Describe, Don't Dictate: Semantic Image Editing with Natural Language Intent

*****Supplementary Materials*****

A. Training Details

Following prior works [1, 11, 10], we adopt `Stable Diffusion v1.5` as the foundational backbone to ensure fair comparisons. Our editing model is trained on the UltraEdit dataset, which contains approximately 4 million text-image pairs. All training images are preprocessed to a resolution of 512×512 using center cropping and resizing.

We use the AdamW optimizer with a learning rate of 1×10^{-5} and a weight decay of 0.01. The model is trained on 8×H20 GPUs with a batch size of 32, without gradient accumulation, for a total of 160K steps. Mixed-precision (fp16) training is employed for efficiency.

During training, the learnable linear layer is zero-initialized and trained from scratch. The newly introduced cross-attention (CA) layers are initialized from the corresponding self-attention (SA) weights and fine-tuned using LoRA, with both the rank and alpha set to 64.

B. Details of Compared Methods

In this section, we provide a detailed description of the main compared methods in our evaluation.

Training-free Methods:

- *MasaCtrl* [2]: MasaCtrl is a tuning-free method based on mutual self-attention, where queries are derived from the edited image while keys and values come from the reference image. This mechanism enables local editing but struggles with global modifications, often resulting in inconsistencies when large-scale changes are required.
- *RF-Solver-Edit* [9]: RF-Solver-Edit is a training-free editing framework based on rectified flow, enabling inversion, reconstruction, and editing across both image and video domains. By leveraging Taylor expansion to improve sampling and inversion accuracy, RF-Solver-Edit achieves high efficiency. However, it requires careful tuning of the sampling step for different cases and tends to have weaker content preservation, leading to instability in certain editing tasks.
- *PnPInversion* [4]: PnPInversion is a diffusion-based text-guided editing method that improves the inversion process by disentangling source and target diffusion branches. This separation enhances both content preservation and edit fidelity, leading to superior editing outcomes. It introduces a lightweight inversion technique implemented in just three lines of code. Evaluated on PIE-Bench with diverse images and editing types, PnPInversion outperforms previous optimization-based methods in accuracy and speed, achieving nearly an order of magnitude acceleration.
- *FPE* [6]: FPE is a tuning-free text-guided image editing method that analyzes the roles of cross-attention and self-attention maps in diffusion models like Stable Diffusion. It finds that cross-attention maps often encode object attribution, which can lead to editing failures, while self-attention maps are crucial for preserving geometric and shape details. Leveraging this insight, FPE simplifies existing editing approaches by modifying only the self-attention maps in specific layers during denoising, resulting in a more stable and efficient editing process.
- *Turboedit* [3]: TurboEdit addresses the challenge of adapting text-based image editing to fast-sampling diffusion models. Focusing on the popular "edit-friendly" DDPM-noise inversion approach, it identifies two main failure modes: visual artifacts and insufficient editing strength. To tackle these, TurboEdit proposes a shifted noise schedule to correct noise mismatches and introduces a pseudo-guidance technique that enhances edit magnitude without causing artifacts. This method enables efficient text-based editing with as few as three diffusion steps, offering both practical speed improvements and novel insights into diffusion-based editing mechanisms.

Training-based Methods:

- *InstructPix2Pix* [1]: InstructPix2Pix is a diffusion-based model trained on a large-scale synthetic dataset, which maps textual instructions to image edits. It introduces additional input channels to encode reference image information. While it enables flexible instruction-driven editing, the reliance on synthetic data introduces potential biases, and the model often exhibits unstable edits — over-modifying some attributes while failing to sufficiently change others.
- *MagicBrush* [11]: MagicBrush is an instruction-based editing model fine-tuned on a manually annotated dataset encompassing various types of edits. Benefiting from

high-quality training data, it achieves improved editing accuracy. However, since the dataset is derived from DALL·E 2 [7] outputs, it may suffer from limited diversity and generalization capabilities.

- *EmuEdit* [8]: EmuEdit is a large-scale instruction-following image editing model trained on a dataset of 10 million examples. It introduces task embeddings to enhance generalization across diverse editing scenarios and establishes a dedicated benchmark, Emu Edit Test, for evaluation. Although it demonstrates strong performance across various tasks, the model and dataset have not been publicly released, limiting reproducibility and further research.
- *AnyEdit* [10]: AnyEdit is a multi-modal instruction-following model trained on a dataset covering 25 types of image edits. It integrates concepts from both instruction-based editing and reference-guided adaptation to enable flexible and high-quality image modifications. However, its relatively complex architecture may pose compatibility challenges for integration with existing frameworks.
- *BrushEdit* [5]: BrushEdit introduces a new paradigm for image editing by integrating multi-modal large language models, object detection modules, and inpainting networks. It reformulates the editing task as an inpainting problem through a pipeline-based workflow. While this design improves accuracy, it also introduces cascading errors, as the overall performance heavily depends on the correctness of intermediate detection and segmentation stages. Moreover, global edits—such as seasonal or stylistic transformations—remain challenging due to the model's reliance on local, mask-based modifications.

