FlowFixer: Towards Detail-Preserving Subject-Driven Generation

Jinyoung Jun[1,2*] Won-Dong Jang[1] Wenbin Ouyang[1] Raghudeep Gadde[1] Jungbeom Lee[2†]
[1]Amazon [2]Korea University
{jyjun, wdjang, wenbinoy}@amazon.com raghudeep.g@gmail.com jbeomlee@korea.ac.kr

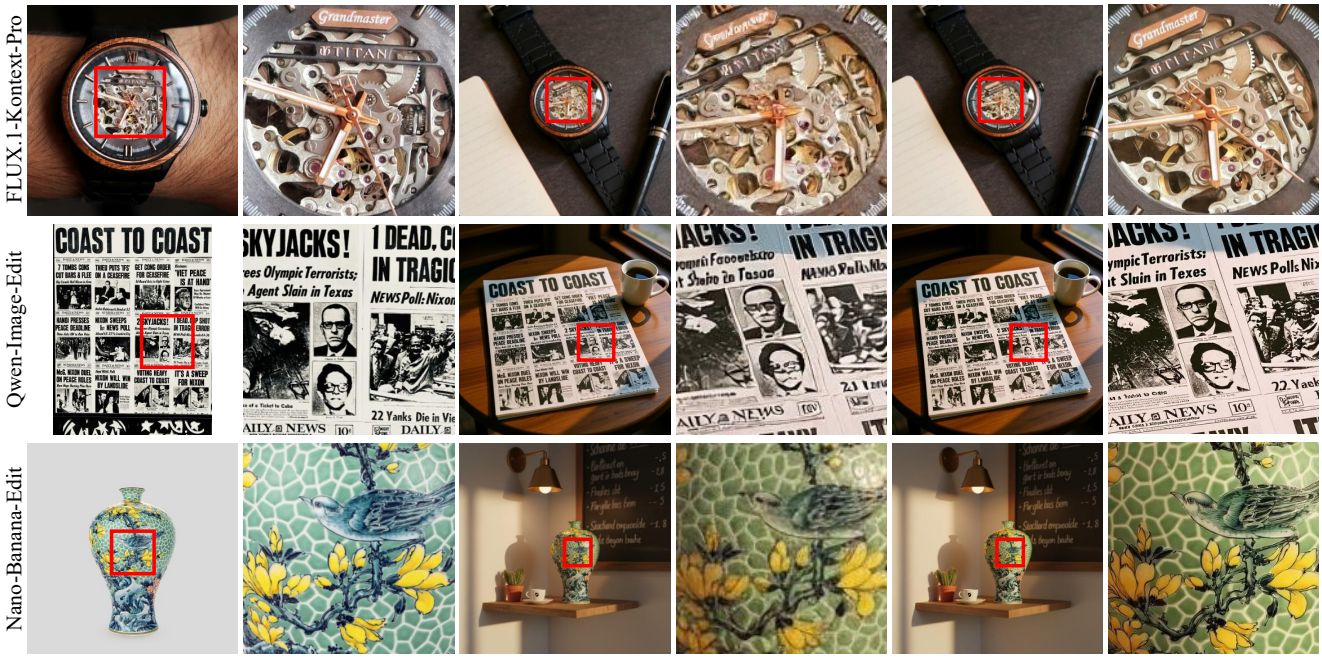

Figure 1. **Detail enhancement on FLUX.1-Kontext-Pro [21], Qwen-Image-Edit [3], and Nano-Banana-Edit [6] using our FlowFixer.** The red boxes indicate the zoomed-in regions. Compared to the baseline subject-driven generations, FlowFixer restores the fine details from the reference, such as complex structures (top and bottom), small text (top and middle), and human identity (middle). It also handles challenging cases involving rotation (top), viewpoint changes (middle), and color shifts (bottom), while preserving the overall scene composition. FlowFixer is a baseline-agnostic, prompt-free model designed to enhance subject fidelity without altering the global layout.

Abstract

*We present **FlowFixer**, a refinement framework for subject-driven generation (SDG) that restores fine details lost during generation caused by changes in scale and perspective of a subject. FlowFixer proposes direct image-to-image translation from visual references, avoiding ambiguities in language prompts. To enable image-to-image training, we introduce a one-step denoising scheme to generate self-supervised training data, which automatically removes high-frequency details while preserving global structure, effectively simulating real-world SDG errors. We further propose a keypoint matching-based metric to properly assess fidelity in details beyond semantic similarities usually measured by CLIP or DINO. Experimental results demonstrate that FlowFixer outperforms state-of-the-art SDG methods in both qualitative and quantitative evaluations, setting a new benchmark for high-fidelity subject-driven generation.*

1. Introduction

Subject-driven generation (SDG) aims to embed a given subject (or an input reference image) into imagery described

*Work done during internship at Amazon.
†Corresponding author.

by an input text prompt while preserving the subject's identity. SDG has received significant attention from the community since it has a number of practical applications, including advertising content generation, short-form content generation, and personalized media creation.

