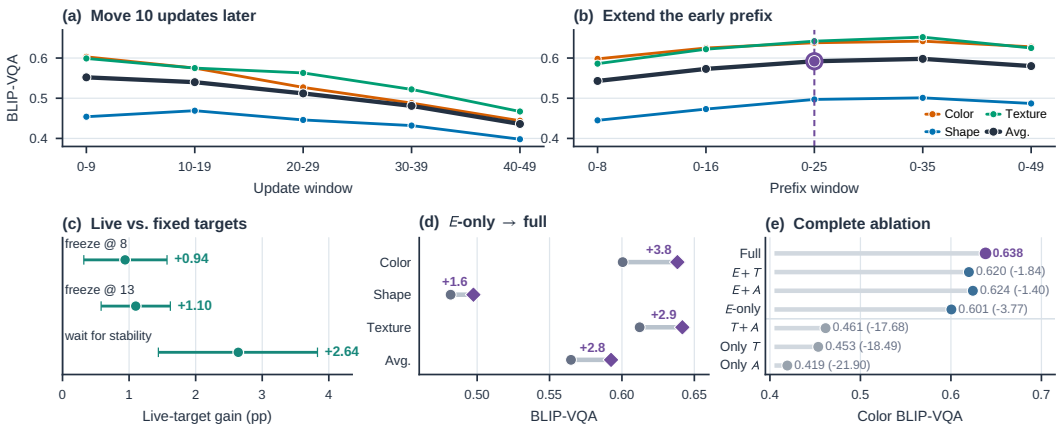


Figure 7: **Design validation from diagnosis to objective.** Matched controls test when to guide (a,b), where to obtain targets (c), and what to optimize (d,e). All use three-seed BLIP-VQA.

3.78/1.57/2.95 points (2.77 on average; Figure 7d). The complete Color ablation in panel (e) resolves the interaction: removing competitor suppression or target attraction costs 1.84 or 1.40 points, and removing both costs 3.77 points. The larger drops without excitation show that ownership guidance also requires object support. Thus noun excitation maintains presence, while the directional objective assigns attributes. Appendices D.14 and D.15 report the full controls.

5.3 Calibration and robustness of the fixed horizon

The upper-tail rule is designed to absorb prompt and initial-noise variation in one schedule. Held-out seeds test transfer across trajectories; a category-specific endpoint tests whether separate schedules provide meaningful efficiency. The $w = 25$ horizon improves vanilla by 28.2/15.9/23.9 color/shape/texture points on two held-out seeds. A category-specific 25/22/25 schedule saves only 3.8% of updates but lowers the mean by 0.25 points, supporting the shared endpoint. Appendices D.3, D.16, and C.4 report the full results.

5.4 Transfer and broader prompts

Transfer should preserve the design rather than reselect it on the target backbone. We therefore keep the objective, normalized step size, and 0–25 window fixed on SD v2.1, and evaluate complex prompts spanning actions, spatial relations, counting, and scenes. CrystalBind improves every SD v2.1 binding category and the average by 14.3 points. At 26 updates, it leads the three-category mean (0.624 versus 0.614 for CONFORM), Shape, Texture, and CLIP-IQA, while running in 4.20 seconds versus 12.48–14.83 for the guided baselines. Across both backbones, it also improves vanilla on complex prompts, ranking second on non-MLLM 3-in-1 and first on MLLM ShareGPT4V-CoT (Table 1; Appendix D.4). The same controller therefore transfers across two Stable Diffusion backbones and beyond simple two-object attribute prompts.

6 Conclusion

Binding crystallization is semantic commitment before visual completion: noun locations keep refining after ownership resists reassignment, so cleaner late maps need not be better guidance targets. Swaps locate this transition, time–strength pulses show that ownership responds less per unit latent write later, and trajectory probes separate presence from ownership. These findings determine CrystalBind's intervention horizon, live noun targets, and joint objective, producing efficient binding gains across two Stable Diffusion backbones. More broadly, a model may still distinguish a condition after that condition loses the ability to redirect the trajectory; semantic counterfactuals can locate this change and guide timely intervention.

AI USE STATEMENT

We used generative AI tools to assist and polish the writing of this paper, including revisions for clarity and concision. We also used these tools to search for and identify relevant literature. The authors have reviewed the AI-assisted material and take full responsibility for the final manuscript.

REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, K. J. Joseph, Apoorv Saxena, Koustav Goswami, and Balaji Vasan Srinivasan. A-STAR: Test-time attention segregation and retention for text-to-image synthesis. In *ICCV*, 2023.

Luca Ambrogioni. The statistical thermodynamics of generative diffusion models. *arXiv preprint arXiv:2310.17467*, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 2024.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*, 2023.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, 2022.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023.

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

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Marvin Li and Sitan Chen. Critical windows: Non-asymptotic theory for feature emergence in diffusion models. *arXiv preprint arXiv:2403.01633*, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *CVPR*, 2023a.

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide and bind your attention for improved generative semantic nursing. In *BMVC*, 2023b.

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

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. CONFORM: Contrast is all you need for high-fidelity text-to-image diffusion models. In *CVPR*, 2024.

Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. *arXiv preprint arXiv:2412.02237*, 2025.

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

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *NeurIPS*, 2024.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *NeurIPS*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences (PNAS)*, 2025.

Takahiro Shirakawa and Seiichi Uchida. NoiseCollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *CVPR*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI*, 2023.

Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. TokenCompose: Text-to-image diffusion with token-level supervision. In *CVPR*, 2024.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *ICML*, 2024.

A Relation to Prior Guidance

Objectives and evaluation. Existing guidance methods emphasize different parts of compositional generation. A&E (Chefer et al., 2023) promotes object presence, Composable Diffusion (Liu et al., 2022) composes concept scores, D&B (Li et al., 2023b) aligns object and attribute attention, and CONFORM (Meral et al., 2024) contrasts paired and competing maps. CrystalBind couples noun excitation with a directional ownership objective, applies both within the measured intervention window, and refreshes its noun targets at each guided step. This design improves binding while attaining the highest CLIP-IQA among the matched 26-update methods.

B Finite-Horizon Assignment Dynamics

The theory follows the causal path measured in the paper. Reverse diffusion accumulates assignment evidence in the latent; an intervention changes one update; and the remaining trajectory either propagates or attenuates that change before the final ownership decision. This section formalizes that finite-horizon response; a bistable normal-form specialization adds a nonlinear-response prediction.

Table 2: Comparison with the closest training-free attention-guidance baselines. CrystalBind changes both the assignment signal and how its noun target is obtained.

Method	Optimized signal	Difference from CrystalBind
A&E	Peak subject/object token activation	Optimizes object presence; CrystalBind additionally models attribute ownership.
D&B	TV-based attendance and symmetric object–attribute map alignment	Aligns paired maps; CrystalBind uses detached live noun targets, competitor suppression, and a measured intervention window.
CONFORM	Contrastive grouping of paired attention maps against other maps	Its released schedule uses repeated refinement; CrystalBind directly attracts and repels attributes with live noun targets.
CrystalBind	Noun excitation with live-target attraction and competitor suppression	Couples presence and ownership control during the measured intervention window.

