WHEN REMOVING MORE GUIDANCE HURTS LESS: TEXT AND SPATIAL CONTROL IN DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlled diffusion models combine text prompts with spatial inputs such as edges or depth maps. We show that retaining text during an early interruption of spatial control can worsen final spatial adherence: temporarily removing both inputs preserves adherence better than removing spatial control alone, even though both are restored afterward. Across four configurations combining two control architectures with edge and depth inputs, this ordering holds in 76–81% of scenes and also yields lower mean perceptual distance to the fully conditioned output. An SDXL experiment reveals a conditional sign reversal: text helps adherence when spatial control is present but harms it during the interruption. Adding the effects of separate removals consequently predicts the wrong ordering of the two schedules. The effects depend on timing and on the outcome measured; a small adherence change can accompany a substantial image change. Full conditioning still gives better mean adherence than joint removal, and the SDXL perceptual-distance estimate does not reproduce the advantage. These results show when separate ablations give the wrong conditional choice: deciding which input to retain requires measuring its effect in the presence or absence of the others, at the intended denoising stage.

1 INTRODUCTION

If a controlled diffusion model temporarily loses its spatial input, should it keep its text prompt? Methods that restrict guidance intervals or cache attention make conditioning schedules practical design choices (Kynkäänniemi et al., 2024; Liu et al., 2025a). Retaining the remaining information seems natural, yet removing the prompt too can leave the final image closer to its spatial target, even after both inputs have been restored.

For ControlNet-Canny, removing spatial control during the first ten of thirty denoising steps reduces final boundary F_1 by 0.318. Additionally removing text over the same window reduces that loss to 0.196. Joint removal also leaves the image closer to the fully conditioned output in mean perceptual distance. The additional active signal therefore increases the final disruption caused by the interruption.

The ordering recurs across four SD1.5 configurations combining ControlNet or T2I-Adapter (Zhang et al., 2023; Mou et al., 2024) with Canny or depth control. Joint removal improves adherence over spatial-control removal in 76–81% of scene averages and reduces mean perceptual distance in all four systems. These are comparisons between temporary interruptions, followed by the same conditioning schedule.

The strongest evidence for why the comparison matters comes from SDXL. Removing text lowers F_1 by 0.238 when spatial control is present, but raises it by 0.096 when control is absent. Text's effect reverses with the conditioning context. An additive prediction from the two separate ablations consequently ranks the early interruption schedules in the wrong order.

We compare all four input combinations, matched by scene and initial latent. Three findings organize the paper: **joint removal can preserve more adherence than spatial-control removal alone**; **timing and the measured outcome matter**; and **conditional sign reversals can invalidate the schedule choice suggested by separate ablations**. This exploratory interaction analysis establishes an empirical ordering and its consequences for evaluation, not a general rule to remove guidance.

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

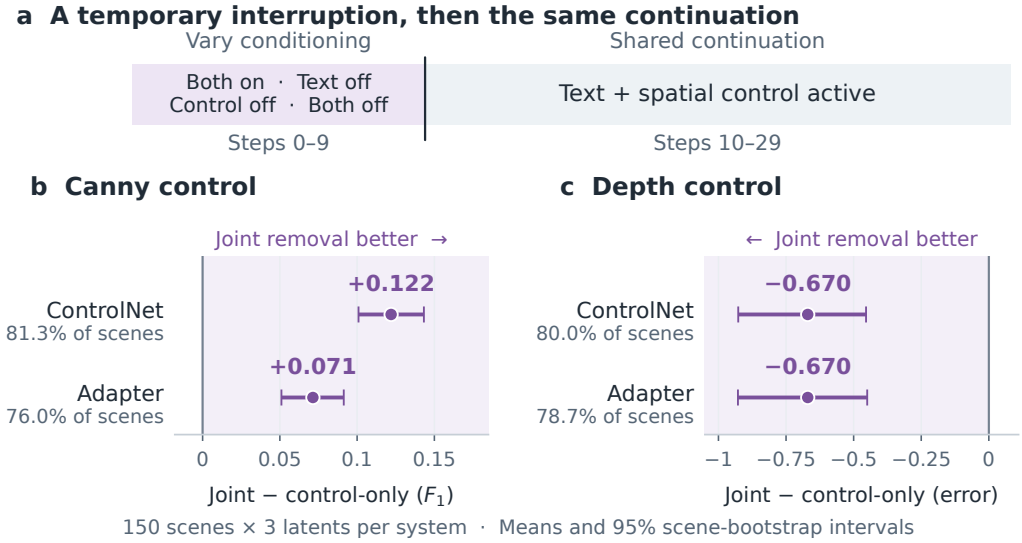


Figure 1: **Removing more conditioning early preserves more final spatial adherence.** (a) The four matched schedules differ only during steps 0–9; all use text and spatial control during steps 10–29. (b,c) Joint removal minus spatial-control removal, measured on the final image. Positive boundary F_1 differences and negative depth-error differences favor joint removal. Points average scenes; bars are pointwise 95% scene-bootstrap intervals. Percentages give scenes with the favorable ordering after averaging three latents. Each system uses the same 150 scenes. Full conditioning still has better mean adherence than joint removal.