Recent foundation models [8, 32] have shown promising improvements in handling subjects with simpler textures (*e.g.*, animals or plain objects) [40, 41, 45–47]. However, preserving complex product-specific details, such as logos, text, and intricate patterns, remains a critical challenge that demands greater attention from the community. This is particularly important for commercial applications where structural fidelity of product details directly impacts the utility of generated content. In advertising, for example, altered logos undermine brand recognition, and distorted text makes the outputs unusable.

There are two key obstacles underlying this difficulty. First, collecting high-quality paired training data for SDG is challenging. Ideally, one would need pairs of subject images and diverse ground truth images containing the same subject to supervise both fidelity and compositional diversity. In practice, however, collecting such data at scale is highly challenging. To address this scarcity, Subjects200K [40] was introduced, yet it is constructed from synthetic images, which often lack fine-grained and realistic alignment of subject details.

Second, existing conditioning mechanisms are often limited in specifying fine-grained geometric and appearance variations of the subject. Text descriptions such as 'a red sports car' or 'a cereal box' convey only coarse appearance and provide limited cues about pose, orientation, or lighting, making precise reproduction of subject details challenging [22, 38]. Even with image-based conditioning (e.g., depth or edge maps), they tend to prioritize global scene coherence over localized detail alignment, which can lead to the loss of high-frequency information in texture-rich or geometrically complex regions [40, 47, 50].

To overcome these challenges, we propose FlowFixer, a novel refinement framework for detail-preserving SDG. Our approach employs a direct image-to-image translation pipeline that learns from visual references. This design choice circumvents the ambiguity inherent in natural language descriptions, enabling precise preservation of diverse visual elements and fine structural details across the image, as illustrated in Figure 1.

At the core of FlowFixer is a self-supervised refinement scheme that leverages pseudo-paired training data. In principle, training a subject refinement framework requires triplets consisting of a clean subject image, a corresponding SDG-generated image, and its ideal ground-truth for refinement. However, collecting such paired data at scale is impractical due to the high cost of annotating subject-scene correspondences and generating controlled SDG outputs.

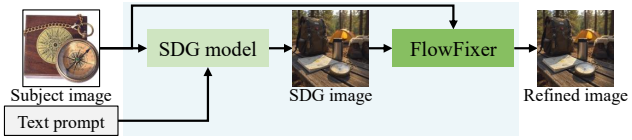

Figure 2. **FlowFixer overview.** FlowFixer enhances SDG images by restoring fine subject details, using the original subject image as reference.

Instead of collecting triplet data for training, we employ a self-supervised approach centered on our one-step denoising strategy. Starting with a clean real image, we synthetically generate its degraded counterpart through a forward diffusion step followed by single-step denoising using an off-the-shelf diffusion model. This process closely mimics SDG artifacts and characteristic distortions, allowing FlowFixer to learn fine detail restoration without expensive human supervision. The resulting framework enables efficient training using web-collected single images while faithfully representing the high-frequency detail loss typical in SDG applications.

Current quantitative evaluation metrics for SDG results have distinct characteristics and constraints. For example, pixel-level similarity measures (e.g., MSE or SSIM) focus primarily on low-level differences, while semantic-level metrics (e.g., FID [13] or LPIPS [51]) may not fully capture fine details. Moreover, many existing metrics require ground truth images, which are often unavailable in real-world generative applications. To overcome these limitations, we propose detail-aware evaluation metrics based on keypoint matching [16, 23]; absolute keypoint increase and keypoint matching gain. These metrics effectively capture structural fidelity by measuring the consistency between a reference image and its generated output, enabling ground-truth-free quantitative evaluation of detail preservation in open-world generative settings. Together with human and VLM evaluation, our metrics provide reliable and reproducible assessments of fine-grained fidelity.

Through extensive experiments, we demonstrate that FlowFixer consistently outperforms existing SDG methods in preserving subject identity, establishing a new baseline for high-fidelity SDG. The key contributions of this work are summarized as follows:

- A novel model-agnostic refinement framework, FlowFixer, that substantially enhances subject fidelity in SDG-generated images.
- An efficient training data curation pipeline based on one-step denoising, which effectively simulates diffusion artifacts to generate high-quality pseudo-paired training data.
- A direct visual translation approach that leverages reference images, enabling precise preservation of visual elements and fine-grained details while eliminating prompt-induced ambiguity.

- A novel ground-truth-free evaluation metric to assess visual fidelity based on keypoint matching, which demonstrates FlowFixer's superior detail preservation capability compared to existing SDG methods.

2. Related Work

Subject-driven generation has received significant attention from the community and builds directly on top of the success of text-to-image foundational diffusion models [20, 21, 37]. While text-to-image models can generate high quality objects, subject driven generation requires faithful rendering of the "subject" (i.e, preserving the identity in the subject image) in a variety of scenes.

