– Supplemental Materials –
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

A. Dataset curation protocol

Data collection. To construct FidelityBench-258K, we first collected 29K product-style subject images from publicly available Amazon product pages, each accompanied by its short textual subject description. We then used a vision–language model (VLM), Claude 3.5, to synthesize SDG prompts from these descriptions. The VLM was instructed to (i) keep the subject identity strictly fixed (type, color, materials, functionality) and (ii) diversify only the surrounding scene, lighting, and photographic style, producing multiple scene-level captions per subject. For each subject–prompt pair, we generate five SDG variants per backbone (FLUX.1-Kontext-Pro, Qwen-Image-Edit, Nano-Banana-Edit), leading to 435K SDG images before filtering.

Prompt generation protocol. We constructed a structured instruction prompt and feed it to the VLM to generate multiple SDG captions per subject, using the model as a description augmenter that turns each original subject description into diverse, scene-level prompts. The instruction enforces two core constraints:

- **Subject accuracy**: the subject type, name, colors, materials, size, and functionality must remain unchanged;
- **Scene diversity**: the surrounding environment, background objects, lighting, season, and photography style should vary across captions (indoor/outdoor, different rooms, different times of day, etc.).

Since our SDG backbones were image editing models, we require each caption to explicitly mention the subject using "`this [subject_name]`" so that the prompt can reliably refer to the subject image for editing. For each subject, the VLM returned a set of seven captions with different lengths and detail levels (one long scene description, three medium-length descriptions, and three concise descriptions), from which we sample prompts for SDG generation. To generate FidelityBench-258K, we used the one long scene description for SDG.

Instruction snippet.

```
You're an expert at creating imaginative and creative image generation prompts.
You are given a <subject_description>, and your task is to create diverse,
creative scenes featuring this exact subject in various contexts.

  CRITICAL RULES:
1.  The SUBJECT must remain exactly as described --- do not change its attributes.
2.  Scenes should be creative and varied.
3.  Every caption must contain ``this [subject_name]''.
  SUBJECT ACCURACY:
type, color, materials, scale, and functionality must match the original description.
  SCENE CREATIVITY:
vary location (indoor/outdoor), decor, lighting, time of day/season, and photography
style.
```

The full template and generation script will be released with the dataset.

Examples of FidelityBench-258K. Figure S1 presents representative examples consisting of (i) the input subject image, (ii) the VLM-generated prompt, and (iii) SDG outputs from all three backbones. These examples demonstrate how the VLM expanded the original product description into diverse scene-level prompts, and how different SDG backbones interpret the same prompt with varying degrees of subject visibility, scale, and fidelity.

Figure S1. **Examples of FidelityBench-258K.** For each subject image-prompt pair, SDG outputs from three backbones-FLUX.1-Kontext-Pro, Qwen-Image-Edit, and Nano-Banana-Edit-are provided.

Filtering using keypoint matching. To ensure that each subject–prompt pair leads to a valid SDG, we apply a quality filtering step based on keypoint matching between the subject image and the generated SDG output. Specifically, we compute the number of matched keypoints using OmniGlue [16] and discard generations whose global matching confidence falls below 0.2 or whose match count is fewer than 20. Generations that pass this filter are those where the subject is clearly visible and not overwhelmed by background clutter or extreme scaling. This fully automatic procedure removes cases where the SDG backbone fails to render the subject, places it at an extremely small scale, or produces scenes dominated by unrelated content.

Table S1 summarizes the number of retained images for each SDG backbone after filtering. In Table S1, we observed substantial variation across models. FLUX.1-Kontext-Pro and Qwen-Image-Edit generally preserve the subject at a reasonable scale. In contrast, Nano-Banana-Edit frequently renders the subject relatively small, as shown in Figure S1. As a result, a significantly larger portion of Nano-Banana-Edit generations falls below our visibility threshold, leading to fewer valid pairs in FidelityBench-258K.

Table S1. **Statistics of retained SDG images after quality filtering.** We report the number of generations that pass our keypoint-based subject visibility threshold for each SDG backbone.

