DreamFuse: Toward Realistic and Seamless Image Fusion Across Diverse Scenarios

Junjia Huang, Pengxiang Yan, Jiyang Liu, Jie Wu, Zhao Wang, Yitong Wang, Xinglong Wu, Liang Lin, *Fellow, IEEE*, and Guanbin Li, *Member, IEEE*

***Abstract*—Image fusion seeks to seamlessly integrate foreground objects with background scenes, producing realistic and harmonious fused images. While existing methods often insert objects directly, adaptive and interactive fusion—requiring contextual adaptation and foreground-background interplay—remains a challenging yet critical task. To address this, we first propose a pipeline for generating high-quality fusion data. By combining iterative in-context learning with existing tools, we curate a diverse cross-scene dataset supporting three core tasks: object integration, replacement, and attribute-referenced editing. Leveraging this, we introduce DreamFuse, a unified diffusion-based approach that jointly optimizes these capabilities. DreamFuse exploits the Diffusion Transformer (DiT) architecture, using its attention mechanism to extract and align foreground-background features for coherent fusion. For flexible control, we incorporate a Positional Affine mechanism, enabling precise spatial and scale adjustments while supporting diverse text-driven fusion. Furthermore, we employ Localized Direct Preference Optimization (L-DPO), refining the model via human feedback to enhance harmony and consistency. Extensive experimental results demonstrate DreamFuse's superiority over state-of-the-art approaches across multiple metrics.**

***Index Terms*—Image fusion, reference-image based editing, diffusion model, multimodal image customization.**

I. Introduction

FOREGROUND-BACKGROUND image fusion is a fundamental and practical task in image editing. Recently, with the rapid development of generation methods based on diffusion models, numerous innovative and imaginative approaches have emerged in this field. Beyond generating images purely driven by text prompts, increasing attention has been directed toward generating customized images guided by specific foreground objects [1], [2], [3], [4] or performing inpainting on designated background regions [5] for image editing and fusion. Going further, several methods aim to achieve harmonious fusion of foreground and background by adjusting details such as lighting [6], shadows [7], and brightness-contrast consistency [8], making the fused images appear more natural. Other approaches [9], [10], [11] strive to enhance fusion by modifying the orientation, pose, or style of the foreground object while preserving its identity attributes, enabling better adaptation to the background. However, most of these methods typically focus on directly placing the foreground object into the background scene. In contrast, real-world scenarios often involve more diverse and interactive conditions, such as partial occlusion, alternating visibility, or integration cases where objects are held, worn, or otherwise physically interacted with. In some cases, the goal may also be to incorporate only specific attributes of the foreground object into the background.

A major challenge in handling such complex fusion scenarios is the lack of suitable datasets. Existing methods typically rely on object segmentation from images or videos [12], [13] followed by inpainting [14], [15] to reconstruct backgrounds. However, this multi-step process frequently suffers from quality degradation due to imprecise segmentation or suboptimal inpainting, which can introduce artifacts like shadows or residual elements in the background. Moreover, handling partially occluded foregrounds is particularly challenging [16], and segmentation-based data often fails to adjust object pose or perspective to align with the background during fusion. To address these challenges, we propose a systematic pipeline for generating high-quality image fusion data, comprising foregrounds, backgrounds, and the fused images, avoiding issues such as incomplete foregrounds and background artifacts. We leverage a Multimodal Large Language Model (MLLM) to dynamically sample and combine foregrounds, backgrounds, and fusion types from a meta-database, ensuring diverse yet coherent fusion scenarios. Corresponding text prompts are then generated, and images of the foreground, background, and fused scene are produced through Self-Inspired In-Context Generation and Tool-Invoking Generation. To ensure quality, we further employ the MLLM for data inspection and filtering, complemented by rule-based criteria to remove substandard samples. This process ultimately constructs a dataset of 160 k high-quality, multi-scene, multi-scale fused image samples.

Received 1 September 2025; revised 11 February 2026; accepted 28 March 2026. Date of publication 7 April 2026; date of current version 6 July 2026. This work was supported in part by the National Key R&D Program of China under Grant 2024YFB3908503 and Grant 2024YFB3908500, in part by the National Natural Science Foundation of China under Grant 62322608, and in part by the Major Key Project of PCL under Grant PCL2025A17. Recommended for acceptance by Y.-Y. Chuang. *(Corresponding author: Guanbin Li.)*

Junjia Huang and Liang Lin are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, also with Peng Cheng Laboratory, Shenzhen 518066, China, and also with the Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China.

Pengxiang Yan, Jiyang Liu, Jie Wu, Zhao Wang, Yitong Wang, and Xinglong Wu are with ByteDance Intelligent Creation, Beijing 100098, China.

Guanbin Li is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, also with Peng Cheng Laboratory, Shenzhen 518066, China, also with the Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China, and also with Shenzhen Loop Area Institute, Shenzhen 518000, China (e-mail: liguanbin@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2026.3680535

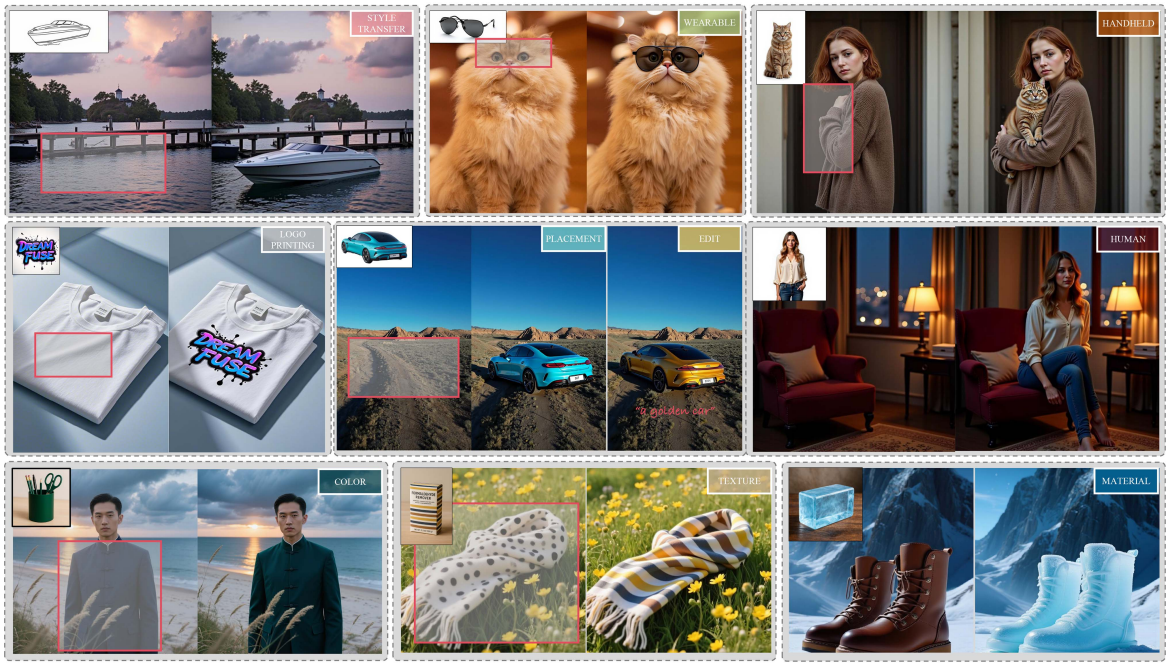


Fig. 1. DreamFuse demonstrates robust adaptability across a wide range of scenarios, including style transfer, integration of wearable items, logo embedding, object placement, handheld interactions, and fine-grained attribute editing involving color, texture, and material. Notably, when given a text prompt, our method effectively responds by further editing the attributes of the foreground object (e.g., a golden car).

Another critical challenge in image fusion is ensuring background consistency and foreground harmony. Some approaches [9], [11] rely on masks or bounding boxes for foreground placement, blending the background outside the mask to maintain consistency. However, these approaches often fail to realistically render effects like shadows or reflections beyond the mask, limiting the realism of the fused results. Other methods [17], [18] reconstruct fused images through inversion, improving foreground harmony but often compromising background consistency. To address this trade-off, we propose DreamFuse, an adaptive image fusion framework based on DiT. By incorporating shared attention, we condition the fused image generation on both the foreground and background while employing positional affine to introduce the foreground's position and scale without restricting its editable regions. Additionally, we employ Localized Direct Preference Optimization (LDPO) to further optimize the foreground and background regions of the fused image, ensuring better alignment with human preferences. Experimental results demonstrate that DreamFuse performs exceptionally well across various scenarios. As shown in Fig. 1, DreamFuse produces highly realistic fusion effects for real-world images. Furthermore, during training, a certain proportion of fused image descriptions is incorporated as prompts. When given a text prompt, DreamFuse effectively responds to the input and enables attribute modifications in the fused scenes, such as turning a car into gold.

In summary, our key contributions are threefold:

- We construct a diverse cross-scene fusion dataset containing 160 k diverse fusion scenarios, including object integration, object replacement, and attribute-referenced editing.
- We propose DreamFuse, a fusion framework based on DiT, which leverages positional affine and LDPO strategies to integrate the foreground into the background more naturally and adaptively.
- Our method outperforms the state-of-the-art methods on various benchmarks and remains effective in real-world and out-of-distribution scenarios.

Difference from our conference version: This manuscript presents significant extensions over the conference version, enabling broader fusion scenarios and providing more comprehensive analyses. (1) We expand the data generation pipeline to a more flexible and automated framework capable of producing a greater variety of fusion data. The framework extends beyond object insertion to support multiple tasks, including object insertion, object replacement, and attribute-referenced editing (see Section III). A more detailed analysis and summary of the data distribution are provided in Section V-A. (2) We further enhance the Positional Affine mechanism to support scenarios where explicit positional input is unnecessary, enabling fusion guided solely by textual descriptions. In addition, we distinguish specific fusion tasks by leveraging both textual instructions and captions. (3) We conduct more extensive and detailed experiments and analyses in Section V-D, discuss additional related studies in Section II, and provide comparisons with a wider range of concurrent methods in Section V-C.

