CareCom: Generative Image Composition with Calibrated Reference Features

Jiaxuan Chen[1], Bo Zhang[1], Qingdong He[2], Jinlong Peng[2], Li Niu[1,3]*

[1]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2]Youtu Lab, Tencent
[3]miguo.ai
{chenjiaxuan, bo-zhang, ustcnewly}@sjtu.edu.cn, {yingcaihe, jeromepeng}@tencent.com

Abstract

Image composition aims to seamlessly insert foreground object into background. Despite the huge progress in generative image composition, the existing methods are still struggling with simultaneous detail preservation and foreground pose/view adjustment. To address this issue, we extend the existing generative composition model to multi-reference version, which allows using arbitrary number of foreground reference images. Furthermore, we propose to calibrate the global and local features of foreground reference images to make them compatible with the background information. The calibrated reference features can supplement the original reference features with useful global and local information of proper pose/view. Extensive experiments on MVImgNet and MureCom demonstrate that the generative model can greatly benefit from the calibrated reference features.

1 Introduction

Image composition is an important image editing operation, aiming to seamlessly insert a given foreground object into a background image. Previous methods (Tsai et al. 2017; Zhang, Wen, and Shi 2020; Hong, Niu, and Zhang 2022) attempted to address different issues in image composition with different sub-tasks (*e.g.*, image blending, image harmonization, and shadow generation). Recently, foundation diffusion models (Rombach et al. 2022; Esser et al. 2024; Labs 2024) have demonstrated powerful image generation ability, and some works have utilized such ability to re-generate the foreground in the background image with all the issues solved simultaneously.

These generative composition methods can be roughly classified into training-free methods and training-based methods. The training-free methods (Lu, Liu, and Kong 2023; Wang et al. 2024; Xu et al. 2025) leverages the prior knowledge in foundation model without the need of training or finetuning. However, they cannot adjust the pose/view of foreground and the generated images are of low quality. In contrast, training-based methods (Chen et al. 2024; Yang et al. 2023; Lu et al. 2023; Kulal et al. 2023; Zhang et al. 2023; Winter et al. 2025, 2024; Canet Tarrés et al. 2025; Yuan et al. 2024; Song et al. 2024; Chen et al. 2025) are more popular and more powerful. They require large-scale training set to train the model or a few images containing the target object to finetune the model. The usual approach (Song et al. 2023; Yang et al. 2023) is extracting foreground information and injecting into denoising UNet of stable diffusion (Rombach et al. 2022). Some subsequent works attempted to better preserve the foreground details by using different strategies, like high-frequency information (Chen et al. 2024) or local enhancement module (Zhang et al. 2023).

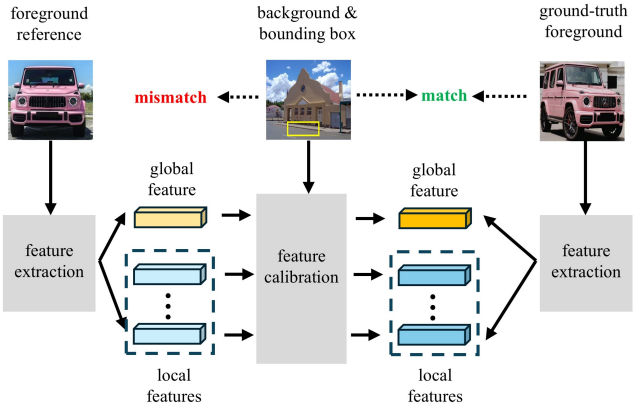

Figure 1: Illustration of feature calibration process. Based on the background and foreground bounding box, we calibrate the features of foreground reference images to match those of ground-truth foreground.

Despite the remarkable progress achieved for generative composition, based on our experimental observation, there is no method adept at both detail preservation and pose/view adjustment of foreground. One reason is that the existing methods only use one reference image for the foreground object, which raises the difficulty of generating the details of dramatically different pose/view. To explore the advantage of using multiple reference images, we extend the existing generative composition model (Song et al. 2023) to support arbitrary number of reference images, by simply concatenating their reference features. Based on the multi-reference generative composition model, we observe that using multiple reference images for the foreground object can greatly

*Corresponding Author.

improve the results, because the model can attend to the reference image with the pose/view relatively matching the background, which can greatly alleviate the task difficulty. Ideally, if the model is provided with abundant reference images covering all views and poses, the model can attend to the one perfectly matching the background. However, it is very costly and even impossible to collect a large number of reference images covering the full range of poses/views.

Another thought is to hallucinate the reference image matching the background based on the given reference images. Instead of hallucinating the compatible reference image, we opt for hallucinating the compatible reference feature, *i.e.*, the feature of compatible reference image, which is more efficient by omitting the transformation process between feature space and image space. Specifically, we design a feature calibration module. This module takes in one reference feature and produces its calibrated reference feature, which is calibrated towards the compatible reference feature based on the background information. We refer to the calibrated reference features as augmented reference features, which are appended to the original reference features and jointly sent to denoising UNet. Among the original reference features, for those relatively compatible with the background, their calibrated versions are expected to be more compatible with the background and the model is expected to attend to these calibrated reference features.

When calibrating the reference features, we consider both global reference features and local reference features, by using global calibration module and local calibration module respectively. The global calibration module relies on background information and current denoised foreground to calibrate the global reference feature, which should be close to the global feature of ground-truth foreground. Similarly, the local calibration module calibrates the local reference feature, which should be close to the corresponding local feature of ground-truth foreground. We name our image **com**position method with **ca**librated **re**ference features as CareCom.