2 Related Work

Conditioning over time. eDiff-I uses changing prompt dependence to motivate expert denoisers (Balaji et al., 2022); T-GATE exploits temporal attention behavior for faster sampling (Liu et al., 2025a); and limited-interval guidance improves sample and distribution quality (Kynkäänniemi et al., 2024). Prompt-conditioned interventions test when concepts persist after a prompt change (Görgün et al., 2026). These studies establish temporal specialization. We ask how removing one condition changes the effect of another, measuring spatial adherence and distance to a matched baseline separately.

Interventions and interactions. Consis-GCPO uses modality-specific intervention trajectories for temporal credit assignment (Jin et al., 2026). DuET uses matched and factorial ablations of source-image removal, interval text, and switch ordering to study image-editing gains (Troeshestova et al., 2026). Its source-image and instruction-to-caption switches differ from emptying text and disconnecting edge/depth residual control here. Vision-language binding experiments trace visual information through text tokens (Ge et al., 2026), and InterSHAP measures cross-modal interactions explicitly (Wenderoth et al., 2025). These established intervention tools support our empirical contribution: the final-image ordering under temporary text and spatial-control removal, its temporal structure, and the conditional sign reversal that changes which input to retain.

Text–control competition. SmartControl learns to relax visual control where it conflicts with text (Liu et al., 2025b); SemanticControl suppresses misaligned visual guidance and strengthens text guidance without training (Joung et al., 2025). We examine condition availability followed by restored full conditioning, using captions and controls associated with the same source scene. This setting exposes competition without deliberately constructing conflicting edit requests, although source association does not guarantee perfect compatibility. ControlNet's original training uses 50% empty prompts to encourage learning from control maps (Zhang et al., 2023); this does not establish the behavior of interrupted trajectories or other checkpoints' training conventions.

3 MEASURING CONDITIONAL GUIDANCE EFFECTS

3.1 FOUR CONDITIONS, MATCHED GENERATIONS

Write X^w_{tv} for the final image when text availability $t \in \{0, 1\}$ and spatial-control availability $v \in \{0, 1\}$ are held fixed during denoising window w. Both inputs are active outside w. Thus X_{11} is the fully conditioned baseline, X^w_{01} removes text, X^w_{10} removes spatial control, and X^w_{00} removes both. Each comparison fixes the source scene, prompt, initial latent, and sampling protocol; the evaluation target also stays fixed.

Text removal uses empty conditioning at every text-input site, including ControlNet. Spatial-control removal disconnects the conditioner's residual features at the denoiser's injection sites and skips its execution for those steps. Text guidance remains 7.5; the unconditional branch uses empty text and the same spatial-control state as the conditional branch. Accordingly, emptying the prompt differs from retaining it at guidance scale 1. Appendix B.1 specifies model-specific null conventions.

We compare early $[0, 10)$, middle $[10, 20)$, and late $[20, 30)$ windows in a 30-step trajectory; an early intervention is followed by twenty fully conditioned steps. Equal-duration windows vary denoising stage without varying the number of interrupted steps. Within a window, spatial-control and joint removal differ only in whether text is retained. All outcomes refer to the final decoded image.

3.2 OUTCOMES AND CONDITIONAL CONTRASTS

Spatial adherence is boundary F_1 for Canny control (higher is better) or aligned absolute relative error against a normalized inverse-depth control (lower is better). Perceptual change is LPIPS distance to X_{11} (Zhang et al., 2018). This measures departure from the baseline, not image quality or prompt correctness.

For an adherence score m, define the effect of removing text with control present, A^w_m, and absent, D^w_m:

$$A^w_m = m(X^w_{01}) - m(X_{11}), \qquad D^w_m = m(X^w_{00}) - m(X^w_{10}). \tag{1}$$

The direct joint-removal advantage is D^w_m: positive for F_1, negative for depth error. The factorial interaction compares these two conditional effects:

$$J^w_m = D^w_m - A^w_m = m(X^w_{00}) - m(X^w_{10}) - m(X^w_{01}) + m(X_{11}). \tag{2}$$

If $V^w_m = m(X^w_{10}) - m(X_{11})$ is the spatial-control removal effect, addition predicts a joint change $A^w_m + V^w_m$; the observed joint change is $V^w_m + D^w_m$. Separate ablations rank joint and spatial-control removal incorrectly when A^w_m and D^w_m have opposite signs.

A favorable D alone need not imply a nonzero J. Interaction magnitudes also depend on the score scale (Rohrer & Arslan, 2021). We therefore distinguish the observed ordering, the additive discrepancy, and a conditional sign reversal. These endpoint contrasts do not identify the network or trajectory mechanism producing them. LPIPS is used for a direct paired comparison, not as the principal interaction measure.

3.3 SCENE-LEVEL INFERENCE

We difference outcomes within source–latent pairs, average latents within scenes (three for SD1.5 evaluation, two for the matched comparison), and weight scenes equally. Contrasts are computed before rounding. Pointwise 95% intervals use 50,000 paired scene-bootstrap resamples; all conditions and repetitions of a scene remain together. An additional family of eight early adherence effects—D and J in each main system—uses 99.375% marginal percentile intervals for approximate 95% Bonferroni family coverage.

