When Text-to-Image Helps Editing: The Effects of Conditioning During Denoising

Anonymous authors
Paper under double-blind review

Abstract

Unified models are trained for both instruction-based image editing and text-to-image (T2I) generation, but standard editing pipelines keep source-image conditioning throughout denoising. We ask whether editing can benefit from T2I, and study how the effects of conditioning vary across edits and denoising stages. In pure editing, source attention declines for some edits over the sampling trajectory. This observation led us to task switching, which lets the model draw on its T2I capabilities. Across three unified editors and four benchmarks, switching to the T2I task for bounded intervals improves edit quality, while mean perceptual preservation remains close to pure editing across all three models. Unified editors therefore benefit from using both conditioning modes they are trained for, and the timing of the switch sets the balance between quality and preservation.

1 Introduction

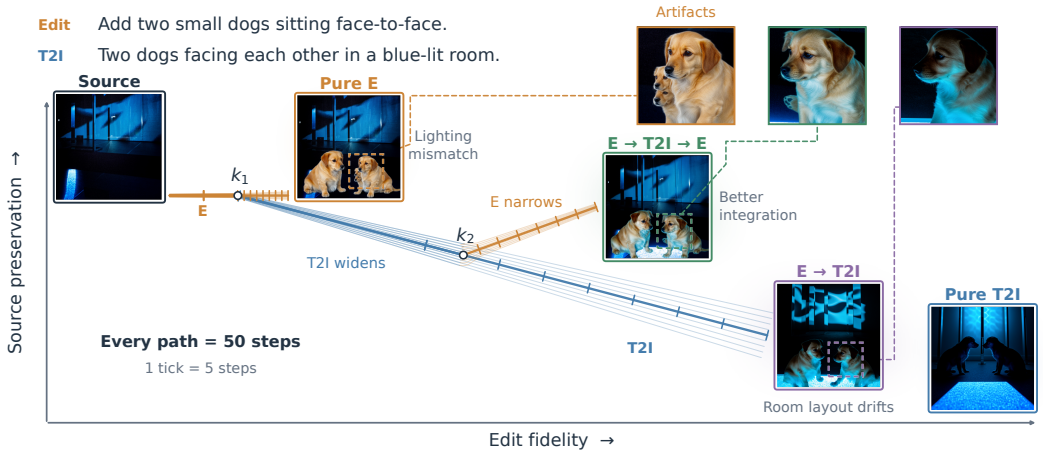

Figure 1: **Switching conditioning helps added objects fit the scene.** Adding two dogs to a blue-lit room with pure editing produces artifacts and mismatched lighting (orange close-up). Switching to caption-only T2I at k_1 improves integration but changes the room. The E $\rightarrow$ T2I $\rightarrow$ E schedule returns to editing at k_2 to retain more of the source.

Unified multimodal models support text-to-image (T2I) generation and instruction-based image editing with the same architecture and weights (Black Forest Labs, 2025; Deng et al., 2025). T2I generation synthesizes an image from a caption describing the desired scene (Rombach et al., 2022), while editing transforms a source image according to a natural-language instruction (Brooks et al., 2023). Within a unified model, the two tasks offer a synergy between synthesizing a coherent target scene and retaining relevant content from an existing image.

Standard pipelines nevertheless use these capabilities separately. In diffusion editors such as FLUX2-Klein (Black Forest Labs, 2025), editing conditions on the source image and instruction throughout denoising, while T2I generation uses a caption alone. Source-image conditioning therefore stays fixed over the course of an edit, and the model's dual capability goes unused within a single trajectory.

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.

Persistent source-image anchoring preserves structure, but it can also suppress intended changes (Hertz et al., 2022; Zhou et al., 2025). When the target scene diverges substantially from the input, editors can under-execute the instruction or integrate the change poorly, with visible seams and mismatched lighting. This raises a question: does every edit benefit from source-image conditioning at every stage of denoising?

To answer it, we examine source-image attention in pure-editing trajectories. With source conditioning available throughout denoising, some edits progressively attend less to it, while localized changes retain strong source attention through the final steps (Section 3.1). This led us to test whether temporarily removing source conditioning can improve edit fidelity and naturalness.

For a fixed interval, we remove source-image conditioning and replace the editing instruction with a target caption generated by a vision-language model (VLM). We then restore the original editing inputs. The model weights and number of denoiser forward passes remain unchanged. Figure 1 shows an example: switching to T2I helps integrate the added objects into the scene, and returning to editing preserves more of the original room.

Contributions.

- We find that the strength of source-image attention depends on the type of edit and varies over the course of denoising. This suggests that for some classes of edits, source conditioning is redundant at certain stages.
- We identify a conditioning regime that consistently improves edit quality across three models and four benchmarks while keeping preservation virtually the same.
- We vary the degrees of freedom of this regime, the conditioning inputs, the interval placement and duration, and the guidance formulation, and characterize how each affects edit quality and preservation.

2 Related Work

Attention and source conditioning. Prompt-to-Prompt reuses cross-attention maps over text tokens from a source generation and limits their injection in time to permit structural changes (Hertz et al., 2022). SaaS identifies source-image dominance over neglected instructions in OmniGen and adaptively scales instruction attention. In its examples, jointly replacing image and text inputs late in denoising has little effect (Zhou et al., 2025). GRAG modulates the relative influence of source-image and instruction tokens to control editing intensity (Zhang et al., 2025). We ask the same question about source conditioning, but instead of rescaling attention we change the conditioning inputs themselves and measure the outcome, using attention maps only as a diagnostic.

Temporal conditioning. eDiff-I observes that reliance on text conditioning decreases during sampling and trains separate denoisers for different noise levels (Balaji et al., 2022). Restricting classifier-free guidance to a limited interval also improves generation (Kynkäänniemi et al., 2024). PCI intervenes on the text prompt during T2I sampling to find when a concept can still be inserted. The resulting temporal profiles differ across concepts and models and carry over to editing (Görgün et al., 2026). We extend this line of work from varying denoisers, guidance, or text prompts across denoising stages to switching between image-conditioned editing and caption-only generation within a unified model.

Combining generation and editing. HiDream-E1's released pipeline provides a direct precedent for an editing-to-generation handoff: it continues the same latent with a target caption, removes source-image conditioning, and finishes the remaining steps with the HiDream-I1 generator weights (HiDream-ai, 2025). For refinement of generated images, RvR argues that pixel-level preservation restricts the space of possible changes, and trains a regeneration model conditioned on target prompts and semantic image tokens without source VAE tokens (Guo et al., 2026). Building on these pipelines, we study how source-image conditioning, the text paired with it, and interval timing affect edit quality and source preservation across unified editors. Conditioning controls and timing sweeps use the models' native editing and T2I modes within one denoising trajectory under unchanged weights.

3 SOURCE CONDITIONING ALONG EDITING TRAJECTORIES

A unified model is trained for two tasks but runs one at a time. Every editing request has a source image, and the pipeline conditions on it at every step regardless of what the request asks for. Editing requests, however, differ widely in nature and in difficulty: enhancing the lighting adjusts an intact scene, whereas replacing a background asks the model to synthesize most of the image from scratch. It would be surprising if one conditioning policy served both equally well. This section examines how an editor actually uses the source image across such edits and across denoising steps.

3.1 SOURCE ATTENTION VARIES ACROSS EDITS AND TIMESTEPS

If the source image matters more for some edits than for others, the most direct place to look is the model's attention to it. We therefore measure how much each output patch attends to the source-image tokens during ordinary editing, and how that changes as denoising proceeds. We examine source attention along 50-step pure-editing trajectories on FLUX2-Klein 4B. For each output latent patch i, we sum attention over source-image tokens and average across heads and transformer blocks, obtaining a_i. At step $k = 9$ (Figure 2), the three edits attend to the source very differently: only on the kept foreground for the background replacement, strongly everywhere for the lighting change, and weakly everywhere for the makeup look. How much the model uses the source depends on what the edit needs from it.

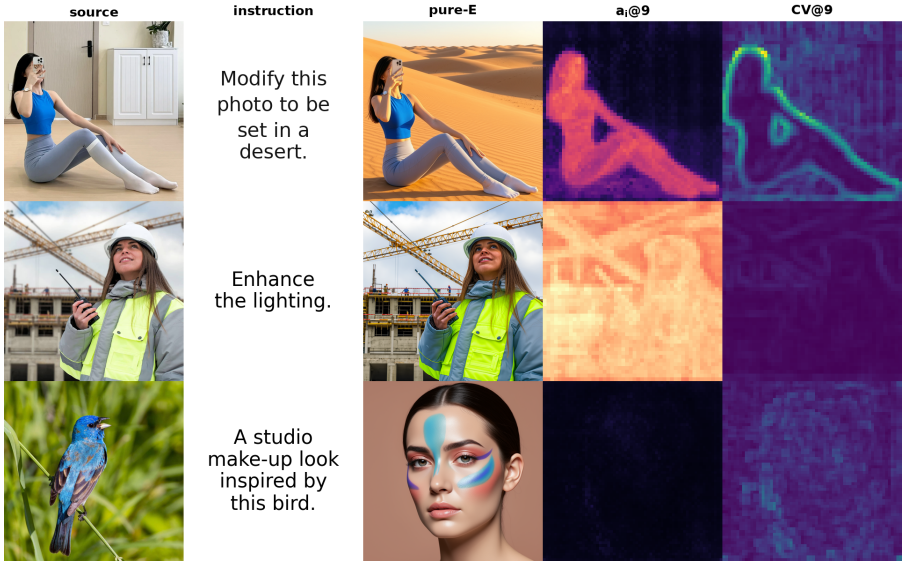

Figure 2: **Source attention varies across edits (FLUX2-Klein 4B, step 9).** Maps average over heads and blocks. a_i@9 shows per-patch source attention; CV@9 shows the coefficient of variation in local 3×3 neighborhoods.

Attention also changes over time (Figure 3a): source-attention mass generally decreases in the examined trajectories, whereas some local edits retain high attention through the final steps. The source image and instruction are identical at every step, so this variation reflects how the model uses fixed inputs. Where attention to the source has faded, conditioning on it constrains the model without contributing much. Since T2I generation is free of that constraint, we hypothesize that it can produce a better scene, and the question becomes where in the trajectory a T2I phase would disturb the edit the least.

