Named Entity Driven Zero-Shot Image Manipulation
Supplementary Materials

A. Additional Implementation Details

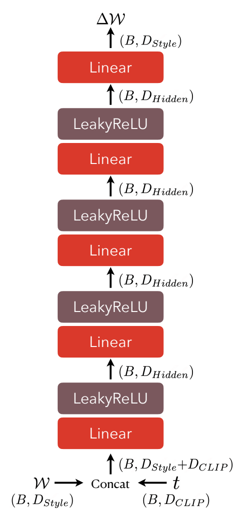

Figure 1. Architecture of the StyleEntity Mapper, where D_{style} represents the dimension of StyleGAN's style space, D_{CLIP} denotes the CLIP space dimension, and D_{hidden} is the dimension of the MLP's hidden layers.

Mapper Architecture Our mapper network is based on a 4-layer Multi-Layer Perceptron (MLP) architecture, as illustrated in Figure 1. We initialize the final linear layer weights and biases to zero, promoting stable training. The hidden layer dimension (D_{Hidden}) is set to 512.

Training Algorithm Our training algorithm, outlined in Algorithm 1, effectively leverages our novel approach to named entity driven image manipulation.

Named Entity Dataset Collection We employed three named entity text datasets in our experiments:

- **Celebrity-Names-90k**: Compiled from MS-Celeb-1M [2], this dataset includes 90,084 names of celebrities like singers, athletes, and actors.
- **Dog-Breeds-354**: Sourced from the Federation Cynologique Internationale (FCI)[1], this dataset encompasses texts from 354 different dog breeds.

[1] https://www.fci.be/en/Nomenclature/

Algorithm 1 Training algorithm for StyleEntity

Require: Pre-trained StyleGAN Generator G
Require: Mapper M initialized for training
Require: Manipulation Strength α
Require: Pre-computed Named Entities Text Embeddings $\{t_1, t_2, \ldots, t_n\}$
Require: Regularization Weighting Factor λ
while not converged **do**
  Sample t_i from $\{t_1, t_2, \ldots, t_n\}$
  Sample style code $\mathcal{W}$ from $\mathcal{W}^+$ space
  $\Delta\mathcal{W} = M(t_i, \mathcal{W})$
  Generate image $x = G(\mathcal{W} + \alpha\Delta\mathcal{W})$
  Compute regularization loss $\mathcal{L}_{regularization} = ||\Delta\mathcal{W}||^2$
  Compute contrastive loss $\mathcal{L}_{contrastive}$ using x and $\{t_1, t_2, \ldots, t_n\}$
  Calculate total loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{contrastive} + \lambda\mathcal{L}_{regularization}$
  Update Mapper M parameters by backpropagation to minimize $\mathcal{L}_{\text{total}}$
end while

- **Cat-Breeds-101**: This dataset contains texts of 101 cat breeds, gathered from Wikipedia [2].

B. Quantitative Evaluation Details

Evaluation Prompts To evaluate text-guided image manipulation methods, we devised 100 diverse prompts, covering hair color, hairstyle, beard style, mood, and more. Examples of these prompts are detailed in Table 1.

Trade-off Curve Construction For constructing trade-off curves, we systematically adjusted the inference hyperparameters for each model to generate a range of FID and CLIP scores. The manipulation strength in our model varied from 0.04 to 0.36, while FFCLIP [5] and DeltaEdit [3] featured a scaling coefficient ranging from 0.5 to 4.4.

[2] https://en.wikipedia.org/wiki/Category:Cat_breeds

	Examples
Hair Color	*Red Hair, Black Hair,* *Purple Hair, Green Hair, ...*
HairStyle	*Mohawk Hairstyle, Bob-cut Hairstyle,* *Curly Hair, Afro Hairstyle, ...*
Beard Style	*Full beard, Goatee,* *Mustache, Sideburns, ...*
Mood	*Angry, Disgust,* *Sad, Surprised, ...*
Others	*Chubby, Tanned Skin,* *Big eyes, Mouth Open, ...*

Table 1. List of prompts used for quantitative evaluation.

StyleCLIP-GD [4]'s manipulation strength was adjusted from 0.5 to 3.4.

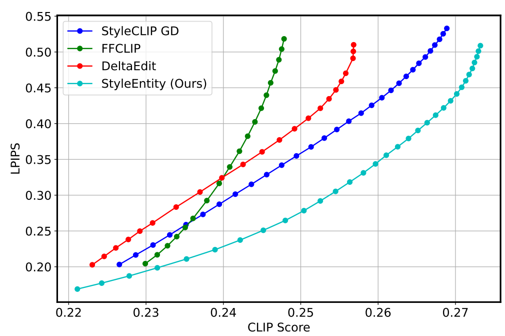

Figure 2. LPIPS-CLIP trade-off curves.

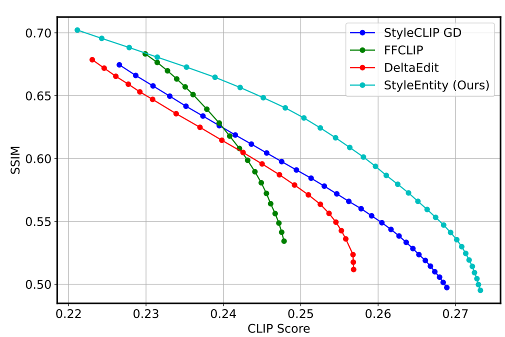

Figure 3. SSIM-CLIP trade-off curves.

Additional Trade-off Curves In addition to using the primary FID-CLIP curve for evaluation, we incorporated LPIPS-CLIP and SSIM-CLIP curves to assess perceptual and structural similarities. Furthermore, following the approach of InstructPix2Pix [1], we integrated CLIP image versus text-image similarity results, replacing direction similarity with cosine similarity. These additional metrics, as shown in Figure 2, Figure 3 and Figure 4, further demonstrate the effectiveness of StyleEntity.

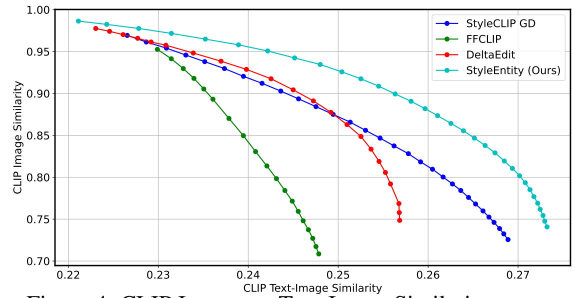

Figure 4. CLIP Image vs. Text-Image Similarity curves.

Human Evaluation Interface The user interface for our human preference study is depicted in Figure 5. Participants used this interface to assess the quality of images generated from various prompts.

References

[1] 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*, pages 18392–18402, 2023. 2

[2] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic imaging*, 2016 (11):1–6, 2016. 1

[3] 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

[4] 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

[5] 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

Figure 5. Human evaluation interface. Participants are required to blindly select the best result from six outputs that have been randomly ordered and generated by different models.