LEARNING MINIMAL SUFFICIENT VISUAL INTERVENTION FOR INSTRUCTION-BASED IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in instruction-based image editing have enabled high-quality edits from source images and language instructions. However, the visual intervention needed to translate user intent into executable editing conditions typically remains implicit, leaving the intended change boundary underspecified and potentially leading to incomplete edits or unintended modifications to unrelated content. In this work, we formulate instruction-based image editing as learning a *minimal sufficient visual intervention* (MSVI) that fulfills the instruction while minimizing changes to instruction-invariant visual content. Specifically, we first use a multimodal large language model as an edit planner to predict a structured edit specification encoding the intervention regime, visual transition, semantic actions, and spatial layout. Based on the predicted specification, a deterministic compiler then invokes SAM 3 only when localized grounding is required. The resulting outputs form the intervention condition supplied to a flow-matching editor. To supervise the planner and the editor, we further construct MSVI-13K with annotated structured edit specifications. The annotations provide prediction targets for the planner and determine the intervention conditions used to train the editor with conditional flow matching and a source-copy suppression loss. Extensive experimental results on MSVI-13K and two external benchmarks show that our method matches or outperforms state-of-the-art approaches, supporting explicit visual intervention learning for balancing instruction fulfillment and content preservation. Code is available at https://anonymous.4open.science/r/MSVI/.

1 INTRODUCTION

Instruction-based image editing aims to modify a source image according to a language instruction while preserving content unrelated to the requested change (Brooks et al., 2023; Yu et al., 2025). However, language instructions typically specify expected deliverables without defining the corresponding visual interventions, *i.e.*, what to change, how far the change should extend, and what to preserve (Geng et al., 2024; He et al., 2025; Huang et al., 2024; Ji et al., 2025). For example, as shown in Figure 1(a), the instruction "*Change the cat not in the mirror to a red panda*" requires altering the target cat's identity, silhouette, and texture while preserving the cat in the mirror, the floor, the wall, and the viewpoint. Other requests may instead warrant image-wide adaptation or a combination of focal and holistic changes (Ye et al., 2026; Yu et al., 2025; He et al., 2025; Huang et al., 2025a; Parihar et al., 2026). Therefore, the primary challenge is to translate an instruction into a scoped visual intervention that realizes the requested edit without introducing unintended changes.

Existing image editing methods broadly follow two paradigms for handling the required decisions, *i.e.*, direct generation and reasoning- or planning-based generation (Huang et al., 2025a). Direct editors map a source image and an instruction to an edited output, as illustrated in Figure 1(b), leaving the edit scope and preservation requirements implicit within the image generator (Brooks et al., 2023; Sheynin et al., 2024; Yu et al., 2025). Recent reasoning- and planning-based approaches make edit intent more explicit through visual reasoning, instruction decomposition, editing blueprints, or grounded sub-instructions before generation (He et al., 2025; Qu et al., 2025; Zeng et al., 2026; Yeh et al., 2026; Geng et al., 2024). However, the boundary of the required visual intervention generally remains a byproduct rather than an explicit prediction objective (Qu et al., 2025; Huang et al., 2025a; Cheng et al., 2021; Zarei et al., 2026; Jia et al., 2026; Carion et al., 2025). The required semantic

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

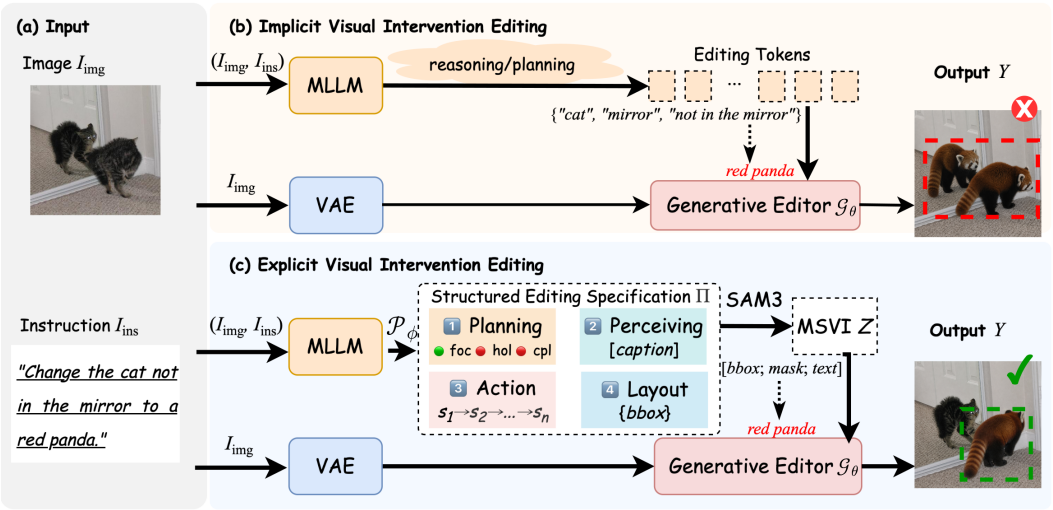

Figure 1: Motivation for MSVI. (a) The source image and instruction form the joint input. (b) An implicit visual intervention converts MLLM reasoning and planning into editing tokens without explicitly representing the modification scope. (c) MSVI predicts an explicit structured intervention and uses SAM 3 to ground editing with bounding-box, mask, and semantic conditions. Dashed boxes in the output images mark edited regions. "foc", "hol", and "cpl" denote FOCAL, HOLISTIC, and COUPLED editing types, respectively. Dashed boxes in the output images mark edited regions.

transition, intervention scope, edit layout, and source-image evidence are therefore typically not jointly represented as an explicit predicted variable.

To address this challenge, we reformulate instruction-based image editing as learning a *minimal sufficient visual intervention* (MSVI). Given a source image I_{img} and an instruction I_{ins}, a visual intervention $\mathbf{Z}$ is *sufficient* if conditioning the editor on $\mathbf{Z}$ produces an output Y that fulfills the requested change (Ji et al., 2025; Soatto & Chiuso, 2014). Among sufficient interventions, $\mathbf{Z}$ is *minimal* if it induces the least change to content that should remain invariant under I_{ins}. Sufficiency thus rules out incomplete edits, whereas minimality limits unintended changes (Qian et al., 2025; Kim et al., 2025). Minimality therefore asks whether each change is necessary to fulfill I_{ins}, rather than whether the edited region or the number of operations is small. Consequently, a holistic or coupled intervention can still be minimal when the instruction requires changes to illumination, style, or scene-wide consistency (Zhang et al., 2025b; He et al., 2025; Huang et al., 2024). As illustrated in Figure 1(c), MSVI makes both minimality and sufficiency explicit.

To instantiate MSVI, we propose a three-stage approach comprising structured edit specification prediction, regime-adaptive intervention, and conditional flow-matching generation. *First*, a multimodal large language model is used as an edit planner $\mathcal{P}_\phi$ to predict a structured edit specification (SES) $\Pi = (q, o, s, i)$ with four components. Within Π, `Planning` identifies the edit regime $q \in \{\text{FOCAL}, \text{HOLISTIC}, \text{COUPLED}\}$ for localized, image-wide, or combined changes; `Perceiving` records the edit-relevant source and target states $o = (o_s, o_t)$; `Action` decomposes the instruction into ordered substeps $s = (s_1, \ldots, s_n)$ that transform the source image into the edited image; and `Layout` i describes the spatial arrangement of relevant entities. *Then*, the regime-adaptive intervention compiler maps (I_{img}, Π) to the visual intervention $\mathbf{Z}$ by integrating the intervention signals in Π with regime-specific grounding outputs and relevant source-image evidence. Guided by q, the compiler invokes SAM 3 (Carion et al., 2025) only when localized grounding is required, whereas holistic edits bypass object masking. *Finally*, $\mathbf{Z}$ serves as the condition for the MMDiT editor $\mathcal{G}_\theta$, which generates Y through conditional flow matching (Lipman et al., 2022). This sequence separates deciding what should change from generating the result, so the editor receives an explicit condition for both instruction fulfillment and content preservation.

To supervise both SES and the conditional flow model, we further construct MSVI-13K by augmenting over 13,000 source-instruction-target triplets drawn from AnyEdit (Yu et al., 2025), CompBench (Jia et al., 2026), and the synthetic complex-instruction data released by SmartEdit (Huang et al., 2024). Each triplet is annotated with a structured edit specification Π under the three-regime taxonomy. Extensive experimental results demonstrate that our approach matches or outperforms

state-of-the-art methods across four evaluation dimensions while remaining competitive in model size. Compared with baselines, our approach averagely improves instruction fulfillment by 71.7%, invariant preservation by 1.4%, intervention accuracy by 11.7%, and visual quality by 1.415.