II. Related Work

Customized Image Generation: Customized image generation aims to create user-specific images based on text prompts or reference images. Some methods [1], [19], [20], [21], [22]

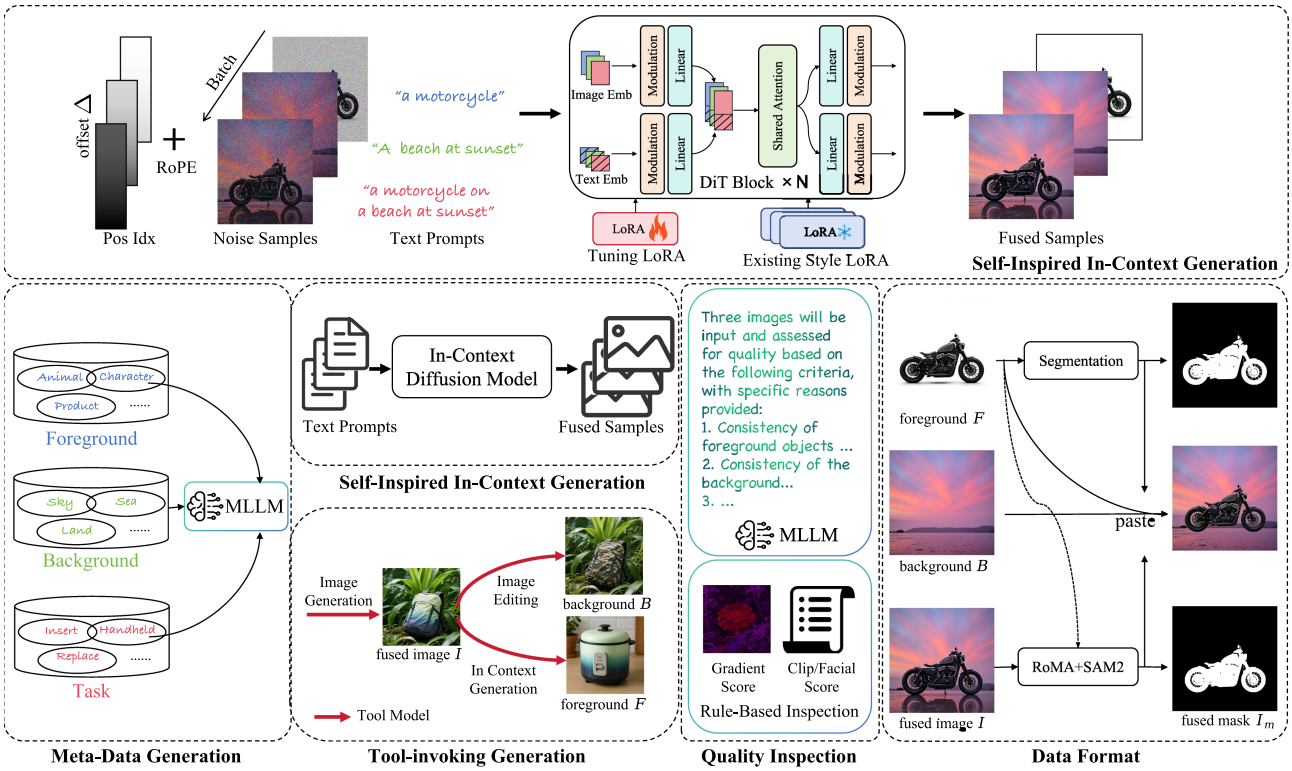


Fig. 2. The framework of the data generation model and position matching process. The left side of the figure illustrates the design structure of our data generation model, while the right side shows the position matching process and data format. We enhance the diversity of fused data generation through flexible and rich prompts combined with various style LoRAs.

incorporate reference concepts into specific text prompts, TokenVerse [23] disentangles complex visual elements and attributes, leveraging the modulation space to generate composite images of intricate concepts. Approaches [24], [25], [26], [27] employ additional encoders to encode reference images as visual prompts, introducing customized representations for generation. Other methods [3], [28] achieve subject-driven generation by directly concatenating reference and target images during generation. Recently, Mod-Adapter [4] and XVerse [29] transform reference images into offsets for token-specific modulation, offering training-free capabilities for complex customized image generation. In this paper, we further focus on customized image editing, enabling the model to modify the background image by referencing objects or even specific attributes from the foreground image.

Image Fusion: The goal of image fusion is to seamlessly integrate an object from a foreground image into a background image. Compared to directly cutting and pasting, some approaches [6], [30], [31], [32] adjust the lighting, shadows, and colors of the pasted foreground region to achieve a more harmonious fusion. Other methods [9], [10], [11], [17], [18], [33], [34], [35], [36] focus on altering the perspective, pose, or style of the foreground object to make it fit more naturally into the background image. ACE++ [37] and InsertAnything [38] follow the inpainting paradigm of FLUX.1-Fill-dev and proposes an inpainting-based generation model capable of referencing objects from the foreground image. However, these methods are largely limited to object placement, and inpainting approaches that restrict editing to the input mask often produce fused images lacking coherence and harmony. Omnigen2 [39] provides a unified solution for multiple generation tasks. Nevertheless, for image fusion, it predominantly adopts in-context generation, which limits its ability to maintain full background consistency. In this paper, we propose a more versatile fusion approach that supports a wide range of scenarios, including interactive object insertion, object replacement, and attribute-referenced editing, thereby enabling more diverse and coherent integrations.

Human Feedback Learning: Many methods now leverage human feedback learning to make generated images more aligned with users' preferences. Some approaches [40], [41], [42], [43], [44] train a reward model to understand human preferences and improve the generation quality through reward feedback learning. Others [45], [46], [47] utilize human comparison data to directly optimize a policy that best satisfies human preferences. For the image fusion task, we propose localized direct preference optimization, which focuses on region-specific optimization to enhance both the background consistency and the harmony in fused images.

III. Data Generation Pipeline

As shown in the data format in Fig. 2, the image fusion task typically involves the following types of images: a foreground image $F \in \mathbb{R}^{H\times W\times 3}$ with mask $F_m \in \mathbb{R}^{H\times W}$, a background

image $B \in \mathbb{R}^{H\times W\times 3}$, a fused image $I \in \mathbb{R}^{H\times W\times 3}$, a fused mask $I_m \in \mathbb{R}^{H\times W}$ associated with the fused object. The masks F_m and I_m are primarily used to indicate the position and size of the foreground object. In practical applications, only the centroid and bounding box of the mask are required, or in some cases, merely a textual prompt.

A. Meta-Data Generation

To generate diverse fused data, we first construct a sufficiently rich set of text prompts by collecting a pool of metadata as foreground and background sources. For different tasks such as insertion, handheld integration, replacement, and attribute reference, we utilize Doubao [48] to select suitable foreground–background combinations and generates task-specific prompts for the foreground, background, and fused images. These prompts are then processed through two pipelines, Self-Inspired In-Context Generation and Tool-Invoking Generation, to produce the corresponding images. Finally, the results undergo quality inspection and are formatted into the finalized data format.

B. Self-Inspired In-Context Generation

Data Startup: Unlike methods [11], [49] that start with a fused image to segment the foreground and generate the background using inpainting, we aim to create higher-quality fusion data with richer scenes and more diverse foreground fusion. To this end, we design an iterative, human-in-the-loop data generation process. We first extract a pair of high-quality foreground F and fused image I from a subject-driven dataset [3]. We then manually refine the inpainting regions to remove the foreground object and its effects, such as reflections and shadows, creating a high-quality background image B. A total of 86 initial samples are curated, and their corresponding descriptions C are generated with GPT-4o. These data are then fed into the data generation model depicted in Fig. 2 for training. We performed a total of three rounds of human inspection and refinement. This iterative process allowed us to progressively scale our dataset from an initial seed of 86 samples to 1,000, and finally to 4,000 high-quality pairs, which were used to train the self-inspired model.

Data Generation Model Design: We adopt Flux-Dev [50] as the base model and input our curated fusion samples $G=(F,B,I)$ with prompts $C_G=(C_F,C_B,C_I)$ in batches. The images and prompts are encoded into image embeddings $E_i \in \mathbb{R}^{h\times w\times d}$ and text embeddings E_c, supplemented by learnable tag embeddings to differentiate the foreground and background. In Flux's RoPE [51] mechanism, which uses a 2D position index $P_{idx}=(i,j), \forall i\in[0,h), j\in[0,w)$ to represent image positions, we optionally introduce an offset Δ to P_{idx} for F, B, and I as follows:

$$P_{idx}=\begin{cases}(i,j), & \text{if } F,\\ (i,j)+\Delta, & \text{if } B,\\ (i,j)+2\Delta, & \text{if } I.\end{cases} \tag{1}$$

During training, adding an offset improves the model's ability to generate diverse fused scenes but performs poorly across resolutions, while omitting the offset produces overly consistent results yet adapts well to multi-scale data, as shown in Fig. 3. Based on this, we use two models—with and without offset—to generate diverse, multi-scale samples.

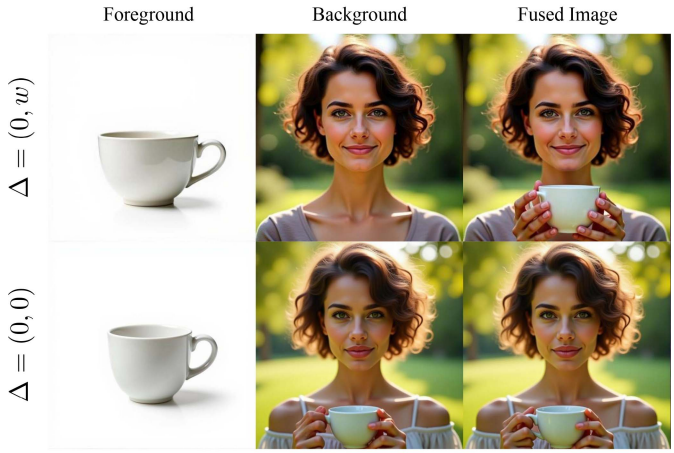


Fig. 3. Comparison of generalization capabilities introduced by offset: Models with offset $\Delta=(0,0)$ tend to generate consistent images, leading to foreground objects appearing in background scenes.

To establish connections between F,B and I, we modify the original independent attention mechanism in the DiT to a shared attention (SA) mechanism [52]. As shown in Fig. 2, after the text embeddings and image embeddings of each sample are processed through modulation and linear layers, they are concatenated to form the attention query $Q_g=[Q_g^c;Q_g^i], g\in G$, where c and i represent the text and image components. We then concatenate the image components of the key and value across all samples: $K_g=[K_g^c;K_F^i;K_B^i;K_I^i]$, $V_g=[V_g^c;V_F^i;V_B^i;V_I^i]$. The shared attention is then computed as:

$$SA=softmax\left(\frac{Q_gK_g^T}{\sqrt{d}}\right)V_g, \tag{2}$$

where d denotes the dimension. After applying shared attention, the model gains an initial ability to generate similar images. We then fine-tune the model with LoRA [53] to enable the generation of high-quality image fusion samples.

Scene Generalizability and Style Variability: Our initial data only consists of object placement scenes. After fine-tuning the data generation model, it demonstrates the ability to generalize to more diverse scene prompts. In subsequent iterations, we leverage GPT-4o with open-source prompts to generate fusion prompts C_G, expanding foreground objects to include animals, pets, products, portraits, and logos, while categorizing background scenes into indoor and outdoor settings. Fusion scenarios are further diversified to include placement, handheld, logo printing, wearable, and style transfer. Furthermore, we observe that the data generation model responds effectively to existing style LoRAs. Therefore, we integrate fine-tuned style LoRAs, such as depth of field, realism, and ethnicity, to further enhance fusion data in both scene variety and artistic style, mitigating the stylistic bias of the Flux base model. This expansion process is iterative: the model is fine-tuned on data generated in the previous step, additional data is generated, and high-quality

fusion samples are manually curated to serve as input for the next round of fine-tuning. After several iterations, the data generation model is able to reliably produce fused data across diverse scenarios.

C. Tool-Invoking Generation

To broaden the coverage of tasks, we further extend the results of Self-Inspired In-Context Generation using advanced existing tools, as shown in Fig. 2. For replacement tasks, we employ editing models such as SeedEdit [54] and Kontext [55] to substitute objects in the fused images with other categories, which are then used as background images. For attribute-referenced editing tasks, we first generate images with specified attributes using image generation models, and then apply editing models to modify the local attributes of objects as backgrounds. In parallel, in-context models such as GPT-4o and Kontext [55] are used to generate foregrounds with the same attributes, thereby constructing paired samples of foregrounds, backgrounds, and fused images. With the assistance of both Self-Inspired In-Context Generation and Tool-Invoking Generation pipelines, we expand the original insertion tasks into more diverse replacement and attribute-referenced editing tasks.

