MINIMUM-SUFFICIENT SELECTIVE CONTROL FOR INSTRUCTION-BASED IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Instruction-based image editing now handles many simple requests well, but difficult edits still expose a fundamental control-allocation problem: weak intervention often under-edits hard cases, while uniformly strong intervention increases spillover and cost. We cast this problem as minimum-sufficient control allocation under a locked bounded-cost action family with explicit end-to-end efficiency accounting. We keep this predeclared intervention family fixed and learn a lightweight controller from deployment-time observables. The policy first learns whether to leave the temporal anchor, then learns the promoted temporal horizon, and finally applies deterministic localization to the escalated branch. On locked MagicBrush and PIE_Bench_pp benchmarks, the full policy **FO** improves over the temporal-only anchor **TO** from 0.7607 to 0.8481 edit success and from 22.9583 to 24.7291 composite score while staying within the same efficiency envelope, with sub-second latency and low VRAM. Matched-budget controls, multi-seed confidence intervals, abstraction-transfer tests, and oracle-based diagnostics further support that the gain comes from better allocation rather than a blanket move to a heavier editing path. Overall, selective escalation and localization make the same bounded intervention set substantially more effective on hard instruction-based edits.

1 INTRODUCTION

Instruction-based image editing allows users to modify an image through natural-language instructions while preserving irrelevant content, and recent work has made this setting increasingly practical in diverse real-world editing scenarios. Prompt-based editing manipulates denoising or attention behavior without retraining the whole editor (Meng et al., 2022; Hertz et al., 2023; Couairon et al., 2023); instruction-tuned models learn direct mappings from image–instruction pairs to edited outputs (Brooks et al., 2023; Zhao et al., 2024); multimodal reasoning improves the interpretation of complex requests (Huang et al., 2024b; Zhou et al., 2025); and structural or spatial control injects stronger conditional signals when geometry or locality matters (Zhang et al., 2023b; Mou et al., 2024). These advances improve endpoint editing quality on many common requests. Yet complex compositional, structural, or tightly localized edits remain difficult. We view these failures through a control-allocation lens: the system must decide not only how strong an editor can be, but how much control each image–instruction pair needs.

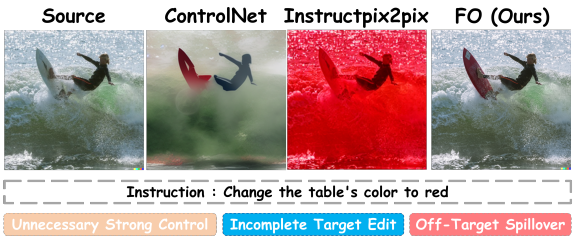

Figure 1: Problem setup. Three representative failure modes in complex instruction-based editing: incomplete target editing under weak control, off-target spillover under strong control, and the resulting allocation tradeoff.

These directions are valuable, but they mainly improve what happens after a stronger editing pathway has been chosen. This leaves the routing decision implicit. If the intervention is too weak, difficult edits may remain incomplete; if strong control is applied uniformly, easy or localized requests may suffer off-target spillover and unnecessary cost. The key challenge is not to replace stronger editors,

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 2: Qualitative benefit overview. Three representative examples show improved target selection, more complete edit execution, and tighter localization of the requested change.

but to decide whether a sample should leave the cheap default path, how much stronger control should be retained after escalation, and how the escalated branch should be spatially constrained.

We address this problem by making control selection explicit. Given a fixed set of bounded-cost interventions, the goal is to choose the least control that is sufficient for each image–instruction pair while measuring latency and memory end to end. Rather than introducing a new editing backbone, we keep this predeclared intervention family fixed and learn a controller for this decision. It first tests whether the cheap default path is sufficient and then selects a retained denoising horizon for escalated samples. The full policy deterministically applies a spatial gate to the escalated branch to limit irrelevant changes. This yields a temporal-only anchor TO and a full policy FO that adds localized refinement on top of the same gate and temporal decisions. Across locked benchmarks, the proposed policy improves edit success and overall quality with nearly unchanged latency and memory, supporting that the gain comes from selective allocation rather than a uniformly heavier editing path.

Figure 2 previews representative outcomes and illustrates the practical benefits targeted by the policy, including more complete target edits and tighter localization of stronger control. Subsequent matched-budget controls, oracle diagnostics, and transfer analyses further test this allocation claim. Our main contributions are as follows:

- We formulate instruction-based image editing as a *minimum-sufficient control allocation* problem over a locked bounded-cost intervention family, and show that a fixed intervention set can be deployed more effectively under explicit end-to-end efficiency accounting.
- We propose a three-stage policy with two learned routing decisions, whether to escalate and how much stronger control to retain, followed by deterministic spatial refinement.
- We establish a decision-focused evaluation protocol using locked benchmarks, matched-budget simplification tests, abstraction-transfer analysis, and oracle-based diagnostics to verify that the observed gains come from improved control allocation rather than uniformly stronger editing.

2 Related Work

2.1 Stronger editors for instruction-based image editing

Instruction-based image editing methods aim to modify an input image according to a textual request while preserving irrelevant content. Recent work has improved endpoint editing quality through prompt-based editing, inversion, text-guided diffusion translation, instruction tuning, and multimodal reasoning (Meng et al., 2022; Hertz et al., 2023; Brooks et al., 2023; Mokady et al., 2023; Couairon et al., 2023; Parmar et al., 2023; Huang et al., 2024b; Xu et al., 2025; Zhou et al., 2025; Das Biswas et al., 2025; Pathiraja et al., 2025). Benchmark and dataset efforts such as MagicBrush and EditBench have also made evaluation more realistic and more demanding (Zhang et al., 2023a; Wang et al., 2023). These advances improve what happens once a strong editing pathway is available, but they largely assume that this stronger pathway has already been invoked. Our focus is different: we study how to decide when stronger control should be used at all, and how much of it should be retained once escalation becomes necessary.

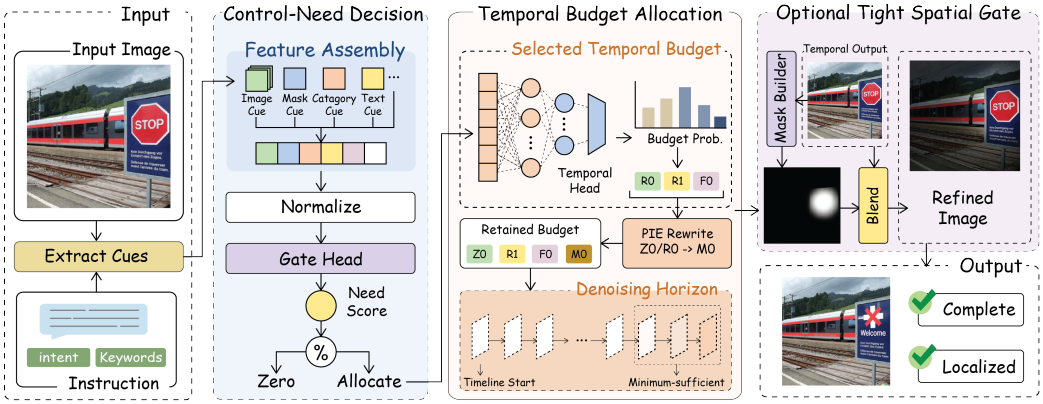

Figure 3: Method overview. Observable image and instruction cues feed a control-need gate; escalated samples then select a denoising horizon, after which the full policy deterministically constructs a tight spatial mask within the same locked intervention family.