2 Related Work

Instruction-Based Image Editing. Early language-conditioned GANs enabled text-relevant attribute manipulation while preserving unrelated content, but were largely limited to restricted domains and modest editing diversity (Nam et al., 2018; Li et al., 2020). To address these problems, diffusion models improved generation quality and controllability through source noising, attention control, and image-specific adaptation (Meng et al., 2022; Hertz et al., 2022), yet often relied on target prompts, inversion, or per-image optimization (Huang et al., 2025a; Kulikov et al., 2025). Paired instruction-editing methods therefore learned direct source–instruction–target mappings (Brooks et al., 2023; Zhang et al., 2023). Recent generalist editors further broaden editing coverage through larger datasets, diffusion-transformer backbones, multi-task training, and expert routing (Yu et al., 2025; Liu et al., 2025; Lin et al., 2025; Wu et al., 2025a; Zhang et al., 2026; Wu et al., 2026). In parallel, reasoning-oriented editors augment this mapping with intermediate reasoning and semantic-spatial guidance (He et al., 2025; Fang et al., 2025; Zeng et al., 2026; Yin et al., 2026). Across both paradigms, the required edit scope and preservation constraints are usually left implicit in the generator, which may blur the boundary between intended and unintended changes. In this work, MSVI instead makes these decisions explicit through a structured edit specification that records the edit regime, source and target states, ordered editing substeps, and spatial layout.

Reasoning- and Planning-Based Image Editing. Existing approaches expose different parts of the editing process before generation. MLLM-based methods derive richer visual-language features, learn from reasoning supervision, decompose instructions, couple reasoning with the editor, or predict target states (Fu et al., 2023; Jin et al., 2024; Huang et al., 2024; He et al., 2025; Zhang et al., 2025a; Zeng et al., 2026). For example, GoT (Fang et al., 2025) makes semantic-spatial reasoning explicit before generation. X-Planner (Yeh et al., 2026) executes grounded sub-instructions sequentially, whereas MSVI uses a single editor invocation, avoiding repeated generation and error propagation between edits. RePlan (Qu et al., 2025) provides region-aligned guidance for complex instructions. Planning-based image-editing methods instead localize changes through masks, text-and-shape guidance, grounded boxes, or general conditional-generation interfaces (Li et al., 2023; Geng et al., 2024; Parihar et al., 2026; Zarei et al., 2026). Preservation-oriented methods retain source content by reusing inverted latents, source features, or attention maps (Ju et al., 2024; Cao et al., 2023; Parmar et al., 2023). Open-vocabulary detectors and promptable segmenters further connect language to image regions, providing spatial support for local control (Liu et al., 2024; Kirillov et al., 2023; Carion et al., 2025). These methods expose semantic or spatial cues to the generator, but each cue captures only part of the required intervention. In particular, these signals do not by themselves specify the required extent of change across localized, image-wide, and combined edits while also identifying what should change and what must be preserved. MSVI addresses this gap by compiling the predicted SES into an explicit visual intervention. The SES predicts the edit regime, which then determines the focal and contextual support supplied to a shared MMDiT flow executor (Lipman et al., 2022; Esser et al., 2024).

3 Minimal Sufficient Visual Intervention (MSVI)

3.1 Overview

MSVI aims to make each edit *sufficient* for the requested transition and *minimal* in its effect on other content through explicit planning and grounding before image synthesis (Brooks et al., 2023; Yu et al., 2025). As shown in Figure 2, given a source image I_{img} and an instruction I_{ins}, our MSVI comprises three stages. Specifically, an MLLM planner predicts an edit specification Π, a regime-adaptive intervention compiler converts the source image and Π into a structured visual intervention $\mathbf{Z}$ (Carion et al., 2025; Liu et al., 2024), and an MMDiT (Esser et al., 2024) flow executor conditioned on $\mathbf{Z}$ then generates the edited image Y. The overall editing pipeline can be expressed as:

$$(I_{\text{img}}, I_{\text{ins}}) \xrightarrow{\mathcal{P}_\phi} \Pi, \qquad (I_{\text{img}}, \Pi) \xrightarrow{\Gamma} \mathbf{Z}, \qquad \mathbf{Z} \xrightarrow{\mathcal{G}_\theta} Y, \tag{1}$$

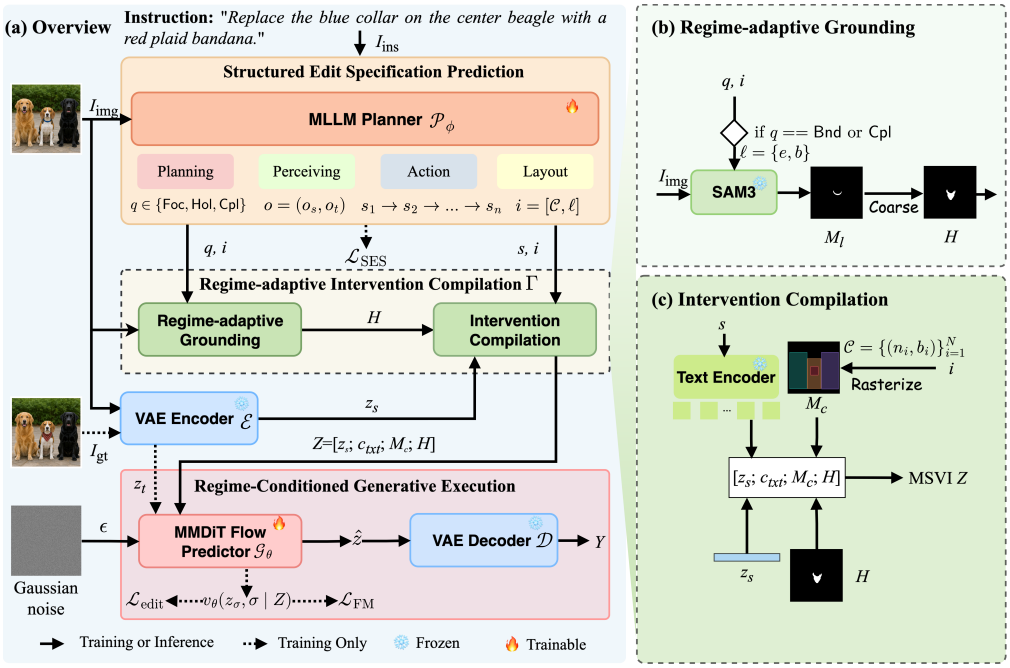

Figure 2: Overview of MSVI. Given a source image I_{img} and an instruction I_{ins}, an MLLM planner $\mathcal{P}_\phi$ first predicts a structured edit specification Π with four components. Regime-adaptive intervention compilation Γ then converts (I_{img}, Π) into the structured intervention $\mathbf{Z}$. The MMDiT flow executor $\mathcal{G}_\theta$ further generates the edited output Y conditioned on $\mathbf{Z}$.

where $\mathcal{P}_\phi$, Γ, and $\mathcal{G}_\theta$ denote the MLLM planner, the regime-adaptive intervention compiler, and the intervention-conditioned flow executor, respectively.

3.2 Structured Edit Specification Prediction

As editing instructions often leave the required transition and preservation scope implicit (Qu et al., 2025; Huang et al., 2025a; Cheng et al., 2021; Carion et al., 2025), we predict an SES Π that specifies both for a *sufficient* and *minimal* intervention. Given $(I_{\text{img}}, I_{\text{ins}})$, an MLLM planner $\mathcal{P}_\phi$ autoregressively predicts four components, which can be formulated as:

$$\Pi = (\underbrace{q}_{\texttt{Planning}}, \underbrace{o}_{\texttt{Perceiving}}, \underbrace{s}_{\texttt{Action}}, \underbrace{i}_{\texttt{Layout}}). \tag{2}$$

where the `Planning` $q \in \{\text{FOCAL}, \text{HOLISTIC}, \text{COUPLED}\}$ predicts the intervention regime for localized, image-wide, or combined changes, respectively. By resolving the intervention scope, the regime determines whether localized grounding is invoked. `Perceiving` provides `Action` with the transition endpoints by estimating the edit-relevant state pair $o = (o_s, o_t)$, where o_s summarizes the source state and o_t describes the target state inferred from $(I_{\text{img}}, I_{\text{ins}})$. Given these states, `Action` decomposes the complex editing instruction into an ordered sequence of simple semantic substeps $s = (s_1, \ldots, s_n)$, which collectively transform o_s into o_t. `Layout` specifies the global target-image layout and the spatial constraints for local edits, expressed as:

$$i = [\mathcal{C}, \ell], \tag{3}$$

where $\mathcal{C} = \{(n_j, b_j)\}_{j=1}^N$ specifies the whole-image context by pairing edit-relevant contextual entities n_j with their normalized bounding boxes b_j, while $\ell = (e, b)$ contains a referring expression e and a bounding box b for focal grounding. The localized control ℓ is provided only for FOCAL and COUPLED specifications.

