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

Yusuf Dalva Pinar Yanardag
Virginia Tech
{ydalva, pinary}@vt.edu

Project webpage: https://noiseclr.github.io

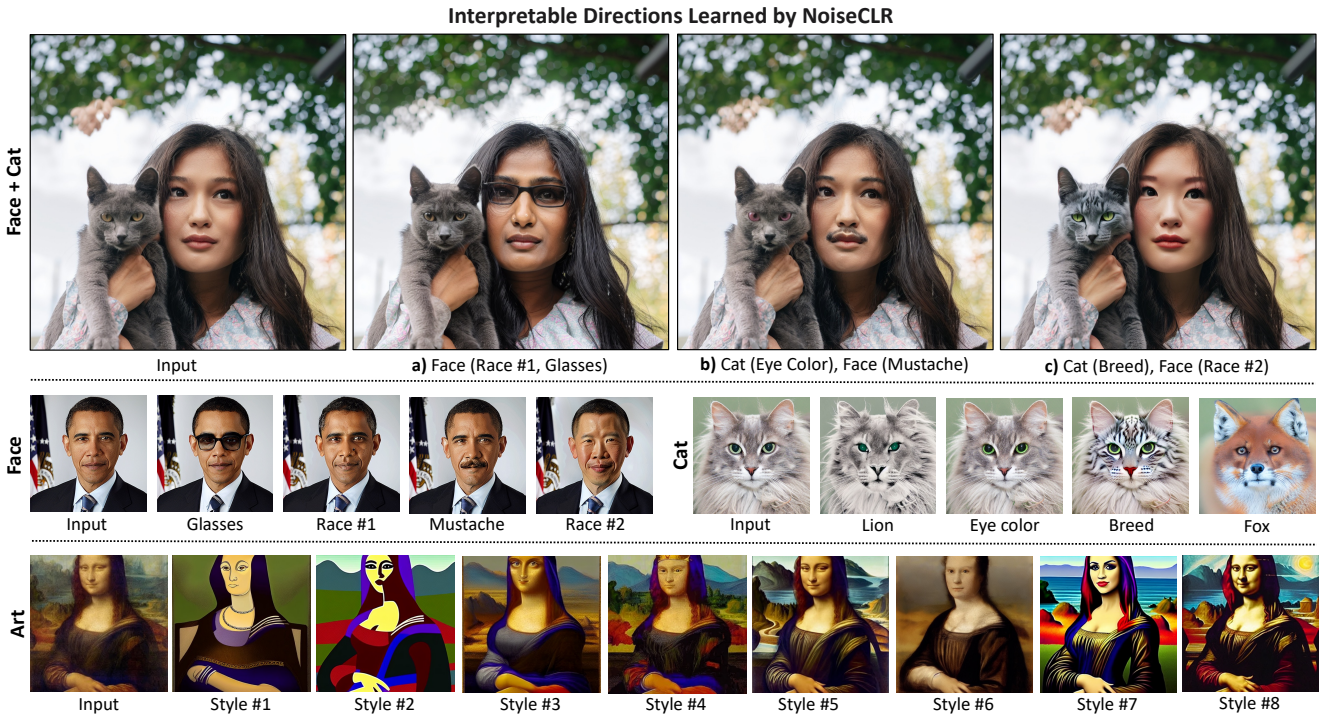

Figure 1. **NoiseCLR.** We propose an unsupervised approach to identify interpretable directions in text-to-image diffusion models, such as Stable Diffusion [30]. Our method finds semantically meaningful directions across various domains like *faces, cats*, and *art*. NoiseCLR can apply multiple directions either within a single domain (a) or across different domains in the same image (b, c) in a disentangled manner. Since the directions learned by our model are highly disentangled, there is no need for semantic masks or user-provided guidance to prevent edits in different domains from influencing each other. Additionally, our method does not require fine-tuning or retraining of the diffusion model, nor does it need any labeled data to learn directions. *Note that our method does not require any text prompts, the direction names above are provided by us for easy understanding.*

Abstract

Generative models have been very popular in the recent years for their image generation capabilities. GAN-based models are highly regarded for their disentangled latent space, which is a key feature contributing to their success in controlled image editing. On the other hand, diffusion models have emerged as powerful tools for generating high-quality images. However, the latent space of diffusion models is not as thoroughly explored or understood. Existing methods that aim to explore the latent space of diffusion models usually relies on text prompts to pinpoint specific semantics. However, this approach may be restrictive in areas such as art, fashion, or specialized fields like medicine, where suitable text prompts might not be available or easy to conceive thus limiting the scope of existing work. In this paper, we propose an unsupervised method to discover latent semantics in text-to-image diffusion models without re-

lying on text prompts. Our method takes a small set of unlabeled images from specific domains, such as faces or cats, and a pre-trained diffusion model, and discovers diverse semantics in unsupervised fashion using a contrastive learning objective. Moreover, the learned directions can be applied simultaneously, either within the same domain (such as various types of facial edits) or across different domains (such as applying cat and face edits within the same image) without interfering with each other. Our extensive experiments show that our method achieves highly disentangled edits, outperforming existing approaches in both diffusion-based and GAN-based latent space editing methods.

