Supplementary Material

This document provides: (i) detailed discussion of limitations (Section A); (ii) generalization experiments of Character Position Encoding (CPE) on multilingual text-to-image synthesis (Section B); (iii) qualitative comparison of subject extension mitigation strategies (Section C); (iv) data efficiency analysis investigating minimal data requirements for eliminating subject extension (Section D); (v) automatic evaluation metrics for subject preservation using DinoV3 (Section E); and (vi) complete text prompts for visualized examples in the main paper (Section F).

A. Limitations

While SimplePoster achieves strong performance in product poster generation, several limitations remain, falling into three main categories:

Dependency on segmentation and matting quality. Our framework assumes the input product image has a clean white background. When this is not the case, we rely on off-the-shelf segmentation and matting models to extract the foreground. However, inaccurate extraction—such as over-segmentation (including non-product regions) or under-segmentation (truncating parts of the product)—can degrade generation quality. This issue is particularly pronounced when the original image contains dense promotional text, which may confuse the extractor and lead to incorrect masks. In such cases, the inpainting model must either reconstruct missing product content or suppress conflicting background elements, potentially compromising structural fidelity. As demonstrated in the last row of Figure 3 of the main paper, under-segmentation forces the model to complete missing sections. In contrast, general image editing models do not require explicit foreground masking and can implicitly localize the subject from context, though they risk misidentification as well.

Inability to modify product attributes under inpainting constraints. The inpainting paradigm enforces strict preservation of the unmasked product region, preventing changes to intrinsic product states. For example, if the input shows a water-filled bottle but the prompt requests an empty one, SimplePoster cannot fulfill this instruction without modifying the masked area. It is a fundamental constraint of the inpainting framework. While this design ensures high subject fidelity, it limits applicability to tasks requiring semantic edits to the product itself, such as state changes, color swaps, or style transfers within the product region.

Text generation accuracy can be further improved. Despite leveraging precise text line coordinate, SimplePoster only slightly outperforms SeedEdit3.0 which relies solely on coarse spatial cues, in overall text rendering quality. As discussed in Section 5.5 of the main paper, we attribute this primarily to differences in base model capabilities. This suggests that while our layout control mechanism is effective, further improvements in text fidelity will depend on stronger pretraining in multilingual data.

B. Generalization Verification of Character Position Encoding

To validate the generalization capability of our Character Position Encoding (CPE) on general text-to-image synthesis and multilingual scenarios, we conduct a pilot experiment extending FLUX.1-dev—originally an English-only model—to generate Japanese and Korean text.

B.1. Dataset

We collect approximately 4w images containing Japanese or Korean text from social media platforms, evenly split between the two languages. Unlike product posters, these images feature relatively simple layouts, predominantly single-line text. We construct two validation sets: (i) *Single-line:* images containing one text line; (ii) *Multi-line:* images containing two or more text lines.

B.2. Implementation.

We first replace the T5 text encoder in FLUX.1-dev with Qwen2.5-VL, following the main paper. We establish a baseline by fine-tuning the DiT with LoRA (rank=256) for 30 epochs. Subsequently, we incorporate our Character Position Encoding during training, applying it with 50% probability per image. During inference, we disable CPE, reverting the model to standard text-to-image generation.

B.3. Results

As shown in Table 1, CPE yields substantial improvements on the multi-line validation set and moderate gains on the single-line set. Notably, multi-line text accuracy (Sen. Acc) improves by 458% relative to the baseline (25.7% vs. 4.6%), demonstrating that CPE is particularly critical for complex layouts where spatial ambiguity is most severe.

Method	Single-line		Multi-line	
	NED ↑	Sen. Acc ↑	NED ↑	Sen. Acc ↑
LoRA (w/o CPE)	0.6326	0.3899	0.1924	0.0461
LoRA (w/ CPE)	**0.6765**	**0.4698**	**0.4813**	**0.2571**

Table 1. Experiments on multilingual text-to-image generation.

C. Qualitative Comparison of Subject Extension Mitigation

Figure 1 compares generation results across four configurations: (1) vanilla FLUX-Fill baseline, (2) ControlNet-augmented, (3) LoRA-tuned, and (4) our full-parameter fine-tuning approach.