3.3 Regime-Adaptive Intervention Compilation

To construct $\mathbf{Z}$ as a *minimal* and *sufficient* visual intervention from the source image and predicted Π, the compiler uses regime-dependent localization, context, and synthesis supports (Ju et al., 2024; Cao

et al., 2023; Parmar et al., 2023), beginning with regime-specific grounding:

$$M_l = \begin{cases} \text{SAM3}(I_{\text{img}}; e, b), & q \in \{\text{FOCAL}, \text{COUPLED}\}, \\ \mathbf{0}, & q = \text{HOLISTIC}. \end{cases} \tag{4}$$

For FOCAL and COUPLED specifications, the referring expression e and bounding box b in ℓ guide SAM 3 (Carion et al., 2025) to produce the localization mask M_l. HOLISTIC specifications set $M_l = \mathbf{0}$ to avoid imposing an object boundary on a scene-wide edit. The compiler then forms the binary contextual support M_c by filling the union of the boxes in $\mathcal{C}$, which marks the planned locations of edit-relevant entities in the edited scene, providing a global layout constraint. Because M_c encodes locations rather than semantic intent, a text encoder maps the ordered action sequence s to semantic tokens c_{txt}. The localization mask may be too restrictive for edits that alter object silhouettes or nearby boundaries. The focal synthesis support H is therefore a padded axis-aligned box around the largest connected component of M_l, clipped to the image boundary. HOLISTIC specifications set $H = \mathbf{0}$. Finally, a frozen VAE encodes the source image as $z_s = \mathcal{E}(I_{\text{img}})$. After resizing M_c and H to the latent resolution, Γ constructs the executor condition, which can be expressed as:

$$\mathbf{Z} = [z_s; c_{\text{txt}}; M_c; H], \tag{5}$$

where the brackets enumerate the four fields rather than feature concatenation. The compiled condition $\mathbf{Z}$ is then passed to the intervention-conditioned flow executor (Esser et al., 2024).

3.4 TRAINING AND INFERENCE

Intervention-Conditioned Flow Execution. Under the compiled condition $\mathbf{Z}$, we use an MMDiT-based conditional flow model (Lipman et al., 2022; Esser et al., 2024) to generate the edited image Y in this stage. The executor predicts the conditional velocity field $v_\theta(z_\sigma, \sigma \mid \mathbf{Z})$, whose training strategy and inference integration are defined below.

Training. To prevent training collapse, our training strategy consists of two sequential optimization phases (Xia et al., 2024; Huang et al., 2025b). In the first phase, we freeze the executor and train the planner under teacher forcing. Given an annotated specification Π^{gt} with response tokens $\Pi^{\text{gt}}_{1:L}$, the response-only objective is formulated as:

$$\mathcal{L}_{\text{SES}} = -\sum_{j=1}^{L} \log p_\phi \left(\Pi^{\text{gt}}_j \mid I_{\text{img}}, I_{\text{ins}}, \Pi^{\text{gt}}_{<j} \right), \tag{6}$$

where j indexes the tokens in the ground-truth SES sequence Π^{gt}, and L denotes its total sequence length. In the second phase, we freeze the trained planner and VAE (Kingma & Welling, 2013), and optimize the executor. For each triplet $(I_{\text{img}}, I_{\text{ins}}, I_{\text{gt}})$, the annotated specification Π^{gt} is passed to Γ to construct $\mathbf{Z}$ as in Sec. 3.3. The executor is therefore trained with ground-truth rather than predicted specifications. The frozen VAE maps the paired target image I_{gt} to $z_t = \mathcal{E}(I_{gt})$. Given Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and flow time $\sigma \sim \mathcal{U}[0, 1]$, the linear path z_σ and target velocity v^* are expressed as:

$$z_\sigma = (1 - \sigma) z_t + \sigma \epsilon, \qquad v^* = \epsilon - z_t. \tag{7}$$

The flow-matching objective is expressed as:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{I_{\text{img}}, I_{\text{ins}}, I_{\text{gt}}, \sigma, \epsilon} \left[\| v_\theta(z_\sigma, \sigma \mid \mathbf{Z}) - (\epsilon - z_t) \|_2^2 \right]. \tag{8}$$

For spatially sparse edits, unchanged regions can dominate $\mathcal{L}_{\text{FM}}$, causing flow matching to favor source copying and yield an *insufficient* edit. We therefore use the source-copy suppression loss (McDougall et al., 2023), which penalizes the no-edit solution by matching the predicted clean endpoint to the paired target within the compiled intervention support. The loss function can be formulated as:

$$\widetilde{z}_0 = z_\sigma - \sigma v_\theta(z_\sigma, \sigma \mid \mathbf{Z}), \qquad \mathcal{L}_{\text{edit}} = \frac{\| (M_c \vee H) \odot (\widetilde{z}_0 - z_t) \|_2^2}{\| M_c \vee H \|_1 + \varepsilon}, \tag{9}$$

where $M_c \vee H$ denotes the union of M_c and H. The overall training objective can be expressed as:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{SES}}}_{\substack{\text{Fixed in} \\ \text{the second phase}}} + \underbrace{\mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{edit}}}_{\substack{\text{Fixed in} \\ \text{the first phase}}}. \tag{10}$$

We fully understand introducing weighting coefficients for sub-loss terms may further improve the performance. However, systematic loss-weight tuning is not the focus of this work.

Inference. In inference, the planner $\mathcal{P}_\phi$ first predicts the edit specification Π, and the compiler Γ converts (I_{img}, Π) into the condition $\mathbf{Z}$ following Equation 1. The executor $\mathcal{G}_\theta$ then initializes z_1 as Gaussian noise and iteratively updates it according to the learned conditional velocity field $v_\theta(z_\sigma, \sigma \mid \mathbf{Z})$. As σ decreases from 1 to 0, these updates transform the noisy latent into the clean endpoint z_0. Finally, the frozen VAE decoder maps z_0 to the edited output $Y = \mathcal{D}(z_0)$.

4 EXPERIMENTS

Datasets. Existing instruction-editing datasets (Yu et al., 2025; Jia et al., 2026; Huang et al., 2024) provide source-instruction-target supervision but lack structured planning annotations as discussed in Sec. 1. To address this gap, we therefore construct MSVI-13K with 11,825 training and 1,312 disjoint testing examples across FOCAL, HOLISTIC, and COUPLED editing types. Due to page limit, dataset construction details are provided in the supplementary material. Besides, for complementary cross-dataset evaluation, we use Reason-Edit (Huang et al., 2024) for visual-attribute and world-knowledge reasoning, and PIE-Bench (Ju et al., 2024) for non-target preservation.

Baselines. The selected baseline methods cover representative image editing approaches and include recent editors with strong reported performance. We divide direct editors into direct instruction-tuned editors (*i.e.*, InstructPix2Pix (Brooks et al., 2023), MagicBrush (Zhang et al., 2023), and AnySD (Yu et al., 2025)) and direct generalist editors (*i.e.*, PixWizard (Lin et al., 2025), ICEdit-MoE (Zhang et al., 2026), Step1X-Edit (Liu et al., 2025), OmniGen2 (Wu et al., 2026), and Qwen-Image-Edit (Wu et al., 2025a)). The remaining groups are reasoning-based editors (*i.e.*, MGIE (Fu et al., 2023) and SmartEdit (Huang et al., 2024)) and planning-based editors (*i.e.*, GoT (Fang et al., 2025) and RePlan (Qu et al., 2025)). For a fair experimental comparison, we use their official inference settings.

