MASKFLOW: PRECISE, CONSISTENT AND SEAMLESS REGIONAL IMAGE EDITING

Anonymous authors
Paper under double-blind review



Figure 1: Qualitative comparison of the proposed **MaskFlow** with representative reference-based and inpainting methods. MaskFlow precisely localizes edits to masked regions, consistently preserves background content, and produces seamless transitions across mask boundaries.

ABSTRACT

Regional image editing has attracted considerable attention for its spatial controllability. Although instruction-based and mask-based editing methods can achieve strong semantic alignment, reliable regional control remains challenging, where an edit must be accurately localized and naturally integrated with the preserved context. We propose **MaskFlow**, a training framework for precise localization, consistent background preservation, and seamless boundary transitions. MaskFlow incorporates the mask into the probability path and flow-matching objective, coordinating generation within the editable region with source preservation outside it. The proposed Soft-Poisson De-seaming module further refines the predicted vector field during both training and sampling to improve the smooth integration of the edited foreground with the preserved background. We also introduce MaskEdit-Benchmark for general scene and infographics editing, where prompts describe the desired edits without localization cues, leaving masks to specify the target regions. Experiments on natural scenes and infographic images demonstrate consistent improvements over competing methods in both quantitative and qualitative evaluations.

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.

1 INTRODUCTION

Recent advances in diffusion and flow-matching models have substantially improved the fidelity and controllability of image generation (Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2022; Peebles & Xie, 2023; Wu et al., 2025). These developments have stimulated growing interest in conditioning mechanisms and controllable generation, particularly in regional image editing, which aims to modify regions specified by the user while preserving the remaining content (Hertz et al., 2022; Brooks et al., 2023; Zhang et al., 2026).

Existing approaches to regional image editing can be broadly categorized according to how the editable region is specified. Instruction-based methods identify the target content from textual instructions, whereas mask-based methods use explicit masks as spatial references. Benefiting from advances in large language models and vision language models (Radford et al., 2021; Achiam et al., 2023; Team et al., 2023; Bai et al., 2025), instruction-based image editors have demonstrated strong semantic understanding and editing capabilities (Koh et al., 2023; Zhou et al., 2025; Huang et al., 2024; Xu et al., 2025), especially when the target can be described unambiguously. However, as illustrated in Figure 1(a,b), specifying an exact target through language becomes cumbersome in complex scenes such as infographic images (Peng et al., 2025; Tang et al., 2026). These images often contain repeated or visually similar elements, dense layouts, or targets whose locations are difficult to describe precisely. Such spatial ambiguity can substantially reduce localization accuracy. Moreover, instruction-based methods commonly predict noise or vector fields over the entire spatial representation without explicit regional constraints (Wu et al., 2025; Labs, 2025; Deng et al., 2025). Consequently, they may unintentionally alter background content that should remain unchanged.

Mask-based methods alleviate spatial ambiguity by allowing users to specify the editable region explicitly (Avrahami et al., 2023; Wang et al., 2023; Xie et al., 2023; Zhuang et al., 2024; Zhou et al., 2026). Given a source image and a mask that identifies the editable region, existing methods typically encode the mask as an additional image condition and process it in a manner similar to the source image. Common strategies introduce mask information through shared attention layers (Vaswani et al., 2017), additional conditioning branches (Zhang et al., 2023; Mou et al., 2024), or parameter efficient fine-tuning (Hu et al., 2022; Zhang et al., 2026). Despite their improved localization capability, the mask-based methods continue to face challenges in edit fidelity and image integration. Generating the masked region independently can lead to semantic or stylistic drift (Rombach et al., 2022; Xie et al., 2023), which reduces the reliability of the edited content, may creating visible seams along the mask boundary and produce suboptimal results, as shown in Figure 1(c). Taken together, *these limitations call for a regional editing framework that provides precise localization, consistent background preservation, and seamless boundary transitions*.

To address these challenges, we propose **MaskFlow**, a training and inference framework for regional image editing. Given a source image, an editing instruction, and a mask of arbitrary shape, MaskFlow firstly samples noise and predict the vector field conditioned on source and mask images. Then, it calculates the clean image and conducts Soft-Poisson De-seaming to produce a refined vector field. The refined vector is used to update the next noisy latent features. Specifically, the proposed MaskFlow solves the problems through the following designs. **(1)** The localization capability is enhanced by encouraging the model to identify the editable regions from the masks via heuristic training on the constructed dataset. **(2)** The background is preserved with the proposed Siamese Probability Paths that recover the source image background while guide foreground editing. **(3)** The seamless boundary is refined by the designed Soft-Poisson De-seaming module that eliminates the visible discontinuities between the foreground and background by solving Poisson equation iteratively. As demonstrated in Figure 1(a), MaskFlow accurately replaces all text instances specified by the masks. Figure 1(b) further shows that semantically related background content remains unchanged. The results in Figure 1(c) illustrates more natural transitions than methods that directly replace masked regions. **(4)** We introduce MaskEdit-Benchmark, a training and evaluation dataset for mask-based image editing, that pairs region masks with concise referential instructions, decoupling spatial localization from editing intent to better reflect users' preference for concise prompts once the editing region is specified. The dataset further covers infographic editing, using masks to specify local regions that are difficult to describe textually and supporting the evaluation of models' ability to jointly interpret spatial guidance and editing instructions.

