WORLD LAUREATES FORUM
世界顶尖科学家论坛
国际工程智能大会
2026 International Conference on AI for Engineering
World Laureates Foundation
上海世界顶尖科学家发展基金会
2026
工程智能研究院
INSTITUTE OF AI FOR ENGINEERING

Learning Minimal Sufficient Visual Intervention for Instruction-Based Image Editing

Lingfeng He[1] Dong Zhang[1] Jinhui Tang[2]

[1] Nanjing University of Science and Technology [2] Nanjing Forestry University

01 ABSTRACT

Using implicit instructions for image editing tasks may leave intended change boundary underspecified and potentially lead 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*** that fulfills the instruction while minimizing changes to instruction-invariant visual content. Experimental results show that our method matches or outperforms SOTA approaches, supporting explicit visual intervention learning for balancing instruction fulfillment and content preservation.

02 METHOD

Our method 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. As illustrated in Fig. 1, our approach mainly comprises: (1) structured edit specification prediction; (2) regime-adaptive intervention; and (3) conditional flow-matching generation.

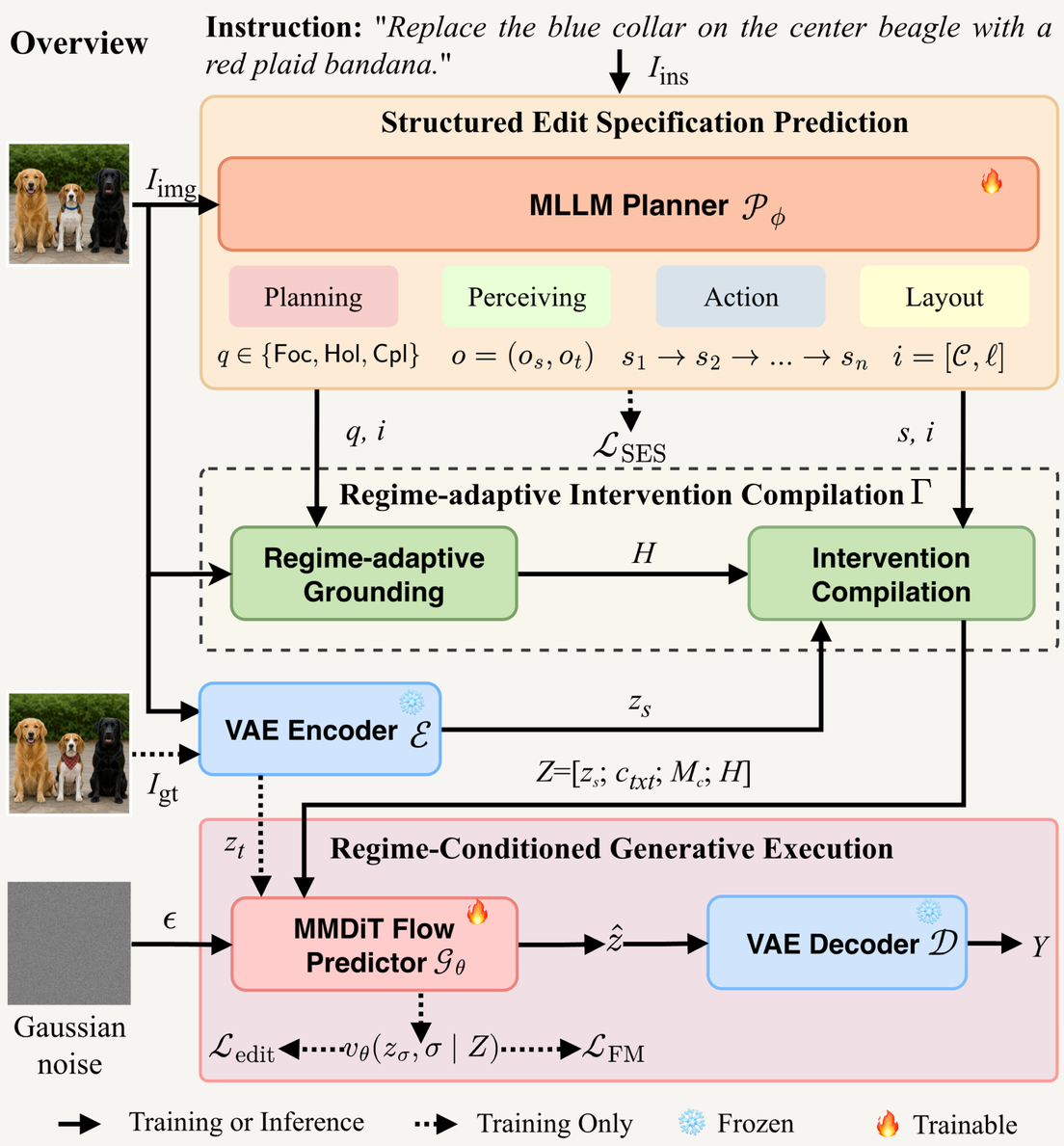


Figure 1. Overview of minimal sufficient visual intervention

(1) ***Structured edit specification prediction***: we use a multimodal planner to predict a structured edit specification that defines the edit regime, source and target states, semantic actions, and spatial layout.