Evaluation Metrics. Following established image-editing evaluation protocols (Huang et al., 2024; Ju et al., 2024; Wu et al., 2025b), we report EditReward-IF for instruction fulfillment, InvPres and BG-LPIPS for invariant and background preservation, Intervention-F1 (Int-F1) for intervention accuracy, and EditReward-VQ for visual quality. Int-F1 measures intervention accuracy as the F1 score between the source-to-output perceptual change mask and the annotated intervention region. To capture source-to-target progress within the intended region, we introduce **Edit Target Proximity (ETP)**, the masked perceptual proximity to the paired target normalized by the source-to-target distance. Let $M_e^\star$ denote the annotated intervention support and $D_M(\cdot,\cdot)$ denote the masked LPIPS distance (Zhang et al., 2018). ETP can be obtained by:

$$\text{ETP} = \frac{1}{1+\bar{r}}, \qquad r = \frac{D_{M_e^\star}(Y, I_t)}{D_{M_e^\star}(I_s, I_t) + \epsilon}, \tag{11}$$

where ϵ is a numerical stabilizer and $\bar{r}$ averages r within each applicable regime and then across regimes. Higher ETP indicates greater target-directed progress and is reported only for paired data. For a fair result comparison, the external benchmarks follow official protocols (Huang et al., 2024; Ju et al., 2024), *i.e.*, BG-PSNR and BG-LPIPS measure pixel-level and perceptual background preservation, while FG-CLIP and CLIP-Edit measure target-region semantic alignment.

Implementation Details. Following the design of GoT (Fang et al., 2025), we instantiate the planner with Qwen3.5-4B (Qwen Team, 2026) and fine-tune it using response-only LoRA (Hu et al., 2021) while freezing its visual backbone. Following the Stable Diffusion 3.5 Medium (Esser et al., 2024) design, we adapt it through LoRA. Executor training uses one NVIDIA RTX PRO 6000 GPU, a global batch size of 64, a learning rate of 5×10^{-6}, 100 warm-up steps, and weight decay 0.01. At inference, SAM 3 (Carion et al., 2025) uses a confidence threshold of 0.25. We use 40 flow-matching steps and classifier-free guidance of 5.0.

4.1 COMPARISONS WITH STATE-OF-THE-ART METHODS

Quantitative Results on MSVI-13K. Table 1 compares edit sufficiency, minimality, intervention accuracy, and visual quality on the testing set. MSVI surpasses RePlan (Qu et al., 2025) in edit sufficiency, reaching 1.630 ER-IF and 0.464 ETP. It also achieves the only positive ER-VQ (0.012), with BG-LPIPS (0.150) second only to GoT (Fang et al., 2025). For *intervention accuracy*, MSVI attains an Int-F1 of 0.464, providing evidence of spatial accuracy in the intermediate intervention. Together with edit sufficiency and invariant preservation, this result supports our formulation of a

Method	Params	Sufficiency		Minimality		Int. Acc.	Quality
		ER-IF ↑	ETP ↑	InvPres ↑	BG-LPIPS ↓	Int-F1 ↑	ER-VQ ↑
Direct instruction-tuned editors							
InstructPix2Pix (Brooks et al., 2023)	1.1B	0.251	0.220	0.872	0.506	0.462	-2.430
MagicBrush (Zhang et al., 2023)	1.1B	0.357	0.242	0.878	0.439	0.458	-2.293
AnySD (Yu et al., 2025)	0.9B	0.386	0.262	**0.908**	0.305	0.419	-2.327
Direct generalist editors							
PixWizard (Lin et al., 2025)	2.0B	1.024	0.406	0.904	0.196	0.351	-1.354
ICEdit-MoE (Zhang et al., 2026)	12.0B	1.213	0.350	0.894	0.246	0.361	-0.840
Step1X-Edit (Liu et al., 2025)	12.0B	1.233	0.422	0.886	0.299	0.426	-0.755
OmniGen2 (Wu et al., 2026)	7.0B	1.256	0.418	0.892	0.223	0.350	-0.814
Qwen-Image-Edit (Wu et al., 2025a)	20.0B	1.382	0.407	0.842	0.313	0.458	-0.522
Reasoning-based editors							
MGIE (Fu et al., 2023)	7.0B	0.888	0.370	0.876	0.313	0.407	-1.425
SmartEdit (Huang et al., 2024)	7.0B	0.555	0.318	0.889	0.484	0.463	-1.895
Planning-based editors							
GoT (Fang et al., 2025)	6.0B	1.278	0.438	0.889	**0.116**	0.409	-2.010
RePlan (Klein) (Qu et al., 2025)	4.0B	1.571	0.441	0.877	0.277	0.422	-0.172
MSVI(ours)	6.5B	**1.630**	**0.464**	0.896	0.150	**0.464**	**0.012**

Table 1: Comparison on the MSVI-13K testing set. "ER-IF", "Int-F1" and "ER-VQ" denote EditReward-IF, Intervention-F1 and EditReward-VQ, respectively.

Method	Params	Reason-Edit (Huang et al., 2024)			PIE-Bench (Ju et al., 2024)		
		BG-PSNR ↑	BG-LPIPS ↓	FG-CLIP ↑	BG-PSNR ↑	BG-LPIPS ↓	CLIP-Edit ↑
Direct instruction-tuned editors							
InstructPix2Pix (Brooks et al., 2023)	1.1B	14.128	0.279	22.303	20.817	0.157	21.982
MagicBrush (Zhang et al., 2023)	1.1B	18.522	0.160	21.833	24.378	0.084	21.532
AnySD (Yu et al., 2025)	0.9B	21.684	0.118	21.771	25.922	0.094	21.071
Direct generalist editors							
PixWizard (Lin et al., 2025)	2.0B	23.326	0.086	19.925	24.909	0.085	22.024
ICEdit-MoE (Zhang et al., 2026)	12.0B	22.776	0.085	20.443	25.411	0.096	22.293
Step1X-Edit (Liu et al., 2025)	12.0B	24.877	0.071	22.767	27.017	0.066	23.397
OmniGen2 (Wu et al., 2026)	7.0B	22.939	0.090	22.199	**28.436**	0.055	22.602
Qwen-Image-Edit (Wu et al., 2025a)	20.0B	13.684	0.306	23.096	15.528	0.252	23.484
Reasoning-based editors							
MGIE (Fu et al., 2023)	7.0B	14.335	0.263	20.861	21.711	0.137	21.908
SmartEdit (Huang et al., 2024)	7.0B	23.872	0.076	22.496	20.876	0.168	22.225
Planning-based editors							
GoT (Fang et al., 2025)	6.0B	26.497	0.080	21.375	26.840	0.079	21.776
RePlan (Klein) (Qu et al., 2025)	4.0B	25.113	**0.039**	22.141	27.333	**0.039**	23.238
MSVI(ours)	6.5B	**27.796**	0.041	**23.566**	27.032	**0.039**	**23.488**

Table 2: Results of the cross-benchmark evaluation.

minimal sufficient visual intervention as one that completes the requested transition while limiting unnecessary changes. Direct instruction-tuned editors, especially AnySD (Yu et al., 2025), preserve source invariants but score poorly on instruction fulfillment, indicating under-execution. Among direct generalist editors, Qwen-Image-Edit (Wu et al., 2025a) follows instructions more strongly but has weaker invariant preservation. Reasoning-based SmartEdit (Huang et al., 2024) nearly matches MSVI in Int-F1, but its lower ETP and higher BG-LPIPS show that spatial overlap alone does not ensure the intended transition or background preservation.

Cross-Benchmark Generalization. Table 2 evaluates editing performance transfer by training all models on MSVI-13K and testing them on Reason-Edit and PIE-Bench without benchmark-specific

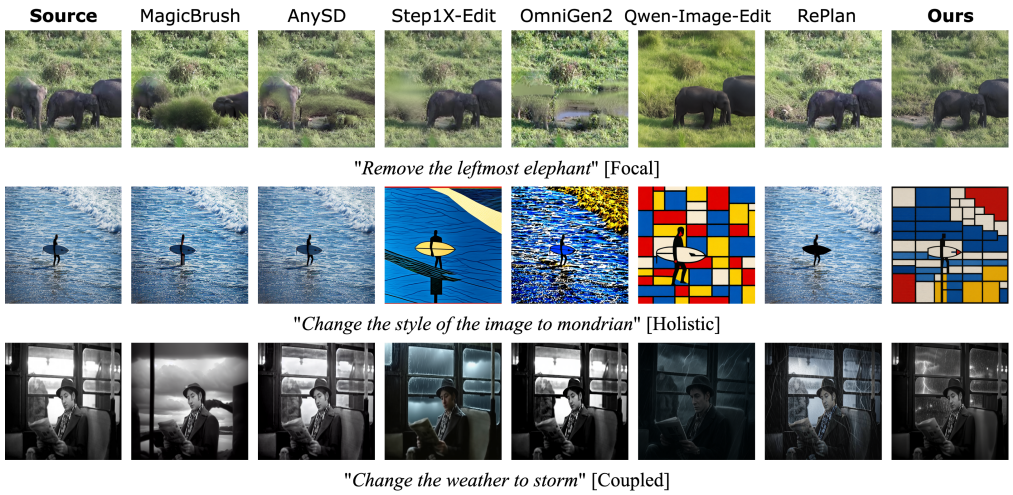

Figure 3: Qualitative comparisons across three intervention regimes.