D. Quality Inspection

To improve the quality of fused data and ensure sample reliability, we conducted a comprehensive quality inspection that combines a MLLM [48] with rule-based evaluation. Following the protocol in [56], the MLLM is used to evaluate the consistency between the foreground and fused images, the consistency between the background and fused images, as well as to detect issues such as limb abnormalities and background clutter, while also providing interpretable feedback. Samples that failed these checks are removed. We then apply additional filtering based on CLIP score [57], AES score [58] and Facial score [59] to eliminate images with poor text–image alignment, low aesthetic quality or dissimilar faces. Furthermore, to mitigate potential misalignment issues in Self-Inspired In-Context Generation, we compute the gradients of the fused and background images and assess their edge similarity using the DICE score, discarding samples with noticeable edge shifts. Through this multi-stage inspection, we curate 160 k high-quality fused samples. A detailed analysis of the dataset is provided in Section V-A.

E. Data Format

To determine the position of the foreground object in the background, we use RoMA [60] to perform feature matching between the foreground and the fused image, converting the results into bounding boxes. Then we utilize SAM2 [61] to segment the foreground object from the fused image based on these bounding boxes, yielding I_m. For the foreground object, we use an internal segmentation model to obtain F_m. The position and size of the foreground object in the background are then calculated using the centroids and bounding boxes of I_m and F_m. To mitigate potential discrepancies between the generated images and the original prompts, we utilize the GPT-4o to regenerate captions for the foreground, background, and fused images, and to construct a corresponding fusion instruction.

IV. Methodology

A. Adaptive Image Fusion Framework

In image fusion tasks, three key aspects need to be considered: (1) how to model the relationship between the background, foreground, and fused image; (2) how to incorporate the position and size information of the foreground object into the background; (3) how to ensure background consistency and foreground harmony in the fused image.

Condition-aware Modeling: Inspired by the work [3], we model the background and foreground as conditions, with the fused image treated as the denoised target. Given a fixed dataset $\mathcal{D} = \{(s_i, c_i, x_f, x_b, x_i)\}$, each sample consists of a textual description of the fused image c_i, a fusion instruction s_i, along with images representing the foreground x_f, background x_b and fused image x_i. We adopt the Flux-based DiT architecture, using the foreground x_f and background x_b as conditions with a fix timestep 0. Additionally, most text prompt c_i are randomly dropped out with a probability p during training, replaced with empty strings, while a portion of the prompts is retained to preserve the network's text-responsive capability. The DiT network is tasked with denoising the noisy fused image x_i^t at timestep t defined as:

$$x_i^t = (1-t)x_i + tx_n, \tag{3}$$

where $x_n \sim q(x_n)$ denotes a noise sample and $t \in [0, 1]$. The DiT model is trained to regress the velocity field $\epsilon_\theta(x_i^t, x_f, x_b, t)$ by minimizing the Flow Matching [62] objective $\mathcal{L}_{noise}(\theta)$:

$$\mathbb{E}_{t,(c_i,x_f,x_b,x_i)\sim\mathcal{D},x_n\sim q(x_n)}[||\epsilon - \epsilon_\theta(c_i, x_i^t, x_f, x_b, t)||], \tag{4}$$

where the target velocity field is $\epsilon = x_n - x_i$. To clearly differentiate between insertion and replacement tasks, we introduce fusion instructions, as shown in Fig. 4. In particular, the indicators <instruction> and <caption> are employed to explicitly denote the task-specific instructions and the fused image caption in the text prompt as $Concat(s_i, c_i)$. Within the DiT attention mechanism, all components are concatenated as $[Concat(s_i, c_i), x_i, x_f, x_b]$, enabling joint attention computation, as illustrated in Fig. 4. This mechanism effectively integrates background and foreground information into the fused image through the attention layers.

Positional Affine: We explore three approaches to incorporate positional information, as shown in Fig. 5. The most straightforward approach is to directly transform the foreground to match the desired position and size in the background (Fig. 5(b)). However, this method compresses the information of foreground during scaling, which is unfavorable for inserting small objects. Another approach involves using the placement information of the foreground, such as the mask after positioning, as a condition. This information is encoded via a tokenizer and introduced into the attention computation (Fig. 5(c)). However, this approach relies heavily on the tokenizer, requiring a large amount of data to optimize its representation of positional information. To leverage the relative positional relationship of the foreground

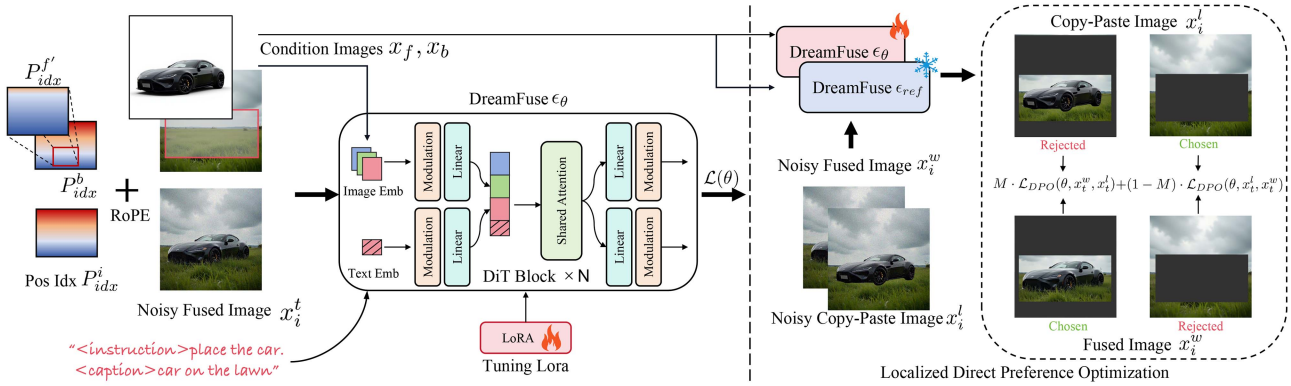


Fig. 4. The framework of the DreamFuse. We apply positional affine transformations to map the foreground's position and size onto the background. The foreground and background are concatenated with the noisy fused image as condition images before DiT's attention layers. Localized direct preference optimization is then used to improve background consistency and foreground harmony.

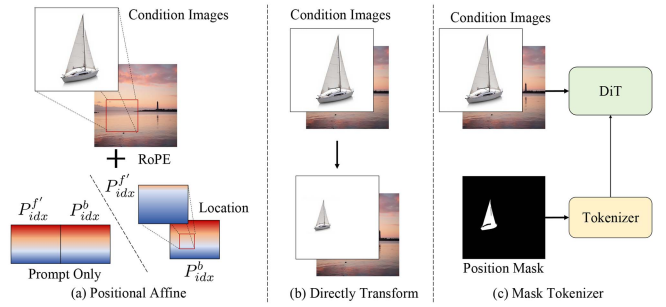


Fig. 5. Three ways for injecting positional conditions: (a) using positional affine to map the foreground's position index to its target placement; (b) directly transforming the foreground object to the target position; (c) encoding position mask information with a tokenizer and integrating it into DiT's attention computation.

more directly and effectively, we propose the positional affine method shown in 5(a).

Specifically, both the foreground and background are assigned 2D position index $P_{idx}^f, P_{idx}^b = (i, j), \forall i \in [0, h), j \in [0, w)$ to represent their spatial relationships within the image. When placing the foreground in a target region $P_{idx}^r = (u, v), \forall u \in [h_r, h_r'), v \in [w_r, w_r')$ of the background, the affine transformation matrix A is computed as follows:

$$A = \begin{bmatrix} \frac{w_r' - w_r}{w} & 0 & w_r \\ 0 & \frac{h_r' - h_r}{h} & h_r \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

For fusion cases that do not require explicit positional input and rely on prompts for guidance, we set the position index of the foreground image to a fixed offset w and the affine transformation matrix A_p is defined as:

$$A_p = \begin{bmatrix} 1 & 0 & w \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{6}$$

Next, the position index P_{idx}^r of the target region is mapped to the foreground with the inverse affine transformation:

$$P_{idx}^{f'} = A_{(p)}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{7}$$

We utilize $P_{idx}^{f'}$ as the new position index of foreground. By employing this positional affine transformation, and leveraging DiT's responsiveness to position index, we directly incorporate the position and size information of the foreground into the target location within the background. This approach eliminates the need to scale or compress the foreground, enabling a more effective and reasonable integration of positional information.

Localized Preference Optimization: In the image fusion process, maintaining background consistency and foreground harmony is crucial. When directly generating the denoised fused image, issues such as inconsistent backgrounds or disharmonious foregrounds can easily arise. To address this, we propose Localized Direct Preference Optimization (LDPO) based on Diffusion-DPO [44], [45], enabling the diffusion network to more effectively learn from human preferences in the context of image fusion.

We construct a dataset $\mathcal{D}' = \{(c_i, x_f, x_b, x_i^w, x_i^l)\}$ consisting of additional fused sample pairs, where x_i^w aligns better with human preferences than x_i^l. For example, as shown in Fig. 4, x_i^l can be a fused image obtained by directly copying and pasting the foreground onto the background. For simplicity, we define the model input as $x_t^{\{w,l\}} = (c_i, x_f, x_b, x_i^{t,\{w,l\}})$. The Diffusion-DPO optimizes a policy to satisfy human preferences via objective $\mathcal{L}_{DPO}(\theta, x_t^w, x_t^l)$:

$$\mathbb{E}\left[\log\sigma\left(-\frac{\beta}{2}(||\epsilon^w - \epsilon_\theta(x_t^w, t)||^2 - ||\epsilon^w - \epsilon_{ref}(x_t^w, t)||^2 - (||\epsilon^l - \epsilon_\theta(x_t^l, t)||^2 - ||\epsilon^l - \epsilon_{ref}(x_t^l, t)||^2))\right)\right], \tag{8}$$

where $\epsilon_\theta(\cdot)$ and $\epsilon_{ref}(\cdot)$ denotes the predictions of optimized model and reference model, respectively, β is the regularization

Algorithm 1: Localized Direct Preference Optimization Loss.

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

coefficient and σ is the sigmoid function. Intuitively, minimizing $\mathcal{L}_{DPO}$ encourages the predicted velocity field ϵ_θ closer to the target velocity ϵ^w of the chosen data, while diverging from ϵ^l (the rejected data). However, not all aspects of x_i^l fail to align with human preferences. For instance, in copy-paste fused images, the consistency of the background better aligns with human preferences. Therefore, we adopt a Localized DPO strategy for x_i^w and x_i^l. A localized foreground region $M(f)$ is defined as:

$$M(f) = \begin{cases} 1, \text{if} f \in \alpha \cdot Bbox(x_f) \\ 0, \text{otherwise}, \end{cases} \tag{9}$$

where f represents a pixel location, $Bbox(x_f)$ denotes the region in the bounding box of the foreground object and α is a dilation factor that moderately expands this region. The optimized objective $\mathcal{L}_{LDPO}(\theta, x_t^w, x_t^l, M)$ is defined as:

$$M \cdot \mathcal{L}_{DPO}(\theta, x_t^w, x_t^l) + (1-M) \cdot \mathcal{L}_{DPO}(\theta, x_t^l, x_t^w). \tag{10}$$

We provide the pseudo-code for LDPO in the Algorithm 1. This strategy ensures background consistency while making the foreground more harmonious and aligned with human preferences.