The sampling protocol, metric panel, and windows were fixed before the main outputs were generated. The interaction interpretation and analyses were developed after inspection. The stated family adjustment does not correct that preceding search. Appendices B.3 and C give the statistical procedures and reproduction checks.

4 EXPERIMENTAL SETUP

The main systems pair an SD1.5 host (Rombach et al., 2022) with ControlNet or SD1.4-labelled T2I-Adapter checkpoints, each using Canny or depth control. Generation uses 30-step deterministic DDIM (Song et al., 2021), text guidance 7.5, visual scale 1.0, an empty negative prompt, and 512^2 output. Reference scenes are from Open Images validation (Kuznetsova et al., 2020), with stored controls and brief captions. Fixed hash ordering selects 100 development and 150 distinct evaluation scenes, each with three initial latents. Evaluation contains $150 \times 3 \times 4 \times 10 = 18{,}000$ images: one baseline and nine interventions per scene–latent–system.

Canny adherence uses a four-pixel boundary tolerance at 512^2. Depth evaluation aligns a DPT prediction (Ranftl et al., 2021) to the stored normalized inverse-depth control by scale and shift. It measures agreement with that supplied target, not physical-depth accuracy. Appendix B.2 gives exact preprocessing and supporting edge measures.

An SDXL-Canny study (Podell et al., 2024) uses 50 development scenes and two latents, yielding 1,000 images across the same ten conditions. Its SD1.5 comparison uses the matching scenes and latent identities. SDXL generates at 1024^2; both systems are evaluated at 512^2. The comparison changes backbone, control checkpoint, output resolution, and native empty-conditioning convention together.

The fully conditioned baselines already vary in adherence. Mean Canny F_1 is 0.695 for ControlNet and 0.581 for Adapter; mean depth error is 0.801 and 0.998, respectively. Preserving more adherence than another interrupted generation therefore does not imply perfect target reproduction. Depth-error means exceed their medians, motivating inspection of the paired distributions alongside averages. Appendix A.3 gives absolute baseline distributions.

5 RESULTS

5.1 JOINT REMOVAL PRESERVES MORE ADHERENCE DURING AN EARLY INTERRUPTION

Removing spatial control early lowers boundary F_1 by 0.318 for ControlNet-Canny and 0.187 for Adapter-Canny. Removing text too reduces these losses to 0.196 and 0.116. For depth, joint removal reduces the error increases from 1.253 to 0.583 and from 1.088 to 0.418 (Table 1). Across the four systems, this averts 38–62% of the mean adherence loss caused by spatial-control removal. The percentage is a ratio of mean losses on each native score scale, not an image-quality improvement.

Table 1: **Joint removal improves both mean adherence and mean baseline similarity in the four SD1.5 systems.** Both schedules remove spatial control during steps 0–9; joint removal also empties text. Adherence loss is relative to full conditioning (positive is worse). Negative paired LPIPS differences favor joint removal; brackets are pointwise 95% scene-bootstrap intervals. The last column counts scene averages improving on both measures (150 scenes, three latents each).

System	Adherence loss control / joint	LPIPS: joint − control (95% CI)	Both improve
ControlNet–Canny	0.318 / 0.196	-0.046 $[-0.060, -0.031]$	59.3%
ControlNet–depth	1.253 / 0.583	-0.041 $[-0.054, -0.028]$	58.7%
Adapter–Canny	0.187 / 0.116	-0.087 $[-0.101, -0.073]$	66.7%
Adapter–depth	1.088 / 0.418	-0.033 $[-0.048, -0.019]$	56.7%

The adherence ordering holds in 76–81% of scene averages (Fig. 1). Mean LPIPS is also lower under joint removal in every system, with all four paired intervals below zero. Better adherence and lower LPIPS coincide in 57–67% of observed scene averages. Figure 2 illustrates the effects; Section 5.4 examines the complete scene distributions and their tradeoffs. These population tendencies do not guarantee improvement for a particular image.

Separate ablations understate the benefit. In the Canny systems, removing text with control present changes F_1 by $+0.019$ and $+0.012$; removing it with control absent changes F_1 by $+0.122$ and $+0.071$. Their differences give $J = +0.103$ [0.082, 0.124] and $+0.059$ [0.035, 0.082]. Depth

