# *CompleteMe*: Reference-based Human Image Completion

Yu-Ju Tsai[1] Brian Price[2] Qing Liu[2] Luis Figueroa[2]
Daniil Pakhomov[2] Zhihong Ding[2] Scott Cohen[2] Ming-Hsuan Yang[1]
[1]UC Merced [2]Adobe Research

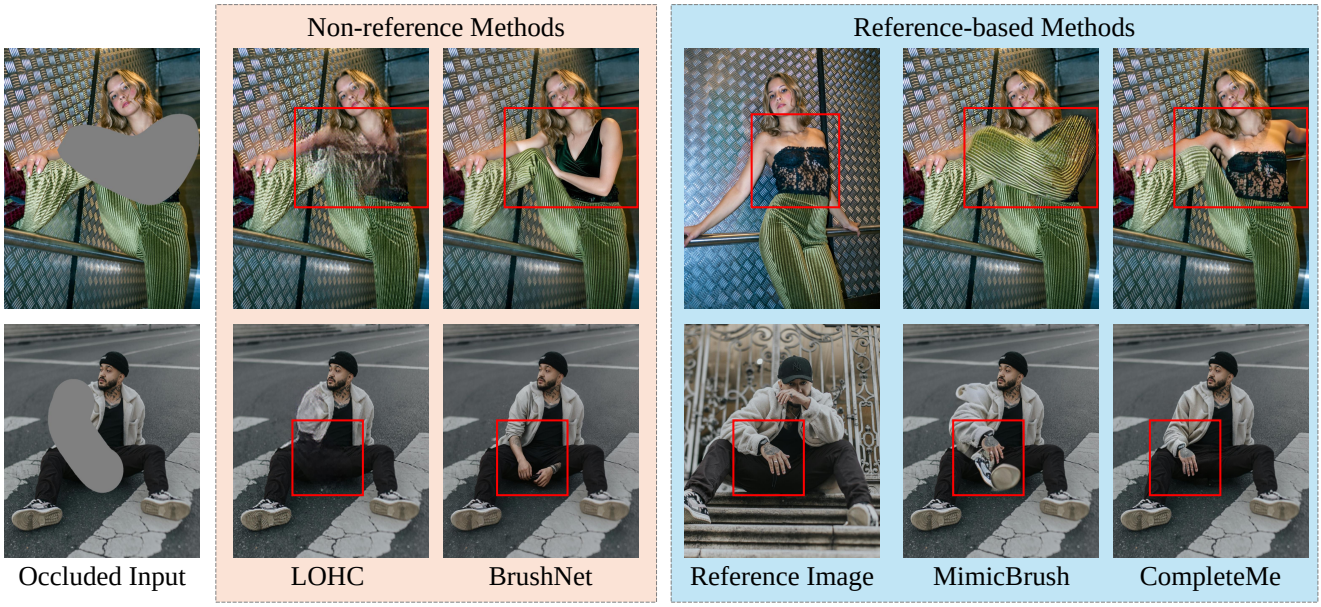


Figure 1. Given occluded human image, non-reference methods, LOHC [41] and BrushNet [12], can generate plausible results but lack the unique information of the person like special clothing and tattoo pattern (highlighted in Red box). Such information can be only acquired by additional reference images. Given the reference image, MimicBrush [3] fails to find the corresponding parts between input and reference. Our *CompleteMe* can preserve identical and fine-detail information from the reference image and generate a consistent result.

## Abstract

*Recent methods for human image completion can reconstruct plausible body shapes but often fail to preserve unique details, such as specific clothing patterns or distinctive accessories, without explicit reference images. Even state-of-the-art reference-based inpainting approaches struggle to accurately capture and integrate fine-grained details from reference images. To address this limitation, we propose CompleteMe, a novel reference-based human image completion framework. CompleteMe employs a dual U-Net architecture combined with a Region-focused Attention (RFA) Block, which explicitly guides the model's attention toward relevant regions in reference images. This approach effectively captures fine details and ensures accurate semantic correspondence, significantly improving the fidelity and consistency of completed images. Additionally, we introduce a challenging benchmark specifically designed for evaluating reference-based human image completion tasks. Extensive experiments demonstrate that our proposed method achieves superior visual quality and semantic consistency compared to existing techniques.*

## 1. Introduction

Human image completion [30, 41–43] is an essential task in computer vision, with a wide range of applications, including photo editing [13, 28], virtual try-on [5, 7, 14, 20], and animation [9, 34]. The ability to accurately reconstruct missing parts of human images has significant implications for enhancing user experience in these areas. Traditional inpainting methods [37, 38] have made strides in generating plausible image completions, but they often fall short in maintaining consistency of complex features like clothing, pose, and human anatomy. These challenges become even

more pronounced when dealing with large or irregular missing regions, which require a comprehensive understanding of both the local and global context of an image.

Amodal completion methods [22, 33, 39] have recently garnered attention for their ability to infer occluded parts of an object beyond visible regions. These approaches aim to reconstruct the entirety of an object even when portions are entirely hidden, relying on learned priors to predict missing information. However, they primarily focus on reconstructing general object shapes obscured by occlusions and often fall short in complex scenarios that involve varied human poses or intricate details, such as unique clothing patterns or distinctive features like tattoos. Without explicit reference information, these methods struggle to generate accurate completions that capture individual characteristics, as people often seek to faithfully restore specific, original details. Amodal completion methods, however, are currently unable to achieve this level of precise restoration.

Reference-based inpainting [1, 3, 4, 26, 27, 35] provides a promising solution by utilizing additional reference images that share similar attributes, offering valuable information for reconstructing missing regions. These methods leverage visual cues from reference images, such as clothing details, textures, or human poses, to fill in missing regions more accurately and consistently. Despite these advancements, these methods mainly focus on object-level insertion or completion, and challenges still remain in terms of effectively capturing fine-grained details, particularly in cases involving intricate clothing patterns and unique parts of the person, where explicit reference information is crucial for generating identical results.

