AddMe: Zero-shot Group-photo Synthesis by Inserting People into Scenes

Dongxu Yue[1,2] ⋆, Maomao Li[1], Yunfei Liu[1], Qin Guo[2], Ailing Zeng[1], Tianyu Yang[1], and Yu Li[1] ✉

[1]International Digital Economy Academy (IDEA), [2]Peking University
http://addme-awesome.github.io/page/

Fig. 1: We present AddMe, a framework for adding a new portrait to any position in an image using just one reference face and user-provided masks.

Abstract. While large text-to-image diffusion models have made significant progress in high-quality image generation, challenges persist when users insert their portraits into existing photos, especially group photos. Concretely, existing customization methods struggle to insert facial identities at desired locations in existing images, and it is difficult for existing local image editing methods to deal with facial details. To address these limitations, we propose `AddMe`, a powerful diffusion-based portrait generator that can insert a given portrait into a desired location in an existing scene image in a zero-shot manner. Specifically, we propose a novel identity adapter to learn a facial representation decoupled from existing characters in the scene. Meanwhile, to ensure that the generated portrait can interact properly with others in the existing scene, we design an enhanced portrait attention module to capture contextual information during the generation process. Our method is compatible with both text and various spatial conditions, enabling precise control over the generated portraits. Extensive experiments demonstrate significant improvements in both performance and efficiency.

⋆ This work was done during Dongxu's internship at IDEA. ✉ Corresponding author.

1 Introduction

Can you imagine being able to take photos with your favorite celebrities at film festivals or travel around the world, capturing iconic landmarks? What about unexpectedly missing a gathering with friends but still being able to appear in the group photo? Personalized portrait insertion has a variety of applications, such as virtual group photo capturing, AI portraits, virtual try-on, *etc*. However, adding customized characters into existing scene images still remains a challenging task.

Early exploration [9] inserts given persons into existing images by three handcrafted stages: generating a semantic map for the new person, rendering the appearance, and refining the generated face. Nevertheless, it struggles to maintain the identity of the reference image and produce unrealistic results since unstable training of GANs [13]. Thanks to the evolution of generative models, the recent developments in text-to-image (T2I) with diffusion models have made it possible to generate high-quality new content [7, 16, 36, 37, 40, 42, 44]. To insert specified new concepts into a T2I model and enable the model for subject-driven generation, a group of methods [10, 24, 39, 49] tries to add special tokens to T2I model for inserting the given concept. Despite personalized image generation has achieved high-quality results, it requires massive computing resources and time-consuming fine-tuning. More recently, FastComposer [50] and IP-Adapter [55] opt to use an image encoder to encode reference images and align them with text embeddings. Although these diffusion-based customization methods can generate realistic customized portraits, they cannot insert the specified person into the desired location in a given scene, especially in group photos.

In this work, we investigate the topic of personalized image composition. Specifically, as illustrated in Fig. 1, given a masked target scene image and a reference face, we aim to incorporate this person into the masked area of the existing group photo while robustly controlling the appearance of the generated portrait. To clarify, we frame our problem as personalized conditional inpainting. Similar to this, Paint-by-Example [52] takes a reference image as a template and edits a specific area of the target image. However, it cannot generate consistent content with reference identity. AnyDoor [5] enhances the consistency between the generated region and the reference image through a more dense encoding. However, it struggles to generate a personalized portrait from the reference face. Moreover, both of them do not support text or other control conditions.

To address the above issues, we propose AddMe, a tuning-free and plug-and-play solution, which can generate high-quality and ID-consistent compositions at desired locations within seconds, while providing controls over the generated portraits. Specifically, we introduce a novel disentangled identity adapter running in parallel with the text condition to preserve facial identity from the reference image. Through our decoupling design, we avoid the blending of extracted identity with other characters in the existing scene image, thus enhancing the fidelity of the facial appearance in the desired location. Besides, we introduce an enhanced portrait attention (EPA) module and train it on multi-person scene data to capture reasonable person interactions. We then connect EPA through residual connections to make ID-Adapter perceive both global semantics and human

interactions, thereby forming a plug-and-play human essence module (HEM). Additionally, we apply identity-centric data pre-processing and augmentation techniques for the efficiency of self-supervised training.