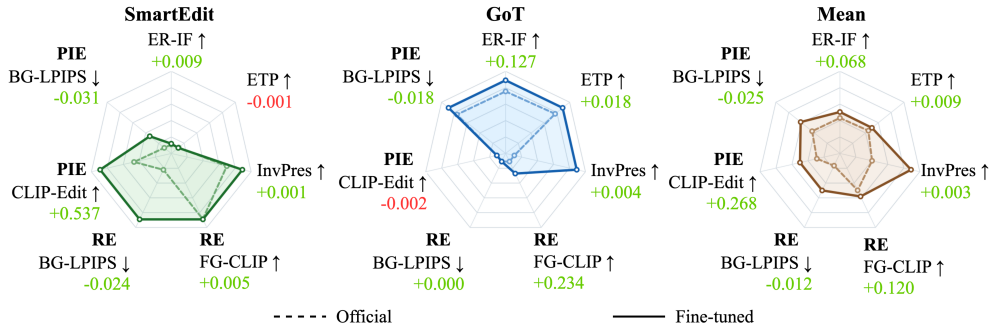

Figure 4: Results on cross-model transfer after fine-tuning on the MSVI-13K dataset.

fine-tuning. On Reason-Edit, MSVI achieves the highest BG-PSNR (27.796) and FG-CLIP (23.566), while its BG-LPIPS (0.041) remains close to RePlan's (Qu et al., 2025) 0.039. RePlan's lower FG-CLIP (22.141) indicates weaker target alignment despite slightly better perceptual background preservation. On PIE-Bench, MSVI ties the best BG-LPIPS (0.039) and reaches the highest CLIP-Edit (Huang et al., 2024) (23.488), but its BG-PSNR (27.032) remains below OmniGen2 (Wu et al., 2026) (28.436). OmniGen2 attains lower CLIP-Edit (22.602), whereas Qwen-Image-Edit (Wu et al., 2025a) reaches 23.484 CLIP-Edit but substantially worse BG-PSNR (15.528) and BG-LPIPS (0.252). MSVI therefore provides the stronger joint semantic alignment and background preservation.

Qualitative Results. Figure 3 visualizes qualitative results across three intervention regimes. Additional results are in the supplementary materials due to page limits. The results show three failure modes motivating MSVI. *First*, when the intended change boundary is underspecified, a focal request can spill into nearby content, *i.e.*, in the elephant-removal example, several baselines alter the surrounding vegetation or the remaining animals, whereas MSVI removes the leftmost elephant while preserving the rest of the scene. *Second*, an intervention can be incomplete even when it preserves the source well. In the holistic example, AnySD (Yu et al., 2025) and RePlan (Qu et al., 2025) largely retain the photographic appearance instead of applying the Mondrian style across the image, while MSVI completes the scene-wide transformation and retains the surfer's layout. *Third*, an editor can satisfy part of the request yet introduce unintended modifications. In the coupled storm example, several outputs change the subject's appearance, or introduce visible artifacts, while MSVI changes the weather while keeping the person and scene recognizable.

4.2 MSVI-13K for Cross-Model Generalization and Multimodal Reasoning

To assess the utility of MSVI-13K, we fine-tune the reasoning-based SmartEdit (Huang et al., 2024) and planning-based GoT (Fang et al., 2025) on MSVI-13K. Figure 4 summarizes their changes

Variant	Sufficiency		Minimality		Int. Acc.	Quality
	ER-IF ↑	ETP ↑	InvPres ↑	BG-LPIPS ↓	Int-F1 ↑	ER-VQ ↑
(a) Structured edit specification prediction						
Direct instruction conditioning	0.682	0.274	0.672	0.179	0.217	-0.009
Unstructured plan (matched length)	0.993	0.284	0.667	0.163	0.312	-0.005
SES w/o `Planning`	1.074	0.420	0.773	0.161	0.405	0.001
SES w/o `Perceiving`	0.997	0.418	0.768	0.162	0.402	0.001
SES w/o `Action`	1.132	0.440	0.821	0.155	0.432	0.008
SES w/o `Layout`	1.021	0.415	0.754	0.165	0.398	0.002
Complete SES (ours)	**1.630**	**0.464**	**0.896**	**0.150**	**0.464**	**0.012**
(b) Regime-adaptive intervention compilation						
w/o localized grounding	1.423	0.437	0.483	0.277	0.408	0.009
w/o contextual support	1.282	0.352	0.311	0.289	0.390	0.010
w/o focal-support expansion	1.589	0.431	0.205	0.337	0.329	0.009
Full compilation (ours)	**1.630**	**0.464**	**0.896**	**0.150**	**0.464**	**0.012**
(c) Source-Copy Suppression Loss (SCSL)						
w/o SCSL	0.503	0.221	**0.906**	**0.033**	0.361	0.012
Full objective (ours)	**1.630**	**0.464**	0.896	0.150	**0.464**	**0.012**

Table 3: Component ablations on the MSVI-13K testing set.

on MSVI-13K, Reason-Edit (Huang et al., 2024), and PIE-Bench (Ju et al., 2024). GoT improves Reason-Edit FG-CLIP by 0.234, suggesting better target-region semantic alignment on tasks requiring visual-attribute and world-knowledge reasoning. This provides evidence of transfer to reasoning-intensive editing beyond the training dataset. SmartEdit improves PIE-Bench CLIP-Edit by 0.537 while reducing BG-LPIPS by 0.031. Stronger alignment with the requested edit thus accompanies better background preservation. These gains suggest that MSVI-13K provides transferable visual reasoning capability and cross-dataset generalization in both editors.

4.3 Ablation and Diagnostic Analysis

Table 3 ablates SES prediction, intervention compilation, and source-copy suppression loss(SCSL). **(a) SES Prediction.** Direct instruction conditioning reduces ER-IF from 1.630 to 0.682. A length-matched unstructured plan recovers only to 0.993 and remains below the complete SES on every metric, attributing the residual gap to typed structure rather than plan length. Removing any SES field also weakens sufficiency and minimality. **(b) Regime-adaptive Intervention Compilation.** Removing M_c produces the largest sufficiency loss, reducing ER-IF from 1.630 to 1.282, consistent with M_c encoding planned post-edit locations. Removing localized grounding affects target proximity less but weakens invariant and background preservation. Replacing H with M_l largely retains ER-IF but reduces Int-F1 from 0.464 to 0.329, indicating that localization alone lacks sufficient synthesis freedom for boundary changes. **(c) SCSL.** Without SCSL, ER-IF falls from 1.630 to 0.503 and ETP from 0.464 to 0.221. SCSL suppresses source-copying no-ops missed by visual quality and restores edit sufficiency. The higher InvPres (0.906) and lower BG-LPIPS (0.033) without SCSL further show that source similarity can mask under-execution and should be assessed jointly with edit sufficiency.

5 Conclusion

In this paper, we formulate instruction-based image editing as learning a minimal sufficient visual intervention through structured edit specification prediction, regime-adaptive intervention compilation, and intervention-conditioned flow execution. We also construct MSVI-13K, a dataset with structured intervention annotations. Across MSVI-13K and the external benchmarks, experiments show that MSVI balances instruction fulfillment, invariant-content preservation, and visual quality. One limitation of our approach is that intervention prediction and execution are trained separately, which may allow specification or grounding errors to propagate during inference. Future work will explore joint, uncertainty-aware optimization and closed-loop refinement, using execution feedback to revise the predicted intervention while preserving the sufficiency and minimality objectives of MSVI.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and readability of the manuscript. We also used a multimodal large language model to generate the structured edit specification annotations for MSVI-13K, for which the dataset construction procedure and annotation prompt are reported in the appendix. All AI-assisted text and generated annotations were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with AI assistance.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pp. 18392–18402, 2023. 1, 3, 6, 7

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 *ICCV*, pp. 22503–22513, 2023. 3, 4

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv*, 2025. 1, 2, 3, 4, 5, 6

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, pp. 15329–15337, 2021. 1, 4

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 3, 5, 6

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv*, 2025. 3, 6, 7, 8

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv*, 2023. 3, 6, 7

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *CVPR*, pp. 12709–12720, 2024. 1, 3