Contributions. This work addresses localization, background preservation, and seamless transitions in regional image editing. Our contributions are summarized below.

- We propose **MaskFlow**, a training and inference framework for regional image editors. Guided by the mask, the Siamese Probability Paths combines the foreground and background generation with clean contexts preview, and Soft-Poisson De-seaming refines the vector fields by iterative optimization at both training and sampling stages.
- We introduce MaskEdit-Benchmark, covering general scenes and infographics with masks and referential instructions that omit object identification and location descriptions. It supports training and evaluation of joint mask and text understanding for localized editing.
- Extensive qualitative and quantitative experiments demonstrate that the proposed approach performs the requested edits correctly while improving localization accuracy, background consistency, and boundary alignment.

2 Related Work

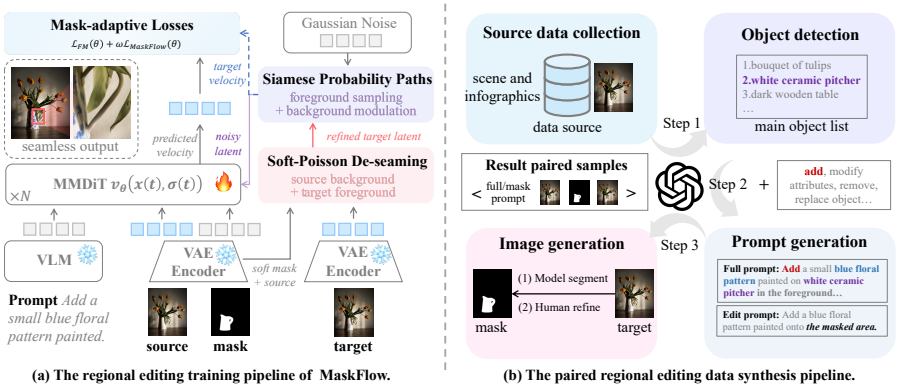
Diffusion models (Ho et al., 2020; Rombach et al., 2022) and flow-matching models (Lipman et al., 2022; Liu et al., 2022), together with scalable transformer backbones (Peebles & Xie, 2023; Ma et al., 2024), provide the generative foundation for regional edit models. A model must accurately localize, keep the unmasked area unchanged, and integrate the generated foreground with its surrounding context. Existing work mainly addresses these requirements through mask-based editing and refinement at region boundaries.

Mask-based regional editing. Masks provide explicit spatial information that is difficult to express through language alone. Early diffusion inpainting methods either blend generated and source latents during denoising or condition the model on masked images, object shapes, and high resolution source content (Avrahami et al., 2023; Wang et al., 2023; Xie et al., 2023). PowerPaint (Zhuang et al., 2024) further uses task prompts to unify filling, insertion, removal, and shape guided inpainting. Recent generalist editors (Mao et al., 2025; Li et al., 2026) encode masks within long context multimodal inputs or combine semantic target recognition with dedicated inpainting branches. Task specific methods (Jiang et al., 2025; Wasserman et al., 2025) improve removal or insertion through guidance from masked regions and purpose built training pairs. RegionE (Chen et al., 2025) and SpotEdit (Qin et al., 2026) instead infer editable or stable tokens from intermediate features and selectively reuse background features. Despite their different implementations, these approaches generally treat the mask or estimated region as a conditioning signal rather than a constraint on the underlying probability path. In contrast, **MaskFlow** incorporates the user mask into both the probability path and the flow-matching objective. This design directly constrains edits with masks of arbitrary shapes while maintaining the source trajectory outside the mask.

Blending and refinement at boundaries. Classical Poisson editing (Pérez et al., 2003) and its modified formulation (Tanaka et al., 2012) reconstruct a pasted region by matching source gradients under destination boundary conditions. Deep Image Blending (Zhang et al., 2020) augments this principle with learned content and style objectives. These methods can suppress abrupt pixel discontinuities, but they operate on a completed composite and may introduce color propagation or fail to repair semantic and structural incompatibility created during generation. Recent approaches move closer to generative models. Long et al. (2025) uses trajectory divergence and feature injection to localize shape changes, whereas RefineAnything (Zhou et al., 2026) and PixPerfect (Zheng et al., 2026) learn refiners that operate at boundaries or in pixel space. However, automatically inferred regions do not provide exact user control. Auxiliary losses and subsequent refinement also cannot directly correct intermediate generative decisions that produce an incompatible foreground. The proposed **MaskFlow** instead applies Soft-Poisson De-seaming at every sampling step to refine the predicted vector field, allowing boundary consistency to influence the entire generation trajectory.

3 Preliminaries

Flow-matching. Define the initial noise distribution $p_1 = \mathcal{N}(0, I)$ and target data distribution $p_1 = p_{\text{data}}$, flow-matching models (Lipman et al., 2022) learn vector field to map the random



(a) The regional editing training pipeline of MaskFlow. (b) The paired regional editing data synthesis pipeline.