Equipped with these techniques, as shown in Fig. 1, AddMe demonstrates extraordinary customization capabilities with just one reference image and achieves state-of-the-art results, showing a significant quality advantage over prior works in a similar setting. Note that AddMe not only performs single-person completion but also extends to multi-subject compositions easily (see the last column of Fig. 1). In summary, our contributions are as follows:

- We introduce a new image editing scenario, *i.e.*, group-photo synthesis by inserting personalized portraits into existing scenes, which requires controls over the generated content through both text and spatial conditions.
- We propose AddMe, an innovative plug-and-play ID-preserving inpainting method, which deals with group-photo synthesis by a disentangled identity adapter and enhanced portrait attention.
- We construct a large-scale human face dataset with instance-level annotations comprising 1.6 million data pairs, and a new evaluation benchmark for measuring the performance of ID-preserved personalized local image editing.
- Extensive experiments demonstrate the excellent performance and efficiency of AddMe, which outperforms existing methods consistently. AddMe can also serve as a general model that can be controlled by ControlNet.

2 Related Work

Text-to-image Diffusion Models. In recent years, text-to-image (T2I) generation has gained more and more attention in computer vision community. Denoising Diffusion Probabilistic Model (DDPM) [17] and its variant Denoising Diffusion Implicit Model (DDIM) [43] are widely used for T2I generation. Based on CLIP [35] text embedding, DALL-E [36, 36] and CogView [8] operate diffusion process on pixel space and train an image generative pre-training transformer to perform T2I. Then, Stable Diffusion [37] proposes to train diffusion models on the latent space of powerful pretrained autoencoders, leading to lower computational complexity.

Customized Image Generation. Customized image generation inserts the new personalized subject into the base T2I model and make the model generate content containing the new concept. DreamBooth [39] and Textual Inversion [11] achieve this by fine-tuning the parameters of the T2I models. Then, FastComposer [50] performs multi-subject customization without fine-tuning, which uses subject embeddings extracted by an image encoder to augment the generic text conditioning. Besides, IP-adapter [55] achieves image prompt capability for the pretrained T2I models by separating cross-attention layers for text features and image features. Further, Face0 [46] and PhotoMaker [29] adopt a similar approach, but enhance performance through different face embedding techniques.

Local Image Editing. Existing local image editing methods can be classified into two categories: text-driven and exemplar-based. [1, 33] combine Style-

GAN [21] with CLIP and edit the latent code or the feature map based on the input mask. Recently, diffusion based text-driven inpainting [2,3,28,32,34,37,51,56] has gained widespread attention. Exemplar-based local image editing is a relatively new topic, some works [9, 23, 26, 27, 30, 45, 58] aim to restore the coarse appearance of the reference image into the scene image without focusing on its identity. Paint-by-Example [52] makes the first attempt by replacing the text condition of the T2I diffusion model with an image condition. AnyDoor [5] employs a detail extractor designed to maintain texture details yet allow versatile local variations. A concurrent work, Unipaint [53] enables the T2I diffusion model to be conditioned on both reference images and text through minutes of test-time fine-tuning. However, these methods struggle to generate customized images while accurately preserving the intricate identity details of human subjects. Different from these methods, our AddMe requires no fine-tuning during inference, and seamlessly injects ID-preserved portrait in existing target scene images while robustly controlling the appearance of the generated portrait.

3 Preliminary

Stable Diffusion. We build our method on top of Stable Diffusion [37], which is a latent diffusion model that integrates diffusion processes in the low-dimensional latent space of a variational autoencoder(VAE) [22]. The denoising network of Stable Diffusion adopts a U-Net [38] architecture, consisting of 16 main blocks that include residual convolutional layers, as well as self-attention and cross-attention modules [47]. Formally, for an input image $x_0 \in \mathbb{R}^{H\times W\times 3}$, the VAE encoder transforms it into a latent representation $z_0 \in \mathbb{R}^{h\times w\times c}$, and the diffusion process then operates in the latent space. The U-Net denoiser is used to predict noise $\hat{\epsilon}_t = \epsilon_\theta\left(z_t; t, C\right)$ from the noisy latent z_t at time step t, given text condition C, where $z_t = \alpha_t z_0 + \sigma_t \epsilon$ and α_t, σ_t are predefined functions of t that determine the diffusion process. The training objective is then defined as:

$$\mathcal{L}_{sd}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(1,T),\epsilon_t\sim\mathcal{N}(0,I)}\left[\left\|\epsilon_t - \epsilon_\theta\left(z_t, t, C\right)\right\|^2\right]. \tag{1}$$

Attention in Stable Diffusion. The U-Net denoiser includes two types of attention mechanisms: self-attention and cross-attention. Self-attention can control layout and texture details when generating images [53]. Meanwhile, the scores in cross-attention maps represent the amount of information from the text token to a latent pixel. Formally, for a given query feature Z and the text feature c_t encoded by CLIP [35], the cross-attention mechanism can be expressed as:

$$Z_{text} = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V, \tag{2}$$

where $Q = ZW_q$, $K = c_t W_k$ and $V = c_t W_v$ are the query, key and values, d is the dimension of embedding and W_q, W_k, W_v are the weight matrices.

