SR-Edit: Region-Aware Image Editing via Self-Refinement

Andong Wang[1,2], Zehua Chen[1]*, Yuxuan Jiang[1], and Jun Zhu[1]*

[1] Tsinghua University, Beijing, China
[2] Renmin University of China, Beijing, China
andong000@ruc.edu.cn {zhc23thuml,dcszj}@mail.tsinghua.edu.cn

Abstract. With the recent rapid progress in generative models, image editing has made remarkable advances, yet achieving faithful edits that precisely modify only the target regions while strictly preserving all other regions remains challenging. Since externally provided region annotations are often difficult to obtain in practice, a growing body of work seeks to improve preservation by automatically inferring edit and non-edit regions, and then enforcing consistency on the latter. However, these approaches still suffer from inaccurate region estimation and heuristic correction strategies that distort the native inference process, making methods designed for fidelity themselves a new source of artifacts. We propose SR-Edit, an image editing framework that overcomes these issues via iterative self-refinement. Specifically, at each iteration, SR-Edit first (i) extracts progressively precise and self-consistent region separation from the model's own predictions by lightweight post-processing, and then (ii) enforces preservation in non-edit areas through correction updates that remain aligned with the original sampling dynamics. Extensive experiments demonstrate that SR-Edit achieves superior preservation and overall image quality compared to existing editing techniques.

Keywords: Image Editing · Generative Models · h-transform theory

1 Introduction

Recent progress in generative models [20,33,36,48,54] has greatly expanded the capabilities of image editing. A variety of editing paradigms have been proposed, including editing with forward noising followed by conditional denoising [40], training-based editors that learn an explicit mapping from a source image and a textual editing instruction to the edited image [5,35,61,62,65], and inversion-based methods that invert an image into a diffusion trajectory and then modify the denoising process to achieve the edit [19,27,37,43,63,71].

Despite impressive results, a persistent bottleneck remains: *faithful editing*, i.e., producing the desired modifications while *strictly preserving* all other content. In real-world edits, users typically expect backgrounds, identity cues, fine

* Corresponding authors: Zehua Chen and Jun Zhu.

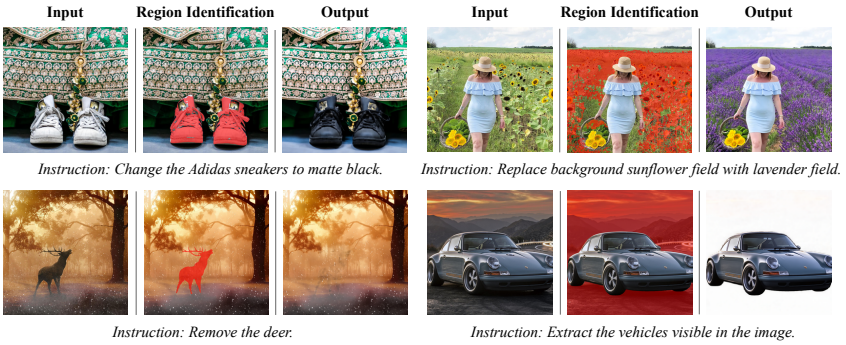

Fig. 1: Examples of edited images produced by SR-Edit. The edit regions are automatically inferred by the model and indicated by red overlays ▢ on original images.

textures, and geometric details outside the target concept to remain unchanged. However, most generative editing pipelines still rely on sampling dynamics that start from a noisy state and evolve under a new condition, which inherently operates on global image content [40]. As a result, even when the semantic edit succeeds, the output may exhibit unintended drift in non-edited regions, such as subtle background changes, texture inconsistencies, or identity degradation.

A common strategy to improve preservation is to explicitly separate the image into *edit* and *non-edit* regions, and then enforce consistency on the latter. When accurate user-provided masks are available, spatially constrained editing can be effective [1, 38, 51], but manual annotation is costly and often unavailable in practice. This motivates mask-free approaches that infer editable regions automatically, for example by manipulating or interpreting attention maps [18, 19, 45, 52, 55] or by computing differences between diffusion predictions under source and target prompts [11, 37].

Nevertheless, automatic region estimation remains challenging. The inferred masks are often insufficiently precise, failing to align with true pixel boundaries, and can also be temporally unstable across sampling steps [6]. Moreover, many preservation mechanisms rely on heuristic interventions such as feature fusion and reuse, or KV [56] interpolation in intermediate representations, which are not consistent with the original sampling dynamics [37, 46]. Consequently, the techniques introduced above to improve fidelity do not reliably enhance preservation. Instead, they can become a new source of imprecision and artifacts.

We propose **SR-Edit**, a *self-refined* image editing framework that addresses both failure modes: inaccurate region estimation and heuristic preservation corrections. The key idea is to *use the model's own output as self-feedback* to construct a reliable pixel-space separation between edited and preserved content, and then to enforce preservation through an update rule that remains aligned with intrinsic sampling dynamics.

Concretely, SR-Edit leverages the model's predictions to build *self-consistent difference maps* and converts them into a precise edit and non-edit decomposi-

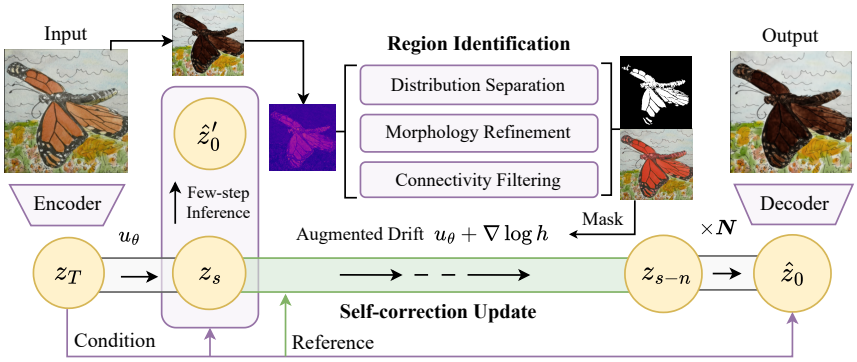

Fig. 2: Overview of SR-Edit. The edit instruction of the case is *Change the animal's fur color to a darker shade*. We use u_θ to denote the sampling vector field that drives the generative trajectory. s denotes the start time of self-correction update while n denotes the time duration. N denotes the number of self-refinement iterations.

tion via lightweight pixel-space post-processing [17, 53]. Unlike latent-space discrimination, pixel-space differences directly reflect decoded visual changes and admit mature refinement operations, yielding sharper and more stable region estimates.

