Named Entity Driven Zero-Shot Image Manipulation

Zhida Feng[1,2,4] Li Chen[1,2,*] Jing Tian[3] JiaXiang Liu[4] Shikun Feng[4]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China.
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System,
Wuhan University of Science and Technology, Wuhan, China.
[3]Institute of Systems Science, National University of Singapore
[4]Baidu Inc.

{fengzhida,chenli}@wust.edu.cn, tianjing@nus.edu.sg, {liujiaxiang,fengshikun01}@baidu.com

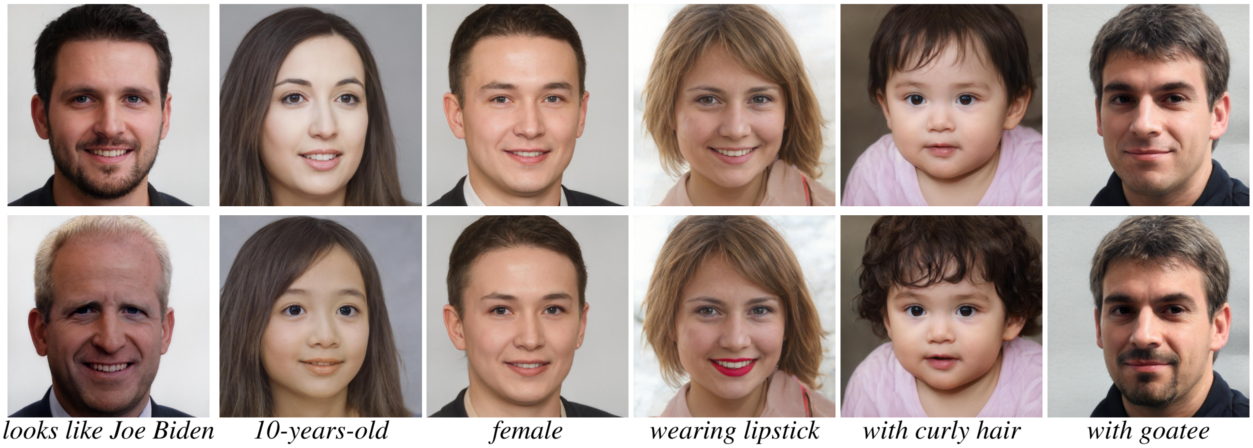

Figure 1. Zero-Shot Manipulation with StyleEntity. Top row: Input images. Bottom row: Manipulation results. All prompts are unseen in the training set.

Abstract

We introduced StyleEntity, a zero-shot image manipulation model that utilizes named entities as proxies during its training phase. This strategy enables our model to manipulate images using unseen textual descriptions during inference, all within a single training phase. Additionally, we proposed an inference technique termed Prompt Ensemble Latent Averaging (PELA). PELA averages the manipulation directions derived from various named entities during inference, effectively eliminating the noise directions, thus achieving stable manipulation. In our experiments, StyleEntity exhibited superior performance in a zero-shot setting compared to other methods. The code, model weights, and datasets are available at https://github.com/feng-zhida/StyleEntity.

1. Introduction

Generative Adversarial Networks (GANs) [17] have significantly advanced synthetic image generation, with StyleGAN [21–24] architectures leading in producing high-resolution, lifelike images. Numerous studies [3, 4, 18, 20, 37–39, 41, 43] have demonstrated the ability to manipulate these images subtly through precise control of style codes within StyleGAN frameworks. Moreover, recent advancements have enabled the modulation of style codes using textual prompts, enhancing GANs' applicability. When combined with GAN Inversion techniques [1, 2, 5, 7, 11, 34, 40, 44], this capability extends GANs' utility, allowing for the detailed refinement and editing of real-world photographs.

Current text-guided image manipulation methods [29, 32, 48] typically utilize the CLIP model [33], which learns to map prompts to the StyleGAN latent space, facilitating image manipulation via textual descriptions. However, these methods have notable limitations. Optimization-based methods require per-image, per-prompt optimization, incurring high inference costs. Conversely, neural network-based

[*] Corresponding Author

mappers need a dedicated mapper trained for each specific prompt, leading to practical inconveniences. FFCLIP [48], despite training with prompts similar to those during inference and thus allowing for adaptability to diverse prompts without retraining, performs poorly with prompts outside the training set's distribution. Amassing a large, varied prompt dataset to cover the entire representational space seems a straightforward solution, but is impractical due to the difficulty in data collection. Our method ingeniously overcomes this challenge by using named entities as proxies during training. These proxies allow our model to transition from named entity prompts to user-provided prompts in a zero-shot fashion during inference, effectively addressing the challenge of collecting diverse prompts.

In this work, we introduce StyleEntity, a novel approach to image manipulation that capitalizes on the inherent compositional properties of named entities within the CLIP text space. Our findings suggest that named entities can serve effectively as conglomerates of descriptive text components. With one training session utilizing this named entity dataset, our model gains zero-shot manipulation capabilities, proficiently employing a variety of descriptive prompts at the inference stage. Figure 1 demonstrates the zero-shot manipulation capability of StyleEntity. Moreover, we propose the Prompt Ensemble Latent Averaging (PELA) technique. PELA significantly improves the generative quality of our model based on named entity data, thus facilitating the creation of manipulated images with exceptional stability and fidelity in a zero-shot scenario.

