A. HP-Image-40K Dataset Statistics

To better demonstrate the diversity and comprehensiveness of the HP-Image-40K dataset, we provide detailed statistics in terms of mask area ratio and product categories. These statistics highlight the broad coverage of the dataset, making it a valuable resource for training robust and generalizable models across various real-world scenarios for generating high-quality human-product images.

Mask Area Ratio. The HP-Image-40K dataset includes a wide range of mask area ratios, as shown in Fig. 7. The mask area ratio, defined as the proportion of the mask area to the total image area, varies significantly across the dataset. This variation ensures that the dataset covers diverse object sizes and spatial distributions, from small, localized objects to large, prominent ones. Such diversity is critical for training models that can handle objects of different scales and spatial contexts, improving their performance across various application scenarios.

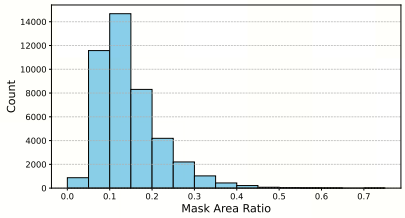

Figure 7. **Histogram of the mask area ratio in HP-Image-40K.** Our dataset exhibits a diverse range of mask area ratios, effectively covering various real-world scenarios.

Product Categories. The HP-Image-40K dataset also features a rich variety of product categories, as visualized in the word cloud in Fig. 8. These categories include bottles, containers, jars, tubes, and dispensers, among others. This diversity not only reflects the dataset's real-world applicability but also enriches the feature representation for model training. By exposing models to a broad spectrum of shapes, materials, and structural characteristics, the dataset enhances the model's ability to generalize across multiple domains and adapt to different product types.

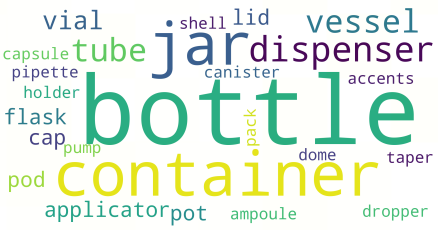

Figure 8. **Word cloud of the product categories in HP-Image-40K.** Our dataset encompasses a wide variety of product categories, providing the model with a wide range of shapes, materials, and structures for training.

B. Information of Internal Real-World Dataset

In addition to the synthetic HP-Image-40K dataset, we further construct a internal real-world dataset collected from publicly available internet images to evaluate model generalization under more realistic conditions. After preprocessing, the real-world dataset is aligned with the synthetic dataset in terms of image resolution and aspect ratio to ensure a fair comparison protocol.

Compared to the synthetic data, the real-world dataset exhibits substantially higher diversity and complexity. It contains a wide range of scenes (indoor and outdoor environments), diverse human subjects with varying poses and interactions, and products with more complex appearances, materials, textures, and branding details. The visual conditions also vary significantly in lighting, viewpoint, occlusion, and background clutter, making the learning problem more challenging than in the controlled synthetic setting.

For training, we utilize approximately 14,000 preprocessed real-world samples. For evaluation, we curate a separate test set consisting of 2,000 preprocessed real-world samples that are not used during training. This split enables a comprehensive assessment of the model's robustness and generalization capability in complex real-world scenarios.

C. More Details of Setups

Training Configuration. We set the LoRA scaling factor α to 256, which is equal to the rank. Although our model supports reference images of arbitrary resolutions, we adopt a padding-then-resizing strategy for both training and evaluation to ensure consistent spatial alignment. Specifically, when a reference image does not match the target resolution of 1024×576, we first pad it to the target aspect ratio while preserving its original content, and then resize it to the fixed resolution. This strategy avoids geometric distortion and maintains structural integrity of the product appearance.