2.2 Localized control and spillover reduction

Localized-control methods align instructions with target regions to reduce off-target changes, using controllable diffusion backbones, localized attention control, learnable region discovery, and plug-and-play conditional mechanisms (Zhang et al., 2023b; Mou et al., 2024; Cao et al., 2023; Zou et al., 2024; Lin et al., 2024; Ju et al., 2024; Ma et al., 2024a; Wu & De la Torre, 2023; Nair et al., 2023; Cheng et al., 2023). They refine where stronger control is applied, but do not decide whether a sample should leave the cheap path, how much stronger control should be retained, or how an escalated branch should be spatially constrained as a final refinement.

2.3 Adaptive inference and cost-aware decision making

A related literature studies how to allocate computation according to input difficulty, for example by accelerating diffusion sampling, reusing cached features, or adapting token and inference depth (Lu et al., 2022; Yu et al., 2024; Zheng et al., 2024; Ma et al., 2024b; Zhang et al., 2025; Rao et al., 2021; Yin et al., 2022; Wang et al., 2021; 2022; Huang et al., 2018; Wu et al., 2018; Wang et al., 2018). These works provide a useful cost-sensitive lens, but their action spaces typically involve generic compute, routing depth, or token allocation. In contrast, we study allocation over a fixed family of editing interventions, where the decision is whether to remain on a cheap temporal anchor, whether to escalate to stronger structural control, how much temporal control to retain, and how the escalated branch should be spatially localized.

3 Method

Figure 3 summarizes the framework, including its observable inputs, adequacy-constrained oracle supervision, factorized controller, shared TO/FO decision contract, and allocation diagnostics.

3.1 Observable Inputs and Locked Intervention Family

The controller operates on a frozen feature vector $\phi(x, q)$ built entirely from deployment-time observables rather than latent image or text embeddings. It contains dataset indicators, instruction and prompt cues, simple source-image statistics, and mask/category cues; Appendix A lists the exact inventory. Restricting the inputs to deployment-time observables keeps routing lightweight, auditable, and transferable across frozen editing stacks.

The intervention family is fixed throughout every sweep. For the reported locked action-family configuration (v21), $\mathcal{Z} = \{\mathrm{Z0}, \mathrm{Z1}, \mathrm{R0}, \mathrm{R1}, \mathrm{F0}\}$. Here Z0 and Z1 are zero-control InstructPix2Pix

anchors with different step and guidance settings; R0 and R1 retain progressively stronger and longer ControlNet guidance; and F0 is the full-control anchor. The controller only routes among these predeclared actions. It does not change the evaluation protocol, image resolution, or metric definition, so differences remain attributable to allocation rather than protocol drift.

3.2 Adequacy-Constrained Oracle Labeling

We operationalize minimum-sufficient allocation as adequacy-first selection over the ordered, locked intervention family. Under a shared protocol, the controller learns to imitate the selected minimum-sufficient intervention.

We define a discrete intervention tuple

$$z = (g, h), \tag{1}$$

where $g \in \{0, 1\}$ is the escalation decision and $h \in \mathcal{H}$ is the selected temporal intervention schedule. Each named intervention in $\mathcal{Z}$ corresponds to a fixed instantiation of this routed pair under the retained action family. Spatial refinement is derived deterministically after routing and is not part of the oracle label.

Let $T_d(x, q, z)$ and $P_d(x, q, z)$ denote target-edit and preservation scores under dataset d, oriented so larger values are better. The retained benchmark protocol combines these two components by a fixed linear rule:

$$Q_d(x, q, z) = \alpha\, T_d(x, q, z) + (1 - \alpha)\, P_d(x, q, z), \tag{2}$$

where the global coefficient α is fixed before held-out evaluation and shared across all datasets, methods, and samples. Only the definitions of T_d and P_d vary by dataset; no post-hoc reweighting is used. The composite ranks offline candidates only after the preceding adequacy and intervention-order criteria, while a separate hard gate enforces adequacy:

$$A_d(x, q, z) = \mathbb{1}\Big[T_d(x, q, z) \geq \theta_d^{\mathrm{tar}} \ \wedge \ P_d(x, q, z) \geq \theta_d^{\mathrm{keep}}\Big]. \tag{3}$$

Offline supervision is therefore reduced to an explicit oracle-labeling problem over the locked family. For each calibration sample, we enumerate the candidate interventions and select the lowest-ranked adequate action. Measured end-to-end latency $\tau(z)$, peak VRAM $\nu(z)$, and composite quality are successive tie-breakers:

$$z^\star(x, q) = \underset{z \in \mathcal{Z}: A_d(x,q,z)=1}{\arg\operatorname{lexmin}} \big(\mathrm{rank}(z),\, \tau(z),\, \nu(z),\, -Q_d(x, q, z)\big). \tag{4}$$

Here $\mathrm{rank}(z)$ encodes the fixed ordering from weak anchors to promoted reduced-control schedules and the strongest retained mode. If no candidate is adequate, the audit protocol records a hard-failure case and assigns $z^\star = F0$ for bookkeeping only. This offline sweep yields the gate and temporal supervision targets

$$y_g(x, q) = \mathbb{1}\big[z^\star(x, q) \in \{R0, R1, F0\}\big], \qquad y_h(x, q) = h^\star(x, q), \tag{5}$$

where $h^\star(x, q)$ denotes the promoted temporal label induced by the selected oracle action, and y_h is used only on non-hard-failure escalated rows. The exact retained benchmark instantiation, dataset thresholds, and implementation details are summarized in Appendix A.

3.3 Factorized Three-Stage Policy

The policy has two learned routing stages followed by deterministic spatial refinement under a fixed intervention family. The gate decides whether a sample may leave the cheap path, the temporal stage decides how much denoising-time control to retain after escalation, and the spatial stage localizes only the promoted branch. This separation supports matched-budget tests, simplification baselines, and direct evaluation of spatial refinement under the same routing decisions.

Stage I: control-need gate. A posterior is estimated over the binary control-need decision:

$$\hat{g}(x, q) = \underset{c \in \{0,1\}}{\arg\max}\, p_\theta(g = c \mid \phi(x, q)). \tag{6}$$

If $\hat{g} = 0$, the sample remains on the frozen zero-control anchor. If $\hat{g} = 1$, the sample is allowed to spend stronger control. This stage is the main safeguard against collapsing to always-heavy

editing. The retained head is a lightweight supervised classifier trained on the offline gate target. We later calibrate the scalar score $p_\theta(g = 1 \mid \phi(x, q))$ per dataset for selective-prediction analysis, but benchmark routing uses the frozen learned gate.

