MADiff: Text-Guided Fashion Image Editing with Mask Prediction and Attention-Enhanced Diffusion

Zechao Zhan[1]*, Dehong Gao[3,4]*, Jinxia Zhang[1,2]†, Jiale Huang[1], Yang Hu[1], Xin Wang[5]
[1]*Key Laboratory of Measurement and Control of CSE, Ministry of Education,*
School of Automation, Southeast University, Nanjing 210096, China
[2]*Advanced Ocean Institute of Southeast University, Nantong 226010, China*
[3]*Northwestern Polytechnical University, School of Cybersecurity, Xi'an, China*
[4]*Binjiang Institute of Artificial Intelligence, ZJUT, Hangzhou, China*
[5]*Alibaba Group, Hangzhou, China*

***Abstract*—Text-guided image editing model has achieved great success in general domain. However, directly applying these models to the fashion domain may encounter two issues: (1) Inaccurate localization of editing region; (2) Weak editing magnitude. To address these issues, the MADiff model is proposed. Specifically, to more accurately identify editing region, the MaskNet is proposed, in which the foreground region, densepose and mask prompts from large language model are fed into a lightweight UNet to predict the mask for editing region. To strengthen the editing magnitude, the Attention-Enhanced Diffusion Model is proposed, where the noise map, attention map, and the mask from MaskNet are fed into the proposed Attention Processor to produce a refined noise map. By integrating the refined noise map into the diffusion model, the edited image can better align with the target prompt. Given the absence of benchmarks in fashion image editing, we constructed a dataset named Fashion-E, comprising 28390 image-text pairs in the training set, and 2639 image-text pairs for four types of fashion tasks in the evaluation set. Extensive experiments on Fashion-E demonstrate that our proposed method can accurately predict the mask of editing region and significantly enhance editing magnitude in fashion image editing compared to the state-of-the-art methods.**

***Index Terms*—Text-guided Image Editing, Fashion Domain, Diffusion Model**

I. Introduction

Due to advances in text-to-image models [1]–[3] and diffusion models [4] [5], text-guided image editing has also experienced rapid development. Blended Diffusion [6] and Blended Latent Diffusion [7] combine noisy versions of the original image with the intermediate results of the diffusion model to perform text-guided editing. Based on these framework, DiffEdit [8] automatically generates masks for the editing regions by comparing the differences between two generation pipelines. To identify an accurate latent space similar to GAN Inversion [9] for effective image editing, DDIM Inversion [10] is proposed for diffusion models. Based on this inversion space, Imagic [11] and Prompt Tuning Inversion [12] optimize text prompts and use interpolation for image editing. Another type of models, such as P2P [13], FPE [14], and PnP [15], control the editing direction and intensity through modifications of the attention map. Instruct-pix2pix [14], based on P2P [13] and GPT-3 [16], proposes an inversion-free model for text-guided editing based on instruction.

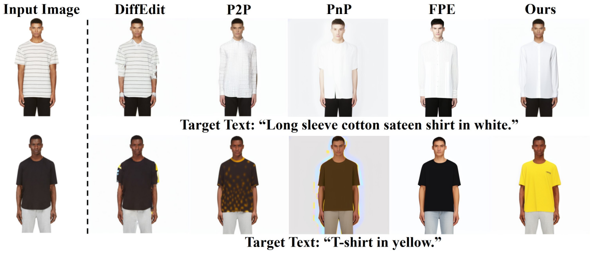
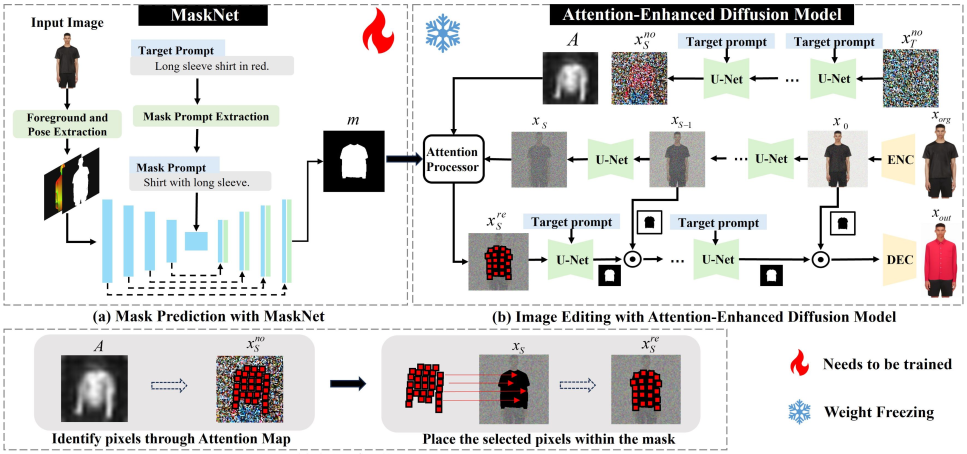