Method	FLUX.1-Kontext-Pro					Qwen-Image-Edit					Nano-Banana-Edit					Sum
Samples	20,361	20,426	20,397	20,473	20,441	22,258	22,306	22,344	22,278	22,255	8,937	8,973	8,855	8,838	8,955	-
Total	102,098					111,441					44,558					258,097

B. Analysis

VLM-as-a-judge instruction. The full template and instruction script will be released with the dataset.

```
  ANALYSIS PROCESS:
1.  INDEPENDENTLY analyze each candidate image against the reference
2.  Document ALL differences for each candidate, focusing on visibility
3.  Classify and weigh each difference using this scoring system:
- CRITICAL (5 points):  Changes subject model/identity/brand
- MAJOR (3 points):  Alters key functionality or primary features
- MODERATE (2 points):  Changes notable design elements
- MINOR (1 point):  Subtle variations in color, texture, or non-defining details
4.  Calculate a weighted difference score for each image
5.  Determine which subject would cause greater disappointment

  PAY SPECIAL ATTENTION TO:
- Logo design, style, and placement
- Brand name presentation and typography
- Brand-specific design signatures or distinctive elements
- Color schemes associated with specific subject

  FINAL VALIDATION CHECKLIST:
- Have you identified all visible differences systematically?
- Did you properly distinguish between cosmetic variations and model-defining differences?
- Is your reasoning focused on subject identity rather than just counting differences?
- Does your conclusion align with the severity of differences you identified?
- Would this assessment protect humans from receiving a fundamentally different subject?

  INSTRUCTIONS:
- Focus ONLY on the subject
- ignore people, backgrounds, and usage contexts
- Analyze candidates independently to avoid comparative bias
- Prioritize differences that change subject identity over mere presentation variations
- Provide specific, observable evidence for your conclusions
- Consider the cumulative effect of all differences for each subject
- Remember:  The key question is which subject would most disappoint a human expecting
the reference

Follow the specified output format exactly, ensuring your reasoning clearly explains which
subject has differences that fundamentally change its identity compared to the reference.
```

Example of VLM-as-a-judge results. We employed Claude-3.7 as a VLM judge to evaluate which generated image exhibited greater differences compared to the reference image. The model demonstrated its ability to analyze and articulate visual distinctions, as illustrated in Figure S2. The results show that the VLM can effectively discern and describe subtle variations between images, providing detailed comparative analysis.

Subject image (reference) | Generated image (Image A) | + FlowFixer (Image B)

Figure S2. **Example of VLM-as-a-judge image triplet.**

VLM response for Figure S2.

```
  DIFFERENCES-IMAGE A:
Text reads 'WÆDING YOU A VERY' instead of 'WISHING YOU A VERY',
Text reads 'HAPT NEW YEAR' instead of 'HAPPY NEW YEAR',
The foliage pattern appears darker and more saturated than in the reference,
The red box has 'HOLIDAY CARDS' text at the bottom which is not visible in the reference,
The box appears to have a gold trim at the top edge

  DIFFERENCES-IMAGE B:
Text reads 'WISHING YOU A VERY' instead of 'WISHING YOU A VERY', (correct spelling but
different font)
The foliage pattern appears darker and more saturated than in the reference,
The red box has 'HOLIDAY CARDS' text at the bottom which is not visible in the reference,
The box appears to have a gold trim at the top edge

  REASONING:
Image A contains more critical differences from the reference subject, particularly
in the text content.
The misspellings ('WÆDING' instead of 'WISHING' and 'HAPT' instead of 'HAPPY')
significantly alter the meaning and readability of the card.
While both generated images have similar differences in color saturation and additional
elements like the 'HOLIDAY CARDS' text, Image A's text errors are more substantial changes
that affect the core message of the Christmas card.

  ANSWER:
Image A contains more critical differences.
```

Diffusion backbone for one-step denoising. Figure S3 illustrates one-step denoising examples using SDXL and FLUX.1-Kontext. We observed that both models apply distortion similarly, introducing high-frequency details while preserving the global scene structure. Based on this observation, we used SDXL to generate $\mathbf{I}_{\text{degraded}}$ for computational efficiency.

Figure S3. **Examples of one-step denoising using SDXL and FLUX.1-Kontext.** Both models successfully preserve the global scene structure while introducing various types of distortions.

Keypoint matching. Table S2 reports AKI and $\mathcal{K}_{\text{Gain}}$ computed with two off-the-shelf keypoint matchers, OmniGlue [16] and LightGlue [23], on FidelityBench-300. The absolute scale of AKI differs substantially across matchers: LightGlue produces significantly more correspondences due to its dense SuperPoint backbone, relative positional encoding, and adaptive inference, which collectively focus attention on high-frequency local structure. As a result, improvements brought by FlowFixer—designed to enhance fine-scale visual detail—lead to markedly higher AKI values under LightGlue.

