Instructing Text-to-Image Diffusion Models via Classifier-Guided Semantic Optimization

Yuanyuan Chang[1*], **Yinghua Yao**[2,3†], **Tao Qin**[1], **Mengmeng Wang**[4,5], **Ivor Tsang**[2,3] and **Guang Dai**[5]

[1]MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University
[2]Center for Frontier AI Research, Agency for Science, Technology and Research, Singapore
[3]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[4]Zhejiang University of Technology
[5]SGIT AI Lab, State Grid Corporation of China
1371306634@stu.xjtu.edu.cn, eva.yh.yao@gmail.com, qin.tao@mail.xjtu.edu.cn,
mengmengwang@zju.edu.cn, {ivor.tsang, guang.gdai}@gmail.com

Abstract

Text-to-image diffusion models have emerged as powerful tools for high-quality image generation and editing. Many existing approaches rely on text prompts as editing guidance. However, these methods are constrained by the need for manual prompt crafting, which can be time-consuming, introduce irrelevant details, and significantly limit editing performance. In this work, we propose optimizing semantic embeddings guided by attribute classifiers to steer text-to-image models toward desired edits, without relying on text prompts or requiring any training or fine-tuning of the diffusion model. We utilize classifiers to learn precise semantic embeddings at the dataset level. The learned embeddings are theoretically justified as the optimal representation of attribute semantics, enabling disentangled and accurate edits. Experiments further demonstrate that our method achieves high levels of disentanglement and strong generalization across different domains of data. Code is available at https://github.com/Chang-yuanyuan/CASO.

1 Introduction

Faithful image reconstruction is a fundamental prerequisite for image editing tasks. Recently, diffusion-based [Ho *et al.*, 2020; Song *et al.*, 2020; Rombach *et al.*, 2022; Song *et al.*, 2021] generative models have demonstrated significant advantages over GANs [Goodfellow *et al.*, 2014] in the field of image generation [Dhariwal and Nichol, 2021]. These approaches have emerged as powerful tools for image editing due to their generative modeling capabilities. However, exploring the decoupled semantic subspace of diffusion models remains a critical yet challenging area of research.

Most of the initial works on diffusion-based image editing require training or fine-tuning the model and focus on a single domain of data [Preechakul *et al.*, 2022; Lu *et al.*, 2024]. The rapid rise in popularity of text-to-image diffusion models, such as Stable Diffusion (SD) [Rombach *et al.*, 2022], DeepFloyd IF [Saharia *et al.*, 2022], and Latent Consistency Models [Luo *et al.*, 2023], has further inspired researchers to achieve decoupled image editing based on text or other conditions (e.g., semantic segmentation) without requiring additional training or fine-tuning [Brack *et al.*, 2023; Lee *et al.*, 2024]. Unlike small, single-domain diffusion models, these advanced models enable cross-domain generation, broadening their applicability across diverse image editing scenarios, have become a hot spot of current research. However, existing solutions are mainly based on text prompts [Brack *et al.*, 2023; Wu and De la Torre, 2023; Lee *et al.*, 2024], which present several challenges. Creating appropriate text prompts can be inherently difficult, and prompts with identical semantics often yield significantly divergent outcomes. Furthermore, achieving disentangled editing through text prompts alone is highly challenging, as they frequently affect unrelated regions of the image. Recent works that eschew text prompts proposed unsupervised ways to discover editing directions in the latent space of diffusion models [Chen *et al.*, 2024; Dalva and Yanardag, 2024] also

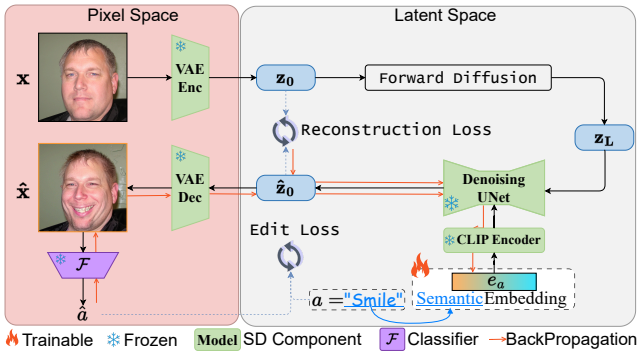


Figure 1: **ClA**ssifier-guide **S**emantic **O**ptimization (CASO). The trainable continuous semantic embedding for the target attribute a, guides Stable Diffusion for desired edits.