Figure 2: Overview of the proposed **MaskFlow** framework and data synthesis pipeline. **(a)** During training, the source image and mask are provided to the MMDiT for vector field prediction. **(b)** The three-stage data synthesis pipeline uses an VLM and an image generation model to produce `<prompt,source,mask,target>` samples.

variables $X_1 \sim p_1$ to $X_0 \sim p_0$. Rectified flow (Liu et al., 2022) parameterizes the probability path using $x_\sigma = (1-\sigma)x_0 + \sigma\epsilon$ with continuous time-shifted $\sigma \in [0,1]$ (Wu et al., 2025). The conditional flow-matching loss can be written as

$$\mathcal{L}_{\text{CFM}}(\boldsymbol{\theta}) = \mathbb{E}\left[\|v_{\boldsymbol{\theta}}(x_\sigma, \sigma) - v_\sigma(x_0, \epsilon|x_\sigma)\|_2^2\right], \tag{1}$$

where $v_{\boldsymbol{\theta}}(x_0, \epsilon|x_\sigma) = \frac{dx_\sigma}{d\sigma} = \epsilon - x_0$ is the conditional velocity while $v_{\boldsymbol{\theta}}(x_\sigma, \sigma)$ is the predicted velocity. Given pre-trained vector predictor $v_{\boldsymbol{\theta}}(x_\sigma, \sigma)$ and the inference timestep sequence $\{T, T-1, \cdots, 1\}$, the predicted clean target at timestep t is $\hat{x}_0^t = x_{\sigma_t} - \sigma_t v_{\boldsymbol{\theta}}(x_{\sigma_t}, \sigma_t)$, where the continuous σ distribution is discretized by timesteps for iterative denoising.

4 METHODOLOGY

We present the proposed MaskFlow framework in Section 4.1 and Soft-Poisson De-seaming (SPD) in Section 4.2. A novel benchmark for mask-based image editing constructed by the data synthesis pipeline will be introduced in Section 4.3.

4.1 MASKFLOW

We now introduce the MaskFlow framework for training and inference. The proposed framework integrates freestyle mask in probability path as well as the supervision losses. Siamese Probability Paths (SPP) provide cleaner background context to guide foreground generation, while the MaskFlow loss strengthens supervision within the editable foreground region.

Siamese Probability Paths. To begin with, we use same flow-matching notations in Section 3 with σ-based probability path $\boldsymbol{x}_\sigma = (1-\sigma)\boldsymbol{x}_0 + \sigma\epsilon$, where $\boldsymbol{x}_0$ denotes the edited image and $\boldsymbol{x}_1 = \epsilon$ is the noise. To eliminate the unwilling background changes, we split the image into the editable region and the background context region. Then, the siamese probability paths are generated with

$$\boldsymbol{x}_\sigma = \boldsymbol{x}_\sigma^{\widetilde{\boldsymbol{M}}} + \boldsymbol{x}_{\sigma,\gamma}^{1-\widetilde{\boldsymbol{M}}}, \tag{2}$$

where $\boldsymbol{x}_\sigma^{\widetilde{\boldsymbol{M}}} = \widetilde{m} \odot ((1-\sigma)\boldsymbol{x}_0^\star + \sigma\epsilon)$ is the mask-cropped rectified flow probability path and $\boldsymbol{x}_0^\star$ is the boundary refined image using SPD. The second item $\boldsymbol{x}_{\sigma,\gamma}^{1-\widetilde{\boldsymbol{M}}}$ is parameterized with $\gamma > 1$ using the background context coming from the source image:

$$\boldsymbol{x}_{\sigma,\gamma}^{1-\widetilde{\boldsymbol{M}}} = (1-\widetilde{m}) \odot ((1-\sigma^\gamma)\boldsymbol{x}_S + \sigma^\gamma\epsilon). \tag{3}$$

As $\gamma > 1$ and $\sigma \in [0,1]$, the background has more cleaner structure and semantic than the editable region for better guiding the generation process.

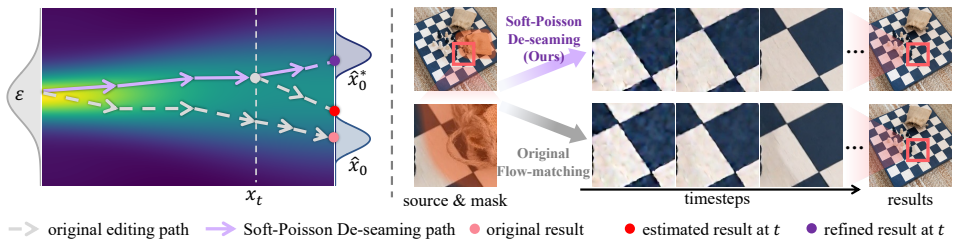


Figure 3: Probability paths for Soft-Poisson De-seaming (SPD) and conventional flow-matching. **Left.** SPD corrects the vector field at time t and moves the sample toward the seamless distribution. **Right.** The upper result from our method contains fewer boundary artifacts.