Techniques for subject driven generation have broadly followed two main directions (a) fine-tuning-based and (b) encoder-based. Early approaches such as DreamBooth [38], Textual Inversion [9], and LoRA [15] in Custom Diffusion [19] adapt pre-trained text-to-image diffusion models to specific subjects using only a few reference images (typically 3 to 5), achieving strong identity preservation but require expensive per-subject fine-tuning. More recent works avoid per-subject finetuning and address the limitation by injecting the reference image of the subject through an encoder directly into the diffusion backbone. For example, IP-Adapter [48] injects image-prompt features via decoupled cross-attention to enable subject conditioning without fine-tuning, while BLIP-Diffusion [22] learns a multimodal subject encoder for tighter subject–prompt alignment. OminiControl [40] further shows that a DiT backbone can encode references natively with minimal additional parameters.

However, these encoder-based approaches, while good at preserving high-level details, struggle to preserve the subject's fine structural details, rendering the synthesized images unusable for real-world applications. In contrast, FlowFixer restores missing low-level details to ensure high-fidelity identity preservation. It employs a reference-guided diffusion refinement process that corrects structural inconsistencies in a generated SDG image by conditioning on the original subject image, as illustrated in Figure 2. This "last mile" refinement makes FlowFixer universally compatible, enhancing identity preservation for any upstream model.

Another line of work that is relevant for subject driven image generation is based on image editing, which focuses on modifying an existing image under additional conditions such as text prompts, reference images, or spatial masks [28, 50]. Existing methods generally fall into two categories: global and local editing. Global editing techniques alter overall image appearance or semantic content through text-based manipulation or latent space interpolation [4, 5, 12, 17, 18, 26, 30, 31, 43]. On the other hand, methods for local editing target specific spatial regions via mask-based selection[25], blending [2], or exemplars [10]. Although effective for coarse transformations, these approaches often fail to preserve fine structural details of the subject and typically require manual inputs such as masks or region-specific prompts [7, 44, 49, 53]. Recent works further reveal that text-driven editors struggle with localized or fine-grained control due to ambiguous conditioning and conflicting attention dynamics [11, 27, 35, 42, 52].

In contrast, FlowFixer performs automatic, reference-guided refinement without requiring manual annotations or textual conditioning. By leveraging cross-image correspondences between the generated and reference images, FlowFixer restores degraded regions while preserving global scene coherence and sharp, subject-consistent details. To further encourage the model to focus on subjects, the proposed FlowFixer exploits automatic subject cropping based on keypoint matching between a subject image and its SDG image.

3. Method

3.1. Diffusion preliminaries

Diffusion models are probabilistic generative frameworks that progressively transform a simple prior p_s (*e.g.*, Gaussian noise) into samples from a target distribution p_t through iterative denoising or continuous flows. Let $\mathbf{x}_t$ (or $\mathbf{z}_t$ in latent diffusion) denote the state at time $t \in [0, 1]$ along this trajectory. The generative process starts from noise and can be formally expressed as

$$\mathbf{x}_1 \sim p_s, \qquad \mathbf{x}_0 = \mathcal{D}(\mathbf{x}_1) \sim p_t, \tag{1}$$

where $\mathcal{D}$ represents a learned denoising or flow-matching process [14, 24, 39]. A key property of diffusion models is their conditioning flexibility:

$$\mathbf{x}_0 = \mathcal{D}(\mathbf{x}_1, \mathbf{c}), \tag{2}$$

where $\mathbf{c}$ denotes auxiliary controls such as text prompts or reference images. In latent diffusion models [20, 37], $\mathbf{x}_t$ corresponds to a latent variable $\mathbf{z}_t$ encoded by a VAE $\mathcal{E}$, and the final image is reconstructed from the latent sample $\mathbf{z}_1$. Diffusion models have achieved highly realistic and semantically coherent image generation, driven by large-scale architectures and training on massive, diverse datasets.

3.2. Problem formulation

Subject-driven generation (SDG) can be viewed as a specific instance of conditional diffusion in Eq. 2. Given a subject reference image $\mathbf{I}_{\text{ref}}$ and a textual description $\mathbf{c}_{\text{text}}$, an SDG model $\mathcal{D}_{\text{SDG}}$ generates a novel scenic image $\mathbf{I}_{\text{gen}}$ conditioned on both inputs:

$$\mathbf{I}_{\text{gen}} = \mathcal{D}_{\text{SDG}}(\mathbf{z}_1, \mathbf{I}_{\text{ref}}, \mathbf{c}_{\text{text}}), \qquad \mathbf{z}_1 \sim p_s, \tag{3}$$

where $\mathbf{c}_{\text{text}}$ provides high-level semantics and $\mathbf{I}_{\text{ref}}$ encodes subject appearance cues.

