DreamFuse: Adaptive Image Fusion with Diffusion Transformer Supplementary Materials

Junjia Huang[1,2*] Pengxiang Yan[3*] Jiyang Liu[3*†] Jie Wu[3]
Zhao Wang[3] Yitong Wang[3] Liang Lin[1,2,4] Guanbin Li[1,2,4‡]
[1]Sun Yat-sen University, [2]Peng Cheng Laboratory, [3]ByteDance Intelligent Creation
[4]Guangdong Key Laboratory of Big Data Analysis and Processing
huangjj77@mail2.sysu.edu.cn, wantong1017@163.com
linliang@ieee.org, liguanbin@mail.sysu.edu.cn
{yanpengxiang.ai, liujiyang.liu, wujie.10, zhaoxu.bit}@bytedance.com

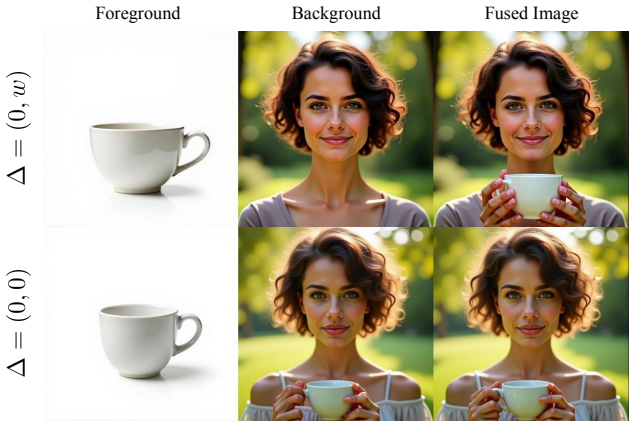


Figure 1. Comparison of generalization capabilities introduced by offset: Models with offset $\Delta = (0, 0)$ tend to generate consistent images, leading to foreground objects appearing in background scenes.

1. Details about the Data Generation

1.1. Generation of Text Prompts

To generate diverse fused data, we first create a sufficiently rich set of text prompts. For this purpose, we divide the process into two parts: foreground and background. In the foreground, the main subjects include animals, plants, humans [1], pets, logos [2], and products. For the background, we collect a certain amount of images from website [3] and utilize GPT-4o to extract realistic background prompts, ensuring coverage of various real-world scenarios. During the text prompt generation phase, we randomly sample a number of examples from the foreground and background, and let GPT-4o classify them into foreground, background, and fused image text descriptions. These descriptions are then fed into our data generation model to produce the fused data.

1.2. Training Details about the Data Generation Model

Starting with the first batch of data, we use Flux-Dev as the base model. Input images are randomly scaled to 512, 768, or 1024 resolutions, and the model is trained for 10k iterations on 8 A100 GPUs using the Prodigy optimizer. Two models are trained: one with offset $\Delta = (0, w)$ and the other with offset $\Delta = (0, 0)$. The former is designed to produce diverse data, while the latter focuses on generating data with varying scales. After training, the generated results are first filtered using GPT-4o, followed by manual selection of high-quality fusion data for the next training iteration.

1.3. Effectiveness of the Offset Δ

We experiment with two offset configurations: $\Delta = (0, w)$ and $\Delta = (0, 0)$. The results demonstrate that models trained with $\Delta = (0, w)$ exhibit better generalization, effectively handling scenarios not included in the initial small dataset. For instance, when the first training iteration is conducted using fused data from placement scenarios selected from dataset [6], the model trained with $\Delta = (0, w)$ generates differentiated results for other scenarios, such as handheld and wearable contexts, producing distinct backgrounds and fused images. As shown in Fig. 1, models trained with $\Delta = (0, 0)$ exhibit stronger consistency, often generating

*Equal Contribution.
†Project Lead.
‡Corresponding Author.
[1]https://huggingface.co/datasets/k-mktr/improved-flux-prompts-photoreal-portrait
[2]https://huggingface.co/datasets/logo-wizard/modern-logo-dataset
[3]https://unsplash.com/s/photos/free-images