To address the above issue, we propose *CompleteMe*, a reference-guided human image completion framework that leverages reference images to guide the completion process. Our model is based on a dual U-Net structure, consisting of the Reference U-Net and the Complete U-Net, which separately handle reference information and completing for the occluded input. To improve correspondence, we divide different parts of human appearance (e.g., hair, face, clothes, shoes) into separate reference images for the Reference U-Net. These reference features are then integrated into the Complete U-Net via our newly designed Region-focused Attention (RFA) Block. The RFA Block explicitly guides attention toward relevant reference regions based on reference masks, effectively establishing precise correspondences and improving the model's ability to produce more realistic and semantically accurate completions, particularly for challenging cases involving complex clothing patterns, body patterns, or unique accessories. As shown in Fig. 1, *CompleteMe* can generate more fine-detail results based on the information provided by the reference image, outperforming other methods. To comprehensively evaluate the performance of various methods on reference-based human completion tasks, we construct a challenging benchmark featuring significant body pose differences and varying scenarios between the occluded input and the reference image. This benchmark tests the model's ability to generate consistent information and establish proper correspondences. Our contributions are summarized as follows:

- We propose *CompleteMe*, a novel reference-based human image completion model employing a dual U-Net architecture enhanced by our Region-focused Attention Block, explicitly designed to preserve fine details and identity consistency with enhanced correspondence.
- We construct a challenging benchmark dataset with significant pose differences and varying scenarios to systematically evaluate the model's ability to find proper correspondences and maintain identical and consistent information from the reference image.
- We conduct comprehensive experiments, including a large user study, to demonstrate the best performance of the proposed method both qualitatively and quantitatively.

## 2. Related Work

**Image Completion.** Recent advancements in object image completion have introduced various methods to address the challenges in reconstructing missing or occluded regions. Xiong *et al*. [32] develop a foreground-aware image inpainting method incorporating explicit contour guidance to enhance object reconstruction. SmartBrush [31] combines text and shape guidance with a diffusion model to fill missing regions with detailed object reconstructions. BrushNet [12] introduces a plug-and-play dual-branch model to embed pixel-level masked image features into any pre-trained text-to-image diffusion model to generate inpainting outcomes. For human-centric image completion, FiNet [40] propose Fashion Inpainting Networks, which reconstruct missing clothing parts in fashion portrait images using parsing maps as priors. Wu *et al*. [30] extend the approach with a two-stage deep learning framework for portrait image completion, utilizing a human parsing network to extract the body structure before filling in unknown regions. Zhao *et al*. [42] propose a prior-based human completion method, incorporating structural and texture correlation priors to recover realistic human forms. LOHC [41] introduces a two-stage coarse-to-fine method and leverages human segmentation maps as a prior, and completes the image and segmentation prior simultaneously.

**Reference-based Inpainting.** Reference-based image inpainting has made significant improvements in recent years, focusing on leveraging external references to improve image completion tasks with enhanced realism and semantic accuracy. TransFill [44] introduces a method that aligns source and target images using multiple homography informed by depth levels. Paint-by-Example [35] leverages diffusion models for exemplar-guided editing, integrating