[*]Completed during the internship at A*STAR and SGIT AI Lab
[†]Corresponding author

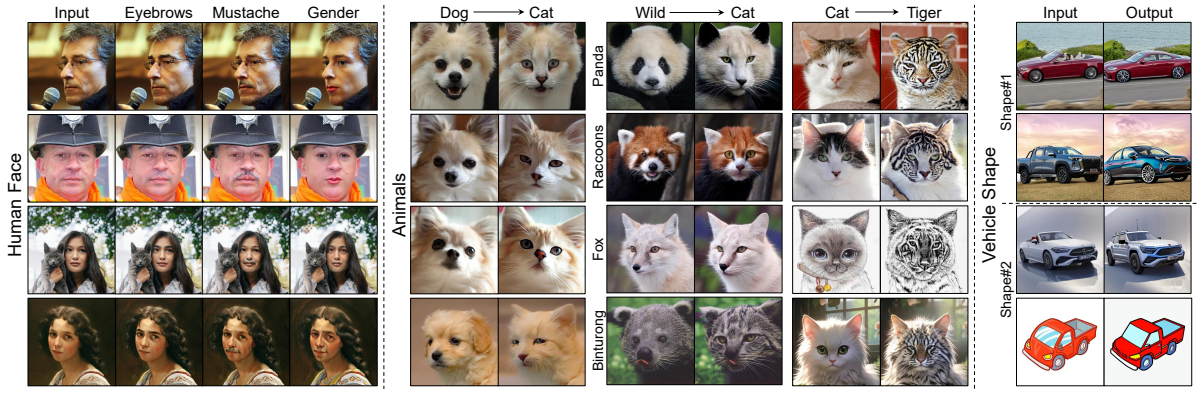


Figure 2: **CASO Edit Result.** Our method generalizes well to data with different styles.

exhibit limitations, particularly in terms of generalization. For instance, while these methods explore editable directions using real face data, their applicability often diminishes when extended to other types of images, such as artistic or anime-style faces. Moreover, they can only obtain a limited number of subtle editing directions, often lack the ability to specify attributes, and rely on the clear structure or inherent variance within the dataset itself.

In this paper, we propose optimizing continuous semantic embeddings through attribute classifier gradients to guide text-to-image models for targeted editing. Our framework is shown in Fig. 1. It can identify disentangled and generalizable editing directions. Our contributions are as follows.

- We propose **C**l**A**ssifier-Guided **S**emantic **O**ptimization (CASO), a lightweight and efficient framework for image editing, which simply tunes trainable semantic embeddings and does not require training or fine-tuning of diffusion models.
- We theoretically and experimentally demonstrate that CASO learns semantic embeddings containing attribute class mean information, thereby enabling precise attribute editing.
- Experimental results show that our editing directions are highly disentangled and exhibit strong generalization, outperforming various existing approaches.

2 Related Work

Deep generative models have achieved great success in image editing. Various types of generative models, such as VAEs [Kingma, 2013], GANs [Goodfellow *et al.*, 2014], flow-based models [Dinh *et al.*, 2014], and diffusion models [Ho *et al.*, 2020; Song *et al.*, 2020], have been adopted as editing frameworks. Among these, the recent state-of-the-art editing models are based on diffusion models. In the following, we will focus on reviewing work on diffusion based image editing.

Diffusion models perform iterative denoising based on a sequence of latent variables, which lack explicit semantic representations like those in GANs. So identifying the disentangled semantic space in diffusion models remains a significant challenge. Some image editing methods based on diffusion models require retraining or fine-tuning a diffusion model [Kim *et al.*, 2022; Wang *et al.*, 2023; Preechakul *et al.*, 2022; Valevski *et al.*, 2022]. For example, DiffAE [Preechakul *et al.*, 2022] and HDAE [Lu *et al.*, 2024] use diffusion model as the representation encoders and enable image editing by modifying representations. These methods are generally only suitable for single-domain (such as human face) diffusion models. DiffusionCLIP [Kim *et al.*, 2022] fine-tunes the diffusion model with the help of the CLIP [Radford *et al.*, 2021]. UniTune [Valevski *et al.*, 2022] fine-tunes the diffusion model on a single base image, encouraging the model to generate images similar to the base image. However, training or fine-tuning the diffusion model can be somewhat expensive.