Baseline Adaptation. For fair comparison, all baselines are evaluated under the same input resolution (1024×576) and identical masked regions. *Paint-by-Example*, *ACE++* and *Insert Anything* natively support reference-based inpainting with multi-image inputs. Therefore, we directly follow their official inference protocols and input formatting without additional modification or prompt engineering. *FLUX-Kontext* is an instruction-based image editing model that does not support explicit multi-image conditioning. To adapt it to our task, we concatenate the product reference image and the masked human image along the width dimension to form a single composite input. Following the official inpainting usage guidelines, we adopt the instruction prompt: "Change the object in the black square to the product in the left image." This enables the model to interpret the left region as the reference product and perform object replacement within the masked area accordingly.

Table 4. **Quantitative comparison on real-world data.** The results of automatic metrics demonstrate HiFi-Inpaint's overall state-of-the-art performance. The best and second-best results are marked in **bold** and underlined.

Method	Text Alignment	Visual Consistency				Generation Quality	
	CLIP-T↑ (%)	CLIP-I↑ (%)	DINO↑ (%)	SSIM↑ (%)	SSIM-HF↑ (%)	LAION-Aes↑	Q-Align-IQ↑
Paint-by-Example [53]	27.1	56.2	24.3	50.8	35.7	**4.34**	2.23
ACE++ [33]	28.2	80.1	74.2	53.5	36.6	3.90	3.47
Insert Anything [44]	28.9	83.1	77.5	55.1	37.8	3.95	**3.48**
FLUX-Kontext [6]	29.0	59.9	55.7	44.6	34.3	4.30	2.91
HiFi-Inpaint (Ours)	**29.7**	**86.8**	**79.8**	**60.5**	**44.1**	4.27	3.29