Input Image
DiffEdit
P2P
PnP
FPE
Ours
Target Text: "Long sleeve cotton sateen shirt in white."
Target Text: "T-shirt in yellow."

Fig. 1. General-domain text-guided image editing models have inaccurate localization of the editing region (e.g. wrongly handle the sleeve length) and weak editing magnitude (e.g. cannot change the color from black to yellow).

Although existing text-guided editing models show promising results in general domain, they primarily focus on global edits, such as changing the object categories and styles, with generally weak editing magnitudes. In the fashion domain, model must possess sufficient editing magnitude to handle editing tasks with significant visual differences, as well as consider local editing tasks. Therefore, directly applying general-domain editing models to the fashion domain may encounter the following issues: (1) Inaccurate localization of editing regions; (2) Weak editing magnitudes. As shown in the first row of Fig. 1, DiffEdit [8] and PnP [15] fail to adequately handle edits related to sleeve length, while FPE [17] makes additional modifications to the face, reflecting inaccurate localization of editing regions. As illustrated in the second row of Fig. 1, the t-shirt edited by general-domain methods still largely retain the original black color rather than the yellow color described in the target text, indicating weak editing magnitude.

To address these issues, we propose the MADiff model, comprising of two main components: MaskNet and Attention-Enhanced Diffusion Model. Specifically, to address the issue of inaccurate localization of editing region, the MaskNet is proposed, in which the foreground region, the densepose map

* Equal Contribution. † Corresponding author. This work was supported by Research Fund for Advanced Ocean Institute of Southeast University, Nantong (General Program;Key Program; Major Program), Guangdong Basic and Applied Basic Research Foundation(2022A1515011435), ZhiShan Scholar Program of Southeast University and the Fundamental Research Funds for the Central Universities, the Natural Science Basic Research Program of Shaanxi (Program No.2024JC-YBMS-513), and Key Research and Development Program of Zhejiang Province under Grants 2024C01025.



(a) Mask Prediction with MaskNet

(b) Image Editing with Attention-Enhanced Diffusion Model

(c) Attention Processor

Fig. 2. Overview of our model. (a) Mask Prediction phase: The input image and target prompt are preprocessed to get the foreground region, densepose map and mask prompt, which is then input into MaskNet to predict the editing region. (b) Image Editing phase: DDIM inversion and sampling are conducted to get the attention and noise maps. Then all these maps and the mask from MaskNet are fed into the Attention Processor to create a refined noise map, which is finally used for blended editing to produce the edited image. (c) Attention Processor: The pixels with higher attention values are identified and then placed within the predicted editing region.

and the mask prompt from large language model are fed into a lightweight UNet to predict the mask for editing region. To strengthen the editing magnitude, the Attention-Enhanced Diffusion Model is proposed, where pixels with higher attention values are used to replace the original pixels within the editing region through the proposed Attention Processor. As shown in Fig. 1, compared to state-of-the-art editing models, our method shows accurate localization of editing region and sufficient editing magnitude.

Given the absence of benchmarks in fashion image editing, the Fashion-E dataset is constructed, which is composed of the training set and the evaluation set. The training set consists of 29380 aligned image-text pairs from Fashion-Gen dataset [18] and the corresponding cloth masks, which serve as the inputs and ground truth for the training of MaskNet respectively. The evaluation set is designed for four types of fashion editing tasks: color, detail, material, and comprehensive editing tasks. Within this set, 2639 images are annotated with task-specific target texts to comprehensively assess the editing models.