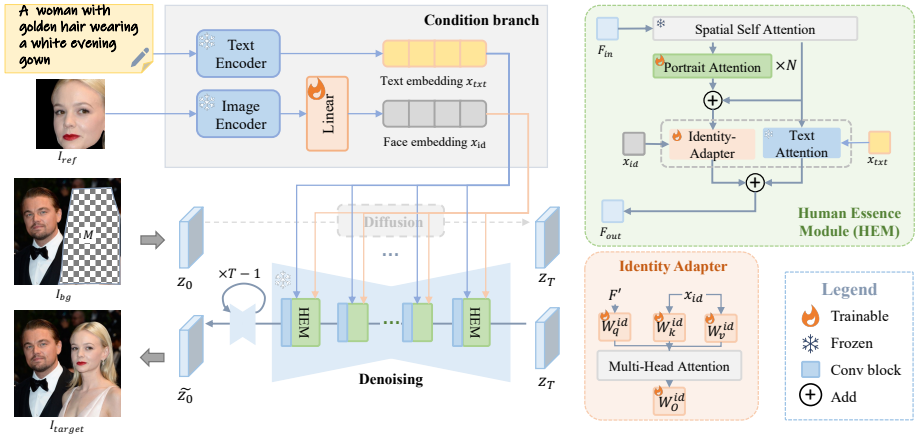

Fig. 2: Framework of the proposed AddMe. Based on the latent diffusion model, AddMe consists of a condition branch and a denoising U-Net. We design a disentangled Identity-Adapter to modulate the face embedding extracted from the reference ID image I_{ref}. Moreover, we propose an Enhanced Portrait Attention (EPA) module to capture reasonable person interaction. We then connect the ID-Adapter and EPA through residual connections to form a plug-and-play Human Essence Module (HEM).

4 Method

4.1 Problem Formulation of Group-photo Synthesis

Given the reference ID image I_{ref}, the target scene image I_{target}, the position mask M, and the text description T, our goal is: (i) generating a portrait in the M region of I_{target} that is consistent with the identity of I_{ref} and the text prompt T; (ii) making the generated character have the proper and natural body interactions with characters and contextual information in $I_{target} \odot \overline{M}$, where $\overline{M} = \mathbf{1} - M$ and $\mathbf{1}$ is the all-ones matrix.

The rest of this section is organized as follows. In Section 4.2, we first introduce the model design of the proposed group-photo synthesizer AddMe. Then, we provide a detailed explanation of our training strategy and data augmentation techniques in Section 4.3. Finally, we present our collected multi-modal instance-level human face dataset for our task in Section 4.4, which we believe would be a new benchmark for the community.

4.2 Model Designs

The pipeline of our proposed group-photo synthesizer AddMe is demonstrated in Fig. 2. It comprises two key components: (1) a lightweight disentangled identity adapter, which is used to modulate the reference face of I_{ref} and map it to an identity-consistent character portrait, and (2) an enhanced portrait attention (EPA) module that ensures the generated character portraits to have proper

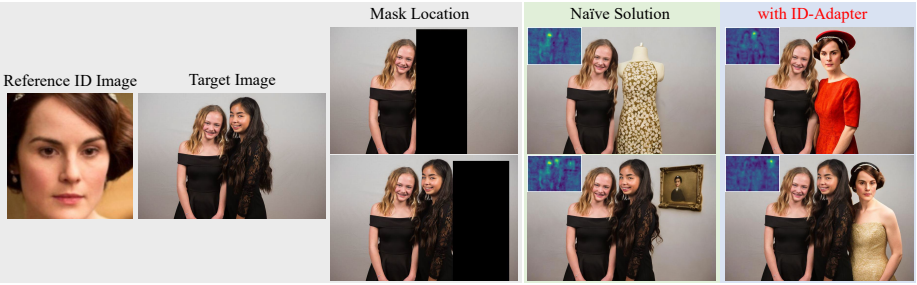

Fig. 3: Illustration of the identity leakage issue of the naive solution. We show the average cross-attention maps of ID tokens at the left-top corner. We also demonstrate our disentangled ID-Adapter makes the model focus on the editing region.

poses and clothing that harmonize with other characters in the context while seamlessly integrating the foreground and background. We then connect them through residual connections to form a plug-and-play human essence (HEM) module for T2I diffusion inpainting models.