Recent advances [Yang *et al.*, 2023; Brack *et al.*, 2023; Tumanyan *et al.*, 2023; Wu and De la Torre, 2023; Hertz *et al.*, 2022a; Mokady *et al.*, 2022] avoid training or fine-tuning diffusion models. Some methods (SEGA [Brack *et al.*, 2023] and OIG [Lee *et al.*, 2024]) use text prompts to guide editing. OIG uses CLIP text encoder [Radford *et al.*, 2021] to ensure semantic correlation and extracts the feature map of UNet to ensure structural similarity. SEGA generates editing noise from text prompts and then extracts a small portion of this noise (by default, the top 5%, based on empirical observations) for editing. Unfortunately, the 5% threshold can only be set empirically, which is an unstable way to edit and can cause image quality loss. Additionally, some work optimize text-based embeddings to faithfully reconstruct the input image, so as to achieve disentangled edits (the editing does not change the irrelevant features) with classifier-free guidance. Null-Text Inverse [Mokady *et al.*, 2022] fine-tunes the null-text embedding to align the sampling and inversion trajectories, enabling accurate image reconstruction, but it relies on a user-provided source prompt. Prompt Tuning Inversion [Dong *et al.*, 2023] introduces a faster approach by encoding image information into a learnable text embedding and performing editing through linear interpolation with the target embedding. However, these methods prioritize reconstruction quality, whereas our goal is to obtain an optimal semantic embedding for effective image editing.

NoiseCLR [Dalva and Yanardag, 2024] and LOCO [Chen *et al.*, 2024] propose unsupervised methods to discover interpretable directions in pre-trained diffusion models, but they suffer from several shortcomings. The unsupervised learned

Figure 3: **Comparison of different methods for attribute "Mustache".** Our method shows the best generalization because it captures the exact semantics at the dataset level.

direction is uncontrollable and may not converge to user-desired direction. And no more than 10% of the directions learned by NoiseCLR are semantic meaningful. LOCO is only suitable for local editing, and some major structural changes such as cat→dog are not feasible.

3 Method

In this section, we describe our proposed method. First, we briefly discuss the background of latent diffusion models, and then introduce our method.

3.1 Latent Diffusion and Classifier-free Guidance

Diffusion models are generative models that produce data samples by fitting the data distribution through an iterative denoising process [Ho *et al.*, 2020]. Stable Diffusion (SD) [Rombach *et al.*, 2022] is a type of Latent Diffusion Model (LDM) that operates in the latent space of image data, where the conversion between latent code z and raw image data x is performed by the VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$. The training loss is defined as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right], \tag{1}$$

where t is uniformly sampled from $\{1, \ldots, T\}$ and z_t is a noisy version of $z_0 = \mathcal{E}(x)$.

Classifier-free guidance [Ho and Salimans, 2022] trains a conditional diffusion model. Since the condition is left blank with a certain probability during training, it also supports unconditional generation. The final predicted noise with classifier-free guidance in SD is defined as:

$$\tilde{\epsilon}_\theta(z_t, c) = \epsilon_\theta(z_t, \phi) + \lambda(\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t, \phi)), \tag{2}$$

where c is the condition, ϕ is null text and λ is guidance scale. For simplicity, we use $\epsilon_\theta(z_t)$ instead of $\epsilon_\theta(z_t, t)$ to represent the predicted noise for timestep t, as t is implicitly denoted with variable z_t.

3.2 Classifier-guided Semantic Optimization

We can set c as a suitable text to guide the editing. However, as shown in Appendix.B, even when sampling with the same random seed, images generated from different text prompts with the same semantic meaning exhibit significant differences (the first four images). Furthermore, adding additional descriptions to the text also leads to substantial changes in the overall appearance of the generated images (the first and last images). Text guidance struggles to achieve disentangled edits, namely, preserving the attribute excluding details. To achieve precise image editing, we propose a lightweight approach to obtain semantic embeddings as conditional guidance based on a ready-to-use attribute classifier.

Suppose the target editing attribute has K classes, i.e., $a \in \{1, \ldots, K\}$. Define the ready-to-use attribute classifier as $\mathcal{F}$ and the images generated by the Stable Diffusion editing process as $\hat{x} = \mathcal{G}(x, e_a)$, where x is the input image and e_a is the embedding for each class a of the attribute. We define an editing loss to optimize $\{e_a\}_{a=1}^K$ as follows:

$$\min_{\{e_a\}_{a=1}^K} \mathcal{L}_{\text{edit}} = \mathbb{E}_{x,a}\left[\ell_{\text{c}}\left(\mathcal{F}(\mathcal{G}(x, e_a)), a\right)\right] \tag{3}$$

where ℓ_{c} is the classification loss that enforces generated images conditioned on e_a to be classified as the corresponding attribute class a.