Text-guided image manipulation methods [14, 25, 28–30, 32, 46–48] aim to edit the content of images based on user-provided prompts, while preserving prompt-irrelevant content. To evaluate the effectiveness of these manipulations, we introduce a trade-off plot that illustrates the model's effectiveness across a spectrum of hyperparameters, providing a comprehensive evaluation of text-guided image manipulation techniques.

2. Related Work

Latent Space Manipulation. Recent years have seen notable progress in latent space manipulation, aimed at controlling the attributes of generated images by adjusting latent variables. One straightforward strategy has been to determine the path in latent space that corresponds to a desired attribute modification, such as changing hair color [13, 16, 38]. GANSpace [18] utilizes principal component analysis in the activation space to discover controllable dimensions within generative adversarial networks (GANs), influencing aspects like perspective, aging, and lighting, including time of day alterations. Collins et al.[12] improved the modification of style vectors for localized, semantically meaningful changes in an image, seamlessly integrating elements from another image. These methods typically adopt an *"invert first, edit later"* approach[34], which requires reverting images to the latent space using GAN Inversion before editing. Most recent techniques manipulate the $\mathcal{W}$ or $\mathcal{W}+$ spaces [34, 40], while Wu et al.[45] extend their work into the $\mathcal{S}$ space. In StyleGAN, the $\mathcal{W}$ space, an intermediate latent space, is derived from the initial random noise vector (z space), facilitating more interpretable and controllable feature manipulation than z space. The $\mathcal{W}+$ space allows even finer control over the image generation by using different $\mathcal{W}$ vectors for each layer (18 layers for 1024×1024 images), thus enabling more precise manipulation of specific image features. We use e4e[40] for GAN inversion, allowing manipulation within the $\mathcal{W}+$ space.

Text-Guided Image Manipulation. Initial methods [14, 26, 30] were often trained on manually annotated datasets like MSCOCO [27] and CUB [42]. The StyleGAN-based TediGAN [46] generated texts by annotating attributes based on predefined rules, resulting in a lack of diversity in the descriptions. The advent of the CLIP model [33], trained on a vast dataset, has produced semantically rich text embeddings. Many studies have since utilized CLIP to bridge between the text and StyleGAN spaces, with impressive outcomes. Optimization-based methods such as StyleCLIP Optimization [32] optimize the latent code for each image and prompt, which can be cumbersome; StyleCLIP's Global method [32] identifies a direction in the StyleGAN latent space using CLIP embeddings, yet this approach does not tailor the manipulation direction for individual images; StyleCLIP's Mapper [32] trains a mapper for each textual prompt, which then modifies the latent code accordingly. FFCLIP [48] enhances this by manually creating prompts and training a universal mapper, but struggles with prompts outside the training distribution. Our paper introduces training with named entities seen in the CLIP space as combinations of various descriptive texts, enabling the zero-shot application of general unseen descriptive texts during inference. In recent years, there have been image editing methods based on diffusion that typically manipulate either the pixel space [31] or the compressed space [35] of a VAE. Our approach, however, differs in that we directly manipulate within the attribute-decoupled StyleGAN latent space to edit images.

3. Method

3.1. Utilizing Named Entities as Proxies

In the field of Natural Language Processing (NLP), named entities refer to phrases that include proper nouns such as individuals' names, organizations, and locations, among others. A unique aspect of human cognition is the ability to associate a variety of visual attributes with these named entities. For example, when one thinks of the celebrity Dwayne "The Rock" Johnson, several visual attributes such

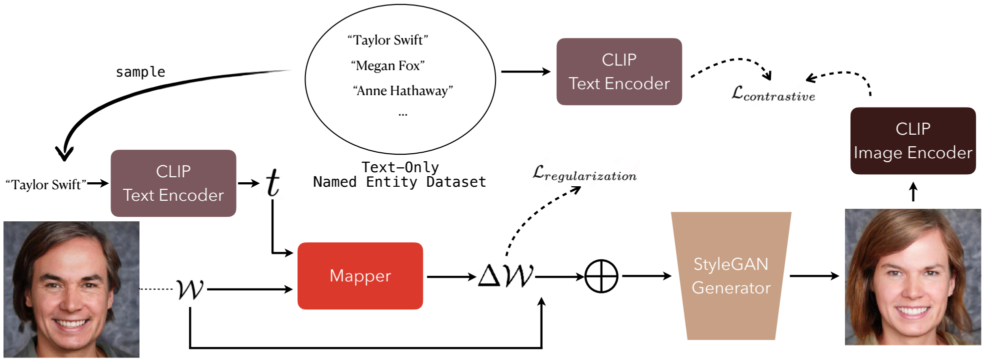

Figure 2. The training pipeline of our proposed StyleEntity. All model parameters are frozen except for the mapper.

as "bald", "dark eyes", and "beardless" come to mind. Similarly, the CLIP model's visual space associates visual attributes with named entities.