D. Ablation on Data Scale

We conduct an ablation study to investigate how much data is required to eliminate subject extension. Following the protocol in Section 3 of the main paper, we fix the total training iterations while varying dataset size: 3K images for 300 epochs, 30K images for 30 epochs, and 300K images for 3 epochs.

As shown in Table 2, surprisingly, even 3K training images reduce the extension rate from 41% to 3.6%, outperforming ControlNet trained on 300K images (23.6%). We further reduce training epochs for the 3K subset to 3 and 10 epochs; performance drops to 17.3% and 9.3%, respectively, indicating that sufficient iterations are necessary for convergence.

Method	Subject Extension Rate ↓
Vanilla FLUX-Fill	41.0%
ControlNet	23.6%
Full tuning (3K, 3 ep)	17.3%
Full tuning (3K, 10 ep)	9.3%
Full tuning (3K, 300 ep)	3.6%
Full tuning (30K, 30 ep)	2.0%
Full tuning (300K, 3 ep)	0.6%

Table 2. Subject extension rate under varying data scales.

E. Automatic evaluation on subject preservation

Since the extension rate relies on human evaluation, we additionally provide an automatic evaluation method here. Specifically, we crop the subject regions, extract embedding using DINO v3 [1], and compute the cosine similarity between the original product and the generated product to automatically measure subject fidelity. We report DINOv3 [1]. similarity on the segmented subject to quantify consistency Table 3. Model-based metrics are less reliable than human evaluation as they may miss fine-grained errors (e.g., subtle text artifacts, minor texture distortions). We recommend treating these as supplementary indicators only.

Method	DINOv3 similarity ↑
FLUX Kontext(pro)	0.9123
Step1x-Edit	0.9050
Gemini2.5Flash	0.9366
SeedEdit3	0.9465
PosterMaker	0.9706
Ours	**0.9811**

Table 3. Subject similarity in DINOv3.

F. Complete Prompts for Visualized Examples

All text prompts corresponding to visualized examples in the main paper are provided below. Original Chinese prompts are translated to English. For general editing models (SeedEdit 3/DreamPoster, Gemini 2.5 Flash, FLUX-Kontext, Step1x-Edit), we prepend the standardized prefix ``Background generation task.'' to all prompts.

Examples without promotional text. See Figure 3.

Examples with promotional text. See Figure 2.