The attribute embedding $\{e_a\}_{a=1}^K$ obtained above contains the attribute class information at the dataset level, enabling

Method	Old	Mustache	Gender	Lipstick
Comp-Diff	0.19	0.40	0.42	0.38
DiffAE	0.20	0.19	0.24	0.23
Cycle-Diff	**0.10**	0.21	0.23	0.41
SEGA	0.11	0.23	0.27	0.42
NoiseCLR	0.17	0.17	**0.20**	0.19
T-LOCO	0.26	0.24	0.28	0.24
Ours	0.16	**0.13**	**0.20**	**0.18**

Table 1: Comparison of various methods with LPIPS(↓) for human face edit.

image editing with precise semantics. We provide a theoretical justification in the following.

Denoting all layers of the classifier except the last as $\mathcal{F}_{-1}$, and the last layer with weight w_a and bias b_a for each attribute class a, the classifier for predicting a class label can be expressed as $\arg\max_{a'} \left(\langle w_{a'}, \mathcal{F}_{-1}(x)\rangle + b_{a'}\right)$.

Definition 1 (Globally-centered attribute class mean μ_a)**.** *We define the globally-centered attribute class mean μ_a as:*

$$\mu_a = \mathbb{E}_i[h_{i,a}] - \mathbb{E}_{i,a}[h_{i,a}],$$

where $a = 1, 2, \ldots, K$ and $h_{i,a} = \mathcal{F}_{-1}(x_{i,a})$ is the last-layer feature of the i-th sample in attribute class a for training the classifier.

Theorem 1. *[Han* et al.*, 2021] For a sufficiently large classifier network, the last layer of the classifier w_a will converge to the globally-centered class mean μ_a, namely,*

$$\frac{w_a}{\|w_a\|_2} - \frac{\mu_a}{\|\mu_a\|_2} \to 0, \tag{4}$$

where $a = 1, 2, \ldots, K$. ■

Theorem 1 describes the neural collapse phenomenon of training a deep classifier network [Han *et al.*, 2021; Papyan *et al.*, 2020; Yang *et al.*, 2022]. We apply it to demonstrate that our classifier-guided semantic embedding encodes the attribute class mean.

Proposition 2. *Give a sufficiently large classifier network $\mathcal{F}$ which is well-trained on the training data. When optimizing $\{e_a\}_{a=1}^K$ using Eq.* (3) *until converge while fixing $\mathcal{F}$, the neural collapse phenomenon still happens, namely,*

$$\frac{w_a}{\|w_a\|_2} - \frac{\mu'_a(e_a)}{\|\mu'_a(e_a)\|_2} \to 0, \tag{5}$$

where $a = 1, 2, \ldots, K$. In particular,

$$\mu'_a(e_a) = \mathbb{E}_i\left[\mathcal{F}_{-1}(\mathcal{G}(x_i, e_a))\right] - \mathbb{E}_{i,a}\left[\mathcal{F}_{-1}(\mathcal{G}(x_i, e_a))\right],$$

which denotes the globally-centered attribute class mean of generated images. ■

The proof is similar to Theorem 1 given the same classifier. We give a detailed proof in the Appendix.

With Eq. (4) and Eq. (5), we can easily obtain $\frac{\mu'_a(e_a)}{\|\mu'_a(e_a)\|_2} \to \frac{\mu_a}{\|\mu_a\|_2}$. This means the optimal attribute embedding e_a in Eq. (3) is determined by the attribute class mean μ_a. We also verify this in our empirical study.

Algorithm 1 Training algorithm

Input: Training data $\mathcal{X} = \{x^i\}_{i=1}^N$, batch size B, denoising UNet ϵ_θ, VAE encoder $\mathcal{E}$, VAE decoder $\mathcal{D}$, classifier $\mathcal{F}$.
Output: Semantic embedding e_a

repeat
  Sample a batch of images $\{x^i\}_{i=1}^B$.
  Get latent codes $\{z_0^i = \mathcal{E}(x^i)\}_{i=1}^B$.
  Following Eq. (8), generate noise latent codes $\{z_L^i\}_{i=1}^B$.
  Following Eq. (10), generate edited images $\{\hat{x}^i\}_{i=1}^B$.
  Train the condition embedding e_a according to Eq. (13).
until converged

Image Editing by Diffusion Models

The typical image editing pipeline using LDMs incorporates conditioning during the denoising process, following these steps [Mokady *et al.*, 2022]:

(i) The input image x is encoded into its latent code $z_0 = \mathcal{E}(x)$ using the VAE encoder $\mathcal{E}$.