With its ability to align visual and textual spaces, a named entity text in the CLIP model can encapsulate various visual attributes. This unique characteristic enables us to utilize named entities to efficiently cover a substantial portion of the representational space, circumventing the necessity to gather numerous individual descriptive phrases. Such broad coverage is vital in a zero-shot setting, where the model needs to accommodate a set of new prompts. By utilizing named entities as proxies during training, we ensure extensive representational coverage. This method helps our model seamlessly transition from named entity prompts to descriptive prompts in zero-shot inference, leading to efficient and high-quality image manipulation.

Moreover, employing named entities significantly streamlines the data collection process. Traditional methods typically necessitate the labor-intensive and time-consuming manual gathering of descriptive phrases. By contrast, named entities, particularly those related to facial descriptors, are easily compiled by aggregating lists of public figures. This approach not only eases data collection but also broadens the model's learning of associations between text and style codes. Crucially, this method requires only the collection of named entity textual data, foregoing the need for corresponding images.

3.2. Training

Figure 2 illustrates the training pipeline of StyleEntity. The mapper takes the style code of the original image and the CLIP text embeddings of a named entity. It outputs a manipulation direction $\Delta\mathcal{W}$, which, when added to $\mathcal{W}$, results in a manipulated style code. The StyleGAN generator uses this to produce an image. The image is then employed to calculate a contrastive loss with text embeddings of all named entities for optimizing the mapper.

Network Design. Our network architecture is founded on a straightforward multi-layer perceptron (MLP) design. This design accepts a style code and a CLIP text embedding as inputs and provides the corresponding manipulation direction as an output. The inputs to our model are created by simply concatenating the style code and the CLIP text embedding. The inclusion of the style code as an input is driven by our intention to enable the model to adaptively generate manipulation directions for various style codes.

Given this configuration, we can express the operation as:

$$\Delta\mathcal{W} = MLP([\mathcal{W}; t]) \tag{1}$$

where $\mathcal{W}$ denotes the style code, and t represents the text embeddings.

Moreover, considering that each style code in the $\mathcal{W}+$ space corresponds to different attributes, we assign a unique MLP to each style code. This implies that for an image of resolution 1024×1024, which has 18 style codes, our network will have 18 corresponding MLPs. This design is predicated on the understanding that different style codes in the $\mathcal{W}+$ space regulate different aspects of the image. By assigning a unique MLP for each style code, we ensure that our model can adaptively manipulate each aspect of the image in response to the given prompt. This strategy enables us to conduct fine-grained, attribute-specific image manipulation, contributing to the overall effectiveness and adaptability of our method.

Contrastive Training. The Mapper's output is fed into a pre-trained StyleGAN generator to produce an image. This operation can be mathematically expressed as:

$$G(\mathcal{W} + \alpha\Delta\mathcal{W}) \tag{2}$$

where the G denotes a pre-trained StyleGAN Generator, α symbolizes the manipulation strength, and is consistently set at 0.20 during both training and inference phases. The generated image is then processed via the CLIP image encoder to produce image embeddings.

These embeddings are optimized using a contrastive loss function. The objective of this function is to minimize the distance between the image embeddings and the input named entity embeddings, while simultaneously maximizing the distance to other entities. This approach ensures that the manipulated image aligns more closely with the desired named entity, thereby enhancing the accuracy and precision of our image manipulation.

The contrastive loss can be expressed as:

$$\mathcal{L}_{contrastive} = \frac{1}{M}\sum_{i=1}^{M} -\log \frac{\exp(sim(u_i, v_i)/\tau)}{\sum_{j=1}^{M}\exp(sim(u_i, v_j)/\tau)} \tag{3}$$

where u_i and v_i are the image and text embeddings respectively, M is the number of samples, and τ is a temperature parameter.

To ensure that the manipulation direction does not deviate too far from the original style code, we also introduce a regularization term, defined as:

$$\mathcal{L}_{regularization} = ||\Delta\mathcal{W}||^2 \tag{4}$$

The overall loss is then a weighted sum of these two terms:

$$\mathcal{L}_{total} = \mathcal{L}_{contrastive} + \lambda\mathcal{L}_{regularization} \tag{5}$$

where λ is a weighting factor.

Many studies [8–10, 19] have highlighted the importance of large negative sampling for contrastive learning. Given that the CLIP model is frozen during our training process, we can precompute the text embeddings for all named entities. This precomputation allows us to perform negative sampling across the entire dataset when calculating the contrastive loss. By doing so, we can ensure a more robust and comprehensive optimization process. This methodology not only accelerates the training process but also improves the generalization ability of our model by exposing it to a broader range of named entities during training. It enables our model to learn more discriminative features by comparing each named entity with all others in the dataset.

3.3. Prompt Ensemble Latent Averaging

During the training phase, only named entity data were utilized, which encapsulate various visual attributes; however, during inference, the user-provided prompts typically exhibit fewer attributes. This disparity between the training and inference data distributions introduces considerable noise during inference. To mitigate this problem, we have introduced the Prompt Ensemble Latent Averaging (PELA) technique.