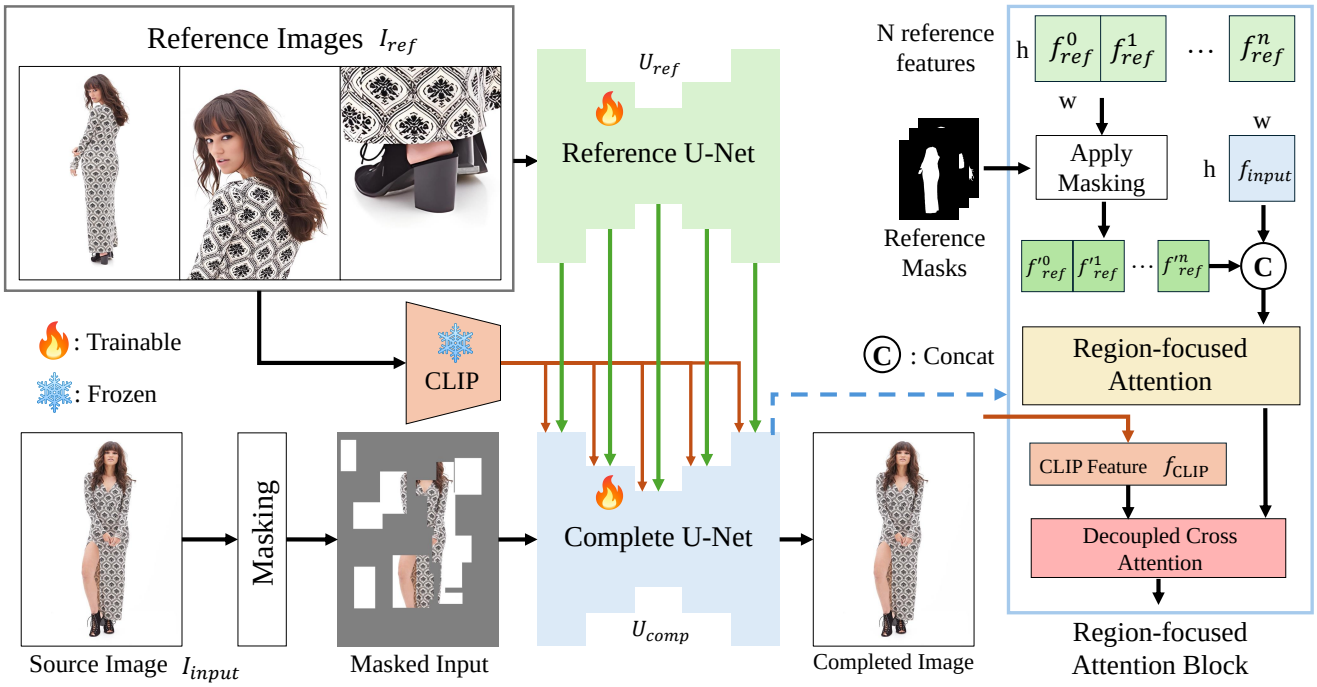


Figure 2. ***CompleteMe* Pipeline.** Our proposed *CompleteMe* utilizes a dual U-Net framework composed of a Reference U-Net ($U_{ref}$) and a Complete U-Net ($U_{comp}$). Given an input image ($I_{input}$) with masked regions, we first encode the input image to latent feature $f_{input}$. The Reference U-Net then extracts detailed visual features ($f^0_{ref}, f^1_{ref}, \ldots, f^n_{ref}$) from multiple reference images ($I_{ref}$), which consist of different human body parts. Along with global semantic features ($f_{\text{CLIP}}$) extracted by CLIP, the reference features are processed within our novel Region-focused Attention (RFA) Block embedded in the Complete U-Net. These reference features are then explicitly masked according to reference masks, producing masked reference features ($f'^0_{ref}, f'^1_{ref}, \ldots, f'^n_{ref}$). This explicit masking and concatenation strategy enables the model to precisely zoom in and focus on relevant human regions, establishing accurate and fine-grained correspondences through the Region-focused Attention mechanism. Finally, decoupled cross-attention integrates these refined local features with the global semantic CLIP features ($f_{\text{CLIP}}$), resulting in a detailed and semantically coherent completion.

example patches into target images. ObjectStitch [26] uses conditional diffusion models and introduces a content adaptor to maintain categorical semantics and object appearance. AnyDoor [4] introduces a zero-shot framework that teleports target objects into new scenes at user-specified locations and orientations. IMPRINT [27] proposes a diffusion model trained with a two-stage learning framework that decouples learning of identity preservation from compositing. LeftRefill [1] presents a strategy that stitches reference and target views as a unified input to a text-to-image diffusion model. MimicBrush [3] offers an approach to locally edit the source region with reference images by training dual diffusion U-Nets in a self-supervised manner with video data.

These methods illustrate the progression of reference-based inpainting, moving from traditional alignment techniques to advanced diffusion-based models prioritizing identity preservation, contextual consistency, and zero-shot learning capabilities. However, human image completion presents a more complex challenge, as current methods primarily focus on object-level completion and struggle to establish accurate correspondences between the source and reference when conditions differ significantly.

## 3. Method

### 3.1. Overall Pipeline

Our proposed *CompleteMe* utilizes a dual U-Net architecture comprising a Reference U-Net ($U_{ref}$) and a Complete U-Net ($U_{comp}$), as illustrated in Fig. 2, explicitly tailored for reference-based human image completion. Given an input source image ($I_{input}$) with masked regions, our masking strategy applies random grid masking (50% probability) 1 to 30 times and employs human body shape masks (50% probability) to ensure complexity and realism. The Reference U-Net ($U_{ref}$) first extracts detailed spatial features ($f^0_{ref}, f^1_{ref}, \ldots, f^n_{ref}$) from multiple reference images ($I_{ref}$), which consist of different human body parts. The reference features are then processed within our novel Region-focused Attention (RFA) Block, embedded in the Complete U-Net ($U_{comp}$). These extracted reference features are explicitly masked using corresponding reference masks, yielding masked reference features ($f'^0_{ref}, f'^1_{ref}, \ldots, f'^n_{ref}$). The RFA block explicitly guides the input feature $f_{input}$ with attention toward relevant human regions inside masked reference features ($f'^i_{ref}$). Along with the global semantic features

($f_{\text{CLIP}}$) from CLIP [23] encoder, the RFA Block enables the model to precisely identify and establish accurate correspondences, significantly enhancing detail preservation and semantic coherence. During inference, our model is flexible, operating effectively even with a single reference image and optionally incorporating textual prompts, enabling practical and versatile human image completion.

### 3.2. Reference Feature Encoding

In reference-based image inpainting tasks, previous approaches [4, 26, 27, 35] typically utilize semantic-level encoders such as CLIP [23] or DINOv2 [21] to extract global features from reference images. However, these methods often lose crucial spatial information, resulting in limited preservation of fine-grained appearance details. Motivated by recent successes in image and video generation conditioned on reference images [3, 9, 10, 34], we propose a specialized Reference U-Net encoder designed for detailed identity preservation across multiple reference images.

Our Reference U-Net ($U_{ref}$) is initialized from pretrained Stable Diffusion 1.5 [24] weights but operates explicitly without the diffusion-based noise step (at timestep zero), directly encoding reference images ($I_{ref}$) into latent visual features ($f^0_{ref}, f^1_{ref}, \ldots, f^n_{ref}$). Each reference image, corresponding to distinct human appearance attributes (e.g., upper body, lower body, shoes), is first transformed into latent representations and then sequentially processed by the Reference U-Net. This sequential encoding strategy ensures flexibility and robustness, effectively managing varying numbers and types of reference images while preserving detailed appearance information. Additionally, global semantic features ($f_{\text{CLIP}}$) are extracted from each reference image using the CLIP [23] image encoder, supplementing the spatially-detailed latent features with global semantic context. These combined reference and semantic CLIP features are cached before feeding to our Region-focused Attention (RFA) Block, facilitating efficient and detail-preserving encoding process.