1. Introduction

Denoising Diffusion Models (DDMs) [14] and Latent Diffusion Models (LDMs) [30] have received considerable attention for their ability in generating high-quality, high-resolution images across a variety of domains. They have achieved remarkable outcomes in the field of generative modeling, particularly with text-to-image models like Stable Diffusion [30] which inspired researchers to employ them for image editing tasks through text prompts or various conditions such as scribble or segmentation maps [44].

A fundamental aspect of image editing in generative models is the disentangled application of semantics, which involves making changes that are semantically significant to specific areas of the image without affecting unintended regions [23, 41]. Previous research has demonstrated that generative adversarial networks (GANs) are particularly effective at disentangled image editing due to their structured latent space, leading to significant research in both supervised and unsupervised exploration of the latent directions in GANs [11, 32, 43].

However, while identifying directions in the latent space of GANs is relatively straightforward, such as using principal component analysis on sampled latent vectors to discover semantically meaningful directions [11], uncovering directions in diffusion models in an unsupervised manner is more challenging. This difficulty arises from the inherent design of diffusion models, which estimates the forward noise independently of the input and manage a significant number of latent variables over several recursive timesteps, unlike the more direct approach in GAN-based models. Therefore, most of the prior work that provides fine-grained control over the generation process in diffusion-based models focus on simple solutions such as blending latent vectors, model fine-tuning, embedding optimization [1, 2, 12]. However, these methods depend on user-provided text prompts to pinpoint specific semantics, e.g. *'A photo of a woman with an eyeglass'*. This approach can be restrictive in areas such as *art, fashion* where appropriate text prompts might not be straightforward to create,

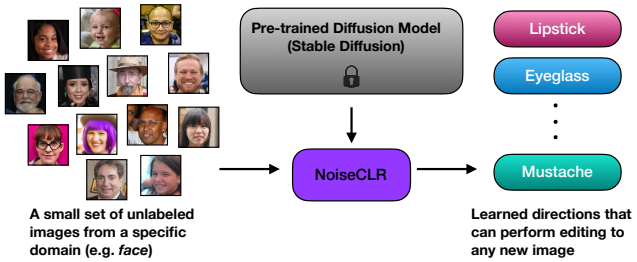

Figure 2. **NoiseCLR in a nutshell**. Our method employs a pre-trained diffusion model such as Stable Diffusion [30], alongside a small collection of *unlabeled* images from a specific domain such as *faces* or *cats* and learns diverse directions in an unsupervised fashion using a contrastive learning objective. The discovered directions can perform disentangled edits, such as allowing for semantically meaningful edits or adding *lipstick* or *eyeglasses* to any new image.

or in specialized fields such as the medical domain, which demand extensive domain knowledge to create appropriate text prompts. This limitation highlights the significance of discovering directions in the latent space in an unsupervised manner.

A number of strategies have been introduced to systematically investigate the latent directions within diffusion models [19, 26]. However, much of the existing research acknowledges limitations when working with large models like Stable Diffusion, often opting for simpler diffusion models such as DDPM [8, 19, 26]. These methods fail to fully exploit the capabilities of large-scale models such as Stable Diffusion and rely on separate DDPM models for each domain to identify directions. Therefore, despite significant advancements, a thorough exploration of the latent space in large diffusion-based models like Stable Diffusion remains an ongoing challenge. Discovering directions in latent space of a diffusion-based models is essential not only in the context of image editing but also for a broad spectrum of other applications. First, it allows more precise control over the image generation process, thereby significantly enhancing the versatility and applicability of the model across a variety of creative and specialized domains. Second, this approach fosters a more transparent and insightful exploration, demystifying what is often seen as a 'black-box' model, thus making its latent space more understandable. Thirdly, these insights enhance trust and reliability in the model and could be instrumental in identifying and mitigating potential biases, thus fostering further research in the ethical domain.

To the best of our knowledge, our approach is the first unsupervised method that successfully discovers directions in the latent space of Stable Diffusion in a disentangled manner to the extent of combining multiple directions within and across various domains (see Fig. 1). Our contri-

butions are as follows:

- We propose NoiseCLR, a contrastive-learning based framework to discover semantic directions in a pre-trained text-to-image diffusion model such as Stable Diffusion. Our approach does not need textual prompts, labeled data, or user-guidance, relying on a relatively small number of images (around 100) related to the target domain (see Fig. 2).
- Our method demonstrates the ability to discover diverse and fine-grained directions across diverse categories, such as face, cars, cats, and artwork.
- Our directions are highly disentangled, can apply multiple directions either within a single domain or across different domains. Our experiments demonstrate that our method can perform edits that are competitive with both state-of-the-art diffusion-based and GAN-based image editing methods.