V. Experiments

A. Data Analysis

As outlined in Section III, we adopt a comprehensive pipeline for image fusion data generation. By constructing diverse databases of foregrounds, backgrounds, and tasks, and leveraging an MLLM to create coherent fusion prompts, we generate fused samples through two distinct generation pipelines. The results are further refined using both a MLLM and rule-based inspection, yielding a total of 160 k fused image samples. As illustrated in Fig. 6(a), we conduct a thorough analysis of the dataset. The distribution of categories covers three major scenarios: replacement, insertion, and attribute-referenced editing, encompassing a wide range of image classes from products to symbols. Within the insertion and replacement scenarios, the dataset includes object placement, style transfer, logo printing, as well as numerous interactive cases such as handheld and wearable objects. For attribute-referenced editing, the data primarily focuses on attributes such as color, texture, and material. We employed GPT-4o and Seed-VL-1.5 to jointly evaluate sample quality across five dimensions: Foreground Consistency, Background Integrity, Global Composition Harmony, Background Cleanliness & Artifacts, and Instruction Adherence. To validate the filtering effectiveness, we report the data retention rates in Fig. 6(b). Notably, the "Symbols" category exhibits the lowest retention, which aligns with Flux's known limitations in text generation. This corroborates that our pipeline effectively identifies and purges low-quality samples, ensuring the high standard of the final dataset. Fig. 6(c) further reports the top-10 foreground categories (excluding human-related classes). In addition, our dataset exhibits substantial scale diversity, as shown in Fig. 6(d), ranging from resolutions of 400 to 1600, making it better aligned with real-world application scenarios.

B. Implementation Details

1) Hyperparameters: For the self-inspired in-context model training in the data generation pipeline, starting with the first batch of data, we use Flux-Dev [50] as the base model. Input images are randomly scaled to 512, 768, or 1024 resolutions, and the model is trained for 10 k iterations on 8 A100 GPUs using the Prodigy [63] 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 and Seed1.5-VL [48], followed by manual selection of high-quality fusion data for the next training iteration.

For DreamFuse training, we adopt Flux-Dev [50] as the base model. The input images are scaled proportionally from their original resolution to a short side of 512 pixels. The training procedure comprises two stages: first, the full DreamFuse 160 k dataset is trained for 20 k iterations with batch size 1,

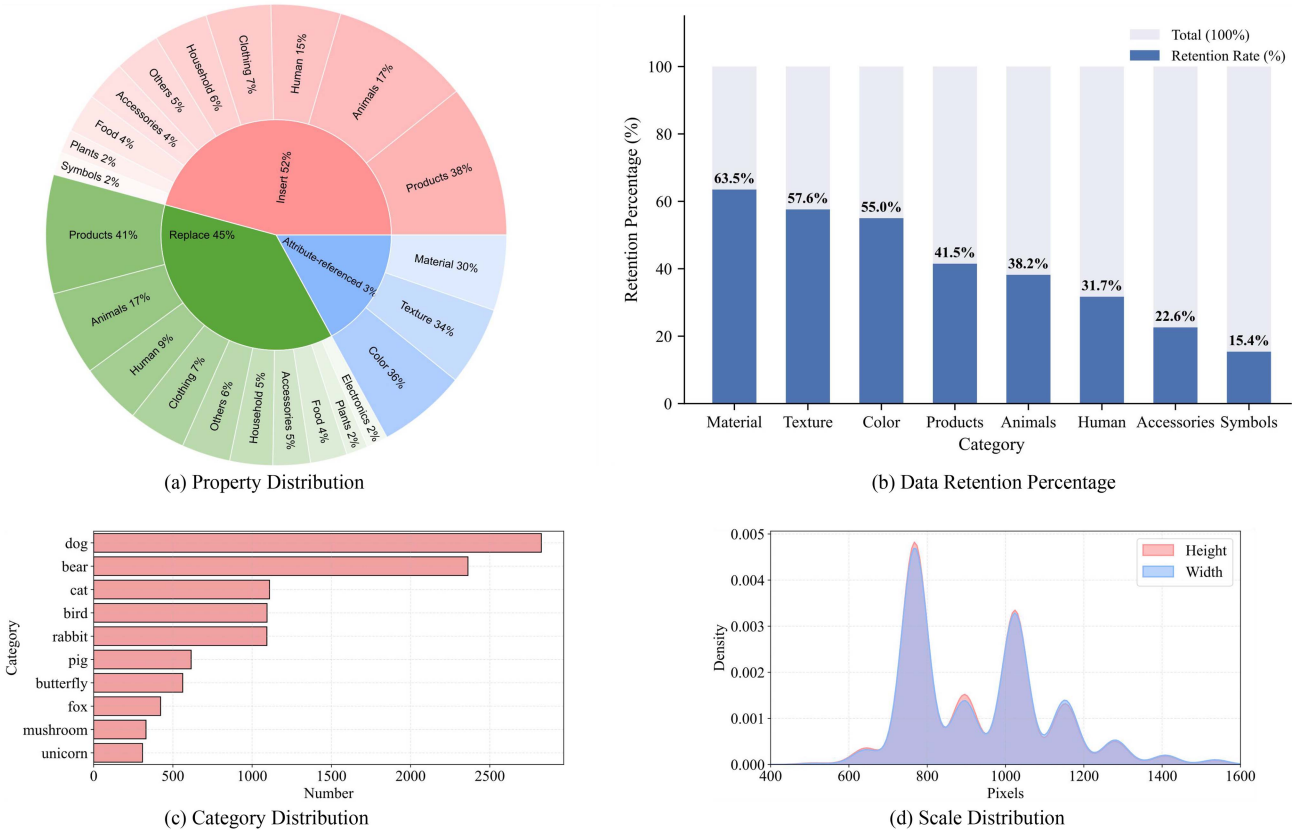


Fig. 6. Visualization of data distribution. (a) Distribution of fusion categories, including insertion, replacement, and attribute-referenced editing, covering diverse image types from products to symbols. Attribute-referenced editing further includes color, material, and texture. (b) Data retention percentage after quality evaluation. (c) Top-10 foreground categories by count (excluding humans). (d) Distribution of image resolutions.

the Prodigy [63] optimizer, and a LoRA rank of 16. Subsequently, we manually select 15 k high-quality samples from the initial dataset for LDPO training. Negative samples x_i^l in LDPO training consist of two types: copy-paste images (15 k) and low-quality inference results generated from the selected samples after stage one (15 k). We apply the $\mathcal{L}_{LDPO}$ loss to the copy-paste data and $\mathcal{L}_{DPO}$ loss to remaining negative samples. The second stage employs the AdamW [64] optimizer (learning rate 5×10^{-5}) for 10 k iterations. During training, we randomly drop the instruction prompts with a probability of 0.01. This strategy is primarily employed to enhance the model's robustness, ensuring it can adapt to complex scenarios or situations where textual inputs might be missing. The dilation factor α is set to 1.5 and regularization parameter is set to 5000. The entire training is conducted on 8 NVIDIA A800 GPUs, requiring approximately 36 hours.

2) Benchmarks: We randomly selected 1500 unseen samples from the generated dataset as DreamFuse test set to evaluate the model's fusion capabilities across multiple scenarios, including object placement, wearing, logo printing, handheld and style transfer, replacement and attribute-referenced editing. Additionally, we evaluated method's performance on out-of-domain data with the TF-ICON [17] dataset, which consists of 332 samples spanning four visual domains: photorealism, pencil sketching, oil painting, and cartoon animation. To further validate the effectiveness of DreamFuse in real-world scenarios, we conducted additional experiments on the FOSCom [65] 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 [70] training set. Furthermore, To rigorously validate our method's effectiveness in real scenarios, we curated a dedicated "Real Test Dataset" consisting of 120 real-world samples. This dataset covers diverse challenging scenarios, including object insertion, replacement, and virtual try-on.

3) Evaluation Metrics: We utilize the CLIP [57] score to evaluate the similarity between the fused images and their corresponding descriptive texts, the AES [58] score to assess the aesthetic quality of the fused images. Additionally, we employ the VisionReward [41] (VR) score, which leverages a vision language model [71] (VLM) to evaluate the fused results from multiple perspectives across various questions, better reflecting human preferences. Since the aforementioned evaluation metrics can only assess individual images in isolation, we utilize VIEScore [56] to introduce an explainable metric for evaluating conditional image generation tasks, specifically image fusion. Concretely, we employ an MLLM to examine the fused foreground, background, and fused image, providing ratings

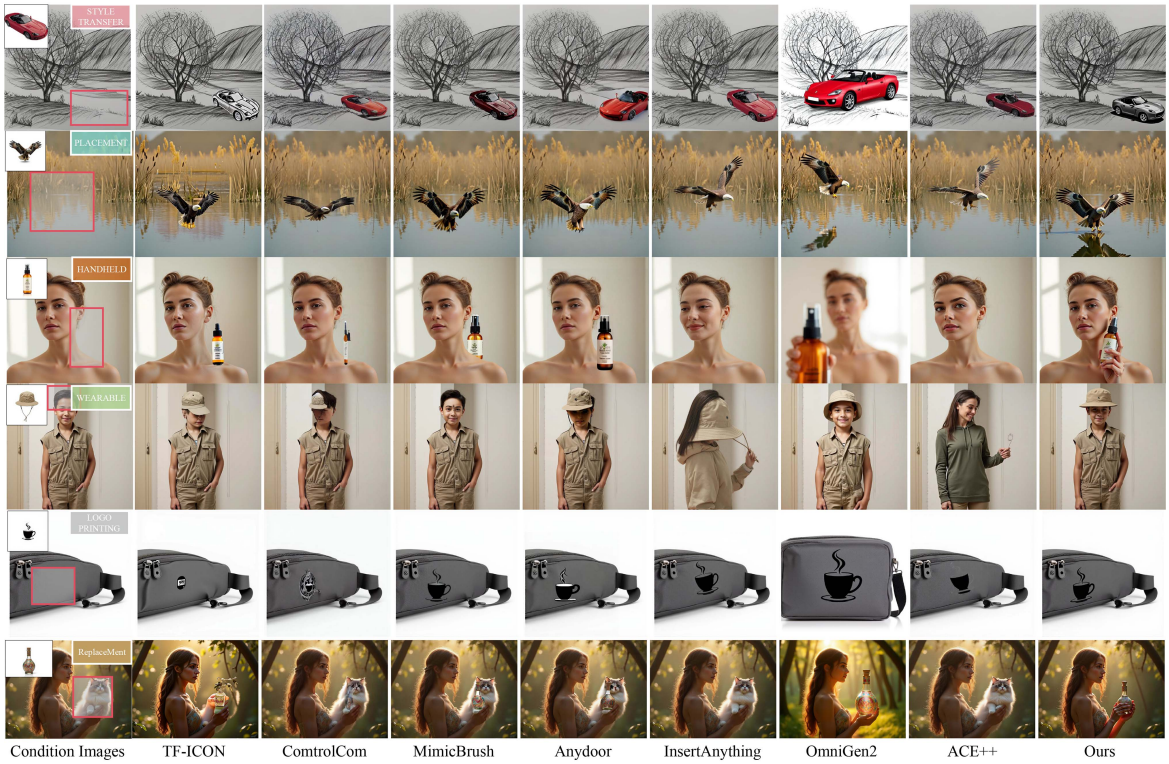


Fig. 7. Qualitative comparisons with existing methods in various fusion tasks, including placement, replacement, logo printing, style transfer, handheld, wearable items. The first row is from the TF-ICON dataset, while the others are from the DreamFuse test set. Our approach achieves a more seamless integration of foreground objects with the background images, resulting in higher visual consistency and realism.

