NoiseCLR: A Contrastive Learning Approach for Unsupervised Discovery of Interpretable Directions in Diffusion Models

Supplementary Material

S.1. Overview

In this appendix we include our supplementary material as follows:

- Additional comparisons with GAN-based and Diffusion-based editing methods in Sec. S.2.
- Re-scoring analysis for car domain in Sec. S.3.
- Compositional editing results in Sec. S.4.
- Edit interpolation results in Sec. S.5.
- Ablation studies in Sec. S.6.

S.2. Additional Comparisons

S.2.1. GAN-based Editing Methods

GAN-based editing methods are known to have superior editing capability due to their disentangled latent space [8]. In our analysis, we also include a comparison of NoiseCLR with several state-of-the-art GAN-based methods that find directions in the latent space in an unsupervised manner, LatentCLR [9], GANSpace [2], and SeFa [6] (see Fig. S.1). As can ben seen from the figure, our diffusion-based edits achieves competitive results when compared with its GAN-based counterparts.

S.2.2. Diffusion-based Editing Methods

In this section, we provide additional results across other methods using qualitative and quantitative comparisons.

S.2.3. Qualitative Comparisons

- **Diffusion-Pullback** [5] proposes an unsupervised direction discovery method in diffusion-based models. Their approach utilizes the pullback metric to identify latent bases for image editing and optionally incorporates text prompts. While they achieved promising results with DDPM-based models (note that this needs a *separate DDPM model for each domain such as face, cats, and so on*), they report that their application to Stable Diffusion didn't fully realize its potential. Specifically, their method uncovered only a limited number of directions in Stable Diffusion, e.g. only two reported for face edits in their paper. They noted that some of the latent vectors they discovered led to sudden and drastic changes during the editing process. This issue was attributed to the complex geometry of the latent space, which poses a challenge for achieving smooth and seamless edits.
  In their study, only two editing directions called 'overweight' and 'gender' were initially reported. For a fair comparison, we used the same input image from their paper and created edited results for these directions using our method. Additionally, we run their source code to discover two more directions, 'Old' and 'Race', and reported the results. Please refer to Fig. S.2.
  The comparisons demonstrate that our method not only executes edits more faithfully compared to [5], but it also uncovers a significantly greater number of directions.
- **Concept sliders** [1], a concurrent work with ours, either rely on text prompts or paired image data for editing images. For instance, to edit *eyebrow shape* of a face image, one would need a text prompt like 'eyebrows' or a pair of images showing the person *before* and *after* their eyebrow shape changes. This reliance on text prompts or paired data for defining edits aligns them with supervised methods in image editing.
  Please see Fig. S.3 for a comparison of their edits (found via providing text prompts defining the edits) vs. ours (found via unsupervised discovery). Although Concept Sliders are capable of accomplishing the intended edit to some degree, our method stands out by remaining more faithful to the original input image and ensuring the edits are disentangled. For instance, Concept Sliders often alter the facial shape (as observed in the Race edit) and mix changes in the face with aging in the mustache edit, leading to entangled edits.
- **Prompt2Prompt** [3] is an image editing method that uses cross-attention, and requires both a source and target text prompt. We compare our editing results with Prompt2Prompt in Fig. S.3. Note that since their method does not discover a direction, they are only able to perform a single edit and do not have the ability to control the scale of the edit applied to the image, which limits their usage. On the other hand, the requirement of a source prompt poses another limitation, which might not be feasible in domains such as art. Nevertheless, our results show that our method is able to perform the desired edits in a much more disentangled way while being faithful to the input image. Notice that Prompt2Prompt (P2P) tends to significantly modify the original image, deviating considerably from the initial input as can be seen from Age edit (Fig. S.3 bottom left) or performs unrealistic edits as in Mustache edit (Fig. S.3 top right).
- **DiffusionDisentanglement** Note that although we intended to compare our results with those from Wu et al. [7], an editing method that optimizes weights to perform disentangled edits given a text prompt, we were unable to