2. Related Work

Latent Space Exploration of GANs. Various techniques have emerged that harness the latent space of GANs for image manipulation [5, 6, 28]. Supervised methods often leverage pre-trained attribute classifiers to guide the optimization, facilitating the discovery of meaningful directions within the latent space. Alternatively, they employ labeled data to cultivate classifiers aiming directly at learning desired directions [7, 32]. Conversely, some studies have demonstrated the potential to identify semantically meaningful directions within the latent space without supervision [15, 31, 36, 38, 43]. More recent work on GAN-based latent space explorations are pivoting towards utilizing image-text alignment methods such as StyleCLIP [27].

Latent Space Exploration of Diffusion Models. As diffusion-based image generation models are able to synthesize images from various domains, they encode semantically rich content in the form of a latent representations. In order to benefit from these representations, studies attempted to make use of the semantics encoded in the latent space. As a natural extension of latent space discovery, some works [19, 39] attempted to apply image editing by modifying the backward diffusion path using representations learned from the latent variables. While [19] formulate their transformation relying on the features learned by the bottleneck block of the denoising model, [39] apply modulation to the latent variables for a target domain, using stochastic diffusion models. In more recent efforts, [26] offered a framework to discover latent-specific directions encoding different semantics, inspired by latent space discovery literature in GANs. Even though their approach succeeds in discovering directions in single-domain diffusion models such as DDPMs, their proposed method fails in large-scale diffusion models, such as Stable Diffusion. Moreover, [21] decomposes images into a set of composable energy functions representing concepts such as *lighting* or *camera position*. However, their approach is particularly limited in terms of the number of concepts that can be learned due to memory constraints of their method, and they can only learn conceptualized representations rather than fine-grained latent directions.

Image Editing with Diffusion Models. The field of image generation has seen a growing interest towards utilizing diffusion models for editing tasks. One common approach involves supplying text prompts describing the intended edit. Yet, many implementations result in entangled edits, where unintended sections of the image are altered alongside the target area. Exceptions to this trend can be seen in works like [12, 44] which demonstrate more precise control over the editing process. For instance, ControlNet [44] leverages conditional diffusion model to allow users to manipulate specific image attributes by providing conditions. Likewise, [37] allows to perform content-preserving edits by overfitting the diffusion model to the input image. Additionally, [10, 24, 40] offer faithful reconstruction of the input image, which makes content-preserving edits with classifier-free guidance possible. Despite being able to preserve the input image while editing, such methods require per-image optimization, which is a bottleneck against real-time image editing. In recent efforts, [39] attempts to perform the image editing task by modifying the denoising process of a stochastic diffusion model for the real-editing task. Even though such approaches promise realistic image editing, constructing the ideal prompt for editing is a bottleneck against achieving realistic edits while staying faithful to the original image. To address the flexibility problem, [2, 20] proposed to compose the desired edit into multiple counterparts. However, these methods face difficulties when applying multiple edits, resulting in entangled results when several changes are made to the same image.

Contrastive Learning. Contrastive learning has gained traction recently, achieving state-of-the-art results in various unsupervised representation learning tasks. Its principle lies in learning representations by contrasting positive pairs against negative ones [9]. This approach has found applications in numerous computer vision tasks, including data augmentation [3], diverse scene generation [34], random cropping, and flipping [25]. In the context of diffusion models, approaches structured with a contrastive setup enabled tasks such as style transfer [42], and representation learning [35]. LatentCLR [43] introduces a method of contrastive learning to identify latent directions in GAN-based models by exploring feature divergences within an intermediate representation of latent vectors. While sharing similar

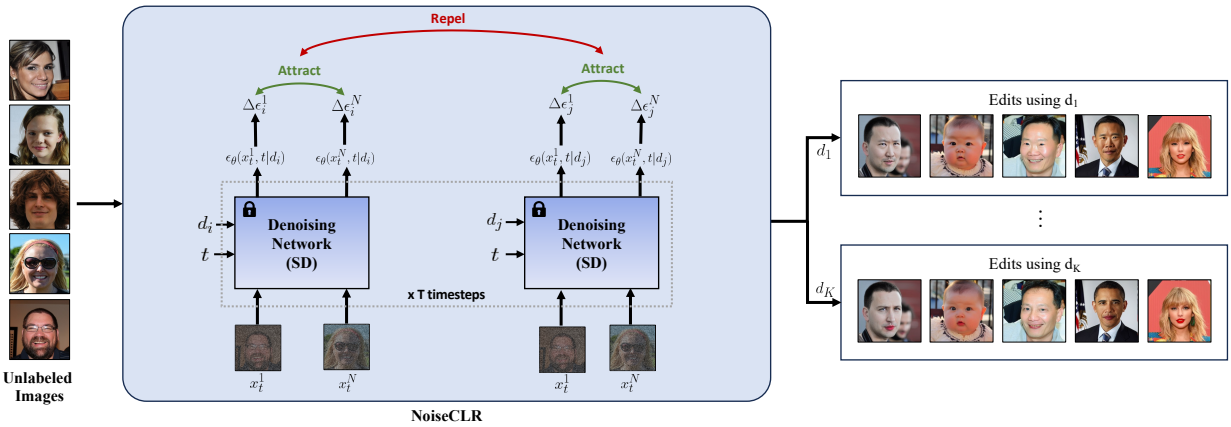

Figure 3. **NoiseCLR Framework.** NoiseCLR employs a contrastive objective to learn latent directions in an unsupervised manner. Our method utilizes the insight that similar edits in the noise space should attract to each other, whereas edits made by different directions should be repelled from each other. Given N unlabeled images from an particular domain such as facial images, we first apply the forward diffusion process for t timesteps. Then, by using the noised variables $\{x_1, ..., x_N\}$, we apply the denoising step, conditioning this step with the learned latent directions. Our method discovers K latent directions $d_1, \ldots, d_K$ for a pretrained denoising network such as Stable Diffusion, where directions correspond to semantically meaningful edits such as *adding a lipstick*.