MaskFlow Loss. With the siamese probability paths, the target vector field is $\boldsymbol{v}_\sigma = \frac{\mathrm{d}\boldsymbol{x}_\sigma}{\mathrm{d}\sigma} = \widetilde{m} \odot (\epsilon - \boldsymbol{x}_0^\star) + (1 - \widetilde{m}) \odot \gamma\sigma^{\gamma-1}(\epsilon - \boldsymbol{x}_S)$. Then the conditional flow-matching loss Equation 1 can be conducted on this target and model predicted vector. However, the proposed MaskFlow is designed to suit masks with different shapes and sizes provided by users. Supervised solely on the vector field defined on the whole image has weak signals on small size masks. To overcome the problem, we design a mask-adaptive weighted loss to add additional supervision on the masked area

$$\mathcal{L}_{\text{MaskFlow}}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{HW}{\sum_{H,W}\widetilde{\boldsymbol{M}}}\widetilde{m} \odot \|(\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_\sigma, \sigma) - \boldsymbol{v}_\sigma)\|_2^2\right], \tag{4}$$

the loss weight is sample-wise and adaptive to the area of soft mask, balancing the contribution between small and large masks. The overall supervision loss is $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{\text{CFM}}(\boldsymbol{\theta}) + \omega_{\widetilde{\boldsymbol{M}}}\mathcal{L}_{\text{MaskFlow}}(\boldsymbol{\theta})$, where $\omega_{\widetilde{\boldsymbol{M}}}$ is a loss weight to adjust the scale of MaskFlow loss.

Training and Inference. As shown in Figure 2(a), training begins by applying SPD to the edited target to obtain the boundary-refined image $\boldsymbol{x}_0^\star$. The noisy input is then constructed using the siamese probability paths in Equation 2, with the source background following Equation 3. Conditioned on the source image, mask, and editing instruction, the model predicts the vector field and is optimized with the overall supervision loss $\mathcal{L}(\boldsymbol{\theta})$. During inference, at timestep t, the model predicts $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_{\sigma_t}, \sigma_t, \boldsymbol{c})$ under the given conditions $\boldsymbol{c}$. The predicted clean sample is computed as $\widehat{\boldsymbol{x}}_0^t = \boldsymbol{x}_{\sigma_t} - \sigma_t \boldsymbol{v}_{\boldsymbol{\theta}} + C(\sigma_t, \gamma)$, where the bias correction $C(\sigma_t, \gamma)$ is derived in Section B.4. SPD then refines this estimate to obtain $\widehat{\boldsymbol{x}}_0^{t,\star}$, and the correction $\frac{1}{\sigma_t}(\widehat{\boldsymbol{x}}_0^t - \widehat{\boldsymbol{x}}_0^{t,\star})$ is added to the predicted vector field to produce $\boldsymbol{v}_{\boldsymbol{\theta}}^\star$ for the next denoising step. The complete algorithms are provided in Section C. Both stages thus rely on Soft-Poisson De-seaming to integrate the edited foreground with the source background, and we will introduce it in the following section.

4.2 Soft-Poisson De-seaming

This section introduces **Soft-Poisson De-seaming** (SPD), which refines the predicted vector field at every training and sampling step through a unified gradient-domain objective as shown in Figure 3. Starting from the binary mask $\boldsymbol{M} \in \{0,1\}^{H\times W}$ where $\boldsymbol{M}_{i,j} = 1$ denotes the editable token, the soft mask is obtained by Gaussian blur $\widetilde{\boldsymbol{M}} = \text{GaussianBlur}(\boldsymbol{M})$. The expanded editable region is $\widetilde{\Omega} = \left\{\boldsymbol{p} = (i,j) \in \{1,\dots,H\} \times \{1,\dots,W\} \,\middle|\, \widetilde{\boldsymbol{M}}_{i,j} > 0\right\}$. Given the source image $\boldsymbol{x}_S$ and a candidate image $\boldsymbol{y}$ (such as the edited image and the predicted clean image), the target of SPD is to paste the masked region from $\boldsymbol{y}$ to $\boldsymbol{x}_S$ with a soft and seamless manner especially in the boundary.

Note that the tokens in the boundary between the editable and background areas is covered by the soft mask, whose values are modulated by the soft mask $\widetilde{\boldsymbol{M}}$. Let $\widetilde{m}(\boldsymbol{p}) = \widetilde{\boldsymbol{M}}_{i,j}$ for $\boldsymbol{p} = (i,j)$. We obtain the boundary refined image by solving

$$\begin{aligned}\boldsymbol{z}^\star = \arg\min_{\boldsymbol{z}} &\int_{\widetilde{\Omega}} \widetilde{m}(\boldsymbol{p}) \|\nabla_{\boldsymbol{p}}\boldsymbol{z}(\boldsymbol{p}) - \nabla_{\boldsymbol{p}}\boldsymbol{y}(\boldsymbol{p})\|_F^2 \,\mathrm{d}\boldsymbol{p} + \int_{\widetilde{\Omega}} \widetilde{m}(\boldsymbol{p}) \|\boldsymbol{z}(\boldsymbol{p}) - \boldsymbol{y}(\boldsymbol{p})\|_2^2 \,\mathrm{d}\boldsymbol{p} \\ &+ \int_{\widetilde{\Omega}} (1 - \widetilde{m}(\boldsymbol{p})) \|\boldsymbol{z}(\boldsymbol{p}) - \boldsymbol{x}_S(\boldsymbol{p})\|_2^2 \,\mathrm{d}\boldsymbol{p} \qquad \text{s.t.} \quad \boldsymbol{z}|_{\partial\widetilde{\Omega}} = \boldsymbol{x}_S|_{\partial\widetilde{\Omega}},\end{aligned} \tag{5}$$

where the final solved $z^\star$ is the blended image with refined seamless boundary. The first term weights gradient alignment by the soft mask to preserve structure while reducing its strength in the transition region. The second term anchors the refined field to the regions with large soft-mask values. The third term progressively restores the source feature as the soft-mask value decreases. Together with the boundary condition, these terms form a continuous transition from the generated foreground to the preserved background within a single optimization objective.