### 3.3. Completion Process

**Complete U-Net.** Our Complete U-Net ($U_{comp}$), initialized from pretrained Stable Diffusion 1.5 [24] inpainting model, receives as input a source image ($I_{input}$) with masked regions represented in the latent space, along with cached latent reference features ($f^0_{ref}, f^1_{ref}, \ldots, f^n_{ref}$) and global CLIP features ($f_{\text{CLIP}}$), as shown in Fig. 2. The Complete U-Net then processes a concatenation of these masked reference features with the input feature ($f_{input}$) inside Region-focused Attention Block, providing detailed context for the completion task.

**Region-focused Attention Block.** To effectively integrate detailed local information from reference images, we introduce the Region-focused Attention (RFA) Block, as illustrated in Fig. 2. Given the encoded latent reference features ($f^i_{ref}$), we explicitly mask irrelevant regions using the corresponding reference masks, generating masked reference features ($f'^i_{ref}$). These masked reference features ($f'^i_{ref}$) are then concatenated with latent input features ($f_{input}$) extracted from the input image to form the concatenated feature ($f_{concat}$). Within the RFA block, we apply region-focused attention to the concatenated features as follows:

$$\text{Region-focused Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (1)$$

where the queries ($Q$), keys ($K$), and values ($V$) are defined as: $Q = f_{input}$, $K, V = f_{concat}$. This region-focused attention allows the model to explicitly identify accurate and fine-grained spatial correspondences between the masked source regions and relevant masked reference regions. After this detailed correspondence establishment via region-focused attention, we utilize the decoupled cross-attention mechanism proposed by IP-Adapter [36] to fuse the refined, detail-focused local features with global semantic features ($f_{\text{CLIP}}$). Specifically, we perform two separate cross-attention operations—one using the refined local features, and the other using the global CLIP features—then sum their outputs to form enriched, semantically consistent feature maps. This explicit integration of visual and textual information results in more detailed, coherent, and contextually accurate completed outcomes.

### 3.4. Evaluation Benchmark

Since no suitable dataset evaluates the reference-based human image completion task, we construct our benchmark to systematically evaluate the performance of different methods. Our main target is to establish the scenario in which reference images are necessary for completing the unique information. We establish the benchmark to meet the following criteria: 1) the same person in the same clothing, 2) a significantly different pose, 3) unique patterns like special clothing, accessories, or tattoos, and 4) different background conditions. To construct the benchmark, we first select image pairs from the Wpose dataset in UniHuman [16], which contains a wide variety of poses, allowing us to test the model's ability to find the proper correspondence. We manually draw the source mask to indicate the inpainting area. Finally, we obtain 417 image groups, each consisting of a source image, inpainting area, and reference image, please refer to supplementary material for more benchmark examples. Additionally, we use LLaVA [17, 18] to generate text prompts describing the source image. For evaluation metrics, we use CLIP [23] to calculate text-to-image and image-to-image similarity, DINO [2] to calculate similarity scores, and the DreamSim [6] metric to better evaluate the generated results. Aside from these metrics, we also use PSNR [8], SSIM [29], and LPIPS [40] as our evaluation metrics for masked regions.

Table 1. **Quantitative Comparison on Our Benchmark (Sec. 3.4).** "CLIP-I" measures the similarity between images. "CLIP-T" measures the similarity between text and image. Red and blue indicate the best and second-best, respectively.

| Method | CLIP-I ↑ | CLIP-T ↑ | DINO ↑ | DreamSim [6] ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ |
|---|---|---|---|---|---|---|---|
| LOHC [41] | 96.03 | 29.46 | 82.52 | 0.0732 | 0.0709 | 28.4884 | 0.9264 |
| BrushNet [12] | 95.90 | **30.69** | 95.08 | 0.0576 | 0.0600 | 28.5764 | 0.9224 |
| Paint-by-Example [35] | 95.04 | 29.79 | 94.98 | 0.0611 | 0.0601 | 28.6441 | 0.9222 |
| AnyDoor [4] | 89.65 | 28.14 | 88.80 | 0.1454 | 0.0812 | 28.1807 | 0.9089 |
| LeftRefill [1] | 96.33 | 29.74 | 95.12 | 0.0574 | 0.0598 | **28.8657** | **0.9283** |
| MimicBrush [3] | 96.98 | 29.48 | 94.37 | 0.0651 | 0.0694 | 28.3598 | 0.9174 |
| *CompleteMe* (Ours) | **97.18** | 29.83 | **96.29** | **0.0419** | **0.0588** | 28.7020 | 0.9239 |

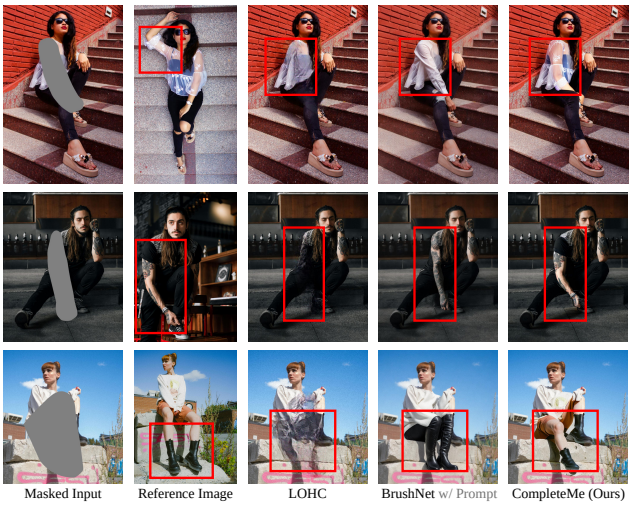


Figure 3. **Qualitative Comparison with Non-reference Methods.** We compare *CompleteMe* with non-reference methods, LOHC[41] and BrushNet [12]. Given masked inputs, these non-reference methods generate plausible content for the masked regions using image priors or text prompts. However, as indicated in the Red box, they cannot reproduce specific details such as tattoos or unique clothing patterns, as they lack reference images to guide the reconstruction of identical information.