intuitions in terms of utilizing contrastive learning for direction discovery, our approach diverges from LatentCLR. Unlike LatentCLR, which operates on latent vectors sampled from the GAN model, we focus on noise estimations, spanning across multiple diffusion steps. Notably, discovering directions in text-to-image diffusion models is considerably more complex than in GANs. This complexity arises because diffusion models independently estimate forward noise, irrespective of the input, and maintains a significant amount of latent variables across several recursive timesteps.

3. Method

In this section, we describe our proposed method, NoiseCLR, on discovering interpretable directions. First, we briefly discuss background on denoising probabilistic diffusion models.

3.1. Denoising Probabilistic Diffusion Models

Diffusion models [14, 30, 33] are generative models that produce data samples through an iterative denoising process, which is often referred as the reverse process. The reverse process involves a set of noise levels $t \in \{1, ..., T\}$, $\epsilon^t = \alpha^t \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. The denoising network, ϵ_θ, is designed to estimate the noisy component ϵ from noised image x_t during the reverse process where x_t refers to the noised version of the real image x_0 with a noise level of ϵ^t. The objective function for training such a denoising network is formulated as shown as:

$$\mathcal{L}_{DM} = \mathbb{E}_{x_0, \epsilon^t \sim \mathcal{N}(0,1), t}\left[||\epsilon^t - \epsilon_\theta(x_t, t)||_2^2\right] \quad (1)$$

To generate an image using the denoising network ϵ_θ, the reverse process is initiated with input $x_T \sim \mathcal{N}(0, 1)$. Throughout the reverse diffusion process, the variable x_t is iteratively denoised to get x_0 where $t \in \{1, ..., T\}$. The iterative denoising process is formulated as Equation 2 for step size γ and timestep t.

$$x_{t-1} = x_t - \gamma \epsilon_\theta(x_t, t) + \xi, \; \xi \sim \mathcal{N}(0, \sigma_t^2 I) \quad (2)$$

Classifier-free guidance [13] offers a way for conditioned sampling through subtle adjustments in both forward and backward diffusion processes with a specified condition c. By training ϵ_θ compatible with classifier-free guidance, conditional image generation becomes possible by modifying the noise prediction $\epsilon_\theta(x_t)$ with conditional noise prediction, to get $\tilde{\epsilon_\theta}(x_t, c)$. For simplicity, we use $\epsilon_\theta(x_t)$ instead of $\epsilon_\theta(x_t, t)$ to represent the predicted noise for timestep t, as t is implicitly denoted with variable x_t. The predicted noise with classifier-free guidance, $\tilde{\epsilon_\theta}(x_t, c)$, is defined by Equation 3:

$$\tilde{\epsilon_\theta}(x_t, c) = \epsilon_\theta(x_t, \phi) + \lambda_g(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)) \quad (3)$$

where ϕ is null-text and λ_g is guidance scale.

3.2. Contrastive Learning Objective

The primary objective of NoiseCLR is to learn K semantically meaningful directions, $D = \{d_1, \ldots, d_K\}$, given a small set of N images, $X = \{x_1, \ldots, x_N\}$ in diffusion models in an unsupervised manner, where we formulate each direction as a conditional embedding with the same dimensionality with text embeddings. The intuition behind