We conduct experiments on MVImgNet (Yu et al. 2023) and MureCom (Lu et al. 2023) datasets. They contain multiple reference images for each foreground object, which is suitable for our task. The results demonstrate that our CareCom excels in detail preservation and pose/view adjustment of foreground at the same time. Our contributions can be summarized as follows: 1) We propose the first multi-reference generative composition model supporting an arbitrary number of foreground reference images. 2) We propose to calibrate the foreground reference features to match the background. Technically, we design a global (*resp.*, local) calibration module to calibrate the global (*resp.*, local) reference features. 3) Comprehensive experiments on two datasets show that our method outperforms other baselines in terms of faithfulness and realism.

2 Related Work

2.1 Image Composition

The goal of image composition is inserting the foreground from one image into another background image. To address the inconsistency issues between foreground and background, previous methods work on different sub-tasks to solve different issues. In particular, image harmonization (Guerreiro, Nakazawa, and Stenger 2023; Tsai et al. 2017; Jiang et al. 2021; Ke et al. 2022; Xue et al. 2022; Tao et al. 2024) focuses on adjusting foreground illumination to be harmonious with the background. Some other methods (Hong, Niu, and Zhang 2022; Liu et al. 2024; Sheng et al. 2022) consider the effect of foreground on the background, like shadow and reflection. Object placement (Zhou et al. 2022; Zhu et al. 2023; Qin et al. 2025) targets at seeking for reasonable location, scale, and perspective for the foreground object.

In recent years, generative composition has emerged for object insertion with one unified model, thanks to the unprecedented potential of foundation generative model like stable diffusion (Rombach et al. 2022) and Diffusion transformer (Esser et al. 2024; Labs 2024). These methods can be categorized into training-based (Yang et al. 2023; Lu et al. 2023; Kulal et al. 2023; Winter et al. 2024; Song et al. 2024; Chen et al. 2024; Canet Tarrés et al. 2025; Tarrés et al. 2025; Huang et al. 2025; Yu et al. 2025) and training-free approaches (Lu, Liu, and Kong 2023; Wang et al. 2024; Pham, Chen, and Chen 2024; Li et al. 2024; Xu et al. 2025). Our method belongs to the training-based group, which has much stronger ability in adjusting the pose and view of foreground object. Among the existing training-based methods, some methods attempt to promote the foreground details (Chen et al. 2024; Zhang et al. 2023; Song et al. 2023) or impose additional controls (Zhang et al. 2023). Some methods (Winter et al. 2024; Canet Tarrés et al. 2025) place emphasis on shadow and reflection generation. Some more recent works (Song et al. 2025; Wang et al. 2025) explore in-context learning or multi-condition control based on DiT architecture. However, the above methods only support one reference image, which limits the performance upper bound when multiple reference images are available.

2.2 Subject-driven Image Generation and Editing

Subject-driven image generation refers to a variety of tasks of generating or editing images in terms of specific object.

With the emergence of diffusion models, many works (Kumari et al. 2023; Ruiz et al. 2024, 2023; Gal et al. 2023a) have explored text-based image customization. They propose to learn new concepts by associating specific object with special text token. However, these methods are only applicable to text-generated background instead of a given background image. Some other works (Wei et al. 2023; Gal et al. 2023b; Tao et al. 2025) suggest using specific encoders to extract visual information and integrating it into customized images. However, they cannot control the placement of the foreground. Some approaches (Gu et al. 2023; Mokady et al. 2023; Choi et al. 2023; Yang et al. 2024; Li et al. 2025; Zhang et al. 2025) use text or image as guidance for image editing. Despite the diversity of subject-driven tasks, our method primarily focuses on generative composition, re-generating compatible foreground object at the designated location in the background image.

3 Our Method

In Section 3.1, we will introduce our multi-reference generative composition framework. In Section 3.2 and 3.3, we will elaborate on our global and local reference feature calibration modules respectively. In Section 3.4, we will introduce the training strategy.

3.1 Multi-reference Composition Framework

In this work, we propose a multi-reference generative composition framework, aiming to insert the foreground object at the specified location in the background image based on several reference images of the foreground object. The inserted foreground is expected to maintain the detail information and match the background *w.r.t.* illumination, pose/view, and so on. Formally, given a complete background image with a bounding box $\boldsymbol{b}$ to place the foreground object and a set of foreground reference images set $\{\boldsymbol{I}_1^f, \boldsymbol{I}_2^f, ..., \boldsymbol{I}_K^f\}$, the generated image should resemble the ground-truth image $\hat{\boldsymbol{I}}$.

Our model is built upon ObjectStitch (Song et al. 2023) considering its compelling ability to adjust foreground pose/view and generate realistic images. We get the foreground mask $\boldsymbol{M}^f$ based on the bounding box $\boldsymbol{b}$ and get the masked background $\boldsymbol{I}^b$ by erasing the content within $\boldsymbol{b}$. Following (Song et al. 2023), we first project $\boldsymbol{I}^b$ and $\hat{\boldsymbol{I}}$ into latent space with VAE encoder, yielding $\boldsymbol{z}^b$ and $\boldsymbol{z}$ respectively. Then, we concatenate $\boldsymbol{z}^b$, noisy $\boldsymbol{z}_t$, and $\boldsymbol{M}^f$ as the input of denoising UNet. For the k-th foreground reference image, we extract its global reference features $\boldsymbol{f}_k^g$ and local reference features $\{\boldsymbol{f}_{k,i}^l|_{i=1}^N\}$ using E^f which includes the pretrained CLIP encoder (Radford et al. 2021) and an adapter. Global features are the CLS token output with dimension 1×1024 from CLIP encoder, while local features are the remaining output tokens with dimension 256×1024 from the last layer.