B.1 Binding state and operational reversibility

Definition 1 (Binding Configuration, State, and Crystallization Point)**.** *For K attribute–noun pairs, a binding configuration is a permutation $\pi \in S_K$; for $K{=}2$ there are two: $\pi_{correct}$ and π_{swap}. Let $u \in \{0, \dots, T\}$ denote denoising progress counted from the noisy end and $z^{(u)}$ the latent after u steps. Let $S_\pi(I)$ denote an evaluator's support score for configuration π in image I. We define the continuation-dependent binding state as $\mathcal{B}_\pi(z^{(u)}, c_{u:T}) \triangleq S_\pi(\mathrm{Decode}(\hat{z}_0(z^{(u)}, c_{u:T})))$ and the preferred terminal configuration as $\arg\max_{\pi \in S_K} \mathcal{B}_\pi(z^{(u)}, c_{u:T})$. A trajectory remains reversible at u when replacing the remaining conditioning by the opposite assignment changes this preferred configuration. Binding crystallization is the boundary beyond which that counterfactual continuation no longer reverses ownership. Experimentally, t^* is the first tested switch point at which the category-mean CLIP forced-choice swap score falls below 0.5 (Figure 2); T_i applies the same definition to an individual observed crossing.*[1]

B.2 Local assignment control authority along a nonlinear trajectory

Write one reverse update as $z^{(j+1)} = F_j(z^{(j)}, c)$ and its state Jacobian as $J_j = \partial F_j / \partial z$. Along a common continuation, $\Phi_{v \leftarrow u+1} = J_{v-1} \cdots J_{u+1}$ transports a perturbation written at step u to a later state v. For any small latent write q_u and ownership readout r,

$$\delta r_v \approx \nabla r(z^{(v)})^\top \Phi_{v \leftarrow u+1} q_u.$$

Setting $q_u = B_u$ recovers Eq. 3. Dividing by $\|B_u\|_2$ gives the assignment control authority $\mathcal{C}_{u \to v}$ in Eq. 4: the signed ownership response per unit assignment-directed latent displacement. The response separates into the conditional-noise difference, the DDIM write magnitude $|\kappa_u| g$, and downstream transfer. Pulses measure these factors at matched states; the time–strength sweep isolates the last by indexing ownership response by realized latent displacement.

A prompt switch applies the opposite assignment throughout the remaining trajectory. Its final effect is the nonlinear accumulation of the corresponding state-dependent responses. As denoising proceeds, evidence for the current assignment can grow in the state while the number and effect of the remaining corrections shrink. The swap experiment evaluates this finite-horizon reversal directly, and the pulse experiments resolve the local response that produces it. Since both quantities depend on $z^{(u)}$, transition times naturally vary across prompts and seeds.

B.3 Why state evidence grows during reverse diffusion

In the two-mode model of Eq. 1, the text term selects a preferred assignment while the latent term records evidence accumulated by earlier updates. Under the shared-covariance Gaussian specialization with latent means $\bar{z}_+$ and $\bar{z}_-$, the expected separation of the two state likelihoods grows as $\bar{\alpha}_u \|\bar{z}_+ - \bar{z}_-\|^2 / \sigma_u^2$. Reverse denoising therefore moves assignment evidence from condition to state: text can redirect an overlapping high-noise state but must increasingly oppose evidence in a low-noise trajectory. All tested model–category pairs show the predicted endpoint decrease.

[1] The color/shape/texture splits use $K_{\text{attr}} = K_{\text{obj}} = 2$; Complex extends to $K > 2$ binding, spatial relations, and counting. Operationally, $\hat{z}_0$ is the final DDIM sample decoded by the base-model VAE.

B.4 LATENT–ATTENTION POSITIVE FEEDBACK

The state evidence is reinforced by a concrete feedback path. An attribute–noun preference changes cross-attention and the conditional noise prediction; the sampler writes that difference into the next latent; the new latent changes the next spatial queries and can strengthen the same preference. Coarse noun-location support can already route this write before the noun map reaches its late-denoising position. This produces the testable order used in the main text: coarse noun-location support, assignment commitment, and continued spatial refinement. The swap, pulse, and noun-map probes measure the three parts separately.

B.5 A BISTABLE NORMAL-FORM SPECIALIZATION

The control-authority framework also admits nonlinear branch formation. A scalar normal form captures this two-assignment regime. Let $\phi = \mathcal{B}_+ - \mathcal{B}_-$ be requested-minus-swapped assignment preference, h an assignment-directed input, and τ an effective control parameter that decreases as commitment strengthens. The lowest-order potential respecting exchange symmetry at $h = 0$ is

$$V(\phi, \tau) = \tfrac{1}{2}(\tau - \tau_c)\phi^2 + \tfrac{b}{4}\phi^4 - h\phi, \qquad \Delta\phi \propto -\partial_\phi V, \quad b > 0. \tag{10}$$

Its stationary points satisfy $(\tau-\tau_c)\phi+b\phi^3-h = 0$. At $h = 0$, two stable branches appear below τ_c, at $\phi = \pm\sqrt{(\tau_c - \tau)/b}$, with barrier $\Delta V = (\tau_c - \tau)^2/(4b)$. Differentiating the stationary condition with respect to h gives the local response on either branch,

$$\chi_\pm = \left.\frac{\partial\phi}{\partial h}\right|_{h=0,\phi_\pm} = [V''(\phi_\pm)]^{-1} = \frac{1}{2(\tau_c - \tau)}.$$

The metastable branch disappears at the spinodal field

$$|h_{\mathrm{sp}}| = \frac{2(\tau_c - \tau)^{3/2}}{3\sqrt{3b}}.$$

This specialization adds an amplitude prediction to the general finite-horizon account: as commitment deepens, the local response slope decreases and a larger directional input is needed for reversal. The time–strength sweep tests this prediction by varying only the magnitude and timing of a two-step assignment perturbation before restoring identical conditioning.

C METHOD DETAILS

C.1 IMPLEMENTATION SEQUENCE

CrystalBind takes prompt $\mathcal{P}$, horizon w, and a 3×3 Gaussian kernel g_σ ($\sigma = 0.5$):

1. Parse $\mathcal{P}$ into $\{(a_k, n_k)\}_{k=1}^K$ and sample $z_T \sim \mathcal{N}(0, I)$.
2. For each guided step $u = 0, \ldots, w$, enable gradients on the current latent and run the conditional UNet to collect $A^{(u)}$.
3. Compute $\hat{a}_k = \mathcal{N}(g_\sigma * A_{a_k}^{(u)})$ and the detached live noun target $\mu_k^{(u)} = \mathrm{stopgrad}(\mathcal{N}(g_\sigma * A_{n_k}^{(u)}))$.
4. Evaluate the loss and apply the normalized update in Eq. 9.
5. Run the standard no-gradient classifier-free-guidance forward on the updated latent and take the DDIM step. Outside the window, execute only this standard step.