A Naïve Solution. A naive solution is to directly integrate a customization module (e.g., IP-Adapter [55]) into the pretrained text-guided image inpainting pipeline. Here, IP-Adapter introduces a new cross-attention dealing with "image prompt". Specifically, given the query Q in the cross-attention of the diffusion model and encoded reference ID image features c_i, additional cross-attention outputs are obtained through the newly introduced W_k^{ip} and W_v^{ip} as

$$Z_{ip} = \text{Attention}\left(Q, K^{ip}, V^{ip}\right), \tag{3}$$

where $K^{ip} = c_i W_k^{ip}$, $V^{ip} = c_i W_v^{ip}$ are key and values for image cross-attention. Note that only W_k^{ip} and W_v^{ip} are trainable parameters, while the rest parameters are shared with the text cross-attention. Z_{ip} then add to Z_{text} to obtain the final output of cross-attention layer.

As illustrated in the fourth column of Fig. 3, although such a naive approach can roughly perform single-person replacement within the target scene image I_{target}, the newly introduced image attention would leak into other facial regions when there are multiple characters in I_{target}. That is, when generating a portrait, the model tends to complete the specified region with a contextual background rather than referring to the reference ID image I_{ref}.

We argue that the identity leakage issue when multiple individuals are involved in customized image completion can be attributed to two reasons. (i) The newly introduced cross-attention is coupled with text cross-attention designed for controlling global semantics. (ii) The pre-trained text-guided image inpainting model lacks prior knowledge to generate full or half-body character portraits when a reference ID image are given.

Disentangled Identity Adapter. To address the identity leakage issue, we introduce an ID-Adapter with disentangled multi-modal cross-attention. Specifically, we utilize a pre-trained CLIP [35] image encoder to encode the reference

ID image I_{ref} into 257 ID tokens to obtain a dense representation of identity. To seek an alignment in the representation space, we add several linear layers to encode the ID token as face embedding c_i, which is later used for decoupled cross-attention. The operation in our cross-attention can be defined as:

$$\begin{cases} Q^{id} = W_q^{id} \cdot Z; K^{id} = W_k^{id} \cdot c_i; V^{id} = W_v^{id} \cdot c_i, \\ Z_{\text{out}} = \text{Attention}\left(Q, K, V\right) W_O + \text{Attention}\left(Q^{id}, K^{id}, V^{id}\right) W_O^{id}. \end{cases} \tag{4}$$

This approach allows the model to effectively control the generated content by treating I_{ref} as the ID prompt alongside the text prompt. Note that we freeze the original U-Net denoiser in Stable Diffusion, and only the linear layers, W_q^{id}, W_k^{id}, W_v^{id}, and W_O^{id} are trainable in our disentangled identity adapter.

Enhanced Portrait Attention (EPA). Although the identity adapter can inject the identity of the reference ID image I_{ref} into specified locations accurately, the model still lacks an understanding of interaction with other characters in $I_{target} \odot \bar{M}$, such as pose, clothing and scale. It would lead to an unnatural fusion of the generated portrait with existing characters. We argue that this drawback stems from the inherent training mechanisms and properties of the pre-trained Stable Diffusion. Specifically, given a reference ID image I_{ref} as an ID prompt, the model tends to learn a trivial mapping function to only reconstruct the face in I_{ref}. That is, instead of understanding the interactions between characters, the model tends to focus on the identity of I_{ref} merely.

To deal with this challenge, we introduce an enhanced portrait attention module to encourage the model to understand contextual information. Specifically, as illustrated in the Fig. 2, we opt to separately train a vanilla transformer as our portrait modeling module in each main block of the U-Net denoiser. The transformer consists of one or more portrait attention blocks. Given the spatial feature map Z, portrait attention block is formulated as:

$$Z_h = \text{Attention}(Q^h, K^h, V^h), \tag{5}$$

where $Q^h = W_q^h \cdot Z$, $K^h = W_k^h \cdot Z$, $V^h = W_v^h \cdot Z$. The output of our enhanced portrait attention module will then be added to the output of the existing self-attention layer as the query feature for our proposed disentangled identity cross-attention in Eq. 4:

$$Q^{id} = W_q^{id} \cdot (Z + Z_h). \tag{6}$$

Note that the cross-attention query feature for text remains unchanged. Inspired by ControlNet [57], to introduce our EPA module during training without harmful effects, we zero-initialize the output projection layer of the proposed module to enable the newly added modules gradually.