Qingdong He, Xueqin Chen, Chaoyi Wang, Yanjie Pan, Xiaobin Hu, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xiangtai Li, and Jiangning Zhang. Reasoning to edit: Hypothetical instruction-based image editing with visual reasoning. *arXiv*, 2025. 1, 2, 3

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv*, 2022. 3

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *TPAMI*, 47 (6):4409–4437, 2025a. 1, 3, 4

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, pp. 8362–8371, 2024. 1, 2, 3, 6, 7, 8, 9, 14, 19

Zhipeng Huang, Shaobin Zhuang, Canmiao Fu, Binxin Yang, Ying Zhang, Chong Sun, Zhizheng Zhang, Yali Wang, Chen Li, and Zheng-Jun Zha. Wegen: A unified model for interactive multi-modal generation as we chat. In *CVPR*, pp. 23679–23689, 2025b. 5

Liya Ji, Chenyang Qi, and Qifeng Chen. Instruction-based image editing with planning, reasoning, and generation. In *ICCV*, pp. 17506–17515, 2025. 1, 2, 14

Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, et al. Compbench: Benchmarking complex instruction-guided image editing. In *CVPR*, pp. 1112–1122, 2026. 1, 2, 6, 14

Ying Jin, Pengyang Ling, Xiaoyi Dong, Pan Zhang, Jiaqi Wang, and Dahua Lin. Reasonpix2pix: instruction reasoning dataset for advanced image editing. In *CVPR Workshop*, 2024. 3

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *ICLR*, volume 2024, pp. 23395–23422, 2024. 3, 4, 6, 7, 9, 19

Yoonjeon Kim, Soohyun Ryu, Yeonsung Jung, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang, and Eunho Yang. Preserve or modify? context-aware evaluation for balancing preservation and modification in text-guided image editing. In *CVPR*, pp. 23474–23483, 2025. 2

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013. 5

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, pp. 4015–4026, 2023. 3

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *ICCV*, pp. 19721–19730, 2025. 3

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *CVPR*, pp. 7877–7886, 2020. 3

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pp. 22511–22521, 2023. 3

Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Junlin Xie, Gao Peng, and Hongsheng Li. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. In *ICLR*, volume 2025, pp. 38680–38711, 2025. 3, 6, 7

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2022. 2, 3, 5

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pp. 38–55, 2024. 3, 14

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1X-Edit: A practical framework for general image editing. *arXiv*, 2025. 3, 6, 7, 17

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv*, 2023. 5

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 *ICLR*, 2022. 3

Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *NeurIPS*, 31, 2018. 3

Rishubh Parihar, Or Patashnik, Daniil Ostashev, Venkatesh Babu Radhakrishnan, Daniel Cohen-Or, and Kuan-Chieh Jackson Wang. Kontinuous kontext: Continuous strength control for instruction-based image editing. In *CVPR*, pp. 37929–37939, 2026. 1, 3

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH*, pp. 1–11, 2023. 3, 5

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. 2

Tianyuan Qu, Lei Ke, Xiaohang Zhan, Longxiang Tang, Yuqi Liu, Bohao Peng, Bei Yu, Dong Yu, and Jiaya Jia. RePlan: Reasoning-guided region planning for complex instruction-based image editing. *arXiv*, 2025. 1, 3, 4, 6, 7, 8, 14, 17, 19

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. 6, 14

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *CVPR*, pp. 8871–8879, 2024. 1

Stefano Soatto and Alessandro Chiuso. Visual representations: Defining properties and deep approximations. *arXiv*, 2014. 2

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv*, 2025a. 3, 6, 7, 8, 17

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *CVPR*, pp. 21964–21975, 2026. 3, 6, 7, 8, 17, 19

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv*, 2025b. 6

Bin Xia, Shiyin Wang, Yingfan Tao, Yitong Wang, and Jiaya Jia. Llmga: Multimodal large language model based generation assistant. In *ECCV*, pp. 389–406, 2024. 5

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. In *NeurIPS*, 2026. 1

Chun-Hsiao Yeh, Yilin Wang, Nanxuan Zhao, Richard Zhang, Yuheng Li, Yi Ma, and Krishna Kumar Singh. Beyond simple edits: X-planner for complex instruction-based image editing. In *AAAI*, volume 40, pp. 11991–11999, 2026. 1, 3

Fukun Yin, Shiyu Liu, Yucheng Han, Zhibo Wang, Peng Xing, Rui Wang, Wei Cheng, Yingming Wang, Aojie Li, Zixin Yin, et al. Reasonedit: Towards reasoning-enhanced image editing models. In *CVPR*, pp. 23283–23293, 2026. 3

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *CVPR*, pp. 26125–26135, 2025. 1, 2, 3, 6, 7, 8, 14, 17

Arman Zarei, Samyadeep Basu, Mobina Pournemat, Sayan Nag, Ryan A Rossi, and Soheil Feizi. Slideredit: Continuous image editing with fine-grained instruction control. In *CVPR*, pp. 14430–14439, 2026. 1, 3

Ziyun Zeng, David Junhao Zhang, Wei Li, and Mike Zheng Shou. Draw-in-mind: rebalancing designer-painter roles in unified multimodal models benefits image editing. In *ICLR*, volume 2026, pp. 84455–84487, 2026. 1, 3

Dong Zhang, Lingfeng He, Rui Yan, Fei Shen, and Jinhui Tang. R-genie: Reasoning-guided generative image editing. *arXiv*, 2025a. 3

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS*, 36:31428–31449, 2023. 3, 6, 7, 17

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *ICLR*, volume 2025, pp. 10728–10745, 2025b. 2

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018. 6

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. *NeurIPS*, 38:139195–139227, 2026. 3, 6, 7

Supplementary Material

Learning Minimal Sufficient Visual Intervention for Instruction-Based Image Editing

In this supplementary material, we describe the construction details of MSVI-13K in Sec. S1. We then summarize the dataset distribution in Sec. S2, evaluate the accuracy of predicted SES of MSVI in Sec. S3, provide additional qualitative comparisons and ablation visualizations in Sec. S4 and Sec. S5, and analyze representative failure cases in Sec. S6.

S1 Dataset Construction

We construct the annotations in two stages that separate visual grounding from structured specification generation. A single unconstrained pass would otherwise need to discover repeated instances, infer the requested transition, and serialize spatial supervision jointly. We first externalize image-pair evidence as detections, then use it to generate four SES labels in a fixed schema.

We draw 9,904 valid triplets from AnyEdit (Yu et al., 2025), 2,905 from CompBench (Jia et al., 2026), and 328 from the synthetic complex-instruction data released by SmartEdit (Huang et al., 2024), for 13,137 triplets in total. We retain records with a readable source image, target image, and instruction, and apply one annotation pipeline to all three sources.

Stage 1: paired-image detection. For each triplet, we present the source and target images jointly to Qwen3.5-9B (Qwen Team, 2026) and ask it to detect the visible instances in both images. The model returns separate source and target lists. For instances that share a category, the prompt requires a referring expression based on location, appearance, or relative position, such as "the car on the left." The descriptions and boxes serve as facts for the next stage.

Stage 2: four-field SES generation. We then invoke Qwen3.5-9B with the source–target image pair, instruction, and both detection lists. It generates the four ordered labels `Planning`, `Perceiving`, `Action`, and `Layout` under the shared annotation prompt in Sec. S1.1. The image pair supplies evidence of the realized transition, the instruction identifies the requested change, and the detections disambiguate entity identity and location. We disable thinking and use greedy decoding to avoid sampling-induced variation. The authors review the outputs for agreement with the image pair and instruction, cross-field consistency, and unambiguous grounding.

S1.1 Annotation Prompt

The annotation prompt is the key control mechanism that turns general multimodal reasoning into trainable SES supervision. Detection supplies object locations, but cannot decide which differences constitute the requested intervention or which content must remain invariant. The prompt makes these decisions explicit by jointly enforcing a closed intervention taxonomy, a fixed dependency among the four fields, and a coordinate-grounded layout. This produces an executable specification rather than an unconstrained description of the image pair, consistent with prior evidence for explicit planning and spatial grounding in image editing (Ji et al., 2025; Qu et al., 2025; Liu et al., 2024).