3.2 COMPARING EDITING AND T2I PREDICTIONS

The two conditioning modes propose different updates to the same latent. Where they agree, switching between them changes little; where they diverge, a switch pulls the trajectory off the edit. To find the low-disturbance region, we compute both predictions at every step of an ordinary edit. At each latent state of a pure-editing trajectory, we compare predictions conditioned on the source image and instruction (v_{E}) with predictions conditioned on a target caption (v_{T2I}). Their directional

agreement and norm ratio (Figure 3b,c) are highly variable early, comparatively stable in the middle, and diverge again late for local edits. This suggests that the effect of switching depends on its timing, so we test both one-way switches and bounded intervals placed in the stable middle region. The comparison is a diagnostic: it reads both predictions at each step without altering the sampled trajectory, whereas the switching schedules below evaluate one mode per step.

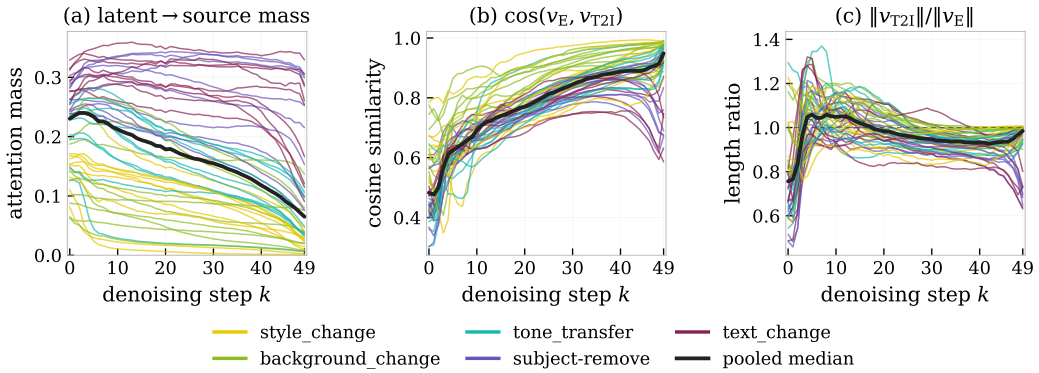

Figure 3: **Conditioning disagreement and source attention along pure editing (FLUX2-Klein 4B).** *(a)* Source-attention mass generally decreases, while local edits retain high attention late in sampling. *(b)* Prediction cosine similarity, initially low (median 0.48). *(c)* T2I/edit prediction norm ratio. Both are evaluated at identical latent states.

3.3 Conditioning interventions and controls

The diagnostics above say where a switch is likely to be safe. To test whether it helps, we intervene directly: for part of the trajectory we hand the latent to the model's T2I mode and let it continue with a caption of the target scene, then return to editing. Edit conditioning E uses a source image x_{src} and instruction p_{edit}; T2I uses a target caption c_{tgt} generated beforehand by a VLM from those inputs. A one-way switch E $\rightarrow$ T2I retains caption-only conditioning from step k onward. A bounded schedule E $\rightarrow$ T2I $\rightarrow$ E uses it on $[k_1, k_2)$ and edit conditioning elsewhere. Both continue denoising the same latent. Only one mode is active per step, so FLUX2-Klein 4B/9B still run the baseline's two denoiser passes per step. The only overhead is one VLM call for the caption (timings in Appendix D).

Table 1: Conditioning inputs.

Mode / control	Image	Text
E	x_{src}	p_{edit}
E*	x_{src}	p^*_{edit}
T2I	$\varnothing$	c_{tgt}
T2I w/ editing prompt	$\varnothing$	p_{edit}
E w/ caption	x_{src}	c_{tgt}

Switching changes both image and text inputs. To separate their effects, we test the original instruction without the source image ($\varnothing$) and the target caption with the source image (Table 1). The E* control supplies a VLM-improved instruction p^*_{edit}, at the cost of the same single VLM call as caption generation. Outside the interval, the controls use ordinary edit conditioning. An additional control uses the target caption throughout and removes the source image only during the interval (Table 6). Together, the schedules (when conditioning changes) and the controls (what it changes to) let us attribute any gain to the removal of the source image, the change of text, or their timing.

4 Effects of conditioning and interval timing

To test whether the effect is a property of unified editors rather than of one model, we evaluate three that differ in size, architecture, and training: FLUX2-Klein 4B and 9B (Black Forest Labs, 2025), two sizes of one diffusion-transformer family, and BAGEL (Deng et al., 2025), a separately developed model. The benchmarks cover complementary aspects of editing quality. GEdit-Bench (Liu et al., 2025) uses real user instructions across task types and scores semantic consistency and perceptual quality separately. ImgEdit (Ye et al., 2025) adds hard categories such as extraction, composition, and actions. MagicBrush (Zhang et al., 2023) supplies a reference caption for every target,

which allows a judge-free alignment score. GIE-Bench (Qian et al., 2025) scores functional correctness and preservation separately, so quality gains can be priced in preservation.

Setup. All models run with unchanged weights and 50 steps ($k = 0, \ldots, 49$); target captions and improved instructions come from Gemini-2.5-pro by default (Comanici et al., 2025) (Qwen3.5 captioners are compared in Appendix D). GEdit is judged by GPT-4.1 (OpenAI, 2025) on a 0–10 scale, ImgEdit by GPT-4o (Hurst et al., 2024) on 1–5, and GIE-Bench correctness by GPT-4o in %; MagicBrush reports CLIP-T cosine similarity to the reference caption, and GIE-Bench preservation is SSIM outside the edit mask and whole-image Unmasked-CLIP similarity to the source. Per GEdit sample, semantic consistency (SC) is the minimum of fidelity and preservation, perceptual quality (PQ) the minimum of naturalness and artifact-free appearance, and overall $O = \sqrt{\mathrm{SC} \cdot \mathrm{PQ}}$; scores are averaged within task types, then across types. The GEdit conditioning matrices and Figures 6 and 7 use Qwen3.5 (Qwen Team, 2026); the extended 4B GIE-Bench sweep uses Gemini-3-Flash (Google DeepMind, 2025).

Variability. Baseline standard deviations come from three seeds with one judge pass each, so they mix generation and judge variability. They describe the baseline only, and the uncertainty of the switching gains is assessed separately. Paired quality tests on 4B use two-sided Wilcoxon signed-rank tests and exact McNemar tests for binary correctness on matched per-edit records. Preservation comparisons are described in Appendix F.

4.1 Task switching improves editing quality

Trading preservation for quality is easy: any editor can be pushed toward the target by loosening its anchor to the source. Improving quality without giving up preservation is the hard part, and this is what a T2I interval achieves. Our sweeps over interval length and position identify $[10, 16)$ as a strong fixed interval: applied to every edit, it raises all six quality scores on all three models (Table 2), while preservation as judged by GPT-4.1 stays within a fraction of a point of pure editing. The pixel-level measures, SSIM and Unmasked-CLIP, do drop, and Section 4.4 examines how much of that drop is perceptible. The interval is not unique: several placements improve GEdit semantic consistency and perceptual quality at once (Figure 4), and which one is best depends on the model and the metric.

We feature $[10, 16)$ because it was consistently strong in the FLUX2-Klein 4B sweeps across all four benchmarks. On 9B and BAGEL we tested a smaller set of intervals, and $[10, 16)$ again ranked among the strongest (Figure 4 and Table 9), so we report it for all three models and use it for the selective variant. The interval was not re-tuned per model or benchmark. Full GIE-Bench, MagicBrush, and GEdit results are reported in Tables 3 to 5.

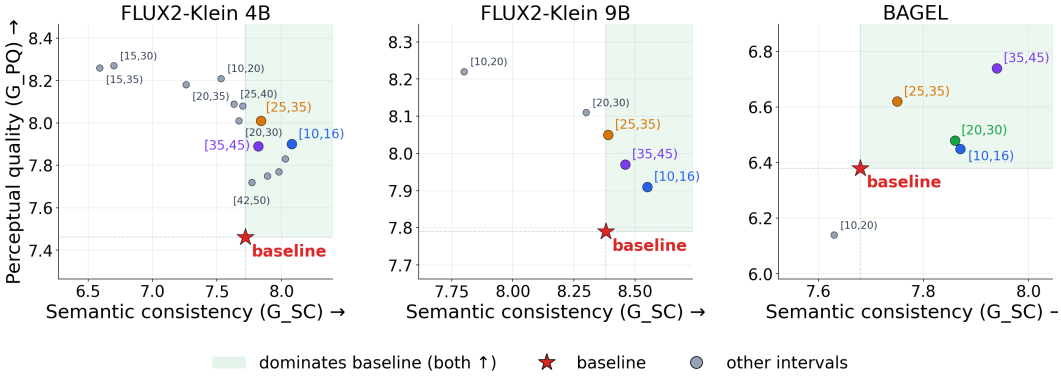

Figure 4: **A fixed T2I interval can improve both GEdit semantic consistency and perceptual quality on each model.** GEdit scores use the GPT-4.1 judge. Points are E → T2I → E schedules, labeled by $[k_1, k_2)$. The shaded region indicates improvement over pure editing on both metrics.

Figure 5 shows what the scores summarize: switching executes edits that pure editing leaves undone or half done, and returning to editing is what keeps the rest of the source intact, as the one-way column shows by contrast (more examples in Appendix A).

Table 2: **A fixed T2I interval improves quality across models and benchmarks.** Fixed and selective switching use the six-step interval $[10, 16)$. The first six scores measure edit quality. Higher is better throughout. Green/red intensity shows the magnitude of increases/decreases from each model's pure-editing baseline. Standard deviations describe baseline variability across three seeds. GPT-4.1 preservation is scored on a 0–10 scale.

		GEdit			ImgEdit	MagicBrush	GIE-Bench		
							Func. Corr.	Preservation	
Model	Sampling	SC	PQ	O	overall	CLIP-T	Overall	SSIM	GPT-4.1
FLUX2-Klein 4B	± std (3 seeds)	0.05	0.03	0.04	0.04	0.001	0.35	0.001	0.05
	Pure editing	7.72	7.46	7.11	3.74	0.317	73.40	0.805	7.65
	Fixed	8.08	7.90	7.67	3.99	0.322	76.46	0.781	7.57
	Selective*	8.01	7.67	7.45	3.88	0.319	77.11	0.786	7.60
FLUX2-Klein 9B	± std (3 seeds)	0.01	0.01	0.01	0.02	0.001	0.72	0.001	0.03
	Pure editing	8.38	7.79	7.74	4.02	0.319	77.94	0.939	8.70
	Fixed	8.55	7.91	7.96	4.23	0.324	81.28	0.899	8.54
	Selective*	8.60	7.89	7.95	4.23	0.323	80.72	0.912	8.61
BAGEL	± std (3 seeds)	0.06	0.09	0.07	0.03	0.001	1.25	0.001	0.02
	Pure editing	7.68	6.38	6.64	3.32	0.320	67.16	0.960	7.90
	Fixed	7.87	6.45	6.81	3.44	0.324	72.38	0.952	7.86
	Selective*	7.86	6.48	6.79	3.40	0.323	71.36	0.954	7.90

Fixed applies the interval to every edit. *Selective uses the attention rule in Section 4.5, with thresholds tuned on GEdit only.