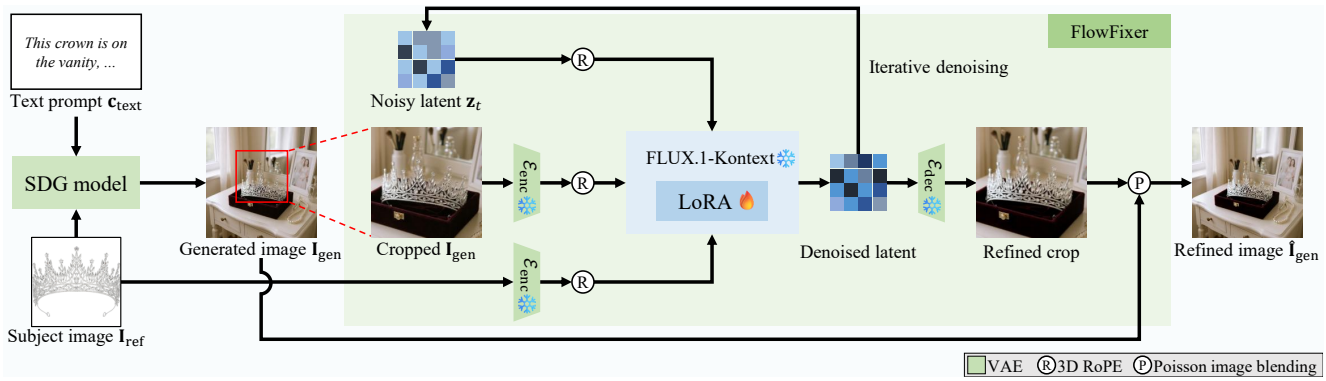

Figure 3. **FlowFixer inference pipeline.** The model takes two conditional inputs: reference subject image $\mathbf{I}_{\text{ref}}$ and the generated image $\mathbf{I}_{\text{gen}}$ from any SDG model. Then the model produces a refined result $\widehat{\mathbf{I}}_{\text{gen}}$ that preserves global layout. For faster inference, we optionally refine only a subject-centric crop of $\mathbf{I}_{\text{gen}}$ and blend it back using Poisson image blending.

While diffusion models achieve strong global realism and semantic consistency, text-conditioned variants often prioritize global coherence over local structural fidelity. This limitation stems from the ambiguity of textual conditioning, which captures broad semantics but lacks precise visual cues such as small textures or logos. As a result, diffusion models tend to favor perceptual plausibility at the expense of subject-specific details [11, 27, 35, 42, 52]. Despite notable progress in large-scale foundation models—including Qwen [3], FLUX.1-Kontext [21], and Nano Banana [6]—fine-grained subject fidelity remains a persistent challenge.

To address this, we design a text-free diffusion-based refiner $\mathcal{D}_{\text{refine}}$ that re-generates $\mathbf{I}_{\text{gen}}$ under the guidance of $\mathbf{I}_{\text{ref}}$ through a conditional diffusion process starting from latent noise $\mathbf{z}_1 \sim p_s$ as follows,

$$\widehat{\mathbf{I}}_{\text{gen}} = \mathcal{D}_{\text{refine}}(\mathbf{z}_1, \mathbf{I}_{\text{gen}}, \mathbf{I}_{\text{ref}}). \tag{4}$$

Here, $\widehat{\mathbf{I}}_{\text{gen}}$ preserves the global layout of $\mathbf{I}_{\text{gen}}$ while restoring subject-consistent details from $\mathbf{I}_{\text{ref}}$. Unlike conventional inpainting methods that rely on explicit masks or user interaction, Eq. 4 denotes fully automatic refinement without external inputs. Furthermore, $\mathcal{D}_{\text{refine}}$ is optimized with self-supervised pseudo pairs, enabling scalable and annotation-free enhancement beyond mask-based editing. We refer to this refiner as **FlowFixer**, reflecting its ability to restore fine structural consistency by correcting disrupted feature flow between $\mathbf{I}_{\text{gen}}$ and $\mathbf{I}_{\text{ref}}$.

3.3. Pseudo-paired training data

A key challenge in training $\mathcal{D}_{\text{refine}}$ is the lack of paired data where only subject details are degraded while global structure remains unchanged. To address this, we construct pseudo pairs $(\mathbf{I}_{\text{degraded}}, \mathbf{I}_{\text{clean}})$ from real images by mocking SDG's degradation using a one-step denoising process as follows:

Figure 4. **Example of one-step denoising distortions.** For each distortion level, pixel-wise variance maps are computed over 10 degraded samples. Insets show example outputs, with distortions concentrated in high-frequency regions.

1. Start from a clean real image $\mathbf{I}_{\text{clean}}$.
2. Add noise to $\mathbf{I}_{\text{clean}}$ and apply a single-step denoising using an off-the-shelf diffusion model [34].
3. Control the degradation level by downscaling $\mathbf{I}_{\text{clean}}$ to $1.0\times$, $0.5\times$, or $0.25\times$ its original resolution before VAE encoding.

To verify that this process mainly affects fine details, we generate 10 variants with different random seeds in step 2 and compute per-pixel variance maps across them. As shown in Figure 4, the variance concentrates in high-frequency regions while remaining low in smooth backgrounds, confirming that the perturbation minimally disturbs global structure. For step 2, we use SDXL [34].