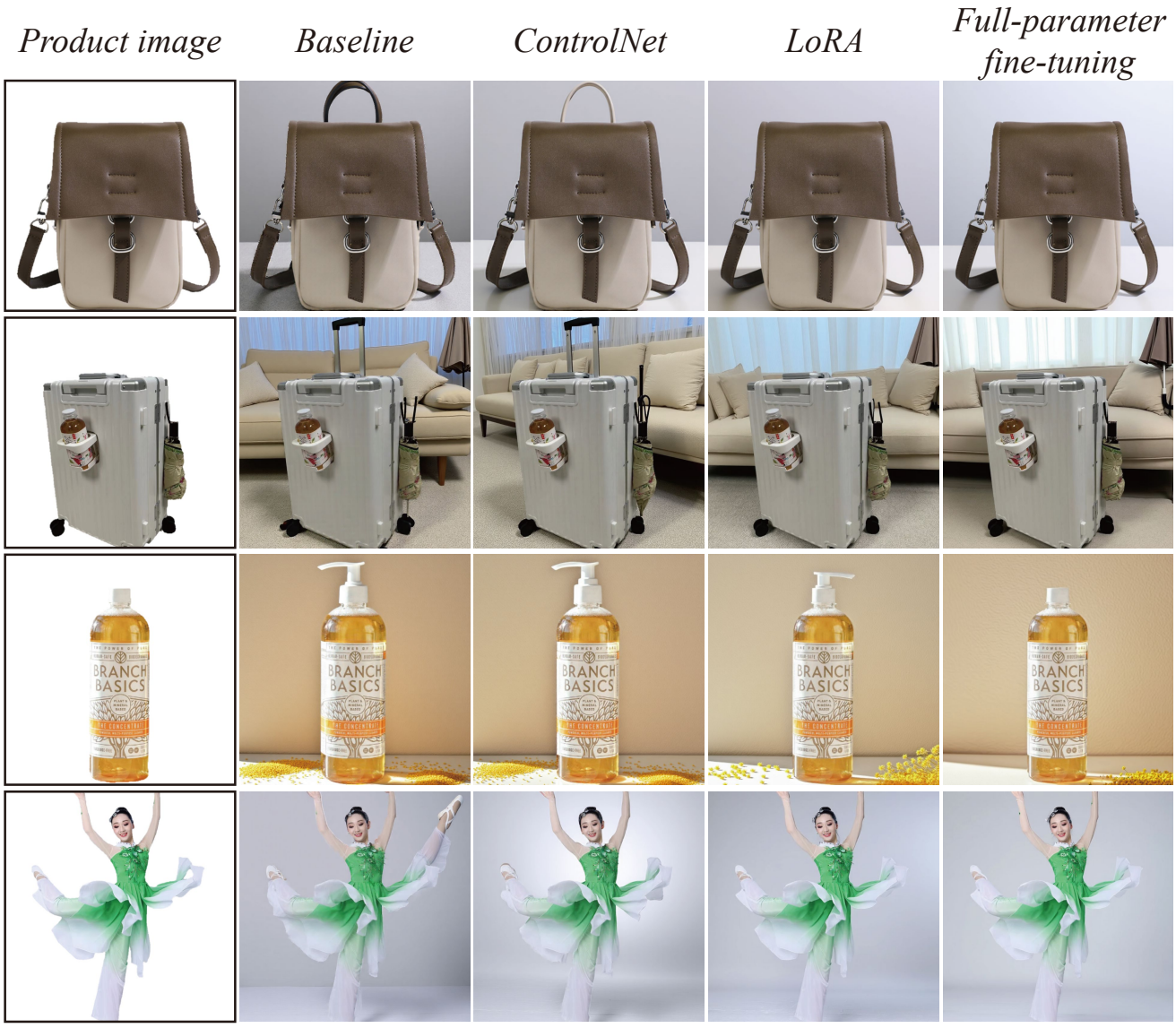

Figure 1. Samples with promotional text.

scene description

In the foreground, a black double-layer disinfection cabinet is displayed. Its doors are transparent, revealing neatly arranged tableware inside. Above the disinfection cabinet is a control panel, and below it are two silver trays. The background is a corner of a modern kitchen. The wall features a gray marble texture. To the right, there is a stainless steel sink and a golden faucet. To the left, there is a wooden shelf adorned with some decorative items and books. The floor is covered with light gray tiles.

In the foreground, a pink facial lifting and firming device is displayed. The device conforms snugly to the model's face and neck. It features purple light spots and yellow arrow indicators. The device's display screen is located on its side, showing the number 88. The background is a solid pink color, with no other complex elements or objects. The overall appearance is simple and clean, highlighting the product in the foreground. The background color is uniform, with no obvious textures or patterns.

In the center of the frame, a fresh meat mooncake is displayed, cut in half to reveal its rich interior filling. The mooncake's surface is golden and crispy. A whole fresh meat mooncake is placed underneath, with the two stacked closely together. Another whole mooncake is faintly visible nearby. The background is a dark, wood-textured tabletop, featuring delicate geometric patterns. Some blurred white objects are also present in the background, possibly packaging boxes or other decorative items. The overall color tone is dark, emphasizing the mooncakes as the main subject.

A dark wooden double piano bench is featured, its surface covered with black PU leather. The seat cushion features evenly distributed square tufting. The background is a light beige wall, unadorned by any decorations. The floor is covered with light brown wooden flooring, where the wood grain is clearly visible. To the right of the frame, a green plant stands in a simple pot. It features broad leaves, some of which have serrated edges.

text line description

Text `全国联保·苏宁售后' covering the area from the top-center to the top-right corner.
Text `超大容量' covering the area from the top-center to the middle-right. \newline
Text `升级热风烘干', `负离子净味消毒', and similar text elements covering the area from the center to the middle-right. \newline
Text `送' in the bottom-middle of the image.
Text `赠送沥水盘' covering the area from the bottom-middle to the bottom-right corner.