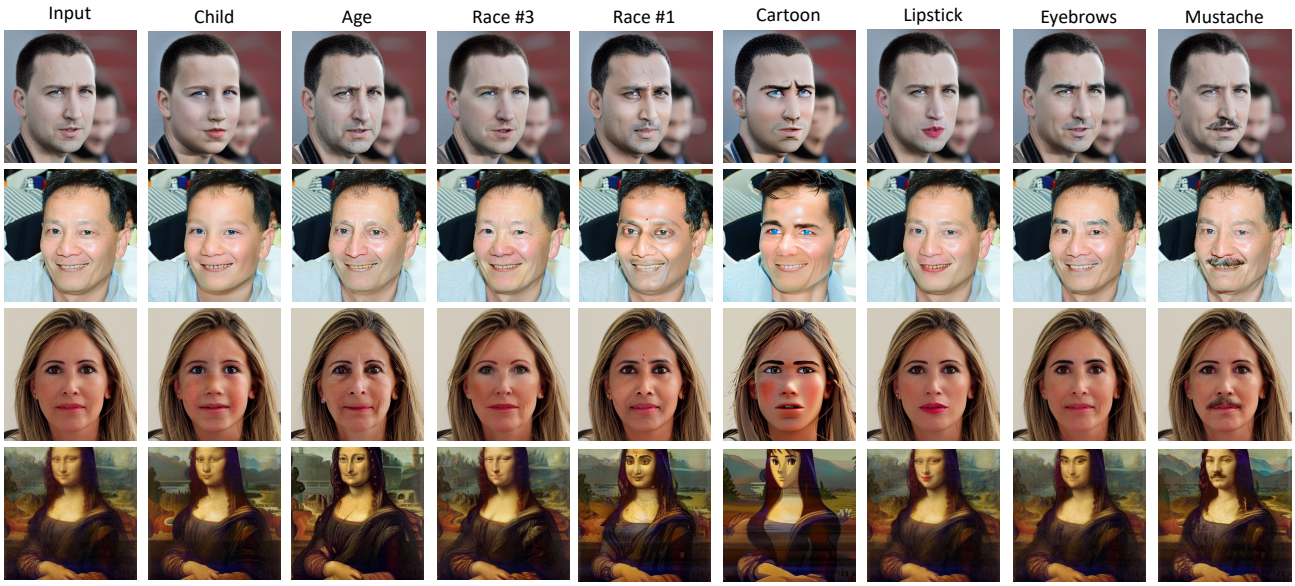

Figure 4. **Directions learned by NoiseCLR on face domain.** Edits are performed using the directions learned by our method in an unsupervised manner. We annotate the discovered directions above for the sake of understandability. The edits learned by our method are both effective in domain examples (e.g. human faces) and out-of-domain images (e.g. paintings).

NoiseCLR is best explained as defining an objective that encourages the similarity of the edits done by an arbitrary direction, while discouraging the similarity of edits performed by different directions. In other words, we want for edits carried out by the same direction to be attracted towards each other, while edits conducted by different directions to repel one another, in line with the core principles of contrastive learning. To formulate such an objective, we first define the feature divergences $\Delta\epsilon_k^n$ caused by an arbitrary direction d_k on an arbitrary data sample x_n as follows:

$$\Delta\epsilon_k^n = \epsilon_\theta(x_t^n, d_k) - \epsilon_\theta(x_t^n, \phi) \tag{4}$$

We define the target feature divergences obtained from d_j and a set of data samples $X' \subset X$ as our positive samples, whereas the target feature divergences for sample $x_i \in X'$ and a set of latent directions $D' \subset D - d_j$ are selected as the negative samples. We formulate our contrastive learning objective in Equation 5:

$$\mathcal{L}_j = -log \frac{\sum_{a=1}^{|X'|} \sum_{b=1}^{|X'|} \mathbf{1}_{[a \neq b]} \text{exp}(\text{sim}(\Delta\epsilon_j^a, \Delta\epsilon_j^b)/\tau)}{\sum_{a=1}^{|X'|} \sum_{i=1}^{|D'|} \mathbf{1}_{[i \neq j]} \text{exp}(\text{sim}(\Delta\epsilon_j^a, \Delta\epsilon_i^a)/\tau)} \tag{5}$$

To express the semantic similarity between a pair of target feature differences, we use cosine similarity which is formulated as:

$$\text{sim}(\Delta\epsilon_j^a, \Delta\epsilon_j^b) = \frac{\Delta\epsilon_j^a \cdot \Delta\epsilon_j^b}{||\Delta\epsilon_j^a|| \; ||\Delta\epsilon_j^b||} \tag{6}$$

Image Editing. Given the set of discovered directions $\{d_1, \ldots, d_K\}$, our editing scheme aims to reflect these semantics to input images in a disentangled manner. To perform such edits, we slightly modify Equation 3 with an editing direction d_e to obtain $\bar{\epsilon_\theta}(x_t, c, d_e)$ as formulated in Equation 7, where c serves as the condition used to generate the original image. Leveraging the observation that the difference $\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)$ encodes semantic information encoded by the condition c, we expand the noise prediction with the difference $\epsilon_\theta(x_t, d_e) - \epsilon_\theta(x_t, \phi)$ for the timestep where editing will be performed:

$$\bar{\epsilon_\theta}(x_t, c, d_e) = \tilde{\epsilon_\theta}(x_t, c) + \lambda_e(\epsilon_\theta(x_t, d_e) - \epsilon_\theta(x_t, \phi)) \tag{7}$$

where λ_e denotes the editing scale.

Editing in Multiple Directions. Since we formulate editing for an arbitrary direction as a summation of predicted noises for a given timestep t, we are able to perform multiple edits to a given input variable x_t. To perform a set L of discovered directions $\{d_1, \ldots, d_L\}$, we formulate the editing term as a sum of noise predictions, $\hat{\epsilon_\theta}(x_t, L)$, which is formulated as:

$$\hat{\epsilon_\theta}(x_t, L) = \sum_{i=1}^{|L|} \lambda_i(\epsilon_\theta(x_t, d_i) - \epsilon_\theta(x_t, \phi)) \tag{8}$$

Figure 5. **Editing results on various domains.** To demonstrate the generalizability of our method across different domains, we provide editing results on artistic paintings, cats and cars. As demonstrated from in the editing results, our method is able to learn and apply latent directions from various domains using a single diffusion model. Note that we label the edits only for the sake of understandability.