During training, we treat $\mathbf{I}_{\text{degraded}}$ as the generated input $\mathbf{I}_{\text{gen}}$ in Eq. 4. The reference $\mathbf{I}_{\text{ref}}$ is a spatially perturbed version of the clean image $\mathbf{I}_{\text{clean}}$, created through random cropping, rotation, or mild color augmentation—and vice versa for data diversity. This setup enables $\mathcal{D}_{\text{refine}}$ to focus on recovering fine details by learning local correspondences between $\mathbf{I}_{\text{degraded}}$ and $\mathbf{I}_{\text{ref}}$, without depending on strict pixel-wise alignment.

3.4. Training pipeline

Network architecture. FlowFixer builds on FLUX.1-Kontext [21] to leverage image-native editing capability. For a text-free pipeline, we discard the original text token and introduce an additional image input, as illustrated in Figure 3. Consequently, FlowFixer takes three inputs, $\mathbf{z}_1$, $\mathbf{I}_{\text{gen}}$, and $\mathbf{I}_{\text{ref}}$. The images $\mathbf{I}_{\text{gen}}$ and $\mathbf{I}_{\text{ref}}$ are encoded by the pretrained VAE into latent tokens, which are concatenated with $\mathbf{z}_1$ before being processed by the DiT backbone. We adopt 3D RoPE with per-stream timestep offsets (0 for $\mathbf{z}_1$, 1 for $\mathbf{I}_{\text{gen}}$, 2 for $\mathbf{I}_{\text{ref}}$), to maintain stream separation while allowing full cross-attention.

To discover dense correspondences between $\mathbf{I}_{\text{gen}}$ and $\mathbf{I}_{\text{ref}}$, we adopt an explicit dual-stream conditioning mechanism operating in a shared spatial space. This design enforces alignment between the two inputs and facilitates localized refinements guided by the reference. The alignment is further strengthened by our self-supervised, pseudo-paired training scheme.

Implementation details. We fine-tune FLUX.1-Kontext using LoRA [15] with rank 192, specializing the model for automatic refinement while keeping the parameter overhead minimal. Training is conducted for 50K iterations with a batch size of 4. For each iteration, a pseudo training pair $(\mathbf{I}_{\text{degraded}}, \mathbf{I}_{\text{clean}})$ is sampled as described in Sec. 3.3. One degraded variant is randomly selected from the three downscaling levels $(1.0\times, 0.5\times, 0.25\times)$ to ensure balanced degradation diversity. The model is trained using a mean squared error (MSE) loss between the refined output and the clean target. We use a guidance scale of 1.0 during training and 2.5 at inference, respectively.

We use 18,450 high-quality real-world photographs from Unsplash [1] to construct the pseudo pairs for training. The images span diverse objects, materials, and lighting conditions, providing sufficient visual diversity for self-supervised refinement.

3.5. Crop-based refinement

While high-resolution generation is critical for subject fidelity, a full-resolution global pass incurs substantial latency and memory cost. Instead, FlowFixer preserves the background layout while selectively restoring subject details, enabling crop-based refinement during inference, as illustrated in Figure 3. We first determine a subject-centric crop using keypoint matching [16] between a subject and its generated image, and then refine only this region and paste the result back into the original image. Since the global structure remains unchanged and only fine details are corrected, simple image-domain blending (*e.g.*, Poisson blending) achieves seamless integration without user-defined masks or inversion. This substantially reduces runtime and memory while retaining subject-level fidelity.

4. Detail-aware Evaluation

4.1. Evaluation metric

While widely used, existing perceptual metrics fall short in evaluating fine-grained details. Common similarity measures, such as CLIP [36] or DINOv2 [29] primarily capture global semantics and overlook high-frequency fidelity, making them unsuitable for assessing detail consistency.

To better quantify subject fidelity, we exploit keypoint matching that finds dense correspondences between the reference and generated images. This approach is based on the observation that images with better subject fidelity yield a higher number of matched keypoints. We define two metrics: i) absolute keypoint increase (AKI) and ii) keypoint matching gain $\mathcal{K}_{\text{Gain}}$. First, we formulate AKI by

$$\text{AKI} = \mathcal{N}(\mathcal{M}(\mathbf{I}_{\text{ref}}, \widehat{\mathbf{I}}_{\text{gen}})) - \mathcal{N}(\mathcal{M}(\mathbf{I}_{\text{ref}}, \mathbf{I}_{\text{gen}})), \quad (5)$$

where $\mathcal{N}(\mathcal{M}(a, b))$ denotes the number of matched keypoints between a and b using the keypoint matching network $\mathcal{M}$. A higher AKI score indicates stronger preservation of subject-specific fine details and structural alignment.