(ii) The clean latent code z_0 is converted to its noisy counterpart z_L through DDIM inversion (Eq. (6)), a deterministic mapping process.

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_\theta(z_t, \phi), \tag{6}$$

where α_t is the time dependent scale.

(iii) Classifier-free guidance (Eq. (2)) is applied during the denoising process using the condition e_a:

$$\begin{aligned}\hat{z}_{t-1} =& \sqrt{\alpha_{t-1}} \left(\frac{\hat{z}_t - \sqrt{1-\alpha_t}\tilde{\epsilon}_\theta(\hat{z}_t, e_a)}{\sqrt{\alpha_t}}\right) \\ &+ \sqrt{1-\alpha_{t-1}}\tilde{\epsilon}_\theta(\hat{z}_t, e_a),\end{aligned} \tag{7}$$

where $t = L, L-1, \ldots, 1$ and $\tilde{\epsilon}_\theta(z_t, e_a)$ is calculated according to Eq. (2). We initialize $\hat{z}_L = z_L$.

(iv) The edited image is obtained by decoding the denoised latent code $\mathcal{D}(\hat{z}_0)$.

For training, the edited image $\mathcal{D}(\hat{z}_0)$ is used as $\mathcal{G}(x, e_a)$ in Eq. (3) to update the embedding. However, steps (ii) and (iii) require multiple iterations, which impose significant time and memory overhead for the training phase. To improve training efficiency, we introduce the following approximations:

① For step (ii), we just add random noise to it directly via diffusion forward process by Eq. (8).

$$z_L = \mathcal{A}(z_0, L) = \sqrt{\alpha_L} z_0 + \sqrt{1-\alpha_L}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{8}$$

where L is a hyperparameter in this paper with $0 < L \leq T$ and $\mathcal{A}$ is the operator that transfers z_0 to its noisy version.

② For steps (iii) and (iv) to obtain a clean edited image for calculating the classification loss (Eq. (3)), we propose computing the loss on the decoded image from $\hat{z}_L$ rather than $\hat{z}_0$. To prevent the classifier from handling noisy data $p(a \mid \mathcal{D}(\hat{z}_L))$, we make the following approximation:

$$\begin{aligned}p(a \mid \mathcal{D}(\hat{z}_L)) &= \mathbb{E}_{\hat{z}_0 \sim p(\hat{z}_0 \mid \hat{z}_L)}\left[p(a \mid \mathcal{D}(\hat{z}_0))\right] \\ &\simeq p\left(a \mid \mathcal{D}\left(\mathbb{E}_{\hat{z}_0 \sim p(\hat{z}_0 \mid \hat{z}_L)}\left[\hat{z}_0\right]\right)\right),\end{aligned} \tag{9}$$

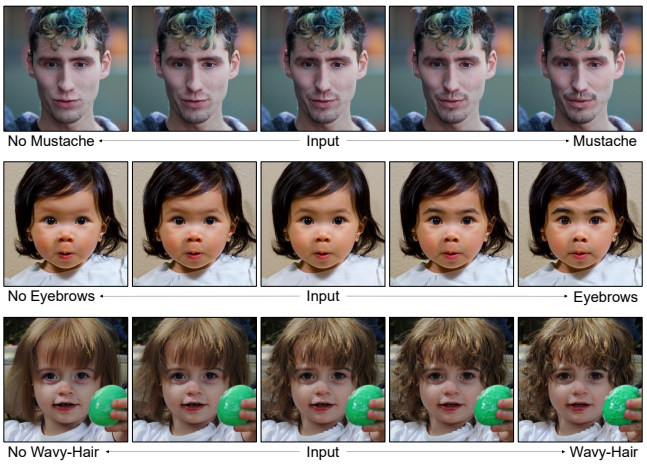


Figure 4: **CASO Interpolation Results.** Our method allow users to implement fine-grained editing and bidirectional editing by simply changing the classifier free guidance scale.

where $\mathbb{E}_{\hat{z}_0 \sim p(\hat{z}_0|\hat{z}_L)}[\hat{z}_0] = \frac{\hat{z}_L - \sqrt{1-\alpha_L}\tilde{\epsilon}_\theta(\hat{z}_L, e_a)}{\sqrt{\alpha_L}}$.
Now the decoded image $\hat{x}$ in Eq. (3) can be:

$$\hat{x} = \mathcal{G}(x, e_a) = \mathcal{D}\left(\frac{\hat{z}_L - \sqrt{1-\alpha_L}\tilde{\epsilon}_\theta(\hat{z}_L, e_a)}{\sqrt{\alpha_L}}\right). \quad (10)$$

Notably, Eq. (10) still allows us to incorporate the conditional guidance e_a.