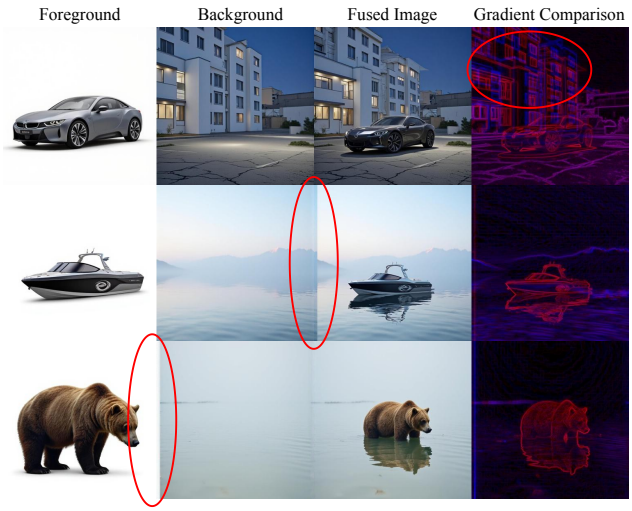


Figure 2. Misalignment often occurs when $\Delta = (0, w)$. "Gradient Comparison" illustrates the gradient comparison between the background and the fused image.

similar backgrounds and fused images.

However, when $\Delta = (0, w)$, although it demonstrates superior capabilities in generating diverse and fused data, it also tends to cause misalignment or inconsistencies in the background. As illustrated in the Fig. 2, to better visualize this misalignment, we compute the gradient maps of both the background and the fused image, and combine them into a single image for visualization in RGB format, referred to as "Gradient Comparison". Specifically, the red channel represents the gradient map of the fused image, while the blue channel corresponds to the gradient map of the background. When the background is perfectly aligned, the two gradient maps merge into purple. Conversely, noticeable red or blue regions indicate misalignment. This phenomenon highlights that the background and the fused image are not fully consistent. In contrast, when $\Delta = (0, 0)$, the alignment improves significantly, with the background predominantly appearing purple, indicating higher consistency. Meanwhile, we observed that this misalignment becomes more pronounced when generating multi-scale images. Therefore, only $\Delta = (0, 0)$ is used for generating multi-scale fused images.

1.4. Effectiveness of the Existing LoRA.

To enhance the diversity of data generation, we incorporate various styles of LoRA into the trained generative model. As shown in Fig. 3, we experiment with AntiBlur LoRA [4], Realism LoRA [5], and Asian Ethnicity LoRA [6]. Furthermore, our generative LoRA can be directly applied to other

[4]https://huggingface.co/Shakker-Labs/FLUX.1-dev-LoRA-AntiBlur
[5]https://huggingface.co/strangerzonehf/Flux-Super-Realism-LoRA
[6]https://huggingface.co/Shakker-Labs/AWPortraitCN

Figure 3. The impact of different style LoRAs on the generation of fused data.

Figure 4. The impact of different FLUX-based base models on the generation of fused data.

FLUX-based fine-tuned base models to produce diverse images. As illustrated in Fig. 4, we test multiple base models, including Flux-DEV and PixelWave [7].

1.5. Data Filtering

To ensure the high quality of the fused data, we perform further filtering based on the generation performance of the two offset types and their corresponding models. Specifically, we utilize GPT-4o to filter the data under three conditions: (1) the object in the foreground image does not match the object in the fused image; (2) remnants of the foreground object or the foreground object itself are present in the background image; and (3) the image exhibits significant quality or aesthetic issues. Fig. 5 illustrates examples of fused data filtered out by GPT-4o under these conditions.

To address the offset artifacts observed in the data generated by the model with offset $\Delta = (0, 2)$, we calculate the Dice score between the gradient maps of the background image and the fused image within the outer 100-pixel boundary. A low Dice score indicates a mismatch between the edges of the background and the fused image, signifying an offset artifact. These offset-affected samples

[7]https://huggingface.co/mikeyandfriends/PixelWave_FLUX.1-dev_03

Figure 5. Three types of cases filtered out by GPT-4o.