Text `瘦脸神器' covering the area from the top-left to the top-center. Text `紧致提拉' covering the area from the top-left to the center.Text `高品微震动提拉' in the middle-left of the image.Text `88' in the middle-right of the image.Text `5重不同效果光源随意切换' covering the area from the bottom-left to the bottom-center.Text `提拉紧致' in the bottom-left corner of the image.
Text `提拉/塑形/修护/美颜' covering the area from the bottom-center to the bottom-right corner.

Text `苏' in the top-left corner of the image. Text `州', `鲜肉月饼', and similar text elements covering the area from the top-left to the middle-left. Text `全程冷链' in the top-right corner of the image. Text `原汁原味' in the middle-left of the image. Text `优惠价', '¥', '45', and similar text elements in the bottom-left corner of the image. Text `层层酥脆新鲜味美55克*6枚' across the bottom of the image. Text `苏州特产正宗鲜肉月饼' covering the area from the bottom-center to the bottom-right.

Text `双人' covering the area from the top-left to the top-center.Text `带箱琴凳' covering the area from the middle-left to the center.Text `优良实木', `PU皮革', and similar text elements are present in the middle-left of the image. Text `聚酯漆' covering the area from the middle-left to the bottom-left. Text `现货包邮' in the bottom-left corner of the image. Text `厂家直供实木烤漆' across the bottom of the image.

Figure 2. Samples with promotional text.

In the foreground, four small portable Bluetooth speakers with gradient colors are displayed, each placed on a wooden step. Each speaker is equipped with a brown carrying strap, and their colors graduate from pink to purple, showcasing a stylish and compact design. The background features a yellow curtain, soft in texture with natural, even folds. To the right, there is a light orange geometric decorative object, resembling an arch, which adds depth to the space. The wooden steps are dark brown with clear, smooth grain, creating a strong contrast with the speakers. The ground is light pink, clean and tidy. In the bottom-right corner, there is a small white deer figurine and a black microphone model, adding a playful touch. There is no obvious text information in the image.

A creative backflow incense burner, with a lotus-shaped body, features an inverted incense stick placed on top, from which ash slowly flows downwards. At its base is a statue of a small monk meditating with closed eyes. The entire setup is displayed inside a transparent glass dome. The background is a dark wooden bookshelf, neatly arranged with books and decorative items, creating a serene and elegant atmosphere. To the left, there is a green plant, adding a touch of vitality. The incense burner is placed on a gray stone countertop with clear, natural textures. Beside it, a string of wooden prayer beads further enhances the Zen ambiance.

A bottle of essential oil is placed on a natural stone slab with irregular edges and a rough texture. To the left, a sprig of rosemary leans against a large, beige, porous rock. The background is a smooth, warm beige gradient. Soft light, high resolution.

A modern-style lamp hangs from a white ceiling, composed of multiple circular structures. The lamp body is black, and the light strip emits a soft white glow. The background is an indoor environment with a smooth, flat ceiling. On the left side of the wall, there is a dark wooden decorative panel with clear grain. On the right side, a light gray translucent curtain allows soft light to filter through. Inside the room, there is also a dark brown door with simple line decorations. The overall environment is warm and modern.

There are two celadon tea canisters in the foreground, decorated with landscape paintings on their surfaces and adorned with brown tassels. They are neatly placed on a wooden tabletop. The background features a light gray wall, on which hangs a piece of calligraphy framed in wood, echoing the color of the tabletop. In front of the wall on the left side, there is a potted green plant with slender leaves, adding a touch of vitality. On the right side, there is a woven basket decorated with blue patterned fabric. The overall color palette is harmonious and unified, creating a tranquil and elegant atmosphere. There is no obvious text visible in the image.

A purple lipstick with a gold and black alternating casing, adorned with delicate patterns, is standing upright on a tabletop. The background is dark and blurred, creating a mysterious and elegant atmosphere. In the background, there are a few pale pink flowers with delicate petals and deep purple stamens. Some petals and small berries are scattered around the flowers, giving an overall impression of nature intertwined with luxury.

Figure 3. Samples without promotional text.

References

[1] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 2