Stage II: retained temporal horizon. Conditioned on $\hat{g} = 1$, the controller estimates a posterior over the promoted temporal family and routes by maximum posterior:

$$\hat{h}(x, q) = \underset{h \in \{R0, R1, F0\}}{\arg\max} \; p_\psi(h \mid \phi(x, q), \hat{g} = 1). \tag{7}$$

The exact normalized temporal schedules for $R0$, $R1$, and $F0$ are fixed implementation choices and are deferred to Appendix A. This stage learns how much promoted control to allocate before spatial refinement. The temporal-only method TO is exactly the retained temporal decision without the spatial refinement stage. The retained temporal head is a lightweight supervised classifier trained on offline escalated rows that are not marked as hard failures.

Stage III: localized spatial refinement. For every sample with $\hat{g} = 1$, FO adds a deterministic local restriction to reduce off-target change:

$$\tilde{m}(\cdot \mid x, q, \hat{h}) = \mathrm{SpatialGate}(x, q, \hat{h}). \tag{8}$$

Spatial activation is therefore fixed as $\hat{s}_{\mathrm{TO}} = 0$ and $\hat{s}_{\mathrm{FO}} = \hat{g}$; no independent spatial classifier is trained. The full policy FO therefore differs from TO by an escalation-aware spatial refinement, not by swapping in a different editor. The spatial stage is deterministic rather than mask-supervised: it uses saliency and directional priors from the source image and instruction together with a budget-dependent coverage schedule and mask-based blending. It never consumes a target image, oracle mask, or post-hoc quality signal at inference. For the reported FO row, the retained `hybrid_saliency_tight` gate instantiates $\mathrm{SpatialGate}$; Table 3 compares it against the trivial full-image gate and other deterministic alternatives under the same temporal predictions, while the exact gate construction is deferred to Appendix A.

3.4 FO as a Spatial Extension of TO

The controller emits a structured decision rather than a final image directly. This yields a clean contract between the promoted temporal line and the full method:

$$\begin{aligned} TO(x, q) &= \mathrm{Edit}(x, q, \hat{g}, \hat{h}), \\ FO(x, q) &= \begin{cases} \mathrm{Refine}\big(TO(x, q), x, \tilde{m}(\cdot \mid x, q, \hat{h})\big), & \hat{g} = 1, \\ TO(x, q), & \hat{g} = 0. \end{cases} \end{aligned} \tag{9}$$

Both methods share the same input, features, gate, and temporal-horizon decision, and FO extends the contract only after the learned gate has selected escalation. Because TO and FO share this frozen gate-temporal policy, the matched-budget and exact-budget controls in Section 4 test allocation rather than a different promotion rule.

3.5 Policy Diagnostics

Endpoint metrics do not directly assess allocation. We therefore report oracle exact match, calibration, transition-level recovery and regression, matched-budget controls, and oracle-label perturbations under the same locked action space.

4 Experiments

4.1 Protocol and Compared Methods

Settings. We evaluate locked held-out splits from MagicBrush (Zhang et al., 2023a) and PIE_Bench_pp (Huang et al., 2024a), with 528 and 525 examples. All methods share resolution, timing, logging, and metric definitions, so performance differences reflect allocation rather than protocol variation or reporting differences across methods.

Table 1: Unified benchmark overview of reproduced baseline methods and the promoted pair. The left block records whether control-need decision, temporal horizon allocation, and the tight spatial gate are explicitly present. The right block reports overall benchmark metrics together with dataset-wise edit success on MagicBrush and PIE_Bench_pp.

		Selective-control components			Overall benchmark					Dataset-wise success	
Method	Source	Need	Horizon	Gate	Succ.	Pres.	Comp.	Lat.	VRAM	MB	PIE
InstructPix2Pix	CVPR 2023	✗	✗	✗	0.750	17.200	22.846	0.768	2.896	0.650	0.851
ControlNet img2img	ICCV 2023	✗	✗	✗	0.770	17.066	21.135	0.740	5.908	0.644	0.897
InstructDiffusion	CVPR 2024	✗	✗	✗	0.743	19.243	24.464	13.336	7.627	–	–
UltraEdit SD3	NeurIPS 2024	✗	✗	✗	0.958	26.522	28.257	7.725	12.344	0.992	0.924
TO	Ours	✓	✓	✗	0.761	18.018	22.958	0.460	5.617	0.623	0.899
FO	Ours	✓	✓	✓	**0.848**	**20.786**	**24.729**	0.467	5.629	**0.782**	**0.914**

Table 2: Diagnostic recovery slice defined by temporal-anchor failure under the locked adequacy criterion (205 held-out cases). The table reports reproduced baselines and the promoted pair under official per-sample end-to-end latency and pipeline-VRAM reruns.

Method	Succ.	Pres.	Comp.	Lat.	VRAM
IP2P	0.263	14.223	16.906	0.768	**2.722**
CN	0.342	15.714	16.921	0.719	5.908
TO	0.000	14.845	16.800	0.561	5.433
FO	**0.371**	**21.430**	**21.110**	**0.650**	5.438

Baselines. Our primary comparison is between the promoted temporal-only policy TO and the promoted full policy FO. We also compare against representative reproduced baselines, including InstructPix2Pix (Brooks et al., 2023), ControlNet img2img (Zhang et al., 2023b), InstructDiffusion (Geng et al., 2024), and UltraEdit SD3 (Zhao et al., 2024). These baselines cover both relatively light and substantially heavier editing pipelines, allowing us to evaluate whether the proposed selective controller improves the quality–cost tradeoff without relying on a uniformly stronger endpoint.

Benchmarks and Metrics. For all methods, we report edit success rate, defined as the fraction of held-out samples that satisfy the locked dataset-specific adequacy gate, together with preservation, composite score, end-to-end latency, and pipeline peak VRAM under the same protocol. Unless otherwise specified, all later decision-quality analyses reuse the same locked outputs and add multi-seed stability checks, matched-budget controls, oracle reconstruction, frozen cross-stack transfer, leave-one-family-out transfer, and a retained run ledger for compute accounting. This shared evaluation basis keeps the main results, robustness checks, and transfer analyses mutually comparable.

4.2 FO Improves the Quality–Cost Tradeoff Without Leaving the Same Cost Envelope

Table 1 shows that the gain stays within the same cost envelope: relative to TO, FO raises success from 0.761 to 0.848 and composite quality from 22.958 to 24.729, while adding only 0.007 s latency and 0.012 GB VRAM. UltraEdit SD3 is stronger in absolute quality but occupies a different efficiency regime (7.725 versus 0.467 s; 12.344 versus 5.629 GB). Success rises more on MagicBrush (0.623 to 0.782) than on PIE_Bench_pp (0.899 to 0.914), indicating that FO mainly recovers edits unresolved by the temporal anchor. Thus, the improvement does not require a uniformly heavier inference regime: FO changes where control is spent while preserving TO's operating envelope. Dataset gains track remaining correction headroom, with larger increases where TO leaves edits unresolved.