To obtain the reference features compatible with the background, we design a global calibration module C^g to produce the calibrated global reference features $\tilde{\boldsymbol{f}}_k^g$. Similarly, we also design a local calibration module C^l to produce the calibrated local reference features $\{\tilde{\boldsymbol{f}}_{k,i}^l|_{i=1}^N\}$. We denote the set of global (*resp.*, local) reference features from all reference images as $\mathcal{F}^g$ (*resp.*, $\mathcal{F}^l$). Besides, we refer to the calibrated reference features as augmented reference features, and denote the set of global (*resp.*, local) augmented reference features from all reference images as $\tilde{\mathcal{F}}^g$ (*resp.*, $\tilde{\mathcal{F}}^l$).

The original reference features $\{\mathcal{F}^g, \mathcal{F}^l\}$ are injected into both encoder and decoder of denoising UNet as in (Song et al. 2023). The augmented reference features $\{\tilde{\mathcal{F}}^g, \tilde{\mathcal{F}}^l\}$ are only injected into the decoder, because the generation of augmented reference features relies on the encoder features of denoising UNet and injecting augmented reference features into the encoder would cause dependency loop.

In the training stage, given a set of training images containing the specific foreground object $\{\boldsymbol{I}_1^o, \boldsymbol{I}_2^o, ..., \boldsymbol{I}_K^o\}$, the foreground images are cropped from training images, followed by geometry and color perturbation. The perturbed foreground images form the set of foreground reference images $\{\boldsymbol{I}_1^f, \boldsymbol{I}_2^f, ..., \boldsymbol{I}_K^f\}$. Then, we take one training image $\boldsymbol{I}_k^o$ as the ground-truth image $\hat{\boldsymbol{I}}$ and prepare the corresponding masked background $\boldsymbol{I}^b$, foreground mask $\boldsymbol{M}^f$.

The latent $\boldsymbol{z}$ of $\hat{\boldsymbol{I}}$ is added with t-step noise, leading to $\boldsymbol{z}_t$. The denoising UNet with parameters θ is trained using the following objective to predict the added noise:

$$\mathcal{L}^{sd} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}\|\epsilon - \epsilon_\theta\left(\boldsymbol{z}^b, \boldsymbol{z}_t, \boldsymbol{M}^f, \{\mathcal{F}^g, \mathcal{F}^l\}, t\right)\|_2^2, \tag{1}$$

where ϵ is the added Gaussian noise and t is the time step ranging from 0 to T.

In the testing stage, given a background image with bounding box, we can obtain $\boldsymbol{z}^b, \boldsymbol{z}_T, \boldsymbol{M}^f$. We can obtain $\mathcal{F}^g, \mathcal{F}^l$ from foreground reference images. Then, we pass through the denoising process to get the denoised latent $\boldsymbol{z}_0$, which is mapped back to image space with VAE decoder.

3.2 Global Reference Feature Calibration

For each global reference feature $\boldsymbol{f}_k^g$ (the global feature of foreground reference image $\boldsymbol{I}_k^f$), we use global reference feature calibration (GRFC) module to calibrate $\boldsymbol{f}_k^g$ to match the ground-truth global feature $\hat{\boldsymbol{f}}^g$ (the global feature of ground-truth foreground). Specifically, we crop the foreground $\hat{\boldsymbol{I}}^f$ from ground-truth image $\hat{\boldsymbol{I}}$ and use foreground encoder to extract its global feature $\hat{\boldsymbol{f}}^g$.

In the GRFC module, we employ the encoder features in denoising UNet to facilitate the calibration process. Since denoising UNet takes the masked background and denoised latent as input, its encoder features should contain the rich information of background and denoised foreground. The background information is helpful to make the calibrated global reference feature compatible with the background. The current denoised foreground may also provide useful hints for the calibration.

Formally, we use each $\boldsymbol{f}_k^g$ as query, while the encoder features $\boldsymbol{F}^{en}$ in denoising UNet are used as keys and values. We pass them through a cross-attention layer to produce the calibrated global reference features $\tilde{\boldsymbol{f}}_k^g$, which can be formulated as

$$\tilde{\boldsymbol{f}}_k^g = \text{Softmax}\left(\frac{\boldsymbol{f}_k^g(\boldsymbol{F}^{en}\boldsymbol{W}^{gk})^T}{\sqrt{d}}\right)(\boldsymbol{F}^{en}\boldsymbol{W}^{gv}) + \boldsymbol{f}_k^g, \tag{2}$$

where $\boldsymbol{W}^{gk}, \boldsymbol{W}^{gv}$ are projection matrices and d is the dimension of query feature.

The calibrated global reference feature $\tilde{\boldsymbol{f}}_k^g$ is forced to match the ground-truth global feature $\hat{\boldsymbol{f}}^g$ using the following loss:

$$\mathcal{L}^{gc} = \sum_{k=1}^{K} ||\tilde{\boldsymbol{f}}_k^g - \hat{\boldsymbol{f}}^g||^2. \tag{3}$$

The calibrated global reference features $\tilde{\boldsymbol{f}}_k^g$ form the augmented global reference feature set $\tilde{\mathcal{F}}^g$, which is injected to the decoder of denoising UNet via cross-attention.