4.3 Training and Inference

Self-supervised Training. Recall the problem formulation in Section 4.1, it is difficult to collect annotated paired data for a reference ID and a target image with mask location. Therefore, we turn to self-supervised training. Specifically,

Fig. 4: Example illustration of the proposed AddMe-1.6M dataset.

given an image I_{target}, its corresponding text description T, and bounding boxes M for each human in the image, we use the cropped face within M as the reference ID image I_{ref} for each character, with the bounding box serving as the location mask. Naturally, the training objective is the original image, and our training data consists of $\left\{\left(I_{ref}, \overline{M} \odot I_{target}, M\right), T, I_{target}\right\}$. During the training process, the denoiser takes both text T and the reference ID image I_{ref} as conditions. Therefore, our training objective can be formulated as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(1,T),\epsilon_t\sim\mathcal{N}(0,I)}\left[\left\|\epsilon_t - \epsilon_\theta\left(z_t, \overline{M} \odot I_{target}, M, t, T, I_{ref}\right)\right\|^2\right]. \quad (7)$$

It is worth noting that we also randomly drop image and text conditions in this stage to enable classifier-free guidance [18] during inference.

Data Augmentations. During self-supervised training, our reference ID image I_{ref} is derived from the target scene image I_{target}, implying a risk of the model learning a simple mapping function: $\overline{M} \odot I_{target} + I_{ref} + b = I_{target}$, where b represents the areas filled in M except for I_{ref}. Therefore, we apply several data augmentation techniques to I_{ref} (including flip, rotation, brightness, contrast, and blur) to enhance the generalization capability of the model, especially in generalizing the reference ID image to an appropriate scale. Further, we generate masks of arbitrary shapes based on the bounding box of the character to achieve fine-grained control over the generated portrait such as height and body shape. Next, constrained by the resolution in the training strategy of Stable Diffusion, we crop the images in the training data. Conventional cropping methods may risk damaging the region corresponding to I_{ref}, which is harmful to the model as it learns incorrect identity correspondences. To deal with it, we design a new cropping method based on the geometric center of I_{ref} in I_{target}, then we expand the cropping area to ensure that I_{ref} and the characters in I_{target} is preserved during training. More details are in Appendix A.

4.4 Dataset with Instance Annotation: AddMe-1.6M

Existing multimodal human face datasets can roughly ensure the presence of faces in the images rather than detailed human positions and facial locations for each person. Driven by the proposed customized editing task, we create a large-scale high-quality multimodal human face dataset with instance-level annotations, which can be used for various customized portrait generation tasks, as shown in Figure 4. Our raw data is filtered from COYO-700M [4] and LAION-2B [41] based on resolution and aesthetic score to ensure the image quality.

Subsequently, our process unfolds in three stages. First, we use YOLOv7-Face [1] to detect all faces in each image and count the number of faces. After filtering out data without faces, we generated position and confidence labels for each face in the image, obtaining the cropped face images. Second, we then use YOLOX [12] to generate bounding boxes corresponding to each character and match each bounding box to a face. Last, we employ BLIP-2 [25] to generate text descriptions for each character within the bounding box. This leads to a total number of 2.3M face-position-text pairs, included in 1.6M scene images, which comprise 1M single-person and 600K multi-person scene images. Equipped with instance-level annotations for high-quality facial images, we are able to achieve high-quality customized character generation with fine-grained control. More details of the dataset are shown in the Appendix D.

5 Experiments

5.1 Experimental Setup

Data and Benchmark. We utilize the collected AddMe-1.6M to train our model, where 10,000 images are left for validation and testing. To the best of our knowledge, there is no prior work targeting the composition of character images based on a reference face. Therefore, we constructed a test benchmark for quantitative analysis. To be specific, we manually selected 500 target scene images from the test set, each with a resolution of 512×512 and containing only one position mask. For diversity considerations, we select 20 reference ID images from the FFHQ [20], CelebA-HQ [19], and our AddMe-1.6M. Consequently, we generated 10,000 images for the combination of characters and scenes. We will make this benchmark publicly available to encourage further engagement.

Training Strategy. We adopt a two-stage training. For the first stage, the total trainable parameters included a linear mapping network and our ID-Adapter. Here, we train our ID-Adapter with a mixed dataset of both single-person and multi-person data. In the second stage, we train the enhanced portrait attention module exclusively with multi-person scene data.

Evaluation Metrics. For our newly proposed task, based on CLIP-IQA [48], we constructed a new evaluation metric named CLIP-H by setting prompt pairs: "with person" and "without person", to assess the likelihood of generating portraits within the location mask. We use Arcface [6] to calculate face similarity, where a higher score indicates higher identity similarity. Furthermore, to measure the distribution difference between generated and real images, we calculate the FID [15] between the generated 10,000 images and the test set images. Besides, we calculate QS [14] and CLIP-IQA (quality, natural) to assess the overall aesthetic quality of the edited images. For efficiency evaluation, we consider the total editing time, including fine-tuning and inference, as well as peak memory usage throughout the entire process. We provide detailed descriptions for each metric in the Appendix B.