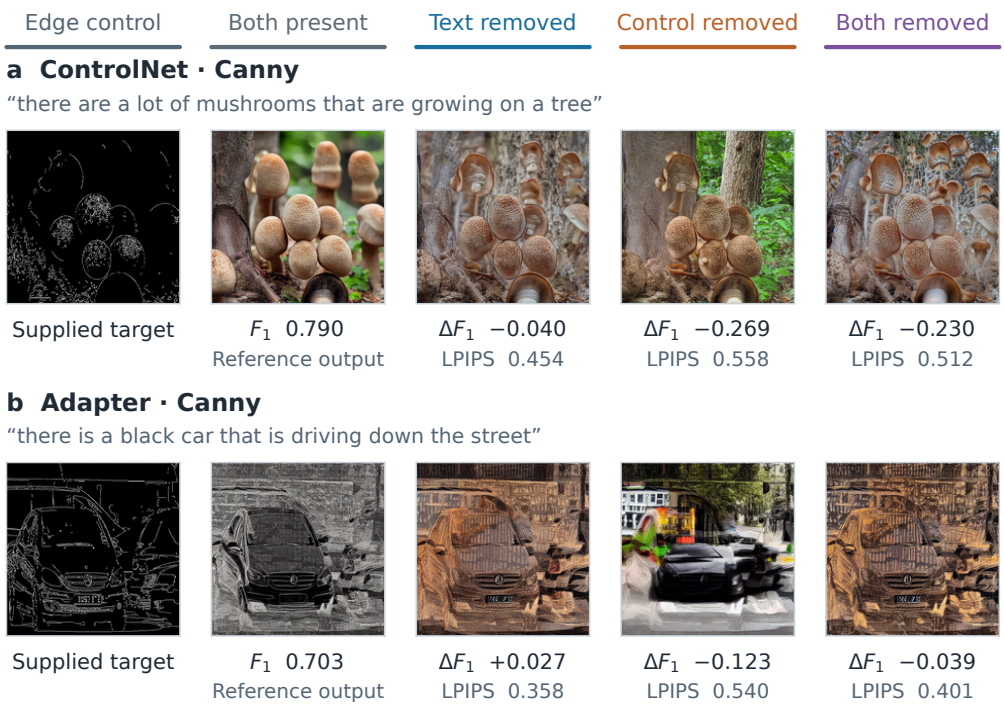


Figure 2: **The same prompt and latent produce different outcomes under early removals.** Each row shows the supplied edge target and four final outputs; both inputs are active again from step 10. The baseline has an absolute F_1 score; interventions report its paired change and LPIPS distance. Each scene is nearest the median favorable scene-level joint-versus-control-removal difference, with latent L0 displayed. This numerical selection is illustrative, not a quality assessment. Appendix A.2 shows counterexamples selected by the corresponding unfavorable rule.

interactions are -0.771 $[-1.051, -0.528]$ and -0.730 $[-1.017, -0.486]$. All eight early D and J intervals retain the favorable sign after the stated family adjustment (Appendix A.1).

5.2 The ordering survives complementary measurement checks

A higher boundary F_1 could reflect a different balance of missing and spurious edges. We therefore examine precision, recall, and symmetric Chamfer distance separately (Table 2). Joint removal improves both precision and recall in the two Canny systems and reduces the average distance between predicted and target edges. Precision improvement weakens an explanation based solely on producing more edges to increase recall. These measures share the same detector, so their agreement supports boundary adherence without independently establishing object-level layout or human preference.

Table 2: **The Canny advantage extends beyond the aggregate F_1 score.** Joint minus spatial-control removal during the early window, with pointwise 95% paired scene-bootstrap intervals (150 scenes, three latents). Higher precision and recall and lower symmetric Chamfer distance are favorable.

Measure	ControlNet–Canny	Adapter–Canny
Precision	$+0.074$ $[+0.059, +0.089]$	$+0.017$ $[+0.003, +0.030]$
Recall	$+0.179$ $[+0.149, +0.209]$	$+0.172$ $[+0.141, +0.203]$
Chamfer (pixels)	-4.40 $[-5.25, -3.55]$	-3.26 $[-4.12, -2.38]$

The result also persists away from the boundaries of the F_1 scale. Requiring all four early scores to lie strictly between 0.1 and 0.9 for every latent retains 97 ControlNet and 114 Adapter scenes. Their