In practice, we solve Equation 5 with Euler–Lagrange equation and discretize it for efficient numerical computation. Let $\mathrm{N}(\boldsymbol{p})$ be the set of neighboring pixels of $\boldsymbol{p} \in \widetilde{\Omega}$, and let $\widetilde{m}_{\boldsymbol{p}}$ denote the corresponding soft-mask value. Define $w_{\boldsymbol{pq}} = (\widetilde{m}_{\boldsymbol{p}} + \widetilde{m}_{\boldsymbol{q}})/2$ and $J_{\boldsymbol{p}} = \sum_{\boldsymbol{q}\in\mathrm{N}(\boldsymbol{p})} w_{\boldsymbol{pq}}$, the discrete equation using Jacobi iteration is

$$\boldsymbol{z}_{\boldsymbol{p}}^{(k+1)} = \frac{1}{J_{\boldsymbol{p}}+1}\Big[\sum_{\boldsymbol{q}\in\mathrm{N}(\boldsymbol{p})\cap\bar{\Omega}} w_{\boldsymbol{pq}}\boldsymbol{z}_{\boldsymbol{q}}^{(k)} + \sum_{\boldsymbol{q}\in\mathrm{N}(\boldsymbol{p})\setminus\bar{\Omega}} w_{\boldsymbol{pq}}\boldsymbol{x}_{S,\boldsymbol{q}} + \sum_{\boldsymbol{q}\in\mathrm{N}(\boldsymbol{p})} w_{\boldsymbol{pq}}\left(\boldsymbol{y}_{\boldsymbol{p}} - \boldsymbol{y}_{\boldsymbol{q}}\right) + \widetilde{m}_{\boldsymbol{p}}\boldsymbol{y}_{\boldsymbol{p}} + \left(1-\widetilde{m}_{\boldsymbol{p}}\right)\boldsymbol{x}_{S,\boldsymbol{p}}\Big]. \tag{6}$$

The complete derivation is provided in supplementary Section B. In practice, $\boldsymbol{z}$ is updated using relaxed Jacobi iteration. After K iterations, the refined image is approximated by $\boldsymbol{y}_{\boldsymbol{p}}^\star \approx \boldsymbol{z}_{\boldsymbol{p}}^{(K)}$.

4.3 MaskEdit Benchmark

We introduce MaskEdit-Benchmark in this section to support training and evaluation by decoupling spatial localization from edit instructions. It includes editing samples from both general scenes and infographics. As shown in Figure 2(b), the pipeline contains object detection, prompt generation, and image generation. Given source images containing multiple concepts and potentially editable objects, **object detection** identifies the salient objects in each image. As there may be many visually insignificant objects whose editing provides limited value for training, we prompt a vision language model to identify the primary objects in each source image. The **prompt generation** stage produces two instructions for regional editing. The first is a complete instruction that specifies the editing operation, target position, and desired result. The second replaces the explicit position description with a demonstrative expression, encouraging the model to obtain localization information from the mask during training. The final **image generation** stage synthesizes target images and constructs their masks. We use the complete instructions with detailed positions to generate the target images. SAM (Kirillov et al., 2023) then provides an initial segmentation of each target region and then the human annotators refine these regions and create masks of arbitrary shapes that resemble practical user inputs. Dataset composition and counting conventions are detailed in Section A.

5 Experiments

5.1 Experimental settings

Implementation details. We use QwenImage-2511 (Wu et al., 2025) as the default backbone and additionally evaluate MaskFlow on FLUX.2-dev (Labs, 2025). We train attention LoRA modules (Hu et al., 2022; Vaswani et al., 2017) for 5K steps with rank 256 using the Prodigy optimizer (Mishchenko & Defazio, 2023) and an adaptive learning rate of 1.0. We set the MaskFlow loss weight to $\omega_{\widetilde{M}} = 2.0$ and the background path exponent to $\gamma = 1.5$. Soft-Poisson De-seaming uses $K = 50$ Jacobi iterations Equation 6. During inference, we use 50 sampling steps and a classifier-free guidance scale of 4.0 (Ho & Salimans, 2022).

Evaluation metrics. We evaluate editing quality and background preservation on the general scene and infographics subsets of MaskEdit-Benchmark. CLIP-I (Radford et al., 2021) and DINO (Oquab et al., 2023) measure semantic and structural similarity between edited results and target images. CLIP-C (Radford et al., 2021) evaluates the foreground semantic similarity. FID (Heusel et al., 2017) evaluates distributional fidelity, while PSNR (Korhonen & You, 2012), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) assess pixel, structural, and perceptual fidelity. Background consistency is measured against the source image in unmasked regions using $\mathrm{SSIM}_{\mathrm{bg}}$ and $\mathrm{LPIPS}_{\mathrm{bg}}$.

Table 1: Editing performance and background consistency on mask-based editing tasks. **Bold** is the best and underline is the second best within each task. Δ denotes the gains from our framework.