[1] `https://github.com/derronqi/yolov7-face`

Fig. 5: Various applications of AddMe. Our method supports different types of locaion mask, text prompt control, and spatial control.

5.2 Qualitative Results

We present qualitative results in various settings to demonstrate the robustness, mask controllability, text prompt editability, and compatibility of our method.

Reference ID Image Only. In the most common application scenario, where only a reference ID image is provided along with a location specified by a bounding box mask and the text prompt is set to be empty. During the generation process, the model guides the creation of a new portrait by leveraging the identity information from the reference ID image and the context information perceived by the enhanced portrait attention module from other characters in the scene (e.g., height, body shape, pose, and clothing). As shown in the first row of Fig. 5, our model demonstrates robustness in maintaining the identity, expression, skin tone, and age of the reference ID image. Additionally, our model can generate portraits with proper scales and seamlessly integrated backgrounds based on the contextual information from the target image.

Reference ID Image + Mask Shape Control. Our method is compatible with masks of arbitrary shapes and can then enable fine-grained control over the height, body shape, and relative position of characters in the generated region based on the shape of the mask. As shown in the second row of Fig. 5, using a more detailed mask makes the generated individual appear thinner and shorter than the girl on the left.

Reference ID Image + Text Prompt. Although our model can generate a natural appearance with the target scene image through the EPA module even without a text prompt, we can achieve more precise control over the generated

Fig. 6: Comparison with text-driven and exemplar-based methods on single portrait insertion. For text-driven methods, we use the name of the reference character in the prompt (such as Joe Biden). We also show setting comparison at the bottom.

portraits through text. In this setting, our model can effectively alter appearances while ensuring identity consistency. Besides, it is compatible with mask shape control without any degradation in identity preservation, or image quality, as shown in the row 3-4 of Fig. 5.

Reference ID Image + Spatial Control. Integrating our method with pretrained tools such as ControlNet [57] and T2I-Adapter [31] achieves more precise control. The last row of Fig. 5 demonstrates control via human pose [54].

5.3 Comparisons

Baseline Models. To the best of our knowledge, there is no prior work on mutimodal personalized character image editing based on both a reference ID image and text. We selected two relevant categories of methods as baselines for our approach, namely exemplar-based and text-driven local image editing, and we demonstrate the difference between our method and the baselines at the bottom of Fig. 6. We make comparisons with the state-of-the-art (SOTA) exemplar-based image editing methods including PbE [52], AnyDoor [5], and Unipaint [53]. PbE replaces the text encoder of SD with an image encoder and realize image editing conditioned on images through fine-tuning. AnyDoor transfers the reference image into the target image through a more dense representation, representing generative image composition methods. Unipaint attempts to capture contextual information through test-time fine-tuning, aiming to generalize the model to support for multi-conditions. Wish You Were Here [9] and Putting People in Their Place [23] lack available implementations, precluding direct comparison.

Qualitative Comparisons. In columns 4-6 of Fig. 6, we present visual comparisons with the previous exemplar-based methods. The results show that these

Table 1: Quantitative comparisons between our method and baseline approaches. We evaluate the edited region using CLIP-H and Face Sim, while FID, QS, and CLIP-IQA (referred to as IQA) are used to assess the overall image quality after editing.

Models	Local		Global			Efficiency	
	CLIP-H(%)↑	Face Sim.(%)↑	FID ↓	QS↑	IQA(%)↑	Time↓	Memory↓
PbE [52]	61.82	15.16	28.73	**14.83**	53.28	2.94s	10.6GB
AnyDoor [5]	64.83	14.21	50.62	7.45	54.00	6.55s	19.2GB
UniPaint [53]	57.35	37.53	41.42	4.16	52.71	64.93s	36.7GB
AddMe(Ours)	**65.52**	**60.15**	**26.11**	12.09	**55.81**	**2.44s**	**6.51GB**

methods can only maintain semantic consistency at the category level, generating an individual with a similar appearance in the edited region. PbE achieves competitive image fusion through CLIP-compressed representations but loses identity information and generates incorrect skin color. AnyDoor tends to copy and paste the reference image into the editing region, failing to extend the reference face to a portrait. Unipaint has noticeable boundary artifacts in the editing region, which may be caused by the flow control in self-attention.

