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

Dongxu Yue<sup>1,2 \*</sup>, Maomao Li<sup>1</sup>, Yunfei Liu<sup>1</sup>, Qin Guo<sup>2</sup>, Ailing Zeng<sup>1</sup>, Tianyu Yang<sup>1</sup>, and Yu Li<sup>1</sup>  $\boxtimes$ 

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

The appendix is organized as follows. In appendix A, we present more implementation details of our method. Then, in appendix B, we show details of our evaluation metrics and effect of mask size in our AddMe. Next, we display more experimental results in appendix C. Last, we give more information on the dataset construction and the corresponding statistics in appendix D.

# A Implementation Details

We train our model based on the Stable Diffusion inpainting model <sup>1</sup>, using OpenCLIP ViT-H/14 [4] as the image encoder. We incorporate the ID-Adapter in all 16 major blocks of the SD model, adding it in the cross-attention layer and including the portrait attention module after each self-attention layer. Our model is trained on a single machine with 8 NVIDIA A100 GPUs (80GB), with a batch size of 16 per GPU for 500k steps. During training, we resize the shortest side of the image to 512 and then process the image to a resolution of  $512 \times 512$ . The trainable parameters in the first and second stages are 46.58M and 198.42M, respectively. Our model is based on SD 1.5. To deal with arbitrary mask shapes, we utilize both bounding box masks and masks of arbitrary shapes with a probability of 0.5 each. It should be noted that we follow the same training details in the naive solution.

### A.1 Inference

During inference, our model requires a reference ID image  $I_{ref}$ , a target scene image  $I_{target}$ , a square or arbitrarily shaped mask M, optional text prompts T, and spatial control conditions such as pose or depth map. Following previous image editing methods [3,7,14,17], we adopt DDIM sampler [11] with 50 steps, and set the guidance scale to 7.5.

**Classifier-free Guidance.** To enable classifier-free guidance [6], during training, we randomly drop the text or reference image condition with a probability of 0.05, and simultaneously drop both text and reference image conditions with

<sup>\*</sup> This work was done during Dongxu's internship at IDEA.

<sup>&</sup>lt;sup>1</sup> https://huggingface.co/runwayml/stable-diffusion-inpainting

a probability of 0.05. The classifier-free guidance during the inference phase can be represented as:

$$\hat{\epsilon}_{\theta}\left(z_{t}, T, I_{ref}, t\right) = w\epsilon_{\theta}\left(z_{t}, T, I_{ref}, t\right) + (1 - w)\epsilon_{\theta}\left(z_{t}, t\right), \tag{1}$$

where w is the guidance scale. If the image condition is dropped, we set  $I_{ref}$  to zero.

### A.2 Mask Shape Augmentation

We aim not only to specify the location of the portrait through the mask but also to achieve fine-grained control over the generated portrait such as height and body shape using masks of arbitrary shapes. To achieve this, during the training phase, we generate masks of arbitrary shapes based on the bounding box of the character. Specifically, following PbE [15], we use Bessel curves to fit the bounding box and uniformly sample 20 points on this curve. Finally, these points are connected with straight lines to form an arbitrary-shape mask.

## **B** Experimental Details

In this section, we first explain the evaluation metrics in Section 5.1. Then, we present ablation studies regarding the size of the mask.

#### **B.1** Evaluation Metrics

We give the details of evaluation metrics as follows.

**CLIP-IQA.** The metric is based on the CLIP model [9], which is a neural network trained on a variety of (image, text) pairs. It extracts a common vector representation of the image and the text if the image and text are semantically similar [13]. Specifically, given antonym prompts (e.g., "Good photo." and "Bad photo.") as a pair for each prediction, and let  $t_1$  and  $t_2$  be the features from the two prompts opposing in meanings, we first compute the cosine similarity between the image x and the prompt pair:

$$s_i = \frac{\boldsymbol{x} \odot \boldsymbol{t}_i}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{t}_i\|}, \quad i \in \{1, 2\},$$
(2)

and then Softmax is used to compute the final score:

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}.$$
(3)

To assess the overall quality of the generated images, we use the dimensions of quality ("Good photo." vs "Bad photo.") and naturalness ("Natural photo." vs "Synthetic photo.") separately. We average the scores of these two dimensions to obtain the final result.



**Ap-Fig. 1:** Visual ablation studies of various mask sizes in our method. We conduct separate analyses based on the width and height of the masks.

**CLIP-H.** For our group photo generation task, one of our main concerns is the probability of generating a portrait within the given region. Therefore, based on the aforementioned CLIP-IQA, we constructed a new evaluation metric called CLIP-H by setting prompt pairs: "With person" and "Without person".