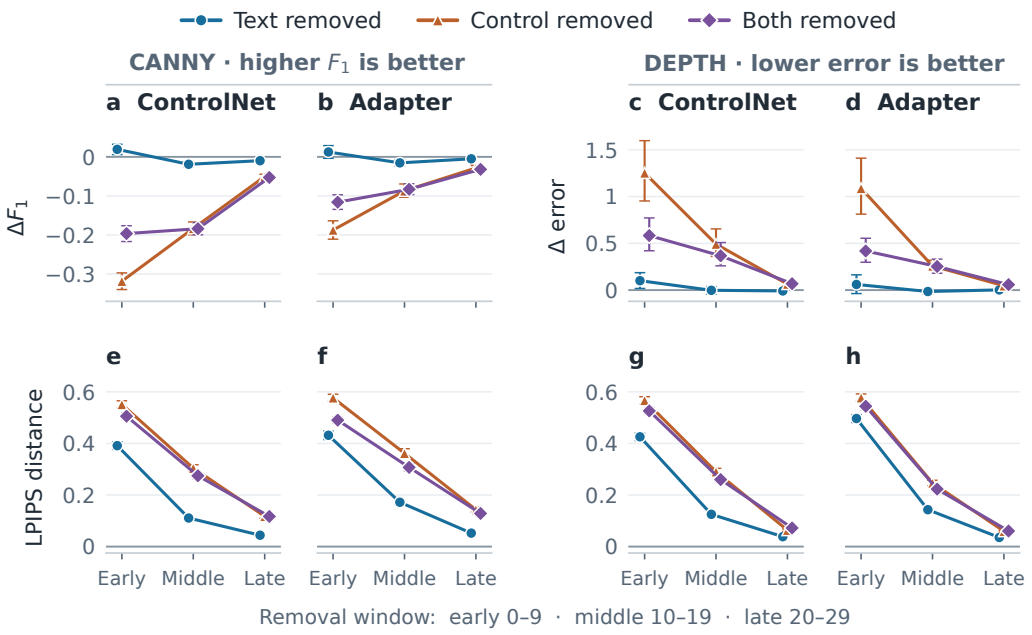


Figure 3: **A small adherence change can accompany a substantial image change, and timing matters.** Top: adherence change from full conditioning; bottom: LPIPS distance to that baseline. Text removal has little mean edge-score effect early but a substantial perceptual effect. Spatial-control and joint removals have smaller effects late. Adherence scales match within control type; all LPIPS panels share a scale. Points average 150 scenes with three latents; bars are pointwise 95% scene-bootstrap intervals. Each window is ten steps, with full conditioning outside it.

direct advantages are $D = +0.112$ and $+0.072$, compared with $+0.122$ and $+0.071$ in the full sets; their interactions remain positive at $J = +0.095$ and $+0.065$. This post hoc restriction selects on intervention outcomes and changes the population. It checks whether the ordering is confined to boundary scores; it does not correct the scale or constitute a confirmatory subgroup test.

Stratifying instead by fully conditioned adherence avoids selecting on intervention outcomes. Within each system, three empirical tertiles contain 50 scenes each. Mean D favors joint removal in all twelve groups, while J has the favorable sign in eleven. Adapter-Canny's lowest baseline tertile has $D = +0.040$ but $J = -0.008$: the direct advantage can persist even where the additive discrepancy changes sign. Together, these descriptive checks separate the recurring ordering from its more variable interaction magnitude. Appendix A.3 gives the full subset and tertile summaries.

5.3 Timing and the measured outcome change the interpretation

Early text removal barely changes mean Canny F_1 in SD1.5 but produces LPIPS distances of 0.391 for ControlNet and 0.432 for Adapter (Fig. 3). A small adherence effect therefore does not mean the input has little influence on the image. Even the edge distribution changes: recall rises while precision falls. In depth, early text removal raises error by 0.101 [0.017, 0.186] for ControlNet and 0.060 [−0.037, 0.163] for Adapter. Text's influence cannot be summarized as appearance alone.

Figure 4a,b compares the two conditional text effects directly. With control present, text removal has small mean effects on SD1.5 adherence; with control absent, its early effect is much larger and favors joint removal. The gap between these curves is J. ControlNet-Canny's J falls from 0.103 early to 0.008 late; ControlNet-depth's moves from -0.771 to 0.013. Paired early-minus-middle and early-minus-late contrasts have pointwise intervals excluding zero in all four systems (Appendix A.1). Noise level, sampler coefficients, and remaining continuation length change together, so these comparisons establish temporal dependence without isolating its cause.