PELA begins by constructing a target and source prompt using named entities and user input, respectively. For example, considering the prompt "with beard", we sample a named entity e_i from dataset, and generate the source as "e_i" and the target as "e_i with beard". Each prompt is then fed into the mapper, yielding manipulation directions for both target and source. The distance between these directions provides a manipulation vector unique to the named entity e_i. The manipulation direction for a prompt is mathematically defined as:

$$\Delta\mathcal{W}(e_i) = MLP([\mathcal{W}; t_{target(e_i)}]) - MLP([\mathcal{W}; t_{source(e_i)}]) \tag{6}$$

where MLP refers to our mapper model and $target(\cdot)$ and $source(\cdot)$ represent the construction methods of target and source prompts respectively.

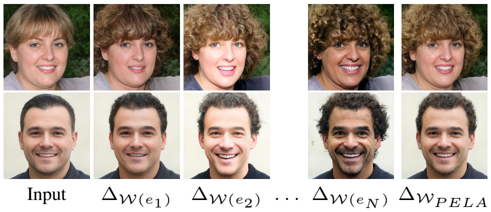
Input
$\Delta_{\mathcal{W}(e_1)}$
$\Delta_{\mathcal{W}(e_2)}$
. . .
$\Delta_{\mathcal{W}(e_N)}$
$\Delta_{\mathcal{W}_{PELA}}$

Figure 3. Comparative visualization of manipulation vectors for individual named entities (columns 2 to the second-to-last) versus the Prompt Ensemble Latent Averaging (PELA) approach (final column). PELA demonstrates enhanced stability in manipulation by averaging out the incidental directional noise. The applied prompt is "curly hair"

	Celebrity-Names-90k	**Cat-Breeds-101**	**Dog-Breeds-354**
Examples	*George Bush* *Taylor Swift* . . .	*Aegean Cat* *Kinkalo Cat* . . .	*English Pointer* *Spanish Water Dog* . . .
Size	90084	101	354

Table 1. Text-Only Named Entity datasets utilized for training.

Assuming that the noise in manipulation directions derived from different named entities is random, as illustrated in Figure 3, we mitigate this by averaging the manipulation directions from N different named entities:

$$\Delta\mathcal{W}_{PELA} = \frac{1}{N}\sum_{i=1}^{N}\Delta\mathcal{W}(e_i) \tag{7}$$

Through this process, PELA leverages the diversity of named entities to effectively neutralize the noise arising from discrepancies in distributions between training and inference phases. This approach significantly enhances our model's image manipulation capabilities in a zero-shot setting.

4. Experiments

4.1. Implementation Details

To illustrate the wide-ranging applicability of our StyleEntity approach, we trained our model across multiple domains, such as faces and animals. In the area of face