4.2 What accounts for the improvements?

Image and text conditioning. Switching changes two things at once, the image input and the text, and the controls in Table 1 pull them apart; Table 8 summarizes the best scores for each conditioning schedule. Rewording alone does help: a VLM-improved instruction early in the trajectory reaches the highest GEdit overall score among the control winners (Figure 7, left). But the perceptual-quality gain is specific to removing the source image. With the same caption, comparing conditioning with and without the source image (Figure 6) shows that removing the image allows larger gains in fidelity, naturalness, and artifact scores on every model. The text matters in two different ways. Removing the image while keeping only the editing instruction loses semantic consistency, because nothing tells the model what the scene should become. Pairing the caption with the source image also loses semantic consistency, whether the pair is used for the whole trajectory or only outside the interval; here the caption does describe the target, so the likely cause is that the editor was trained to combine the source image with an instruction, not with a description of the result. The improvement therefore comes from switching the task, source-free generation toward a described target, and not from either change alone.

The strongest single competitor is the improved instruction used for the whole trajectory, pure E^*, which costs the same one VLM call and the same two denoiser passes per step as switching (Table 10). It is a strong baseline on the two FLUX models, where it wins GEdit semantic consistency and keeps SSIM higher. But on every model the best T2I interval beats it on GIE-Bench functional correctness, MagicBrush alignment, and GEdit perceptual quality, and the perceptual-quality margin is large. Where E^* wins semantic consistency, it does so through the preservation part of that score; fidelity is matched or exceeded by the T2I interval. On BAGEL, the improved instruction alone does not help at all, while the T2I interval improves every quality score. Better prompts make the editor follow the instruction more faithfully; only switching makes the result look better.

Continuous guidance comparisons. Continuous guidance provides a natural alternative to switching. Reducing image guidance over the same intervals barely changes GEdit overall and lowers ImgEdit, because image-conditioned predictions still drive the update even at zero guidance. Adding a caption-conditioned branch instead blends edit and T2I velocities at every step. At a quarter editing weight, this improves semantic consistency, correctness, and preservation over the pure $E \rightarrow T2I \rightarrow E$ switch, but requires four denoiser passes per step rather than two (Appendix C). Switching is the endpoint of this family at the cost of pure editing and captures most of the benefit.



Figure 5: **Returning to edit conditioning retains more of the source (FLUX2-Klein 4B).** A one-way T2I switch can improve instruction fidelity while changing unrelated content; resuming edit conditioning after the interval limits these changes.

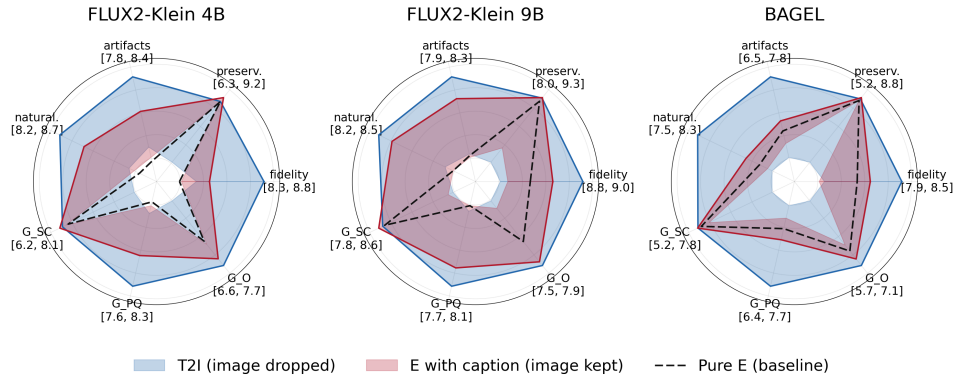

Figure 6: **Removing the source image allows larger quality gains with the same caption (GEdit, Qwen3.5).** Bands span per-metric extrema across switch timings, with and without a return to editing, so one band edge can combine several configurations. Axes normalized independently.

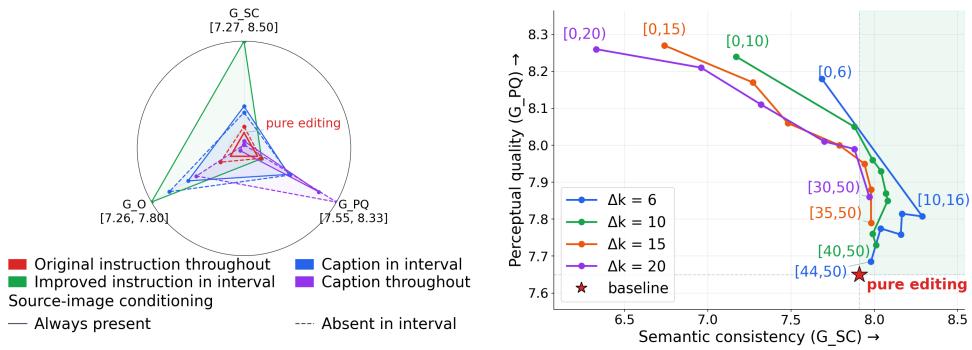

Figure 7: **Conditioning and timing affect edit quality (FLUX2-Klein 4B, GEdit, Qwen3.5).** *Left:* best-G_O configurations per conditioning group over shared placements in Tables 6 and 7; red: pure editing. Colors denote text schedules; solid/dashed lines indicate source kept/removed in the interval. Axes use independent linear scales above and below pure editing. *Right:* T2I sweeps by duration Δk; shading marks gains in both metrics.

4.3 When does switching help?

Interval placement and duration are not independent knobs. Across all four durations, the sweeps trace one curve (Figure 7, right): moving an interval earlier or making it longer buys perceptual quality and spends semantic consistency, and every curve returns to the pure-editing point as the interval slides toward the end of the trajectory. What matters is how much of the trajectory is spent in T2I and how early. The semantic-consistency cost comes from its preservation half; because SC takes the per-sample minimum of fidelity and preservation, a fidelity gain is hidden whenever preservation drops further. Only short intervals in the early-middle of the trajectory land in the region that improves both scores, which is why $[10, 16)$ is the featured setting. Beyond that region the choice is a trade: an editor that values naturalness should switch earlier or longer, one that values consistency later or shorter.

4.4 Quality and source preservation

The quality gains are not free, but the two kinds of preservation metric price them very differently. SSIM outside the edit mask and Unmasked-CLIP drop for every switching configuration, and significantly so (Appendix F). A judge asked only about unintended changes sees far less. We score preservation with GPT-4.1 using GEdit's preservation criterion with the fidelity part removed, so the judge penalizes only changes that the instruction did not ask for (Appendix F). On this measure the fixed interval moves mean preservation by at most 0.16 points on a ten-point scale across the three models, and none of these shifts is significant. Figure 8 (right) puts the two sides together: the featured intervals gain correctness at judged preservation indistinguishable from pure editing, while the early one-way switches that gain the most correctness are the ones that pay for it. Withholding the interval from part of the edits (Section 4.5) recovers a portion of the SSIM loss on every model and brings judged preservation closer still to pure editing.

The two metrics disagree because they measure different things. Figure 8 (left) shows how far the pixel drop overstates the cost: across the 4B sweep, judged preservation forms a plateau at the pure-editing level while SSIM varies, and declines only below it. SSIM is known to fall sharply under small translations, rescaling, or rotations that leave an image perceptually unchanged (Wang & Simoncelli, 2005; Ding et al., 2022), and regenerating a scene from a caption for a few steps is likely to introduce exactly such shifts. SSIM therefore tracks pixel displacement, while the instruction-aware judge is the closer estimate of what a viewer would notice.

4.5 Selective application as an illustration

Section 3.1 showed that source attention reflects how much an edit still depends on the source. Read at the step before the interval, the same signal can decide whether to switch: edits that still depend on the source are more likely to be hurt by removing it. As an illustrative application, we summarize

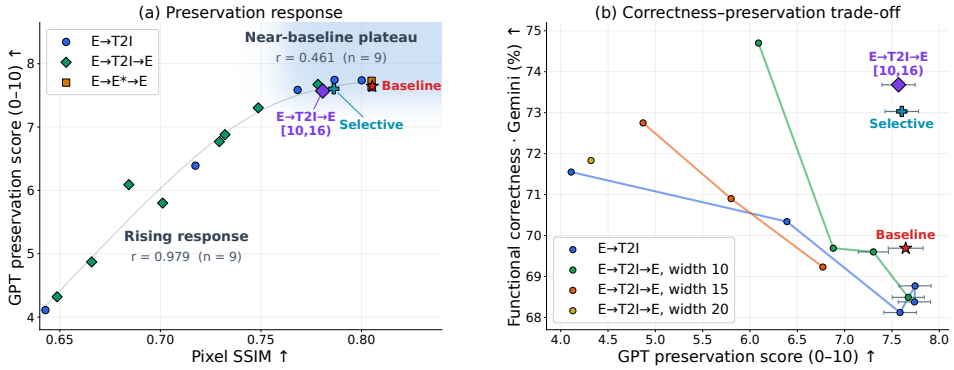

Figure 8: **Preservation and correctness on GIE-Bench (FLUX2-Klein 4B).** *Left:* mean GPT-4.1 preservation versus SSIM outside the edit mask; shading: near-baseline plateau; faint curve: visual guide. Pearson r is computed separately below and above SSIM 0.76. *Right:* Gemini-3-Flash correctness versus mean GPT preservation with 95% source-clustered confidence intervals (Appendix F).

source attention at $k = 9$ by its strength and its spatial unevenness:

$$\pi = Q_{0.75}(\{a_i\}), \qquad \mathrm{cv} = \sigma(\{a_i\})/\mu(\{a_i\}). \tag{1}$$

Edits with strong or uneven source attention stay in pure editing; the rest are switched. Both statistics come from the ordinary edit pass at no extra denoiser cost. Thresholds were tuned on GEdit outcomes for 4B and on GEdit probe distributions for 9B and BAGEL, then shared across all benchmarks (Appendix E).