FID. The calculation of FID [5] is given by:

$$FID = |\mu - \mu_w| + \operatorname{tr}\left(\Sigma + \Sigma_w - 2\left(\Sigma\Sigma_w\right)^{\frac{1}{2}}\right),\tag{4}$$

where  $\mathcal{N}(\mu, \Sigma)$  is the multivariate normal distribution estimated from CLIP features calculated on the test set of AddMe-1.6M and  $\mathcal{N}(\mu_w, \Sigma_w)$  is estimated from CLIP features calculated on generated images.

**Quality Score (QS).** Our quality scores are provided through Generated Image Quality Assessment (GIQA), which quantitatively evaluates the quality of each generated image. Assuming the feature of a generated sample is  $\mathbf{x}$  and its k-th nearest real sample's feature is  $\mathbf{x}^k$ , the probability of generated image can be calculated as:

$$p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\|\mathbf{x} - \mathbf{x}^k\|^2},$$
(5)

where we use Inception v3 [12] as the feature extractor for the feature x.

#### B.2 Effect of Mask Size

Recall that AddMe is compatible with masks of various shapes, and allows users to achieve precise control over the generated portraits with handcrafted masks

**Ap-Tab. 1:** Quantitative comparison between our AddMe and baseline methods on text alignment. CLIP similarity (*i.e.*, CLIP-T) is calculated for the metric.

Method	BLD [1]	SD-Inpaint [10]	Unipaint [16]	${ m AddMe} \ ({ m Ours})$
CLIP-T $\uparrow$	24.54%	23.27%	20.60%	$\underline{23.88\%}$

(See Fig. 8 in our main paper). Here, we want to further explore the impact of different mask sizes. As shown in Ap-Fig. 1, our method can inject identity information into the desired location for masks of different sizes. Based on the input mask, our AddMe generates meaningful interactions with existing individuals, which demonstrates the robustness of our method and flexible mask control.

# C Additional Results

### C.1 Additional Qualitative Results

In this section, we provide additional results of our method in various application scenarios. In Ap-Fig. 3, we demonstrate that our method can insert a portrait into any region of real images. In Ap-Fig. 4, we demonstrate the ability of AddMe to freely control the clothing and pose of the generated portrait through text. In Ap-Fig 5, we display the results when using the same text prompt "A woman wearing a gorgeous gown" with different random seeds. From left to right, the random seeds are 42-46. The results demonstrate the robustness of our method and show the potential application of our method for virtual try-on.

#### C.2 Validation of Image-Text Alignment for Our AddMe

As described in Sec 5.3 in our main paper, despite having a similar setup, most of our baseline methods cannot take both images and text as input simultaneously. Here, we present a quantitative comparison of image-text alignment with the text-driven inpainting model. Specifically, for fair and statistical comparison, we prepare a list of text templates, where for each text prompt, we perform inference on 500 images for each method. Then, we calculate the CLIP similarity [8] between the masked region and the given text (denoted as CLIP-T). As shown in Ap-Tab. 1, our method obtains comparable performance to the current SOTA methods. The list of used text templates are as follows:

- "A person wearing a rainbow scarf"
- "A person in a chef outfit"
- "A person wearing a yellow shirt"
- "A person in a firefighter outfit"
- "A person holding a glass of wine"
- "A person baking cookies"
- "A person reading a book"
- "A person holding a piece of cake"



Ap-Fig. 2: Statistics of our proposed dataset AddMe-1.6M.

# D Details of AddMe-1.6M Dataset

In this section, we first introduce how we obtain the raw images for AddMe-1.6M, and then provide statistics about the dataset distribution.

### D.1 Data Selection and Standardization

We first filter the raw images with the criteria 'aesthetic > 5 & pwatermark < 0.5 & width > 512 & height > 512'. The raw image files are renamed using a seven-digit number to avoid duplication. The shortest side of the image files is resized to 512 and converted to jpg format. All conversion operations are performed based on img2dataset [2]. After standardization and conversion, the raw image files are turned into a raw dataset.

### D.2 Statistics of AddMe-1.6M

We count the number of faces in each scene image in AddMe-1.6M, and over 70% of the scene images contain a single face, as shown in Ap-Fig. 2 (a). We also examined the distribution of face confidence scores, as depicted in Ap-Fig 2 (b).



**Ap-Fig. 3:** Our method supports the insertion of portraits at arbitrary locations in the target scene image.



**Ap-Fig. 4:** Additional uncurated samples generated with text prompts. Our method can maintain facial identity fidelity while demonstrating excellent text-editing capabilities, allowing for free editing of clothing, poses, or introducing other objects.



**Ap-Fig. 5:** Uncurated samples of portrait variations created using different random seed. While maintaining facial fidelity, our method exhibits outstanding generative diversity, paving the way for potential applications such as virtual try-on.

### References

- Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) 42(4), 1–11 (2023)
- 2. Beaumont, R.: img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset (2021)
- 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)
- Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
- 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)
- Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
- 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)
- 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)
- 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)
- 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)
- 11. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
- Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)
- Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuningfree multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023)
- 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)
- 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)
- 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)