C. Evaluation Metrics and Protocol

C.1. Evaluation Metrics

To ensure a comprehensive evaluation of semantic image editing performance, we employ multiple metrics covering instruction adherence, image consistency, and image quality.

Instruction Adherence:

- *CLIP-T*: CLIP-T quantifies alignment between the target description and the edited image using CLIP text-image similarity. A higher score indicates better adherence to the given instruction.

Image Consistency:

- *L1 and L2 Distance*: L1 and L2 distances measure pixel-wise differences between the original and edited images. Lower values indicate better consistency and minimal unwanted alterations.
- *LPIPS*: LPIPS assesses perceptual similarity by comparing deep feature representations extracted from neural networks. A lower score suggests that the edited image retains more structural and textural similarity to the original image.

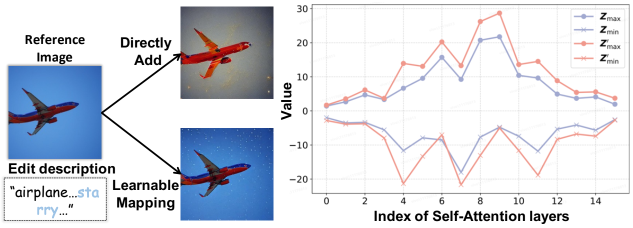


Figure 1. Comparison of different fusion methods for Z (self-attention output) and Z' (attention bridge output), along with their respective ranges.

- *SSIM*: SSIM captures structural integrity and luminance consistency between original and edited images. Higher values indicate superior preservation of image structure.
- *DINO-I and CLIP-I*: DINO-I and CLIP-I measure high-level feature similarities between the original and edited images, leveraging self-supervised and vision-language representations, respectively. Higher scores suggest that the edited image preserves semantic attributes more effectively.

Image Quality:

- *PSNR*: PSNR evaluates the pixel-wise reconstruction quality by comparing the difference in intensity values. A higher PSNR implies that the edit maintains high fidelity to the original content.

C.2. Evaluation Protocol

During evaluation, we use the official open-source implementations and publicly available pretrained weights of baseline methods whenever possible. To ensure a fair and reproducible comparison, all methods are evaluated under a same protocol, with fixed random seeds, consistent preprocessing steps, and recommended hyperparameters where applicable.

D. Analysis of Feature Magnitude Imbalance in Attention Fusion

As shown in Fig. 1 (right), Z' exhibits a larger value range than Z, often overpowering the original features and nullifying the editing effect (Fig. 1 (left)). To mitigate this imbalance, we replace the naive addition with a linear projection that aligns the magnitude of Z' with Z. This simple yet effective adjustment preserves the residual structure and leads to better editing performance, as visually confirmed in the Fig. 1 (left).

E. User Study

To assess the perceptual quality of different semantic image editing methods, we conducted a user study involving 30 participants, including AI researchers and general users. Participants were shown a series of edited images generated by various methods and asked to rate them on a Likert scale from 1 to 5, with higher scores indicating better performance.

Ratings were based on three key aspects:

- Instruction Adherence – How well the edited image aligns with the given textual instruction.
- Image Consistency – The extent to which the edited image preserves the structure and semantics of the original image.
- Image Quality – The overall visual realism and aesthetic appeal of the edited image.

Figure 2 presents the results, including a bar chart of average scores and pie charts showing the vote distributions across the three criteria. Our method achieves the highest ratings across all aspects, indicating strong alignment with user intent, superior consistency and quality, highlighting its robustness in maintaining structural fidelity while pro-

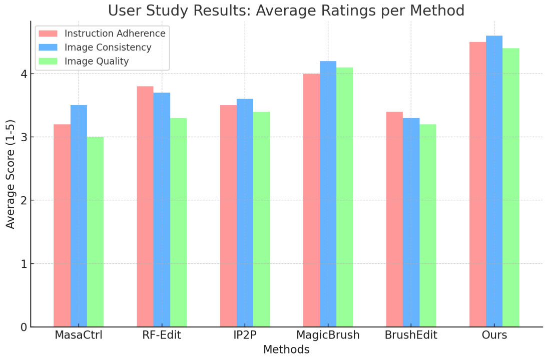


Figure 2. User Study Results.