# 4. Experiments

## 4.1. Experimental Setting

**Implementation Details.** In this work, we employ the Reference U-Net and the Complete U-Net, initialized with pre-trained weights from Stable Diffusion-1.5 [24] and Stable Diffusion-1.5 inpainting model. Our image encoder uses CLIP [23] Vision Model, along with projection layers. For training, we use Adam [15] optimizer and set an initial learning rate of $2 \times 10^{-5}$ with a total batch size of 64. Training is performed on 8 NVIDIA A100 GPUs for 30,000 iterations. We apply mean square error (MSE) loss as our supervision. To enhance the robustness of the model, we employ a random drop strategy, where all reference image features are randomly dropped with a probability of 0.2. This helps the model learn to handle cases with partial information from reference images. Additionally, to increase the flexibility of the completion process, each reference condition is randomly dropped with a probability of 0.2, allowing image completion to be conditioned on various reference images. During inference, we adopt the DDIM sampler [25] with 50 steps and set the guidance scale to 7.5 to improve output quality and identity.

**Training Dataset.** To train our *CompleteMe* model, we modify a multi-modal human dataset based on [10], which is constructed from the DeepFashion-MultiModal [11, 19] dataset. To meet our requirements, we rebuild the training pairs by using occluded images with multiple reference images that capture various aspects of human appearance along with their short textual labels. Each sample in our training data includes six appearance types: *upper body clothes*, *lower body clothes*, *whole body clothes*, *hair or headwear*, *face*, and *shoes*. For the masking strategy, we apply 50% random grid masking between 1 to 30 times, while for the other 50%, we use a human body shape mask to increase masking complexity. After the construction pipeline, we obtained 40,000 image pairs for training.

## 4.2. Comparison with Other Methods

In this section, we compare our *CompleteMe* with other approaches capable of performing similar functions in the reference-based human image completion task. Among non-reference methods, we select LOHC [41], the state-of-the-art in non-reference human image completion, and BrushNet [12], a leading model for image inpainting with text prompts. For reference-based methods, we include Paint-by-Example [35], AnyDoor [4], LeftRefill [1], and MimicBrush [3] for a comprehensive comparison. We also provide additional inputs where applicable for previous methods. For instance, we include extra prompts for BrushNet [12] and supply reference region masks for Paint-by-Example [35] and AnyDoor [4].

**Quantitative Comparison.** To assess the effectiveness of *CompleteMe*, we perform a quantitative comparison with other state-of-the-art methods for human image completion. We evaluate both non-reference and reference-based inpainting approaches using several metrics: CLIP-I [23] (image-to-image), CLIP-T [23] (text-to-image), DINO [2], DreamSim [6], PSNR [8], SSIM [29], and LPIPS [40]. As shown in Table 1, *CompleteMe* demon-