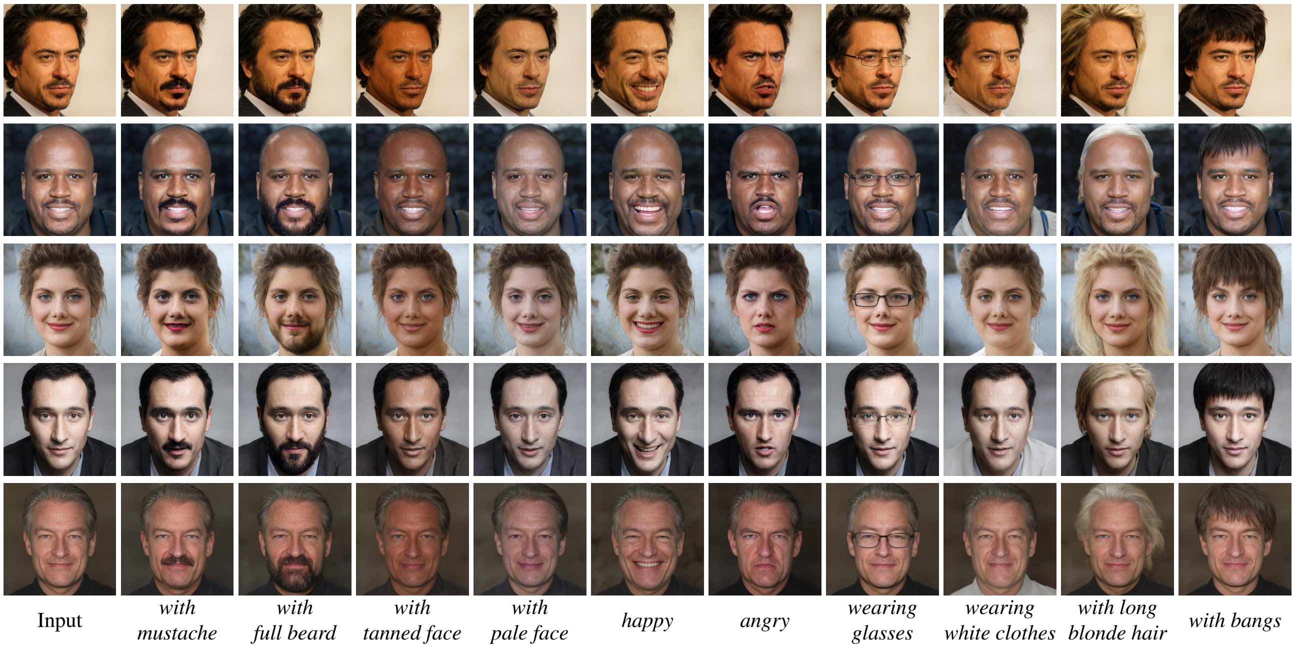

Figure 4. Diverse facial manipulations by StyleEntity. The input images are shown in the first column with subsequent columns displaying manipulated RESULTS. Prompts are indicated below each column.

manipulation, we utilized the Celebrity-Names-90k dataset in conjunction with the StyleGAN2 architecture, which was pretrained on the FFHQ dataset. For non-face domains, we leveraged the Cat-Breeds-101 and Dog-Breeds-354 datasets, using the StyleGAN-ada model pretrained on the AFHQ dataset. The specific Named Entity datasets employed during our training are detailed in Table 1. Our mapper training was not reliant on real images. We generated style codes from a Gaussian distribution using StyleGAN, thereby ensuring a bias-free learning process. The entire training was conducted on a single NVIDIA Tesla A100 40GB GPU, utilizing an Adam optimizer with a batch size of 8, a learning rate of 0.2, and $\beta_1 = 0.9$, $\beta_2 = 0.999$.

4.2. Various Text-guided Manipulation Results

In a comprehensive evaluation of StyleEntity's manipulation capabilities, we present a broad range of text-driven image editing results within the facial domain, as shown in Figure 4. This collection demonstrates StyleEntity's proficiency in executing accurate edits according to the given prompts while preserving the features not specified by the prompts. The presented edits cover a varied set of prompts, from simple attributes like "beard" to more intricate constructs such as "mustache", illustrating the model's detailed comprehension and implementation of text-guided manipulations. StyleEntity's robustness is further emphasized by its capability to manage a multitude of facial attributes—whether skin tone, facial expressions, accessories, or hairstyles—producing results that are not only realistic and stable but also adaptable to the text prompts.

Corresponding to the facial manipulations, Figure 6 shows the model's expertise in the domain of cats and dogs. Reflecting the high fidelity and relevance observed in facial image manipulation, StyleEntity effectively modifies single attributes such as fur color and hair length while also skillfully handling complex attribute prompts like "Golden Retriever". The model's success in non-facial domains reinforces its versatility and the broad applicability of the StyleEntity framework across various Style-Based generators.

The experimental evidence consolidates StyleEntity's capability in generating lifelike and textually coherent manipulations, thereby establishing its superiority in the realm of zero-shot image manipulation across varied domains.

4.3. Evaluation

Metrics. To evaluate text-guided image manipulation methods, we concentrate on two factors: the precision of the content modifications in response to user prompts and the preservation of image aspects that are unrelated to the prompts. The former is quantifiable via the CLIP-score, designed to gauge the semantic correlation between the generated image and the textual prompt. For the latter, we utilize metrics such as the Fréchet Inception Distance (FID). FID is used to evaluate the similarity in the feature space between the original and manipulated images, thereby reflecting changes to prompt-irrelevant features.