(a) General scene image editing

	Edit Performance							Background Consistency	
Model	CLIP-I ↑	CLIP-C ↑	DINO ↑	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	$SSIM_{bg}$ ↑	$LPIPS_{bg}$ ↓
Gemini 3 (Team et al., 2023)	0.9715	0.8726	0.9533	20.32	21.55	0.7184	0.1990	0.8606	0.1048
GPT Image 2 (Achiam et al., 2023)	0.9728	0.9111	0.9361	24.16	16.74	0.5039	0.2594	0.6074	0.1819
RefineAnything (Zhou et al., 2026)	0.9625	0.9321	0.9060	26.35	20.85	**0.7435**	0.1856	**0.9999**	**0.0001**
RegionE (Chen et al., 2025)	0.96545	0.8637	0.9314	25.89	20.55	0.7065	0.2465	0.8690	0.1378
SpotEdit (Qin et al., 2026)	0.9199	0.8807	0.8899	56.80	10.83	0.3049	0.6113	0.4161	0.5122
QwenImage-Inpaint (Lugmayr et al., 2022)	0.9714	0.9226	0.9403	23.07	21.41	0.7380	0.1489	0.9161	0.0440
BAGEL-7B-MoT (Deng et al., 2025)	0.9436	0.8969	0.9056	27.69	20.30	0.7201	0.2488	0.8810	0.1382
HiDream-O1-Image (Cai et al., 2026)	0.9116	0.8329	0.8423	45.16	13.09	0.3732	0.5623	0.4887	0.4544
FLUX.2-dev (Labs, 2025)	0.9546	0.8932	0.9138	29.96	19.20	0.6473	0.2296	0.7911	0.1246
+ MaskFlow (ours)	0.9669	0.9136	0.9291	24.11	20.94	0.7397	0.1216	**0.9999**	**0.0001**
Δ(%)	+1.29	+2.28	+1.67	+19.53	+9.06	+14.27	+47.04	+26.41	+100.0
QwenImage-2511 (Wu et al., 2025)	0.9455	0.8672	0.9103	29.83	18.80	0.6208	0.2804	0.7425	0.1793
+ MaskFlow (ours)	**0.9790**	**0.9408**	**0.9579**	**18.44**	**21.95**	0.7424	**0.1210**	**0.9999**	**0.0001**
Δ(%)	+3.54	+8.49	+5.23	+38.18	+16.76	+19.59	+56.85	+34.68	+100.0

(b) Infographics image editing

	Edit Performance							Background Consistency	
Model	CLIP-I ↑	CLIP-C ↑	DINO ↑	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	$SSIM_{bg}$ ↑	$LPIPS_{bg}$ ↓
Gemini 3 (Team et al., 2023)	0.9806	0.8874	0.9581	26.12	19.31	0.8129	0.1165	0.9005	0.0553
GPT Image 2 (Achiam et al., 2023)	0.9558	0.8693	0.9508	37.81	14.85	0.6367	0.3390	0.6824	0.2942
RefineAnything (Zhou et al., 2026)	0.9846	0.8836	0.9595	22.36	19.96	0.8509	0.0818	0.9998	0.0002
RegionE (Chen et al., 2025)	0.9622	0.8449	0.9278	37.76	17.38	0.8044	0.1767	0.8972	0.1133
SpotEdit (Qin et al., 2026)	0.9729	0.8583	0.9527	27.10	18.79	0.8228	0.1342	0.9164	0.0686
QwenImage-Inpaint (Lugmayr et al., 2022)	0.9871	0.8973	0.9686	20.80	20.49	0.8534	0.0875	0.9620	0.0193
BAGEL-7B-MoT (Deng et al., 2025)	0.9662	0.8536	0.9464	31.85	17.99	0.8123	0.1582	0.9113	0.0919
HiDream-O1-Image (Cai et al., 2026)	0.9303	0.8324	0.8722	61.73	12.16	0.5580	0.4368	0.6188	0.3844
FLUX.2-dev (Labs, 2025)	0.9689	0.8696	0.9388	36.06	16.40	0.7373	0.1588	0.8041	0.1037
+ MaskFlow (ours)	0.9720	0.8750	0.9578	25.37	19.51	0.8444	0.0939	**0.9999**	**0.0001**
Δ(%)	+0.32	+0.62	+2.02	+29.65	+18.96	+14.53	+40.87	+24.36	+100.0
QwenImage-2511 (Wu et al., 2025)	0.9693	0.8765	0.9395	37.95	17.49	0.7921	0.1387	0.8710	0.0805
+ MaskFlow (ours)	**0.9903**	**0.9137**	**0.9758**	**17.20**	**21.10**	**0.8575**	**0.0725**	**0.9999**	**0.0001**
Δ(%)	+2.17	+4.24	+3.86	+54.68	+20.64	+8.26	+47.73	+14.81	+100.0

5.2 IMAGE EDITING COMPARISONS

Quantitative analysis. Table 1 shows that MaskFlow delivers strong editing quality together with consistent background preservation across both tasks. On general scenes, MaskFlow with QwenImage-2511 leads in semantic similarity, structural similarity, distributional fidelity, PSNR, and perceptual fidelity, while remaining competitive with the strongest SSIM result. On infographics, it achieves the best results across all reported metrics, demonstrating its effectiveness in images with dense layouts and intertwined visual and textual content. Both MaskFlow variants attain the optimal reported background scores, showing that high-quality edits can be achieved while faithfully retaining unmasked content. The improvements also hold across backbones. Applying MaskFlow to either FLUX.2-dev or QwenImage-2511 improves every reported metric on both tasks, with particularly substantial gains in distributional and perceptual fidelity. These consistent improvements demonstrate that the framework strengthens regional editing across different base models and image domains.