Selection does what it was meant to do. It keeps the quality gain over pure editing on every score and every model (Table 2), and relative to the fixed interval it moves the operating point toward preservation: SSIM improves significantly on all three models and judged preservation moves closer to pure editing, at the price of somewhat smaller quality gains. The natural question is whether the attention rule is better than choosing edits at random. In a counterfactual replay on GEdit we therefore compare the rule with random routing tuned to spend the same mean preservation (Appendix E). The rule wins on 4B and, with thresholds transferred by matching probe distributions alone, again on 9B: at equal preservation it delivers higher overall, perceptual-quality, and fidelity scores. The attention signal identifies edits that can afford the interval.

5 Discussion and Conclusion

Unified editors are trained for two tasks but use one at a time; this study puts the unused mode to work within an edit. An editor's attention to the source image depends on what the edit needs from it and fades over denoising for edits that rebuild the scene, leaving a stretch of the trajectory where source conditioning constrains without contributing. Switching to the model's own T2I mode on that stretch improves edit quality on every model and benchmark we tested, while an instruction-aware judge finds preservation essentially unchanged. The gain is a property of the switch rather than of the prompt: removing the source image is what unlocks perceptual quality, the caption is what keeps the scene on target, and continuous guidance reproduces the effect only at twice the cost. Placement and duration set where on the quality–preservation curve an editor lands, and the same attention signal that motivated the study can decide per edit whether to switch at all.

We use a simple switching schedule to establish that task switching can improve editing quality. More sophisticated switching and conditioning schemes may offer further gains. Our interval sweeps are coarse, so $[10, 16)$ is one effective setting among those tested. A caption prepared before sampling cannot react to the trajectory, and its errors produce coherent but unintended changes. Preservation is measured by image similarity and a GPT judge, and how these track human judgments is untested.

AI USE STATEMENT

AI assistants helped implement switching conditions, perform statistical tests of preservation, prepare figures, search the literature, polish author-written text, and identify inconsistencies in reported numerical results. The authors verified citations, experimental code, and numerical outputs and take responsibility for the final text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.5 specify the conditioning schedules and attention statistics, while Section 4.1 and Appendix E describe interval selection and model-specific thresholds. The experimental setup and evaluation metrics are described in Section 4, with extended results and guidance formulas in Appendices B and C. Appendix D reports captioner comparisons and timings, and Appendix F describes the preservation judge protocol. Generation code and example notebooks for all three models are available at `https://github.com/pbnc5r6/conditioning-switching`.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Black Forest Labs. FLUX.2: Frontier visual intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. PAIR diffusion: A comprehensive multimodal object-level image editor. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8609–8618. IEEE, 2024.

Google DeepMind. Introducing Gemini 3 Flash: Benchmarks, global availability. `https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/`, 2025.

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*, volume 2026, pp. 66640–66669, 2026.

Jiayi Guo, Linqing Wang, Jiangshan Wang, Yang Yue, Zeyu Liu, Zhiyuan Zhao, Qinglin Lu, Gao Huang, and Chunyu Wang. Refinement via regeneration: Enlarging modification space boosts image refinement in unified multimodal models. *arXiv preprint arXiv:2604.25636*, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

HiDream-ai. HiDream-E1. https://github.com/HiDream-ai/HiDream-E1, 2025. Official editing and refinement pipeline.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025.

Yusu Qian, Jiasen Lu, Tsu-Jui Fu, Xinze Wang, Chen Chen, Yinfei Yang, Wenze Hu, and Zhe Gan. GIE-Bench: Towards grounded evaluation for text-guided image editing. *arXiv preprint arXiv:2505.11493*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.

Zhou Wang and Eero P Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pp. 573–576. IEEE, 2005.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. ImgEdit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

Xuanpu Zhang, Xuesong Niu, Ruidong Chen, Dan Song, Jianhao Zeng, Penghui Du, Haoxiang Cao, Kai Wu, and An-an Liu. Group relative attention guidance for image editing, 2025. URL https://arxiv.org/abs/2510.24657.

Chao Zhou, Tianyi Wei, and Nenghai Yu. Scale your instructions: Enhance the instruction-following fidelity of unified image generation model by self-adaptive attention scaling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15171–15181. IEEE, 2025.

A ADDITIONAL QUALITATIVE EXAMPLES

We extend Figure 5 with 15 GEdit examples, grouped by the observed failure of pure editing: missed edits, incorrect edits, unnatural results, and incomplete edits. The column layout follows Figure 5: edit instruction, source, pure editing, target caption, a one-way E→T2I switch at step 10, and E→T2I→E on $[10, 16)$. All examples use FLUX2-Klein 4B and 50 denoising steps.



Figure 9: **Missed edits (FLUX2-Klein 4B, GEdit).** Pure editing leaves the requested change largely unexecuted. A one-way switch uses T2I from step 10 onward; E→T2I→E returns to editing after $[10, 16)$.



Figure 10: **Incorrect edits (FLUX2-Klein 4B, GEdit).** Pure editing changes the image but does not produce the requested result. Columns as in Figure 9.

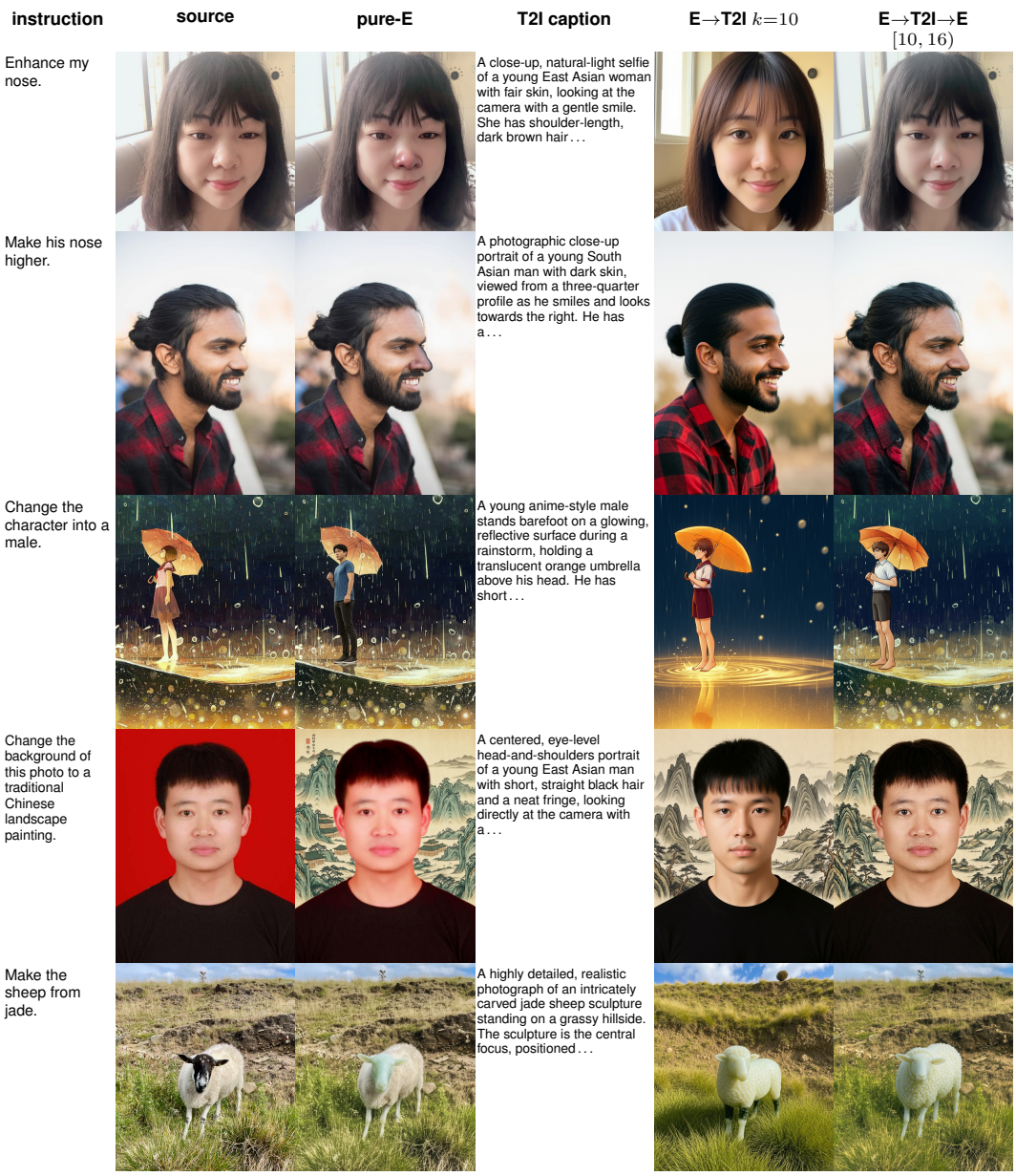

Figure 11: **Unnatural and incomplete edits (FLUX2-Klein 4B, GEdit).** The first four rows illustrate unnatural appearance or artifacts under pure editing; the final row illustrates an incomplete transformation. Columns as in Figure 9.

B EXTENDED EDITING RESULTS

The following tables report benchmark and conditioning sweeps for FLUX2-Klein 4B, together with ImgEdit category breakdowns for all three models. A single switch E $\rightarrow$ T2I retains T2I conditioning from step k onward; E $\rightarrow$ T2I $\rightarrow$ E uses it only on $[k_1, k_2)$. Selective rows apply $[10, 16)$ according to the attention rule, with thresholds in Appendix E. Conditioning controls are defined in Section 3.3.

Reading the tables. Green and red shading indicate improvement and deterioration relative to baseline, scaled separately for each metric within a comparison block. Higher is better except for MSE. GPT-4.1 and Qwen3.5 GEdit results use different judges and are presented separately.

Table 3: **GIE-Bench results (FLUX2-Klein 4B, GPT-4o correctness judge).** Functional correctness by edit category and source-preservation metrics for fixed schedules, the improved-instruction control, and selective application.