C.2 CRYSTALBIND AS A FEEDBACK INTERVENTION

At each guided step, CrystalBind observes the current noun and attribute maps and writes the normalized direction $q_u = -2\nabla_z\mathcal{L}/(\|\nabla_z\mathcal{L}\|+10^{-8})$ into the latent. Its effect on ownership follows the same propagated response as Eq. 4. Recomputing the noun target after each state transition makes this a feedback intervention: the next correction follows the scene produced by the previous write. The measured intervention horizon determines when these updates are applied, while the directional ownership objective aligns them with requested ownership.

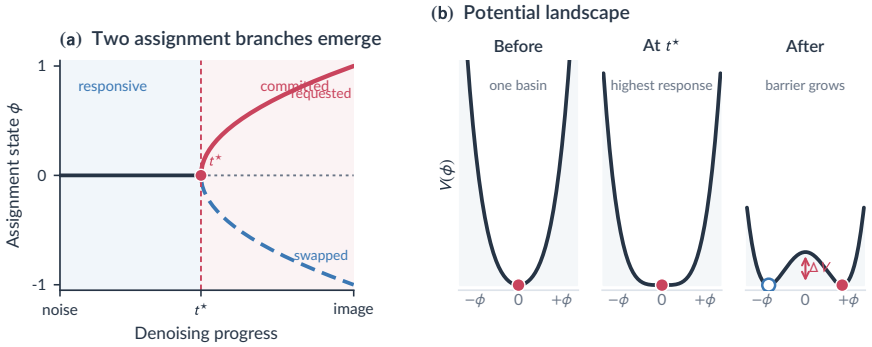


Figure 8: **A bistable specialization of assignment commitment.** Denoising moves the effective state from weak preference toward one of two assignment branches. The directional input h favors one assignment; increasing branch curvature lowers local response, while the growing barrier raises the input required for reversal.

C.3 Contrastive form of the directional ownership objective for $K{=}2$

For $K{=}2$, the target-attraction and competitor-suppression terms in Eq. 8 reduce exactly to

$$\mathcal{L}_T + \mathcal{L}_A = \langle \hat{a}_1 - \hat{a}_2,\ \mu_2^{(u)} - \mu_1^{(u)} \rangle.$$

The sum is therefore a direct contrast between the two attribute maps and the two noun targets. This structure explains the coupled ablation pattern: removing A or T costs 1.84 or 1.40 points, whereas removing both costs 3.77 points (Figure 7e). The explicit attraction and suppression terms extend the same directional ownership objective to $K > 2$, where each entity has multiple competitors.

C.4 Why unit-mass normalization

The directional ownership objective compares how an attribute distributes attention across noun regions. Unit-mass normalization preserves this allocation and makes each dot product an overlap between spatial distributions. Per-map min–max normalization forces even a low-contrast map to span $[0, 1]$, amplifying weak fluctuations and making overlap depend on the number of high-valued pixels. In the three-seed Color control, replacing unit mass with min–max normalization reduces BLIP-VQA by 12.6 points. This result supports interpreting Eq. 8 as spatial allocation.

C.5 Computational cost

CrystalBind adds one gradient-enabled UNet forward and backward pass at each guided step. The repeated single-device timing measurement in Table 5 gives 3.17 ± 0.08 seconds per prompt, or $2.01\times$ the 1.579 ± 0.004-second vanilla reference.

D Experiment Details

D.1 Qualitative examples

Figure 9 compares all five methods on fixed examples from color, shape, and texture. Table 1 reports quantitative results over all prompts and seeds.

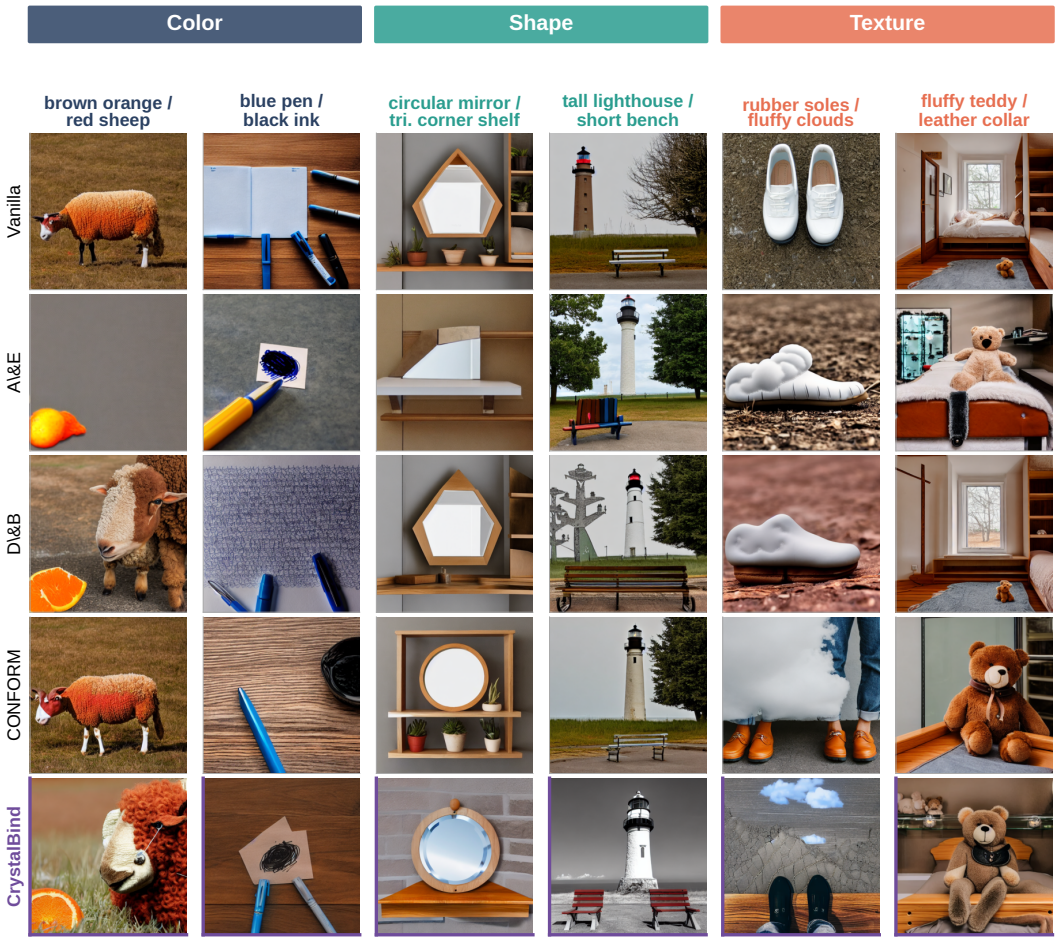


Figure 9: **Qualitative comparison on T2I-CompBench.** All methods share six fixed prompt–noise pairs.

D.2 Baseline operating points

The released CONFORM configuration (Meral et al., 2024) combines 30 guided denoising steps with three embedded refinement loops, for approximately 90 attention-loss updates. The 26-update operating point evaluates the guidance objectives at CrystalBind's schedule. CrystalBind matches CONFORM's color mean and improves shape/texture by $5.8/4.3$ points at $w = 25$, with corresponding gains of $0.8/5.1/5.7$ points at $w = 16$.

Table 3: Comparison at the shorter CrystalBind endpoint $w = 16$ on SD v1.5 (three-seed BLIP-VQA). A&E and D&B retain iterative refinement in their released schedules. CrystalBind is shaded; maxima are bold.