We have noted a trade-off between these metrics: ma-

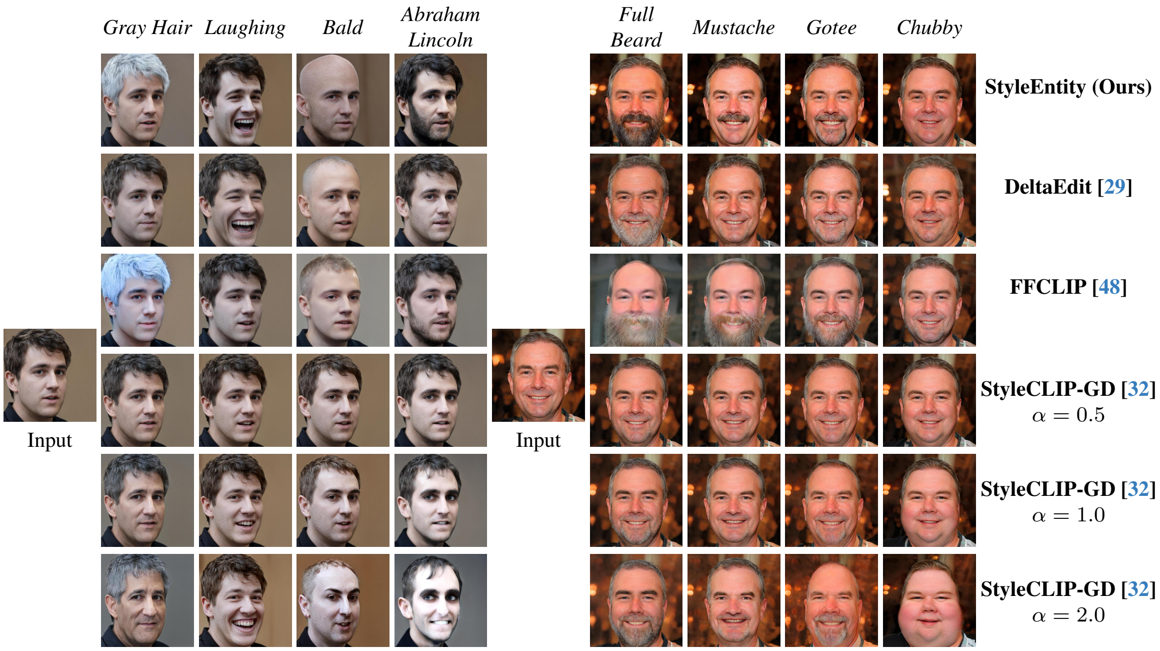

Figure 5. Qualitative comparison of StyleEntity with DeltaEdit, FFCLIP, and StyleCLIP-GD across diverse prompts. Prompts used for each column are displayed above the images.

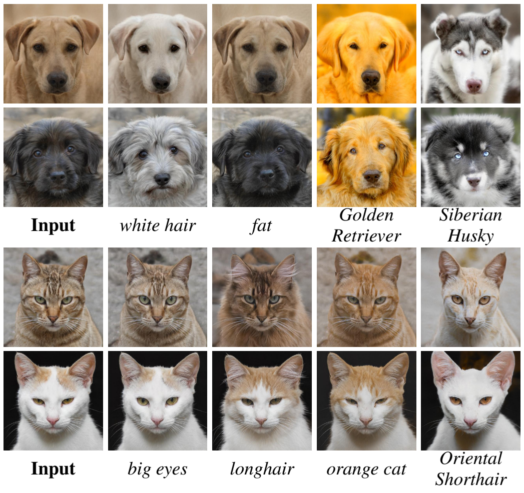

Figure 6. Manipulation results on cat and dog domain by StyleEntity. The input images are in the first column, with the remaining columns showing the manipulated results. Prompts used are indicated below each column.

nipulations that enhance relevance to the prompt often result in differences from the original image, negatively impacting other metrics. To navigate this trade-off, drawing inspiration from recent large-scale text-to-image models [6, 15, 36], we plot FID-CLIP curves. These curves evaluate models based on the resemblance of the edited image to the original at a specific level of prompt similarity (CLIP-score). By adjusting inference hyperparameters, we plot different points to shape these curves. Models with curves leaning towards the bottom right are considered superior as they suggest a smaller deviation from the original image at higher CLIP-scores, indicating better preservation of prompt-irrelevant details while aligning with the prompt.

Qualitative Evaluation. This study compared StyleEntity with state-of-the-art text-guided image manipulation methods, namely StyleCLIP-GD [32], FFCLIP [48], and DeltaEdit [29]. In all results presented, StyleEntity employed a static hyperparameter $\alpha = 0.20$. For DeltaEdit and FFCLIP, we used the pre-trained models provided by them with default parameters. For StyleCLIP-GD, we followed the approach of DeltaEdit by setting the β to 0.03. Owing to its sensitivity to parameters, we presented its results at manipulation intensities of $\alpha = 0.5, 1.0, 2.0$. Each model will edit 10 same images for each prompt.

As depicted in Figure 5, the performance of StyleCLIP-GD heavily relies on the selection of hyperparameters. This necessitates prompt-specific fine-tuning for optimal results, which could be impractical in real-world applications. FFCLIP responds to all test cases but has a tendency to misinterpret the target prompt. For instance, the prompt "gray hair" might inadvertently lead to images with lighter skin tones alongside the gray hair. DeltaEdit struggles to manipulate certain prompts accurately; it has difficulty with prompts such as "gray hair" and "Abraham Lincoln".

In contrast, StyleEntity consistently executed precise manipulations across a wide range of prompts. Remarkably, in cases involving similar beard styles like "Full Beard",

Methods	Human Preference
StyleCLIP GD [32] $\alpha = 0.5$	15.0%
StyleCLIP GD [32] $\alpha = 1.0$	12.0%
StyleCLIP GD [32] $\alpha = 2.0$	8.2%
FFCLIP [48]	11.0%
DeltaEdit [29]	14.2%
StyleEntity (Ours)	**39.6%**

Table 2. Results of Human Evaluation.

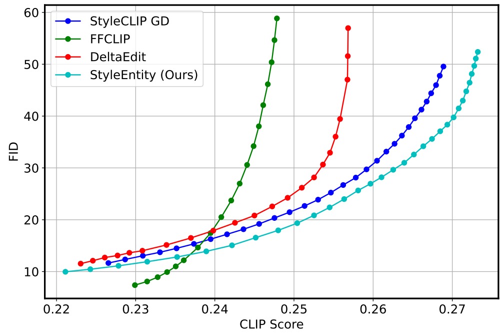

Figure 7. Quantitative comparison of StyleEntity, DeltaEdit, FFCLIP, and StyleCLIP-GD using FID-CLIP trade-off curves.

