

# Fashion Style Editing with Generative Human Prior

Chaerin Kong\*, Seungyong Lee\*, Soohyeok Im\*, Wonsuk Yang\*  
NXN Labs  
{chaerin, seungyong, soohyeok, wonsuk}@nxn.ai



Figure 1. Example of text-driven fashion style edition using our framework (FaSE). Given an input image, a text prompt is used to drive change in garment style while preserving its overall attributes.

## Abstract

Image editing has been a long-standing challenge in the research community with its far-reaching impact on numerous applications. Recently, text-driven methods started to deliver promising results in domains like human faces, but their applications to more complex domains have been relatively limited. In this work, we explore the task of fashion style editing, where we aim to manipulate the fashion style of human imagery using text descriptions. Specifically, we leverage a generative human prior and achieve fashion style editing by navigating its learned latent space. We first verify that the existing text-driven editing methods fall short for our problem due to their overly simplified guidance sig-

nal, and propose two directions to reinforce the guidance: textual augmentation and visual referencing. Combined with our empirical findings on the latent space structure, our **Fashion Style Editing** framework (**FaSE**) successfully projects abstract fashion concepts onto human images and introduces exciting new applications to the field.

## 1. Introduction

Recent years have witnessed unprecedented progress in generative models [8, 13, 14, 24, 43], with the language guidance (*i.e.*, text-to-image) being the cornerstone for their enhanced controllability and generalizability [3, 23, 31, 38, 40, 42, 45]. Image editing task [25, 26, 44] also started to draw great attention from the public thanks to the intuitive

\* Equal contributions.

editing interface text-to-image frameworks provide [2, 12, 36]. One of the most popular ways to incorporate text condition to the image editing pipeline is to use a pretrained vision-language foundation model [1, 15, 32, 37, 51] as in StyleCLIP [36], CLIPstyler [26] or diffusionCLIP [22], where a joint embedding model like CLIP [37] guides the editing direction in the form of similarity score. Although this approach has demonstrated impressive editing capability in some visual domains like human faces [18], they are vulnerable to the blindspots of CLIP features. Specifically, CLIP exposes weakness in domain-specific applications that requires special knowledge [6, 48], and its tendency to focus on the global semantic often renders it unsuitable for guiding local editing operation, especially when the image contains complex semantics [23].

Unfortunately, these vulnerabilities are greatly highlighted in our problem setting: *fashion style editing of human imagery*. In an attempt to introduce an intuitive interface for human fashion restyling, we explore the creative task of text-driven fashion style editing, where we steer the *style* of full-body human portrait with language guidance such as “street fashion” or “floral pattern” (Fig. 1). With our unique problem scope arises a new set of challenges, namely the modeling complexity of whole human body imagery and the elusiveness of fashion concepts, *i.e.*, how some terms like ‘street fashion’ can induce different mental images to individuals. These two distinguish our approach from previous endeavors that tried to manipulate human faces with text guidance [26, 36], in that we need to explicitly reformulate our guidance (*i.e.*, fashion style) into a more illustrative and visually grounded learning signal.

To this end, we propose two directions to visually reinforce the guidance signal. First, we enrich the text prompt by querying a language model [4, 46] to generate descriptions for the concept. Second, we retrieve reference images for the target text from a fashion database we constructed through web crawling, and guide our model to consult them for more descriptive guidance. Leveraging the enhanced visual clarity, our **Fashion Style Editing (FaSE)** framework adeptly transforms abstract fashion concepts into tangible human imagery across a variety of applications.

## 2. Related Work

**Image editing** poses unique technical challenges stemming from the famous distortion-editability trade-off, *i.e.*, making necessary changes while preserving the original. The most prominent approach is leveraging a pretrained generative model to constrain the output to the learned data manifold and executing edits in its latent space. StyleGAN family [18–21] has been a popular choice for the generative prior for long, thanks to its remarkable generation quality and well-behaved latent space [29, 49]. Many works [5, 26, 36, 39, 44, 47] have demonstrated impressive

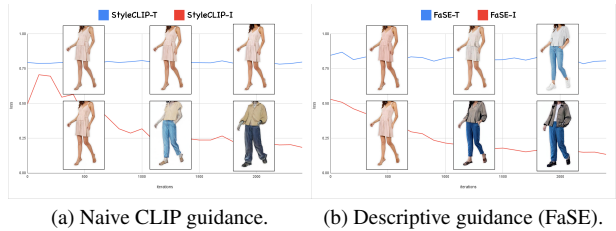


Figure 2. Using CLIP text guidance as in StyleCLIP [36] is insufficient to steer fashion concepts in full-body images, and image guidance in CLIP feature space often distorts the semantics (*left*). FaSE, in contrast, successfully projects non-trivial fashion styles (‘street-fashion’ in this example) onto human images with both type of guidances (*right*).

editing performance by exploring its learned latent space, but their application was mainly confined to narrow semantic domains (*e.g.*, human faces [18]) that can be readily steered by simple words (*e.g.*, “smile”) [36, 39]. Recently, diffusion-based image editing [12, 22, 25, 27, 34, 53] is also being widely explored.