F. Comparison with LoRA-tuned IP2P

To further evaluate the advantage of our architecture over existing editing paradigms, we fine-tuned Instruction-Pix2Pix using LoRA on the same description supervision. As shown in Table 1, while LoRA-tuned Instruction-Pix2Pix (IP2P$_{+\text{LoRA}}$) achieves improved performance compared to the original Instruction-Pix2Pix baseline, it still falls short of our method across most metrics.

In particular, our model achieves lower distortion and higher perceptual alignment. These results demonstrate that simply fine-tuning a generic instruction-based editing model, even with LoRA and rich descriptions, is insufficient to fully capture semantic intentions and preserve visual fidelity. By contrast, our method benefits from the architectural design of cross-attentive UNets with an attention bridge, leading to more consistent and effective edits.

Table 1. Comparison with LoRA-tuned Instruction-Pix2Pix on EmuEdit test set.

Method	**L1↓**	**L2↓**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**DINO-I↑**	**CLIP-I↑**	**CLIP-T↑**
IP2P	0.150	0.048	0.319	14.99	0.529	0.564	0.739	0.268
IP2P$_{+\text{LoRA}}$	0.091	0.017	0.218	19.09	0.596	0.703	0.811	**0.315**
Ours	**0.065**	**0.011**	**0.139**	**20.99**	**0.661**	**0.843**	**0.874**	**0.315**

G. Preservation of Base Model's Generative Capability

Our approach leverages the robust generative abilities of pre-trained T2I models by integrating our editing framework without altering the base architecture. This design choice ensures that the inherent generative capabilities of the original model remain unaffected. To substantiate this, we present the outputs of the base model after removing our additional network components, demonstrating that the model's original generative performance remains intact.

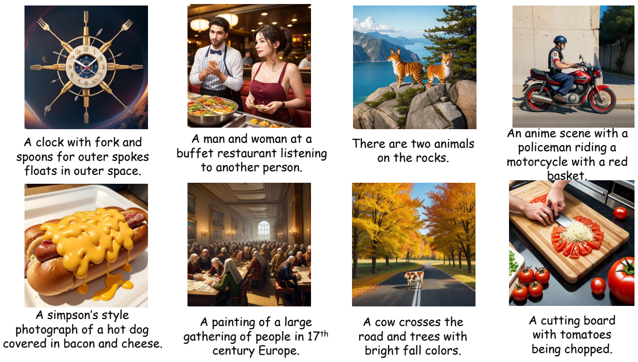
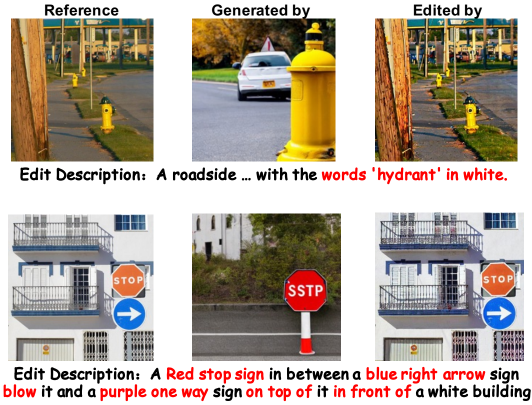

A clock with fork and spoons for outer spokes floats in outer space.
A man and woman at a buffet restaurant listening to another person.
There are two animals on the rocks.
An anime scene with a policeman riding a motorcycle with a red basket.
A simpson's style photograph of a hot dog covered in bacon and cheese.
A painting of a large gathering of people in 17th century Europe.
A cow crosses the road and trees with bright fall colors.
A cutting board with tomatoes being chopped.

Figure 3. Generated images from the base T2I model after removing our editing framework. The consistent quality of these images demonstrates that the base model's generative capabilities remain intact, validating the non-intrusive nature of our approach.

This analysis underscores that our method preserves the original strengths of the base T2I model while enhancing its functionality with advanced editing capabilities.

H. Limitation



Edit Description: A roadside ... with the words 'hydrant' in white.
Edit Description: A Red stop sign in between a blue right arrow sign blow it and a purple one way sign on top of it in front of a white building.

Figure 4. Limitation of DescriptiveEdit.

DescriptiveEdit depends on the generative capabilities of the underlying T2I model. Refer to Fig. 4: Row 1 shows text rendering failure, where `Stable Diffusion v1.5` fails to generate "hydrant". Row 2 shows challenges in multi-object generation and spatial arrangement, producing edits similar to the reference images.