Using $\hat{\epsilon_\theta}(x_t, L)$, the overall noise prediction for timestep t is formulated as $\bar{\epsilon_\theta}(x_t, c, L) = \tilde{\epsilon_\theta}(x_t, c) + \hat{\epsilon_\theta}(x_t, L)$.

Real Image Editing. In addition to performing edits on generated images, we expand our editing approach such that the discovered edits are applicable to real images. Different than sampling fake images from $x_T \sim \mathcal{N}(0, 1)$, we initially apply DDIM Inversion [24] to obtain this initial variable x_T. Using this inverted variable, we reformulate $\bar{\epsilon_\theta}(x_t, c, d_e)$ as $\bar{\epsilon_\theta}(x_t, d_e)$ since the image is conditioned by the initial variable x_T only. The formulation for real image editing with a single direction is provided as follows:

$$\bar{\epsilon_\theta}(x_t, d_e) = \epsilon_\theta(x_t, \phi) + \lambda_e(\epsilon_\theta(x_t, d_e) - \epsilon_\theta(x_t, \phi)) \quad (9)$$

Note that, our approach in editing with multiple directions is also applicable for real images.

4. Experiments

To assess the effectiveness of NoiseCLR in identifying semantically meaningful latent directions and demonstrate the generalizability of our method, we conducted evaluations across various domains, including human faces, cats, cars, and paintings.

Experimental Setup We used Stable Diffusion-v1.5 for all of our experiments. We used several diverse datasets including FFHQ [16], AFHQ-Cats [4], and Stanford Cars datasets [18]. For the artistic domain, we perform our experiments on a small subset of paintings to discover latent directions corresponding to artistic styles. In our default setting, we train NoiseCLR with $N = 100$, $K = 100$ and $\tau = 0.5$. To optimize the directions, we use a learning rate of 10^{-3} and batch size of 6 for AdamW optimizer [22]. Throughout these experiments, we train our directions with relatively modest dataset sizes such as 100 for each domain. For face and painting domains, we set the size the set of directions to be learned as $|D| = 100$ and for the domains of cats and cars, we set the number of directions to be $|D| = 50$. In each experiment, we set the size of the subset of directions to be used $|D'|$ at every iteration as 20. To ensure the reproducibility of our experiments, we conduct all experiments with a fixed random seed of 0. Moreover, training our method on a single domain requires approximately 7 hours to learn 100 directions in face domain, and once trained, performing any edit in a zero-shot manner takes about 5 seconds using a single NVIDIA L40 GPU.

4.1. Qualitative Results

Unlike existing methods for exploring the latent space of diffusion models, our approach can identify latent directions of various domains using a single diffusion model. Since face images carry significant amount of variance in terms of facial features and is one of the most popular type of edits in both GAN and diffusion-based models, we first investigate the face editing capabilities of the directions discovered by NoiseCLR. Fig. 4 displays a variety of distinct directions, including broad edits that can change the overall structure of the face, like *aging* or *race*, as well as more detailed directions that modify fine-grained facial features,