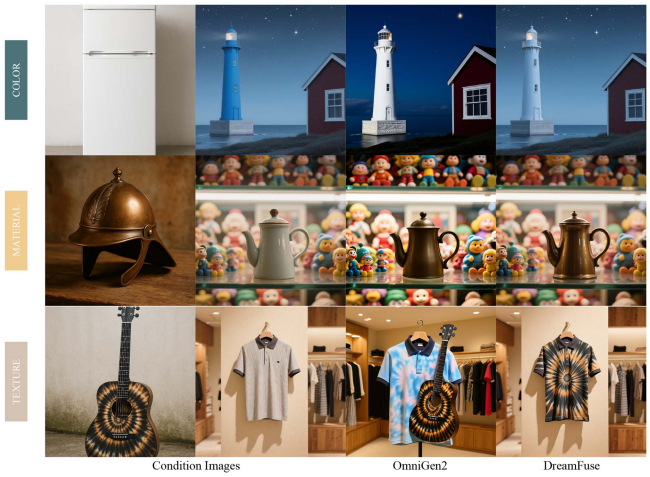


Fig. 8. Qualitative comparisons with OmniGen2 in attribute-referenced editing task, including color, material and texture.

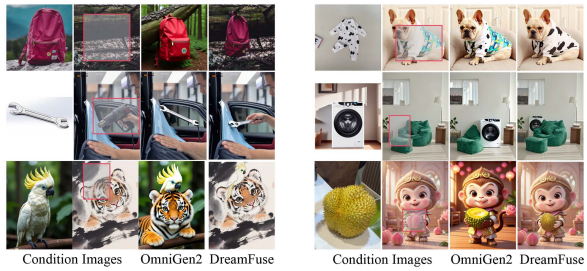


Fig. 9. Qualitative comparison on the collected challenging Real-world Test Dataset.

from two perspectives: (1) Semantic Consistency (SC), which evaluates foreground consistency and the degree of background over-editing; and (2) Perceptual Quality (PQ), which evaluates naturalness and the presence of artifacts.

C. Comparisons With Existing Methods

1) Quantitative Results: We evaluate several existing state-of-the-art image fusion methods on the TF-ICON and DreamFuse test datasets, including image composition methods such as TF-ICON [17], Anydoor [9], MADD [11] and ControlCom [65], InsertAnything [38], as well as the reference-based fusion method MimicBrush [66], ACE++ [37], OmniGen2 [39]. As shown in Table I, our method (based on Flux) outperforms existing approaches across multiple metrics on DreamFuse test dataset, with a notable 1.14 improvement in the SC score compared to the second-best method. Our method consistently demonstrates better fusion quality across all scenarios. Similarly, evaluations on the FOSCom dataset and TF-ICON dataset in Table II, which comprises both realistic and diverse-domain images, further demonstrate the robustness and strong generalization ability of our approach.

Additionally, Figs. 10 and 11 present the PQ and SC scores of different methods on the DreamFuse test dataset across

TABLE I
QUANTITATIVE EVALUATION OF OUR METHOD ACROSS DIFFERENT DIT-BASED BACKBONES (E.G., SD3, FLUX, QWEN-IMAGE) ON THE DREAMFUSE AND REAL-WORLD TEST DATASETS

Base Model	DreamFuse Test					Real Test				
	CLIP ↑	AES ↑	VR ↑	PQ ↑	SC ↑	CLIP ↑	AES ↑	VR ↑	PQ ↑	SC ↑
ControlCom [65]	31.66	5.64	1.53	7.62	5.51	32.66	5.50	1.28	7.41	6.22
TF-ICON [17]	33.78	5.95	3.94	7.90	5.89	33.88	5.76	2.90	7.18	5.42
Anydoor [9]	32.52	5.95	2.89	7.80	6.61	33.11	5.77	2.37	7.23	6.77
MADD [11]	31.17	5.61	0.69	6.65	4.21	31.68	5.25	0.93	6.44	4.57
MimicBrush [66]	32.35	5.98	2.94	7.48	5.56	31.59	5.72	1.68	6.08	4.25
ACE++ [37]	30.92	6.10	4.81	8.24	4.93	33.23	6.03	4.81	8.12	7.04
OmniGen2 [39]	32.33	6.26	5.93	8.48	6.86	33.89	6.13	5.70	8.38	7.56
InsertAnything [38]	32.45	6.08	4.52	8.22	6.55	33.36	6.05	4.33	7.98	7.60
Ours (SD3 [67])	34.30	6.22	6.35	8.55	7.80	34.65	**6.19**	5.54	8.46	7.13
Ours (Flux [50])	34.94	6.27	6.26	**8.59**	8.00	34.19	6.13	**6.12**	**8.69**	7.51
Ours (Qwen-Image [68])	**35.19**	**6.33**	**6.36**	8.55	**8.06**	**34.68**	6.15	5.81	8.61	**7.91**
Ours + Real Data [69]	35.20	6.30	6.35	8.56	8.11	34.71	6.12	5.80	8.63	7.93

TABLE II
QUANTITATIVE EVALUATION RESULTS ON FOSCOM DATASET AND TF-ICON DATASET

Method	ControlCom					TF-ICON [17]				
	CLIP ↑	AES ↑	VR ↑	PQ ↑	SC ↑	CLIP ↑	AES ↑	VR ↑	PQ ↑	SC ↑
ControlCom [65]	32.20	5.14	0.72	7.05	7.01	31.28	6.09	1.33	8.09	6.56
TF-ICON [17]	32.95	5.51	4.10	7.32	6.91	**32.28**	6.51	2.37	8.31	7.21
Anydoor [9]	32.58	5.31	1.40	7.08	7.04	31.80	6.27	1.31	7.93	**7.70**
MADD [11]	31.32	4.96	0.21	6.57	6.29	30.36	5.94	0.61	7.55	6.59
MimicBrush [66]	31.13	5.04	1.78	6.89	6.00	31.71	6.27	0.83	7.17	7.28
ACE++ [37]	29.98	4.93	2.71	7.48	4.38	30.03	6.22	2.01	7.84	6.07
OmniGen2 [39]	33.12	**5.54**	**4.20**	8.26	6.95	31.73	6.55	**4.74**	8.37	6.08
InsertAnything [38]	32.55	5.37	1.73	6.53	**7.35**	31.77	6.41	1.71	7.59	7.25
Ours (Flux [50])	**33.55**	**5.54**	4.14	**8.35**	7.10	31.98	**6.56**	3.40	**8.41**	7.34

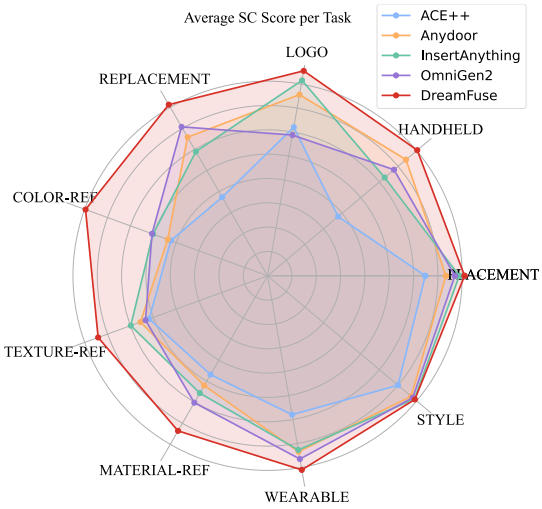


Fig. 10. Average Semantic Consistency (SC) Score for Various Tasks.

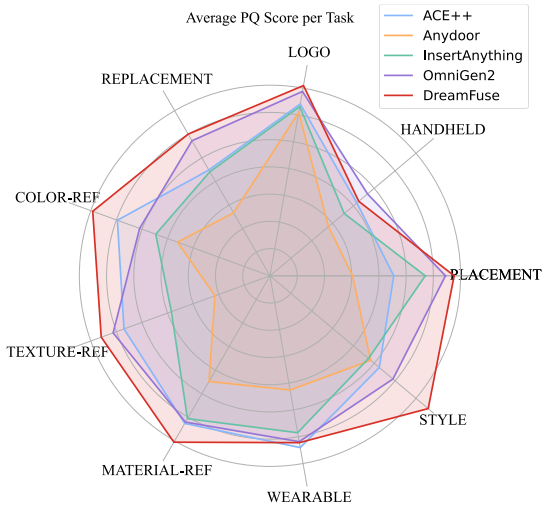


Fig. 11. Average Perceptual Quality (PQ) Score for Various Tasks.

nine fusion scenarios, including placement, replacement, logo printing, style transfer, handheld, wearable items, and material-, color-, and texture-referenced editing. As shown, our method consistently achieves superior results in SC across all scenarios, demonstrating its strong capability in maintaining editing consistency and faithfully following the given instructions.

2) Qualitative Results: Fig. 7 presents the qualitative visualization results of our method compared with other methods. As shown, our method not only performs well across various fusion scenarios but also surpasses inpainting approaches that require strict mask inputs, by generating more harmonious integrations beyond the target region. For instance, in the second row with the eagle fusion, our method adapts to the surrounding environment and automatically generates a realistic shadow of the eagle on the surface of water. Moreover, our approach is not limited to insertion tasks; it also supports replacement tasks, such as substituting the cat in the person's hand with the input bottle (sixth row).

TABLE III
USER STUDY RESULTS BASED ON PREFERENCE RANKING (1-5, LOWER IS BETTER). THE EVALUATION COVERS THREE PERSPECTIVES: HARMONY, CONSISTENCY, AND OVERALL QUALITY. **KENDALL'S W** INDICATES THE INTER-RATER AGREEMENT. STATISTICAL SIGNIFICANCE IS VERIFIED, WHERE * INDICATES $p < 0.01$ COMPARED TO OURS.

Methods	Harmony↓	Consistency↓	Overall↓
TF-ICON [17]	3.98*	4.35*	4.39*
Anydoor [9]	3.72*	3.86*	3.95*
InsertAnything [38]	3.17*	2.95*	3.30*
OmniGen2 [39]	2.42*	3.45*	3.10*
DreamFuse (Ours)	**1.85**	**2.54**	**2.13**
Inter-rater Agreement	*0.45*	*0.43*	*0.48*

In Fig. 8, we further compare our method with OmniGen2 [39] in terms of attribute-referenced editing, focusing on color, material, and texture. As shown, our method demonstrates stronger background preservation and more accurate attribute transfer from the referenced foreground compared to OmniGen2.

In Fig. 9, we present the qualitative visual comparisons on the Real Test dataset. As evident from the results, our method significantly outperforms OmniGen2 [39] in these real-world scenarios, generating more harmonious and consistent results. This advantage is observed across diverse challenging situations, ranging from complex tasks like object replacement and style transfer, to intricate scenes involving occlusions (e.g., placing a fridge behind a sofa).