To enforce preservation, we adopt a principled conditioning mechanism based on *Doob's h-transform* theory [4, 14]. We formulate preservation as a constraint on the non-edit region by modeling it as a vanishing-noise observation, and derive an additive guidance term that enhances fidelity while preserving the structure of the sampling procedure. Under a plug-in approximation, this guidance reduces to a masked residual on the model's clean-image prediction, which can be injected into diffusion samplers as an additive drift correction and analogously into flow-matching samplers as a velocity correction.

In contrast to heuristic state interventions (e.g., direct feature replacement or blending) [46, 56], our update is derived from a distribution transform and therefore tends to preserve native inference dynamics while suppressing non-edit drift. This complements existing methods that improve fidelity to the source image through more accurate inversion [23, 43, 57] and structural conditioning approaches that anchor geometry of the source via external controls [61, 68].

SR-Edit operates in an *iterative self-correction* loop: region separation is inferred from current predictions, preservation guidance is applied to suppress non-edit drift, and the model produces a progressively refined result across iterations. This self-refinement gradually improves region localization and preservation while maintaining edit intent. Extensive experiments demonstrate that SR-Edit yields superior non-edit preservation and improved overall visual quality compared to prior region-aware and mask-free editing techniques [11, 37, 46].

Our contributions can be summarized as follows:

- We introduce **SR-Edit**, an image editing framework that improves faithfulness by jointly enhancing region identification and preservation in an iterative self-refinement process.
- We propose **self-consistent region identification** derived from the model's own predictions and show that **pixel-space** post-processing enables more precise and stable edit and non-edit separation than latent or attention based discrimination.
- We develop a **Doob's h-transform** formulation for region preservation and derive a practical masked guidance rule that integrates naturally with both diffusion and flow-matching samplers.
- We validate that SR-Edit improves non-edit preservation and overall image quality across diverse editing scenarios, outperforming existing editing techniques.

2 Related Work

2.1 Image Editing with Generative Model Backbones

Generative model backbones. Modern generative models, such as diffusion models [20, 48, 54], flow matching [2, 33, 36], and bridge models [7, 10, 32, 70], have shown a strong capability of faithfully reconstructing a target distribution with learned time-dependent scores or vector fields. Given a scalable network architecture [34, 39, 48], these generative frameworks have been able to capture complex data distributions defined by large-scale datasets, enabling high-fidelity generation across data modalities, such as image [15, 48], audio [9, 25, 29], video [21, 42, 59], or time-series signals [3, 8, 41]. In the inference process, these frameworks usually start from a prior distribution, *e.g.*, Gaussian noise or a clean representation, and gradually generate the target with iterative refinement steps, showing a noise-to-data [12, 24, 58] or data-to-data sampling trajectory [31, 66].

Image editing paradigm. Following the remarkable breakthroughs of generative models, a growing body of research has focused on adapting these models for editing tasks [22,26]. Image editing methods can be roughly grouped by how they incorporate an input image and enforce edit intent. *Forward-and-backward* approaches inject noise into the input and denoise under new conditions, using the noise level to regulate the edit magnitude [40]. *Training-based instruction editing* methods learn an explicit mapping from the source image and textual instruction to the edited output through supervised or synthetic training pipelines, allowing direct and scalable manipulation [5, 35, 61, 62, 65]. *Inversion-based* methods first invert a real image into a diffusion latent trajectory and then apply prompt changes while trying to preserve reconstruction [19, 27, 37, 43, 63, 71].

Preservation of the source. Preserving the source image during edits is still difficult because noisy initialization in generative backbones can induce global drift, causing unintended changes in irrelevant regions. Existing approaches improve preservation through several typical mechanisms. [23, 43, 57] achieve high-fidelity

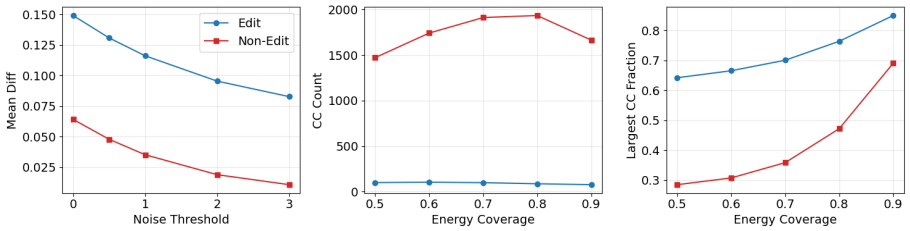

Fig. 3: Quantitative evidence of model behavior. On edit cases and human-annotated region partitions from MagicBrush test split [67], editing model [62] shows a clear pattern: edit regions concentrate strong changes in a coherent area, while non-edit regions produce low-magnitude and weakly structured changes. **Left:** mean pixel change inside and outside the annotated region after applying a MAD-based threshold[1]; the inside-to-outside ratio increases from 2.32 to 7.71, indicating much higher change energy in edit regions. **Center:** number of connected components among pixels covering a top fraction of total change energy, extremely large ratios (e.g., 1472 vs. 100 at 0.5 coverage) indicate that non-edit differences are highly fragmented whereas edits remain spatially concentrated. **Right:** fraction of selected pixels contained in the largest connected component, further confirming stronger spatial concentration for edits (e.g., 64% vs. 29% at 0.5 coverage).

inversion and reconstruction, then the conditional sampling can start from a more faithful state and thus reduce global deviation. External or automatically estimated masks separate edit regions and non-edit regions, which explicitly improve preservation of the source [1, 11]. We will further discuss the method in Sec. 2.2. Manipulating attention features [56] or injecting intermediate features can also enforce spatial alignment between the source image and edited output [6, 19, 45, 55]. Moreover, external structural constraints can be conditioned on edges, poses, segmentation, or depth, providing an explicit control to preserve original structure even when appearance changes [61, 68].