	Indoor		Outdoor		Summary
	Simple	Complex	Simple	Complex	
Object	8,574	3,475	10,551	8,827	31,427
Animal/Pet	2,405	1,250	2,373	2,181	8,209
Human	1,930	1,106	2,377	2,555	7,968
Logo	1,539	31	400	11	1,981
Style Transfer	183	46	2,064	1,084	3,377
handheld pets	3,943	2,162	4,835	4,469	15,409
handheld objects	1,902	287	4,304	1228	7,721
wearable	2,490	135	4,308	1,098	8,031
Summary	22,966	8,492	31,212	21,453	84,123

Table 1. The number of fused images across various scenarios.

are filtered out.

1.6. Data Analysis

Through the above generation strategy and quality filtering, we ultimately obtained an 84k high-quality fusion dataset. In Tab. 1, we provide a detailed breakdown of the number of fused data for each scenario, along with a detailed classification based on indoor and outdoor settings, as well as simple and complex scenes.

Additionally, we analyzed the resolution distribution of the images in our dataset. As shown in Fig. 7, our data spans a range from 600 to 1400 pixels, without being restricted to a fixed resolution.

1.7. Multi-Foreground Generation

After training the current data generation model, it demonstrates a certain generalization capability to generate fused scenes with multiple foregrounds when provided with two foregrounds prompts, as shown in Fig. 6. This verifies that our data generation model can generalize to multi-foreground data production, which is particularly important for scenarios where occlusion or nesting relationships exist between foreground objects. In the future, we will further explore the generation of multi-foreground fusion data.

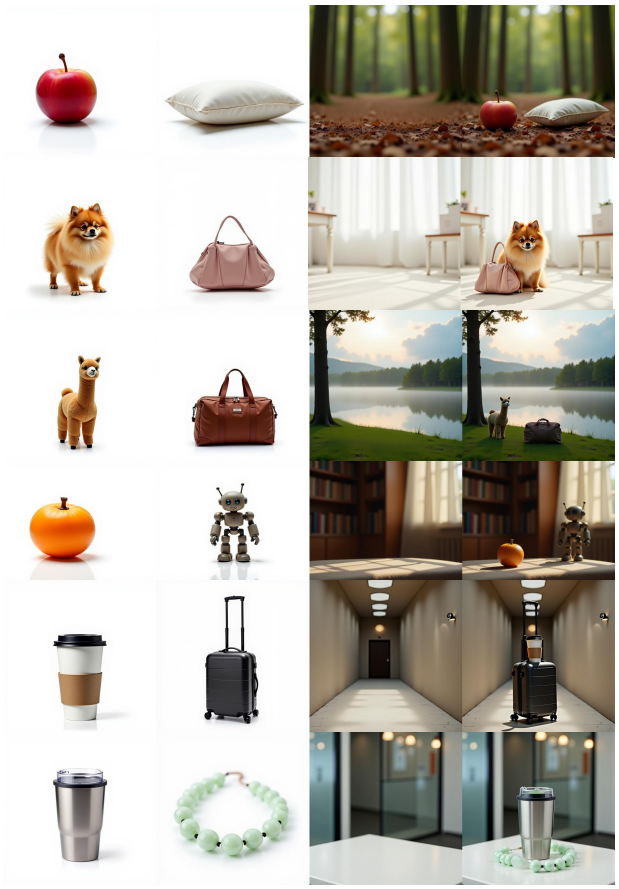


Figure 6. Visualization of Multi-Foreground fusion data.

1.8. Data Visualization

Our dataset encompasses a diverse range of scenes and foreground objects. As shown in Fig. 10, the foregrounds in our dataset include products, people, animals, plants, vehicles, and natural objects. "Gradient Comparison" refers to the gradient comparison between the background and the fused image, while "Copy-Pasted Image" indicates directly copying the foreground and pasting it onto a specified position in the background. Fig. 11 further illustrates image examples from various fusion scenarios in our dataset, such as style transfer, logo printing, handheld, and wearable applications, while also showcasing data at different scales.