		Functional correctness										Preservation				
Schedule	Interval	Add Object	Attribute Change	Color Change	Layout Modif.	Object Replace	Remove Object	Scene / Bkgd Change	Size Change	Textual Edit	Overall	SSIM↑	CLIP↑	Unmasked CLIP↑	PSNR↑	MSE↓
Baseline	—	91.67	88.33	90.76	44.17	93.33	68.33	95.83	20.83	67.50	73.40	0.805	0.97	0.913	20.88	1882.32
E $\rightarrow$ E* $\rightarrow$ E	$[0, 10)$	91.67	93.33	95.80	43.33	93.33	71.43	96.67	27.5	61.67	74.95	0.801	0.970	0.911	20.32	2117.10
E $\rightarrow$ E* $\rightarrow$ E	$[25, 35)$	89.52	88.00	92.86	45.45	94.29	66.34	93.48	23.58	67.65	73.13	0.81	0.97	0.914	20.76	1897.70
E $\rightarrow$ E* $\rightarrow$ E	$[35, 45)$	90.83	89.17	91.60	48.33	93.28	67.50	95.83	25.00	64.17	73.93	0.81	0.97	0.913	20.84	1878.95
E*	$[0, 50)$	90.83	93.33	94.12	41.67	95	70.59	96.67	30.83	70	75.88	0.799	0.970	0.910	20.25	2174.08
E$\rightarrow$T2I	$k = 5$	95.00	95.83	92.44	59.17	94.17	82.50	96.67	38.33	82.50	81.84	0.54	0.93	0.861	9.21	8892.47
E$\rightarrow$T2I	$k = 15$	93.33	90.83	94.12	35.00	94.12	79.17	95.83	20.83	77.50	75.60	0.64	0.95	0.879	9.78	8177.44
E $\rightarrow$ T2I $\rightarrow$ E	$[10, 16)$	94.17	94.17	94.96	45.00	95.00	70.83	97.50	27.50	69.17	76.46	0.78	0.97	0.906	18.45	2569.91
E $\rightarrow$ T2I $\rightarrow$ E	$[10, 20)$	93.33	97.50	93.28	39.17	92.50	80.83	94.17	31.67	68.33	76.74	0.68	0.95	0.889	12.54	5568.90
E $\rightarrow$ T2I $\rightarrow$ E	$[20, 35)$	93.33	89.17	92.44	35.83	95.00	80.00	95.83	23.33	75.00	75.53	0.70	0.96	0.891	10.95	7090.06
E $\rightarrow$ T2I $\rightarrow$ E	$[25, 35)$	91.59	90.65	92.55	40.19	94.79	75.73	95.33	22.22	72.28	75.03	0.75	0.96	0.900	12.93	5518.98
E $\rightarrow$ T2I $\rightarrow$ E	$[35, 45)$	92.65	88.52	91.30	39.71	96.47	77.11	93.75	24.14	64.29	73.13	0.78	0.97	0.905	13.96	5005.39
Selective	$[10, 16)$	94.17	93.33	94.96	46.67	94.17	74.17	96.67	26.67	73.33	77.11	0.79	0.97	0.908	19.00	2370.95

Table 4: **MagicBrush CLIP-T-gen (FLUX2-Klein 4B).** Metric reported in Table 2.

Schedule	Interval	CLIP-T-gen
Baseline	—	0.317
E $\rightarrow$ E* $\rightarrow$ E	$[0, 10)$	0.317
E*	$[0, 50)$	0.317
E$\rightarrow$T2I	$k = 5$	0.322
E$\rightarrow$T2I	$k = 10$	0.324
E$\rightarrow$T2I	$k = 15$	0.325
E$\rightarrow$T2I	$k = 25$	0.321
E$\rightarrow$T2I	$k = 35$	0.318
E$\rightarrow$T2I	$k = 40$	0.318
E$\rightarrow$T2I	$k = 45$	0.317
E $\rightarrow$ T2I $\rightarrow$ E	$[10, 16)$	0.322
E $\rightarrow$ T2I $\rightarrow$ E	$[10, 20)$	0.324
E $\rightarrow$ T2I $\rightarrow$ E	$[15, 30)$	0.325
E $\rightarrow$ T2I $\rightarrow$ E	$[15, 35)$	0.325
E $\rightarrow$ T2I $\rightarrow$ E	$[20, 30)$	0.322
E $\rightarrow$ T2I $\rightarrow$ E	$[20, 35)$	0.323
E $\rightarrow$ T2I $\rightarrow$ E	$[25, 35)$	0.321
E $\rightarrow$ T2I $\rightarrow$ E	$[25, 40)$	0.321
E $\rightarrow$ T2I $\rightarrow$ E	$[35, 45)$	0.318
Selective	$[10, 16)$	0.319

Table 5: **GEdit sweep (FLUX2-Klein 4B, GPT-4.1 judge).** Single-switch and return-to-edit sweeps, including early and full-trajectory improved-instruction controls.

Schedule	Interval	G_SC	G_PQ	G_O
Baseline	—	7.72	7.46	7.11
E $\rightarrow$ E* $\rightarrow$ E	$[0, 10)$	8.16	7.49	7.54
E $\rightarrow$ E* $\rightarrow$ E	$[25, 35)$	7.77	7.47	7.16
E $\rightarrow$ E* $\rightarrow$ E	$[35, 45)$	7.73	7.46	7.09
E*	$[0, 50)$	8.17	7.49	7.54
E$\rightarrow$T2I	$k = 5$	5.31	8.50	6.15
E$\rightarrow$T2I	$k = 10$	5.79	8.45	6.47
E$\rightarrow$T2I	$k = 15$	6.39	8.28	6.72
E$\rightarrow$T2I	$k = 25$	7.26	8.12	7.12
E$\rightarrow$T2I	$k = 35$	7.36	7.89	7.02
E$\rightarrow$T2I	$k = 40$	7.28	7.73	6.86
E$\rightarrow$T2I	$k = 45$	7.18	7.59	6.71
E $\rightarrow$ T2I $\rightarrow$ E	$[10, 16)$	8.08	7.90	7.67
E $\rightarrow$ T2I $\rightarrow$ E	$[10, 20)$	7.53	8.21	7.53
E $\rightarrow$ T2I $\rightarrow$ E	$[15, 30)$	6.70	8.27	6.89
E $\rightarrow$ T2I $\rightarrow$ E	$[15, 35)$	6.59	8.26	6.92
E $\rightarrow$ T2I $\rightarrow$ E	$[20, 30)$	7.67	8.01	7.43
E $\rightarrow$ T2I $\rightarrow$ E	$[20, 35)$	7.26	8.18	7.33
E $\rightarrow$ T2I $\rightarrow$ E	$[25, 35)$	7.84	8.01	7.54
E $\rightarrow$ T2I $\rightarrow$ E	$[25, 40)$	7.63	8.09	7.47
E $\rightarrow$ T2I $\rightarrow$ E	$[35, 45)$	7.82	7.89	7.41
Selective	$[10, 16)$	8.01	7.67	7.45

Table 6: **GEdit interval-conditioning matrix (FLUX2-Klein 4B, Qwen3.5).** Rows specify image conditioning before, during, and after $[k_1, k_2)$; columns specify the text schedule. Here $p = p_{\text{edit}}$, $p^* = p^*_{\text{edit}}$, $c = c_{\text{tgt}}$, and $x = x_{\text{src}}$. All interval variants share the same 12-interval grid. SC/PQ/O denote semantic consistency, perceptual quality, and overall score.

Image schedule		Text (p,p,p)			Text (p,c,p)			Text (c,c,c)			Text (p,p^*,p)		
(x,x,x)	Interval	SC	PQ	O	SC	PQ	O	SC	PQ	O	SC	PQ	O
	$[0,10)$	7.91	7.65	7.33	7.83	8.06	7.46	7.57	8.18	7.28	8.50	7.69	7.80
	$[0,15)$				7.87	8.09	7.49				8.48	7.64	7.76
	$[10,20)$				8.13	7.79	7.53				8.17	7.67	7.50
	$[15,30)$				8.03	7.75	7.45				8.05	7.65	7.43
	$[15,35)$				8.13	7.77	7.54				8.14	7.63	7.47
	$[20,30)$				8.01	7.70	7.41				8.07	7.59	7.38
	$[20,35)$				8.02	7.67	7.40				8.07	7.59	7.38
	$[25,35)$				8.02	7.64	7.38				8.03	7.62	7.37
	$[25,40)$				7.98	7.62	7.37				8.02	7.63	7.36
	$[35,45)$				7.96	7.69	7.39				8.04	7.60	7.37
	$[35,50)$				7.96	7.71	7.39				7.99	7.63	7.34
	$[40,50)$				7.99	7.65	7.39				8.00	7.63	7.37
$(x,\varnothing,x)$	Interval	SC	PQ	O	SC	PQ	O	SC	PQ	O	—		
	$[0,10)$	1.15	7.58	1.36	7.17	8.24	7.26	6.95	8.33	7.20			
	$[0,15)$	0.39	7.74	0.51	6.74	8.27	7.04	6.59	8.31	6.94			
	$[10,20)$	5.21	7.81	5.47	7.88	8.05	7.67	7.42	8.33	7.53			
	$[15,30)$	4.94	7.92	5.46	7.48	8.06	7.45	7.29	8.26	7.39			
	$[15,35)$	4.46	7.97	4.99	7.32	8.11	7.37	7.16	8.26	7.33			
	$[20,30)$	7.36	7.85	7.19	8.04	7.93	7.69	7.51	8.26	7.46			
	$[20,35)$	6.31	7.85	6.51	7.79	8.00	7.57	7.34	8.24	7.36			
	$[25,35)$	7.61	7.75	7.32	8.07	7.87	7.63	7.54	8.22	7.40			
	$[25,40)$	7.13	7.78	7.02	7.94	7.95	7.61	7.38	8.22	7.34			
	$[35,45)$	7.78	7.69	7.33	7.99	7.76	7.50	7.52	8.18	7.32			
	$[35,50)$	7.74	7.70	7.32	7.98	7.79	7.48	7.52	8.22	7.36			
	$[40,50)$	7.86	7.63	7.34	8.01	7.73	7.48	7.58	8.15	7.34			

Green/red denotes improvement/degradation relative to pure E (gray), with shared per-metric scales in Tables 6 to 8. Constant-conditioning results are independent of placement; — marks an unavailable or untested combination.

Table 7: **GEdit single-switch conditioning matrix (FLUX2-Klein 4B, Qwen3.5).** Rows specify image conditioning before/after k; columns specify the text schedule. The reverse switch T2I→E is shown on the right. Notation and colors follow Table 6.