**Human generation** fundamentally differs from human face modeling as the underlying complexity is much greater to synthesize realistic rendering of whole human body with natural garments, body proportions and poses. StyleGAN-Human [10] pioneered this direction of research from a data-centric perspective, with following works [11, 17, 52] further pushing the boundary. In our work, this learned prior serves as the starting point for fashion-specific style editing. **Fashion image generation**, due to its high practical impact, has been studied from various angles, such as virtual try-on [7, 30, 33, 35, 50], clothing synthesis [16, 41] and clothing editing [25]. However, manipulating human-fitted fashion imagery is under-explored due to its innate complexity. FE-GAN [9] proposes an image-based editing framework that relies on adversarial training, but the editing operation is confined to few cases (*e.g.*, sleeve length) and the results are generally unsatisfactory. VPTNet [28] introduces an attribute-based editing technique, which suffers from similar drawback that only limited operations (sleeve and length) predefined at the training phase are supported. In this work, we aim to steer the fashion *style* of human imagery with high-level user input keywords (*e.g.*, “formal”) in a highly versatile and compute-efficient manner.

## 3. Approach

### 3.1. Preliminaries

Our framework is based on StyleGAN-Human [10], which serves as a pretrained prior for human image generation. For editing, we employ the latent mappers introduced in StyleCLIP [36]. Latent mappers are fully-connected networks that transform any latent vector in StyleGAN’s  $\mathcal{W}+$

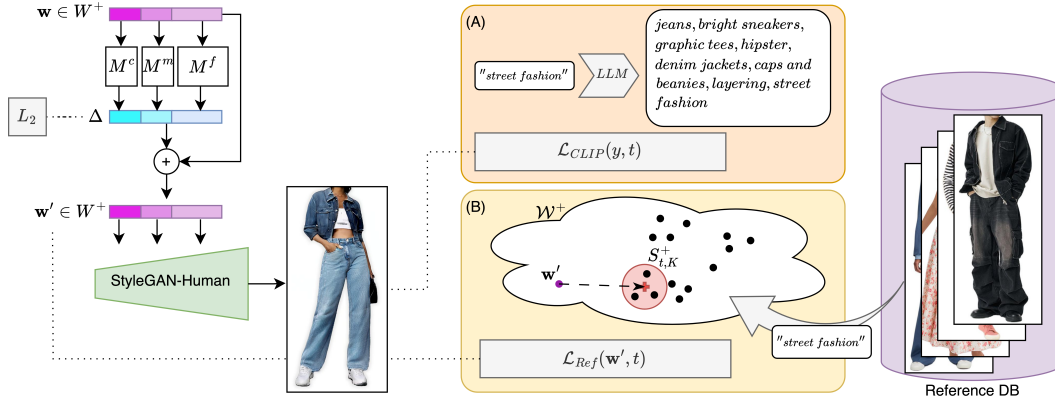


Figure 3. Summary of our fashion style editing framework (FaSE). We learn a latent mapper [36] with two types of illustrative guidance. (A) We transform fashion style concepts into visual descriptions with a pretrained language model. (B) From our preconstructed fashion image database, we retrieve top-k reference images that both suit the target text prompt and resemble the source image, which are used as additional visual guidance.

space to the direction that better aligns with the provided text prompt. Specifically, the latent vector is split into three parts (coarse, medium, and fine), each responsible for different level of detail. Each part undergoes mapping through distinct networks ( $M^c$ ,  $M^m$ ,  $M^f$ , respectively) and is subsequently added to the original latent vector before being fed to the generator. The network is trained to minimize the cosine distance between the generated image and the text prompt in CLIP joint embedding space along with an  $L_2$  regularization to preserve the original image. Although it has delivered impressive results in human face editing, its coarse guidance signal is insufficient for controlling non-trivial fashion styles on full-body human images (Fig. 2).