Remark 1. *The approximation error of Eq.* (9) *can be quantified with the Jensen gap [Chung* et al.*, 2022]. Its upper bound is:*

$$\mathcal{J} \leq \frac{d}{\sqrt{2\pi\sigma^2}} e^{-1/2\sigma^2} \left\|\nabla_z \mathcal{F}(\mathcal{D}(z))\right\| \mathbf{Q}, \quad (11)$$

where d is environmental dimension (dimension of z), σ^2 is the ambient noise when the classifier predicts the label. And when $\sigma \to 0$, $\frac{d}{\sqrt{2\pi\sigma^2}} e^{-1/2\sigma^2} \to 0$. $\left\|\nabla_z \mathcal{F}(\mathcal{D}(z))\right\| := \max_{z'} \left\|\nabla_{z'} \mathcal{F}(\mathcal{D}(z'))\right\|$ and $\mathbf{Q} := \int \|z_0 - \hat{z}_0\| p(z_0|z_L)\, dz_0$. $\left\|\nabla_z \mathcal{F}(\mathcal{D}(z))\right\|$ is generally finite [Chung et al.*, 2022] and when we choose the value of L carefully, the upper bound on $\mathcal{J}$ can be sufficiently small. When L is too large, $\|z_0 - \hat{z}_0\|$ in $\mathbf{Q}$ becomes excessively large, reducing classifier accuracy. Conversely, when L is too small, z_L is close to clean data, making it difficult to edit the image at this time.*

Reconstruction loss To preserve the attribute excluding details, we add a latent reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{z,a}\left[\ell_{\text{r}}\left(\hat{z}_0, z_0\right)\right]. \quad (12)$$

where ℓ_{r} is the reconstruction loss. The final training objective is as follows:

$$\min_{\{e_a\}_{a=1}^K} \mathcal{L}_{\text{edit}} + \gamma \mathcal{L}_{\text{rec}}. \quad (13)$$

We show the impact of this part of the loss function on the results in Appendix. Our training algorithm is Algorithm.1.

Method	cat→dog	dog→cat	cat→tiger	tiger→cat
Comp-Diff	0.48	0.59	0.59	0.44
SEGA	0.57	0.50	0.56	0.55
Cycle-Diff	**0.40**	<u>0.48</u>	<u>0.47</u>	<u>0.40</u>
Ours	<u>0.43</u>	**0.42**	**0.45**	**0.34**

Table 2: Comparison of various methods with LPIPS(↓) for animal type transfer.

4 Experiment

To demonstrate the advancement and generalization of our method, we perform the following experiments.

4.1 Experimental Setup

We use Stable Diffusion-v1.5[1] for all experiments following [Dalva and Yanardag, 2024]. The datasets used include: FFHQ [Karras *et al.*, 2019], AFHQ [Choi *et al.*, 2020], CelebAHQ [Karras, 2017] and Stanford Cars datasets [Krause *et al.*, 2013]. For more details, please refer to Appendix.

Timesteps are also critical for editing quality and disentanglement [Wu *et al.*, 2023; Hertz *et al.*, 2022b]. During training process, L is set to $0.3T$ for human face and $0.4T$ for others. During editing, for subtle features like eyebrows, we start to apply our direction from $t \in [0.1T, 0.3T]$, while for some coarse-grained changes like species, editing at earlier timesteps is required ($t \in [0.8T, 0.9T]$). The results we show in the main text are all done with timesteps $T = 50$. Our classifiers are all obtained by simply fine-tuning a VGG16 model [Simonyan and Zisserman, 2014]. With a well-trained classifier, only 100-200 images are needed to train the embedding.

We compare our method with: DiffAE [Preechakul *et al.*, 2022], Cycle-Diff [Wu and De la Torre, 2023], SEGA [Brack *et al.*, 2023], Comp-Diff [Liu *et al.*, 2022], NoiseCLR [Dalva and Yanardag, 2024] and T-LOCO[2] [Chen *et al.*, 2024].