While AKI effectively quantifies instance-level improvements, its absolute values depend on the choice and calibration of the keypoint matcher, and thus are not strictly comparable across settings. Moreover, when averaged over a large set, many small yet consistent improvements can be diluted by a few large changes, obscuring the overall trend. Therefore, we also calculate the keypoint matching gain, $\mathcal{K}_{\text{Gain}}$, which averages the fraction of cases that improve. Formally, we define

$$\mathcal{K}_{\text{Gain}} = \frac{1}{N} \sum_{i=1}^{N} \delta(\text{AKI}_i, \tau) \quad (6)$$

where δ is a binary indicator function that becomes 1 when AKI_i is higher than τ and 0 otherwise. AKI_i is an AKI score of i-th image sample in a dataset. We report $\mathcal{K}_{\text{Gain}}$ in percentage and set τ=0 as default. These metrics effectively capture enhancement of local fidelity. For evaluation, we employ an off-the-shelf keypoint matching network, OmniGlue [16].

4.2. Evaluation dataset

Existing SDG benchmarks [33, 38] primarily focus on global realism and semantic alignment rather than preserving fine-grained subject fidelity. As a result, their subject categories are often visually simple and contain limited texture or detail (*e.g.*, rubber ducks, plants, or cartoon figures). To achieve rigorous evaluation of subject fidelity, we introduce **FidelityBench-258K**, a large-scale, subject-diverse benchmark structured by subject–prompt pairs. To construct the dataset, we first collected 29K subject images and generated prompts for SDG using a vision-language

Table 1. **Refinement performance on the FidelityBench-258K.** For all metrics, higher numbers indicate better performance.

Method	FLUX.1-Kontext-Pro				Qwen-Image-Edit				Nano-Banana-Edit			
	AKI ↑	$\mathcal{K}_{\text{Gain}}$ ↑	CLIP-I ↑	DINO ↑	AKI ↑	$\mathcal{K}_{\text{Gain}}$ ↑	CLIP-I ↑	DINO ↑	AKI ↑	$\mathcal{K}_{\text{Gain}}$ ↑	CLIP-I ↑	DINO ↑
Baseline	-	-	0.776	0.663	-	-	**0.777**	**0.668**	-	-	**0.796**	**0.711**
Text-based editing [21]	7.5	52.7%	0.763	0.647	11.1	54.1%	0.762	0.647	61.7	77.0%	0.782	0.691
OminiControl [40] + F-Dev	29.0	53.9%	0.724	0.551	0.48	43.8%	0.724	0.552	**108.0**	70.7%	0.747	0.605
OminiControl [40] + F-Kontext	0.49	45.7%	0.766	0.649	-3.37	46.8%	0.765	0.649	53.0	56.6%	0.786	0.696
FlowFixer (ours)	**66.5**	**77.9%**	**0.778**	**0.668**	**54.0**	**74.8%**	**0.777**	**0.668**	64.7	**79.2%**	**0.796**	**0.711**

Figure 5. **Qualitative comparison on Subject fidelity refinement on the FidelityBench-258K dataset.** The insets in the full images show the reference subject images and the red and green boxes indicate the zoomed-in regions. The regions for zoomed-in views are found on the SDG baseline images and cropped the same area for all methods.

model (VLM), Claude 3.5. For each prompt-subject pair, we generated five variants per SDG baseline. We used three SDG baselines (FLUX.1-Kontext-Pro [21], Qwen-Image-Edit [3], and Nano Banana-Edit [6]), which led to 435K SDG images in total. Finally, we applied a coarse quality filter to ensure that the subject is clearly present in the SDG image. After the filtering, the final dataset consists of 258K subject - SDG image pairs.

For controlled and reproducible studies, we also curate a fixed subset, **FidelityBench-300** from the FidelityBench-258K dataset, collecting 100 images from each backbone. FidelityBench-300 preserves the distribution of baseline match counts while balancing categories, and we reuse this fixed subset for all ablations and human evaluations to ensure comparability and reproducibility. More detailed data curation protocol is provided in the supplemental document.

5. Experiments

5.1. Subject fidelity refinement

Table 1 reports refinement results on FidelityBench-258K dataset under the three SDG baselines (*i.e.*, FLUX.1-Kontext-Pro, Qwen-Image-Edit, and Nano-Banana-Edit). In Table 1, we compare four different refinement models, including the proposed FlowFixer. The compared refinement models are:

- **Text-based editing:** FLUX.1-Kontext, which is a text-based editing model, accepting the subject and SDG images concatenated on the x-axis while refinement is guided by an input prompt.
- **OminiControl + FLUX.1 (Dev/Kontext):** OminiControl [40] fine-tuned on FLUX.1-Dev and FLUX.1-Kontext, respectively, using our pseudo-paired dataset. We use the same training data as FlowFixer.

Since there is no algorithm tailored for refinement of SDG, we finetuned OminiControl [40] with state-of-the-art backbones [20, 21], which can accept a subject as a conditional input, and used the text-based editing method [21] for benchmarking.

Note that AKI and $\mathcal{K}_{\text{Gain}}$ are not obtainable for the baselines since these metrics quantify the changes relative to the baseline's SDG output. We additionally report CLIP-Image (CLIP-I) and DINOv2 similarity as complementary perceptual metrics. To isolate subject fidelity from background content, similarities are computed only on the subject regions by detecting the bounding box of the subject.