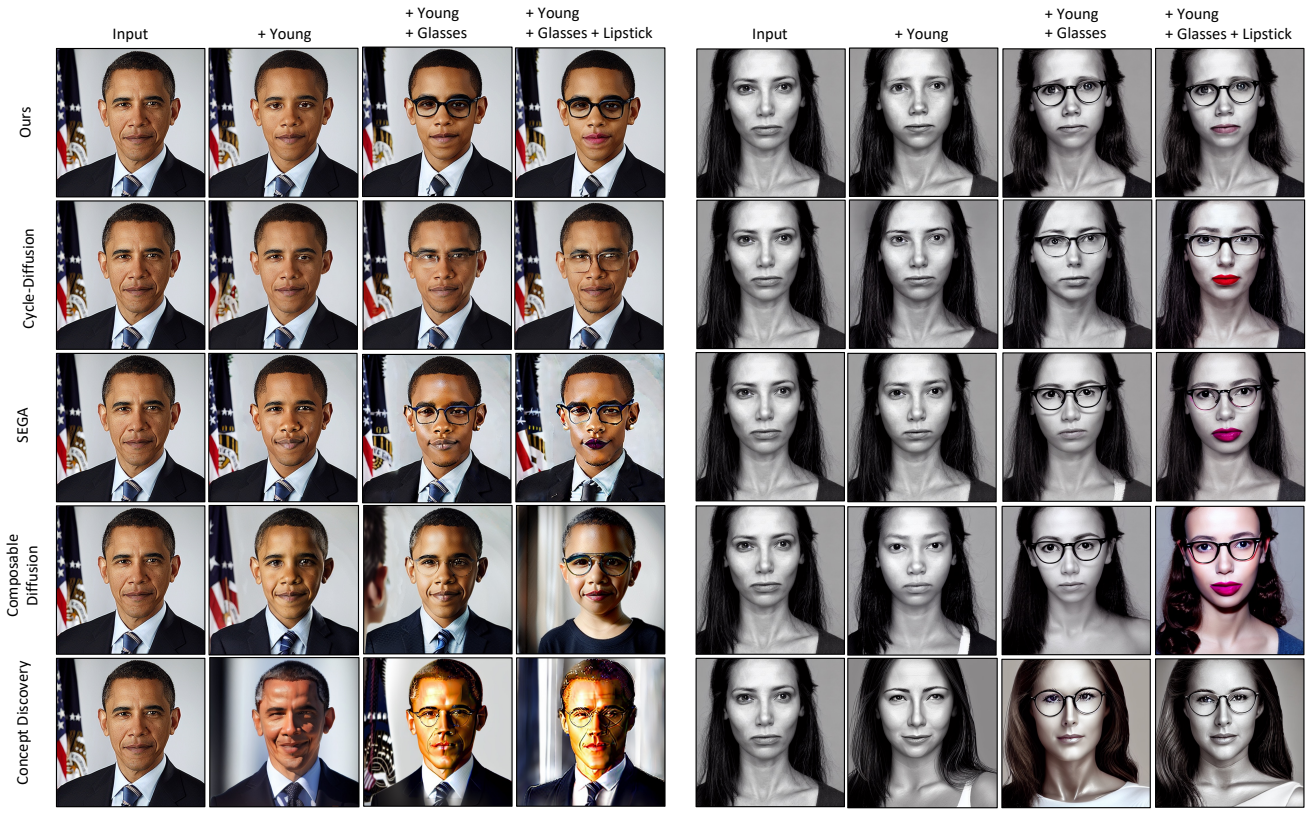

Figure 6. **Qualitative Comparisons.** We compare our method with Cycle-Diffusion[39], SEGA [2] and Composable Diffusion [20] in terms of image editing capabilities and Unsupervised Concept Discovery [21] to assess the quality of the representations learned by NoiseCLR. We present our comparisons for both real-image editing task and conditional image generation with the provided semantics. As it can be observed from the presented qualitative results, NoiseCLR succeeds over competing methods both in terms of disentangled image editing and learning fine-grained latent directions.

such as *lipstick* or a *mustache*. We emphasize that our edits are highly disentangled, meaning they achieve the intended modification without affecting any unintended parts. Given that our method conducts direction discovery in an entirely unsupervised manner, it has the freedom to explore the semantic space of the diffusion model during training. Consequently, NoiseCLR can discover directions not explicitly represented in the input dataset but still compatible with the domain of the training images, such as *cartoon* direction. Additionally, as NoiseCLR leverages the semantic understanding of the diffusion model, face edits generalize well to artistic paintings, as demonstrated in the last row of Fig. 4. Besides the facial domain, we also showcase the efficacy of our approach in the domains of art, cats, and cars. Qualitative results are presented in Fig. 5. As evident from the figure, NoiseCLR is capable of learning diverse semantics across various domains. This includes identifying multiple *artistic* styles, directions for transforming cats into foxes or lions, and directions for converting cars into trucks.

Qualitative comparisons. We compare our method with recent approaches, namely, Cycle-Diffusion [39], SEGA [2], Composable Diffusion [20], and Unsupervised Concept Discovery [21] methods. Over the compared methods, it is evident that [39] and [2] struggles with the real image editing task, when multiple semantics are modified even though the edits are performed in a disentangled manner. On the other end, the edits performed by [20] manages to preserve the edit quality, whereas they suffer preserving the image contents. As can be seen in Fig. 6, NoiseCLR outperforms the competing approaches both in terms of semantic faithfulness and disentanglement capabilities. For comparisons with diffusion and GAN-based editing methods, please refer to the supplementary material.

4.2. Quantitative Results

In evaluating the efficacy of discovered edits, we conduct a re-scoring analysis on the directions representing the semantics of "Indian", "Asian", "mustache", "child" and "lipstick" which are arbitrarily selected from the directions discovered by NoiseCLR. In our analysis, we assess the change in classification probability of the CLIP classifier [29] in the desired attribute and examine if the directions are dis-

	Indian	Asian	Mustache	Child	Lipstick
Indian	**29.8**	16.1	8.1	5.6	4.7
Asian	-10.5	**27.5**	0.0	-2.0	1.2
Mustache	3.6	-7.6	**48.9**	-13.2	1.7
Child	-37.7	-14.1	2.3	**32.8**	11.7
Lipstick	-8.9	-3.4	7.3	-0.7	**11.0**

Table 1. **Re-scoring Analysis.** The change in classification probability of the CLIP classifier for various attributes. Bold numbers indicate that NoiseCLR consistently enhances the target semantics across all attributes. Additionally, our approach achieves disentangled editing by minimizing its influence on other attribute scores when modifying a single attribute.

entangled between each other (see Table 1). In the presented scores, an increase corresponds to increased classification confidence for the subjected semantic, whereas a decrease implies the decrease of the presence of the semantic. Ideally, we expect the scores of the unedited semantics to change minimally, whereas the confidence for the edited semantic should increase. Relying on the re-scoring analysis performed, we consider our edits disentangled as the semantics that are not naturally related do not change significantly. However, we also acknowledge that applying the edit of "Child" significantly changes the race-based edits ("Indian", "Asian"). We relate this due to internal biases in Stable Diffusion. Moreover, we compared LPIPS [45] scores to measure how well the similarity to the original image distribution is maintained. Table 3 shows the results for several edits. The LPIPS metrics clearly demonstrate that our method consistently achieves lower LPIPS scores compared to other approaches, signifying improved coherence during the editing process.