2.2 Region-Aware Image Editing

External mask. Accurate edit and non-edit region separation offers an important pathway to faithful image editing. Early and widely-adopted approaches rely on binary masks provided by users to explicitly specify where edits should occur [1, 38, 51]. This design provides direct spatial control and often yields strong results when accurate masks are available. However, it requires manual annotation, which limits applicability in real-world usage. Moreover, the manual mask often has an inaccurate boundary, leading to incorrect modifications on irrelevant regions and therefore degrading edit faithfulness.

[1] MAD [28,50] (median absolute deviation), a robust noise scale estimator that reduces the impact of outliers by using the median and absolute deviations.

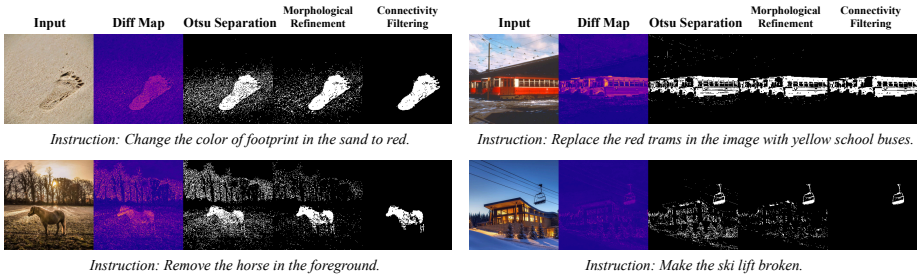

Fig. 4: Visualization of the incremental effects of each component in the SR-Edit post-processing pipeline for region identification.

Text-driven localization. To reduce reliance on external masks, subsequent work investigates inferring editable regions from text-conditioned signals alone. DiffEdit and Follow-Your-Shape [11, 37] derive edit regions by computing differences between diffusion processes conditioned on the source and target prompts. Image editing frameworks based on manipulation of cross-attention maps associate textual tokens with spatial regions, enabling localized edits without external masks [18, 19, 45, 52]. However, the editable region is inferred from semantic representations instead of pixel-level evidence, and the subsequent projection from semantic differences to image region separation is only approximate, often resulting in imprecise and unstable localization [6].

Region identification via inference behavior. More recent methods aim to localize editable regions by directly leveraging the model's inference behavior, which is naturally aligned with the editing process. SpotEdit [46] further explores region control from the inference trajectory by leveraging single-step reconstruction and performing linear interpolation between KV features [56] of intermediate latents and references. Although effective in practice, the single-step reconstruction is typically sensitive, while the reliance on KV-level masking or interpolation introduces heuristic engineering choices that could be refined toward more principled and precise localization.

3 SR-Edit

At inference time, we organize editing into a short *stabilization* stage followed by a stack of *SR-Edit blocks*. Each SR-Edit block performs a lightweight *probe-and-refine* cycle: (i) a few-step inference produces a provisional edit $\mathbf{I}_{\text{ref}}$, from which we infer the edit and non-edit regions via the pixel-space pipeline in Sec. 3.1; (ii) conditioned on this newly estimated region, we run a few guidance steps using the Doob's h-transform term in Sec. 3.2 to enforce region preservation while continuing the edit. Stacking these blocks yields an iterative self-refinement process: the model repeatedly *re-estimates* where changes occur and *re-applies* principled preservation guidance, so the region identification is continually refined

along with the prediction, progressively sharpening region separation for more accurate boundaries while keeping the edit flexible.

3.1 Precise Region Identification

Observation. Modern training-based image editors [2, 35, 62, 65] often demonstrate reasonably good fidelity in practice: meaningful modifications tend to concentrate in the intended edit region, while non-edit areas are largely preserved, with residual deviations typically small and weakly structured (Fig. 3). This empirical behavior makes the editor output itself a useful self-feedback cue for region separation: large and spatially coherent differences indicate edited areas, whereas scattered low-amplitude differences are more consistent with preserved regions.

Moreover, while many editing models operate in the latent space, performing region discrimination in decoded pixel space rather than latent space enables more precise region identification and allows us to leverage mature image post-processing techniques [17, 53] to suppress scattered artifacts, since latent features are typically patch-level and their variations are not a stable proxy for pixel changes due to highly nonlinear decoding [48]. Therefore, following prior practices [11, 37, 46] that exploit output-driven cues for refinement and typically operate in pixel space, we explicitly treat the model's own prediction as a stable self-feedback signal in pixel space to infer edit regions.

Region identification pipeline. Given an input image $\mathbf{I}_{\text{in}}$ and an edited image $\mathbf{I}_{\text{ref}}$ produced by few-step inference, both defined on the same pixel grid $\Omega = \{1, \dots, H\} \times \{1, \dots, W\}$ with C channels, we estimate a binary change mask $\mathbf{M} : \Omega \to \{0, 1\}$, where $\mathbf{M}(\mathbf{p}) = 1$ indicates edited pixels and $\mathbf{M}(\mathbf{p}) = 0$ indicates non-edited pixels. The method involves four steps: (i) constructing a difference map, (ii) binarization using Otsu's threshold, (iii) morphological refinement, and (iv) connected-component filtering. The incremental effects of each component can be viewed in Fig. 4.