Image schedule		Text (p,p)			Text (p,c)			Text (c,c)			Text (p,p^*)		
(x,x)	k	SC	PQ	O	SC	PQ	O	SC	PQ	O	SC	PQ	O
	0	7.91	7.65	7.33	7.57	8.18	7.28	7.57	8.18	7.28	8.44	7.53	7.72
	5				8.01	7.97	7.50				8.28	7.66	7.61
	10				8.08	7.92	7.58				8.11	7.66	7.49
	15				8.14	7.86	7.55				8.10	7.62	7.44
	20				8.08	7.72	7.45				8.02	7.62	7.35
	25				8.05	7.67	7.41				8.11	7.60	7.42
	30				8.04	7.68	7.39				8.03	7.63	7.37
	35				7.96	7.71	7.39				7.99	7.63	7.34
	40				7.99	7.65	7.39				8.00	7.63	7.37
	45				7.87	7.63	7.29				7.94	7.60	7.32
$(x,\varnothing)$	k	SC	PQ	O	SC	PQ	O	SC	PQ	O	—		
	0	0.11	7.77	0.15	5.53	8.36	6.16	5.53	8.36	6.16			
	5	0.61	8.38	0.82	6.09	8.27	6.53	5.93	8.34	6.52			
	10	2.03	8.18	2.55	6.73	8.23	7.04	6.46	8.30	6.85			
	15	4.12	7.97	4.70	7.12	8.14	7.23	6.88	8.26	7.14			
	20	5.81	7.90	6.17	7.62	8.02	7.51	7.02	8.24	7.17			
	25	6.88	7.80	6.88	7.85	7.96	7.55	7.24	8.21	7.25			
	30	7.51	7.70	7.20	7.97	7.86	7.56	7.44	8.19	7.33			
	35	7.74	7.70	7.32	7.98	7.79	7.48	7.52	8.22	7.36			
	40	7.86	7.63	7.34	8.01	7.73	7.48	7.58	8.15	7.34			
	45	7.95	7.68	7.39	8.02	7.66	7.41	7.58	8.16	7.32			

Reverse switch
T2I → E
Image $(\varnothing, x)$
Text (c, p)

k	SC	PQ	O
0	7.91	7.65	7.33
5	7.79	8.07	7.52
10	7.17	8.24	7.26
15	6.74	8.27	7.04
20	6.33	8.26	6.73
25	6.04	8.29	6.54
30	5.98	8.28	6.48
35	5.90	8.33	6.45
40	5.66	8.35	6.27
45	5.73	8.36	6.35

The per-interval and single-switch results show that:

- Editing with the improved instruction (E^*) raises semantic consistency, with essentially no gain in perceptual quality.
- Removing image conditioning gives the highest perceptual quality in both the interval and single-switch sweeps.
- The text should follow the task: use the target caption while the image is absent, and the editing instruction at the source-conditioned start and end. Keeping the original instruction during the image-free interval, or extending the caption outside it, sacrifices semantic consistency.

Table 8: **GEdit best scores over the displayed placements (FLUX2-Klein 4B, Qwen3.5).** Each row pools interval and single-switch results for one conditioning schedule. Metrics are optimized separately; parentheses give a maximizing interval, so entries in a row need not describe one configuration. Image/text triples specify conditioning before, during, and after the interval. Forward and reverse switches are written as $[k, 50)$ and $[0, k)$, respectively. Ties use the earliest start, then the earliest end.

Schedule	Image	Text	Best SC	Best PQ	Best O
E	(x, x, x)	(p, p, p)	7.91	7.65	7.33
$\mathrm{E} \to \mathrm{E}^* \to \mathrm{E}$	(x, x, x)	(p, p^*, p)	8.50 ($[0, 10)$)	7.69 ($[0, 10)$)	7.80 ($[0, 10)$)
Caption in interval; image retained	(x, x, x)	(p, c, p)	8.14 ($[15, 50)$)	8.18 ($[0, 50)$)	7.58 ($[10, 50)$)
Image off in interval; original instruction	$(x, \varnothing, x)$	(p, p, p)	7.95 ($[45, 50)$)	8.38 ($[5, 50)$)	7.39 ($[45, 50)$)
Caption throughout; image off in interval	$(x, \varnothing, x)$	(c, c, c)	7.58 ($[40, 50)$)	8.36 ($[0, 50)$)	7.53 ($[10, 20)$)
$\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$	$(x, \varnothing, x)$	(p, c, p)	8.07 ($[25, 35)$)	8.36 ($[0, 45)$)	7.69 ($[20, 30)$)

Table 9: **ImgEdit per-category scores (GPT-4o judge).** Category breakdown for FLUX2-Klein 4B, 9B, and BAGEL: baseline, improved-instruction controls, T2I schedules, and selective application where reported.

Schedule	Interval	background	adjust	style	extract	remove	add	replace	compose	action	overall
FLUX2-Klein 4B											
Baseline	—	4.21	3.85	4.80	1.47	2.99	4.40	4.18	3.26	4.70	3.74
$E \to E^* \to E$	$[0, 10)$	4.11	3.63	4.80	3.74	4.33	4.49	4.16	3.68	4.73	4.18
$E \to E^* \to E$	$[25, 35)$	4.08	3.83	4.74	1.38	2.98	4.36	4.14	3.19	4.91	3.69
$E \to E^* \to E$	$[35, 45)$	4.15	3.81	4.80	1.55	3.01	4.40	4.20	3.26	4.80	3.73
E^*	$[0, 50)$	4.08	3.73	4.75	3.64	4.28	4.42	4.19	3.55	4.73	4.15
T2I→E	$k = 25$	4.40	3.04	4.70	3.90	3.78	4.45	3.55	2.01	4.36	3.94
E→T2I	$k = 5$	4.43	3.18	4.76	3.58	4.21	4.47	3.70	2.14	3.65	3.96
E→T2I	$k = 10$	4.46	3.66	4.83	2.58	3.58	4.58	3.80	1.99	3.93	3.87
E→T2I	$k = 15$	4.49	4.06	4.88	2.00	3.40	4.64	4.04	2.91	4.23	3.93
E→T2I	$k = 25$	4.46	4.62	4.93	1.94	3.30	4.65	4.30	3.70	4.69	4.08
E→T2I	$k = 35$	4.28	4.43	4.92	1.63	3.20	4.53	4.28	3.68	4.75	3.93
E→T2I	$k = 40$	4.31	4.19	4.91	1.61	3.15	4.51	4.27	3.51	4.83	3.88
E→T2I	$k = 45$	4.28	4.06	4.87	1.48	3.15	4.44	4.21	3.36	4.89	3.82
$E \to T2I \to E$	$[10, 16)$	4.51	4.12	4.89	1.88	3.41	4.57	4.24	3.74	4.86	3.99
$E \to T2I \to E$	$[10, 20)$	4.50	4.31	4.95	2.37	3.79	4.71	4.46	3.48	4.77	4.16
$E \to T2I \to E$	$[15, 30)$	4.57	4.62	4.93	2.00	3.59	4.70	4.14	3.48	4.73	4.12
$E \to T2I \to E$	$[15, 35)$	4.56	4.40	4.91	2.13	3.42	4.65	4.11	3.35	4.69	4.06
$E \to T2I \to E$	$[20, 30)$	4.44	4.57	4.90	1.91	3.34	4.62	4.30	3.74	4.84	4.07
$E \to T2I \to E$	$[20, 35)$	4.47	4.49	4.93	1.90	3.27	4.66	4.22	3.77	4.75	4.04
$E \to T2I \to E$	$[25, 35)$	4.43	4.54	4.92	1.86	3.32	4.64	4.27	3.75	4.81	4.04
$E \to T2I \to E$	$[25, 40)$	4.42	4.68	4.94	1.74	3.36	4.69	4.28	3.65	4.73	4.07
$E \to T2I \to E$	$[35, 45)$	4.33	4.35	4.93	1.66	3.21	4.50	4.30	3.43	4.85	3.92
Selective	$[10, 16)$	4.36	3.99	4.89	1.73	3.26	4.40	4.26	3.68	4.81	3.88
FLUX2-Klein 9B											
Baseline	—	4.22	4.17	4.92	1.94	4.14	4.61	4.47	3.29	4.81	4.02
$E \to E^* \to E$	$[25, 35)$	4.29	4.21	4.90	2.05	4.12	4.55	4.51	3.58	4.81	4.08
$E \to E^* \to E$	$[35, 45)$	4.23	4.24	4.90	2.09	4.05	4.57	4.48	3.32	4.74	4.07
E^*	$[0, 50)$	4.09	4.27	4.88	3.63	4.55	4.57	4.46	3.72	4.81	4.34
$E \to T2I \to E$	$[10, 16)$	4.34	4.29	4.93	2.50	4.36	4.62	4.58	3.61	4.92	4.23
$E \to T2I \to E$	$[10, 20)$	4.50	4.39	4.87	2.98	4.50	4.64	4.36	3.46	4.82	4.33
$E \to T2I \to E$	$[20, 30)$	4.39	4.56	4.96	2.36	4.26	4.66	4.49	3.74	4.88	4.26
$E \to T2I \to E$	$[25, 35)$	4.42	4.52	4.95	2.33	4.27	4.65	4.43	3.80	4.87	4.24
$E \to T2I \to E$	$[35, 45)$	4.34	4.42	4.94	2.23	4.17	4.64	4.40	3.93	4.78	4.18
BAGEL											
Baseline	—	3.38	3.51	4.43	1.38	3.00	3.93	3.64	2.53	4.22	3.32
$E \to E^* \to E$	$[25, 35)$	3.47	3.54	4.43	1.35	3.00	3.92	3.63	2.77	4.24	3.33
$E \to E^* \to E$	$[35, 45)$	3.43	3.43	4.49	1.42	3.01	3.85	3.61	2.59	4.19	3.30
E^*	$[0, 50)$	3.44	3.37	4.41	2.10	3.25	3.89	3.69	2.67	4.36	3.44
$E \to T2I \to E$	$[10, 16)$	3.53	3.45	4.65	1.54	3.21	4.05	3.77	2.75	4.28	3.44
$E \to T2I \to E$	$[10, 20)$	3.55	2.75	4.67	1.63	3.29	3.97	3.69	2.35	4.20	3.35
$E \to T2I \to E$	$[20, 30)$	3.51	3.36	4.68	1.51	3.29	4.10	3.84	2.63	4.31	3.46
$E \to T2I \to E$	$[25, 35)$	3.59	3.46	4.70	1.50	3.38	4.19	3.90	2.49	4.41	3.51
$E \to T2I \to E$	$[35, 45)$	3.62	3.73	4.60	1.39	3.70	4.18	3.95	2.78	4.54	3.57

Table 10: **Improved-instruction controls and best observed T2I intervals across three models.** The evaluation protocols match Table 2: GPT-4.1 for GEdit and GPT-4o for ImgEdit and GIE correctness. The best E→T2I→E score is selected separately for each column from the available interval sweep; the row below reports one winning interval. These maxima need not share one configuration, and sweep coverage varies by model and benchmark. Higher is better; green/red intensity shows changes from each model's pure-editing baseline, using the per-metric scales of Table 2. The ± rows are three-seed baseline standard deviations, as in Table 2.