### 3.2. Textual Augmentation

The key to successful fashion style editing lies in providing specific visual guidance that represents the intended concept. As StyleCLIP utilizes generic CLIP loss with a naive text prompt  $t$ , training often fails for non-trivial fashion concepts, as shown in Fig. 2. We hypothesize that this failure can be attributed to the entanglement in StyleGAN-Human latent space and the innate elusiveness of fashion terminologies. A straightforward yet effective solution would be enriching the text prompt with detailed descriptions, thereby enabling its decomposition into specific components that correspond to various visual aspects of the concept. For example, in visualizing the concept of ‘formal fashion’, identifying specific components like ‘suit,’ ‘trouser,’ and ‘loafer’ proves to be beneficial. To this end, we leverage a Large Language Model (LLM) to augment the text prompt  $t$ , making the context more illustrative by concatenating component words associated with the given concept. Formally, we train the latent mapper to minimize CLIP loss with respect

to the augmented text prompt as follows:

$$\mathcal{L}_{CLIP}(y, t) = \mathcal{D}(\mathcal{E}_{CLIP}^{img}(y), \mathcal{E}_{CLIP}^{text}(g(t)))$$

where  $\mathcal{D}$  denotes the cosine distance,  $\mathcal{E}_{CLIP}$  denotes the CLIP encoder,  $g$  stands for the textual augmentation operation using LLM and  $y$  is the edited image.

### 3.3. Editing with Visual Reference

Upon observing the effectiveness of textual augmentation, a natural progression would be utilizing visual references corresponding to the target concept, since *an image is worth a thousand words*. To this end, we first construct a reference database by collecting clean human images corresponding to various fashion style concepts from the web. Our high level idea is to retrieve an array of relevant images and train our model not only to increase the cross-modal CLIP similarity but also to maximize the image-level similarity with the reference set. Our initial investigation into the CLIP image space reveals its unsuitability for our task, prompting us to transition into a more illustrative feature space, *i.e.*, the  $\mathcal{W}^+$  space [36] of the generator. Formally, our objective can be written as:

$$\mathcal{L}_{Ref}(\mathbf{w}', t) = \frac{1}{K} \sum_{s \in S_{t,K}^+} \mathcal{D}(\mathbf{w}', s),$$

where  $\mathbf{w}'$  is the translated latent vector and  $S_{t,K}^+$  is the collection of  $\mathcal{W}^+$  embeddings of top-K reference images based on similarity.

As mentioned earlier, we empirically find that naively optimizing image similarity in CLIP image space leads to sub-optimal results with undesired distortions, mainly because CLIP visual features are too coarse to capture the subtle nuances of fashion. Hence, we first invert the retrieved

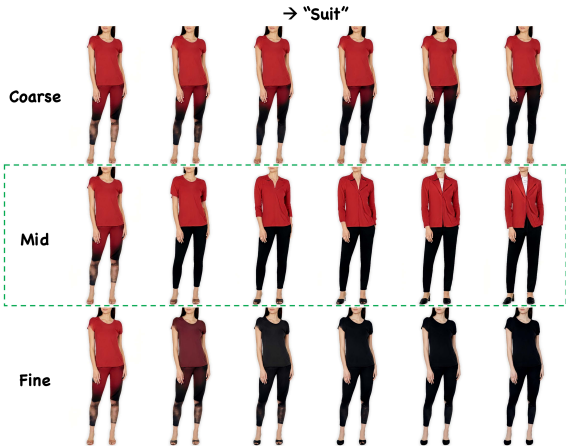


Figure 4. From the top, we train  $M^c$ ,  $M^m$ , and  $M^f$ , respectively, for the prompt ‘suit’. When editing the garment shape as in ‘suit’,  $M^m$  needs to be modified to transform the mid part of  $w$ .

reference images to  $\mathcal{W}+$  space and train the mapper to push latent codes towards them, injecting a more direct and illustrative learning signal. Top-K retrieval is performed to suppress irrelevant guidance coming from overly dissimilar reference images.

### 3.4. Navigating the Hierarchical Latent Space

StyleGAN2 [20] is known for its hierarchical latent space, which has motivated many applications that leverage its coarse-to-fine structure. StyleCLIP [36], for example, divides the latent space into three groups and empirically discovered which level corresponds to specific facial editing operation. We look for a similar empirical guideline for fashion editing on StyleGAN-generated human images.

Following [36], we first compartmentalize the  $\mathcal{W}+$  space into three levels, and test editing operations of various granularity, *e.g.*, changing the color, neckline, garment category, etc. We discover that the coarse group has little association with fashion as it is mainly responsible for human pose, and mid/fine group controls shape/texture respectively. From Fig. 4, for example, it is evident that in order to modify garment category, mid-level transformation is necessary.

## 4. Experiments

We choose StyleCLIP with StyleGAN-Human-v2 checkpoint<sup>1</sup> as our baseline. To construct the reference DB, we collect fashion images from online marketplaces. We select 12 categories (*e.g.*, boxy fit, floral pattern, street fashion, etc) and gather 35 images for each category.

### 4.1. Qualitative Results

We first qualitatively compare our method against StyleCLIP baseline using both text (StyleCLIP-T, FaSE-T) and

<sup>1</sup><https://github.com/stylegan-human/StyleGAN-Human>



Figure 5. Qualitative comparison with StyleCLIP using text and image guidance signal.