Difference map. We compute a difference map $d(p)$ by taking the per-pixel absolute difference between the input and reference images, averaging over the C channels:

$$d(p) = \frac{1}{C} \sum_{c=1}^{C} \left| \mathbf{I}_{\text{in}}(p, c) - \mathbf{I}_{\text{ref}}(p, c) \right|, \qquad p \in \Omega. \tag{1}$$

This map $d(p)$ quantifies the local change at each pixel, with larger values indicating greater deviation between the images.

Otsu thresholding. To binarize d without manual tuning, we apply Otsu's separation method [44]. Its key idea is to choose a single global threshold t that best separates the histogram of d into two classes (low differences as *unchanged* and high differences as *changed*). Specifically, for each candidate threshold t,

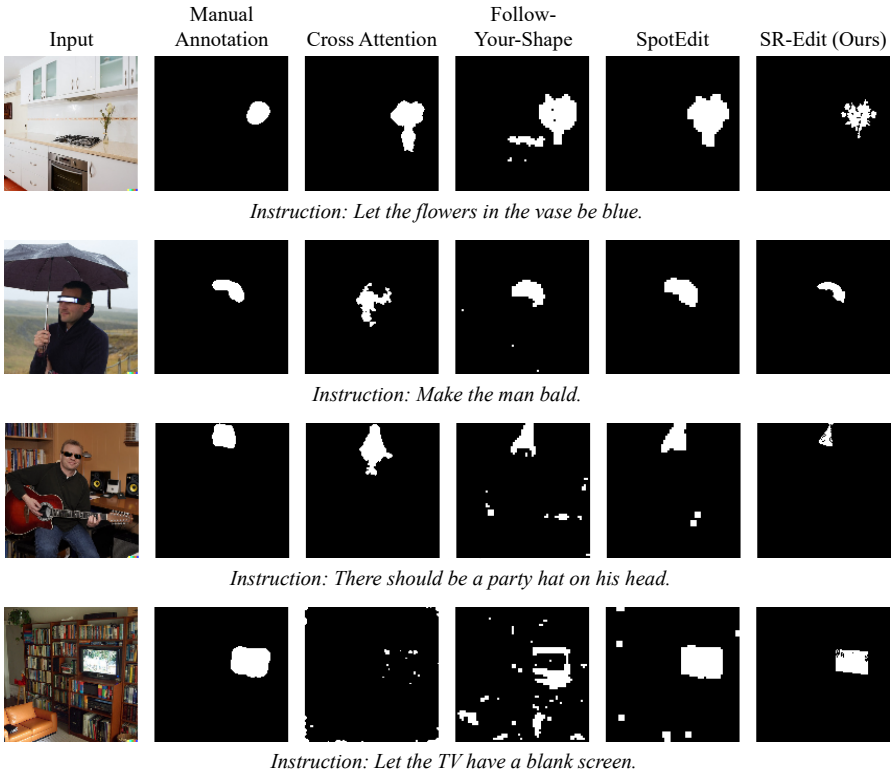

Fig. 5: Comparison on region identification between different methods. We include region separation results from manual annotations in MagicBrush [67], cross-attention maps [56] on token with largest L2 energy, Follow-Your-Shape [37], SpotEdit [46] and SR-Edit. All evaluated methods use Qwen-Image-Edit 2511 [62] as the backbone.

Otsu defines class probabilities $\omega_0(t), \omega_1(t)$ and class means $\mu_0(t), \mu_1(t)$ from the histogram, and selects the threshold that maximizes the between-class variance:

$$t^\star = \arg\max_t \ \omega_0(t)\,\omega_1(t)\big(\mu_0(t) - \mu_1(t)\big)^2. \tag{2}$$

We then obtain the initial mask by thresholding: $\mathbf{M}_0(p) = \mathbb{I}[d(p) \geq t^\star]$, where $\mathbb{I}[\cdot]$ denotes the indicator function that returns 1 if the condition holds and 0 otherwise.

Morphological refinement. The initial mask $\mathbf{M}_0$ may contain small isolated artifacts and small holes or breaks inside true change regions. We refine it using binary morphology [17, 53]. Specifically, we use a small circular kernel on the pixel grid: we first apply *opening* with kernel size s_{open} to remove small foreground speckles, and then apply *closing* with size s_{close} to fill small holes and

connect narrow gaps. This morphological refinement yields a cleaner and more spatially coherent mask $\mathbf{M}_1$.

Connected-component filtering. Finally, we run connected-component labeling on $\mathbf{M}_1$ [49]. Each connected component corresponds to a maximal set of foreground pixels that are mutually reachable through neighbor-to-neighbor steps. We remove components whose area (pixel count) falls below a minimum threshold $A_{\min}$, treating them as residual noise, and keep the remaining components as the final change mask $\mathbf{M}$. Our method shows improved precision in region identification compared with prior approaches, as illustrated in Fig. 5.

3.2 Region preservation via Doob's h-transform

Our method achieves region preservation by conditioning the sampling dynamics through an additive Doob's h-transform [4, 14] guidance term. In contrast to heuristic feature fusion such as KV reuse [56], the correction is derived from a distribution transform and thus more consistent with the underlying generative process.

Diffusion model. We consider a continuous-time diffusion model [20] defined by the forward SDE [54]

$$d\boldsymbol{x}_t = f(\boldsymbol{x}_t, t)\, dt + g(t)\, d\boldsymbol{w}_t. \tag{3}$$

Here $\boldsymbol{w}_t$ is standard Brownian motion, f is the drift, and $g(t)$ controls the noise scale. Let p_t denote the marginal density of $\boldsymbol{x}_t$ at time t. For conditional editing, we condition on a control signal $\mathcal{C}$ and write the corresponding marginal as $p_t(\boldsymbol{x} \mid \mathcal{C})$. Its score $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x} \mid \mathcal{C})$ is the conditional log-density gradient that drives the reverse-time denoising dynamics.