Listing 1 presents the production prompt. It preserves the constraints that determine the supervision while omitting demonstrations and implementation boilerplate. Its order is intervention regime, source-to-target transition, required actions, and normalized layout; indentation exposes the output hierarchy.

```
You are an expert image editing assistant. Your task is to analyze the editing
process from source image to edited image based on provided detection hints.
Generate a structured Chain—of—Thought (CoT) explanation using the following
four XML tags strictly.
1. <Planning>
   Analyze whether this is a Holistic, Focal, or Coupled editing task based on
   the instruction and visual comparison. Provide clear reasoning.

   Output format:
     This is a [Holistic/Focal/Coupled] editing task. [Your reasoning]
2. <Perceiving>
   Analyze and describe the content and style of both source and edited images.

```

```
   Source Content: Describe objects and elements in the source image.
   Source Style: Describe the visual style, atmosphere, and lighting of the
                 source image.
   Target Content: Describe objects and elements in the edited image.
   Target Style: Describe the visual style, atmosphere, and lighting of the
                 edited image.
   Key Differences: Highlight the main changes observed.
3. <Action>
   Break down the editing steps in chronological order.

   — Only describe steps that produce changes.
   — For objects or regions that remain unchanged, simply state:
       All other objects and regions will be kept unchanged.
   — Format each step as:
       Step X: [Action description with referring expression and region,
                if applicable]

4. <Layout>
   Output the final layout and editing information with two fields:
   contextual_prior and localized_intervention.

   For Holistic editing tasks:
     contextual_prior:
       composition:
         — object: [object label or referring expression]
           bounding_box: [x1, y1, x2, y2]
         — ...
       style: [Description of the output image's visual style, atmosphere,
               and lighting]
     localized_intervention:
       referring_expression: whole image
       bounding_box: [0, 0, 0, 0]

   For Focal or Coupled editing tasks:
     contextual_prior:
       composition:
         — object: [object label or referring expression]
           bounding_box: [x1, y1, x2, y2]
         — ...
       style: [Description of the output image's visual style, atmosphere,
               and lighting]
     localized_intervention:
       referring_expression: [Clear description identifying the edited region,
                              e.g., the dog on the left, the woman wearing a
                              hat and striped dress, or the red car in the
                              background]
       bounding_box: [x1, y1, x2, y2]

   Notes for <Layout>:
     — composition is a list of bounding boxes for all objects in the edited
       image.
     — Each box contains object (an object label for a single instance, or a
       referring expression for multiple instances) and bounding_box
       [x1, y1, x2, y2].
     — For multiple instances of the same class, use a referring expression
       such as the dog on the left or the red car in the background.
     — bounding_box uses normalized [x1, y1, x2, y2] coordinates in [0, 1].
     — For focal edited regions, choose the primary one or the most significant
       change region.
     — localized_intervention.referring_expression should be unambiguous and
       descriptive.
———
Now analyze the following editing task.
User Instruction: {instruction}
Source Image Detections (Fact):
  {src_hints}
Edited Image Detections (Fact):
  {edit_hints}
```

Listing 1: The production SES annotation prompt.

Two coupled constraints are particularly important. First, `Action` lists only operations that produce a change and collapses all preservation requirements into a single unchanged-content statement. Second, `Layout` records both the intended final composition and the primary intervention region. The prompt therefore prevents a common mismatch in which a semantically correct plan is paired with an ambiguous or overly broad spatial target. For HOLISTIC edits, it instead disables localized intervention so that an arbitrary object box cannot restrict a scene-wide transition. Figure S1 shows one four-field annotation produced under these constraints.

S2 DATASET DISTRIBUTION

Figure S2 summarizes MSVI-13K along the intervention-regime and semantic-editing dimensions. HOLISTIC edits account for 60.62% of the dataset, followed by FOCAL edits at 33.39% and COUPLED



Figure S1: Example SES for a concrete edit.

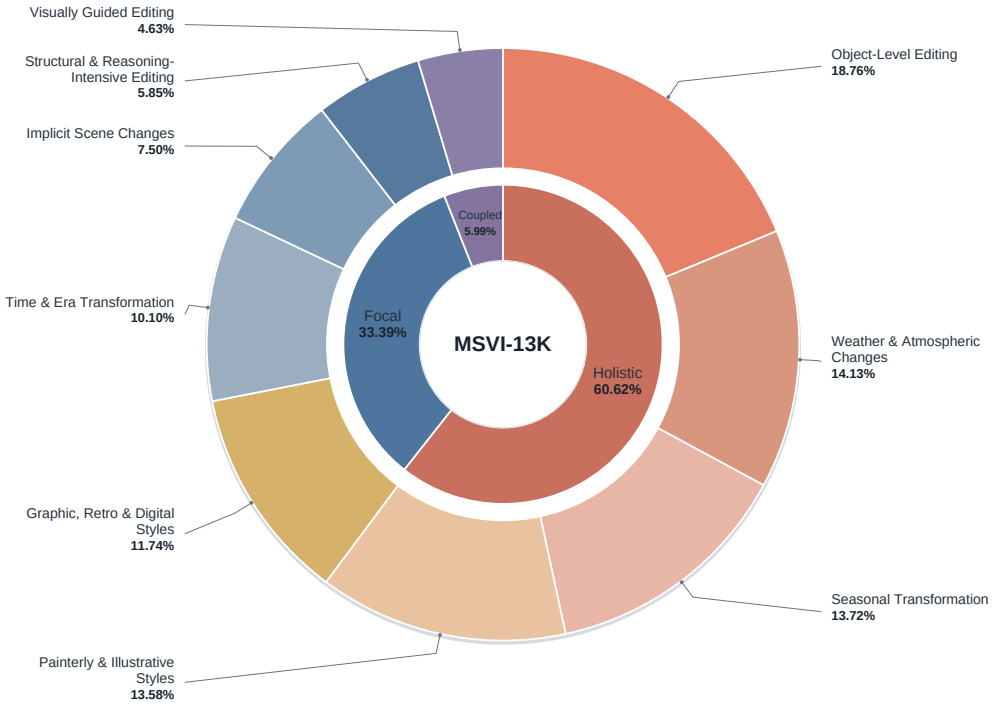

Figure S2: Distribution of the 13,137 examples in MSVI-13K. The inner ring reports intervention regimes: HOLISTIC for image-wide changes, FOCAL for localized changes, and COUPLED for combined changes. The outer ring groups instructions into nine semantic editing categories.

edits at 5.99%. This distribution covers the image-wide, localized, and combined changes used throughout the main paper.

At the semantic level, Object-Level Editing is the largest category at 18.76%, followed by Weather & Atmospheric Changes at 14.13% and Seasonal Transformation at 13.72%. The remaining examples span painterly, illustrative, graphic, retro, and digital styles; time and era transformations; implicit scene changes; reasoning-intensive edits; and visually guided edits. Together, these categories cover both object-local operations and broader scene transformations.

S3 Field-Level Evaluation of Structured Edit Specifications

We evaluate the four fields of the planner's structured edit specification $\Pi = (q, o, s, i)$ on the MSVI-13K test set, before intervention compilation and image generation, as shown in Table S1. `Planning`, `Perceiving`, and `Action` are evaluated using CLIP text similarity, while `Layout` is evaluated using mean intersection over union (mIoU).

SES field	Metric	Score (%)
`Planning`	CLIP similarity	90.123
`Perceiving`	CLIP similarity	88.765
`Action`	CLIP similarity	94.321
`Layout`	mIoU	96.907
SES composite	Equal-weight mean	**92.529**

Table S1: Field-level SES evaluation on the MSVI-13K test set.

S4 Additional Qualitative Result Comparisons

Figure S3 extends the main-paper comparison with one example from each intervention regime. In this section, we compare MSVI with six baselines: the instruction-tuned editors MagicBrush (Zhang et al., 2023) and AnySD (Yu et al., 2025); the generalist editors Step1X-Edit (Liu et al., 2025), OmniGen2 (Wu et al., 2026), and Qwen-Image-Edit (Wu et al., 2025a); and the planning-based editor RePlan (Qu et al., 2025). These methods are the same baselines qualitatively evaluated in the main paper.

The focal example requires adding a person at the specified side and orientation while preserving the street scene. Several baselines omit the person or place the subject away from the requested location, whereas MSVI confines the addition to the intended side. In the holistic example, the methods differ in whether they establish an afternoon appearance without replacing the cyclists or scene layout. The coupled example is more demanding because the weather transition and muddy path must be coordinated while retaining the hikers and forest structure. MSVI applies both changes without reconstructing the scene.