We also make a comparison with text-driven local image editing methods, where they can only control the semantic content of the region through textual prompts. Due to different task settings, we use a text prompt similar to the reference ID image here as input to make a comparison. As shown in column 2-3 of Fig. 6, Blended Latent Diffusion (BLD) [2] utilizes CLIP to provide gradients to guide the diffusion sampling process. The results show that BLD exhibits noticeable artifacts in generating photorealistic human images, which may be due to information bottlenecks caused by the gradient-guided strategy. When compared to Stable Diffusion, we used its inpainting version. As illustrated in Fig. 6, column 3, SD can generate more realistic results, but it is limited for text that is not informative to express complex scenes or concepts. In contrast, our method achieves photorealistic results, generating portraits in the desired location that interact naturally with the context while ensuring identity similarity to the reference ID image.

Quantitative Comparisons. As shown in Table 1, our method achieves state-of-the-art results in terms of CLIP-H and Face Similarity, demonstrating that our approach is able to generate portraits in the desired location and best preserves the identity information of the reference ID image. Notably, our method shows improved performance over baseline models in FID and CLIP-IQA, demonstrating the superiority of our method in the overall quality and harmony of the image. It is noteworthy that our method exhibits significant performance improvements compared to the baseline while maintaining the highest efficiency, indicating better device compatibility. Given that our baseline method cannot support both images and text as input, we place the quantitative comparison of text alignment in the Appendix C.

User Study. We conducted a user study to compare with PbE, AnyDoor, Unipaint. We have 30 participants with diverse backgrounds to rate 20 sets of images.

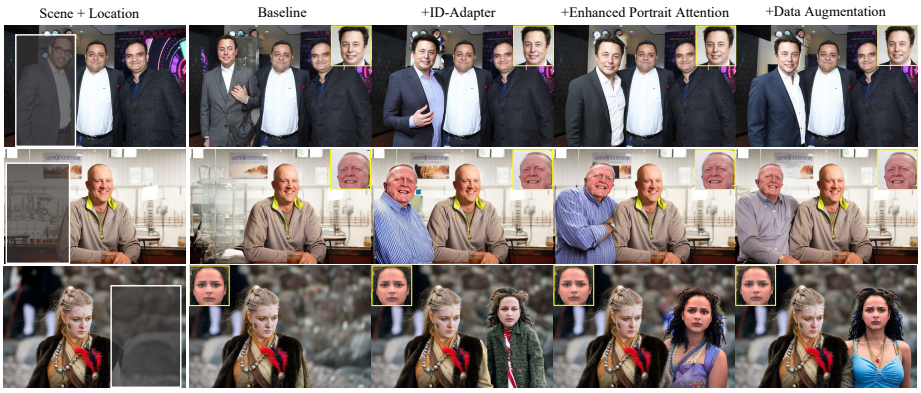

Fig. 7: Visual ablation studies of individual components in our method. From left to right, we progressively make the generation of the portrait more reasonable, while improving facial similarity and image fidelity.

Table 2: User study on the comparison between our AddMe and baseline methods.

Settings	PbE [52]	AnyDoor [5]	Unipaint [53]	AddMe (Ours)
Face Sim.	5.95%	4.77%	5.95%	**83.33%**
Quality	1.20%	13.10%	3.55%	**82.15%**
Natural	9.55%	2.40%	7.10%	**80.95%**

For each set, we provided a reference ID image and 5 target scene images, and each of the four models generates 5 predictions. Participants are asked to choose the best model based on identity fidelity, image quality and naturalness of image fusion. As shown in Table 2, our proposed AddMe is preferred by more participants over baseline methods in all four metrics, especially in terms of identity fidelity, where over 80% of participants perceive our method to have a more similar identity to the reference ID image. This underscores the superiority of our AddMe in terms of identity fidelity and image fusion.

5.4 Ablation Study

We conduct ablation studies to validate the effectiveness of three key techniques of this paper: ID-Adapter, Enhanced Portrait Attention, and Data Augmentation. We use the naive solution mentioned in Section 4.2 as the baseline. We show visual ablation results in Fig. 7, which includes character replacement and generation scenarios. The baseline solution struggles to maintain the identity information of the reference ID image and tends to use the background to complete the editing region. We argue that this is due to cross-attention leakage in multi-person scenes, as depicted in Fig. 3.

With our proposed ID-Adapter, the ID prompt can be correctly injected into a local editing region, significantly increasing the likelihood of generating individuals with improved identity fidelity, as also demonstrated in Table 3. However,

Table 3: Quantitative ablation studies on different variants of our method. We achieve the best performance by leveraging all these techniques.