Method	Color ↑	Shape ↑	Texture ↑	Avg. ↑
$w = 16$ **schedules without iterative refinement**				
CONFORM	0.617 ± 0.029	0.422 ± 0.027	0.565 ± 0.015	0.535
CrystalBind	$\mathbf{0.625 \pm 0.017}$	$\mathbf{0.473 \pm 0.011}$	$\mathbf{0.622 \pm 0.008}$	**0.573**
Released schedules with iterative refinement				
A&E	0.575 ± 0.027	0.470 ± 0.027	0.586 ± 0.009	0.544
D&B	0.585 ± 0.009	0.453 ± 0.027	0.608 ± 0.027	0.549

D.3 Held-out seed robustness

On two additional held-out seeds, the fixed $w = 25$ horizon improves vanilla by 28.2 ± 1.1, 15.9 ± 2.6, and 23.9 ± 1.4 points on color, shape, and texture, respectively.

D.4 COMPLEX-PROMPT RESULTS

T2I-CompBench++ identifies Complex 3-in-1 as its strongest non-MLLM composite for the heterogeneous complex split and ShareGPT4V-CoT as a multimodal large language model (MLLM)-based evaluator. Table 4 reports both the decomposed automatic composite and the MLLM-based evaluator. A&E, D&B, and CONFORM use their released schedules on each backbone.

Table 4: Complex split on two backbones (three-seed mean ± sample std). Shading identifies CrystalBind; bold marks the highest mean within each backbone and evaluator.

Backbone	Method	Complex 3-in-1 ↑	ShareGPT4V-CoT ↑
SD v1.5	Vanilla	0.309 ± 0.007	76.711 ± 0.500
	A&E	0.323 ± 0.006	76.889 ± 0.278
	D&B	0.320 ± 0.006	77.022 ± 0.336
	CONFORM	$\mathbf{0.343 \pm 0.007}$	76.778 ± 0.168
	CrystalBind	0.330 ± 0.004	$\mathbf{77.133 \pm 0.176}$
SD v2.1	Vanilla	0.327 ± 0.013	77.689 ± 0.315
	A&E	0.339 ± 0.011	77.911 ± 0.204
	D&B	0.327 ± 0.011	77.511 ± 0.434
	CONFORM	$\mathbf{0.349 \pm 0.008}$	77.578 ± 0.269
	CrystalBind	0.344 ± 0.011	$\mathbf{78.178 \pm 0.192}$

D.5 RUNTIME AND IMAGE QUALITY

Runtime is measured on one NVIDIA A800 GPU with batch size 1 and 50-step DDIM. We average three 20-prompt runs after warm-up and exclude evaluator time; SD v1.5 and SD v2.1 use their native 512^2 and 768^2 resolutions, respectively.

Table 5: Wall-clock and color CLIP-IQA on SD v1.5. Time is mean ± sample std over three repeats; CLIP-IQA is mean ± sample std over three random seeds. CrystalBind is shaded; protocol maxima are bold.

Method	Updates ↓	Time (s/prompt) ↓	CLIP-IQA ↑
Vanilla	0	1.579 ± 0.004	0.828 ± 0.002
Released schedules			
A&E	25+IR	$\mathbf{4.64 \pm 0.13}$	$\mathbf{0.865 \pm 0.014}$
D&B	25+IR	5.96 ± 0.16	0.851 ± 0.012
CONFORM	90	7.70 ± 0.08	0.846 ± 0.010
26-update schedules (steps 0–25)			
A&E	26	4.48 ± 0.07	0.864 ± 0.014
D&B	26	4.55 ± 0.09	0.852 ± 0.018
CONFORM	26	4.42 ± 0.05	0.858 ± 0.006
CrystalBind	26	$\mathbf{3.17 \pm 0.08}$	$\mathbf{0.876 \pm 0.012}$

D.6 PER-CATEGORY CLIP-IQA OVER THREE SEEDS

Table 5 reports the three-seed color CLIP-IQA used alongside the timing comparison; Table 6 gives all three binding categories. We use CLIP ViT-B/32 and the quality-prompt forced choice $\mathrm{softmax}(100[\cos(I, \text{“Good photo.”}), \cos(I, \text{“Bad photo.”})])_1$, averaging images within each seed before computing the across-seed mean and sample standard deviation.

Among the 26-update schedules, CrystalBind attains the highest average CLIP-IQA (0.812).

D.7 TRANSFER TO SD V2.1

CrystalBind transfers to SD v2.1 with the same loss, step size, and intervention window. We use the backbone's native 24×24 cross-attention resolution and evaluate the three 300-prompt T2I-CompBench binding splits over three random seeds with official BLIP-VQA. All guided methods apply one update per step over steps 0–25.

CrystalBind improves vanilla by $+15.5/+11.3/+16.1$ points on Color/Shape/Texture ($+14.3$ points on average). At the same 26-update budget, CONFORM leads on Color, while CrystalBind

Table 6: Per-category CLIP-IQA on SD v1.5 (three-seed mean ± sample std). Avg. is the mean of the displayed category means. Bold and underline mark the highest and second-highest means within each protocol group.

Method	Color ↑	Shape ↑	Texture ↑	Avg. ↑
Vanilla	0.828 ± 0.002	0.757 ± 0.034	0.752 ± 0.014	0.779
(a) Released schedules with iterative refinement				
A&E	$\mathbf{0.865 \pm 0.014}$	$\mathbf{0.790 \pm 0.012}$	$\mathbf{0.807 \pm 0.018}$	**0.821**
D&B	$\underline{0.851 \pm 0.012}$	0.774 ± 0.021	$\underline{0.773 \pm 0.014}$	$\underline{0.799}$
CONFORM	0.846 ± 0.010	$\underline{0.778 \pm 0.022}$	0.767 ± 0.016	0.797
(b) 26-update schedules (steps 0–25)				
A&E	$\underline{0.864 \pm 0.014}$	$\mathbf{0.782 \pm 0.022}$	$\mathbf{0.785 \pm 0.027}$	$\underline{0.810}$
D&B	0.852 ± 0.018	0.759 ± 0.027	0.764 ± 0.013	0.792
CONFORM	0.858 ± 0.006	$\underline{0.782 \pm 0.029}$	0.772 ± 0.010	0.804
CrystalBind	$\mathbf{0.876 \pm 0.012}$	0.777 ± 0.016	$\underline{0.784 \pm 0.022}$	**0.812**

Table 7: SD v2.1 comparison with 26 updates per guided method. Binding and CLIP-IQA use three seeds; time is mean ± sample std over three 20-prompt runs. Bold and underline mark the top two guided results.