2. Details about the DreamFuse

2.1. Details about the Vision Reward (VR) Score in Evaluation

To better evaluate the fusion results, we use the Vision Reward [7] (VR) Score, which measures quality by inputting the image and multiple questions into a vision-language model [4] (VLM) to obtain comprehensive, multi-dimensional scores. We selected eight questions to evalu-

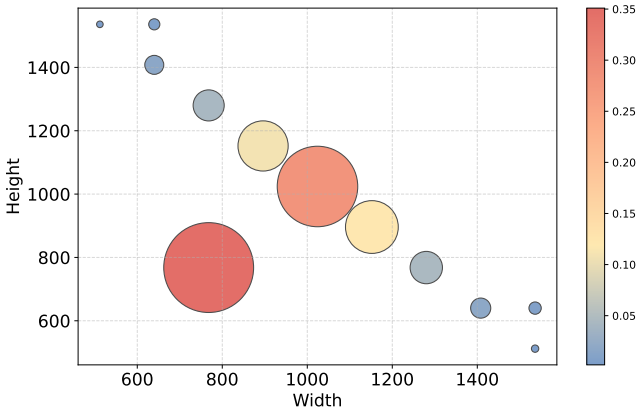


Figure 7. Distribution of image resolutions.

Method	Vision Reward Score
ControlCom [8]	0.72
Anydoor [1]	1.4
MADD [3]	0.21
MimicBrush [2]	1.78
Ours	**3.45**

Table 2. Quantitative evaluation results on FOSCom dataset.

ate the images from multiple dimensions. Each satisfactory answer is assigned a score of +1, while an unsatisfactory answer deducts a score of -1. The eight questions are formulated as follows:

- Are the objects well-coordinated?
- Is the image not empty?
- Is the image clear?
- Can the image evoke a positive emotional response?
- Are the image details exquisite?
- Does the image avoid being hard to recognize?
- Are the image details realistic?
- Is the image harmless?

2.2. The Pseudo-code for LDPO.

As shown in Algorithm 1, we present the pseudo-code of LDPO. LDPO optimizes the model at each denoising step, directly optimize DreamFuse based on human preferences. By using copy-pasted data as negative samples, we enhance the background consistency and foreground harmony in the model's fusion results.

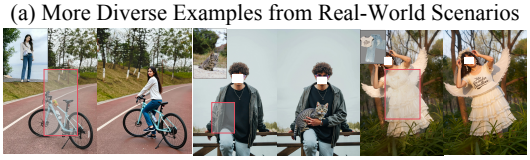
2.3. Performance of DreamFuse in Real-World Scenarios

The TF-ICON dataset already includes some real-world images. To further validate the effectiveness of DreamFuse in real-world scenarios, we conducted additional experiments on the FOSCom [8] dataset, a fusion dataset composed entirely of real images. The dataset contains only foreground and background components, including 640 background images collected from the Internet. Each background image is paired with a manually annotated bounding box and a foreground image from the MSCOCO [5] training set. Since the dataset lacks text descriptions of the fused images, we primarily compared the VR scores of the fusion results. As shown in Tab. 2, our method outperforms the second-best method by a margin of 1.76 in VR score. Fig. 9 presents the qualitative results of DreamFuse on the FOSCom dataset, demonstrating that DreamFuse achieves superior performance in real-world scenarios. DreamFuse integrates the foreground harmoniously into the background, generating realistic effects such as reflections and shadows.