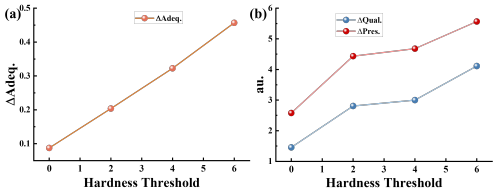

Figure 4: Hardness-stratified gain of FO over TO. (a) Success-rate difference ΔSucc. across hardness thresholds. (b) Composite-score difference ΔComp. and preservation difference ΔPres. across hardness thresholds.

Table 3: Method-line summary under the locked benchmark protocol. The temporal-oracle row is pre-refinement only.

Track	Row	Role	Succ.	Pres.	Comp.	Lat.	VRAM
Ablation	Heuristic	selective baseline	0.639	16.787	22.052	0.585	5.610
	Temporal oracle	pre-refinement	0.852	17.783	23.212	0.528	5.608
	TO	temporal anchor	0.761	18.018	22.958	0.460	5.617
	FI	full-image gate	0.761	18.018	22.958	0.469	5.620
	DB	directional box	0.844	19.112	23.939	0.463	5.630
	HS	hybrid saliency	0.840	20.142	24.166	0.466	5.630
	FO	promoted full method	0.848	20.786	24.729	0.467	5.629
	AG	attention grounded	0.844	19.162	24.302	0.474	6.192

Table 4: Complete-set paired spatial analysis for FO versus TO. "Repair/new" denotes fail→pass/pass→fail under the locked criterion. Continuous differences use sample-level paired-bootstrap 95% confidence intervals.

Dataset	N	Spatial active	ΔPres. [95% CI]	Under-edit	Spillover	Adequacy
MagicBrush	528	93 (17.61%)	+3.4715 [2.6181, 4.4326]	40 / 0	72 / 0	84 / 0
PIE-Bench++	525	35 (6.67%)	+2.0591 [1.1184, 3.1537]	2 / 0	11 / 0	8 / 0
Overall	1,053	128 (12.16%)	–	42 / 0	83 / 0	92 / 0

Table 5: Key decision-quality diagnostics.

Signal	Value
ΔSucc. CI_{95}, FO–TO	0.0852 [0.0762, 0.0950]
ΔComp. CI_{95}, FO–TO	1.4777 [1.3656, 1.5937]
Gate ECE raw → Platt	0.2567 → 0.0104
Learned vs Rule-3Stage ΔSucc. CI_{95}	0.0617 [0.0380, 0.0855]
Learned vs Rule-3Stage ΔComp. CI_{95}	0.4122 [0.2666, 0.5624]
Recovery / regression	36.5% / 1.2%

Table 2 isolates the threshold-5 hard slice, testing recovery where temporal allocation alone is insufficient rather than averaging over easy requests. Here TO records no successful edits, whereas FO reaches 0.371 success with the best preservation and composite scores, at lower latency and VRAM than ControlNet.

Harder subsets reveal whether the benefit of FO scales with edit difficulty, as shown in Figure 4. The gap grows under stricter thresholds, showing that the spatial branch helps most where the cheap temporal path loses control. This supports the allocation view: extra control is applied where it is most needed, rather than uniformly across samples. Across both evaluations, difficult-edit gains remain within the controller's low-cost operating regime.

4.3 Ablations Confirm the Temporal Anchor and Spatial Refinement

Ablations on the policy structure clarify whether the retained gain requires spatial control everywhere, as shown in Table 3. The temporal anchor already improves over the heuristic baseline, so horizon allocation contributes before spatial refinement. The temporal-oracle row is pre-refinement only. Full-image gating stays effectively tied with TO, while localized variants recover the missing gain and FO gives the best preservation and composite score. Thus, spatial refinement helps by constraining stronger control, not by blindly increasing editing strength.

Table 4 separates repair from regression over all 1,053 held-out samples. FO repairs 42 under-edit and 83 spillover failures while introducing none under either criterion, yielding 92 Failure→Success and no reverse transition. The spatial stage is active on only 12.16% of samples, so the gains do not arise from uniformly applying the strongest intervention. Paired counts isolate localized recovery rather than blanket escalation or a preservation-for-success trade.

Table 6: Abstraction-transfer summary over external editor pairs and leave-one-family-out instruction splits. The external-pair rows report endpoint-contract transfer; the held-out-family rows report frozen gate transfer under instruction-family shift.

Setting	Contract signal	Matched simple baseline	Recovery / regression	Auxiliary signal
Primary pair: IP2P → UltraEdit SD3	Succ. 0.884 vs TO 0.750	+0.0399 vs temporal-topk	0.591 / 0.179	ΔComp. +0.566
Support pair: ControlNet → UltraEdit SD3	Succ. 0.843 vs TO 0.770	+0.0163 vs temporal-topk	0.377 / 0.280	ΔComp. +0.282
Held-out add_object family	Prec. 0.5000	Recall 0.2857	Lift 0.4266	ECE 0.0439
Held-out attribute/appearance family	Prec. 0.4020	Recall 0.8119	Lift 0.2931	ECE 0.0177
Held-out background/scene family	Prec. 0.1667	Recall 0.8421	Lift 0.1293	ECE 0.0662

Table 7: Matched-budget strategy comparison using the locked v21 policy rows. Random and top-k baselines use the same number of Z0→F0 promotions as the learned selective policy.

Strategy	Succ.	Pres.	Comp.	Lat.	VRAM
Random matched	0.709	17.212	22.652	0.649	5.328
Top-k matched	0.697	17.143	22.614	0.638	5.328
Always escalate	0.640	16.955	21.509	1.174	5.348
Learned selective	**0.722**	**17.342**	**22.853**	**0.639**	**5.328**

Table 8: Cross-dataset transfer with the controller frozen on one dataset and evaluated on the other without retuning. MB denotes MagicBrush and PIE denotes PIE_Bench_pp.

Train	Eval	Scope	Succ.	Pres.	Comp.	Lat. (s)
MB	MB	in-dom.	0.5473	17.8727	18.3431	0.5186
MB	PIE	cross	0.8152	16.4917	26.6334	0.5297
PIE	PIE	in-dom.	0.7810	15.4753	26.5356	0.4489
PIE	MB	cross	0.5095	17.4432	18.0960	0.4707

4.4 Decision-Quality Evidence and Oracle Robustness Support the Allocation Claim

Table 5 summarizes stability, calibration, and residual errors, with full diagnostics in Appendix Table 19. Multi-seed intervals keep the FO–TO gaps positive, while calibration reduces gate ECE from 0.2567 to 0.0104. Under exact-budget simplification, the learned controller still beats Rule-3Stage by 0.0617 success and 0.4122 quality, attributing the gain to selective promotion rather than promotion count alone under the fixed budget.