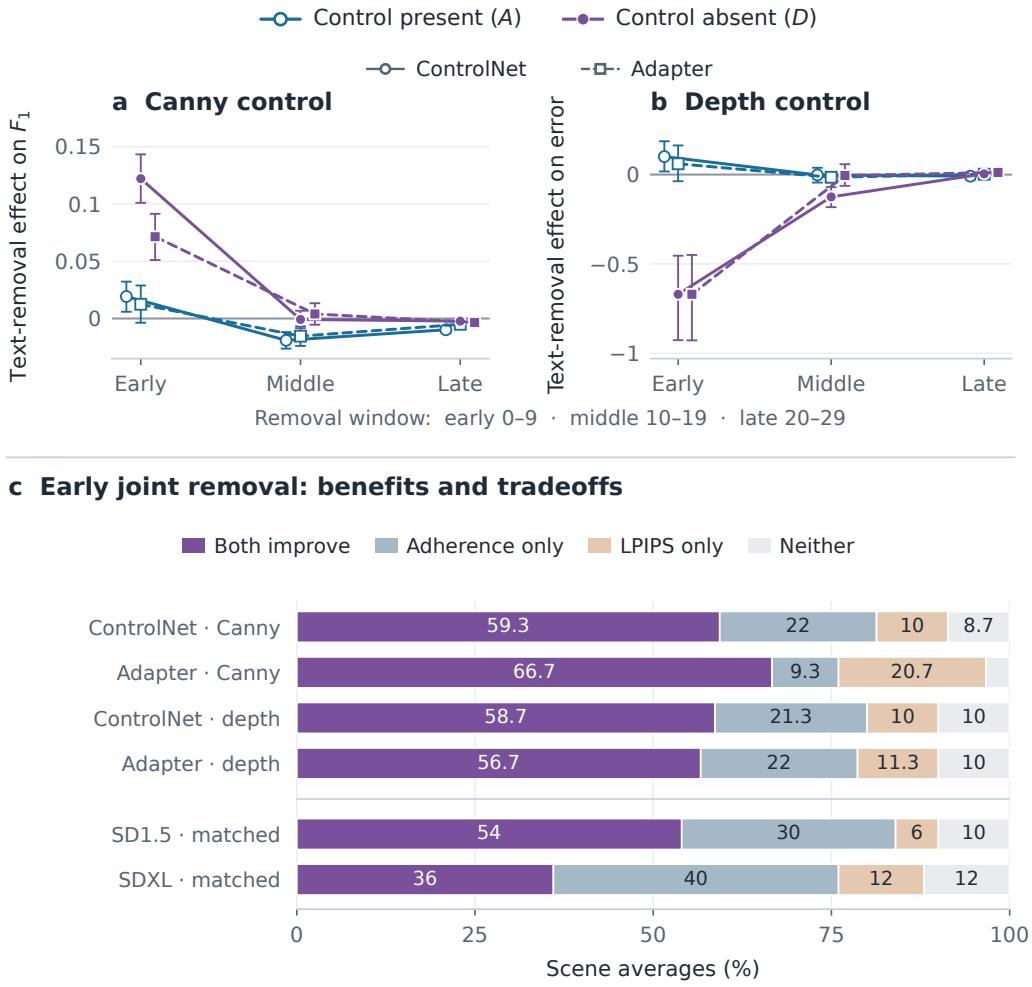


Figure 4: **Text's effect depends on control availability, timing, and the measured outcome.** (a,b) Mean text-removal effects with control present (A, open blue) or absent (D, filled purple); $D - A = J$. Higher F_1 and lower error favor text removal. Bars are pointwise 95% scene-bootstrap intervals (150 scenes, three latents). (c) Early joint versus spatial-control removal, classified by improved adherence and lower LPIPS. Bar segments give percentages; segments below 6% are unlabeled. Table 8 gives all values and intervals. The first four rows use the evaluation scenes; the matched rows use 50 development scenes and two latents.

The direct ordering also changes. In the late window, mean D favors retaining text during control removal in all four systems, by small amounts. ControlNet-Canny, for example, moves from $D = +0.122$ early to -0.002 late. Thus the early advantage is not a time-independent property of jointly removing the inputs. Both the direct choice D and the additive discrepancy J must be evaluated at the intended window.

5.4 Scene distributions distinguish a recurring effect from a removal rule

Figure 4c summarizes the early outcomes without combining adherence and baseline distance into a single score. Each scene is classified after averaging its latents, matching the unit of inference used for the means and intervals. The complete paired scatter in Appendix A.2 retains every observation, including outliers and cases favoring spatial-control removal.

For the four main systems, 57–67% of scene averages improve on both measures, while 9–22% improve adherence at the cost of higher LPIPS. Thus the two favorable population means do not imply that the same scenes always benefit. The lower confidence bounds for joint improvement range from 48.7% to 58.7%, so a population majority is not established in every system. Depth differences also have long favorable tails: larger mean gains coexist with many smaller scene-level changes. The positive ordering in 79–80% of depth scenes shows that its direction is not confined to those large-magnitude cases.

The matched rows show that the perceptual tradeoff changes in SDXL. Both measures improve in 54% of matched SD1.5 scenes and 36% of SDXL scenes; another 30% and 40%, respectively, gain adherence but move farther from the fully conditioned image. Matched SD1.5's mean LPIPS difference is -0.023 [-0.050, 0.005], and SDXL's is $+0.022$ [-0.004, 0.048]. Neither interval excludes zero. Section 5.5 examines the conditional adherence effects in this matched comparison. Appendix A.2 gives quadrant intervals, strict-reversal frequencies, and unfavorable examples selected by the same rule as Figure 2.

5.5 SDXL REVERSES TEXT'S EFFECT AND THE ADDITIVE SCHEDULE RANKING

On the matched 50-scene set, fully conditioned mean F_1 is 0.686 for SD1.5 ControlNet-Canny and 0.633 for SDXL. Removing text early changes these scores by $+0.027$ [-0.002, 0.058] and -0.238 [-0.292, -0.186], respectively. Yet SDXL retains the joint-removal adherence advantage: spatial-control removal costs 0.368 F_1, and joint removal costs 0.272, a paired gain of $+0.096$ [0.047, 0.144] that is favorable in 76% of scene averages (Fig. 4c; Table 8). Text removal thus *harms* adherence with control present and *helps* with control absent (Fig. 5a). This strict sign reversal occurs in 70% [58, 82] of scene averages and 58% [48, 68] of individual latent pairs, with both intervals resampling scenes.

This reversal makes the decision error explicit. If spatial control must be absent early, addition predicts that retaining text is better: its predicted joint change is -0.607, versus -0.368 for spatial-control removal. The observed joint change is -0.272, so removing text too preserves more adherence. The additive discrepancy is $J = +0.335$ [0.271, 0.401]. ControlNet-depth supplies a second mean reversal: addition predicts a joint error increase of 1.353, exceeding the 1.253 spatial-control-only increase, while the observed increase is 0.583. In SD1.5 Canny, addition instead gets the direction right but underestimates the benefit.