Observation model and Doob function. Let $\bar{\mathbf{M}}$ denote the complement of the mask defined earlier, and let $\boldsymbol{y} \triangleq \bar{\mathbf{M}}\boldsymbol{x}_{\text{src}}$. We enforce the preservation constraint in the form $\bar{\mathbf{M}}\boldsymbol{x}_0 = \boldsymbol{y}$. We encode preservation with the vanishing-noise observation model $\boldsymbol{Y} = \bar{\mathbf{M}}\boldsymbol{x}_0 + \boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{obs}}^2 I)$, and take $\sigma_{\text{obs}} \to 0$. The corresponding likelihood is $p(\boldsymbol{y} \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{y}; \bar{\mathbf{M}}\boldsymbol{x}_0, \sigma_{\text{obs}}^2 I)$. Define the Doob function

$$h(t, \boldsymbol{x}) \triangleq \mathbb{E}[p(\boldsymbol{y} \mid \boldsymbol{x}_0) \,|\, \boldsymbol{x}_t = \boldsymbol{x}, \mathcal{C}]\,. \tag{4}$$

By the Doob's h-transform, conditioning on $\boldsymbol{Y} = \boldsymbol{y}$ yields the factorization $p_t(\boldsymbol{x} \mid \boldsymbol{y}, \mathcal{C}) \propto p_t(\boldsymbol{x} \mid \mathcal{C})\, h(t, \boldsymbol{x})$, and therefore the conditional score decomposes additively as

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x} \mid \boldsymbol{y}, \mathcal{C}) = \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x} \mid \mathcal{C}) + \nabla_{\boldsymbol{x}} \log h(t, \boldsymbol{x}). \tag{5}$$

Consequently, any reverse-time diffusion sampler that uses the reference score can be made region-preserving by augmenting its drift with the guidance term $\nabla_{\boldsymbol{x}} \log h(t, \boldsymbol{x})$.

Plug-in and masked Gaussian guidance. Exact evaluation of $h(t, \boldsymbol{x})$ is generally intractable. Let $\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ be the model prediction of the clean image at time t. We adopt a plug-in approximation that replaces the latent $\boldsymbol{x}_0$ inside the likelihood by $\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$, which gives $h(t, \boldsymbol{x}_t) \approx \mathcal{N}(\boldsymbol{y}; \bar{\mathbf{M}}\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \sigma_{\text{obs}}^2 I)$. Equivalently, up to an additive constant independent of $\boldsymbol{x}_t$,

$$\log h(t, \boldsymbol{x}_t) \approx -\frac{1}{2\sigma_{\text{obs}}^2} \left\| \bar{\mathbf{M}}\big(\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) - \boldsymbol{x}_{\text{src}}\big) \right\|_2^2. \tag{6}$$

Differentiating yields

$$\nabla_{\boldsymbol{x}_t} \log h(t, \boldsymbol{x}_t) \approx -\frac{1}{\sigma_{\text{obs}}^2} J_{\widehat{\boldsymbol{x}}_0}(\boldsymbol{x}_t)^\top \bar{\mathbf{M}}\big(\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) - \boldsymbol{x}_{\text{src}}\big), \tag{7}$$

where $J_{\widehat{\boldsymbol{x}}_0}(\boldsymbol{x}_t)$ denotes the Jacobian of $\widehat{\boldsymbol{x}}_0$ with respect to $\boldsymbol{x}_t$. In implementation, the local sensitivity is absorbed into a time-dependent scalar schedule, which yields the practical masked-residual form

$$\nabla_{\boldsymbol{x}_t} \log h(t, \boldsymbol{x}_t) \approx -\lambda(t)\, \bar{\mathbf{M}}\big(\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) - \boldsymbol{x}_{\text{src}}\big). \tag{8}$$

Combining Equations (5) and (8) shows that, under the plug-in approximation, the h-transform guidance reduces to a masked residual on the model prediction $\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$. This residual can be viewed as a time-consistent error signal, so the correction acts as a conditional energy tilt toward preservation rather than a heuristic state intervention [4, 14, 30]. Since it enters only through the drift, it does not modify the diffusion coefficient $g(t)$ and therefore typically leaves the noise schedule and sampling structure largely intact. In the ideal case where the base conditional dynamics already match the desired conditioned process, the guidance term vanishes.

Preservation guidance for flow matching Flow models [33, 36] sample by the ODE $\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{u}_\theta(\boldsymbol{x}_t, t)$ and do not come with a canonical Doob's h-transform tied to an SDE. After the same plug-in reduction, preservation is encoded by the quadratic energy [16] $E_t(\boldsymbol{x}_t) \triangleq \frac{1}{2\sigma_{\text{obs}}^2} \|\bar{\mathbf{M}}(\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) - \boldsymbol{x}_{\text{src}})\|_2^2$, whose gradient yields the same masked residual direction. Because this guidance depends only on $\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ and not on a diffusion coefficient, it can be injected as an additive velocity correction without changing the ODE form:

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{u}_\theta(\boldsymbol{x}_t, t) - \gamma(t)\, \nabla_{\boldsymbol{x}_t} E_t(\boldsymbol{x}_t) \;\approx\; \boldsymbol{u}_\theta(\boldsymbol{x}_t, t) - \gamma(t)\, \lambda(t)\, \bar{\mathbf{M}}\big(\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) - \boldsymbol{x}_{\text{src}}\big), \tag{9}$$

which retains the flow matching transport structure while encouraging $\bar{\mathbf{M}}\widehat{\boldsymbol{x}}_0 = \boldsymbol{y}$.