Oracle-label perturbations in Appendix Table 14 retain the `rf`/`rf` head and vary composite quality by less than 0.05. The head-family check in Appendix Table 20 also reselects RandomForest over linear and tiny-MLP alternatives. These checks probe complementary sensitivities: label perturbations change the supervision rule, whereas head-family selection changes the fitted hypothesis class used by the routing controller.

A matched-promotion counterfactual isolates selective allocation from promotion count (Table 7). At the same promotion budget, the learned selector reaches 0.722 success versus 0.709 for random and 0.697 for top-k matching at similar latency; always-escalate is slower and worse. Thus, which samples are promoted matters more than the raw promotion count.

4.5 Transfer and Generalization Beyond a Single Dataset

Table 8 gives a qualified cross-dataset result: MagicBrush→PIE_Bench_pp reaches 0.815 success, whereas the reverse direction reaches 0.510. The asymmetry rules out a dataset-invariant interpretation. Instead, the learned difficulty signals transfer in one direction while remaining shaped by dataset composition and task difficulty across the evaluated benchmarks.

Table 6 moves beyond dataset shift. Endpoint-pair gains are larger for InstructPix2Pix→UltraEdit SD3 than for the stronger ControlNet img2img→UltraEdit SD3 starting point, but remain positive

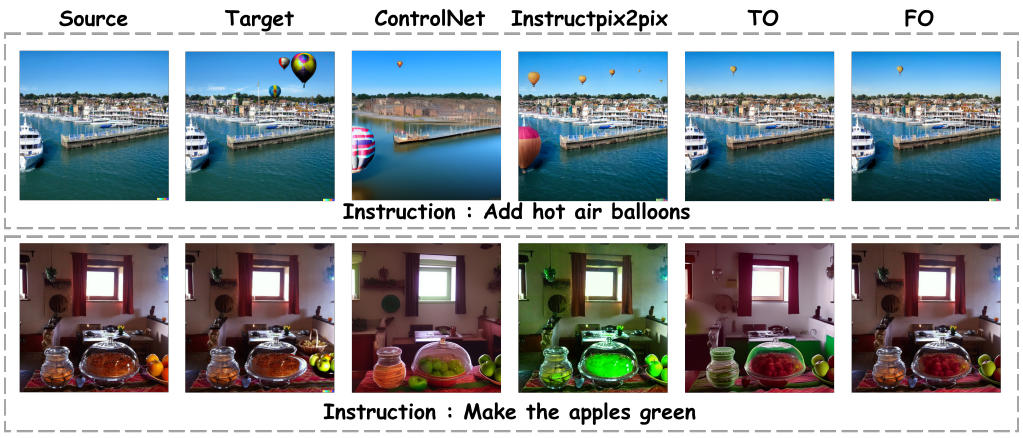

Figure 5: Final qualitative comparison on held-out examples. Each row shows source, target, reproduced low-cost baselines, TO, and FO.

for both evaluated endpoint pairs without retraining. This experiment tests reuse of the allocation contract across endpoint pairs, not unchanged transfer of the complete three-stage policy.

Under semantic shift, the frozen gate retains positive lift across all three held-out families despite uneven precision–recall tradeoffs. Complexity-stratified results appear in Appendix Table 21.

Localized control is selected for 42 of 240 UltraEdit SD3 samples (17.5%; Table 9). Against free-form action, composite quality and preservation rise by 0.4576 and 0.6830; a $+0.0083$ success change satisfies the -0.02 non-inferiority margin. Joint step-and-mask variation supports coupled over uniform intervention.

Table 9: Coupled localized-action test within UltraEdit SD3 on 240 held-out samples.

Signal	Estimate [95% CI]
Localized selection	42/240 (17.5%)
ΔComp.	$+0.4576$ [0.2268, 0.7179]
ΔPres.	$+0.6830$ [0.4208, 0.9789]
ΔSucc.	$+0.0083$ [-0.0083, 0.0250]

Human preference results. The 500 comparison instances provide an additional check on whether the quantitative gains are perceptible in practice, as reported in Table 10. Each instance was independently rated by all five annotators. FO wins overall preferences against every opponent. Against TO, instruction success is close to tied, but preservation and overall preference favor FO by 65.3% and 64.0%, consistent with stronger edits that remain better localized. The same preference pattern against reproduced baselines and the temporal anchor suggests that the preservation gains are visible to human raters, not only to the automatic metrics.

Table 10: Anonymous pairwise human-preference results from 500 comparison instances.

Opp.	Instr.	Pres.	Overall
CN	68.6%	66.7%	73.1%
IP2P	54.5%	63.2%	65.1%
TO	50.7%	65.3%	64.0%

Figure 5 mirrors the quantitative failure modes: baselines under-edit, uncontrolled edits alter irrelevant regions, and the displayed FO outputs complete requests while keeping added control spatially focused on requested regions in the shown examples.

5 Conclusion

We cast instruction-based image editing as minimum-sufficient control allocation over a bounded intervention family. The policy combines learned escalation and temporal allocation with deterministic spatial refinement. Across locked benchmarks, it improves the temporal-only anchor within the same latency/VRAM envelope. Matched-budget, paired localization, and transfer analyses support selective allocation while showing that complete policy transfer remains dataset dependent. Overall, selective control improves a locked intervention family without uniformly routing all inputs through a heavier editing path at inference.

REFERENCES

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

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22560–22570, 2023.

Bin Cheng, Zuhao Liu, Yunbo Peng, and Yue Lin. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22736–22746, 2023.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=3lge0p5o-M-.

Shristi Das Biswas, Matthew Shreve, Xuelu Li, Prateek Singhal, and Kaushik Roy. PIXELS: Progressive image xemplar-based editing with latent surgery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2663–2671, 2025. doi: 10.1609/aaai.v39i3.32270.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, Dong Chen, and Baining Guo. InstructDiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12709–12720, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hk2aImxAb.

Mingzhen Huang, Jialing Cai, Shan Jia, Vishnu Suresh Lokhande, and Siwei Lyu. ParallelEdits: Efficient multi-aspect text-driven image editing with attention grouping. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22569–22595, 2024a. doi: 10.52202/079017-0710.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. SmartEdit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371, 2024b.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. BrushNet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 150–168, 2024. doi: 10.1007/978-3-031-72661-3_9.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via learnable regions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7059–7068, 2024. doi: 10.1109/CVPR52733.2024.00674.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787, 2022. doi: 10.52202/068431-0418.

Wan-Duo Kurt Ma, Avisek Lahiri, J. P. Lewis, Thomas Leung, and W. Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4098–4106, 2024a. doi: 10.1609/aaai.v38i5.28204.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15762–15772, 2024b. doi: 10.1109/CVPR52733.2024.01492.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. NULL-Text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047, 2023.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226.