(a) Binding performance

Configuration	Updates	Color ↑	Shape ↑	Texture ↑	Avg. ↑
Vanilla SD v2.1	0	0.518 ± 0.068	0.420 ± 0.030	0.506 ± 0.035	0.481
A&E	26	0.644 ± 0.021	$\underline{0.501 \pm 0.024}$	0.592 ± 0.029	0.579
D&B	26	0.524 ± 0.066	0.424 ± 0.027	0.513 ± 0.037	0.487
CONFORM	26	$\mathbf{0.684 \pm 0.015}$	0.501 ± 0.025	$\underline{0.657 \pm 0.025}$	$\underline{0.614}$
CrystalBind	26	$\underline{0.673 \pm 0.025}$	$\mathbf{0.533 \pm 0.020}$	$\mathbf{0.667 \pm 0.012}$	**0.624**

(b) Efficiency and image quality

Configuration	Updates	Time (s/prompt) ↓	CLIP-IQA Avg. ↑
Vanilla SD v2.1	0	2.47 ± 0.02	0.814
A&E	26	12.74 ± 0.01	0.822
D&B	26	14.83 ± 0.01	0.813
CONFORM	26	$\underline{12.48 \pm 0.03}$	$\underline{0.823}$
CrystalBind	26	$\mathbf{4.20 \pm 0.12}$	**0.826**

leads on Shape and Texture and has the highest three-category binding mean (0.624 versus 0.614). CrystalBind also has the lowest guided runtime (4.20 seconds per prompt) and highest average CLIP-IQA (0.826), preserving its advantages in binding, efficiency, and image quality after transfer from SD v1.5 to SD v2.1.

The per-category CLIP-IQA means (Color/Shape/Texture) are $0.841/0.801/0.799$ for vanilla, $0.856/0.802/0.809$ for A&E, $0.839/0.801/0.799$ for D&B, $0.856/0.800/0.813$ for CONFORM, and $0.869/0.809/0.800$ for CrystalBind. Each value averages 300 images within each seed and then averages over three random seeds.

On the SD v2.1 complex split (Table 4), CrystalBind improves over vanilla on every paired seed under both metrics. CONFORM has the highest 3-in-1 mean, whereas CrystalBind has the highest ShareGPT4V-CoT mean, providing complementary non-MLLM and MLLM evaluations.

D.8 Swap dynamics on SD v2.1

Figure 2(c) pairs SD v1.5 (UNet with CLIP) with PixArt-α (DiT with T5). We additionally run the full four-category swap experiment on SD v2.1, a second UNet with OpenCLIP ViT-H/14. The protocol follows Section 3.3: a single random seed, 50-step DDIM, 26 switch points $t_{\text{sw}} \in \{0, 2, \ldots, 50\}$, and the same attribute–noun parser, with the backbone's native 768^2 resolution. The four splits retain 261–287 prompts per category.

Model	Category	n	t^* (% of 50)	Swap score at 0	$s(0) - s(50)$
SD v2.1	Color	279	8 (16%)	0.543	0.099
	Shape	261	4 (8%)	0.523	0.102
	Texture	287	16 (32%)	0.565	0.115
	Complex	272	6 (12%)	0.524	0.080

All four SD v2.1 crystallization points ($t^* \in \{4, 6, 8, 16\}$) fall within the first 32% of denoising, matching the SD v1.5 pattern qualitatively; every category retains the same overall descending trend. Figure 10 overlays the color curve against SD v1.5 and PixArt-α on a shared axis. Together, CLIP, OpenCLIP, and T5 text encoders, four attribute types, and two backbone classes (UNet and DiT) provide the cross-backbone evidence cited in §3.3.

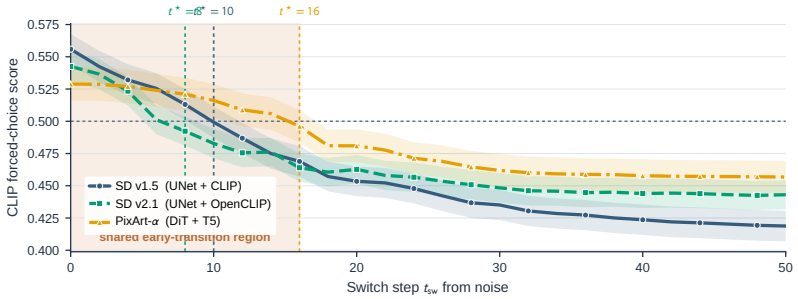


Figure 10: Cross-architecture swap curves (color category) for SD v1.5, SD v2.1, and PixArt-α. All three backbones exhibit an overall decline with $t^* \in [8, 16]$ of 50 DDIM steps.

D.9 Per-prompt crystallization: distribution of T_i

§3.4 visualizes the empirical coverage curves of per-prompt crystallization points T_i. Table 8 gives the full four-category tally and quantiles. For each prompt i, we linearly interpolate the first 0.5 crossing between sampled switch steps $t_{\text{sw}} \in \{0, 2, \ldots, 50\}$. Curves below 0.5 at the first switch point are left-censored, and curves above 0.5 at the last are right-censored; the quantiles summarize trajectories with an observed crossing.

Table 8: Per-prompt T_i distribution, full tally (SD v1.5). Counts are out of the total parsed prompts per category. The two censoring columns report trajectories whose crossing is not observed within the sampled interval; quantiles summarize the observed crossings. Their distribution is right-skewed (median $\approx$ 7, $q_{95} \approx 25$); shading identifies the quantile used to select the w=25 window (§5.2).

	Prompt counts				Observed crossing step T_i					
Category	Total	Crossed	Low-cens.	High-cens.	Median	Mean	q_{75}	q_{90}	q_{95}	q_{99}
Color	279	138	78	63	7.4	9.3	13.2	18.9	24.8	33.7
Shape	261	113	93	55	7.3	9.3	13.3	19.2	21.2	32.0
Texture	287	145	79	63	6.9	9.3	12.2	19.8	25.1	40.2
Complex	272	103	99	70	6.0	8.1	11.1	18.3	21.7	31.3

Calibration in new settings. The long right tail ($q_{99} \in [31, 40]$) identifies a minority of prompts with delayed crossings. New prompt distributions, samplers, and backbones can calibrate their horizon from the corresponding transition-time distribution.

Calibration on held-out prompts. We test whether the shared endpoint is stable to the prompts used for calibration. We partition each 300-prompt split into five prompt folds; for each fold, one global endpoint is selected from the maximum color/shape/texture q_{95} on the 240 calibration prompts, then applied to the 60 held-out prompts in every category over three random seeds. The five prompt-fold endpoints are 25/25/27/26/26. Across the held-out folds, they cover 95.5% of observed crossings and preserve the fixed-$w = 25$ operating point (Table 9).

Effect of a two-step assignment change. For every prompt, adjacent switch conditions are paired: the $t_{\text{sw}} = t$ and $t + 2$ trajectories share the same prefix through t, receive opposite assignment

Table 9: Five-fold prompt-held-out window calibration. Crossing coverage is computed over held-out prompts with an observed crossing; BLIP-VQA is the official three-seed mean on the complete held-out union.