4.2 Qualitative and Quantitative Results

Our qualitative results are shown in Fig. 2. Although our approach is trained on real world images, it exhibits remarkable generalization on images of other styles, demonstrates that our method generalizes beyond previous works. Notably, our edits are not confined to the data seen by the classifier. For example, if we use a well-trained dog and cat classifier learn semantic embedding for "cat", it enables edits not only for cat and dog, but also all of others. The results demonstrate that our embeddings can capture precise semantics for the target attribute.

We compare the generalization performance with other methods for attribute "Mustache". This is shown in Fig. 3. Regardless of the type of image, our model consistently achieves superior performance.

[1]https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[2]LOCO is purely unsupervised editing method. We found it is hard to obtain specified attribute editing directions for comparison. So we adopt its text extension version (T-LOCO) for the experiments.

(a) Muti-attribute edit result. We compare the effect of the position order of the embedding stitching (the second row) and whether single-step editing is used (the last two images in each row are single-step edited)) on the editing effect. When target area "eyebrow" is occluded by hair, and our method can still complete the editing well.

(b) Multi-attribute edits in a cross-domain context. The last input image for "woman and cat" in the first column is generated by SD.

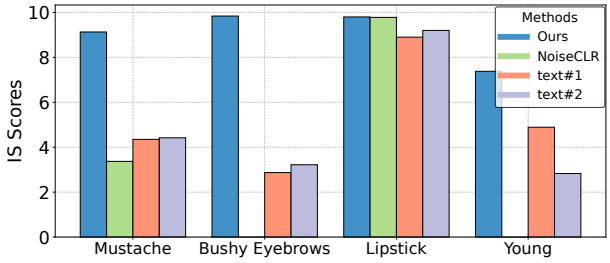
Figure 5: **CASO Muti-attribute Edit**. In complex and challenging scenarios, our method can still achieve perfect editing.

Table 1 and Table 2 shows quantitative results. We calculate the LPIPS metric (lower is better) [Zhang *et al.*, 2018] for editing different attributes on human face data and species conversion between animal data[3]. Our method achieve lower LPIPS than the other methods, indicating greater coherence while performing the edits.

Interpolating Edit Our approach can achieve fine-grained attribute editing and reverse editing by simply changing the classifier free guidance scale λ. For instance, using the "Bushy Eyebrows" embedding, users can create the effect of "Sparse Eyebrows" by setting the λ to a negative value.

[3]For pecies conversion between animal data, we do not compare our method with NoiseCLR, DiffAE and T-LOCO. This is because NoiseCLR and DiffAE do not support such editing. And T-LOCO performs poorly on this structure changes greatly task and the statistical results are not meaningful. We show examples in the Appendix.

(a) Generation with (1):ours, (2):NoiseCLR guidance.



(b) Image quality of generation under different guidance, measured by IS (↑). We do not count the IS score for some directions of NoiseCLR since it is not a desirable editing direction or the authors do not open source weights.

Figure 6: **Generation Result.** We use different direction: learned from NoiseCLR, learned from ours, and two kinds of text prompt (see Appendix) to guide the generation of images from Gaussian noise with corresponding attributes and calculate their IS scores.

Some examples are shown in Fig. 4. A similar effect can be achieved by changing the time at which the condition embedding is added, but this does not allow for reverse editing. We show it in the Appendix.

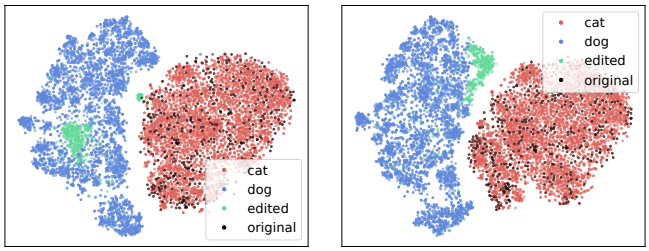
Multi-attribute Edit Our method allows for multi-attribute editing by simply concatenating multiple embeddings. Embedding available on different datasets can be combined (e.g. human face and animal). Experiments in Fig. 5a show that the position order of embedding stitching does not affect the result of image editing, which also proves the semantic decoupling of the learned orientation. We show editing results in difficult scenarios: In Fig. 5a, the target region to be edited has occlusion. In Fig. 5b middle, an artwork image, the face region to be edited is only a small part of the whole image. These are complex scenarios where our model still performs the editing task perfectly.