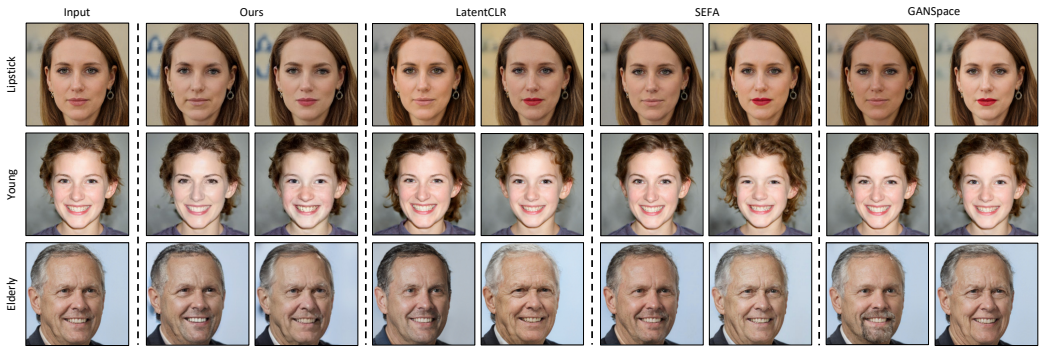

Figure S.1. **Comparisons with GAN-based Latent Discovery Methods.** We also compare NoiseCLR with latent direction discovery methods on GANs. As it is also demonstrated in our results, our editing & direction discovery method produces competitive results with GAN-based methods, in terms of fine-grained face editing.

do so due to the high GPU requirements of their method (namely, 48GB for a single edit). [1]

Method	Race	Mustache	Age	Gender
Prompt2Prompt	0.24	0.22	0.28	0.25
Concept Sliders	0.18	0.13	0.21	0.25
Ours	**0.15**	**0.12**	**0.18**	**0.21**

Table S.1. **LPIPS** [10] metric which measures how well the similarity to the original image distribution is maintained (lower is the better). Our method is able to achieve lower LPIPS than the other methods, indicating greater consistency with the input image while performing the edits.

S.2.4. Quantitative Comparisons

We also compared Concept Sliders [1] and Prompt2Prompt [3] methods in a quantitative way using LPIPS [10] metric which measures how well the similarity to the original image distribution is maintained. Table S.1 shows the results for *Race, Mustache, Age, Gender* edits. As can be seen from the LPIPS metrics, our method is able to achieve lower LPIPS than the other methods, indicating greater coherence while performing the edits.

S.3. Re-scoring Analysis on Car Domain

In addition to the domain of face images, we also conduct a re-scoring analysis for car domain. We present the analysis results in Table S.2, where we evaluate the variation in CLIP classification probability for the attributes listed as the columns in response to the edits indicated in the rows. As expected, the changes that alter the car type affect the scores for each other. For instance, transforming the car type into *sport* leads to a reduced probability score for car type *pickup*. However, when the apply the discovered *background* (Nature) direction, we notice a noticeable entanglement, particularly with edits related to *sport* and *muscle* car body types. While acknowledging that our method discovers more entangled directions in car domain compared to other domains, we refer to the inherent biases in SD for such entanglement issues.

	Nature	Pickup	Sport	Muscle	Wagon
Nature	**8.3**	-6.8	-10.3	-18.4	5.1
Pickup	-11	**22.9**	-36.3	-12.1	39
Sport	-7.1	11.1	**31.8**	5.1	-18
Muscle	-2.1	17.3	19.7	**33.9**	8.2
Wagon	-2.9	-3.1	-34.1	-18.6	**43.1**

Table S.2. **Re-scoring Analysis for car Domain.** The change in classification probability of the CLIP classifier for various attributes in the car domain. The numbers shown in bold indicate that NoiseCLR successfully enhances the target semantics across all other attributes. For car domain, we perform our analysis on edits related to car types (pickup, sport, muscle and wagon), and background (nature).

S.4. Composing Edits

Since NoiseCLR can learn latent directions from different domains within the shared latent space of Stable Diffusion, it is capable of executing both intra-domain and inter-domain edits (see Fig. S.4):

1. **Within the same domain**, where multiple face edits can be applied simultaneously to a single image, as depicted in Fig. S.4 (a). Using the same Stable Diffusion model, edits in the face domain can be simultaneously applied to a single image, allowing for changes like altering *race* and adding a *mustache*, as illustrated in Fig. S.4 (a), top row. Similarly, in the cat domain, our method can concurrently apply edits affecting *eye color* and *lion*, as

[1] https://github.com/UCSB-NLP-Chang/DiffusionDisentanglement/issues/6.

Figure S.2. Our comparison with Diffusion-Pullback [5] focuses on the *overweight* and *gender* edits, the sole directions provided in [5] (utilizing the unsupervised version of their method on Stable Diffusion for face edits). The additional directions, *Old* and *Race*, were identified by ourselves after applying their method to 50 directions. The comparisons clearly demonstrate that our method not only executes edits more faithfully compared to [5], but it also uncovers a significantly greater number of directions, as detailed in the main paper.

shown in Fig. S.4 (a), middle row. Our method can also combine multiple styles in Art domain, as shown in Fig. S.4 (a), bottom row.