Methods	Inference Time (ms)
StyleCLIP GD [32]	43
FFCLIP [48]	34
DeltaEdit [29]	61
StyleEntity (w/o PELA)	34
StyleEntity (w/ PELA, $N = 256$)	44

Table 3. The inference time consumption for each method.

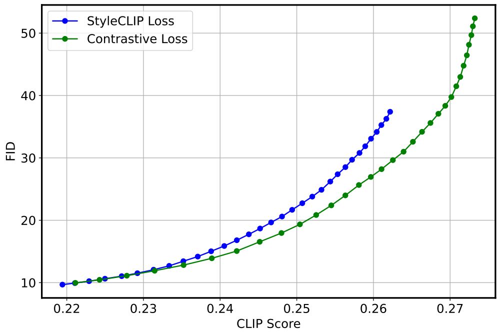

Figure 8. Quantitative comparison using FID-CLIP curves between Contrastive Loss and StyleCLIP Loss.

"Mustache", and "Gotee", StyleEntity accurately distinguished and manipulated each style. Furthermore, StyleEntity skillfully responded to a diverse set of prompts while preserving attributes unrelated to the prompt, demonstrating its robustness and versatility in text-guided image manipulation.

Quantitative Evaluation. A vital part of our evaluation involves the use of trade-off curves. These curves provide a visual representation of the relationship between the FID and CLIP scores. For models such as FFCLIP and DeltaEdit, we included a manipulation strength parameter, which allowed us to scale this parameters uniformly across models. This scaling leads to corresponding FID and CLIP scores, which are then plotted to form the FID-CLIP trade-off curves. We utilized a test set of 100 facial description prompts, encompassing aspects like hair color, hairstyle, beard style, mood, and more. Details of the test set prompts and additional SSIM-CLIP and LPIPS-CLIP curves are provided in the *Supplementary Materials*.

As illustrated in Figure 7, our analysis indicates that the StyleEntity model significantly outperforms competing models in terms of trade-off curve profiles. This implies that for a specific level of textual-visual correlation, StyleEntity introduces fewer modifications to the original image. This superior performance confirms the efficacy of our approach in maintaining a balance between manipulation and preservation of original image attributes.

We also conducted a human evaluation on the 100 prompts we had collected, the results of which are presented in Table 2. This study involved five participants, each tasked with assessing the quality of images generated from 100 prompts. Each prompt resulted in the generation of nine images. Participants evaluated the images based on their textual relevance and visual similarity to the original image, choosing the model they deemed superior subjectively. The findings corroborate the conclusions drawn from the previously presented FID-CLIP curve, with our model consistently preferred in subjective evaluations. This further substantiates the performance of our model and emphasizes its superiority in generating high-fidelity images that align contextually and visually with the provided prompts. Details of the human evaluation can be found in the *Supplementary Materials*.

Table 3 shows the inference speeds of different models. Note that variations in the underlying frameworks of each method could cause deviations in these results. Nonetheless, a general observation reveals that all models demonstrate similar inference speed. This parity in speed, along with the superior performance of our model, highlights the efficiency of StyleEntity in producing high-quality manipulated images without compromising computational efficiency.

4.4. Ablation Study

Constrastive Training In this section, we evaluated the effectiveness of our contrastive training approach. For a comparative analysis, we replaced the loss function used in our model from the contrastive loss (Eq. (3)) to the loss defined in StyleCLIP:

$$\mathcal{L}_{StyleCLIP} = 1 - \frac{sim(ui, v_i)}{100} \quad (8)$$

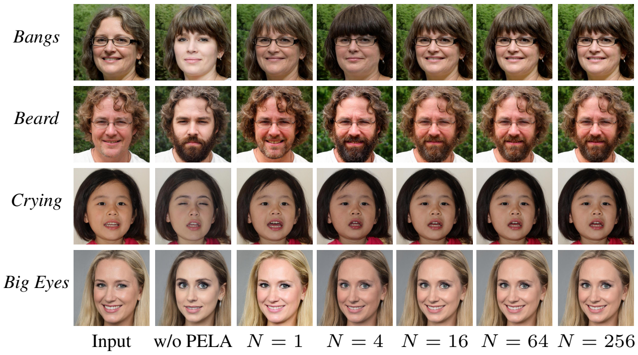

Figure 9. Qualitative ablation study of PELA. It illustrates the manipulation results without PELA and with varying ensemble sizes N.

The StyleCLIP loss diverges from our contrastive formulation by directly aiming to maximize the similarity score without a contrastive denominator, which influences the model's ability to distinguish among the nuanced variations of text-guided manipulations.

Our quantitative analysis is supported by the FID-CLIP curves illustrated in Figure 8. These curves clarify the performance difference between models using contrastive loss and those using StyleCLIP loss. The superior performance of the contrastive loss is apparent, with a notable improvement across the trade-off metrics. This outlines the contrastive loss's superior capability to balance image fidelity and adherence to textual descriptions, thereby validating the usefulness of our chosen loss function.