Qualitative analysis. Figure 4a illustrates MaskFlow's precise regional control across modification, removal, replacement, and addition. In scenes with similar objects, competing methods may modify the wrong instance or spread the edit beyond the mask. MaskFlow recolors the selected motorcycle and replaces the selected banana with an apple while preserving neighboring objects. It also removes the masked item from the tabletop and places a lemon slice at the specified location with natural integration into the scene. Figure 4b highlights the same advantages in more demanding layouts. MaskFlow accurately updates the masked heading, replaces the selected food item, and inserts the requested vase while retaining surrounding text and illustrations. Other methods exhibit misspelled text, changes to unrelated elements, or inaccurate placement. Together, these examples show that MaskFlow combines reliable instruction following with precise localization, background preservation, and coherent visual integration. More visual comparisons are provided in Appendix E.



(a) Qualitative comparisons on general scene image editing task



(b) Qualitative comparisons on infographics image editing task

Figure 4: Qualitative comparisons on general scenes and infographics. MaskFlow accurately edits the masked regions while preserving surrounding content and maintaining natural visual transitions. Competing methods exhibit inaccurate localization, unintended background changes, or inconsistent edited content. Please zoom in for details.

5.3 Ablation Studies

Effectiveness of the proposed modules. Table 2 evaluates the progressive addition of Soft-Poisson De-seaming (SPD), Siamese Probability Paths (SPP), and MaskFlow loss (MFL). The baseline is QwenImage-2511 fine-tuned on the same data for the same number of training steps as the other configurations. Introducing SPD yields the improvement over this baseline, substantially increasing foreground CLIP-C while improving CLIP-I and reducing FID. Adding SPP provides further gains across all three metrics. Incorporating MFL further improves foreground semantic similarity, as reflected in CLIP-C, and produces the best results across all three metrics. This progression supports the complementary roles of boundary refinement, coordinated foreground and background trajectories, and focused supervision of editable regions.

The visual ablation in Figure 6 further clarifies these roles. The fine-tuned baseline alters the surrounding bedding when replacing the masked subject with a golden retriever puppy and recolors birds outside the mask. MaskFlow without SPD better localizes the edits, but visible discontinuities remain in the bedding texture and sky near the edited regions, as highlighted in the enlarged crops. The full model integrates the puppy more naturally with the bedding and recolors the selected bird while preserving the surrounding sky and other birds. These results show that SPD improves boundary continuity while maintaining regional control.

Localization precision. We examine whether additional textual position descriptions are necessary when an explicit mask is provided. Both settings use editing instructions and masks, the proposed MaskFlow omits extra position descriptions during training, whereas the comparison model includes them. Table 3 shows that omitting these descriptions yields consistent improvements across all reported metrics. This supports our training design, where the mask provides spatial guidance,

Table 2: Ablation of the proposed modules.

SPD	SPP	MFL	CLIP-I ↑	CLIP-C ↑	FID ↓
-	-	-	0.9726	0.9206	21.46
✓	-	-	0.9775	0.9377	19.06
✓	✓	-	0.9788	0.9380	18.91
✓	✓	✓	**0.9790**	**0.9408**	**18.44**

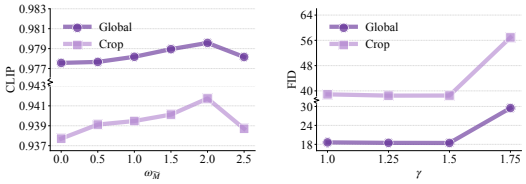


Figure 5: Parameter sensitivity of MaskFlow.

Table 3: Effect of *additional* textual position descriptions on regional image editing given the masks.

Additional Position	CLIP-I ↑	DINO ↑	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
✓	0.9779	0.9567	17.81	21.08	0.6993	0.1680
✗ **(Ours)**	**0.9798**	**0.9595**	**17.18**	**21.27**	**0.7001**	**0.1664**



Figure 6: Qualitative ablation comparing the fine-tuned baseline, MaskFlow without SPD, and the full framework. Red arrows indicate visible errors. Please zoom in for details.

allowing the instruction to focus on the desired edit. MaskFlow thus achieves effective regional control without requiring users to formulate detailed location descriptions.

5.4 Sensitivity of MaskFlow Loss and Siamese Probability Paths

Figure 5 evaluates the loss weight $\omega_{\widetilde{M}}$ and background path exponent γ on full images (Global) and foregrounds cropped using the masks (Crop). As shown in Figure 5, increasing $\omega_{\widetilde{M}}$ improves CLIP-I, particularly on foreground crops, with both evaluations peaking at 2.0. A larger weight reduces performance, highlighting the importance of balancing regional and global supervision. In SPP γ settings, FID remains favorable up to $\gamma = 1.5$ but deteriorates beyond it in both evaluations. This suggests that moderately faster background denoising provides useful context for foreground generation, while excessive separation between the paths weakens editing quality. We therefore use $\omega_{\widetilde{M}} = 2.0, \gamma = 1.5$, the best setting in the comparisons.