Held-out result	Color	Shape	Texture	Aggregate
Crossing coverage (%)	95.7	96.5	94.5	95.5
Fold-calibrated BLIP-VQA	0.639	0.497	0.648	0.595
Fixed $w = 25$ BLIP-VQA	0.638	0.497	0.642	0.592

conditions only during the next two steps, and then share the swapped condition for the remainder. Figure 3(b) reports $\Delta_2(t) = 100[s_i(t) - s_i(t+2)]$ averaged over prompts, where s_i is the swap score in Eq. 5 for prompt i; 95% intervals use 10,000 prompt-level bootstrap resamples. Comparing equal-duration regions, the cumulative response during steps 0–19 exceeds that during steps 30–49 by 8.61 points for color (95% CI $[6.12, 11.20]$), 9.46 for shape ($[6.85, 12.18]$), 11.23 for texture ($[8.65, 13.86]$), and 8.09 for complex prompts ($[5.72, 10.54]$).

D.10 Assignment perturbation and propagation

The pulse experiment in Figure 11 uses two-object, one-attribute-per-object subsets containing 275 Color, 287 Shape, and 236 Texture prompts. Each branch starts from the same saved latent; the perturbed branch receives the opposite assignment for two DDIM updates and then resumes the base condition. Repeating both assignment orientations and orienting the response toward the active base assignment removes fixed wording-order preference.

We measure the conditional-noise response, latent write, fixed-lag change in the attribute–noun overlap ratio, and final CLIP and targeted BLIP-VQA assignment effects; noun-only BLIP-VQA tracks object presence. The full temporal sweep covers pulse starts $\{2, 6, 10, 16, 25, 40\}$ and readout lags $\{0, 4, 12\}$ with one seed; key early, transition, and late starts $\{6, 10, 40\}$ are replicated over three seeds. Confidence intervals use a cluster bootstrap over benchmark prompts.

Table 10: Three-seed pooled step-10 versus step-40 pulse comparison. Brackets in the Δ column are 95% intervals from a cluster bootstrap over prompts; Δ is the paired early-minus-late contrast. CLIP, BLIP-VQA, and object-presence effects are percentage points.

Readout	Step 10	Step 40	$\Delta(10-40)$
Attention cosine-margin difference	0.0770	0.0307	0.0463 [0.0409, 0.0518]
Relative denoiser difference (%)	0.492	0.701	−0.210 [−0.232, −0.188]
Analytic DDIM $\lvert\kappa_u\rvert$	0.0893	0.0414	0.0479
Two-step latent displacement (RMS)	0.00691	0.00373	0.00317 [0.00303, 0.00333]
Change in overlap ratio / write RMS	0.1596	0.0194	0.1402 [0.1219, 0.1589]
Final CLIP assignment effect	0.694	0.044	0.650 [0.455, 0.857]
Final BLIP-VQA assignment effect	0.633	0.040	0.593 [0.435, 0.759]
Object-presence effect	0.008	−0.019	0.027 [−0.227, 0.284]

At the matched four-update readout lag, the step-10 pulse retains more overlap-ratio change per unit initial write than the step-40 pulse: the paired difference is 0.155 with 95% CI $[0.112, 0.198]$. This comparison isolates the temporal decline in ownership change per unit write from the length of post-pulse evolution. Pooled targeted BLIP-VQA follows CLIP, with positive step-10 versus step-40 contrasts in both orientations (CLIP: $0.47/0.83$ points; BLIP-VQA: $0.65/0.54$ points).

D.11 Time–strength response versus realized latent displacement

The binary opposite-assignment pulse above uses a fixed perturbation strength. To test the response prediction of Section 3.2, we vary the magnitude of the same assignment-directed perturbation while holding its duration, continuation, and readout time fixed. The sweep evaluates both orientations for all 275 Color prompts in the two-object, one-attribute-per-object subset with one seed.

At each pulse step, we evaluate both conditional predictions at the same latent and define

$$\epsilon_\gamma = \epsilon_{\text{base}} + \gamma\big(\epsilon_{\text{opp}} - \epsilon_{\text{base}}\big), \qquad \gamma \in \{0, 0.25, 0.5, 1, 2, 4\}. \tag{11}$$

Thus $\gamma = 0$ is the unperturbed update, $\gamma = 1$ is the opposite-assignment pulse, and larger values continue along the same assignment direction. We apply this perturbation for two updates beginning

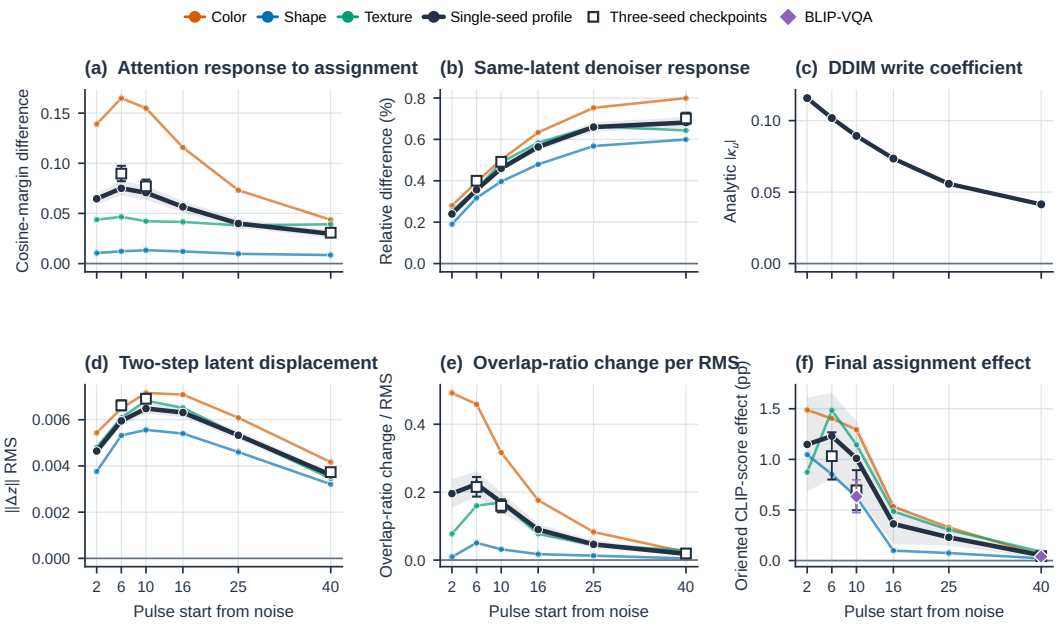


Figure 11: **Reading an assignment and rewriting ownership are different capabilities.** Curves trace the full six-time sweep; open squares show three-seed replications at steps 6, 10, and 40 (95% intervals from a cluster bootstrap over prompts). At the same latent, the denoiser remains assignment-sensitive (a,b), while the DDIM write coefficient, latent displacement, and ownership change under the shared continuation all decrease later (c–e). Targeted BLIP-VQA corroborates the final CLIP pattern (f).

at steps $\{2, 6, 10, 16, 25, 40\}$ and then restore the base conditioning. After four updates under identical conditioning, we measure the decrease in the base assignment's attribute–noun overlap ratio. We also record the RMS latent displacement at the pulse endpoint. The within-batch $\gamma = 0$ branch provides the numerical reference, so every zero-strength response is exactly zero; responses for the two assignment orientations are computed before alignment.