3.3 Local Reference Feature Calibration

Previous methods (Chen et al. 2024; Zhang et al. 2023) have demonstrated that local features of foreground object play an important role in keeping the object details, so we also

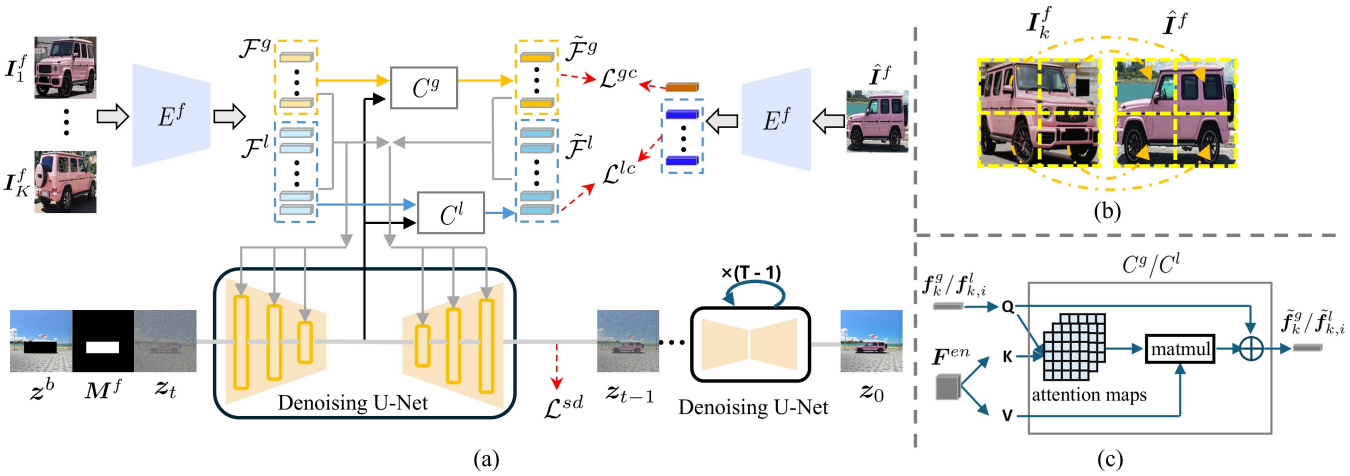

Figure 2: (a) Given multiple foreground reference images, we extract their global/local features $\mathcal{F}^g/\mathcal{F}^l$, which are passed through the calibration module C^g/C^l. The calibrated features $\tilde{\mathcal{F}}^g/\tilde{\mathcal{F}}^l$ are injected into the decoder of denoising UNet. (b) Illustration of seeking for the spatial correspondence of local patches between foreground reference $\boldsymbol{I}_k^f$ and ground-truth foreground $\hat{\boldsymbol{I}}^f$. (c) The structure of calibration module C^g/C^l.

utilize the local features of foreground reference images. For each local reference feature $\boldsymbol{f}_{k,i}^l$ (the i-th local feature of foreground reference image $\boldsymbol{I}_k^f$), we adopt local reference feature calibration (LRFC) module to calibrate $\boldsymbol{f}_{k,i}^l$ to match its corresponding ground-truth local feature.

Different from global reference feature calibration, how to obtain the ground-truth local feature is not straightforward. Considering that the local feature represents the information of the corresponding local patch, we seek for the spatial correspondence between the patches in foreground reference image and the patches in ground-truth foreground image. Given the ground-truth foreground image $\hat{\boldsymbol{I}}^f$, we extract its local features $\{\hat{\boldsymbol{f}}_{k,i}^l|_{i=1}^N\}$. Given the local reference features $\{\boldsymbol{f}_{k,i}^l|_{i=1}^N\}$ of the k-th reference image, for each local reference feature $\boldsymbol{f}_{k,i}^l$, we find its nearest feature in $\{\hat{\boldsymbol{f}}_{k,i}^l|_{i=1}^N\}$ as its ground-truth local feature. Specifically, we calculate the similarity between $\{\hat{\boldsymbol{f}}_{k,i}^l|_{i=1}^N\}$ and $\{\boldsymbol{f}_{k,i}^l|_{i=1}^N\}$. Based on the $N \times N$ similarity matrix, we can associate the i-th patch in the foreground reference image with the most similar $\delta(i)$-th patch in the ground-truth foreground image.

We employ a calibration module C^l to produce the calibrated local reference feature $\tilde{\boldsymbol{f}}_{k,i}^l$, which should be close to $\hat{\boldsymbol{f}}_{k,\delta(i)}^l$. To facilitate the calibration process, similar to Section 3.2, we use the encoder features $\boldsymbol{F}^{en}$ in the denoising UNet to provide auxiliary information, because the information of background and denoised foreground could help determine how the local patch should be warped or transformed. Formally, we use each local reference feature $\boldsymbol{f}_{k,i}^l$ as query, and the encoder features $\boldsymbol{F}^{en}$ in the denoising UNet as keys and values. We pass them through a cross-attention layer to produce the calibrated local reference feature $\tilde{\boldsymbol{f}}_{k,i}^l$, which can be formulated as

$$\tilde{\boldsymbol{f}}_{k,i}^l = \text{Softmax}\left(\frac{\boldsymbol{f}_{k,i}^l(\boldsymbol{F}^{en}\boldsymbol{W}^{lk})^T}{\sqrt{d}}\right)(\boldsymbol{F}^{en}\boldsymbol{W}^{lv}) + \boldsymbol{f}_{k,i}^l, \tag{4}$$

where $\boldsymbol{W}^{lk}, \boldsymbol{W}^{lv}$ are projection matrices and d is the dimension of query feature.

The calibrated local reference features are supervised by

$$\mathcal{L}^{lc} = \sum_{k=1}^{K}\sum_{i=1}^{N} ||\tilde{\boldsymbol{f}}_{k,i}^l - \hat{\boldsymbol{f}}_{k,\delta(i)}^l||^2. \tag{5}$$

The calibrated local reference features $\tilde{\boldsymbol{f}}_{k,i}^l$ form the augmented local reference feature set $\tilde{\mathcal{F}}^l$, which is injected to the decoder of denoising UNet via cross-attention.

3.4 Training Strategy

Our model requires pretraining and few-shot finetuning. 1) We first pretrain our designed GRFC and LRFC using a large-scale training set (*e.g.*, MVImageNet (Yu et al. 2023)) which provides multiple images for each object. The other modules including denoising UNet and VAE are borrowed from the pretrained ObjectStitch (Song et al. 2023) model. 2) After pretraining, given a few training images containing a specific foreground object, we finetune the whole model based on these training images. Given test background images with bounding boxes, we can apply the finetuned model to insert the specific foreground object into background.