User Study. We assess the editing capabilities of our model through an user study conducted on 50 participants using Amazon Mturk platform. For each of the methods that we compare with, we show volunteers several edits performed with common semantics and asked to determine if they consider the performed edit successful in terms of the given semantic, and if the edit is performed in a disentangled way. For each question, the users are asked to give a rating between 1-5 to indicate their preference where 5 means the highest score (see Table 2). Our results demonstrate that NoiseCLR attained higher scores in both edit quality and disentanglement evaluations, underscoring the superior performance of our approach.

5. Limitations

Our method is built upon the pre-trained Stable Diffusion model. Consequently, its manipulation capabilities

Method	Edit Quality ↑	Disentanglement ↑
Composable D.	2.19	2.15
Concept D.	1.20	1.28
SEGA	1.52	1.81
Cycle-Diffusion	1.87	2.73
Ours	**2.65**	**3.05**

Table 2. **User Study Results.** The average response score of the participants are provided in the table. The scoring is performed within the scale of 1-to-5.

Method	Age	Mustache	Gender	Race
Composable D.	0.19	0.40	0.42	0.40
Cycle-Diffusion	**0.10**	0.21	0.23	0.26
SEGA	0.11	0.23	0.27	0.27
Ours	0.17	**0.17**	**0.20**	**0.13**

Table 3. **LPIPS [45] scores (lower is the better).** Our method is able to achieve lower LPIPS than the other methods, indicating greater coherence while performing the edits.

are heavily dependent on the datasets Stable Diffusion was trained on, as well as the language model CLIP utilized by Stable Diffusion. While the joint representation capabilities of CLIP are impressive, they also have limitations and can exhibit biases towards certain attributes (e.g. entanglement between *background* direction and *car shape* attribute in Fig. 5). Furthermore, similar to other image synthesis tools, our framework raises concerns about potential misuse for malicious purposes [17].

6. Conclusion

We present an approach to discover to discover latent directions in large text-to-image diffusion models in an unsupervised way using a novel contrastive learning framework. Our method can combine multiple directions within and across various domains, such as face, cars, cats and artwork. Our experiments demonstrate that our method can perform edits that are competitive with both state-of-the-art diffusion-based and GAN-based image editing methods. Our approach not only provides more precise control over the image generation process, greatly expanding the model's versatility and usability in diverse creative and specialized fields, but it also promotes a more transparent and insightful exploration. This helps to demystify what is often perceived as a 'black-box' model. Furthermore, these insights increase trust and reliability in the model and could play a crucial role in identifying and addressing potential biases, thereby encouraging further research in ethical considerations.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 2

[2] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 7

[3] 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. 3

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 6

[5] Yusuf Dalva, Said Fahri Altındiş, and Aysegul Dundar. Vecgan: Image-to-image translation with interpretable latent directions. In *European Conference on Computer Vision*, pages 153–169. Springer, 2022. 3

[6] Yusuf Dalva, Hamza Pehlivan, Oyku Irmak Hatipoglu, Cansu Moran, and Aysegul Dundar. Image-to-image translation with disentangled latent vectors for face editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[7] 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. 3

[8] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. *arXiv preprint arXiv:2303.11073*, 2023. 2

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 1735–1742. IEEE, 2006. 3

[10] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Yuxiao Chen, Di Liu, Qilong Zhangli, et al. Improving negative-prompt inversion via proximal guidance. *arXiv preprint arXiv:2306.05414*, 2023. 3

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 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, 3

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[15] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. 3

[16] 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. 6

[17] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 8

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 6

[19] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2, 3

[20] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 3, 7

[21] Nan Liu, Yilun Du, Shuang Li, Joshua B. Tenenbaum, and Antonio Torralba. Unsupervised compositional concepts discovery with text-to-image generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2095, 2023. 3, 7

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[23] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International conference on machine learning*, pages 4402–4412. PMLR, 2019. 2

[24] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3, 6

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[26] 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. 2, 3

[27] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 3

[28] Hamza Pehlivan, Yusuf Dalva, and Aysegul Dundar. Styleres: Transforming the residuals for real image editing with stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2023. 3

[29] 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. 7

[30] 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, 2, 4

[31] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 3

[32] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 3

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[34] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 3

[35] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2023. 3

[36] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017. 3

[37] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. 42(4), 2023. 3

[38] 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. 3

[39] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023. 3, 7

[40] 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. 3

[41] 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. 2

[42] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22873–22882, 2023. 3

[43] 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. 2, 3

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[45] 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. 8