Nithin Gopalakrishnan Nair, Anoop Cherian, Suhas Lohit, Ye Wang, Toshiaki Koike-Akino, Vishal M. Patel, and Tim K. Marks. Steered diffusion: A generalized framework for plug-and-play conditional image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20850–20860, 2023.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023. doi: 10.1145/3588432.3591513.

Bimsara Pathiraja, Maitreya Patel, Shivam Singh, Yezhou Yang, and Chitta Baral. RefEdit: A benchmark and method for improving instruction-based image editing model on referring expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15646–15656, 2025.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949, 2021.

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

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen Editor and EditBench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18359–18369, 2023.

Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. SkipNet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 420–436, 2018. doi: 10.1007/978-3-030-01261-8_25.

Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16249–16258, 2021.

Yulin Wang, Yang Yue, Yuanze Lin, Haojun Jiang, Zihang Lai, Victor Kulikov, Nikita Orlov, Humphrey Shi, and Gao Huang. AdaFocus V2: End-to-end training of spatial dynamic networks for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20062–20072, 2022. doi: 10.1109/CVPR52688.2022.01943.

Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7378–7387, 2023.

Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S. Davis, Kristen Grauman, and Rogerio Feris. BlockDrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8817–8826, 2018.

Yingjing Xu, Jie Kong, Jiazhi Wang, Xiao Pan, Bo Lin, and Qiang Liu. InsightEdit: Towards better instruction following for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2694–2703, 2025.

Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10809–10818, 2022.

Zihao Yu, Haoyang Li, Fangcheng Fu, Xupeng Miao, and Bin Cui. Accelerating text-to-image editing via cache-enabled sparse diffusion inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16605–16613, 2024. doi: 10.1609/aaai.v38i15.29599.

Hui Zhang, Zuxuan Wu, Zhen Xing, Jie Shao, and Yu-Gang Jiang. AdaDiff: Adaptive step selection for fast diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9914–9922, 2025. doi: 10.1609/aaai.v39i9.33075.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31428–31449, 2023a. doi: 10.52202/075280-1365. Datasets and Benchmarks Track.

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 (ICCV)*, pp. 3836–3847, 2023b.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 3058–3093, 2024. doi: 10.52202/079017-0100. Datasets and Benchmarks Track.

Qingping Zheng, Yuanfan Guo, Jiankang Deng, Jianhua Han, Ying Li, Songcen Xu, and Hang Xu. Any-Size-Diffusion: Toward efficient text-driven synthesis for any-size HD images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7571–7578, 2024. doi: 10.1609/aaai.v38i7.28589.

Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. FireEdit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13093–13103, 2025.

Siyu Zou, Jiji Tang, Yiyi Zhou, Jing He, Chaoyi Zhao, Rongsheng Zhang, Zhipeng Hu, and Xiaoshuai Sun. Towards efficient diffusion-based image editing with instant attention masks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7864–7872, 2024. doi: 10.1609/aaai.v38i7.28622.

A Supplementary Controller Details

A.1 Controller Construction

Observable feature groups. The retained schema has 36 scalar features: dataset indicators (3), instruction features (14), source/target prompt-length features (6), source-image statistics (7), mask-geometry features (4), and category cues (2). Instruction features comprise character/word counts, punctuation, lexical diversity, average word length, and indicators for add/remove/replace/style/background/color/pose/local terms. Image features are RGB/gray mean and standard deviation, mean saturation/value, and Canny edge density; mask features record presence, coverage, and normalized bounding-box height/width. The dataset coordinates encode MagicBrush, PIE_Bench_pp, and HQ-Edit. MagicBrush provides a mask but no prompt or category fields; PIE_Bench_pp provides source/target prompts, a mask, and category, with its target prompt also serving as the edit request. Missing optional fields are zero-filled. Consequently, the controller can learn dataset-conditioned behavior under this benchmark-specific input contract, although it never receives the target image, generated output, oracle action, or evaluation score. The deterministic spatial refinement itself does not consume the benchmark mask. The 21-D deployment-core ablation removes benchmark-specific identity, mask, and category coordinates and is reported as a stricter deployment-oriented check.

Offline supervision and retained protocol. MagicBrush calibration uses the 1,024 training examples with smallest SHA-1 hashes of their sample identifiers; all 528 local development examples form the evaluation set. PIE_Bench_pp is partitioned independently within category by SHA-1 hash: 25% (at least 10 examples per category) is used for calibration and the remaining 525 examples for evaluation. Oracle labeling and controller selection use only calibration rows, after which the policy is frozen. The action registry, adequacy thresholds, and global composite weighting $Q_d = \alpha T_d + (1-\alpha)P_d$ are shared across all datasets, methods, and samples and fixed as part of the locked protocol before held-out evaluation. The composite cannot trade away preservation failure: adequacy requires T_d and P_d to pass separate hard thresholds, and Q_d enters oracle labeling only as the final tie-break after intervention rank, latency, and VRAM. If no candidate is adequate, it is recorded as a hard failure and assigned $F0$ for gate bookkeeping only; such rows are excluded from temporal-head fitting.

Benchmark metrics and thresholds. For MagicBrush, T_d is output–target PSNR inside the benchmark edit mask and P_d is output–source PSNR outside it; adequacy requires $T_d \geq 15$ and $P_d \geq 16$. For PIE_Bench_pp, T_d is CLIP similarity between the output and target prompt multiplied by 100, and P_d is outside-mask output–source PSNR; adequacy requires $T_d \geq 25$ and $P_d \geq 16$. If the PIE mask leaves no outside region, P_d falls back to whole-image output–source PSNR with threshold 10. For both datasets and all methods, $Q_d = \alpha T_d + (1-\alpha)P_d$ uses the same fixed global coefficient. For N evaluation samples, we report $\mathrm{Succ.} = N^{-1}\sum_i A_d^{(i)}$, $\mathrm{Pres.} = N^{-1}\sum_i P_d^{(i)}$, and $\mathrm{Comp.} = N^{-1}\sum_i Q_d^{(i)}$. Lat. is mean per-sample wall-clock time and VRAM is mean per-sample pipeline peak memory. Overall entries are sample-count weighted over the 1,053 evaluation examples. These success and composite summaries are protocol-derived rather than official metrics supplied by either benchmark.

Controller fitting and model selection. The gate target is $y_g = 1$ exactly for oracle actions in $\{R0, R1, F0\}$; the temporal target is the selected action in that set on escalated, non-hard-failure rows. For each head, we compare a linear classifier, a two-hidden-layer MLP (48 and 24 units), and a random forest by stratified five-fold cross-validation. Selection is lexicographic by accuracy, balanced accuracy, and lower 10-bin expected calibration error. The linear and MLP candidates use class-weighted cross-entropy: the linear model uses AdamW (5×10^{-3} learning rate, 10^{-3} weight decay) for 180 full-batch steps, and the MLP uses Adam (10^{-2} learning rate) for 260 full-batch steps. Both selected heads are random forests with 300 trees, unrestricted depth, minimum leaf size 2, balanced subsampling, and fixed seeds; they are refit on all eligible calibration rows and frozen. Because the selected controller is a tree ensemble, we specify its capacity as two 300-tree forests rather than a neural-network parameter count. It has no gradient optimizer or learning-rate schedule, and no parameter of an editing backbone is optimized.