Settings	Local		Global		
	CLIP-H(%) ↑	Face Sim.(%) ↑	FID ↓	QS ↑	IQA(%) ↑
Baseline	58.42	29.03	34.87	7.04	54.01
+ID-Adapter	64.87	56.79	35.45	6.68	52.90
+EPA Module	65.39	59.73	27.16	11.13	55.67
+Data Augmentation	**65.52**	**60.15**	**26.11**	**12.09**	**55.81**

the ID-Adapter encounters issues of unnatural image fusion. We further address this by introducing the EPA module to make the generated portraits interact more naturally with other characters in the context, leading to improved image quality based on better FID, QS, and CLIP-IQA. Finally, we utilize data augmentation techniques to avoid the copy-and-paste issue in generating portraits.

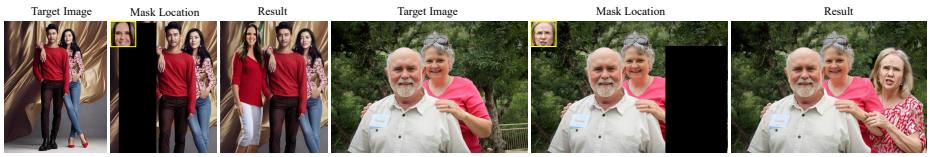

Fig. 8: Limitations of AddMe. Left: When the generated face is small, the fidelity of the identity is affected. Right: We use the prompt: "A woman with black hair is smiling", but facial attributes could not be precisely controlled.

6 Conclusion and Limitation

In this paper, we introduce a novel image editing scenario: group photo synthesis, which aims to insert identity-preserved portraits into a desired location in an existing scene image while ensuring control over the generated content. We achieve this goal by two major upgrades to text-to-image diffusion model. First, an ID-Adapter is designed to learn a facial representation, which is decoupled from existing characters in the scene. Second, an enhanced portrait attention is proposed to ensure the generated portrait interacts naturally with others in the existing scene. Our method enables users to precisely control the editing, and achieves an impressive performance on in-the-wild images.

While our work shows some promising results, there are still some limitations, as illustrated in Fig. 8. For example, there is a noticeable disparity in the generation where the proportion of the face is relatively small compared to larger ones, which we believe is constrained by the limitations of the VAE. Moreover, facial attributes such as hair color are coupled with the ID token, which may pose challenges for facial attribute editing.

Acknowledgements

We thank Changyin Zhou and Fei Yu from Vozo.ai for their contributions to the initial proposal and discussions of this project. We regret their names were not included as co-authors and hope this acknowledgment reflects our appreciation.

References

1. Andonian, A., Osmany, S., Cui, A., Park, Y., Jahanian, A., Torralba, A., Bau, D.: Paint by word. arXiv preprint arXiv:2103.10951 (2021)
2. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)
3. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
4. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022)
5. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
8. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems **34**, 19822–19835 (2021)
9. Gafni, O., Wolf, L.: Wish you were here: Context-aware human generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7840–7849 (2020)
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
12. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
14. Gu, S., Bao, J., Chen, D., Wen, F.: Giqa: Generated image quality assessment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 369–385. Springer (2020)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

17. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
19. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., Efros, A.A., Singh, K.K.: Putting people in their place: Affordance-aware human insertion into scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17089–17099 (2023)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion (2023)
25. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
26. Li, T., Ku, M., Wei, C., Chen, W.: Dreamedit: Subject-driven image editing. arXiv preprint arXiv:2306.12624 (2023)
27. Li, Y., Liu, H., Wen, Y., Lee, Y.J.: Generate anything anywhere in any scene. arXiv preprint arXiv:2306.17154 (2023)
28. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
29. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. arXiv preprint arXiv:2312.04461 (2023)
30. Lu, L., Zhang, B., Niu, L.: Dreamcom: Finetuning text-guided inpainting model for image composition. arXiv preprint arXiv:2309.15508 (2023)
31. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
32. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
33. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
34. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
39. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
45. Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Aliaga, D.: Objectstitch: Object compositing with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18310–18319 (2023)
46. Valevski, D., Lumen, D., Matias, Y., Leviathan, Y.: Face0: Instantaneously conditioning a text-to-image model on a face. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
48. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)
49. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
50. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023)
51. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023)

52. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)
53. Yang, S., Chen, X., Liao, J.: Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3190–3199 (2023)
54. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. pp. 4210–4220 (2023)
55. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
56. Zhang, B., Duan, Y., Lan, J., Hong, Y., Zhu, H., Wang, W., Niu, L.: Controlcom: Controllable image composition using diffusion model. arXiv preprint arXiv:2308.10040 (2023)
57. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
58. Zhang, X., Guo, J., Yoo, P., Matsuo, Y., Iwasawa, Y.: Paste, inpaint and harmonize via denoising: Subject-driven image editing with pre-trained diffusion model. arXiv preprint arXiv:2306.07596 (2023)