4 Experiments

4.1 Dataset

We evaluate on **ImgEdit-Bench**, a benchmark derived from the ImgEdit dataset for instruction-based image editing [64]. We focus on the **single-turn** setting and

Method	CLIP↑	SSIM↑	PSNR↑	DISTS↓	LPIPS↓	Judge↑
InstructPix2Pix [5]	**26.68**	0.67	16.58	0.20	0.46	2.69
+**SR-Edit (Ours)**	25.07	**0.68**	**16.94**	**0.18**	**0.43**	**2.75**
AnyEdit [65]	25.15	0.70	18.95	0.17	0.41	2.82
+**SR-Edit (Ours)**	**25.21**	**0.85**	**19.19**	**0.13**	**0.35**	**2.99**
Qwen-Image-Edit 2511 [62]	<u>26.16</u>	0.61	14.92	0.23	0.46	3.84
+**Follow-Your-Shape** [37]	26.03	0.61	14.96	0.23	0.47	3.59
+**SpotEdit** [46]	26.11	**0.70**	<u>16.55</u>	<u>0.20</u>	<u>0.32</u>	<u>3.91</u>
+**SR-Edit (Ours)**	**26.80**	<u>0.67</u>	**17.40**	**0.18**	**0.31**	**3.94**
Step1X-Edit v1p2 [35]	25.89	0.68	15.96	0.20	0.39	4.00
+**Follow-Your-Shape**	25.84	0.67	15.93	0.20	0.39	<u>4.03</u>
+**SpotEdit**	<u>25.91</u>	<u>0.75</u>	**16.77**	<u>0.17</u>	<u>0.31</u>	**4.08**
+**SR-Edit (Ours)**	**26.09**	**0.77**	<u>16.48</u>	**0.14**	**0.27**	4.01

Table 1: Comparison between different methods on ImgEdit-Bench. Bold indicates the best result for each metric within each backbone group, and <u>underlining</u> indicates the second best. For the first two backbone groups, we do not mark second best results.

consider editing tasks confined to a specific area of the image, including *replace*, *adjust*, *background*, *remove*, *add*, and *action*, resulting in 497 valid cases in total.

4.2 Backbone Editors

We test our method on four open-source backbones from two editor families. For *diffusion-based* backbones, we include **InstructPix2Pix**, a standard instruction-following diffusion editor [5], and **AnyEdit**, a strong unified diffusion image editor that shows competitive editing performance across diverse instructions [65]. For *flow-based* editors, **Qwen-Image-Edit** 2511 and **Step1X-Edit** v1p2 are recent flow-style editors that are commonly adopted as modern backbones for general instruction-based editing [35, 62].

4.3 Baselines

We compare with two representative training-free region-aware methods that automatically infer editable regions and enforce preservation in non-edit areas during inference. **Follow-Your-Shape** uses a trajectory divergence map to localize editable regions and applies scheduled KV [56] injection to preserve non-target content during editing [37]. **SpotEdit** identifies editable regions via reconstruction-based stability estimation and preserves context using KV caching with interpolation from the source image [46].

4.4 Metrics

We report a CLIP-based text–image similarity score to reflect whether the edited result is semantically consistent with the instruction [47]. We report PSNR, SSIM, DISTS, and LPIPS between the edited image and the source image to measure preservation and fidelity, where PSNR and SSIM emphasize structure similarity, while DISTS and LPIPS capture perceptual distance [13, 60, 69]. We also include the ImgEdit judge-model score as an automatic evaluator that is designed to align with human preference for instruction-based editing [64].

4.5 Inference Configuration

Unless otherwise specified, we use the *default* settings provided by each baseline model and method. For SpotEdit [46] and Follow-Your-Shape [37], when the inference step count differs from the backbone editor's setting, we *linearly rescale* the editing schedule to match the new number of steps. For SR-Edit, we instantiate two iterations and set few-step inference stride to 3. The post-process parameters are set to $s_{\text{open}} = 3, s_{\text{close}} = 5, A_{\min} = 1000$. We set $\sigma_{\text{obs}} = 3 \times 10^{-2}$, and employ a linearly scheduled h-transform guidance strength. We enable self-refinement starting at 30% of the model's inference steps.

5 Results

5.1 Quantitative results

Tab. 1 shows a consistent pattern: SR-Edit mainly improves *non-edit preservation*, yielding higher structural fidelity (e.g., SSIM) and lower perceptual deviation (e.g., LPIPS), while largely keeping instruction alignment unchanged. This indicates SR-Edit primarily suppresses unintended drift rather than simply amplifying edit strength.

This behavior is consistent across backbones. On AnyEdit, SR-Edit markedly improves structure (SSIM 0.70 $\rightarrow$ 0.85) and reduces perceptual deviation (LPIPS 0.41 $\rightarrow$ 0.35). Moreover, for flow-based editors, the refinement may also strengthen semantic alignment: Qwen-Image-Edit achieves the best CLIP score (26.80) while improving fidelity, and Step1X-Edit attains the strongest preservation (SSIM 0.77) alongside a CLIP increase (26.09). On weaker baselines (e.g., InstructPix2Pix), CLIP may slightly drop, but the overall Judge Score still improves.

5.2 Qualitative results

Fig. 6 shows that our method achieves higher fidelity to the input and more realistic, higher-quality edits than competing methods.

Fig. 6: Qualitative comparison. Many baselines (*e.g.* the default setting) show weaker preservation to the source image, and Follow-Your-Shape can yield insufficiently realistic edits with residual blur. SpotEdit often introduces boundary discontinuities (edge jumps). In contrast, our method is more faithful to the input and produces realistic, high-quality edits. The instructions of the two cases are *Add a seagull perched on the edge of the wooden pier* and *Change the object color to a soft blue.* All editing methods use Qwen-Image-Edit 2511 [62] as backbone.