We conduct extensive experiments on Fashion-E and evaluate our model using a range of metrics. The experimental results indicate that our method not only predicts more accurate editing regions but also significantly enhances editing magnitude. Additionally, an ablation study is performed to demonstrate the effect of each component in our method. The contributions of this work can be summarized as follows:

- A MaskNet is proposed to accurately predict the mask of editing region, which significantly enhances the model's capability in handling local edits.
- A novel Attention-Enhanced Diffusion Model is proposed to address the issue of weak editing magnitude in text-guided fashion image editing, enabling effective editing based on the target prompt.
- Given the absence of benchmarks in fashion image editing, a new dataset named Fashion-E is proposed, which supports the evaluation of different models on various fashion editing tasks.
- The experimental results on Fashion-E demonstrate that our model outperforms state-of-the-art models in both text alignment and preservation of original information.

II. Method

Our method consists of two main phases: mask prediction and image editing, as shown in Fig. 2. In the mask prediction phase, input image and target prompt are first processed by Graphonomy [19], DensePose [20] and LLAMA3-8b [21] to get the foreground region, densepose map and mask prompt, which are then input into MaskNet to predict the mask of editing region. In the image editing phase, DDIM inversion from input image and DDIM sampling from random noise are first conducted to obtain the attention map, and noise maps. Then the attention map, noise maps and the mask from MaskNet, are input into the Attention Processor to produce a refined noise map. Finally, the refined noise map performs blended editing with the mask to obtain the edited image.

A. *Mask Prediction with MaskNet*

MaskNet is proposed to accurately predict the editing region based on the target prompt and the input image. As illustrated in Fig. 2(a), MaskNet adopts a lightweight UNet as framework, in which the spatial attention layers are used in the middle block to incorporate text information. Graphonomy and DensePose are utilized to extract the foreground region and densepose map of the input image, which are then concatenated and applied as the input of the MaskNet. Additionally, considering that the editing region is only decided by the shape-related vocabulary in the target prompt, LLAMA3-8b is utilized to process the target prompt, yielding a mask prompt that only retains the words describing the shape of the fashion object. The MSE Loss between the predicted mask and the cloth mask, which is aligned with the text, is used to train MaskNet on the training set of Fashion-E.

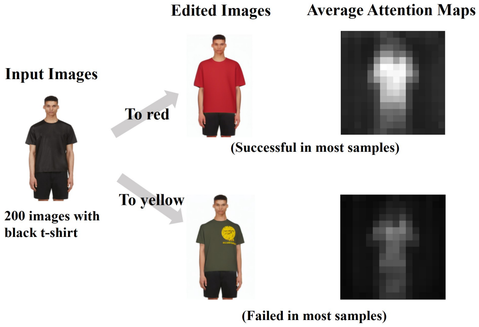


Fig. 3. Comparison of attention maps from successful and failed fashion editing cases. A total of 200 black t-shirts are edited to red and yellow, respectively, and the corresponding average attention maps are calculated.

The key distinction between MaskNet and other text-guided segmentation models [22]–[24] is that MaskNet does not simply segment objects from fashion images. As shown in Fig. 2, with the help of the target prompt, MaskNet can accurately predict the editing region, such as the long sleeve region requested in the target text, rather than only the short sleeve t-shirt region. This enables our model to perform editing tasks that involve altering the shape of fashion objects, such as modifying sleeve length and collar shape.

B. Image Editing with Attention-Enhanced Diffusion Model

As shown in Fig. 1, when confronted with edits necessitating significant visual alterations, such as transitioning from a black t-shirt to yellow, many previous editing methods struggle to generate images that align with the target prompt. To identify the reasons of editing failures, an experiment, in which 200 black t-shirts are edited to red and yellow respectively, is conducted, and the corresponding average attention maps are calculated. As presented in Fig. 3, editing is more likely to succeed when the value of the attention map is high, and vice versa. Therefore, our method enhances editing magnitude by increasing the attention map values.