Table 11: Feature ablation over eight fixed retraining seeds.

Features	Mean Succ. [min, max]
Full (36-D)	0.8481 [0.8443, 0.8519]
No dataset ID	0.8471 [0.8433, 0.8509]
Deployment core (21-D)	0.8437 [0.8405, 0.8490]

The eight-seed feature ablation in Table 11 separates the retained 36-D controller from robustness variants. Removing dataset identity changes mean success only slightly, while the 21-D deployment core retains most of the full controller's success using only instruction and source-image features. The 21-D result is an ablation and is not substituted for the 36-D main policy.

Oracle labeling rule. Within the locked supervision family, candidate actions are ordered from weaker anchors to stronger promoted modes, with $\mathrm{rank}(Z0) = \mathrm{rank}(Z1) = 0$, $\mathrm{rank}(R0) = 1$, $\mathrm{rank}(R1) = 2$, and $\mathrm{rank}(F0) = 3$. Among adequate actions, oracle labeling first prefers lower-rank interventions, followed by lower latency and VRAM, before rewarding additional quality. The following rule is the retained tie-broken realization of the oracle ordering in the main text:

$$z^\star(x,q) = \operatorname*{arg\,lexmin}_{z\in\mathcal{Z}:\, A_d(x,q,z)=1} \big(\mathrm{rank}(z),\, \tau(z),\, \nu(z),\, -Q_d(x,q,z)\big). \tag{10}$$

This is the rule used to derive the offline gate and temporal targets from the sweep summaries.

Temporal control schedules. The promoted temporal labels correspond to fixed reduced-control schedules over the denoising horizon, with $R0$ and $R1$ representing progressively stronger retained control and $F0$ denoting full retained control. All controlled branches use Stable Diffusion v1.5 (Rombach et al., 2022) with the Canny ControlNet and Euler ancestral scheduling; the source image is resized to 512×512, and the Canny condition uses thresholds 100/200. The retained registry is summarized below and remains fixed for held-out evaluation.

Table 12: Retained temporal intervention settings. The control window is normalized over the denoising trajectory.

Budget	Endpoint	Steps	Strength	Text cfg	Control scale	Control window
$Z0$	IP2P	10	–	6.00	–	–
$Z1$	IP2P	20	–	7.50	–	–
$R0$	Canny-CN	14	0.80	6.25	0.55	$[0.00, 0.30]$
$R1$	Canny-CN	20	0.80	7.00	0.80	$[0.00, 0.60]$
$F0$	Canny-CN	20	0.80	7.50	1.00	$[0.00, 1.00]$

$Z1$ is the formal zero-control IP2P anchor. $R0$ uses a short front-loaded schedule, $R1$ retains a longer early-to-mid schedule, and $F0$ keeps control for the full denoising trajectory.

Deterministic spatial refinement. The retained FO refinement uses the deterministic `hybrid_saliency_tight` gate. It first computes a source-image saliency map by combining normalized gradient magnitude, saturation, and value statistics under a single fixed configuration shared across all datasets and samples, followed by Gaussian smoothing. Instruction tokens provide lightweight priors: global/background/style tokens request broader support, while left/right/top/bottom tokens create multiplicative directional ramps. The target coverage is budget dependent, with broader coverage for $F0$ and tighter coverage for reduced-control budgets; directional requests combine a directional box with a sparse saliency support, while non-directional local requests use only the tight sparse saliency support. The sparse support keeps the top saliency mass under the target coverage, applies 3×3 opening/closing, keeps the largest connected components, and falls back to a small directional box if the support is nearly empty. At inference, this mask blends the promoted temporal output with the source image, so changes outside the support are copied from the source. The refinement uses only deployment-time observables and never consumes target images, oracle masks, or post-hoc quality signals.

Table 14: Endpoint sensitivity under oracle-threshold and rank perturbations.

Variant	Succ.	Pres.	Comp.	Lat. (s)
Base	0.6534	16.6557	22.4225	0.4844
Relaxed thr.	0.6439	16.5471	22.4372	0.4570
Strict thr.	0.6629	16.6477	22.3931	0.5049
Strict pres.	0.6619	16.6775	22.4264	0.5017
Rank $R1 = R0$	0.6524	16.6583	22.4218	0.4845
Rank $F0 = R1$	0.6590	16.6834	22.4353	0.4865
Comp. $\succ$ lat.	0.6534	16.6557	22.4225	0.4844

A.2 Additional Diagnostics

Table 13: Routing robustness on 192 held-out samples with four fixed controller seeds.

Dataset	Paraphrase agreement	Mismatch response
Overall	98.70%	94.64%
MagicBrush	97.66%	90.06%
PIE-Bench++	99.74%	99.22%

Table 13 evaluates 192 held-out samples, 96 per benchmark, under four fixed controller seeds. Rule-generated non-verbatim paraphrases preserve the complete route in 98.70% of cases. The mismatch response combines a calibrated paired probability diagnostic with route changes and serves as a negative control against a controller that ignores the instruction.

Table 15: Detailed matched-budget allocation results with frozen modern endpoints. Learned allocation and temporal top-k use the same strong-endpoint call budget.

Endpoint pair	Comparator	ΔSuccess [95% CI]	ΔComp. [95% CI]
IP2P $\rightarrow$ SD3	TO (0% SD3)	$+0.1339$ [0.1130, 0.1548]	$+2.3862$ [2.0757, 2.7049]
IP2P $\rightarrow$ SD3	Temporal top-k	$+0.0399$ [0.0247, 0.0551]	$+0.5662$ [0.3024, 0.8398]
ControlNet $\rightarrow$ SD3	TO (0% SD3)	$+0.0731$ [0.0560, 0.0912]	$+2.3406$ [2.0609, 2.6255]
ControlNet $\rightarrow$ SD3	Temporal top-k	$+0.0163$ [0.0109, 0.0227]	$+0.2824$ [0.1121, 0.4598]
IP2P $\rightarrow$ CosXL	TO (0% CosXL)	$+0.0484$ [0.0266, 0.0693]	$+1.4905$ [1.2418, 1.7508]
IP2P $\rightarrow$ CosXL	Temporal top-k	$+0.0380$ [0.0237, 0.0532]	$+0.4624$ [0.2357, 0.6983]
ControlNet $\rightarrow$ CosXL	TO (0% CosXL)	$+0.0203$ [0.0095, 0.0332]	$+1.4449$ [1.2221, 1.6854]
ControlNet $\rightarrow$ CosXL	Temporal top-k	$+0.0104$ [0.0028, 0.0228]	$+0.2436$ [0.1772, 0.3124]

Table 16: Scope of the three complementary evaluations. The modern-editor routing experiment is not presented as a complete three-stage transfer.