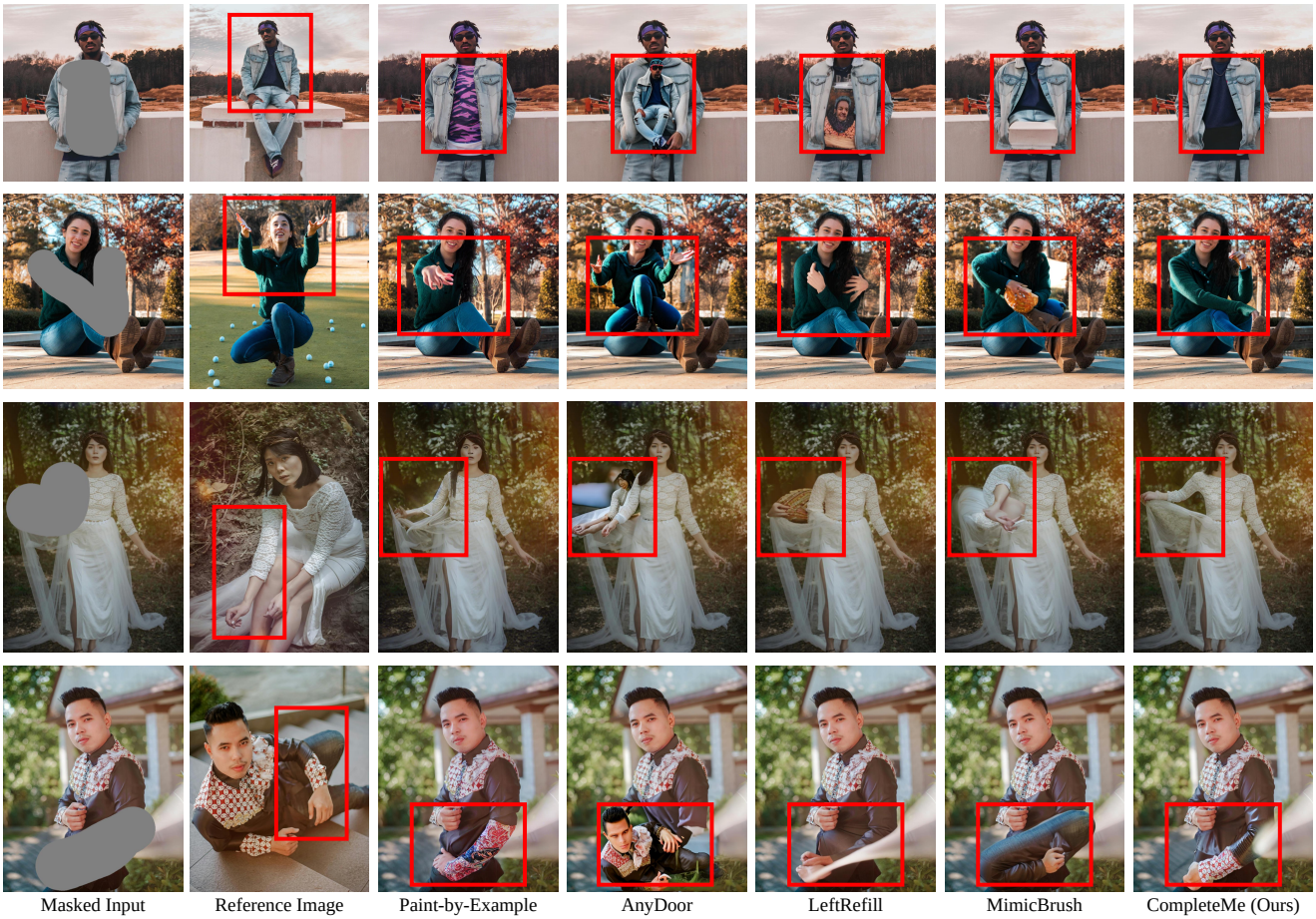


Figure 4. **Qualitative Comparison with Reference-based Methods.** Our *CompleteMe* can generate more realistic and preserve identical information from the reference image. Please refer to the Red box region for a more detailed comparison.

strates strong performance across various perceptual metrics, outperforming other methods in CLIP-I, DINO, DreamSim, and LPIPS, which reflect our ability to maintain semantic alignment and appearance fidelity with the reference image. In terms of image quality metrics, *CompleteMe* achieves competitive PSNR and SSIM scores, demonstrating its high-fidelity reconstructions. These quantitative results illustrate that *CompleteMe* achieves better performance across semantic similarity, structural fidelity, and perceptual quality, positioning it as a robust solution for reference-based human image completion.

**Qualitative Comparison.** For qualitative comparison, we first compare our *CompleteMe* with non-reference methods, LOHC [41] and BrushNet [12], as shown in Fig. 1 and Fig. 3. Given masked inputs, these non-reference methods generate plausible content for the masked regions by leveraging image priors or additional text prompts. However, as highlighted in the red box, they are unable to replicate specific details, such as tattoos or unique clothing patterns, due to the absence of reference images to guide the reconstruction of identical features.

As shown in Fig. 4, we compare *CompleteMe* with reference-based methods: Paint-by-Example [35], AnyDoor [4], LeftRefill [1], and MimicBrush [3]. For the setting of comparison, we use only one reference image and text prompt for our method. Given a masked human image and a reference image, other methods can generate plausible content but often fail to preserve contextual information from the reference accurately. In some cases, they generate irrelevant content or incorrectly map corresponding parts from the reference image. In contrast, *CompleteMe* effectively completes the masked region by accurately preserving identical information and correctly mapping corresponding parts of the human body from the reference image.

**User Study.** Recognizing that metrics alone may not fully capture human preferences, we conducted a user study, as shown in Table 2. We asked 15 annotators to evaluate the generated results from various models on our benchmark (described in Sec. 3.4) and acquire 2,895 groups of data points. We construct the "one-to-one" evaluation pair between *CompleteMe* and other four methods (Paint-by-Example [35], AnyDoor [4], LeftRefill [1], and MimicBrush [3]) with masked input and reference image. Each group sample is assessed based on two primary criteria:

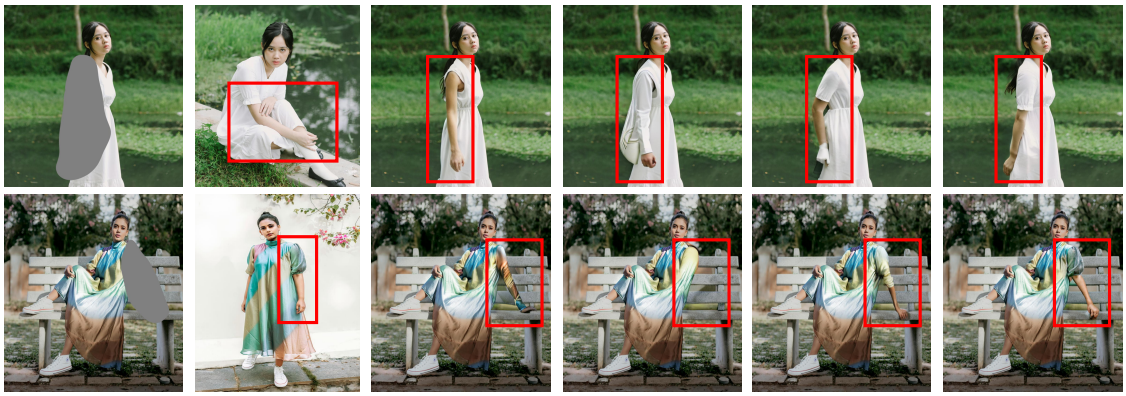


Figure 5. **Qualitative Comparison on Different Reference Image Encoder.** We conduct the ablation study for different encoders to extract the feature from reference images. CLIP [23] and DINOv2 [21] can find the correspondence between masked input and the reference image, but they can not preserve the detailed information compared to the U-Net encoder. For the effectiveness of our Region-focused Attention (RFA), this design further helps preserve the identical information. Please zoom in for the detail inside the Red box.