2. **Across different domains**, enabling the application of edits from various domains on the same image simultaneously. For instance, a face edit and a cat edit can be combined together. Moreover, our method can apply edits in the *car* domain to transform a car into a *sports car*, while keeping its original color and preserving the *background* (Fig. S.4 (b), top row). Concurrently, it can alter the *gender* of a person in the image using a face edit. In the same vein, within the fashion domain, our method can change the *color* of a dress, while a face domain edit can modify the *race* of the person wearing the dress (Fig. S.4 (b), middle row). Our method can also combine face and Art directions, as shown in Fig. S.4 (b), bottom row.

S.5. Interpolating Edits

Our method enables users to modulate the editing effect using a scale parameter. As illustrated in Fig. S.5, it can perform edits along both negative and positive scales. This feature allows users to either diminish or amplify the effect of the editing direction. For example, with the 'Age' direction, users can reduce the aging effect or increase it when applied with a positive scale. Additionally, our method achieves these interpolations in a disentangled manner, ensuring that the edits in both positive and negative directions remain faithful to the original image.

Figure S.3. Comparison on Race, Mustache, Age and Gender attributes with our method, Concept Sliders[1] and Prompt2Prompt (p2p) [3]. Although all methods are capable of accomplishing the intended edit to some degree, our method stands out by remaining more faithful to the original input image and ensuring the edits are disentangled. For instance, Prompt2Prompt (P2P) tends to significantly modify the original image, deviating considerably from the initial input. Similarly, Concept Sliders often alter the facial shape (as observed in the Race edit) and mix changes in the face with age attributes in the mustache edit, leading to entangled edits.

S.6. Ablation Study

In this section, we perform ablation regarding learning from fake/real data, ablation on number of input images N, ablation on number of directions K and ablation on timesteps that edit is applied.

S.6.1. Ablations on Learning from Fake/Real Data

In our ablation study, we explored the potential of our method to learn from synthetic images. We used single text prompts, such as *a face of a person*, to generate these fake images, focusing on the domain without specifying particular attributes. Our experiments revealed that images

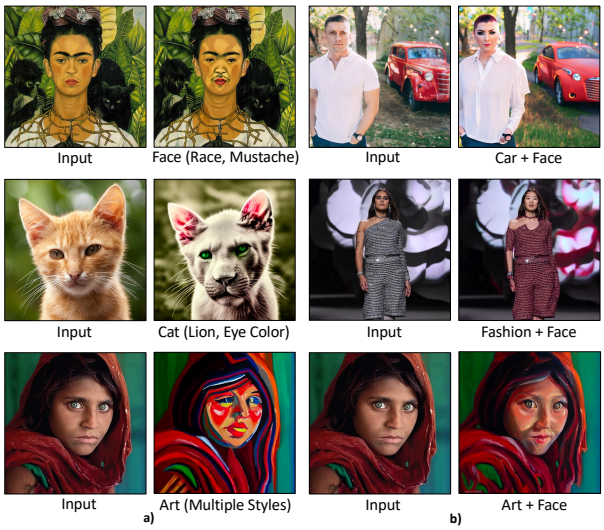

Figure S.4. **Intra-domain and Cross-domain Editing.** Our method can find domain-specific edits that can be composed either a) intra-domain where edits from the same domain can be applied simultaneously, b) cross-domain, where edits from different domains can be combined and applied simultaneously.

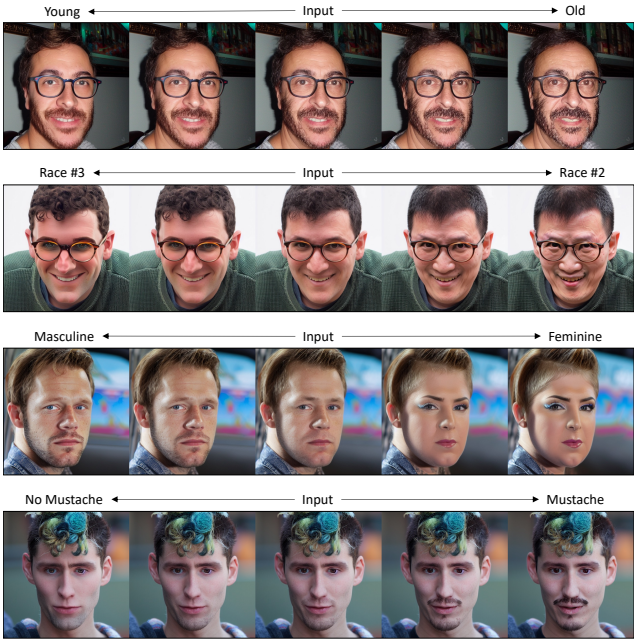

Figure S.5. **NoiseCLR Interpolation Results.** Our method allow users to control the editing effect using a scale parameter. This feature allows users to either diminish or amplify the effect of the editing direction. For example, with the 'Age' direction, users can reduce the aging effect (Young) or increase it when applied with a positive scale (Old).