OmniGlue, on the other hand, leverages coarse region-level guidance from a DINOv2 vision transformer and explicitly separates spatial position from appearance during matching. This makes it more conservative and semantically grounded, prioritizing robust correspondences aligned with global structure rather than purely local texture. Consequently, it yields fewer matches in high-frequency regions, but remains stable across domains. Despite these architectural differences, both matchers produce consistent method rankings and unanimously show that FlowFixer substantially improves subject-level correspondence quality, supporting the robustness of our keypoint-based evaluation.

Table S2. **Performance of different keypoint matching networks on FidelityBench-300.**

Method	OmniGlue [16]		LightGlue [23]	
	AKI	$\mathcal{K}_{\text{Gain}}$	AKI	$\mathcal{K}_{\text{Gain}}$
Text-based editing [21]	1.87	45.9%	10.1	46.6%
OminiControl [40] + FLUX.1-Dev	22.7	46.6%	14.0	49.7%
OminiControl [40] + FLUX.1-Kontext	11.1	38.4%	-54.5	16.2%
FlowFixer (ours)	**67.3**	**91.2%**	**143.3**	**99.3%**

Copy-pasting on OminiControl variants. Figure S4 illustrates the copy-pasting behavior of two OminiControl variants, 'OC+FLUX.1-Dev' and 'OC+FLUX.1-Kontext.' Since 'OC+FLUX.1-Dev' is trained based on a text-to-image model, it fails to perform refinement and tends to directly copy the subject image, generating a single instance of the subject with high matching confidence, which results in elevated AKI scores. On the other hand, 'OC+FLUX.1-Kontext,' trained based on an image editing model, preserves the original scene but struggles to effectively handle multiple conditions simultaneously, often generating multiple instances of the same subject within a single scene. This duplication reduces the matching confidence for each instance, leading to lower AKI scores. In contrast to both variants, FlowFixer consistently refines the image without altering the scene structure, demonstrating more stable and reliable performance across different conditions.

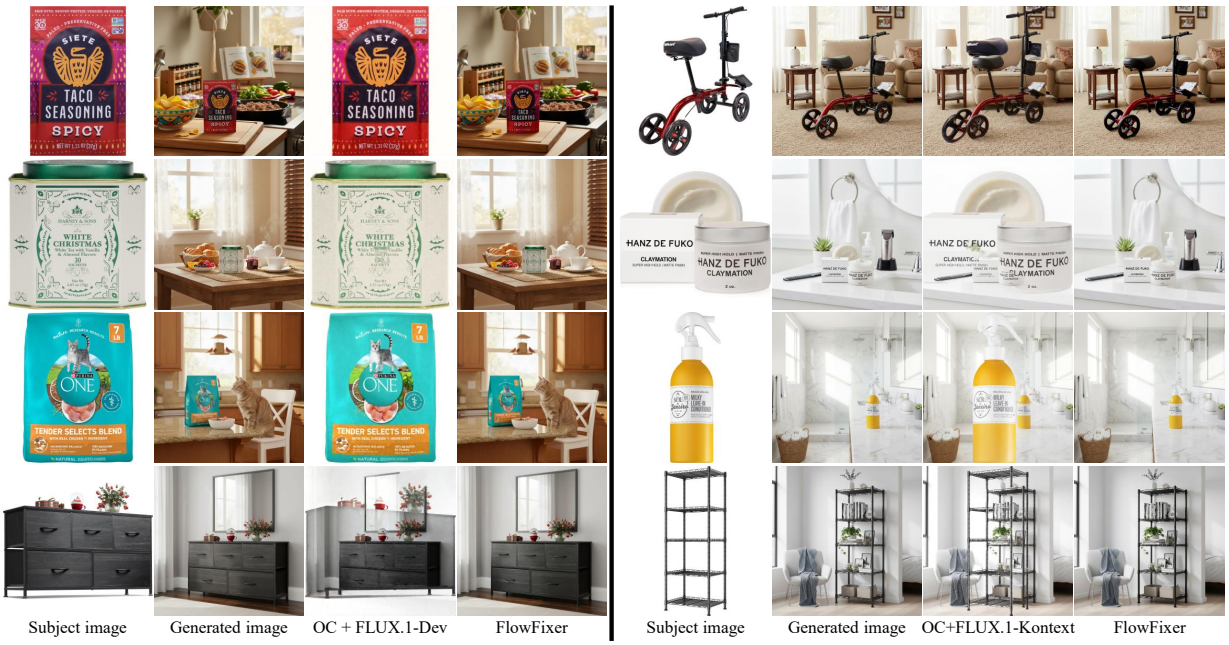

Figure S4. **Copy-pasting examples on OC+FLUX.1-Dev and OC+FLUX.1-Kontext.** Compared to other methods, FlowFixer consistently maintains the global scene structure while effectively refining subject details without introducing duplications or structural artifacts.