Table 2. **User Study on Our Benchmark.** We conduct a user study on our proposed benchmark (see Sec. 3.4). "Quality" and "Identity" measure the completion quality and preservation of identical information from the reference image. We report the one-to-one comparison between these methods and *CompleteMe*. For "a/b", a is the percentage where the compared method is considered better than *CompleteMe*, and b is the percentage where *CompleteMe* is considered better than the compared method.

| Evaluation | Quality | Identity |
|---|---|---|
| Method | CompleteMe | |
| Paint-by-Example [35] | 6.12%/**93.88%** | 3.48%/**96.52%** |
| AnyDoor [4] | 0.55%/**99.45%** | 1.13%/**98.87%** |
| LeftRefill [1] | 14.97%/**85.03%** | 4.58%/**95.42%** |
| MimicBrush [3] | 5.14%/**94.86%** | 6.05%/**93.95%** |

Table 3. **Ablation on Different Masking Ratios.** We conduct experiments with different ratios between human shape and random mask (0% to 100%) and evaluate performance using CLIP-I, DINO, and DreamSim.

| Random Mask Ratio | 0 % | 25 % | 50 % | 75 % | 100 % |
|---|---|---|---|---|---|
| CLIP-I ↑ | 97.09 | 97.02 | **97.18** | 97.07 | 96.78 |
| DINO ↑ | 96.22 | 96.26 | **96.29** | 96.10 | 95.60 |
| DreamSim ↓ | 0.0426 | 0.0419 | **0.0419** | 0.0434 | 0.0495 |

"Quality" and "Identity". The "Quality" criterion examines whether the completed regions contain high-quality fine details, while the "Identity" criterion evaluates the model's ability to preserve the identity of the reference region. As shown in Fig. 4, the annotators will judge the results generated by these reference-based methods and report their preference based on the two criteria. Table 2 shows the significant preference on *CompleteMe*. We will provide more visual comparisons in the supplementary material.

### 4.3. Ablation study

**Different Masking Ratios.** We conducted an ablation study to analyze the impact of varying masking ratios between human shape and random mask on our model's performance. Specifically, we experimented with random mask ratios ranging from 0% to 100% and evaluated the results using three metrics: CLIP-I, DINO, and DreamSim. As shown in Table 3, our model achieves the best overall performance at a 50% random mask ratio, obtaining the highest CLIP-I (**97.18**) and DINO (**96.29**) scores and the lowest DreamSim (**0.0419**) score. This indicates that a balanced masking ratio of 50% effectively enhances our model's robustness and ability to handle diverse occlusions, enabling visual and semantic consistency.

Table 4. **Ablation on Different Reference Image Encoder and Effectiveness of Region-focused Attention.** We conduct an ablation study using various image encoders to process reference images. The U-Net encoder consistently outperforms both CLIP [23] and DINOv2 [21] encoders across all perceptual metrics. We further compare the effectiveness of Region-focused Attention Block, which demonstrate the best performance among all comparisons.

| Method | Region-focused | CLIP-I ↑ | DINO ↑ | DreamSim [6] ↓ |
|---|---|---|---|---|
| CLIP Encoder | | 96.96 | 96.06 | 0.0457 |
| DINOv2 Encoder | | 96.20 | 94.30 | 0.0639 |
| U-Net | | 97.05 | 96.17 | 0.0437 |
| Ours (U-Net) | ✓ | **97.18** | **96.29** | **0.0419** |

**Different Reference Image Encoder.** Several recent methods [3, 9, 10, 34] have shown that an additional U-Net can effectively capture fine-grained details from reference images. Paint-by-Example [35] uses a CLIP [23] encoder to extract features from reference images, while AnyDoor [4] employs DINOv2 [21] for the same purpose. In our study, we investigate whether these encoders can effectively learn feature correspondences and alignment across multiple reference images. To do so, we replace our reference U-Net with CLIP and DINOv2 image encoders, using their token features in the cross-attention layer of the Complete U-Net.

As shown in Fig. 5, both CLIP and DINOv2 successfully identify relevant reference regions, but the U-Net demonstrates clear advantages in preserving fine details. Additionally, quantitative results in Table 4 show that U-Net out-

Table 5. **Quantitative Comparison on Our Benchmark for Ablation Study on Different Training Strategies.** "Train Ref U-Net" indicates whether to train the Reference U-Net. "Prompt" means using the text prompt as additional input for the Complete U-Net. "Reference Mask" stands for whether using reference masks for the Region-focused Attention Block.

| Exp. | Train Ref U-Net | Prompt | Reference Mask | CLIP-I ↑ | DINO ↑ | DreamSim [6] ↓ | LPIPS ↓ |
|---|---|---|---|---|---|---|---|
| (a) Freeze U-Net | | | | 96.02 | 95.54 | 0.0513 | 0.0596 |
| (b) Freeze U-Net+Prompt | | ✓ | | 96.13 | 95.48 | 0.0521 | 0.0598 |
| (c) Freeze U-Net+Prompt+Ref Mask | | ✓ | ✓ | 97.02 | 96.08 | 0.0444 | 0.0600 |
| *CompleteMe* (Ours) | ✓ | ✓ | ✓ | **97.18** | **96.29** | **0.0419** | **0.0588** |

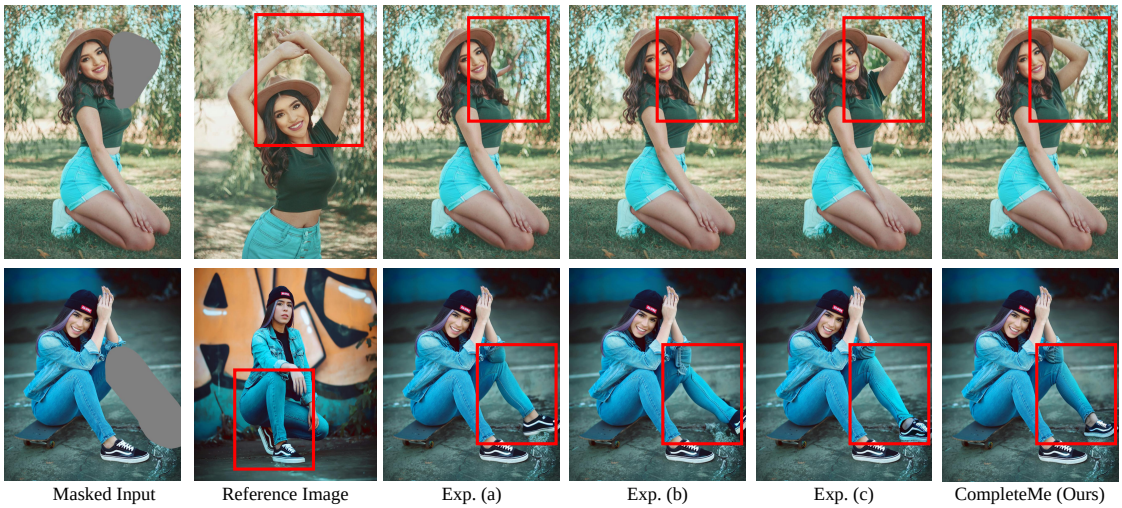


Figure 6. **Qualitative Comparison on Different Training Strategies.** The experimental index follows configurations in Table 5. The Red box highlights the finely detailed regions where different models exhibit varying performance based on distinct training strategies.