6 Conclusion

This paper presents **MaskFlow**, a training and inference framework for regional image editing with precise localization, consistent background preservation, and seamless boundary transitions. MaskFlow incorporates the mask into the flow-matching probability path and uses a mask-aware objective to improve localization accuracy and preserve unmasked content. Training on the prompts without additional position descriptions pushes the model to learn localization from masks, further strengthening the attentions to editable regions. The proposed Soft-Poisson De-seaming module refines the vector field during sampling to produce smooth transitions between edited and preserved regions. Moreover, the constructed MaskEdit-Benchmark provides a new regional editing dataset covering both general scene and infographics editing tasks for further study. Quantitative and qualitative experiments demonstrate the effectiveness of MaskFlow on both natural scenes and infographic images, showing promise for practical application.

AI USE STATEMENT

We used AI tools for language refinement and synthetic data construction. For manuscript preparation, large language models assisted with sentence rephrasing, grammar correction, and improvements to readability and clarity. For data construction, AI models supported salient object identification, editing instruction generation, and image synthesis in the pipeline described in Section 4.3. The resulting MaskEdit-10k dataset includes COCO-derived natural scene images and synthetically generated infographics, as described in Appendix A. The authors take full responsibility for the content of this work, including the text, claims, and data produced with AI assistance.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our research focuses on regional image editing using COCO-derived natural scene images and synthetically generated infographics, with data sources and construction procedures described in the paper. We are committed to respecting the applicable usage guidelines and intellectual property rights of source datasets and models, protecting privacy, and maintaining transparency and integrity throughout the research process. We recognize that source data and generative models may contain biases and that image editing methods can be misused to produce misleading content. We encourage responsible use of MaskFlow and MaskEdit-Benchmark and discourage applications that violate privacy, promote discrimination, or facilitate deception.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the MaskFlow framework and Soft-Poisson De-seaming in Sections 4.1 and 4.2, respectively. The training configuration, inference settings, and evaluation metrics are reported in Section 5.1. Appendix B provides detailed mathematical derivations, and Appendix C presents the training and inference algorithms. The data synthesis pipeline is described in Section 4.3, with dataset sources, composition, and training/test split sizes documented in Appendix A. The code is included in the supplementary materials.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Pengtao Chen, Xianfang Zeng, Maosen Zhao, Mingzhu Shen, Peng Ye, Bangyin Xiang, Zhibo Wang, Wei Cheng, Gang Yu, and Tao Chen. Regione: Adaptive region-aware generation for efficient image editing. *arXiv preprint arXiv:2510.25590*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2024.

Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24452–24462, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36:21487–21506, 2023.

Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pp. 37–38, 2012. doi: 10.1109/QoMEX.2012.6263880.

Black Forest Labs. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, Junhao Zhuang, Ying Shan, Yuexian Zou, and Qiang Xu. Brushedit: All-in-one image inpainting and editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Zeqian Long, Mingzhe Zheng, Kunyu Feng, Xinhua Zhang, Hongyu Liu, Harry Yang, Linfeng Zhang, Qifeng Chen, and Yue Ma. Follow-your-shape: Shape-aware image editing via trajectory-guided region control. *arXiv preprint arXiv:2508.08134*, 2025.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1958–1966, 2025.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Yuyang Peng, Shishi Xiao, Keming Wu, Qisheng Liao, Bohan Chen, Kevin Lin, Danqing Huang, Ji Li, and Yuhui Yuan. Bizgen: Advancing article-level visual text rendering for infographics generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23615–23624, June 2025.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22 (3):313–318, July 2003. ISSN 0730-0301.

Zhibin Qin, Zhenxiong Tan, Zeqing Wang, Songhua Liu, and Xinchao Wang. Spotedit: Selective region editing in diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18683–18692, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Masayuki Tanaka, Ryo Kamio, and Masatoshi Okutomi. Seamless image cloning by a closed form solution of a modified poisson problem. In *SIGGRAPH Asia 2012 Posters*, pp. 1–1. 2012.

Yinghao Tang, Xueding Liu, Boyuan Zhang, Tingfeng Lan, Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, et al. Igenbench: Benchmarking the reliability of text-to-infographic generation. *arXiv preprint arXiv:2601.04498*, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18313–18324, 2025.

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 preprint arXiv:2508.02324*, 2025.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22428–22437, 2023.

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

Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 231–240, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. *Advances in Neural Information Processing Systems*, 38:139195–139227, 2026.

Haitian Zheng, Yuan Yao, Yuqian Zhou, Jiebo Luo, Zhe Lin, et al. Pixperfect: Seamless latent diffusion local editing with discriminative pixel-space refinement. *Advances in Neural Information Processing Systems*, 38:128266–128292, 2026.

Dewei Zhou, You Li, Zongxin Yang, and Yi Yang. Refineanything: Multimodal region-specific refinement for perfect local details. *arXiv preprint arXiv:2604.06870*, 2026.

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

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2024.