3) User Study: We randomly select 100 samples and conduct a user study with 50 participants to compare our method with InsertAnything [38], OmniGen2 [39], Anydoor [9] and TF-ICON [17]. As presented in Table III, we adopted a preference ranking protocol to evaluate the results across three dimensions: Harmony, Consistency, and Overall Quality. The results demonstrate the superiority of our approach, which achieves the best average rankings compared to state-of-the-art methods (e.g., achieving an average rank of 1.85 in Harmony). Furthermore, we validated the reliability of our study using Kendall's Coefficient of Concordance (Kendall's W) [72], which indicates a reasonable level of inter-rater agreement. Subsequent statistical significance tests also confirm that our method significantly outperforms competing methods ($p < 0.01$).

D. Ablation Study

We carry out extensive ablation studies to verify the effectiveness of our DreamFuse. Notably, owing to the sequence of data acquisition, our ablation studies on Positional Affine, Localized Preference Optimization, and the dilation factor α are primarily performed on the insertion subset, which guides the determination of the final architectural choices and parameter settings.

1) Effectiveness of the Offset Δ for Self-Inspired In-Context Generation Model: 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 [3], 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. 3, models trained with $\Delta = (0, 0)$ exhibit stronger consistency, often generating similar backgrounds and fused images.

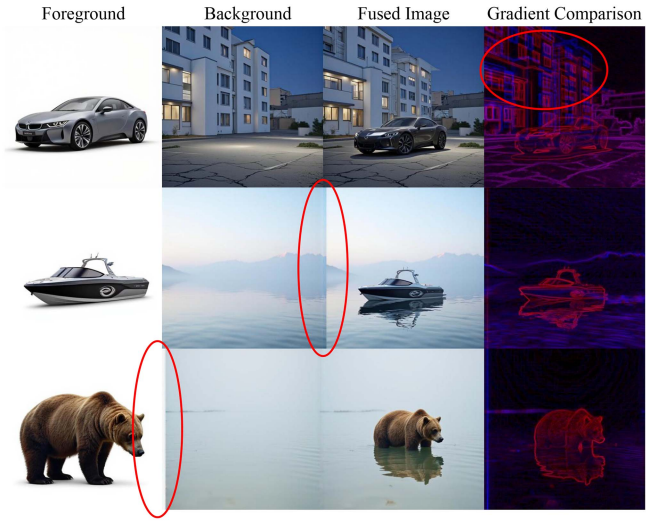


Fig. 12. Misalignment often occurs when $\Delta = (0, w)$. "Gradient Comparison" illustrates the gradient comparison between the background and the fused image.

TABLE IV
QUALITATIVE ANALYSIS OF THREE POSITIONAL INCORPORATION STRATEGIES ON THE DREAMFUSE INSERTION TEST SET AFTER THE FIRST-STAGE TRAINING

Method	CLIP	AES	VR	PQ	SC
Direct Trans.	35.04	6.10	4.72	8.27	7.65
Mask Token.	35.14	6.11	4.93	8.29	7.80
Pos. Affine	**35.22**	**6.13**	**5.28**	**8.35**	**7.95**

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. 12, 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.

2) Positional Affine: We compare the three positional incorporation strategies discussed in Section IV-A. As shown in

TABLE V
QUALITATIVE ANALYSIS OF LOCALIZED PREFERENCE OPTIMIZATION ON THE INSERTION SUBSET

DreamFuse	CLIP	AES	VR	PQ	SC
First-Stage Training	35.22	**6.13**	5.28	8.35	7.95
w/ $\mathcal{L}_{DPO}$	35.36	6.07	5.52	8.39	8.05
w/ $\mathcal{L}_{DPO}$ + $\mathcal{L}_{LDPO}$	35.51	6.10	**5.58**	8.39	8.10
w/ $\mathcal{L}_{DPO}$ + $\mathcal{L}_{LDPO}$ + $\mathcal{L}_{noise}$	**35.52**	6.10	5.56	**8.41**	**8.13**

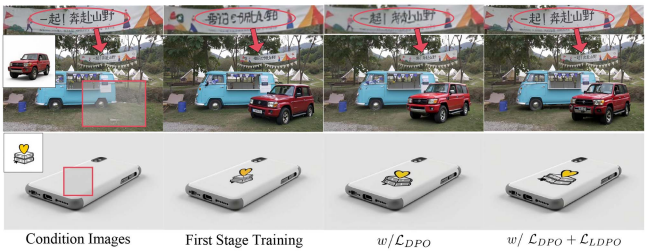


Fig. 13. Qualitative results about the LDPO. Compared to "$w/\ \mathcal{L}_{DPO}$", "$w/\ \mathcal{L}_{LDPO}$" leverages copy-paste data to better help the model understand perspective changes in the foreground while maintaining background consistency as much as possible.

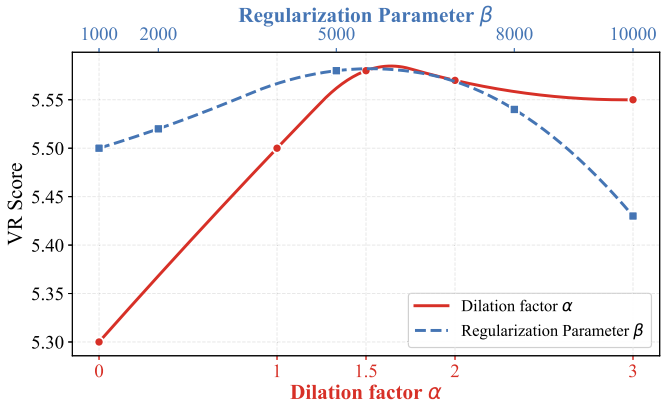


Fig. 14. The effectiveness about the dilation factor α and regularization parameter β.

TABLE VI
THE EFFECTIVENESS OF DIFFERENT TEXT PROMPT. $\checkmark_p$ REPRESENTS THE ADOPTION OF A_p AS THE AFFINE TRANSFORMATION MATRIX IN THE POSITIONAL AFFINE MODULE, WHERE NO EXPLICIT FOREGROUND POSITION OR SCALE IS PROVIDED, ENABLING THE MODEL TO AUTONOMOUSLY SELECT SUITABLE PLACEMENT AND SIZING.

Instruction	Caption	CLIP	AES	VR	PQ	SC
		33.82	6.26	6.13	8.55	7.45
✓		34.40	6.24	6.25	8.59	7.86
	✓	**35.04**	6.27	6.26	8.59	7.86
$\checkmark_p$	$\checkmark_p$	34.93	6.26	**6.30**	8.59	7.90
✓	✓	34.96	**6.27**	6.26	**8.59**	**8.00**

Table IV, the results on the DreamFuse test set after the first-stage training demonstrate that the positional affine approach achieves the highest CLIP and VR scores, reaching 35.22 and 5.28, respectively. This suggests that the positional affine method better preserves the information of foreground objects, outperforming the direct transform and mask tokenizer strategies, which are more prone to information loss of fine-grained details.

3) Localized Preference Optimization: In the second stage of training, we employ Localized Direct Preference Optimization (LDPO) to further enhance the model's performance by utilizing two types of paired data: the first type consists of poorly performing and well-performing results generated after the first training stage with different seeds and optimized with the $\mathcal{L}_{DPO}$, while the second type involves copy-paste data treated as negative samples and optimized with the $\mathcal{L}_{LDPO}$. As shown in Table V, "$w/\ \mathcal{L}_{DPO}$" refers to training with only the first type of data, "$w/\ \mathcal{L}_{LDPO}$" refers to training with copy-pasted data. Results indicate the LDPO significantly improves fusion performance by achieving an approximate 0.11 increase in IR scores compared to the first-stage results. Further incorporating the $\mathcal{L}noise$ loss results in a comparable performance. LDPO primarily enhances background consistency and foreground harmony, as visually demonstrated in Fig. 13, where using copy-paste data as negative samples allows the model to better capture fusion-related transformations, such as perspective and affine adjustments, rather than simply copying and pasting the foreground, while also improving the consistency of other background regions.

4) Dilation Factor α and Regularization Parameter β: We conduct a detailed ablation study on α and β, as shown in Fig. 14. For α, setting it to 0 treats copy-paste data as positive samples, leading to a significant drop in VR scores. Conversely, $\alpha = 1$ restricts the model's focus strictly to the bounding box, neglecting essential external context such as shadows. Increasing α to 1.5 appropriately relaxes this constraint, enabling better blending while maintaining global harmony. Regarding β, which balances reward optimization and KL divergence, we find that a low β leads to reward overfitting and instability, while an excessively high β over-regularizes the model, hindering improvement. Based on these results, we select $\alpha = 1.5$ and $\beta = 5000$ as the optimal configuration.

5) Effectiveness of Different Text Prompt: In our image fusion dataset, three types of tasks are considered: insertion, replacement, and attribute-referenced editing. Providing only the caption of the fused image makes it difficult for the model to accurately distinguish among these tasks. Therefore, as discussed in Section IV-A, we introduce a combined text prompt with both `<instruction>` and `<caption>` to enhance task differentiation. Table VI presents the results under different text prompt settings. As shown, when `<instruction>` is omitted, the model often confuses insertion and replacement tasks, leading to a decline in SC scores; in particular, when both indicators are absent, the SC score drops to 7.45. Moreover, in the Positional Affine module described in Section IV-A, we address cases where explicit foreground position and scale are unnecessary by adopting A_p as the affine transformation matrix. As shown in the fourth row of Table VI, this setting achieves results that remain competitive with those obtained when positional information is included, demonstrating that our model can generate coherent fusion results even without explicit foreground placement and sizing.

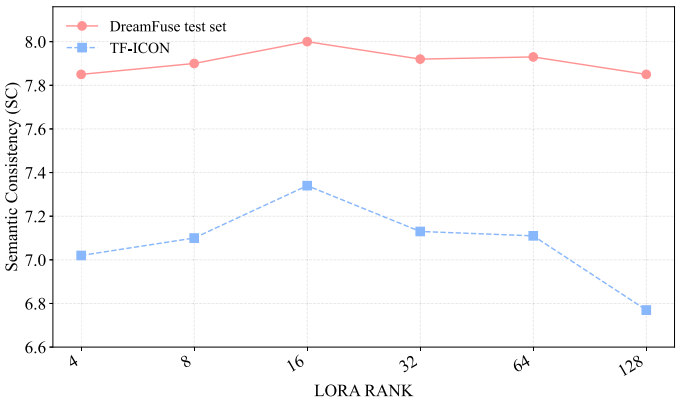


Fig. 15. The effectiveness about the LoRA Rank on dreamfuse test set and TF-ICON dataset.