The shared continuation is central to this interpretation. Both interrupted schedules receive the same prompt and spatial control for the final twenty steps, without resetting their latent states. Their final mean adherence still differs. Joint removal therefore does not describe an unconditional final generation: the experiment measures how an early change in conditioning persists through subsequent fully conditioned denoising.

The individual-pair reversals have a stronger scale interpretation than a nonzero interaction. A strictly increasing transformation of adherence preserves which of two outputs scores higher, so it preserves the opposite signs of A and D for a fixed source–latent pair. It need not preserve J, or even the signs of contrasts after averaging transformed scores. Reporting both the 58% latent-pair prevalence and the 70% scene-average prevalence makes this distinction explicit; neither is a guarantee for a new scene.

The conditional sign reversal is also time-specific in SDXL. Removing text with control absent changes mean F_1 by -0.090 [-0.115, -0.067] in the middle window and -0.021 [-0.032, -0.011] late. With control present, mean A is negative in both windows (Appendix A.1). Separate ablations then suggest the same direction of choice as the observed conditional comparison, although the additive magnitudes can differ. The incorrect early ranking is thus specific to a conditioning context and denoising stage.

The adherence advantage thus recurs despite text's changing structural role, while the mean perceptual advantage does not recur in SDXL (Fig. 4c). To evaluate a spatial-control interruption, compare retaining and removing text within it: the text ablation with control active answers a different question. Four matched conditions identify both the decision contrast D and the additive discrepancy J. Methods that measure interactions need not make this additive prediction. Measuring adherence and baseline distance separately exposes their tradeoffs.

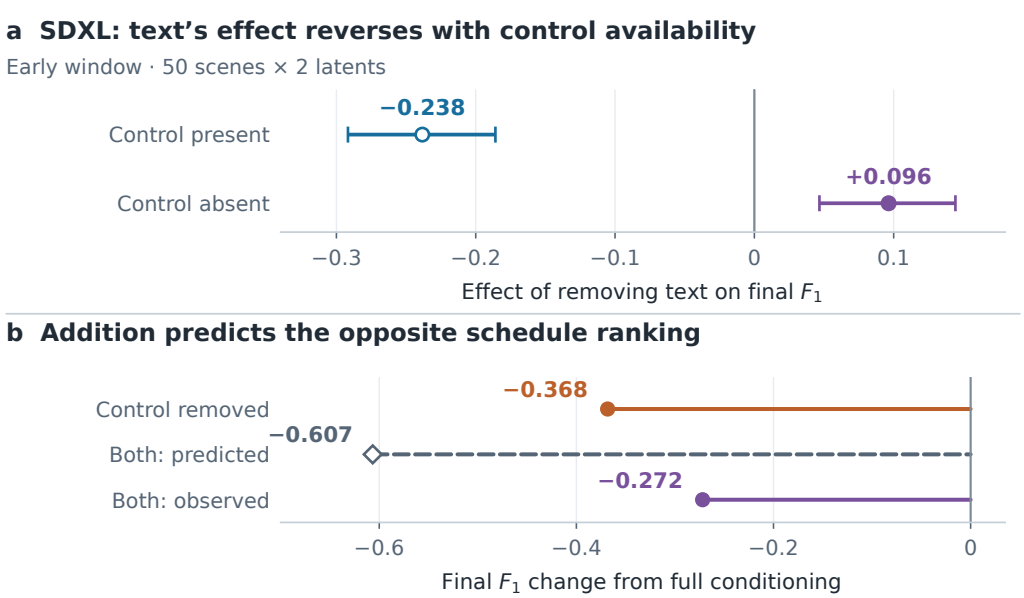


Figure 5: **Separate ablations predict the wrong early guidance choice in SDXL.** (a) Text removal harms final boundary F_1 with control present and helps with control absent. Bars are pointwise 95% scene-bootstrap intervals (50 development scenes, two latents). (b) Adding separate effects predicts joint removal to be worse than control removal; observed means reverse that ordering. The dashed value is a prediction, not a generated condition. Zero denotes full conditioning; panel (a) gives uncertainty in the observed difference between the two removal schedules.

6 Implications and Limits

The conditional ordering does not select an efficient schedule. Late spatial-control removal causes less adherence loss and perceptual change than early joint removal in all four SD1.5 systems. ControlNet-Canny loses 0.050 versus 0.196 F_1, with LPIPS 0.118 versus 0.505. Optimization must compare such alternatives and measure runtime: equal numbers of omitted steps need not save equal cost, especially with cached Adapter features.

Empty text makes the two classifier-free guidance branches coincide, so text removal also removes guidance extrapolation (Appendix A.4). Endpoints cannot separate prompt semantics from guidance strength; retaining the prompt at scale 1 would help distinguish them. Text may steer an interrupted trajectory toward a composition that restored control cannot fully correct, but this mechanism requires trajectory probes. Nor does a local denoiser interaction imply an adherence-improving direction.

Generality is limited by exploratory analysis at one guidance scale, a shared SD1.5 backbone, and a smaller SDXL comparison within the U-Net lineage that changes several choices together. Transfer to transformer denoisers remains untested. Normalized depth error is skewed. The study measures adherence and baseline similarity without human assessment.