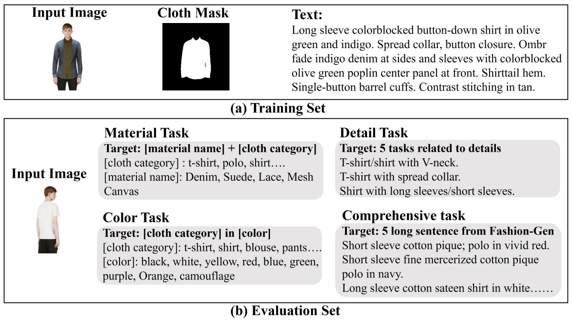
Attention-Enhanced Diffusion Model. Attention-Enhanced Diffusion Model is a training-free model, which utilizes Stable Diffusion as its backbone. During editing process, the input image x_{org} is first encoded into a latent space feature x_0 by the encoder. Then DDIM Inversion, starting from x_0, is executed:

$$x_{t+1} = \sqrt{1-\alpha_{t+1}}\epsilon_\theta(x_t, t) + \sqrt{\alpha_{t+1}} f_\theta(x_t, t), \tag{1}$$

where t is the time step, α_t is a coefficient that decreases over time steps, ϵ_θ is the noise prediction from the U-Net in diffusion model and f_θ is calculated by $(x_t - \sqrt{1-\alpha_t}\epsilon_\theta)/\sqrt{\alpha_t}$. The entire inversion process will take S steps, encoding x_0 into x_S, which is a noise map containing spatial information of the input image. During this process, the DDIM trajectory $x_0, x_1, ..., x_{S-1}$ and the inversion noise map x_S are collected.

Additionally, DDIM sampling, starting from a random noise x_T^{no} with the same size as x_0, is also conducted:

$$x_{t-1}^{no} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \left(x_t^{no} - \frac{\alpha_{t-1} - \alpha_t}{\alpha_{t-1}\sqrt{1-\alpha_t}} \epsilon_\theta(x_t^{no}, t, c) \right), \tag{2}$$


Input Image

Cloth Mask

Text:
Long sleeve colorblocked button-down shirt in olive green and indigo. Spread collar, button closure. Ombr fade indigo denim at sides and sleeves with colorblocked olive green poplin center panel at front. Shirttail hem. Single-button barrel cuffs. Contrast stitching in tan.

(a) Training Set

Input Image

Material Task
Target: [material name] + [cloth category]
[cloth category] : t-shirt, polo, shirt….
[material name]: Denim, Suede, Lace, Mesh Canvas

Detail Task
Target: 5 tasks related to details
T-shirt/shirt with V-neck.
T-shirt with spread collar.
Shirt with long sleeves/short sleeves.

Color Task
Target: [cloth category] in [color]
[cloth category]: t-shirt, shirt, blouse, pants….
[color]: black, white, yellow, red, blue, green, purple, Orange, camouflage

Comprehensive task
Target: 5 long sentence from Fashion-Gen
Short sleeve cotton pique; polo in vivid red.
Short sleeve fine mercerized cotton pique polo in navy.
Long sleeve cotton sateen shirt in white……

(b) Evaluation Set

Fig. 4. Fashion-E dataset. (a) Training Set. The aligned image-text pair is used as the input, while the cloth mask is used as the ground truth for training MaskNet. (b) Evaluation Set. The input image and target text of the four fashion editing tasks are presented, which are used to evaluate editing models.

where α_t is the same as in DDIM Inversion, c represents the target prompt, and ϵ_θ is the noise prediction from the U-Net. After $(T-S)$ steps of DDIM sampling, x_T^{no} will be denoised to x_S^{no} and the attention map A can be calculated by averaging all attention maps with resolution of 16×16 in the U-Net.

After DDIM Inversion and sampling, the attention map A, noise maps x_S, x_S^{no} and the mask m from MaskNet are all input into the attention processor, in which a refined noise map x_S^{re} with higher value of attention map can be obtained.

Finally, guided by the target prompt, the refined noise map x_S^{re} is utilized to conduct DDIM sampling, in which the intermediate results are blended with the DDIM trajectory:

$$x_{t-1}^{re} = DDIM\left(x_t^{re}\right) \odot m + x_{t-1} \odot (1-m), \tag{3}$$

where x_t^{re} is the refined noise map at time step t, m is the mask from MaskNet. Applying Equation (3) for S steps, the refined noise map x_S^{re} can be denoised to x_0^{re}. The edited image x_{out} can be obtained by inputting x_0^{re} into the decoder.