The response contracts sharply with denoising progress (Figure 12). Averaged over steps 6 and 10, the local fixed-lag slope is 0.00272 per unit γ, versus 0.000343 over steps 25 and 40, a $7.9\times$ reduction. This ordering remains after accounting for how much the sampler writes. The local ownership response per unit latent RMS falls from 0.394 at step 10 to $0.205/0.097/0.021$ at steps 16/25/40. Matching the step-10, $\gamma = 1$ response requires $\gamma = 1.88$ at step 16; neither step 25 nor step 40 reaches it within $\gamma \leq 4$. At step 40, $\gamma = 4$ writes $2.28\times$ the step-10 reference displacement but retains only 12.2% of its ownership response. The final CLIP readout gives the same temporal ordering near the unperturbed branch: its mean slope over steps 6/10 is $6.7\times$ that over steps 25/40.

The fixed-lag comparison equalizes post-pulse evolution, and indexing response by realized latent displacement isolates assignment control authority from the DDIM write coefficient. The sweep validates the local response consequence of the commitment model: after assignment becomes established, a state change of the same size has substantially less effect on ownership.

CrystalBind therefore concentrates its updates in the earlier, responsive interval.

D.12 Noun-location stability probe

We run vanilla SD v1.5 with the same 50-step DDIM convention as the swap experiment and extract two noun tokens with CrystalBind's parser. The probe retains $290/295/300/300$ prompts for color/shape/texture/complex, more than the swap experiment because it needs only two valid nouns rather than two complete swappable attribute–noun pairs. At every step, we record each noun's conditional 16×16 cross-attention map while advancing the latent with standard classifier-free guidance. Each map is Gaussian-smoothed, clamped nonnegative, and normalized to unit mass. Its late-state reference is the mean of the same noun's final five normalized maps. We average current-to-reference cosine similarity across the two nouns and define noun-map stability as the first threshold crossing sustained at all later steps; L_i denotes this value at 0.90.

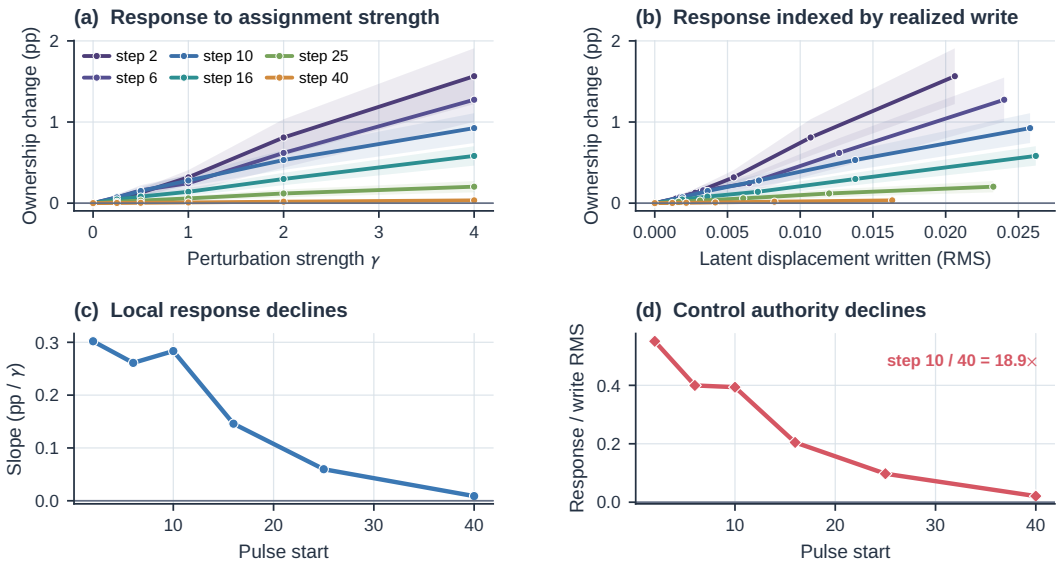


Figure 12: **Time–strength response of attribute ownership.** (a) Fixed-lag ownership response as perturbation strength increases. (b) The same curves indexed by the latent displacement actually written during the pulse. (c) Local slope for $\gamma \leq 1$. (d) Local ownership response per unit latent write. Means pair the two base-assignment orientations within each of 275 prompts.

Figure 3 compares category-level diagnostics: t^* is the category-mean swap crossing, and noun-map stability is the mean prompt-level stability step. For prompts with both an observed T_i and a defined L_i, the mean $L_i - T_i$ at threshold 0.90 is positive in every split: $+2.14$ (color, 137 prompts), $+0.83$ (shape, 111), $+4.67$ (texture, 145), and $+1.22$ (complex, 103).

Table 11: Threshold scan for the noun-location stability probe. Entries are mean noun-map stability steps; lower thresholds detect coarse noun-location support, whereas higher thresholds require closer agreement with the late-denoising noun maps. Shading marks the 0.90 high-stability threshold used in the main analysis.

Cosine threshold	Color	Shape	Texture	Complex
0.85	5.99	4.68	6.86	2.93
0.89	10.10	8.74	11.28	7.00
0.90	11.61	10.19	12.96	8.72
0.95	22.72	21.64	24.03	21.65

D.13 Intermediate-trajectory diagnostics

The trace analysis in Figure 5 uses all three seeds and retains 292/298/296 color/shape/texture prompts. The same pattern holds for the 275/287/236 two-object, one-attribute-per-object prompts.

At each step, activation features summarize noun strength, concentration, and entropy; assignment features summarize requested-versus-competing overlap and cosine margins. Final object presence is measured with noun-only BLIP-VQA, and correct binding uses the normalized correct-versus-swapped VQA score. For each outcome, a logistic-regression probe with standardized features and five-fold cross-validation classifies the top versus bottom quartile within each category–seed group. All seeds of the same benchmark prompt remain in one fold, and we report step-8 out-of-fold AUC.

Table 12: Step-8 out-of-fold AUC from five-fold grouped cross-validation, with 95% intervals from a cluster bootstrap over prompts. The outcome labels are computed from final images, independently of the intermediate attention features. Bold and underline mark the highest and second-highest AUC in each column.

Features	Object presence AUC ↑	Correct binding AUC ↑	Official BLIP-VQA AUC ↑
Activation	<u>0.651</u> [0.611, 0.686]	0.587 [0.551, 0.623]	0.683 [0.647, 0.718]
Assignment	0.589 [0.553, 0.626]	<u>0.720</u> [0.689, 0.752]	<u>0.704</u> [0.671, 0.735]
Combined	**0.664** [0.625, 0.697]	**0.731** [0.699, 0.762]	**0.741** [0.708, 0.772]

The intervention comparison evaluates the same final-image quantities on paired outputs over three seeds. Table 13 reports hierarchical paired-bootstrap intervals over seeds and prompt identities. The first contrast isolates the object-presence effect of CrystalBind's noun-excitation component. The second compares the full controller with noun-excitation-only CrystalBind, showing that directional ownership improves assignment while leaving presence essentially unchanged.

Table 13: Effects on final-image object presence and binding over three seeds for the 292 Color prompts with valid parser outputs. Entries are paired changes in percentage points with 95% CIs. "Correct binding" is the BLIP-VQA correct-versus-swapped forced choice defined above; shading marks the dominant effect.