randomly generated by Stable Diffusion (SD) often exhibit artifacts and biases, potentially affecting learning stability. Despite this, when using 100 fake images, our method successfully identified diverse directions, such as race. Fig. S.6 (rightmost column) compares Gender direction learned from real (left) and fake (right) images. Both models effectively performed disentangled edits. However, the range of discovered directions using fake images was significantly narrower compared to real images, in particular we were only able to discover *age, race (asian), race (indian), gender, mustache, chin shape, child* and *cartoon* directions. This limitation could be attributed to Stable Diffusion's tendency to produce flawed images with issues like crooked teeth or other artifacts, which can obstruct the learning process.

S.6.2. Ablations on N

Our method requires only a small set of images to learn domain-specific directions. We have found that $N = 100$ images are generally sufficient for learning a rich and diverse range of directions. To explore the impact of the number of images on the discovery of directions, we conducted an ablation study using $N = 10, 100, 1000$ images randomly selected from the FFHQ [4] dataset, aiming to learn face-specific edits while keeping the number of directions $K = 100$ constant. Our findings indicate that our method can still learn directions with as few as $N = 10$ images, but the resulting directions often perform more coarse-grained edits, as shown in "Race" edit in Fig. S.6 (first column). We believe this is due to the limited number of samples ($N = 10$) available for learning our contrastive loss, providing too few positive and negative pairs to effectively learn $K = 100$ directions. Conversely, when comparing $N = 100$ and $N = 1000$ samples, our method demonstrates the ability to learn the same directions in both cases. This indicates that a sample size of $N = 100$ is sufficient for effectively learning directions.

S.6.3. Ablations on K

Our method includes a hyperparameter, K, which determines the number of directions to be learned. For varied domains like faces or art, we typically set $K = 100$, while for simpler domains like cats and cars, we choose $K = 50$. In this section, we conducted an ablation study on the impact of the K parameter in the face domain, keeping $N = 100$ constant, and experimenting with $K = 10, 50, 100$. We observed that when the model is constrained to learn a smaller set of directions, such as $K = 10$, it tends to focus on coarse-grained edits that edit the overall structure of the face, like race, age, overweight, or cartoon style. In contrast, increasing the number of directions to $K = 50$ or $K = 100$ leads to the discovery of more fine-grained edits, such as adjustments to lipstick, chin, eyebrows, etc. Fig.

S.6 (middle column) showcases "Race" edit discovered using $K = 10$ and $K = 100$. We also noticed that edits learned with $K = 10$ directions are slightly more entangled than those learned with $K = 100$. This could be due to the fact that directing the model to differentiate $K = 100$ directions from each other enforces disentanglement, whereas a smaller number of directions may lead the model to learn more entangled edits.

S.6.4. Ablations on timesteps

Prior work such as [7] and [3] has shown that timesteps are crucial factors affecting the disentangled editing capability of Stable Diffusion. Our method, while learning directions by considering all timesteps of the diffusion model, specifically modifies the noise prediction for a certain interval of timesteps to achieve more disentangled edits. As a rule of thumb, we apply the discovered edits starting from $t = 0.5T$ to achieve disentangled edits. However, for edits that require changes in the coarse structure of the input (e.g. eyeglasses), editing at earlier timesteps are required (within the interval $[0.9T, 0.8T]$). To demonstrate the effect of timesteps on the disentanglement property of the edits, we conduct an ablation study where we apply selected edits on different timestep intervals in Fig. S.7. In our ablations, we select the number of denoising steps as 50 and demonstrate the edited images w.r.t. denoising step indices where the edit is applied. As shown in our results, applying edits at earlier denoising steps result in more significant changes in the input image whereas edits at later iterations succeed in preserving the coarse structure of the input.