Evaluation	Stage I	Stage II	Stage III	Evidence scope
Original FO	Yes	Yes	Yes	Full three-stage evaluation on the original stack
SD3/CosXL endpoint pairs	Yes	No	No	Routing portability under a matched endpoint budget
Within-SD3 localized action	Yes	Coupled	Yes	Coupled localized-action evidence

Table 17: Additional complete-set spatial diagnostics. Mask coverage is summarized among spatial-active samples. MagicBrush has no official category taxonomy.

Analysis	MagicBrush	PIE-Bench++
Categories with positive strict outside-mask ΔPres.	–	5/9
Categories with positive ΔSuccess	–	6/10
Mask coverage, median [IQR]	11.48% [7.91%, 13.71%]	15.12% [11.10%, 18.02%]

Table 18: Detailed paired FO–TO analysis on all 528 held-out MagicBrush samples. Confidence intervals use 20,000 paired-bootstrap resamples with seed 20260724.

Metric	TO	FO	Difference [95% CI]
Target score	19.3791	22.1084	+2.7293 [2.2956, 3.1753]
Outside-mask preservation	18.6769	22.1484	+3.4715 [2.6181, 4.4326]
Composite	19.1684	22.1204	+2.9520 [2.4396, 3.5191]
Success	0.6231	0.7822	+0.1591 [0.1288, 0.1913]

Table 19: Complete decision-quality diagnostics for the retained controller.

Signal	Value
Multi-seed FO–TO stability	
ΔSucc. CI_{95}	0.0852 [0.0762, 0.0950]
ΔComp. CI_{95}	1.4777 [1.3656, 1.5937]
ΔPres. CI_{95}	2.6106 [2.3632, 2.8658]
ΔLat. CI_{95}	0.0062 [0.0054, 0.0070]
Matched-budget and gate quality	
Learned vs top-k ΔSucc. CI_{95}	0.0247 [0.0133, 0.0361]
Learned vs best-random ΔSucc. CI_{95}	0.0133 [0.0028, 0.0247]
Learned vs Rule-3Stage ΔSucc. CI_{95}	0.0617 [0.0380, 0.0855]
Learned vs Rule-3Stage ΔComp. CI_{95}	0.4122 [0.2666, 0.5624]
Gate AURC gap vs temporal-F0	-0.0111
Gate ECE raw → Platt	0.2567 → 0.0104
Oracle and transition diagnostics	
Oracle exact-match rate	56.3%
Complete-set adequacy repair / new failures	92 / 3

Table 20: Held-out head-family comparison for the observable-only controller.

Setting	Gate	Temp.	Succ.	Pres.	Comp.	Lat. (s)
linear	linear	linear	0.6448	16.7231	22.2884	0.5276
mlp	mlp	mlp	0.6391	16.6196	22.2456	0.4970
rf	rf	rf	0.6534	16.6557	22.4225	0.4844
auto	rf	rf	0.6534	16.6557	22.4225	0.4844

Table 21: Instruction-complexity generalization table under the retained three-band partition, reporting the gain of the promoted full method over the promoted temporal-only method in each band.

Band	N	ΔSucc.	ΔPres.	ΔComp.
Low	230	+0.070	+2.409	+1.393
Mid	556	+0.092	+2.160	+1.683
High	267	+0.094	+4.341	+2.279

A.3 Compute Environment

All experiments were run on a single NVIDIA RTX 4090 GPU with 24GB VRAM. The software environment used Python 3.12.6 and PyTorch 2.11.0.dev+cu126 with CUDA 12.6 support. End-to-end latency and peak VRAM were measured on this setup and are reported in the benchmark tables. The retained run ledger accounts for at least 5.4079 GPU-hours: 2.7204 for reproduced external baselines, 1.2362 for main policy benchmarks, and 1.4513 for selection and promotion sweeps.

A.4 Assets, Licenses, and Human Evaluation

Existing assets and licenses. We use public research releases and do not redistribute raw external datasets, target images, or model checkpoints in the anonymous supplement. Table 22 summarizes the public license labels on the release pages used for this submission; users of the released scripts must obtain each asset from its original source and comply with the corresponding terms.

Table 22: Existing assets used by the experiments and the public license labels checked for this submission.

Asset	Role	Public license or terms label
MagicBrush (Zhang et al., 2023a)	Held-out benchmark	CC-BY 4.0
PIE_Bench_pp (Huang et al., 2024a)	Held-out benchmark	CC-BY-SA 4.0
InstructPix2Pix (Brooks et al., 2023)	Zero-control editor and baseline	MIT model-card label; Stable-Diffusion-derived terms may apply
Stable Diffusion v1.5 (Rombach et al., 2022)	Controlled-branch base checkpoint	CreativeML OpenRAIL-M
ControlNet Canny (Zhang et al., 2023b)	Canny control backbone	OpenRAIL / CreativeML OpenRAIL-M release terms
InstructDiffusion (Geng et al., 2024)	Reproduced external baseline	MIT code license; derived-checkpoint restrictions may apply
UltraEdit SD3 (Zhao et al., 2024)	Reproduced external baseline	Dataset label CC-BY 4.0; model card lists license as other / more information needed

Human-preference protocol. The preference study used 500 comparison instances and five external annotators, with every comparison independently rated by all five. Each interface presented the same source image and edit request with candidates labeled Output A/B. The compared method pair was identified, while its randomized mapping to A/B was hidden and fixed by seed 20260411. Annotators recorded A, B, Tie, or Invalid separately for instruction success, preservation of irrelevant content, and overall quality. In aggregation, a Tie contributes 0.5 vote to each output, whereas Invalid responses are excluded from the denominator for that criterion. This yields 2,500 individual annotations per criterion before Invalid exclusions. No personally identifying information was collected or reported.

B Limitations

Our conclusions are scoped to the stated benchmark input contract, fixed intervention family, and two evaluation datasets. The 36 scalar features favor low overhead and auditability but include dataset indicators and optional prompt, mask-geometry, and category metadata; the fitted policy can therefore exploit dataset regularities and may be brittle to paraphrases, unusual syntax, or settings where such metadata is absent. The asymmetric cross-dataset result shows that it is not dataset invariant. Likewise, the deterministic spatial rule uses saliency and simple directional terms rather than compositional spatial reasoning, so ambiguous references, multiple candidate targets, conflicting directions, and global structural edits can yield incomplete or misplaced support. Endpoint-pair experiments test reuse of the allocation contract, not unchanged transfer of one fitted policy to every SDXL- or diffusion-transformer editor; new action families may require new calibration and fitting. Stronger control can still leak after escalation. More reliable localization can also facilitate misleading image manipulation, so deployment should preserve provenance, consent, disclosure, and the underlying editor's safety controls.