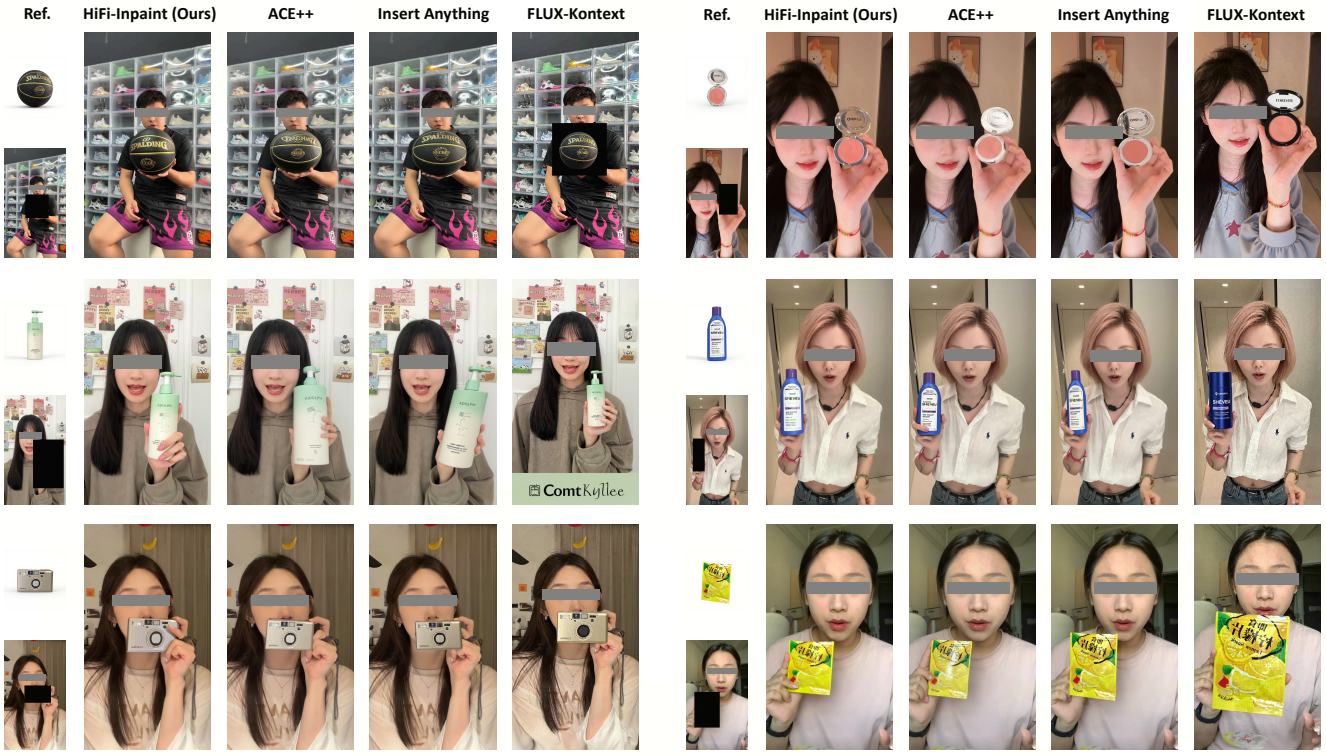

Figure 9. **Qualitative comparison on real-world data.** Compared to existing methods, our HiFi-Inpaint exhibits remarkable performance in generating high-quality human-product images, enabling high-fidelity preservation of fine-grained details. The eyes have been obscured to protect the identity of real humans. *Zoom in for better view.*

D. Evaluation on Real-World Data

In the main paper, experiments were conducted on synthetic data due to their well-controlled alignment with our task definition. To further verify the generalizability of the models, we additionally evaluate HiFi-Inpaint and other baselines on an internal real-world test set containing 2,000 diverse human–product samples, which presents significantly more variation in lighting conditions, pose configurations, and product appearance.

D.1. Quantitative Comparison

Tab. 4 reports quantitative comparisons across all metrics, showing that HiFi-Inpaint remains highly competitive under the more challenging real-world setting. For text alignment, HiFi-Inpaint achieves the best CLIP-T, indicating that the generated content generally stays faithful to textual instructions even when scenes exhibit greater complexity. In terms of visual similarity, our model attains the highest CLIP-I (86.8) and DINO (79.8), suggesting that HiFi-Inpaint can better preserve both global product identity and local appearance details when the alignment between reference and target is less constrained than in synthetic cases. Structural similarity follows a similar trend: HiFi-Inpaint obtains the top SSIM (60.5) and SSIM-HF (44.1), reflecting strong preservation of object structure and high-frequency characteristics such as text, logos, and fine patterns. For aesthetic and perceptual quality, HiFi-Inpaint delivers compet-

Figure 10. **Qualitative ablation analysis.** The results demonstrate the effectiveness of both Shared Enhancement Attention (SEA) and Detail-Aware Loss (DAL) in improving the quality of generated human-product images. Our HiFi-Inpaint, integrating these techniques, achieves the best overall performance with superior detail preservation. *Zoom in for better view.*

itive results, ranking third on both LAION-Aes (4.27) and Q-Align-IQ (3.29). Although these scores are slightly lower than the best-performing baselines, they still indicate visually appealing outputs that remain technically coherent with the reference products. By comparison, FLUX-Kontext exhibits lower CLIP-I (59.9) and DINO (55.7), suggesting that it has more difficulty grounding the reference product under real-world conditions. ACE++ and Insert Anything achieve moderate to strong performance, with Insert Anything showing relatively good structural and detail preservation, yet both are still outperformed by HiFi-Inpaint on most visual consistency metrics. Paint-by-Example lags behind recent methods in this challenging setup, especially on visual consistency and structural similarity. Overall, the real-world evaluation suggests that HiFi-Inpaint generalizes well beyond synthetic data and remains robust in preserving product fidelity under realistic variations, while maintaining competitive aesthetic and perceptual quality.