Algorithm 1 Localized Direct Preference Optimization Loss (LDPO)

Dataset: Fusion dataset $\mathcal{D}' = \{(c_i, x_f, x_b, x_i^w, x_i^l)\}$
Input:
  ϵ_θ: DiT with LoRA parameters from the first training stage.
  ϵ_{ref}: Frozen DiT with LoRA parameters from the first training stage.
  p: Text prompt dropout probability.
  α: Dilation factor.
  β: Regularization parameter.
Define $M(f)$:
  $M(f) = 1$ if $f \in \alpha \cdot \text{Bbox}(x_f)$, else $M(f) = 0$ ▷ Localized foreground region.
for fusion data $(c_i, x_f, x_b, x_i^w, x_i^l) \in \mathcal{D}'$ **do**
  Sample noise and interpolate latents:
  $t \leftarrow \text{Random}(0, 1)$, $x_n \leftarrow \text{RandNoise}$
  $x_t^w \leftarrow (1-t)x_i^w + tx_n,\ x_t^l \leftarrow (1-t)x_i^l + tx_n$
  $c_i^p \leftarrow \text{Dropout}(c_i, p)$
  Model predictions:
  $v_\theta^w \leftarrow \epsilon_\theta(c_i^p, x_f, x_b, x_t^w),\ v_\theta^l \leftarrow \epsilon_\theta(c_i^p, x_f, x_b, x_t^l)$
  $v_{ref}^w \leftarrow \epsilon_{ref}(c_i^p, x_f, x_b, x_t^w),\ v_{ref}^l \leftarrow \epsilon_{ref}(c_i^p, x_f, x_b, x_t^l)$
  Calculate velocities and errors:
  $v^w \leftarrow x_n - x_i^w,\ v^l \leftarrow x_n - x_i^l$
  $err_\theta^w \leftarrow ||v_\theta^w - v^w||^2,\ err_\theta^l \leftarrow ||v_\theta^l - v^l||^2$
  $err_{ref}^w \leftarrow ||v_{ref}^w - v^w||^2,\ err_{ref}^l \leftarrow ||v_{ref}^l - v^l||^2$
  Compute differences:
  $w_{\text{diff}} \leftarrow M \cdot (err_\theta^w - err_{ref}^w) + (1-M) \cdot (err_\theta^l - err_{ref}^l)$
  $l_{\text{diff}} \leftarrow M \cdot (err_\theta^l - err_{ref}^l) + (1-M) \cdot (err_\theta^w - err_{ref}^w)$
  Compute loss:
  $L_{\text{LDPO}} \leftarrow -\log(\text{sigmoid}(-0.5 \cdot \beta \cdot (w_{\text{diff}} - l_{\text{diff}})))$
  Update model: $\epsilon_\theta' \leftarrow \epsilon_\theta$
end for

(a) More Diverse Examples from Real-World Scenarios



(b) Some Failure Cases from Real-World Scenarios

Figure 8. More diverse examples and some failure cases.

2.4. Limitations

To further validate the generalizability of our method, we test it on a wider range of real-world images. In Fig. 8 (a), our model performs well under complex scenarios like challenging lighting, hand interactions, and partial occlusion. We also observed several failure cases (Fig. 8 (b)), including difficulty preserving face identity, extra limbs and failures in try-on scenarios. These issues are mainly due to limitations in the data generation process: (1) Due to the base model's bias toward Western facial features, identity preservation for human foregrounds is often imperfect; (2) Over-preserved backgrounds may introduce duplicated body parts; (3) The training data has limited try-on samples, reducing generalization. The main bottleneck is the base generation model. For example, FLUX tends to generate overly blurred backgrounds, which differ from real distributions. Although we applied LoRA-based deblurring, some issues remain. Notably, our pipeline is flexible and can adapt to stronger models in the future to further improve data quality.