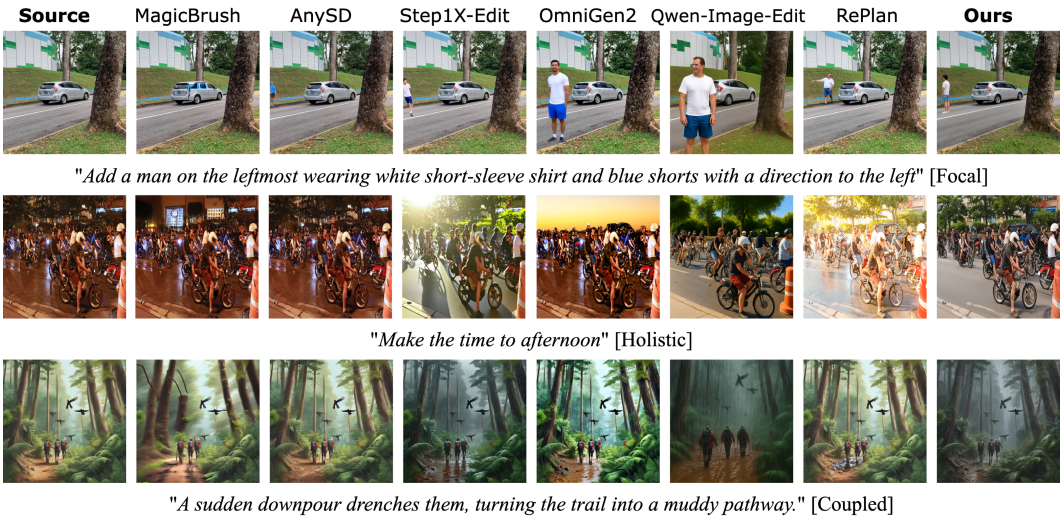

Figure S3: Additional qualitative comparisons across Focal, Holistic, and Coupled interventions.

S5 Ablation Visualization

Figure S4 provides qualitative counterparts to the component ablations. The columns compare the input, a length-matched unstructured plan, variants without grounding or SCSL, and full MSVI. The rows cover two holistic appearance transformations and two localized object-removal instructions.

Structured prediction and grounding. The unstructured plan leaves the dog scene close to the source instead of producing a gothic rendering, and it fails to establish a dusk appearance while altering the foreground riders. Its two removal outputs also contain conspicuous fill artifacts. Without grounding, the appearance changes reduce to broad color or contrast shifts, while the removal cases apply large rectangular patches that spill beyond the target objects. These examples suggest that a textual plan alone does not reliably specify either the requested transition or its spatial extent.

Source-copy suppression. Without SCSL, the gothic instruction remains largely source-like, and the dusk case is darkened without acquiring the warm appearance produced by the full model. The top rabbit is recolored rather than removed, and the white-duck case retains visible corruption near the edited region. This incomplete execution is consistent with the higher InvPres (0.906) and lower BG-LPIPS (0.033) of the variant without SCSL, despite its lower ER-IF (0.503) and ETP (0.221). In the displayed examples, full MSVI produces the requested appearance changes and removes the specified rabbit and duck while retaining the neighboring animals and the main scene structure.

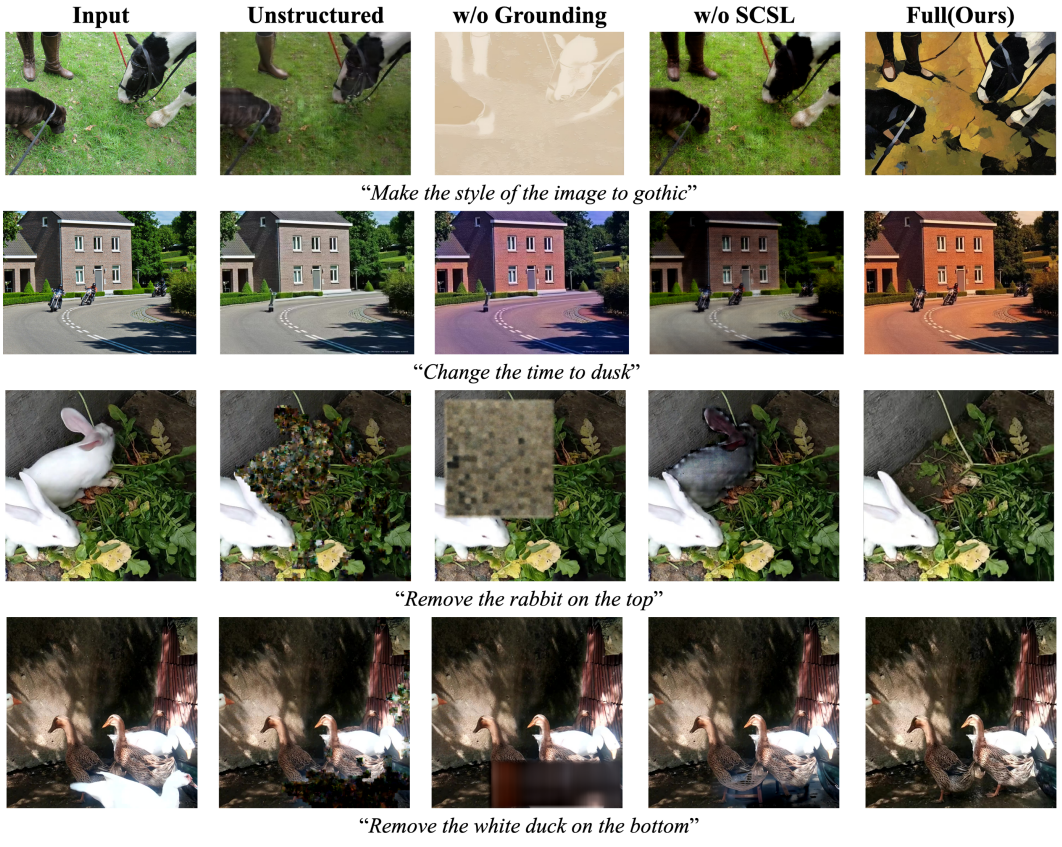

Figure S4: Qualitative visualization of the component ablations.

S6 Failure Cases

Although MSVI achieves the strongest overall balance across the evaluation metrics, two limitations remain. First, its InvPres of 0.896 on MSVI-13K remains below that of the most conservative direct editors. Their lower ER-IF and ETP partly explain this difference, but the result also indicates that

MSVI can alter invariant content while executing a stronger edit. Second, low-level background fidelity is not uniformly optimal under cross-dataset transfer. On Reason-Edit (Huang et al., 2024), MSVI obtains a BG-LPIPS of 0.041, compared with 0.039 for RePlan (Qu et al., 2025). On PIE-Bench (Ju et al., 2024), its BG-PSNR of 27.032 is below OmniGen2's (Wu et al., 2026) 28.436, despite stronger semantic alignment. Thus, small unintended changes can remain outside the intended intervention even when the requested semantics are satisfied.

Figure S5 shows two representative failures from our MSVI outputs. The first scene contains several visually similar cars at different locations and depths. Resolving the "upper-left white car" among these candidates is difficult, and the intended car remains unchanged. In the second scene, the two chickens toward the back overlap, obscuring their instance boundaries and the structure of the group. This occlusion makes both the existing objects and the intended insertion position harder to identify, resulting in an incorrect addition.

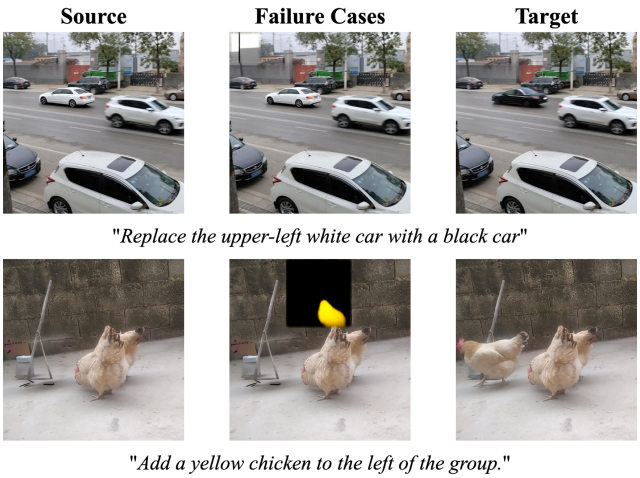

Figure S5: Representative failure cases.

The component ablations in Figure S4 help interpret these outcomes. The omitted car replacement resembles the source retention observed without SCSL. The failed chicken insertion shows a different limitation: selecting a compact editing region does not ensure correct edit-region placement or coherent object synthesis. Thus, preserving the background does not ensure that the instruction is executed, and localization alone does not guarantee a well-integrated edit. Future work will calibrate SES predictions and extend MSVI to iterative image editing and temporally consistent video editing.