4 Experiments

4.1 Datasets

Since multiple reference images are needed, we conduct experiments on two datasets which have multiple reference images for each foreground object.

MureCom (Lu et al. 2023) contains 32 foreground categories. Each foreground category has 3 objects and 20 background images. Each object has 5 images with different poses and viewpoints. Each background image has a bounding box to specify where the foreground should be inserted. For each object, we use its 5 images for few-shot finetuning and 20 background images belonging to its category for evaluation.

MVImgNet (Yu et al. 2023) contains 222,929 objects from 238 foreground categories. Each object has a set of images captured from different camera viewpoints. We select one object from each category to form the test objects. The images of the remaining objects are used to pre-train GRFC and LRFC as described in Section 3.4. Each test object is associated with 5 images, in which 4 images are used for few-shot finetuning and the last image is used for evaluation.

4.2 Evaluation Metrics

We use DINO Score (Caron et al. 2021) to assess the fidelity of generated foreground. Since MVImgNet has ground-truth foreground while MureCom does not, DINO score is calculated based on generated foreground and ground-truth foreground (*resp.*, its nearest foreground reference) on MVImgNet (*resp.*, MureCom). For the background, we use SSIM (Wang et al. 2004) to evaluate background preservation. We choose FOSScore (Zhang, Sui, and Niu 2023) to evaluate the pose/view compatibility between foreground and background. Quality Score (QS) (Wang et al. 2004) is used to evaluate the overall quality of generated images.

4.3 Implementation Details

Our training process consists of two stages: pretraining and fine-tuning. In the first stage, the model is pretrained on MVImgNet dataset. The number of training epochs is set to 50, with a batch size 64. This stage is conducted on 16 V100 GPUs. In the second stage, the model is finetuned using up to five images containing specific foreground objects. Fine-tuning stage is conducted on 1 A6000 GPU that takes about ten minutes for 150 epochs.

4.4 Baselines

We compare our CareCom with recent and open-sourced generative composition methods including ObjectStitch (Song et al. 2023), Anydoor (Chen et al. 2024), ControlCom (Zhang et al. 2023), Unicombine (Wang et al. 2025) and Insert Anything (Song et al. 2025). Among the baselines, (Song et al. 2023; Zhang et al. 2023) support multiple reference images. For these methods, the extension to support multiple reference images is similar to our method: the features of multiple reference images are concatenated along the sequence dimension and fed into the denoising UNet. (Chen et al. 2024; Song et al. 2025; Wang et al. 2025) only supports single reference image. For fair comparison, we perform few-shot finetuning for all baselines.

During inference, for the baselines supporting multiple reference images, we use all reference images. For the baselines only supporting single reference image, we feed the reference images one by one and get multiple results, from which the best result is selected.

Method	Metrics			
	$\text{DINO}_{\text{fg}}\uparrow$	$\text{SSIM}_{\text{bg}}\uparrow$	FOSScore↑	QS↑
ControlCom (2023)	64.92	0.858	0.856	42.27
Anydoor (2024)	68.65	0.857	0.815	43.40
ObjectStitch (2023)	65.04	0.854	0.867	44.39
Insert Anything (2025)	**68.78**	0.854	0.822	45.40
UniCombine (2025)	65.72	0.856	0.819	44.87
CareCom	68.60	**0.859**	**0.883**	**47.07**

Table 1: Quantitative comparison on MureCom dataset. The best results are highlighted in boldface.

Method	Metrics			
	$\text{DINO}_{\text{fg}}\uparrow$	$\text{SSIM}_{\text{bg}}\uparrow$	FOSScore↑	QS↑
ControlCom (2023)	63.93	0.857	0.824	37.72
Anydoor (2024)	69.56	0.853	0.791	40.85
ObjectStitch (2023)	66.21	0.857	0.841	42.73
Insert Anything (2025)	**69.88**	0.857	0.804	41.22
UniCombine (2025)	64.22	0.853	0.812	41.86
CareCom	69.49	**0.858**	**0.874**	**45.79**

Table 2: Quantitative comparison on MVImgNet dataset. The best results are highlighted in boldface.

4.5 Quantitative Comparison

We evaluate different approaches on MureCom and MVImgNet datasets. The results are reported in Table 1 and Table 2 respectively. The quantitative metrics details have been introduced in Section 4.2.

Anydoor and Insert Anything achieve high DINO scores on both datasets, which shows their ability to preserve foreground details. However, they exhibit obvious copy-and-paste effect and lack the ability to adjust pose/view according to background, indicated by lower FOSScore. In contrast, although our method performs slightly worse than them on DINO score, it outperforms all the methods for all other metrics.

4.6 Visual Comparison

In Fig. 3, we provide visual comparison results of different methods on MureCom dataset. The baselines (Zhang et al. 2023; Song et al. 2023) and our method use five reference images. Since AnyDoor, Insert Anything and UniCombine only support single reference image, we use each of the five reference images separately as input and select the visually best result.

It can be observed that the images generated by (Chen et al. 2024; Song et al. 2025) exhibit noticeable copy-paste artifacts from the reference images, resulting in incompatible lighting and perspective between foreground and background. ControlCom (Zhang et al. 2023) performs well in composite images with simple foreground objects but tends to produce some artifacts when the foreground objects are complex. ObjectStitch (Song et al. 2023) retains most of the foreground information, but still misses or alters some

Figure 3: Visual comparison of different methods on MureCom dataset. From left to right, we show background, 5 reference images, the results of Anydoor (Chen et al. 2024), ControlCom (Zhang et al. 2023), ObjectStitch (Song et al. 2023), Insert Anything (Song et al. 2025), UniCombine (Wang et al. 2025) and our CareCom.

details (*e.g.*, the pattern on the guitar in row 1, unrealistic horse legs in row 2). UniCombine (Wang et al. 2025) fails to preserve the detailed information of the foreground objects effectively. In contrast, our method is adept at preserving the foreground details and simultaneously adjusting the foreground pose/view to fit the background.