C. Qualitative examples

More visual results. Figure S5 illustrates the results before and after refinement using FlowFixer. We observe that FlowFixer restores various kinds of distortions, including typographical errors, missing parts, and texture artifacts.

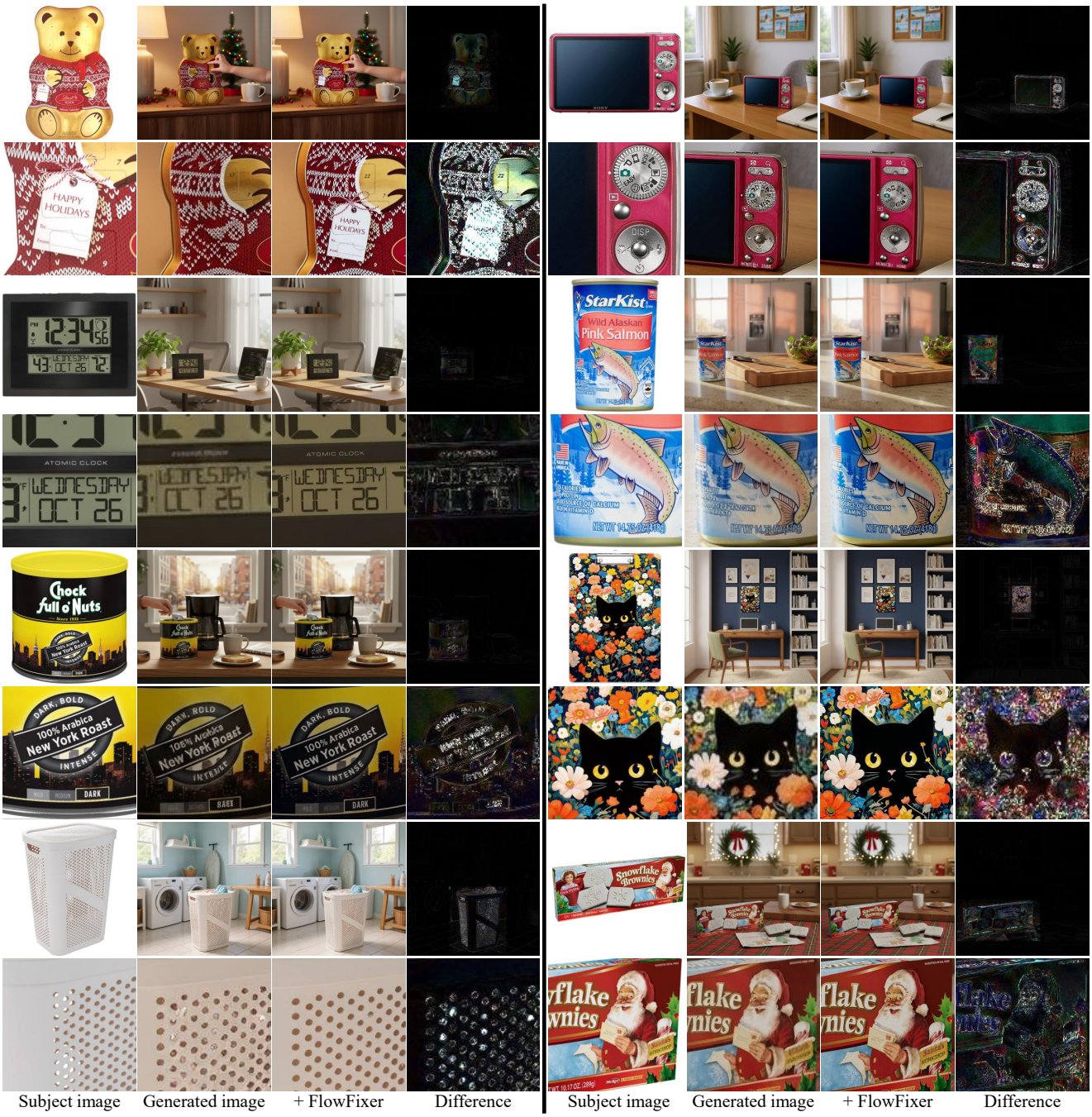

Figure S5. **Examples of FidelityBench-258K.** For each refinement, a difference map between the original and refined images is provided to highlight the modifications. FlowFixer effectively refines distorted subject details while preserving the global scene layout, demonstrating precise localized corrections without introducing structural artifacts.