	GEdit			ImgEdit	MagicBrush	GIE-Bench		
Sampling	SC	PQ	O	overall	CLIP-T	FC	SSIM	U-CLIP
FLUX2-Klein 4B								
± std (3 seeds)	0.05	0.03	0.04	0.04	0.001	0.35	0.001	0.002
Pure E	7.72	7.46	7.11	3.74	0.317	73.40	0.805	0.913
E^* →E (10)	8.16	7.49	7.54	4.18	0.317	74.95	0.801	0.911
E^*	8.17	7.49	7.54	4.15	0.317	75.88	0.799	0.910
E→T2I→E (best)	8.08	8.27	7.67	4.16	0.325	76.74	0.781	0.906
Winning interval	[10, 16)	[15, 30)	[10, 16)	[10, 20)	[15, 30)	[10, 20)	[10, 16)	[10, 16)
FLUX2-Klein 9B								
± std (3 seeds)	0.01	0.01	0.01	0.02	0.001	0.72	0.001	0.0003
Pure E	8.38	7.79	7.74	4.02	0.319	77.94	0.939	0.928
E^*	8.79	7.80	8.10	4.34	0.319	80.43	0.937	0.926
E→T2I→E (best)	8.55	8.22	7.96	4.33	0.326	81.28	0.899	0.922
Winning interval	[10, 16)	[10, 20)	[10, 16)	[10, 20)	[10, 20)	[10, 16)	[10, 16)	[10, 16)
BAGEL								
± std (3 seeds)	0.06	0.09	0.07	0.03	0.001	1.25	0.001	0.001
Pure E	7.68	6.38	6.64	3.32	0.320	67.16	0.960	0.928
E^*	7.53	6.47	6.62	3.44	0.321	66.42	0.959	0.922
E→T2I→E (best)	7.94	6.74	6.99	3.57	0.324	72.38	0.952	0.924
Winning interval	[35, 45)	[35, 45)	[35, 45)	[35, 45)	[10, 16)	[10, 16)	[10, 16)	[10, 16)

C CONTINUOUS GUIDANCE COMPARISONS

On FLUX2-Klein 4B, we test whether lowering image guidance reproduces the effect of a T2I interval, and introduce a multimodal guidance formulation that continuously combines editing with target-caption-conditioned generation. Let U, T, I, and E denote denoiser predictions conditioned on neither image nor text, the target caption alone, the source image alone, and the source image with the edit instruction, respectively.

Image guidance. Following InstructPix2Pix (Brooks et al., 2023), we use

$$U + s_i(I - U) + s_t(E - I), \tag{2}$$

where s_i and s_t control image and text guidance. We hold s_t fixed and reduce $s_i = \alpha$ from the editing baseline $\alpha = 1$, either on a limited interval or throughout sampling. Even at $\alpha = 0$ the image-conditioned predictions E and I still enter the update, so image guidance and caption-only T2I stay distinct controls.

Multimodal guidance. We extend multimodal CFG (Goel et al., 2024) to combine source-conditioned editing with a separate target-caption branch:

$$U + w_t(T - U) + w_i(I - U) + w_{ti}(E - T - I + U). \tag{3}$$

Here E uses the edit instruction, whereas T uses the target caption. Setting $w_t = s_t$, $w_i = \alpha$, and $w_{ti} = \alpha s_t$ gives the interpolation $\alpha v_{\mathrm{E}} + (1 - \alpha)v_{\mathrm{T2I}}$, where $v_{\mathrm{E}} = s_t E + (1 - s_t)I$ and $v_{\mathrm{T2I}} = s_t T + (1 - s_t)U$. We apply multimodal guidance on a bounded interval and retain edit conditioning elsewhere: $\alpha = 0$ recovers $\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$ and $\alpha = 1$ recovers pure editing.

Figure 12 illustrates the two guidance paths in the same weight plane.

Table 12 compares these controls on $[10, 20)$ and $[25, 35)$. At $[10, 20)$, multimodal guidance with $\alpha = 0.25$ improves GEdit semantic consistency and overall score, GIE-Bench functional correctness, and preservation over the pure T2I interval, while reducing perceptual quality, ImgEdit, and CLIP-T scores. At $[25, 35)$, the pure T2I endpoint has the highest GEdit overall, ImgEdit, CLIP-T, and functional-correctness scores among the multimodal-guidance settings, while intermediate weights retain more source structure. Lowering image guidance on either interval changes GEdit overall by at most $+0.06$ and reduces ImgEdit. Applied over the full trajectory, it lowers both further.

For FLUX2-Klein 4B/9B, the denoiser forward counts in Table 11 follow directly from the required predictions: two at either endpoint, three for image guidance with $\alpha < 1$, and four for multimodal guidance with $0 < \alpha < 1$.

Table 11: **Forward passes within the denoising trajectory (FLUX2-Klein 4B/9B).** Counts are per sampling step, counting batched conditioning branches separately. Pure editing and the pure T2I endpoint each require two passes.

Conditioning	Forwards / step	Required predictions
Pure E (baseline)	**2**	E, I
Pure T2I	**2**	T, U
Image guidance ($\alpha < 1$)	3	E, I, U
Multimodal guidance ($0 < \alpha < 1$)	4	E, T, I, U

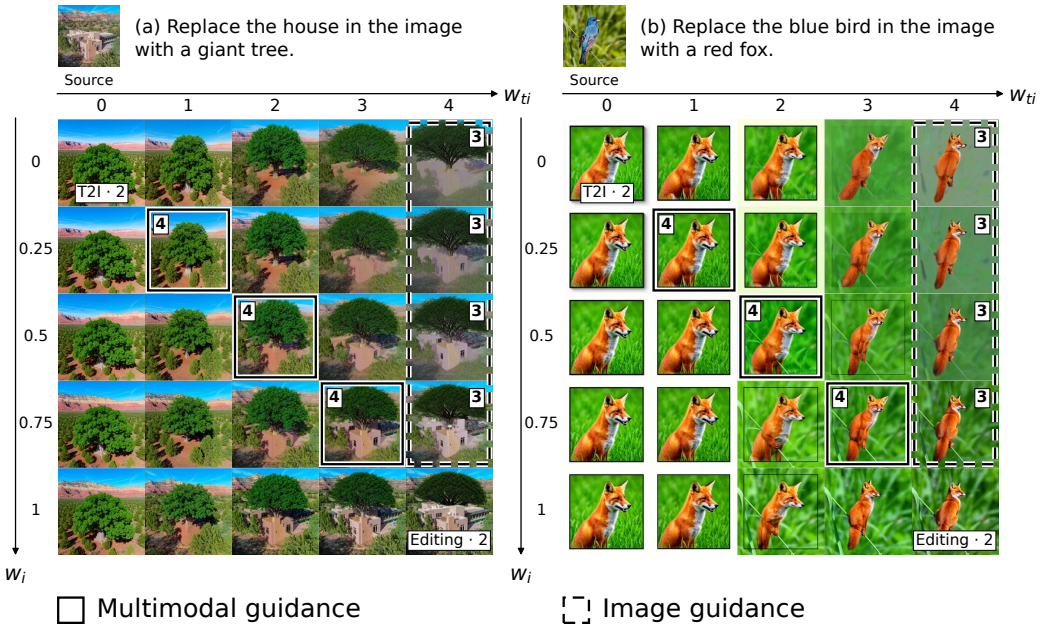

Figure 12: **Continuous guidance in the factorial weight plane (FLUX2-Klein 4B).** We fix $w_t = 4$ and apply each weight pair throughout 50 denoising steps. Multimodal guidance follows the diagonal $(w_i, w_{ti}) = (\alpha, 4\alpha)$, interpolating from caption-only T2I at $\alpha = 0$ to editing at $\alpha = 1$; solid outlines mark $0 < \alpha < 1$. Image guidance follows the rightmost column, $(w_i, w_{ti}) = (\alpha, 4)$; dashed outlines mark $0 \leq \alpha < 1$. Even at $\alpha = 0$, image guidance uses the image-conditioned predictions E and I. Numbers indicate forward passes per denoising step: four for intermediate multimodal guidance, three for image guidance with $\alpha < 1$, and two for the T2I and editing endpoints.

Table 12: **Image guidance and multimodal guidance (FLUX2-Klein 4B).** GEdit uses GPT-4.1; ImgEdit and GIE correctness use GPT-4o. $\alpha = 1$ is editing. For multimodal guidance, $\alpha = 0$ is the T2I endpoint; for image guidance, it sets the image guidance scale to zero. Shading uses one baseline-centered scale per metric across all blocks.

		GEdit			ImgEdit	MagicBrush	GIE-Bench		
α	Fwd/ step	SC	PQ	O	overall	CLIP-T	FC	SSIM	U-CLIP
Image guidance									
Interval $[10, 20)$									
0.0	3	7.69	7.27	7.10	3.66	0.319	74.84	0.782	0.891
0.25	3	7.74	7.36	7.15	3.71	0.319	74.98	0.790	0.898
0.5	3	7.70	7.43	7.15	3.73	0.319	73.35	0.797	0.904
0.75	3	7.79	7.48	7.17	3.72	0.318	73.88	0.802	0.910
1.0 (Editing)	2	7.72	7.46	7.11	3.74	0.317	73.40	0.805	0.913
Interval $[25, 35)$									
0.0	3	7.69	7.38	7.13	3.73	0.317	73.40	0.768	0.890
0.25	3	7.72	7.32	7.10	3.73	0.317	73.57	0.785	0.896
0.5	3	7.77	7.44	7.17	3.73	0.318	74.14	0.799	0.905
0.75	3	7.78	7.43	7.13	3.73	0.317	71.90	0.803	0.911
1.0 (Editing)	2	7.72	7.46	7.11	3.74	0.317	73.40	0.805	0.913
Full trajectory $[0, 50)$									
0.5	3	7.35	6.80	6.71	3.55	0.319	75.21	0.755	0.860
0.0	3	6.12	5.42	5.45	3.07	0.311	72.76	0.700	0.803
Multimodal guidance									
Interval $[10, 20)$									
0.0 (T2I)	2	7.53	8.21	7.53	4.16	0.324	76.74	0.684	0.889
0.25	4	8.10	7.92	7.73	3.97	0.323	77.48	0.766	0.903
0.5	4	7.99	7.79	7.51	3.86	0.321	74.49	0.798	0.910
0.75	4	7.88	7.61	7.32	3.78	0.318	73.40	0.804	0.913
1.0 (Editing)	2	7.72	7.46	7.11	3.74	0.317	73.40	0.805	0.913
Interval $[25, 35)$									
0.0 (T2I)	2	7.84	8.01	7.54	4.04	0.321	75.03	0.749	0.900
0.25	4	7.89	7.87	7.46	3.96	0.319	74.24	0.773	0.903
0.5	4	7.90	7.68	7.36	3.83	0.318	73.86	0.800	0.910
0.75	4	7.71	7.57	7.14	3.76	0.317	73.96	0.804	0.913
1.0 (Editing)	2	7.72	7.46	7.11	3.74	0.317	73.40	0.805	0.913

D END-TO-END TIMING AND CAPTIONER CHOICE

Switching reduces image-generation time because T2I omits source-image tokens while keeping the same denoiser pass count as editing on the FLUX models (Tables 11 and 13). End-to-end latency depends on the captioner: with the Qwen captioners (Qwen Team, 2026), total time stays close to pure editing, whereas the Gemini API call dominates the measured runtime.