Attention Processor. The steps of Attention Processor are illustrated in Fig. 2(c). Firstly, given the mask m comprising N pixels, the pixels inside the mask m of the map x_S can be represented as:

$$G_{ed} = \{pix | m(pix) > 0, pix \in x_S\}, \tag{4}$$

where pix represents the pixel of noise map and G_{ed} is a set of pixels within the mask m in the noise map x_S. The Attention Processor then identify $N/2$ pixels with the largest values of attention map in x_S^{no}:

$$G_{pr} = \{pix | A(pix) > V_{min}, pix \in x_S^{no}\}, \tag{5}$$

where G_{pr} is a set of pixels with higher attention values from x_S^{no} and V_{min} is the $(0.5N)$th highest value in attention map A. Finally, all the pixels in G_{ed} are substituted with the pixels in G_{pr} in order. If all the pixels in G_{pr} are used once, the surplus pixels in G_{ed} are replaced by randomly selected ones from G_{pr}.

III. Experiments

A. Dataset and Metrics

The Fashion-E dataset is constructed to comprehensively train and evaluate editing models. As shown in Fig. 4(a), the training set of Fashion-E contains 29380 aligned image-text pairs sourced from the Fashion-Gen dataset, along with the corresponding cloth masks. With this set, the image-text

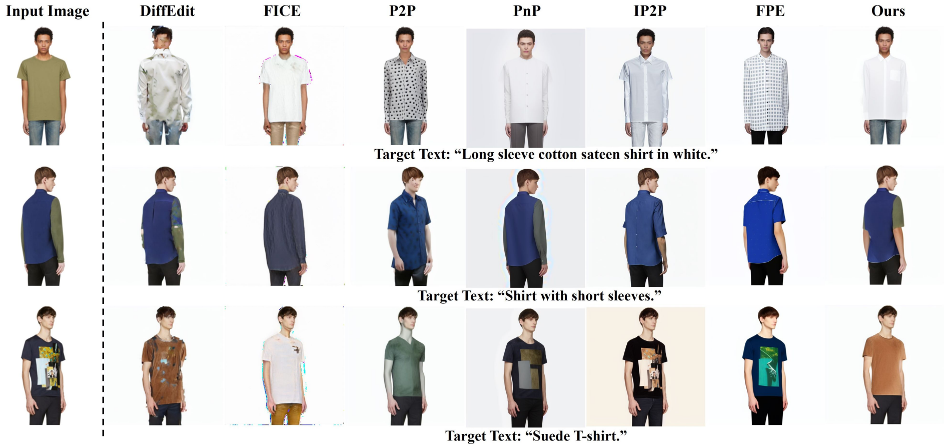

Input Image
DiffEdit
FICE
P2P
PnP
IP2P
FPE
Ours
Target Text: "Long sleeve cotton sateen shirt in white."
Target Text: "Shirt with short sleeves."
Target Text: "Suede T-shirt."

Fig. 5. Qualitative comparison with State-of-the-Art methods on Fashion-E. Target texts are placed below the row of images.

pairs serve as input and the cloth masks serve as the ground truth for the training of MaskNet. Additionally, considering that existing fashion-related datasets cannot comprehensively evaluate models on different fashion tasks, we construct the evaluation set of Fashion-E. Comprising 2639 fashion images, the evaluation set is designed for four types of fashion editing tasks: color editing, detail editing, material editing, and comprehensive editing. Each sample in this set is annotated with a target text for a specific editing task, as shown in Fig. 4(b).

For the evaluation metrics, we employ CLIP Text Score (CLIP-T) [25] to assess the alignment between edited image and target text, CLIP Image Score (CLIP-I) [25] to evaluate the preservation of original image information, and LPIPS [26] to evaluate the perceptual similarity between the original and edited images.

B. Comparison with State-of-the-Art Methods

The proposed method is compared with six text-guided editing models on Fashion-E dataset: Diffedit [8], FICE [27], P2P [13], PnP [15], IP2P [14] and FPE [17]. The experiments for these models are conducted based on their official implements.

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON FASHION-E. BOLD FONT INDICATES THE BEST VALUE FOR THE METRIC, WHILE UNDERLINED FONT INDICATES THE SECOND-BEST VALUE.