(2) ***Regime-adaptive intervention compilation*****:** we selectively use SAM for Regime-adaptive Grounding, then combine source latents, semantic tokens, and spatial supports into the intervention condition through Intervention Compilation.

(3) ***Conditional flow-matching generation*****:** we condition MMDiT editor on the compiled visual intervention and train it with flow-matching and source-copy suppression to generate target.

03 RESULTS

We compare MSVI with 12 baselines on the MSVI-13K, Reason-Edit, and PIE-Bench. As show in Fig. 2, compared with the baseline average on MSVI-13K, gains are **71.7%** in instruction fulfillment, **1.4%** in invariant preservation, **11.7%** in intervention accuracy, and **1.415** in visual quality. These results support minimal sufficient visual intervention learning for effective editing with content preservation.

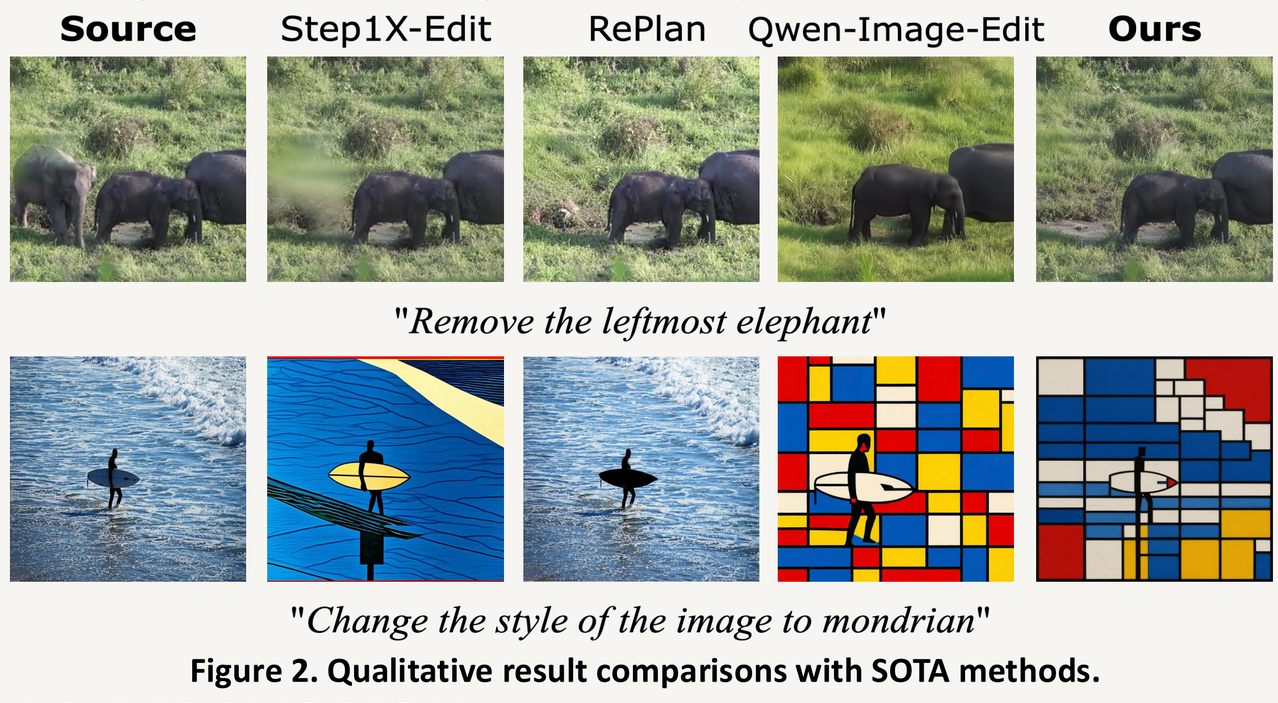


Figure 2. Qualitative result comparisons with SOTA methods.

04 DISCUSSION

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 feedback-based visual intervention refinement.

05 CONCLUSIONS

We introduce **MSVI** and MSVI-13K for learning minimal sufficient visual interventions, balancing instruction fulfillment and content preservation for more precise, controllable image editing.

CONTACT

Main authors: helingfeng@njust.edu.cn, dongzhang@njust.edu.cn

Corresponding author: tangjh@njfu.edu.cn

Scan to connect on WeChat and discuss this work.