TABLE VII
THE EFFECTIVENESS AND LATENCY ABOUT THE INFERENCE RESOLUTION

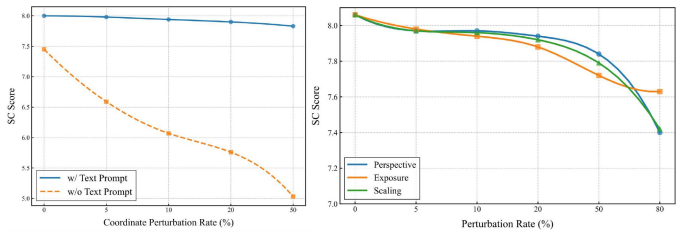
resolution	CLIP	AES	VR	PQ	SC	Latency (s)	Memory (GB)
512	34.77	6.19	6.17	8.49	7.88	18.48	32.28
768	**34.98**	6.26	6.14	8.59	7.98	35.88	33.24
1024	34.96	6.27	6.26	**8.59**	**8.00**	58.98	34.58
1280	34.81	**6.29**	**6.32**	8.58	7.95	113.24	36.93
1536	34.55	6.25	6.21	8.55	7.74	174.21	38.40

6) Effectiveness of LoRA Rank: We investigate the effectiveness of LoRA Rank on the training of the DreamFuse model with the DreamFuse and TF-ICON datasets. The model is trained exclusively on the DreamFuse training set, while inference is performed on both the DreamFuse test set and the TF-ICON dataset. As illustrated in Fig. 15, we report the SC scores on both datasets. When the LoRA Rank is set to a relatively low value of 4, the model lacks sufficient capacity to fit the diverse scenarios of image fusion, resulting in low SC scores. As the LoRA Rank increases, the model's fitting ability improves. However, when the LoRA Rank reaches 128, performance drops significantly, particularly on the out-of-domain TF-ICON dataset, indicating signs of overfitting. Based on these observations, we adopt a LoRA Rank of 16 for training.

7) Effectiveness of Inference Resolution: Due to limited training resources, we trained the model using images with a resolution based on a 512-pixel short side. However, we observe that employing higher resolutions during inference can effectively enhance the quality of fused image generation. As shown in Table VII, we compare inference results across resolutions ranging from 512 to 1536. At 512, the results are suboptimal, with an AES score of only 6.19. As the inference resolution increases, the quality of fusion improves. Nevertheless, when the resolution becomes excessively high, such as 1536, the large gap from the training resolution leads to a significant performance drop, with the SC score falling to 7.74. Furthermore, to analyze the impact of input resolution, we report the processing times across various resolutions in Table VII. As observed, while inference time naturally increases with higher resolutions, the GPU memory footprint remains remarkably stable, showing no significant increase.

TABLE VIII
RELIABILITY ANALYSIS OF MLLM-BASED METRICS. WE REPORT THE CORRELATION WITH HUMAN JUDGMENTS (CALIBRATION) AND THE STATISTICAL STABILITY ACROSS 5 RUNS (MEAN ± STD).

Metrics	Human-Human	MLLM-Human	MLLM Score
VR	0.6907	0.6335	6.27 ± 0.001
PQ	0.5855	0.5775	8.57 ± 0.002
SC	0.5644	0.5626	8.04 ± 0.001



(a) Bounding box noise (b) More edge cases

Fig. 16. Robustness analysis on edge cases. We evaluated performance under challenging edge conditions, including bounding box noise, extreme object scaling, unusual perspectives, and overexposure.

8) Performance on Different DiT Backbone: Our method exhibits remarkable extensibility and can be seamlessly adapted to various Diffusion Transformer (DiT) backbones. As presented in Table I, we evaluated our approach on three architectures: SD3 [32], Flux [50], and Qwen-Image [68]. Empirical results confirm that our method remains robust across different architectures. Furthermore, the fusion performance scales with the capacity of the base model. The more advanced Qwen-Image sets a higher performance ceiling, achieving a top VR score of 6.36 on the DreamFuse Test dataset.

9) Reliability of the MLLM: To validate the reliability of our MLLM-based metrics (VR, PQ, SC), we conducted a comprehensive analysis of human alignment and statistical stability. We compared MLLM ratings against three professional annotators on a subset of 240 samples. Table VIII reveals that Human-to-MLLM correlations match inter-human agreement, validating the MLLM's accuracy. Furthermore, repeating the evaluation five times yielded negligible variance (≈ 0.002), confirming the stability of the metrics. These results ensure that our evaluation protocol is a robust proxy for human preference.

10) Edge Case Study: To comprehensively assess model robustness, we evaluated performance under challenging edge conditions, including bounding box noise, extreme object scaling, unusual perspectives, and overexposure. First, regarding bounding box sensitivity, we introduced random perturbations to scale and position (0%-50%) under settings with and without text prompts. As shown in Fig. 16(a), the inclusion of text prompts significantly enhances resilience; even at a severe perturbation of 50%, the drop in Semantic Consistency (SC) remains marginal ($\approx$ 0.2). Furthermore, we conducted a stress test on extreme scaling, unusual perspectives, and overexposure by varying perturbation intensities from 5% to 80% (Fig. 16(b)). The results demonstrate strong stability, with significant performance degradation occurring only at extreme levels (50%-80%). This

TABLE IX
EVALUATION OF TRY-ON (GARMENT) AND IDENTITY PRESERVATION (HUMAN) ON THE ANYINSERTION BENCHMARK

Methods	AnyInsertion (Garment)				AnyInsertion (Human)				
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	Face-Score ↑
AnyDoor [9]	17.98	0.8202	0.2065	72.98	14.71	0.6807	0.3613	217.17	28.93
MimicBrush [66]	18.73	0.8455	0.1708	76.73	20.58	0.7654	0.2125	108.26	25.42
ACE++ [37]	18.11	0.7507	0.1086	35.62	19.21	0.7513	0.1529	66.84	48.73
InsertAnything [38]	**23.78**	0.8665	**0.0522**	28.54	**23.85**	**0.8457**	**0.1269**	**52.77**	**54.19**
Ours	23.23	**0.8733**	0.0754	**28.16**	23.47	0.8373	0.1347	54.65	52.62

TABLE X
COMPARISON OF COMPUTATIONAL EFFICIENCY AND MODEL STATISTICS

Methods	Base Model	Resolution	Latency (s)	Memory (GB)	Params (B)
ControlCom [65]	SD2.1	512	15.81	10.69	1.37
TF-ICON [17]	SD2.1	512	49.81	17.96	1.30
Anydoor [9]	SD2.1	512	21.22	15.80	2.45
MADD [11]	-	512	35.53	10.15	0.87
MimicBrush [66]	SD1.5	512	19.44	8.12	1.07
ACE++ [37]	Flux	1024	28.94	34.89	11.90
OmniGen2 [39]	OmniGen2	768	125.25	16.11	3.98
InsertAnything [38]	Flux	768	38.84	36.85	12.83
Ours (SD3)	SD3	768	35.9	28.47	8.43
Ours (Flux)	Flux	768	18.48	32.28	11.92
Ours (Qwen-Image)	Qwen-Image	768	64.72	55.71	20.43

(a) More Diverse Examples from Real-World Scenarios

(b) Some Failure Cases from Real-World Scenarios

Fig. 17. More diverse examples and some failure cases from real-world scenarios.

indicates that our model can effectively withstand substantial input variations before failure.

11) Computational Efficiency: We provide a detailed comparison in Table X, covering the base model, inference resolution, inference latency, GPU memory footprint and model parameters for all compared methods. As evidenced by the results, our method maintains competitive efficiency without incurring significant computational overhead. Despite the performance gains, our inference latency and memory usage remain within a reasonable range compared to baseline methods.

12) Syntheic vs. Real Data Gap: To investigate potential synthetic-to-real domain gaps, we evaluated our model on the 'Real Test Dataset'. The results confirm excellent generalization, with our method outperforming the runner-up by a margin of 0.31 in Semantic Consistency (SC). Furthermore, we explored a hybrid training strategy by incorporating 1,300 real-world pairs from OmniPaint [69] (using Qwen-Image). As shown in Table I, while mixing real data improves consistency (SC) and text alignment (CLIP), it leads to a slight decline in aesthetic metrics (VR). This is likely because raw real-world samples often lack the high aesthetic quality inherent in synthetic data. This finding suggests that while real data is beneficial, its scarcity and variable quality make large-scale synthetic data a more scalable and cost-effective solution for high-quality generation.

E. Limitations

1) Challenging Case Analysis: To evaluate generalizability, we tested our method on a diverse set of real-world images. As shown in Fig. 17(a), the model handles complex scenarios—such as harsh lighting, hand interactions, and occlusions—robustly. However, we also observe failure cases (Fig. 17(b)), particularly involving identity consistency, anatomical artifacts (e.g., extra limbs), and virtual try-on scenarios. We quantified these limitations on the AnyInsertion benchmark [38] (Table IX), reporting PSNR, SSIM, LPIPS, FID, and Face-Score (via AdaFace [73]). The results confirm that performance drops in identity-sensitive and try-on tasks. These issues stem primarily from the data generation process: (1) The base model (e.g., FLUX) exhibits bias towards Western facial features, affecting identity preservation; (2) Excessive background preservation can lead to duplicated body parts; and (3) The scarcity of domain-specific training data limits generalization in try-on tasks. Notably, these limitations are largely tied to the base model rather than our framework. To demonstrate this, we compared identity preservation across different backbones in Table XI. A strong correlation is observed: upgrading to a more capable model like Qwen-Image [68] significantly boosts the Face-Score to 50.25. This indicates that our pipeline is flexible and will naturally benefit from future advancements in foundation models.

TABLE XI
QUANTITATIVE COMPARISON OF IDENTITY PRESERVATION ACROSS DIFFERENT BASE MODELS

	SD3 [32]	Flux [50]	Qwen-Image [68]
Face-Score [73]	43.03	47.89	50.25

2) Base Model: In our experiments, we observe that the final fused images largely inherit the style of the base model. As shown in Fig. 18, training with SD3 [32] and Flux [50] results in noticeable artifacts, such as facial overexposure and unnatural waxy textures, which are consistent with the known limitations of these models. Conversely, employing Qwen-Image [68] mitigates these overexposure issues, yielding significantly more realistic details (e.g., hair texture). These findings demonstrate the extensibility of our method: it can be seamlessly applied

Fig. 18. Performance of different base models on image fusion after training.

to more powerful foundation models, leveraging their advanced generative capabilities to achieve superior image fusion.

VI. Conclusion

In this paper, we first propose an comprehensive data generation pipeline to create fusion scenarios that are relatively rare in traditional fusion tasks, making the fusion process more natural and flexible. Using this pipeline, we generated a dataset of 160 k fusion data that encompass various scenarios, including placement, handheld, wearable, style transfer, replacement and attribute-referenced editing tasks. Additionally, we introduce DreamFuse, an adaptive image fusion framework with the diffusion transformer. This framework incorporates positional affine transformations to encode the position and size of foreground, employs shared attention mechanisms to establish connections between the foreground and background, and ultimately leverages localized direct preference optimization to further enhance the quality of the fused images. Experimental results show that our method outperforms existing approaches on multiple benchmarks.

References

[1] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22500–22510.

[2] W. Chen et al., "Subject-driven text-to-image generation via apprenticeship learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 30286–30305.

[3] Z. Tan, S. Liu, X. Yang, Q. Xue, and X. Wang, "Ominicontrol: Minimal and universal control for diffusion transformer," 2024, *arXiv:2411.15098*.

[4] W. Zhong et al., "Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter," 2025, *arXiv:2505.18612*.

[5] B. Yang et al., "Paint by example: Exemplar-based image editing with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18381–18391.

[6] L. Zhang, A. Rao, and M. Agrawala, "Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport," in *Proc. Int. Conf. Learn. Representations*, 2025.