4.7 Ablation Study

In this section, we study the impact of global and local feature calibration. We progressively add the proposed modules and report the results on MureCom dataset in Table 3.

The first row shows the results of original ObjectStitch without reference feature calibration. In the second (*resp.*, third) row, we add global (*resp.*, local) reference feature calibration. It can be seen that all four metrics are improved, indicating the enhancement of detail preservation and overall quality. In the fourth row, we only inject calibrated features into the decoder of denoising UNet. By comparing the first row and the fourth row, we can see that only using calibrated feature can achieve satisfactory performance, which justifies the effectiveness of calibrated features. In the last row, we report the results of our full method. Our full method achieves the best results for all metrics.

We also provide the visualization results of ablation study in Fig. 4. Without the LRFC module, the ability to preserve foreground details is degraded. Without the GRFC module, the ability to adjust the foreground pose/view is impaired. Without using the uncalibrated features, some artifacts can be observed on the foreground object, probably because that the calibration process causes the detail information loss and the original reference features can effectively supplement more details.

GRFC	LRFC	UCF	CF	$\text{DINO}_{\text{fg}}\uparrow$	$\text{SSIM}_{\text{bg}}\uparrow$	FOSScore↑	QS↑
		✓		65.04	0.854	0.867	44.39
✓		✓	✓	65.21	0.856	0.871	45.72
	✓	✓	✓	66.64	0.857	0.873	46.95
✓	✓		✓	66.73	0.856	0.874	46.87
✓	✓	✓	✓	**68.60**	**0.859**	**0.883**	**47.07**

Table 3: Ablation study of the impact of global reference feature calibration module (GRFC), local reference feature calibration module (LRFC), and whether to use calibrated features (CF) and uncalibrated features (UCF).

4.8 Effectiveness of Feature Calibration

To validate the effectiveness of feature calibration, that is, the calibrated reference features are closer to ground-truth features, we calculate the L2 distance between reference fea-

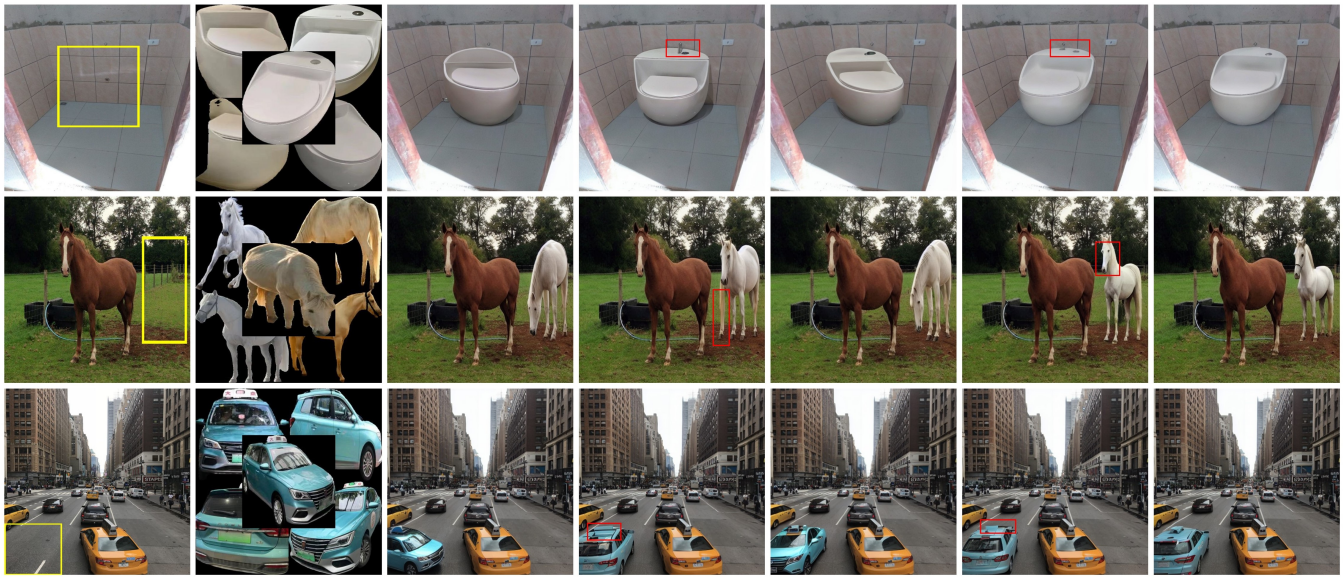

Figure 4: Ablation study of our GRFC/LRFC modules and uncalibrated features (UCF). From left to right, we show background image, five reference images, the results of ObjectStitch, three ablated versions of our method, and our full method. The red boxes indicate the changed details.

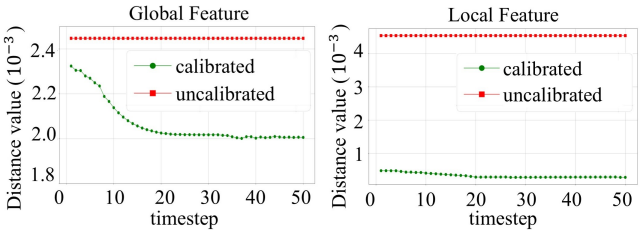

Figure 5: The distance between uncalibrated/calibrated reference features and ground-truth reference features along with the denoising step.