Prompt	GPT-4V		User Study	
	FaSE-T	FaSE-I	FaSE-T	FaSE-I
<b>T-shirts</b>	55.9	58.8	56.9	52.0
<b>Floral</b>	72.5	67.6	63.7	62.7
<b>Boxy</b>	91.2	73.5	61.8	76.5
<b>Formal</b>	91.2	96.1	90.2	94.1
<b>Street</b>	95.1	88.2	79.4	80.4

Table 1. Text-image alignment using GPT-4V and user test. We report 1v1 win rate against [36] on diverse text prompts.

image (StyleCLIP-I, FaSE-I) signal. We can see from Fig. 5 that while naive CLIP signal, whether it be textual or visual, fails to modify human images, ours successfully alters them towards the target direction. FaSE-I, in particular, incurs little change to the original image semantics thanks to the retrieval mechanism.

### 4.2. Quantitative Results

For the quantitative evaluation, we report the 1v1 win rate as assessed by GPT4-V and human raters. For GPT4-V evaluation, we prompt the model with the template: “Between the two images, which do you think better aligns with the fashion concept ‘[concept name]’?” Tab. 1 clearly shows that our methods have comparative advantage over the baseline across various text prompts.

## 5. Conclusion

Our work extended text-driven image editing to fashion style editing by exploring two techniques that visually clarify the guidance signal. We hope our framework facilitates new and exciting applications in fashion field.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2
- [2] 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*, pages 18208–18218, 2022. 2
- [3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 1
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [5] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. *arXiv preprint arXiv:2103.10426*, 2021. 2
- [6] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022. 2
- [7] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 2
- [8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1
- [9] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8120–8128, 2020. 2
- [10] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint, arXiv:2204.11823*, 2022. 2
- [11] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Wayne Wu, and Ziwei Liu. Unitedhuman: Harnessing multi-source data for high-resolution human generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7301–7311, 2023. 2
- [12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1
- [15] Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 980–988, 2023. 2
- [16] Shuhui Jiang and Yun Fu. Fashion style generator. In *IJCAI*, pages 3721–3727, 2017. 2
- [17] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 2
- [18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2
- [19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.
- [20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 4
- [21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 2
- [22] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2
- [23] Chaerin Kong and Nojun Kwak. Analyzing multimodal objectives through the lens of generative diffusion guidance. *arXiv preprint arXiv:2302.10305*, 2023. 1, 2
- [24] Chaerin Kong, Jeeseok Kim, Donghoon Han, and Nojun Kwak. Few-shot image generation with mixup-based distance learning. In *European conference on computer vision*, pages 563–580. Springer, 2022. 1
- [25] Chaerin Kong, Donghyeon Jeon, Ohjoon Kwon, and Nojun Kwak. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 848–857, 2023. 1, 2
- [26] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 1, 2
- [27] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2

- [28] Youngjoong Kwon, Stefano Petrangeli, Dahun Kim, Hao-liang Wang, Viswanathan Swaminathan, and Henry Fuchs. Tailor me: An editing network for fashion attribute shape manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3831–3840, 2022. [2](#)
- [29] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. [2](#)
- [30] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. [2](#)
- [31] Seungwoo Lee, Chaerin Kong, Donghyeon Jeon, and Nojun Kwak. Aadiff: Audio-aligned video synthesis with text-to-image diffusion. *arXiv preprint arXiv:2305.04001*, 2023. [1](#)
- [32] Junnan Li, Dongxu Li, Caimeing Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. [2](#)
- [33] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22788–22797, 2023. [2](#)
- [34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. [2](#)
- [35] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: high-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022. [2](#)
- [36] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. [2](#), [3](#), [4](#)
- [37] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [2](#)
- [38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. [1](#)
- [39] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. [2](#)
- [40] 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*, pages 10684–10695, 2022. [1](#)
- [41] 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. [2](#)
- [42] 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. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. [1](#)
- [43] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pages 30105–30118. PMLR, 2023. [1](#)
- [44] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1532–1540, 2021. [1](#), [2](#)
- [45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. [1](#)
- [46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. [2](#)
- [47] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. [2](#)
- [48] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. [2](#)
- [49] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12863–12872, 2021. [2](#)
- [50] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Oot-diffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024. [2](#)
- [51] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. [2](#)
- [52] Kaiduo Zhang, Muyi Sun, Jianxin Sun, Binghao Zhao, Kunbo Zhang, Zhenan Sun, and Tieniu Tan. Humandiffusion: a coarse-to-fine alignment diffusion framework for

controllable text-driven person image generation. *arXiv preprint arXiv:2211.06235*, 2022. [2](#)

- [53] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. [2](#)