[7] G. C. Tarrés et al., "Thinking outside the BBox: Unconstrained generative object compositing," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 476–495.

[8] K. Wang, M. Gharbi, H. Zhang, Z. Xia, and E. Shechtman, "Semi-supervised parametric real-world image harmonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5927–5936.

[9] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao, "Anydoor: Zero-shot object-level image customization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6593–6602.

[10] D. Winter et al., "Objectmate: A recurrence prior for object insertion and subject-driven generation," 2024, *arXiv:2412.08645*.

[11] J. He, W. Li, Y. Liu, J. Kim, D. Wei, and H. Pfister, "Affordance-aware object insertion via mask-aware dual diffusion," 2024, *arXiv:2412.14462*.

[12] A. Kirillov et al., "Segment anything," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[13] J. Miao et al., "Large-scale video panoptic segmentation in the wild: A benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21033–21043.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[15] R. Suvorov et al., "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2149–2159.

[16] P.-D. Tudosiu et al., "Mulan: A multi layer annotated dataset for controllable text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22413–22422.

[17] S. Lu, Y. Liu, and A. W.-K. Kong, "TF-ICON: Diffusion-based training-free cross-domain image composition," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 2294–2305.

[18] Y. Wang, W. Zhang, J. Zheng, and C. Jin, "PrimeComposer: Faster progressively combined diffusion for image composition with attention steering," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 10824–10832.

[19] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "SVDiff: Compact parameter space for diffusion fine-tuning," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 7323–7334.

[20] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *Proc. Int. Conf. Learn. Representations*, 2023.

[21] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1931–1941.

[22] J. Zhuang, D. Kang, Y.-P. Cao, G. Li, L. Lin, and Y. Shan, "Tip-editor: An accurate 3D editor following both text-prompts and image-prompts," *ACM Trans. Graph.*, vol. 43, no. 4, pp. 1–12, 2024.

[23] D. Garibi et al., "Tokenverse: Versatile multi-concept personalization in token modulation space," *ACM Trans. Graph.*, vol. 4, pp. 1–11, 2025.

[24] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023, *arXiv:2308.06721*.

[25] Y. Zhang et al., "SSR-encoder: Encoding selective subject representation for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8069–8078.

[26] Q. Wang et al., "InstantID: Zero-shot identity-preserving generation in seconds," 2024, *arXiv:2401.07519*.

[27] Z. He, P. Chen, G. Wang, G. Li, P. H. Torr, and L. Lin, "WildVidFit: Video virtual try-on in the wild via image-based controlled diffusion models," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 123–139.

[28] L. Huang et al., "In-context lora for diffusion transformers," 2024, *arXiv:2410.23775*.

[29] B. Chen et al., "XVerse: Consistent multi-subject control of identity and semantic attributes via DiT modulation," 2025, *arXiv:2506.21416*.

[30] J. Peng, Z. Luo, L. Liu, and B. Zhang, "Frih: Fine-grained region-aware image harmonization," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 4478–4486.

[31] J. Zhou et al., "Foreground harmonization and shadow generation for composite image," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 8267–8276.

[32] Q. Meng, Q. Liu, Z. Li, X. Lan, S. Zhang, and L. Nie, "High-resolution image harmonization with adaptive-interval color transformation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 13769–13793.

[33] W. Tao, X. Yang, M. Cui, and G. Lin, "MotionCom: Automatic and motion-aware image composition with LLM and video diffusion prior," 2024, *arXiv:2409.10090*.

[34] D. Winter, M. Cohen, S. Fruchter, Y. Pritch, A. Rav-Acha, and Y. Hoshen, "ObjectDrop: Bootstrapping counterfactuals for photorealistic object removal and insertion," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 112–129.

[35] Z. He et al., "Vton 360: High-fidelity virtual try-on from any viewing direction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 26388–26398.

[36] Z. He et al., "Efficient and robust video virtual try-on via enhanced multi-garment alignment," *IEEE Trans. Multimedia*, vol. 27, pp. 9009–9021, 2025, doi: 10.1109/TMM.2025.3613169.

[37] C. Mao et al., "Ace : Instruction-based image creation and editing via context-aware content filling," 2025, *arXiv:2501.02487*.

[38] W. Song, H. Jiang, Z. Yang, R. Quan, and Y. Yang, "Insert anything: Image insertion via in-context editing in dit," 2025, *arXiv:2504.15009*.

[39] C. Wu et al., "OmniGen2: Exploration to advanced multimodal generation," 2025, *arXiv:2506.18871*.

[40] J. Xu et al., "ImageReward: Learning and evaluating human preferences for text-to-image generation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 15903–15935.

[41] J. Xu et al., "VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation," 2024, *arXiv:2412.21059*.
[42] J. Zhang et al., "UniFL: Improve stable diffusion via unified feedback learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2025, pp. 67355–67382.
[43] J. Zhang et al., "Treereward: Improve diffusion model via tree-structured feedback learning," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 4533–4542.
[44] J. Liu et al., "Improving video generation with human feedback," 2025, *arXiv:2501.13918*.
[45] B. Wallace et al., "Diffusion model alignment using direct preference optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8228–8238.
[46] S. Li, K. Kallidromitis, A. Gokul, Y. Kato, and K. Kozuka, "Aligning diffusion models by optimizing human utility," 2024, *arXiv:2404.04465*.
[47] T. Zhang et al., "Diffusion model as a noise-aware latent reward model for step-level preference optimization," 2025, *arXiv:2502.01051*.
[48] D. Guo et al., "Seed1. 5-VL technical report," 2025, *arXiv:2505.07062*.
[49] X. Tian, W. Li, B. Xu, Y. Yuan, Y. Wang, and H. Shen, "Mige: A unified framework for multimodal instruction-based image generation and editing," 2025, *arXiv:2502.21291*.
[50] B. F. Labs, "Flux: Inference repository," 2024, Accessed: Oct25, 2024. [Online]. Available: https://github.com/black-forest-labs/flux
[51] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, 2024, Art. no. 127063.
[52] J. Huang et al., "DreamLayer: Simultaneous multi-layer generation via diffusion mode," 2025, *arXiv:2503.12838*.
[53] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.
[54] Y. Shi, P. Wang, and W. Huang, "Seededit: Align image re-generation to image editing," 2024, *arXiv:2411.06686*.
[55] S. Batifol et al., "FLUX. 1 kontext: Flow matching for in-context image generation and editing in latent space," 2025, *arXiv:2506.15742*.
[56] M. Ku, D. Jiang, C. Wei, X. Yue, and W. Chen, "VIEScore: Towards explainable metrics for conditional image synthesis evaluation," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 12268–12290.
[57] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[58] C. Schuhmann et al., "Laion-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 25278–25294.
[59] S. Serengil and A. Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *J. Inf. Technol.*, vol. 17, no. 2, pp. 95–107, 2024. [Online]. Available: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077
[60] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 19790–19800.
[61] N. Ravi et al., "SAM 2: Segment anything in images and videos," 2024, *arXiv:2408.00714*.
[62] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," 2022, *arXiv:2210.02747*.
[63] K. Mishchenko and A. Defazio, "Prodigy: An expeditiously adaptive parameter-free learner," 2023, *arXiv:2306.06101*.
[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[65] B. Zhang et al., "Controlcom: Controllable image composition using diffusion model," 2023, *arXiv:2308.10040*.
[66] X. Chen et al., "Zero-shot image editing with reference imitation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2025, pp. 84010–84032.
[67] P. Esser et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 12606–12633.
[68] C. Wu et al., "Qwen-image technical report," 2025, *arXiv:2508.02324*.
[69] Y. Yu, Z. Zeng, H. Zheng, and J. Luo, "OmniPaint: Mastering object-oriented editing via disentangled insertion-removal inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 17324–17334.
[70] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Eccv*, Springer, 2014, pp. 740–755.
[71] W. Hong et al., "CogVLM2: Visual language models for image and video understanding," 2024, *arXiv:2408.16500*.
[72] S. Sidney, "Nonparametric statistics for the behavioral sciences," *J. Nervous Ment. Dis.*, vol. 125, no. 3, 1957, Art. no. 497.
[73] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18750–18759.

Junjia Huang received the BEng degree from Sun yat-sen University, in 2022. He is currently working toward the PhD degree with Sun yat-sen University. His current research interest including computer vision and image processing.

Pengxiang Yan received the BE degree in software engineering from Sun Yat-sen University, in 2018 and the MS degree in computer science from Sun Yat-sen University, in 2021. He is currently a computer vision researcher with ByteDance Intelligent Creation Lab. His research interests mainly cover deep learning, computer vision, dense image understanding, aigc-based image editing.

Jiyang Liu received the BS and MS degrees in department of automation from Tsinghua University, in 2017 and 2020, respectively. He is currently a computer vision researcher with ByteDance Intelligent Creation Lab. His research interests mainly cover deep learning, computer vision, dense image understanding, and aigc-based image editing.

Jie Wu is an algorithm researcher with ByteDance, mainly interested in RLHF, diffusion and Visual Generation. He has authored and co-authored approximately 30 papers published in top-tier academic journals and conferences.

Zhao Wang received the bachelor's degree from Nanjing University and the master's degree from Shanghai Jiaotong university. He is a senior engineer with Bytedance Inc. His research interests include computer vision and deep learning, particularly in image generative models.

Yitong Wang received the BS and MS degrees in computer science from Peking University, Beijing, China, in 2013 and 2016, respectively. From 2017 to 2019, he was formally a researcher with Tencent AI Laboratory, Shenzhen, China. He is currently with ByteDance Inc. His research interests include image processing, computer vision, and machine learning.

Xinglong Wu received the MS degree from the Department of Mathematics, Jilin University. He is currently the Leader of Creation CV team with ByteDance. His main research areas include multimodality, segmentation, intelligent editing, AIGC, etc.

Liang Lin (Fellow, IEEE) is a full professor of computer science with Sun Yat-sen University. He served as the executive director and Distinguished Scientist of SenseTime Group from 2016 to 2018, leading the R&D teams for cutting-edge technology transferring. He has authored or co-authored more than 300 papers in leading academic journals and conferences, and his papers have been cited by more than 34,000 times. He is an associate editor of *IEEE Transactions On Neural Networks and Learning Systems* and *IEEE Transactions on Multimedia*, and served as area chairs for numerous conferences such as CVPR, ICCV, SIGKDD and AAAI. He is the recipient of numerous awards and honors including Wu Wen-Jun Artificial Intelligence Award, the First Prize of China Society of Image and Graphics, ICCV Best Paper Nomination, in 2019, Annual Best Paper Award by Pattern Recognition (Elsevier), in 2018, Best Paper Dimond Award in IEEE ICME 2017, Google Faculty Award, in 2012. His supervised PhD students received ACM China Doctoral Dissertation Award, CCF Best Doctoral Dissertation and CAAI Best Doctoral Dissertation. He is a Fellow of IAPR.

Guanbin Li (Member, IEEE) received the PhD degree from the University of Hong Kong, in 2016. He is currently a full professor in School of Computer Science and Engineering, Sun Yat-sen University. His current research interests include computer vision, image processing, and deep learning. He is a recipient of the ICCV 2019 Best Paper Nomination Award and CVPR2024 best paper candidate. He has authorized and co-authorized on more than 200 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of CVPR and ICCV.