6 Ablation Study

Few-step inference versus reconstruction for region estimation. We compare few-step inference under different strides with reconstruction in terms of region-identification quality (Fig. 7). Overall, few-step inference consistently outperforms reconstruction in both IoU and F1. Moreover, reconstruction shows noticeable instability (e.g., a sharp drop around step 32 in both metrics), whereas the few-step curves remain smooth. Among the few-step settings, stride-2 yields the best performance; stride-3 is slightly lower but close; and stride-4 tends to underperform, as recall degrades more substantially with increasing steps. In practical self-refined inference, using regions inferred via reconstruction results in less faithful edits (Tab. 3).

Method	CLIP↑	SSIM↑	PSNR↑	DISTS↓	LPIPS↓	Judge↑
SR-Edit	26.80	**0.67**	**17.40**	**0.18**	**0.31**	**3.94**
Reconstruction	**26.95**	0.64	16.90	0.21	0.35	3.88
Single Iter	25.90	0.60	16.10	0.24	0.41	3.60

Table 3: Ablation results on ImgEdit-Bench. All methods use Qwen-Image-Edit 2511 [62] as the backbone. **Reconstruction** uses the reconstruction output for region estimation, without the few-step inference, while **Single Iter** uses only one iteration, meaning the edit region is identified once and then kept fixed.

Effect of iterative refinement. Next, we ablate the number of iterative self-refinement blocks. Using two SR-Edit blocks improves region alignment, as the mask and preservation guidance are re-estimated from progressively refined predictions. With a single block, the mask is fixed early and is typically over-inclusive and inaccurate, which weakens non-edit preservation and may also reduce instruction adherence by over-constraining the model (Tab. 3). These results highlight the importance of iterative mask re-estimation for simultaneously maintaining preservation and edit expressiveness.

Region post-processing pipeline. We ablate each component of the mask post-processing pipeline and evaluate mask quality against the output of the full pipeline. Tab. 2 indicates that the four steps are complementary: removing Otsu produces an almost non-informative mask with extremely low precision and IoU; removing opening increases false positives and lowers precision; removing

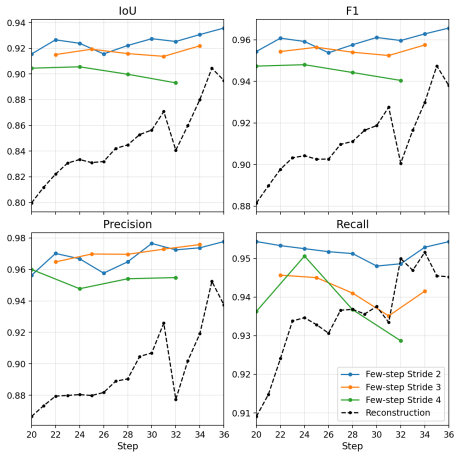

Fig. 7: Comparison between few-step inference and reconstruction. Few-step inference is better at almost all steps.

Setting	P	R	IoU	F1
w/o Otsu	0.10	1.00	0.10	0.16
w/o Open	0.69	1.00	0.69	0.76
w/o Close	1.00	0.91	0.91	0.95
w/o CC	0.76	1.00	0.76	0.84

Table 2: Ablation results on post-processing pipeline of region identification. The results indicate that each component contributes in a distinct and complementary manner. For example, removing the *opening* module may increase recall, as fewer constraints allow more positive predictions, but this comes at the expense of precision and overall balance. These variations confirm that each module plays a specific role in improving the overall accuracy and robustness.

closing harms spatial coherence and reduces overlap; and removing connected-component filtering retains small noisy fragments and degrades the overall balance. Together, these results suggest that the complete pipeline is necessary for accurate and stable region separation.

7 Conclusion

In this paper, we present SR-Edit, a region-aware image editing method that performs faithful edits without external masks. It uses iterative self-refinement to infer and sharpen edit regions from pixel-space predictions, and applies a Doob's h-transform–based masked residual correction to improve preservation for diffusion and flow-based backbones. Experiments across multiple editors and metrics show better non-edit preservation and visual quality than representative baselines while maintaining instruction consistency. Ablation studies further verify the benefits of few-step probing for robust region estimation and the necessity of iterative mask re-estimation for balancing preservation and edit expressiveness. Overall, SR-Edit offers a simple, training-free framework that can be applied to modern backbones to reduce unintended drift during inference.

8 Acknowledgment

This work is supported by the National Natural Science Foundation of China (62550004, U24A20342, U25B6003, 92570001).

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18208–18218 (2022)
2. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)
3. Bolton, A., Zhou, W., Chen, Z., Iacovides, G., Mandic, D.: Refinebridge: Generative bridge models improve financial forecasting by foundation models. In: ICASSP (2026)
4. Bortoli, V.D., Thornton, J., Heng, J., Doucet, A.: Diffusion schrödinger bridge with applications to score-based generative modeling. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
6. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)