PELA. Our ablation study offers a thorough evaluation of the PELA technique's effectiveness. The qualitative results, as illustrated in Figure 9, show that, without PELA, the manipulation direction is burdened with significant noise. For example, in the absence of PELA, the prompt "Bangs" unintentionally removed glasses from the image, suggesting an unwanted modification of unrelated attributes. This noise is reduced with a small ensemble size ($N = 1$ and $N = 4$), yet it continues to affect the manipulation result, as observed with the prompts "Beard" and "Big eyes". As we increase the ensemble size, a clear trend becomes evident: the manipulation results gradually stabilize. Considering both effectiveness and efficiency, we use $N = 64$ as our default setting.

The quantitative results, illustrated in Figure 10, corroborate these findings. The trade-off curve reveals that at $N = 1$ and $N = 4$, without PELA, the performance is suboptimal. As N increases, the curve advances towards the preferred bottom-right quadrant, indicating enhanced fidelity and stability in the manipulated images, which correspond to higher CLIP scores and lower FID values.

Table 3 shows the time consumption related to PELA. Utilizing an ensemble of 256 named entities leads to a slight increase of 10 ms/image during inference. This minor increase in inference time is an acceptable trade-off for the substantial improvements in manipulation quality provided by PELA.

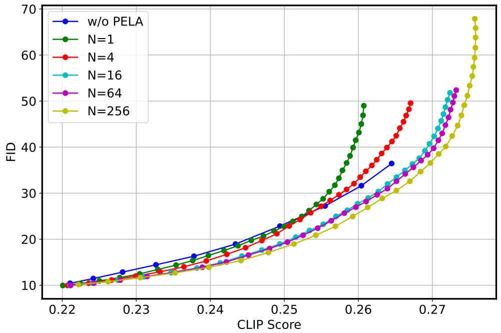

Figure 10. Quantitative ablation study of PELA on FID-CLIP curves.

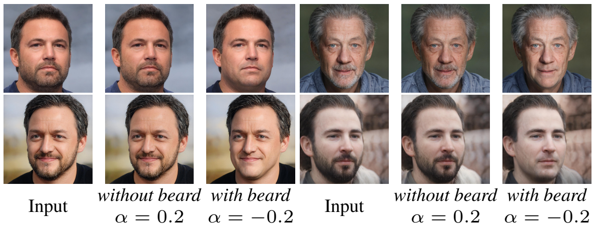

Figure 11. Results of negative manipulation for the "beard" prompt, illustrating the precision of attribute removal.

Negative Manipulation. The existing text-guided image manipulation methods and StyleEntity do not perform well when tasked with removing certain attributes, such as the prompt "without beard". In such cases, our experiments have shown that employing a positive prompt "with beard" alongside a negative manipulation strength has proven to be more efficacious than the original method. Figure 11 shows the results of the negative manipulation, demonstrating that the attributes have been correctly removed.

5. Conclusion

We introduced StyleEntity, a framework for zero-shot image manipulation that utilizes named entities as proxies during training. This framework was augmented by our proposed PELA technique, which significantly enhanced the stability of the results. Following a single training phase, our model demonstrated remarkable adaptability to a broad range of textual prompts not encountered during training. Subsequent qualitative and quantitative experiments validated the superior performance of our model.

Acknowledgment

This work was supported by National Natural Science Foundation of China (62271359).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 1

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 1

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 1

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40 (4):1–12, 2021. 1

[5] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 1

[6] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 6

[7] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 1

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 4

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 1

[12] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 2

[13] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019. 2

[14] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017. 2

[15] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023. 6

[16] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[18] Erik Härkönen, Aaron Hertzman, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *IEEE Conference on Neural Information Processing Systems;*, 2020. 1, 2

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 4

[20] Xianxu Hou, Xiaokang Zhang, Hanbang Liang, Linlin Shen, Zhihui Lai, and Jun Wan. Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. *Neural Networks*, 145:209–220, 2022. 1

[21] 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. 1

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *IEEE Conference on Neural Information Processing Systems;*, 2020.

[23] 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.

[24] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 1

[25] 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

[26] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[28] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1357–1365, 2020. 2

[29] Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6894–6903, 2023. 1, 6, 7

[30] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 42–51, 2018. 2

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[32] 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. 1, 2, 6, 7

[33] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2

[34] 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. 1, 2

[35] 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. 2

[36] 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. 6

[37] 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

[38] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2

[39] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 1

[40] 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. 1, 2

[41] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020. 1

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[43] Binxu Wang and Carlos R Ponce. The geometry of deep generative image models and its applications. *arXiv preprint arXiv:2101.06006*, 2021. 1

[44] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 1

[45] 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

[46] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 2

[47] Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, and Errui Ding. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18229–18238, 2022.

[48] Yiming Zhu, Hongyu Liu, Yibing Song, Ziyang Yuan, Xintong Han, Chun Yuan, Qifeng Chen, and Jue Wang. One model to edit them all: Free-form text-driven image manipulation with semantic modulations. *Advances in Neural Information Processing Systems*, 35:25146–25159, 2022. 1, 2, 6, 7