Methods	Pub./Year	CLIP-T (↑)	CLIP-I (↑)	LPIPS (↓)	Time (↓)
DiffEdit [8]	ICLR$_{23}$	26.85	83.60	0.152	42.69
FICE [27]	Arxiv$_{23}$	27.85	81.54	0.448	20.15
P2P [13]	ICLR$_{23}$	27.13	85.89	0.146	52.61
PnP [15]	CVPR$_{23}$	28.02	87.77	0.139	174.08
IP2P [14]	CVPR$_{23}$	28.74	87.85	0.167	**6.01**
FPE [17]	CVPR$_{24}$	27.51	84.60	0.160	17.83
Ours	-	**29.20**	**90.37**	**0.137**	12.82

Qualitative Comparisons. Visualization results of different models are shown in Fig. 5. As seen in the first and second rows of Fig. 5, our method can accurately predict the editing regions, preserving more original information. Additionally, as shown in the first and third rows of Fig. 5, our method demonstrates sufficient editing magnitude compared to other models, ensuring higher alignment between the edited image and the target prompt.

Quantitative Comparisons. As presented in TABLE I, our method achieves the highest CLIP-T score, indicating that our model has strongest editing capability. Our model also achieves significantly higher CLIP-I score and lowest average LPIPS score, which means our model can predict more accurate editing regions and achieve better results in local editing tasks. Our model is the second fastest among all models. While our model is slightly slower than IP2P, it significantly outperforms IP2P in terms of overall performance.

TABLE II
ABLATION STUDIES ON ATTENTION PROCESSOR (AP) AND MASKNET.

MaskNet	AP	CLIP-T (↑)	CLIP-I (↑)	LPIPS (↓)
		27.21	84.58	0.161
✓		27.11	89.76	0.133
	✓	28.74	89.03	0.151
✓	✓	29.20	90.37	0.137

C. Ablation Study

MaskNet. Ablating the MaskNet and using masks from local blending [13] in the diffusion Model lead to a notable decline in CLIP-I and LPIPS, as shown in the third line of TABLE II. This suggests that MaskNet improves the accuracy of editing regions and preserves original information outside the region. **Attention Processor (AP).** Ablating the Attention Processor and directly using the inversion noise map during editing result in a significant drop in the CLIP-T score on Fashion-E, as seen in the second line of TABLE II. This indicates that the proposed Attention Processor significantly enhances the editing magnitude, aligning the edited image with the target prompt effectively.

IV. CONCLUSION

This paper proposes a novel text-guided fashion image editing mothod named MADiff, which comprises two components: MaskNet and Attention-Enhanced Diffusion Model. The MaskNet addresses the issues of inaccurate localization of editing regions and Attention-Enhanced Diffusion Model enhances the editing magnitude. Quantitative and qualitative results on Fashion-E dataset indicate that our model outperforms other text-guided image editing methods in both text alignment and preservation of original information.

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in neural information processing systems*, 2022, vol. 35, pp. 36479–36494.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in neural information processing systems*, 2020, vol. 33, pp. 6840–6851.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020.

[6] Omri Avrahami, Dani Lischinski, and Ohad Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18208–18218.

[7] Omri Avrahami, Ohad Fried, and Dani Lischinski, "Blended latent diffusion," in *ACM Transactions on Graphics (TOG)*, 2023, vol. 42, pp. 1–11.

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," in *The Eleventh International Conference on Learning Representations*, 2023.

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019.

[10] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10619–10629.

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.

[12] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han, "Prompt tuning inversion for text-driven image editing using diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7430–7440.

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or, "Prompt-to-prompt image editing with cross-attention control," in *The Eleventh International Conference on Learning Representations*, 2023.

[14] Tim Brooks, Aleksander Holynski, and Alexei A Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18392–18402.

[15] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1921–1930.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 1877–1901.

[17] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang, "Towards understanding cross and self-attention in stable diffusion for text-guided image editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7817–7826.

[18] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal, "Fashion-gen: The generative fashion dataset and challenge," *arXiv preprint arXiv:1806.08317*, 2018.

[19] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7450–7459.

[20] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[21] AI@Meta, "Llama 3 model card," 2024.

[22] Artur Shagidanov, Hayk Poghosyan, Xinyu Gong, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi, "Grounded-instruct-pix2pix: Improving instruction based image editing with automatic target grounding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6585–6589.

[23] Tongxin Wang and Mang Ye, "Texfit: Text-driven fashion image editing with diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 10198–10206.

[24] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al., "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman, "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*, 2023, pp. 22500–22510.

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2018, pp. 586–595.

[27] Martin Pernuš, Clinton Fookes, Vitomir Štruc, and Simon Dobrišek, "Fice: Text-conditioned fashion image editing with guided gan inversion," *arXiv preprint arXiv:2301.02110*, 2023.