7. Chen, Z., He, G., Zheng, K., Tan, X., Zhu, J.: Schrodinger bridges beat diffusion models on text-to-speech synthesis. arXiv preprint arXiv:2312.03491 (2023)
8. Chen, Z., Miao, Y., Wang, L., Fan, L., Mandic, D.P., Zhu, J.: Versatile cardiovascular signal generation with a unified diffusion transformer. Nature Machine Intelligence **8**(1), 6–19 (2026)
9. Chen, Z., Tan, X., Wang, K., Pan, S., Mandic, D., He, L., Zhao, S.: Infergrad: Improving diffusion models for vocoder by considering inference in training. In: ICASSP (2022)
10. Chen, Z., Yang, Y., Yuan, B., Zheng, K., Liu, J.S., Zhu, J.: Guidedbridge: Training-freely improving bridge models with prior guidance. In: ICML (2026)
11. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
12. Dai, Y., Chen, Z., Jiang, Y., Ke, Q., Cai, J., Zhu, J.: Omni2sound: Towards unified video-text-to-audio generation. In: CVPR (2026)
13. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
14. Doob, J.L.: Conditional brownian motion and the boundary limits of harmonic functions. Bulletin de la Société mathématique de France **85**, 431–458 (1957)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
16. Feng, R., Yu, C., Deng, W., Hu, P., Wu, T.: On the guidance of flow matching. arXiv preprint arXiv:2502.02150 (2025)
17. Goyal, M.: Morphological image processing. IJCST **2**(4), 59 (2011)
18. Helbling, A., Meral, T.H.S., Hoover, B., Yanardag, P., Chau, D.H.: Conceptattention: Diffusion transformers learn highly interpretable features. arXiv preprint arXiv:2502.04320 (2025)
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
21. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
22. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
23. Jeong, D., Kang, D., Park, J., Lee, H., Paik, J.: Structure-preserving zero-shot image editing via stage-wise latent injection in diffusion models. arXiv preprint arXiv:2504.15723 (2025)
24. Jiang, Y., Chen, Z., Ju, Z., Dai, Y., Dou, W., Zhu, J.: Controlaudio: Tackling text-guided, timing-indicated and intelligible audio generation via progressive diffusion modeling. In: ACL (2026)
25. Jiang, Y., Chen, Z., Ju, Z., Li, C., Dou, W., Zhu, J.: Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation. In: ACM MM (2025)
26. Jiang, Y., Han, M., Dai, Y., Wang, A., Zhou, T., Ye, J., Wang, D., Shi, H., Li, B., Song, J., Yu, C., Zheng, B., Dou, W., Chen, Z., Zhu, J.: Freesonic: Training-free temporal-aware decoupled attention for precise audio editing. In: Interspeech (2026)

27. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6007–6017 (2023)
28. Law, J.: Robust statistics—the approach based on influence functions (1986)
29. Leng, Y., Chen, Z., Guo, J., Liu, H., Chen, J., Tan, X., Mandic, D., He, L., Li, X.Y., Qin, T., Zhao, S., Liu, T.Y.: Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. In: NeurIPS (2022)
30. Léonard, C.: A survey of the schr\" odinger problem and some of its connections with optimal transport. arXiv preprint arXiv:1308.0215 (2013)
31. Li, C., Chen, Z., Bao, F., Zhu, J.: Bridge-sr: Schrödinger bridge for efficient sr. In: ICASSP (2025)
32. Li, C., Chen, Z., Wang, L., Zhu, J.: Audio super-resolution with latent bridge models. In: NeurIPS (2025)
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
34. Liu, H., Chen, Z., Yuan, Y., Xinhao, M., Liu, X., Mandic, D., Wang, W., Plumbley, M.: Audioldm: Text-to-audio generation with latent diffusion models. In: ICML (2023)
35. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
36. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
37. Long, Z., Zheng, M., Feng, K., Zhang, X., Liu, H., Yang, H., Zhang, L., Chen, Q., Ma, Y.: Follow-your-shape: Shape-aware image editing via trajectory-guided region control. arXiv preprint arXiv:2508.08134 (2025)
38. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
39. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV (2024)
40. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022)
41. Miao, Y., Chen, Z., Li, C., Mandic, D.: Respdiff: An end-to-end multi-scale rnn diffusion model for respiratory waveform estimation from ppg signals. In: ICASSP (2025)
42. Mo, S., Chen, Z., Bao, F., Zhu, J.: Diffgap: A lightweight diffusion module in contrastive space for bridging cross-model gap. In: ICASSP (2025)
43. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)
44. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296) (1979)
45. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)

46. Qin, Z., Tan, Z., Wang, Z., Liu, S., Wang, X.: Spotedit: Selective region editing in diffusion transformers. arXiv preprint arXiv:2512.22323 (2025)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
49. Rosenfeld, A., Pfaltz, J.L.: Sequential operations in digital picture processing. Journal of the ACM (JACM) **13**(4), 471–494 (1966)
50. Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. Journal of the American Statistical association **88**(424), 1273–1283 (1993)
51. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8871–8879 (2024)
52. Simsar, E., Tonioni, A., Xian, Y., Hofmann, T., Tombari, F.: Lime: Localized image editing via attention regularization in diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 222–231. IEEE (2025)
53. Soille, P., et al.: Morphological image analysis: principles and applications, vol. 2. Springer (1999)
54. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
55. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
57. Wallace, B., Gokul, A., Naik, N.: Edict: Exact diffusion inversion via coupled transformations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22532–22541 (2023)
58. Wang, J., Chen, Z., Yuan, B., Zheng, K., Li, C., Jiang, Y., Zhu, J.: Audiomog: Guiding audio generation with mixture-of-guidance. In: ICME (2026)
59. Wang, Y., Chen, Z., Chen, X., Zhu, J., Chen, J.: Framebridge: Improving image-to-video generation with bridge models. In: ICML (2025)
60. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
61. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: The Thirteenth International Conference on Learning Representations (2024)
62. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
63. Yan, Z., Ma, Y., Zou, C., Chen, W., Chen, Q., Zhang, L.: Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17474–17484 (2025)

64. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
65. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
66. Zhang, C., Chen, Z., Zheng, K., Zhu, J.: Voicebridge: Designing latent bridge models for general speech restoration at scale. arXiv preprint arXiv:2509.25275 (2025)
67. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
68. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
69. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
70. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024 (2024)
71. Zhu, T., Zhang, S., Shao, J., Tang, Y.: Kv-edit: Training-free image editing for precise background preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16607–16617 (2025)