We summarize our statistical and empirical findings from Table 1 and Figure 5 as follows:

(i) As illustrated in Figure 5, FlowFixer restores fine details of the subject while preserving the original image layout. In contrast, the other refinement models either shift the scene or fail to improve local structure. For example, 'Text-based editing' often maintains semantics but alters composition, undermining subject consistency. In contrast, FlowFixer avoids such layout drift while increasing local correspondences.

(ii) Quantitatively, across all SDG backbones, FlowFixer consistently outperforms its alternatives in AKI and achieves an average $\mathcal{K}_{\text{Gain}}$ of 77.3%, demonstrating model-agnostic robustness.

(iii) Interestingly, these keypoint-based gains are not reflected in CLIP-I or DINOv2 scores, which remain nearly unchanged. This indicates that common perceptual metrics overlook fine-grained structural fidelity, reinforcing the need for specialized metrics like AKI and $\mathcal{K}_{\text{Gain}}$.

(iv) While alternative fine-tuned models (OminiControl + FLUX.1-Dev and Kontext) occasionally increase AKI, their $\mathcal{K}_{\text{Gain}}$ often drops below 50%, meaning such methods show no consistent pattern of improvement.

Table 2. **Refinement performance compared to original SDG images on the FidelityBench-300.** For all metrics, higher numbers indicate better performance.

Method	AKI ↑	$\mathcal{K}_{\text{Gain}}$ ↑	VLM ↑
Text-based editing [21]	1.87	45.9%	41.3%
OminiControl [40] + F-Dev	22.7	46.6%	4.2%
OminiControl [40] + F-Kontext	11.1	38.4%	25.2%
FlowFixer (ours)	**67.3**	**91.2%**	**79.0%**

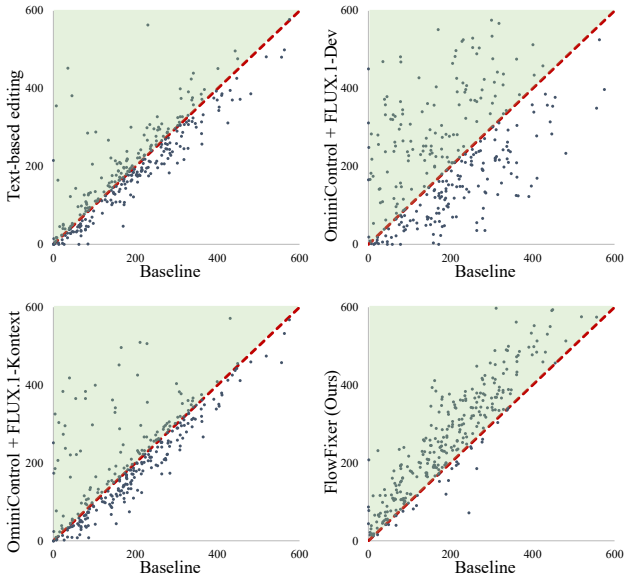

Figure 6. **Scatter plots of the number of keypoint matches on FidelityBench-300.** Each dot represents a sample; points above the red dashed line indicate an increase in keypoint matches (positive AKI, green region), suggesting improved structural alignment and subject fidelity. Samples below the line show decreased correspondence after refinement.

(v) On Nano Banana, certain methods achieve inflated keypoint metrics by copy-pasting the subject or synthesizing a new scene with larger subject rather than refining the given output. This results in the elevated AKI scores but reduced global consistency, as reflected in the lower CLIP-I and DINOv2 similarities on cropped subject regions. Representative examples for copy-pasting are shown in the supplementary document.

Figure 6 shows scatter plots of keypoint changes before and after refinement on FidelityBench-300. Among all methods, only FlowFixer reveals a consistent and directional pattern, reliably increasing the number of matched keypoints (AKI) across most samples. In contrast, alternative methods exhibit no clear trend, with improvements occurring sporadically and often accompanied by regressions. This further highlights the robustness and generalizability of FlowFixer's refinements.

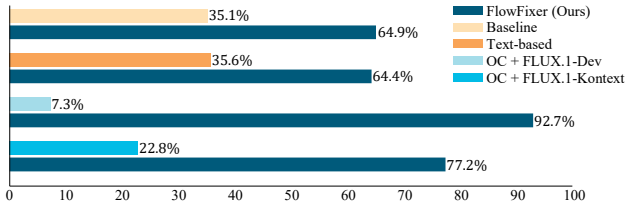

Figure 7. **A/B study results comparing the FlowFixer against four alternatives.** FlowFixer is consistently preferred by human raters.

Figure 8. **Impact of training distortion levels on refinement performance.** Using a range of distortion levels during training enhances the model's ability to handle diverse degradation at inference time, resulting in more robust restoration.