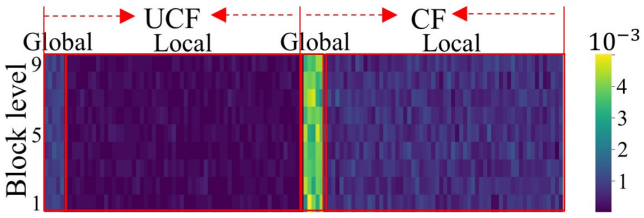

Figure 6: Visualization of cross-attention map in different decoder blocks. Brighter colors indicate larger values.

tures before/after calibration and ground-truth features on MVImgNet test set, because MVImgNet test set has ground-truth features. For each test example, we first extract the global reference features $\{\boldsymbol{f}_k^g|_k\}$ and local reference features $\{\boldsymbol{f}_{k,i}^l|_{k,i}\}$, followed by calculating their distances to ground-truth global/local features. The averaged distances are plotted in red in Fig. 5, which serves as the baseline. When going through the denoising process, we record the calibrated global reference features $\{\tilde{\boldsymbol{f}}_k^g|_k\}$ and local reference features $\{\tilde{\boldsymbol{f}}_{k,i}^l|_{k,i}\}$ at each timestep, followed by calculating their distances to ground-truth global/local reference features. The averaged distance values are plotted in green in Fig. 5. When compared with uncalibrated reference features, calibrated reference features are progressively getting closer to ground-truth reference features as the denoising procedure advances, demonstrating the effectiveness of feature calibration.

To further investigate the role of different features in the calibration process, we visualize the cross-attention maps between decoder features and reference features in different decoder blocks in Fig. 6. The model assigns higher attention to the calibrated features (CF), with the calibrated global features receiving the highest attention, which shows that calibrated features can provide crucial guidance for foreground generation. Uncalibrated features (UCF) are assigned relatively low weights, while they still contribute to the generation process to some extent.

5 Conclusion

In this paper, we have proposed a multi-reference generative composition framework, which can utilize arbitrary number of foreground reference images. Under our framework, we have further proposed to calibrate the foreground reference features to be compatible with the background. Comprehensive experiments have verified the effectiveness of our framework equipped with calibrated reference features.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (Grant No. 62471287).

References

Canet Tarrés, G.; Lin, Z.; Zhang, Z.; Zhang, J.; Song, Y.; Ruta, D.; Gilbert, A.; Collomosse, J.; and Kim, S. Y. 2025. Thinking Outside the BBox: Unconstrained Generative Object Compositing. In *ECCV*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. Anydoor: Zero-shot object-level image customization. In *CVPR*.

Chen, Z.; Wang, W.; Yang, Z.; Yuan, Z.; Chen, H.; and Shen, C. 2025. FreeCompose: Generic Zero-Shot Image Composition with Diffusion Prior. In *ECCV*.

Choi, J.; Choi, Y.; Kim, Y.; Kim, J.; and Yoon, S. 2023. Custom-edit: Text-guided image editing with customized diffusion models. In *CVPR workshop*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023b. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG*, 42(4): 1–13.

Gu, J.; Wang, Y.; Zhao, N.; Fu, T.-J.; Xiong, W.; Liu, Q.; Zhang, Z.; Zhang, H.; Zhang, J.; Jung, H.-S.; and Wang, X. 2023. Photoswap: Personalized Subject Swapping in Images. In *NeurIPS*.

Guerreiro, J. J. A.; Nakazawa, M.; and Stenger, B. 2023. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *CVPR*.

Hong, Y.; Niu, L.; and Zhang, J. 2022. Shadow generation for composite image in real-world scenes. In *AAAI*.

Huang, J.; Yan, P.; Liu, J.; Wu, J.; Wang, Z.; Wang, Y.; Lin, L.; and Li, G. 2025. DreamFuse: Adaptive Image Fusion with Diffusion Transformer. In *ICCV*.

Jiang, Y.; Zhang, H.; Zhang, J.; Wang, Y.; Lin, Z.; Sunkavalli, K.; Chen, S.; Amirghodsi, S.; Kong, S.; and Wang, Z. 2021. Ssh: A self-supervised framework for image harmonization. In *ICCV*.

Ke, Z.; Sun, C.; Zhu, L.; Xu, K.; and Lau, R. W. 2022. Harmonizer: Learning to Perform White-Box Image and Video Harmonization. In *ECCV*.

Kulal, S.; Brooks, T.; Aiken, A.; Wu, J.; Yang, J.; Lu, J.; Efros, A. A.; and Singh, K. K. 2023. Putting People in Their Place: Affordance-Aware Human Insertion into Scenes. In *CVPR*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*.

Labs, B. F. 2024. FLUX. Accessed: 2024-11-12.

Li, P.; Nie, Q.; Chen, Y.; Jiang, X.; Wu, K.; Lin, Y.; Liu, Y.; Peng, J.; Wang, C.; and Zheng, F. 2024. Tuning-Free Image Customization with Image and Text Guidance. In *ECCV*.

Li, Y.; Li, X.; Zhang, Z.; Bian, Y.; Liu, G.; Li, X.; Xu, J.; Hu, W.; Liu, Y.; Li, L.; Cai, J.; Zou, Y.; He, Y.; and Shan, Y. 2025. IC-Custom: Diverse Image Customization via In-Context Learning. *arXiv preprint arXiv:2507.00000*.

Liu, Q.; You, J.; Wang, J.; Tao, X.; Zhang, B. Z.; and Niu, L. 2024. Shadow Generation for Composite Image Using Diffusion Model. In *CVPR*.

Lu, L.; Li, J.; Zhang, B.; and Niu, L. 2023. Dreamcom: Finetuning text-guided inpainting model for image composition. *arXiv preprint arXiv:2309.15508*.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *ICCV*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 6038–6047.