Intervention	Δ object presence (pp) $\uparrow$	Δ correct binding (pp) $\uparrow$
CrystalBind (E-only) $-$ Vanilla	$+\mathbf{13.78}$ $[11.91, 15.77]$	$+6.03$ $[4.64, 7.46]$
Full CrystalBind $-$ E-only	-0.12 $[-0.78, 0.54]$	$+\mathbf{1.98}$ $[1.09, 2.87]$

At step 8, the requested-versus-competing overlap ratio is positively associated with final correct-versus-swapped binding in color, shape, and texture ($\rho = 0.436/0.123/0.266$). On the two-object, one-attribute-per-object subset, activation/assignment AUC is $0.663/0.600$ for presence and $0.601/0.730$ for binding; official BLIP-VQA AUC is $0.654/0.708$ separately and 0.727 combined.

The three-seed trajectories reproduce the noun-map timing result. The mean step at which current-to-late cosine similarity remains above 0.90 is $10.77/10.33/12.74$ for color/shape/texture; at 0.95 it is $21.66/21.75/23.45$. Thus noun maps keep refining after coarse noun-location support emerges.

D.14 Window and component ablations

Table 14: Full three-seed window comparison for CrystalBind (SD v1.5, official BLIP-VQA, mean $\pm$ sample std). Avg. is the mean of the three category means. The shaded row is the distribution-calibrated $w = 25$ setting; bold marks the highest measured mean.

Endpoint w	Color $\uparrow$	Shape $\uparrow$	Texture $\uparrow$	Avg. $\uparrow$	Interpretation
8	0.598 ± 0.031	0.445 ± 0.018	0.586 ± 0.020	0.543	rapid early gain
16	0.625 ± 0.017	0.473 ± 0.011	0.622 ± 0.008	0.573	continued gains
25	0.638 ± 0.014	0.497 ± 0.015	0.642 ± 0.008	0.592	measured endpoint
35	$\mathbf{0.642 \pm 0.020}$	$\mathbf{0.501 \pm 0.008}$	$\mathbf{0.652 \pm 0.006}$	**0.598**	small additional gain
49	0.628 ± 0.017	0.487 ± 0.018	0.625 ± 0.011	0.580	no consistent benefit

Window robustness. Most improvement appears early, and extending $w = 16$ to the upper-tail intervention horizon $w = 25$ improves every category. Extending to $w = 35$ adds less than one average point, while guidance through step 49 mildly regresses all three scores. The measured tail identifies a useful intervention horizon without spending updates on low-effect later steps.

Component interactions. Figure 7(e) reports the complete seven-way, three-seed CrystalBind component ablation. The noun-excitation component sharpens the noun peaks used by the target term and establishes object presence; the directional ownership objective then assigns attributes across those noun regions. T-only and A-only remain close to the unguided regime, while the full controller produces the strongest official score. Table 13 separates presence and assignment gains.

Directional ownership across categories. Figure 7(d) compares the full controller with CrystalBind's noun-excitation-only ablation over three seeds. Its directional ownership objective improves color/shape/texture by $3.78/1.57/2.95$ points, respectively, and every seed has a positive three-category mean. The pooled gain is $+2.77$ points with a 95% interval of $[1.84, 3.73]$.

D.15 Live and fixed noun targets

Target refresh. For starts $s \in \{8, 13\}$, the fixed-target variant reuses the noun maps from the first guided step, while the live-target variant recomputes them along the evolving trajectory. Across three seeds and 900 prompts, live noun targets improve the mean by $+0.94$ points for s=8 (95% CI $[0.32, 1.57]$) and $+1.10$ for s=13 ($[0.58, 1.62]$); the gain is positive for every seed at both starts.

Table 15: Ablation of live versus fixed noun targets (three-seed BLIP-VQA means). Shading identifies live noun targets; paired average gains are computed before rounding.

Target schedule	Color ↑	Shape ↑	Texture ↑	Avg. ↑	Δ Avg.
Fixed noun targets, 8–25	0.594	0.490	0.589	0.558	—
Live noun targets, 8–25	**0.605**	**0.495**	**0.601**	**0.567**	+0.94 pp
Fixed noun targets, 13–25	0.563	0.472	0.572	0.536	—
Live noun targets, 13–25	**0.578**	**0.477**	**0.585**	**0.547**	+1.10 pp

Delaying guidance until noun-map stability. We test a two-stage alternative that first denoises without guidance and then guides against noun maps fixed at a chosen start step. Starting at step 13, immediately after the largest category-mean 0.90 noun-map stability step in Figure 3, the best endpoint is step 35. CrystalBind with live noun targets is higher for each paired seed by $3.02/2.32/2.59$ points, with a 95% interval of $[1.44, 3.83]$ points. Starts at 16 and 20 reduce the three-category averages further. Delaying guidance for a more stable noun-map anchor therefore forgoes an earlier interval in which the assignment remains easier to change.

Table 16: Delayed guidance with fixed noun targets. Entries are changes relative to CrystalBind with live noun targets in the same three-seed run (percentage points); underline marks the strongest delayed schedule.

Target schedule	Δ Color	Δ Shape	Δ Texture	Δ Avg.
Live noun targets, 0–25	0.0	0.0	0.0	0.0
Fixed noun targets, 13–25	−6.6	−1.4	−5.8	−4.6
Fixed noun targets, 13–35	−5.2	+0.3	−3.1	−2.7
Fixed noun targets, 13–49	−6.4	−0.8	−5.0	−4.1
Fixed noun targets, 16–49	−7.1	−2.0	−6.2	−5.1
Fixed noun targets, 20–49	−9.2	−3.4	−6.3	−6.3

D.16 Shared versus category-specific horizons

The prompt-level upper tails for Color, Shape, and Texture are $24.8/21.2/25.1$, so a category-conditioned schedule changes the shared $w = 25$ rule only for Shape. We therefore test a Shape-specific endpoint of $w = 22$ against the shared schedule.

Table 17: Shape-specific endpoint on all 300 prompts and three random seeds. Changes are measured relative to the shared $w = 25$ schedule.

Shape schedule	Δ BLIP-VQA (pp)	Updates	Time (s/prompt)
Shared $w = 25$	0.00	26	3.021
Shape-specific $w = 22$	−0.76	23	2.849

The $w = 22$ Shape schedule uses 11.5% fewer guided updates and 5.7% less time, with a -0.76-point BLIP-VQA change. Across all three categories, the resulting $25/22/25$ schedule saves 3.8% of updates while lowering the matched mean by 0.25 points. These results favor the shared endpoint.

E Scope, Broader Impact, and Reproducibility

Scope. Binding crystallization appears across the tested UNet and DiT backbones. CrystalBind applies the resulting intervention principles using the explicit cross-attention maps available in CLIP-conditioned UNets and focuses on two-object attribute binding. For a new model, sampler, or prompt distribution, the same swap measurement calibrates its intervention horizon. These measurements also guide extensions to multi-object assignments, spatial relations, and counting.

Broader impact. CrystalBind is a training-free inference-time method that improves prompt faithfulness without introducing new model weights or training data. It inherits the constructive uses and misuse risks of the base model.

Reproducibility. The anonymous supplement includes the implementation, primary experiment scripts, benchmark prompts, and dependency specifications.