5.2. Human Evaluation and VLM Judgment

To assess how well our metrics reflect human perception, we conducted an A/B tests on the FidelityBench-300 subset using Amazon Mechanical Turk. For each test case, human evaluators were shown the reference subject alongside two candidate images, *i.e.*, FlowFixer vs. one alternative method. Then, we asked the evaluators to choose the one that better preserves subject-specific details. Each pair was evaluated by five independent evaluators, and responses were aggregated across the dataset.

Figure 7 shows that the human evaluators strongly prefer FlowFixer over the baseline and the other refinement methods, which align with our proposed metrics, AKI and $\mathcal{K}_{\text{Gain}}$. Notably, FlowFixer's advantages over the baseline and the text-based editing [21] are comparable (64.9% and 64.4%), suggesting that a text prompt for editing only makes a negligible difference in terms of subject fidelity. Moreover, FlowFixer is favored over OminiControl variants [40] by even larger margins (92.7% and 77.2%).

In addition to the human evaluation, we also evaluate metric agreement with a Vision-Language Model (VLM), Claude 3.7, serving as an automated judge. For each case, the VLM receives the reference image and two subject-region crops (Baseline vs. one alternative). To mitigate order bias, we present two images in both A/B and B/A orders and average the decisions.

As shown in Table 2, VLM judges FlowFixer to be the best restoration method in terms of subject fidelity. In addition to that, VLM judgments exhibit strong alignment with AKI and $\mathcal{K}_{\text{Gain}}$, cross-validating their effectiveness in capturing perceptual improvements in subject fidelity.

Figure 9. **Efficacy of cropped refinement in comparison with full image refinement.** While full image refinement moderately enhances subject fidelity, cropping further improves legibility.

5.3. Distortion levels for training

To assess the impact of degradation diversity during training, we compare FlowFixer models trained with different subsets of distortion levels: (i) only slight noise ($1.0\times$), (ii) moderate and slight noise ($1.0\times$, $0.5\times$), and (iii) the full range ($1.0\times$, $0.5\times$, $0.25\times$). As illustrated in Figure 8, including various levels of distortion during training significantly boosts robustness, especially under large-scale artifacts, highlighting the importance of diverse degradation simulation for effective refinement.

5.4. Crop-based refinement

Figure 9 compares a single global refinement pass against our crop-based refinement strategy. In both cases, the global scene composition remains unchanged, highlighting FlowFixer's inherent stability with respect to layout drift. Notably, even under the same evaluation resolution, crop-based refinement yields more accurate recovery of fine-grained subject details, thanks to its focused and localized processing. This allows better fidelity in details without compromising global coherence.

6. Conclusion

We introduced FlowFixer, a model-agnostic detail refiner for subject-driven generation that recovers fine structural details while preserving global layout. Trained on self-supervised pseudo pairs simulating high-frequency degradation, FlowFixer scales to in-the-wild references without paired subject–scene data. Our text-free, direct image-to-image formulation avoid prompt ambiguity and consistently improve fidelity. Paired with keypoint-matching-based metrics for ground-truth-free evaluation, FlowFixer demonstrates superior performance across diverse SDG methods. Future directions include (i) multi-reference refinement that leverages multiple reference images, and (ii) user-interactive correction using auxiliary control signals, such as scribble masks.

References

[1] Unsplash. https://unsplash.com/. 5

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 3

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 4, 6

[4] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *CVPR*, 2024. 3

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 3

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 4, 6

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3

[10] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Trans. Graph.*, 43(4):1–15, 2024. 3

[11] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *CVPR*, 2024. 3, 4

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 3

[13] 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. In *NeurIPS*, 2017. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 5

[16] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *CVPR*, 2024. 2, 5

[17] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 3

[18] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 3

[19] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[20] Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024. 3, 7

[21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX. 1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 1, 3, 4, 5, 6, 7, 8

[22] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 2, 3

[23] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, 2023. 2

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 3

[25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3

[26] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 3

[27] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DiffEditor: Boosting accuracy and flexibility on diffusion-based image editing. In *CVPR*, 2024. 3, 4

[28] 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 *AAAI*, 2024. 3

[29] 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. 5

[30] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH conference proceedings*, 2023. 3

[31] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *ICCV*, 2023. 3

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[33] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. DreamBench++: A human-aligned benchmark for personalized image generation. In *ICLR*, 2025. 5

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 4

[35] Yuming Qiao, Fanyi Wang, Jingwen Su, Yanhao Zhang, Yunjie Yu, Siyu Wu, and Guo-Jun Qi. BARET: Balanced attention based real image editing driven by target-text inversion. In *AAAI*, 2024. 3, 4

[36] 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 *ICML*, 2021. 5

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2, 3, 5

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[40] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *ICCV*, 2025. 2, 3, 6, 7, 8

[41] Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*, 2025. 2

[42] Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *CVPR*, 2025. 3, 4

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 3

[44] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, 2023. 3

[45] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2

[46] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *ICCV*, 2025.

[47] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *CVPR*, 2025. 2

[48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[49] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 3

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[52] 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 *CVPR*, 2025. 3, 4

[53] 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 *ECCV*, 2024. 3