The quality gains do not depend on Gemini captions (Table 14). At the same switching interval, the evaluated Qwen alternatives retain improvements over pure editing in GEdit overall, ImgEdit, MagicBrush alignment, and GIE-Bench correctness, with a similar preservation cost. A faster and cheaper captioner can therefore retain the benefit of task switching while avoiding the large captioning overhead.

Table 13: **End-to-end time per edit, including caption generation.** The $\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$ schedule uses the interval $[10, 20)$. Timings use FLUX2-Klein 4B on $1{,}000$ samples; all times are means in seconds. For Gemini, the VLM-call time is the API-call latency. In addition to image generation and the VLM call, total time includes VAE encoding and decoding and image resizing.

Sampling	Caption model	Image generation	VLM call	Total
Pure editing	—	5.325	—	5.475
$\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$	Gemini-2.5-pro	4.900	26.760	31.810
$\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$	Qwen3.5-9B	4.900	0.460	5.510
$\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$	Qwen3.5-27B	4.900	1.143	6.193
$\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$	Qwen3.5-35B-A3B	4.900	0.239	5.289

Table 14: **Captioner choice for $\mathrm{E} \to \mathrm{T2I} \to \mathrm{E}$ on $[10, 20)$ (FLUX2-Klein 4B).** The interval matches the timing measurements in Table 13. Pure editing reproduces Table 2; Gemini-caption results use the same $[10, 20)$ interval as the Qwen results. Caption models supply target captions; the benchmark metrics follow Section 4. Higher is better. Green/red intensity shows gains/losses from pure editing, normalized by the largest absolute change per metric within this table. The $\pm$ row gives the three-seed standard deviation of the pure-editing baseline.

	GEdit	ImgEdit	MagicBrush	GIE-Bench	
Caption model	O	Overall	CLIP-T	FC (%)	SSIM
$\pm$ std (3 seeds)	0.04	0.04	0.001	0.35	0.001
Pure editing (no caption)	7.11	3.74	0.317	73.40	0.805
Gemini-2.5-pro	7.53	4.16	0.324	76.74	0.684
Qwen3.5-27B	7.51	4.10	0.324	77.02	0.684
Qwen3.5-35B-A3B	7.48	4.11	0.323	77.37	0.683

E SELECTIVE-RULE SETTINGS

All models use the rule in Section 4.5, with the thresholds below shared across benchmarks. The probe step ($k = 9$), T2I interval ($[10, 16)$), and statistic definitions are unchanged: source attention is averaged across heads and all transformer layers before computing spatial CV and the 75th percentile π. The probe sits at the last step before the interval because attention patterns grow more differentiated over steps 0–9.

Model	τ_{cv}	τ_{π}
FLUX2-Klein 4B	0.50	0.22
FLUX2-Klein 9B	0.50	0.209
BAGEL	0.166	0.469

For 4B, we replayed pure-editing and fixed-interval outcomes on the GEdit dataset, seeking a task-averaged fidelity gain of at least 0.25 with a preservation loss of at most 0.05. Of 111 candidate threshold rules, 8 met both criteria, and $(0.50, 0.22)$ was selected as one feasible operating point.

For 9B and BAGEL, we transferred the 4B operating point using probe distributions on the GEdit dataset. On 9B, we retained the CV threshold and adjusted π to approximately match 4B's gate

and overall switching rates. On BAGEL, both thresholds were obtained by matching each gate's exceedance rate separately to 4B (13.5% for CV and 18.5% for π), then rounding. ImgEdit, MagicBrush, and GIE-Bench were not used to select these thresholds.

Comparison with random routing. To test whether the attention rule outperforms choosing edits at random, we replay per-sample GEdit scores (seed 1337) under pure editing and under the fixed $[10, 16)$ interval, and route each edit to one or the other. Random routing with probability p has expected macro-averaged scores equal to the p-weighted mix of the two; we set p so that its mean preservation equals that of the attention rule, and compare quality at that matched preservation. Table 15 reports the results for FLUX2-Klein 4B and 9B. On both models the rule exceeds the preservation-matched random baseline on overall, perceptual-quality, and fidelity scores, with the larger margin on 4B, where the thresholds were selected, and a smaller but same-signed margin on 9B, where they were transferred by distribution matching alone. BAGEL was not included in this replay.

Table 15: **Attention rule versus preservation-matched random routing (GEdit, GPT-4.1, macro averaging).** Random routing sends each edit to the $[10, 16)$ interval with a probability chosen to match the rule's mean preservation. Differences are rule minus random; positive favors the rule.

Model	Routing	Preservation	Fidelity	PQ	O
FLUX2-Klein 4B	Pure editing	8.902	8.395	7.421	7.112
	Attention rule (68% switched)	8.860	8.696	7.698	7.459
	Random, preservation-matched ($p = 0.33$)	8.860	8.546	7.578	7.295
	Difference	0	$+0.149$	$+0.120$	$+0.164$
FLUX2-Klein 9B	Pure editing	9.127	9.033	7.746	7.778
	Attention rule (69% switched)	9.085	9.201	7.869	7.949
	Random, preservation-matched ($p = 0.39$)	9.085	9.110	7.813	7.856
	Difference	0	$+0.091$	$+0.056$	$+0.093$

F PRESERVATION ANALYSIS

Judge protocol. We use GPT-4.1, the same judge as in our main GEdit evaluation, providing the source image, edited image, and editing instruction. The prompt and scoring scale are adapted from GEdit's semantic-consistency evaluation by retaining the preservation assessment and removing the fidelity assessment. The preservation-only prompt asks the judge to score unnecessary changes to unrelated content from 0 to 10, without penalizing changes needed to carry out the instruction. A score of 0 indicates substantial unrelated changes; 10 indicates that all unrelated content is preserved.

Paired comparisons. We evaluate preservation on GIE-Bench using these GPT scores and SSIM outside the edit mask. For each comparison, we match edits by identifier and average paired score differences with equal weight per edit. Because multiple edits can share a source image, confidence intervals and significance tests use standard errors clustered by source image.

Table 16 reports paired mean differences with confidence intervals and significance tests adjusted for multiple comparisons.

Table 16: **Paired preservation effects on GIE-Bench.** Positive differences indicate better preservation. Shading marks $p < 0.05$ (green: gain; red: loss). Fixed and selective application use $[10, 16)$. For each metric, Bonferroni 95% confidence intervals and Holm-adjusted p values account for multiple comparisons within the 4B baseline sweep and, separately, the pooled 9B/BAGEL comparisons. †The exploratory 4B selective–fixed comparison uses an ordinary 95% interval and an unadjusted p value.

	GPT preservation			SSIM		
Comparison	Δ	95% CI	p	Δ	95% CI	p
FLUX2-Klein 4B: comparisons against pure editing						
E → T2I, $k = 35$	-0.059	$[-0.207, +0.088]$	0.946	-0.0371	$[-0.0404, -0.0338]$	< 0.001
E → T2I, $k = 40$	$+0.099$	$[-0.046, +0.244]$	0.391	-0.0187	$[-0.0207, -0.0168]$	< 0.001
E → T2I, $k = 45$	$+0.093$	$[-0.043, +0.228]$	0.391	-0.0053	$[-0.0061, -0.0044]$	< 0.001
E → T2I → E, $[25, 35)$	-0.341	$[-0.521, -0.161]$	< 0.001	-0.0567	$[-0.0613, -0.0521]$	< 0.001
E → T2I → E, $[35, 45)$	$+0.030$	$[-0.125, +0.184]$	0.993	-0.0270	$[-0.0297, -0.0244]$	< 0.001
Fixed $[10, 16)$	-0.075	$[-0.246, +0.096]$	0.946	-0.0246	$[-0.0283, -0.0209]$	< 0.001
Selective $[10, 16)$	-0.041	$[-0.202, +0.121]$	0.993	-0.0191	$[-0.0225, -0.0157]$	< 0.001
FLUX2-Klein 4B: follow-up comparison						
Selective − fixed†	$+0.034$	$[-0.047, +0.116]$	0.410	$+0.0055$	$[+0.0041, +0.0069]$	< 0.001
FLUX2-Klein 9B						
Fixed − pure editing	-0.158	$[-0.338, +0.023]$	0.127	-0.0391	$[-0.0442, -0.0340]$	< 0.001
Selective − pure editing	-0.084	$[-0.235, +0.067]$	0.561	-0.0267	$[-0.0312, -0.0223]$	< 0.001
Selective − fixed	$+0.073$	$[-0.049, +0.195]$	0.561	$+0.0124$	$[+0.0093, +0.0155]$	< 0.001
BAGEL						
Fixed − pure editing	-0.044	$[-0.226, +0.139]$	1.000	-0.0082	$[-0.0096, -0.0068]$	< 0.001
Selective − pure editing	$+0.005$	$[-0.165, +0.174]$	1.000	-0.0055	$[-0.0067, -0.0042]$	< 0.001
Selective − fixed	$+0.048$	$[-0.076, +0.173]$	0.920	$+0.0027$	$[+0.0019, +0.0035]$	< 0.001

Among the paired comparisons in Table 16, only the 4B $[25, 35)$ interval shows a significant mean GPT-preservation reduction after correction. Every fixed and selective configuration lowers SSIM significantly relative to pure editing. Selective application recovers part of that loss, with a significant SSIM gain over fixed $[10, 16)$ on all three models and no significant GPT-preservation reduction relative to pure editing.

Figure measurements. Generations use fixed seeds, and preservation and correctness are assessed in independent judge runs. Figure 8 combines GPT-4.1 preservation and Gemini-3-Flash correctness by configuration. Its horizontal error bars are ordinary 95% confidence intervals for individual preservation means, using source-clustered standard errors and a t critical value with degrees of freedom equal to the number of source groups minus one. Configurations without per-edit preservation records are shown as means without error bars. The paired differences and simultaneous confidence intervals in Table 16 provide the comparisons against pure editing and between fixed and selective switching.