D.2. Qualitative Comparison

Qualitative comparisons of the four strongest competing methods are provided in Fig. 9. FLUX-Kontext frequently fails to perform correct inpainting, often generating an isolated product instead of integrating it into the masked region. This suggests that generic instruction-based editing offers limited capability for grounding the reference product within complex inputs. Even when successful, FLUX-Kontext tends to lose high-frequency details, leading to noticeable inconsistencies in structure and texture. ACE++ shows a stronger ability to associate the product with the masked region, preserving overall shape and partially retaining textual or patterned elements. However, fine-scale details such as small characters or intricate logos are often not accurately reconstructed. Insert Anything performs better in detail preservation but tends to introduce artifacts when the masked region becomes smaller, degrading realism and compositional quality. In contrast, HiFi-Inpaint produces clean, realistic, and naturally composited results. The model faithfully preserves product appearance, including text, patterns, and branding elements, while aligning the inpainted region seamlessly with the surrounding context. Importantly, HiFi-Inpaint remains robust even when the mask is small, maintaining structural integrity and fine-grained details without introducing noticeable artifacts.

E. Additional Results of Ablation Analysis

To further validate the effectiveness of key components in HiFi-Inpaint, we conduct a systematic ablation study, with additional qualitative results shown in Fig. 10. The examples show that removing individual components leads to noticeable degradation in the detail preservation performance. In contrast, the complete HiFi-Inpaint consistently produces superior results, faithfully preserving critical details and achieving seamless integration with the background. These results further highlight the contributions of our proposed techniques in tackling challenging inpainting tasks.

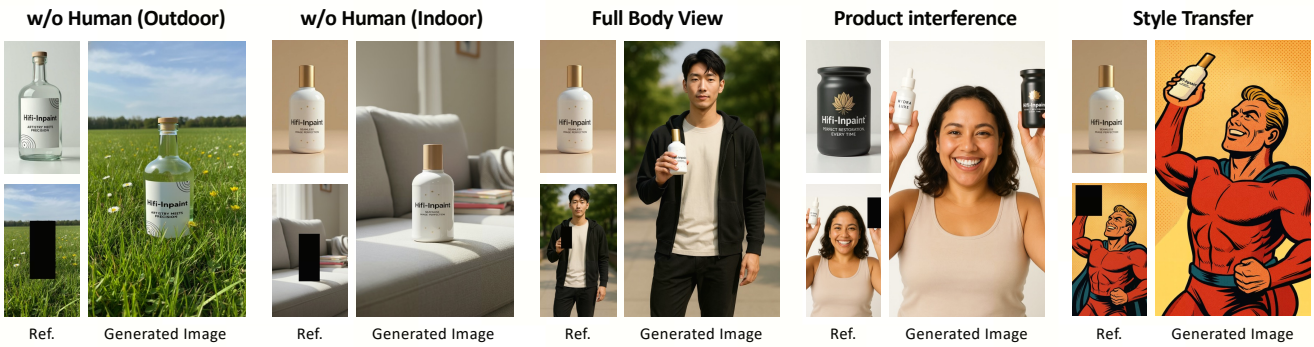

Figure 11. **Generalizability analysis of HiFi-Inpaint.** We further evaluate our HiFi-Inpaint on several hard cases, demonstrating its potential to generalize to a broader range of scenarios. *Zoom in for better view.*

F. Generalizability Analysis

In Fig. 11, HiFi-Inpaint is further evaluated on a collection of challenging real-world cases designed to examine the model's behavior beyond standard inpainting conditions. These examples cover a wide span of difficult scenarios, including images without humans in both outdoor and indoor environments, full-body human views with large pose variations, situations with product interference in the masked image, and cases requiring substantial style adaptation. As illustrated in Fig. 11, HiFi-Inpaint consistently produces coherent and context-aware completions across these heterogeneous inputs. Even when object scale, lighting conditions, or style distributions deviate significantly from the training distribution, the model is able to integrate the target product naturally into the scene while preserving its key visual attributes. Although certain extreme cases still reveal room for improvement, these results highlight the model's potential to generalize toward a broader range of practical applications and more diverse deployment conditions.