Pham, K. T.; Chen, J.; and Chen, Q. 2024. TALE: Training-free Cross-domain Image Composition via Adaptive Latent Manipulation and Energy-guided Optimization. In *ACM MM*, 3160–3169.

Qin, Y.; Xu, J.; Wang, R.; and Chen, X. 2025. Think Before Placement: Common Sense Enhanced Transformer for Object Placement. In *ECCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*.

Ruiz, N.; Li, Y.; Jampani, V.; Wei, W.; Hou, T.; Pritch, Y.; Wadhwa, N.; Rubinstein, M.; and Aberman, K. 2024. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *CVPR*.

Sheng, Y.; Liu, Y.; Zhang, J.; Yin, W.; Oztireli, A. C.; Zhang, H.; Lin, Z.; Shechtman, E.; and Benes, B. 2022. Controllable Shadow Generation Using Pixel Height Maps. In *ECCV*.

Song, W.; Jiang, H.; Yang, Z.; Quan, R.; and Yang, Y. 2025. Insert Anything: Image Insertion via In-Context Editing in DiT. *arXiv preprint arXiv:2504.15009*.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; and Aliaga, D. 2023. ObjectStitch: Object Compositing With Diffusion Model. In *CVPR*.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; Zhang, H.; Xiong, W.; and Aliaga, D. 2024. Imprint: Generative object compositing by learning identity-preserving representation. In *CVPR*.

Tao, J.; Zhang, Y.; Wang, Q.; Cheng, Y.; Wang, H.; Bai, X.; Zhou, Z.; Li, R.; Wang, L.; Wang, C.; et al. 2025. InstantCharacter: Personalize Any Characters with a Scalable Diffusion Transformer Framework. *arXiv preprint arXiv:2504.12395*.

Tao, X.; Qiu, T.; Cao, J.; and Niu, L. 2024. Diverse Image Harmonization. *arXiv preprint arXiv:2407.15481*.

Tarrés, G. C.; Lin, Z.; Zhang, Z.; Zhang, H.; Gilbert, A.; Collomosse, J.; and Kim, S. Y. 2025. Multitwine: Multi-Object Compositing with Text and Layout Control. In *CVPR*.

Tsai, Y.-H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; and Yang, M.-H. 2017. Deep image harmonization. In *CVPR*.

Wang, H.; Peng, J.; He, Q.; Yang, H.; Jin, Y.; Wu, J.; Hu, X.; Pan, Y.; Gan, Z.; Chi, M.; et al. 2025. UniCombine: Unified Multi-Conditional Combination with Diffusion Transformer. In *ICCV*.

Wang, Y.; Zhang, W.; Zheng, J.; and Jin, C. 2024. Primecomposer: Faster Progressively Combined Diffusion for Image Composition with Attention Steering. In *ACM MM*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*.

Winter, D.; Cohen, M.; Fruchter, S.; Pritch, Y.; Rav-Acha, A.; and Hoshen, Y. 2024. ObjectDrop: Bootstrapping Counterfactuals for Photorealistic Object Removal and Insertion. In *ECCV*.

Winter, D.; Shul, A.; Cohen, M.; Berman, D.; Pritch, Y.; Rav-Acha, A.; and Hoshen, Y. 2025. ObjectMate: A Recurrence Prior for Object Insertion and Subject-Driven Generation. In *ICCV*.

Xu, Y.; Tang, F.; Wu, Y.; Gao, L.; Deussen, O.; Yan, H.; Li, J.; Cao, J.; and Lee, T.-Y. 2025. In-Context Brush: Zero-shot Customized Subject Insertion with Context-Aware Latent Space Manipulation. *arXiv preprint arXiv:2505.20271*.

Xue, B.; Ran, S.; Chen, Q.; Jia, R.; Zhao, B.; and Zhao, B. 2022. DCCF: Deep Comprehensible Color Filter Learning Framework for High-Resolution Image Harmonization. In *ECCV*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*.

Yang, Y.; Peng, H.; Shen, Y.; Yang, Y.; Hu, H.; Qiu, L.; Koike, H.; et al. 2024. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. In *NeurIPS*.

Yu, X.; Xu, M.; Zhang, Y.; Liu, H.; Ye, C.; Wu, Y.; Yan, Z.; Liang, T.; Chen, G.; Cui, S.; and Han, X. 2023. MVImgNet: A Large-scale Dataset of Multi-view Images. In *CVPR*.

Yu, Y.; Zeng, Z.; Zheng, H.; and Luo, J. 2025. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. In *ICCV*.

Yuan, Z.; Cao, M.; Wang, X.; Qi, Z.; Yuan, C.; and Shan, Y. 2024. CustomNet: Zero-shot Object Customization with Variable-Viewpoints in Text-to-Image Diffusion Models. In *ACM MM*.

Zhang, B.; Duan, Y.; Lan, J.; Hong, Y.; Zhu, H.; Wang, W.; and Niu, L. 2023. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*.

Zhang, B.; Sui, J.; and Niu, L. 2023. Foreground Object Search by Distilling Composite Image Feature. In *ICCV*.

Zhang, L.; Wen, T.; and Shi, J. 2020. Deep image blending. In *CVPR*.

Zhang, Z.; Xie, J.; Lu, Y.; Yang, Z.; and Yang, Y. 2025. In-Context Edit: Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer. In *NeurIPS*.

Zhou, S.; Liu, L.; Niu, L.; and Zhang, L. 2022. Learning object placement via dual-path graph completion. In *ECCV*.

Zhu, S.; Lin, Z.; Cohen, S.; Kuen, J.; Zhang, Z.; and Chen, C. 2023. TopNet: Transformer-based Object Placement Network for Image Compositing. In *CVPR*.