References

[1] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023. 1, 2, 4

[2] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 1

[3] 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. 1, 2, 4, 6

[4] 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. 5

[5] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *arXiv preprint arXiv:2307.12868*, 2023. 1, 3

[6] 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

[7] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2023. 1, 6

[8] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2022. 1

[9] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14263–14272, 2021. 1

[10] 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 conference on computer vision and pattern recognition*, pages 586–595, 2018. 2

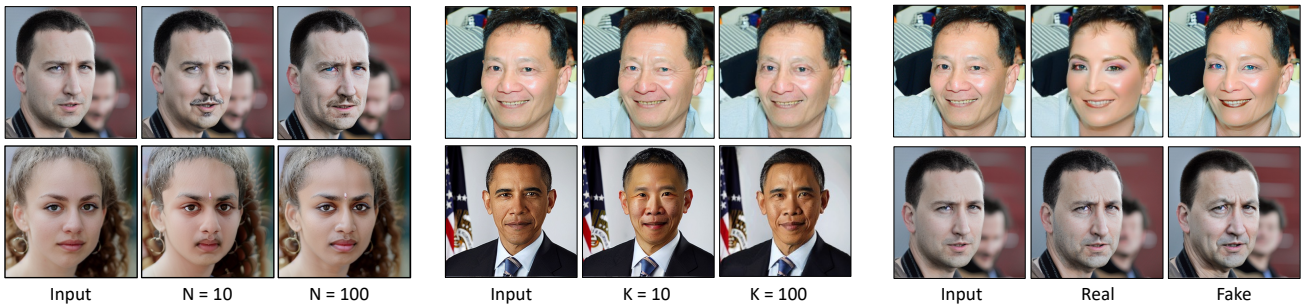

Figure S.6. **Ablation study results.** We perform our ablations on three different variables, which are the usage of real/fake samples, different number of directions(K) and number of samples used for training the model(N). For each of our ablations, we demonstrate qualitative results on two different edits learned by each variant.

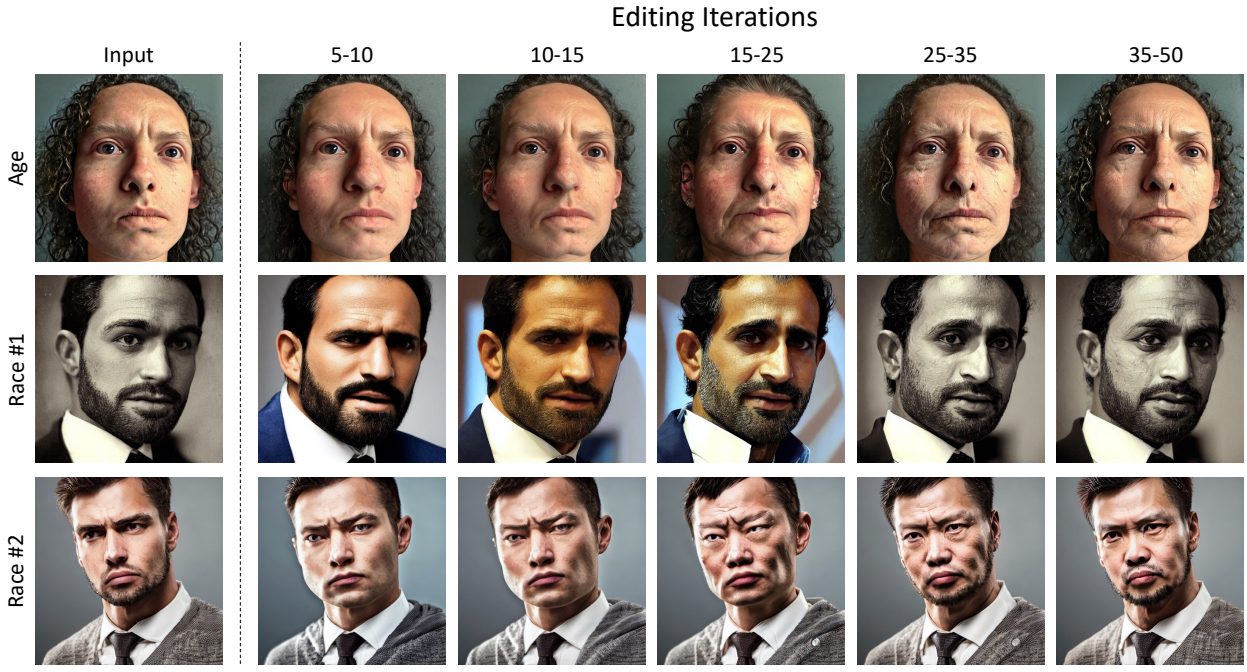

Figure S.7. **NoiseCLR Ablation study on editing interval.** We demonstrate the effect of editing timesteps with ablations over age and race edits. We perform our experiments over 50 denoising steps over images generated with Stable Diffusion. For clarity, we demonstrate the applied edits on the editing iterations where iteration 0 corresponds to $t = T$ whereas iteration 50 stands for $t = 0$.