7 Conclusion

Removing both text and spatial control early can preserve more final adherence than removing spatial control alone, despite restoring both afterward. Across the four SD1.5 systems, it also reduces mean perceptual departure from full conditioning. In SDXL, the adherence advantage persists and text's effect reverses with control availability, causing separate ablations to predict the wrong schedule ranking; the mean perceptual advantage does not recur. The resulting lesson is precise: to decide which condition to retain, measure its effect in the conditioning context, denoising window, and outcome that the decision concerns.

REPRODUCIBILITY STATEMENT

The supplementary PDF contains the supporting results and algebra in Appendix A, intervention and evaluation procedures in Appendix B, and model identifiers, software, and verification coverage in Appendix C. Two aggregate CSV tables supply the complete pointwise estimates and intervals for removal effects and paired contrasts.

We have prepared a fuller reproducibility package and will make it available with the final submission upon acceptance, or earlier through an anonymous link during rebuttal. It will include individual paired numerical outcomes and source/latent identifiers; unchanged example tiles; CPU analysis and plotting code; derived statistical tables; editable figure and manuscript sources; software and protocol configurations; file and pixel-hash manifests; and verification-coverage documentation. The code regenerates the numerical results, figures, and manuscript without model access. This numerical package is distinct from the approximately 600 GB full project archive. Repeating image synthesis and neural scoring additionally requires the original inputs, models, and generation implementation; those are outside the review supplement.

ETHICS STATEMENT

This study analyzes existing image-generation systems using public-image-derived controls. No human-subject study was conducted. Displayed outputs are synthetic. Understanding conditioning can support legitimate editing and can also facilitate image manipulation. Source images and controls remain subject to their original licensing and attribution requirements.

AI USE STATEMENT

We used OpenAI ChatGPT 6 Pro and Codex Astra-6, Anthropic Claude Code and the Claude web interface, and Google Gemini 3.1 Pro. The authors established the research question, core mathematical formulation, experimental design, figure concepts, and initial manuscript, and made the final scientific and editorial decisions. AI tools provided methodological feedback and simulated peer review, checked algebra and assumptions, suggested corrections and notation, and helped assess interpretations and limitations. Identified deficiencies sometimes motivated additional experiments or checks.

AI assistance supported literature discovery and novelty assessment, citation verification, and drafting and substantial revision of manuscript and supplementary text, including the abstract and captions. Coding assistance, primarily Codex, contributed to implementation, debugging, numerical analysis and validation scripts, algorithmic figure production, reformatting numerical records, and organizing and anonymizing the supplementary package.

Source photographs and supplied annotations came from existing datasets. The diffusion models under study generated the experimental images. AI assistance with figure production operated through plotting and layout code; displayed experimental pixels were preserved rather than regenerated or retouched.

The authors inspected the mathematics, code, results, and figures. Implementations were checked using correctness tests and numerical tests on controlled cases; reported values and figures were checked against experimental outcome records. The authors examined cited sources and checked bibliographic details, supplemented by AI-assisted verification against official sources. The authors reviewed the final manuscript and supplementary materials and take responsibility for their text, claims, code, data provenance, figures, artifacts, and verification statements.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. URL https://arxiv.org/abs/2211.01324.

Chris Ge, Rohit Gandikota, Antonio Torralba, and Tamar Rott Shaham. Vision-language binding in in-context image generation. *arXiv preprint arXiv:2605.24624*, 2026. URL https://arxiv.org/abs/2605.24624.

Ada Görgün, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ABjaSsrYPD.

Qiaoqiao Jin, Dong She, Siming Fu, Hualiang Wang, and Jidong Jiang. Consis-GCPO: Consistency-preserving group causal preference optimization for vision customization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf?id=OswqOlTYR2.

Woosung Joung, Daewon Chae, and Jinkyu Kim. SemanticControl: A training-free approach for handling loosely aligned visual conditions in ControlNet. In *Proceedings of the British Machine Vision Conference*, 2025. URL https://bmvc2025.bmva.org/proceedings/138/.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128:1956–1981, 2020. doi: 10.1007/s11263-020-01316-z.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2404.07724.

Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion via temporal attention decomposition. *Transactions on Machine Learning Research*, 2025a. URL https://openreview.net/forum?id=xXs2GKXPnH.

Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. SmartControl: Enhancing ControlNet for handling rough visual conditions. In *Computer Vision – ECCV 2024*, pp. 1–17, 2025b. doi: 10.1007/978-3-031-73195-2_1.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.01952.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021. URL https://arxiv.org/abs/2103.13413.

Julia M. Rohrer and Ruben C. Arslan. Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, 4(2), 2021. doi: 10.1177/25152459211007368.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Lidia Troeshestova, Alexander Ustyuzhanin, and Sergey Kastryulin. DuET: Dual expert trajectories for diffusion image editing. *arXiv preprint arXiv:2606.13303*, 2026. URL https://arxiv.org/abs/2606.13303v2.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21501–21509, 2025. doi: 10.1609/aaai.v39i20.35452.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. URL https://arxiv.org/abs/2302.05543.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://arxiv.org/abs/1801.03924.