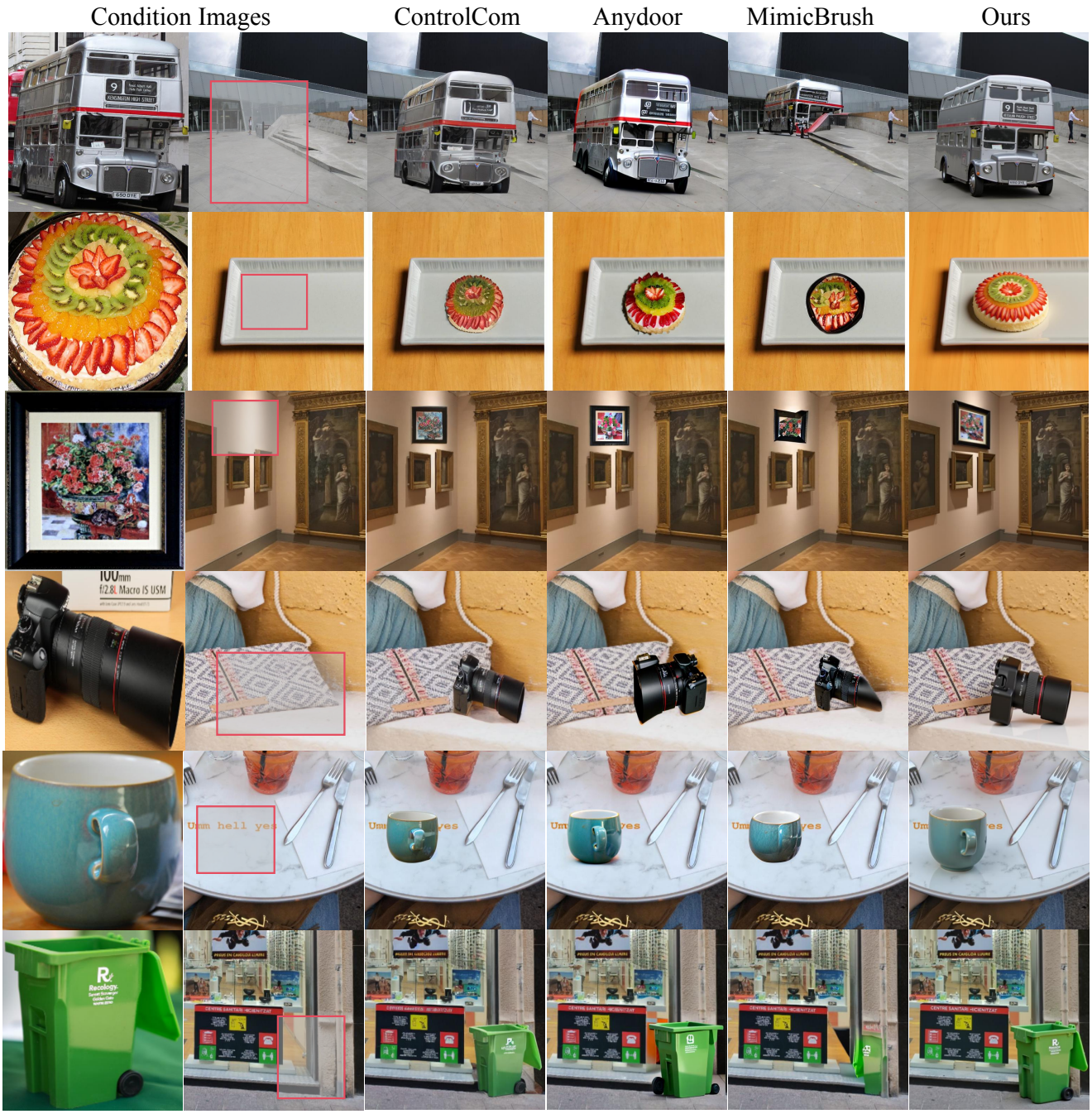


Figure 9. Qualitative comparisons on FOSCom dataset.

Figure 10. Visualization about different foreground in DreamFuse dataset. "Gradient Comparison" refers to the gradient comparison between the background and the fused image, while "Copy-Pasted Image" indicates directly copying the foreground and pasting it onto a specified position in the background.