performs CLIP and DINOv2 on all evaluation metrics. The Reference U-Net encoder provides multi-level representations at higher resolutions, and its feature space aligns naturally with the Complete U-Net, leading to improved results as a reference feature extractor.

**Effectiveness of Region-focused Attention.** We conducted an ablation study to investigate the effectiveness of our proposed Region-focused Attention (RFA) mechanism. As shown in Table 4, integrating our proposed RFA mechanism with the U-Net encoder further enhances performance, yielding the highest CLIP-I (**97.18**) and DINO (**96.29**) scores and the lowest DreamSim (**0.0419**). This clearly demonstrates that the RFA effectively captures detailed correspondences and enhances semantic coherence by explicitly focusing attention on relevant masked regions.

**Different Training Strategies.** We conduct the ablation study to verify the training strategy and different training input sources. We validate the ablation study on the following three aspects: 1) whether to train the Reference U-Net, 2) text prompt input for Complete U-Net, and 3) reference mask for the Region-focused Attention Block.

Table 5 presents the results of our ablation study, demonstrating that *CompleteMe* achieves the highest evaluation scores across all metrics, showing its robustness and effectiveness in the reference-based human image completion task. To further illustrate the impact of our design choices, we provide visual comparisons in Fig.6, showing how each variation affects the quality of generated images. These visuals highlight the strengths of *CompleteMe* in preserving fine details, maintaining identity consistency, and achieving high-quality completions, underscoring the contributions of each component in our model architecture.

## 5. Conclusion

In this paper, we propose *CompleteMe*, a novel reference-based human image completion framework explicitly designed to reconstruct missing regions in human images with high fidelity, detail preservation, and identity consistency. Our approach employs a dual U-Net architecture consisting of a Reference U-Net and a Complete U-Net integrated with our Region-focused Attention (RFA) Block, which explicitly guides attention toward relevant regions in reference images, thus significantly enhancing spatial correspondence and detailed appearance during completion. Extensive experiments on our benchmark demonstrate that *CompleteMe* outperforms state-of-the-art methods, both reference-based and non-reference-based, in terms of quantitative metrics, qualitative results and user studies. Particularly in challenging scenarios involving complex poses, intricate clothing patterns, and distinctive accessories, our model consistently achieves superior visual fidelity and semantic coherence.

## References


[1] Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu. Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In *CVPR*, 2024. 2, 3, 5, 6, 7, 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 4, 5

[3] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. In *NeurIPS*, 2024. 1, 2, 3, 4, 5, 6, 7

[4] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 2, 3, 4, 5, 6, 7, 1

[5] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. In *ECCV*, 2024. 1

[6] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2024. 4, 5, 7, 8

[7] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *ACM MM*, 2023. 1

[8] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010. 4, 5

[9] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 1, 4, 7

[10] Zehuan Huang, Hongxing Fan, Lipeng Wang, and Lu Sheng. From parts to whole: A unified reference framework for controllable human image generation. *arXiv preprint arXiv:2404.15267*, 2024. 4, 5, 7

[11] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM TOG*, 2022. 5

[12] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *ECCV*, 2024. 1, 2, 5, 6

[13] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 1

[14] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024. 1

[15] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Nannan Li, Qing Liu, Krishna Kumar Singh, Yilin Wang, Jianming Zhang, Bryan A Plummer, and Zhe Lin. Unihuman: A unified model for editing human images in the wild. In *CVPR*, 2024. 4, 1

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 4, 1

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 4, 1

[19] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 5

[20] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACM MM*, 2023. 1

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 4, 7

[22] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *CVPR*, 2024. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 5, 7, 2

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4, 5, 2

[25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[26] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *CVPR*, 2023. 2, 3, 4

[27] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *CVPR*, 2024. 2, 3, 4

[28] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, 2023. 1

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 4, 5

[30] Xian Wu, Rui-Long Li, Fang-Lue Zhang, Jian-Cheng Liu, Jue Wang, Ariel Shamir, and Shi-Min Hu. Deep portrait image completion and extrapolation. *TIP*, 2019. 1, 2

[31] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023. 2

[32] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *CVPR*, 2019. 2

[33] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *CVPR*, 2024. 2

[34] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 2024. 1, 4, 7

[35] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023. 2, 4, 5, 6, 7, 1

[36] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 4

[37] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 1

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 1

[39] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *CVPR*, 2024. 2

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 4, 5

[41] Hengrun Zhao, Yu Zeng, Huchuan Lu, and Lijun Wang. Large occluded human image completion via image-prior cooperating. In *AAAI*, 2024. 1, 2, 5, 6

[42] Zibo Zhao, Wen Liu, Yanyu Xu, Xianing Chen, Weixin Luo, Lei Jin, Bohui Zhu, Tong Liu, Binqiang Zhao, and Shenghua Gao. Prior based human completion. In *CVPR*, 2021. 2

[43] Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. Human de-occlusion: Invisible perception and recovery for humans. In *CVPR*, 2021. 1

[44] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In *CVPR*, 2021. 2