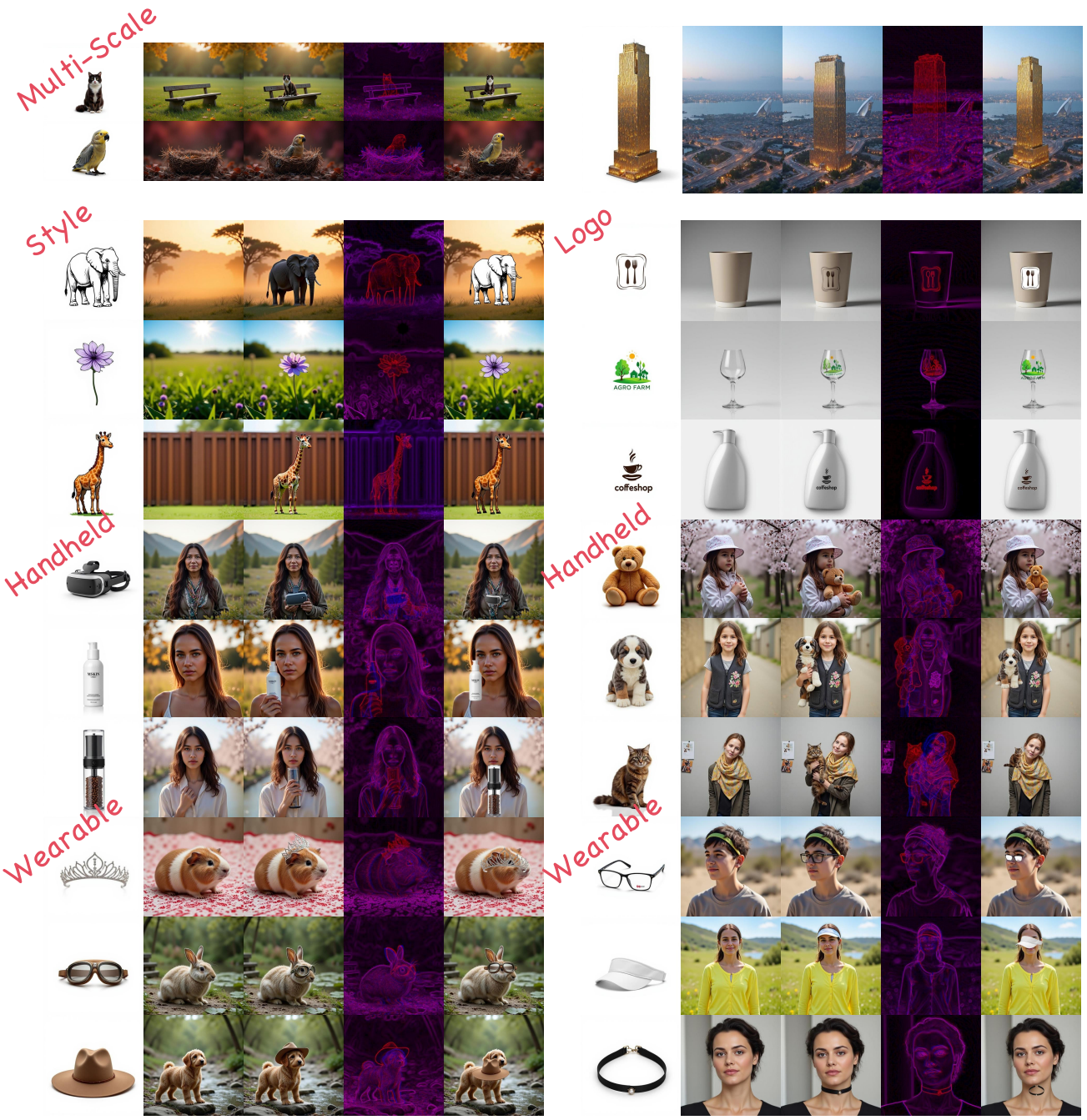


Figure 11. Visualization about different fusion scenarios in DreamFuse dataset. “Gradient Comparison” refers to the gradient comparison between the background and the fused image, while “Copy-Pasted Image” indicates directly copying the foreground and pasting it onto a specified position in the background.

References

[1] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 4

[2] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *NeurIPS*, 37:84010–84032, 2025. 4

[3] Jixuan He, Wanhua Li, Ye Liu, Junsik Kim, Donglai Wei, and Hanspeter Pfister. Affordance-aware object insertion via mask-aware dual diffusion. *arXiv preprint arXiv:2412.14462*, 2024. 4

[4] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 3

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 4

[6] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024. 1

[7] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 3

[8] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 4