Single-Step Edit To further verify the effectiveness and consistency of our method, we conduct a single-step editing experiment, as shown in Fig. 5a. In this experiment, guidance information is applied in only one step of the denoising process, while all other steps before and after rely on unconditional embeddings. In contrast, most existing methods require conditional guidance across multiple denoising steps. Experimental results in Fig. 5a show that our outcomes are highly consistent, regardless of whether single-step editing is used. This demonstrates that our editing method is powerful enough that images generated by a single-step editing already exhibit the target properties.

DDIM Steps	AFHQ Cat		AFHQ Dog	
	$\lambda = 0$	$\lambda = 3$	$\lambda = 0$	$\lambda = 3$
$T = 10$	85.39	**39.00**$\downarrow$ 54.3%	88.52	**52.03**$\downarrow$ 41.2%
$T = 20$	70.34	**32.41**$\downarrow$ 53.9%	47.19	**45.13**$\downarrow$ 4.37%
$T = 50$	72.35	**43.20**$\downarrow$ 40.4%	**50.55**	50.89$\uparrow$ 0.67%

Table 3: Reconstruction quality on AFHQ w/o or w/ guidance, measured by FID ($\downarrow$). $\lambda = 0$ means reconstruction w/o guidance.

4.3 Accurate Semantics



(a) Ours (b) SEGA

Figure 7: T-SNE of cat→dog image edit. Red dot: real cat images; blue dots: real dog images; black dots: real cat images randomly sampled for editing; green dots: the generated images (to be edited as dog from the sampled cat images).

To demonstrate that our model can learn the most precise semantic direction, we directly sample from Gaussian noise with different guidance (ours, NoiseCLR and text). The results are shown in Fig. 6a. The generated images demonstrate that our learned embeddings more accurately capture semantic features. For instance, comparing the generated images guided by "Bushy Eyebrows" learned using our method and NoiseCLR separately, the faces generated by our method consistently exhibit the "Bushy Eyebrows" property, while the images generated by NoiseCLR do not have the desired semantics. We attribute the ability of NoiseCLR to perform attribute editing despite learning incorrect semantics to the coincidence of feature subspaces in the data space, which is also mentioned in SEGA [Brack *et al.*, 2023]. Additionally, we evaluate the IS (Inception Score) [Salimans *et al.*, 2016] of different guided sampling images (higher is better), which is shown in Fig. 6b. Some IS scores are not reported, because it is not an ideal edit direction (Fig. 6a) or the authors do not open source its weight [4]. Experiments show that our guidance can effectively improve the quality of image generation.

We conduct another experiment in which we randomly sample from the cat dataset, edit the images into dogs using different methods, including our approach and SEGA [Brack *et al.*, 2023] (the leading text-based image editing work at present). We visualize the features obtained by the classifier (i.e., $\mathcal{F}_{-1}(x)$) using T-SNE, as shown in Fig. 7. Notably, the features of images edited by our CASO locate at the dog class

[4]We tried to reproduce it. But because the direction of convergence obtained by such unsupervised methods is not controllable, we did not get the direction of interest based on their method.

Figure 8: **Improve image quality during editing and reconstruction.** Our guidance can not only faithfully reconstruct the structure of the original image, but also achieves the editing of the target attribute.

mean, consistent with our theoretical result. This demonstrate that our embedding indeed captures accurate attribute semantics at the dataset level, enabling superior editing performance.

4.4 Not Only Edit

Many editing works based on pre-trained diffusion models struggle to achieve satisfactory results due to the model's limited ability to accurately reconstruct certain images. In our experiments, we demonstrate that our method not only enables effective image editing but also enhances reconstruction quality, as shown in Fig. 8. We argue that the embeddings learned by our approach effectively capture both the overall structural information of the edited object and its target attributes.

When editing an image using attributes identical to those in the original (e.g., guiding the denoising of a cat image with the embedding of "cat"), the task effectively becomes reconstruction. This leads to higher-quality results and faster convergence. As shown in Fig. 8, our method enhances image quality in tasks such as car color editing and cat image reconstruction. This also explains the high quality of the generated images in 4.3. However, we acknowledge that an excessively large λ may introduce distortions in the reconstructed image. More examples can be found in the Appendix.

We calculated the FID metrics [Heusel *et al.*, 2017] for AFHQ Cat and Dog datasets under unconditional reconstruction and guided reconstruction with "cat" and "dog" embeddings, respectively. The results are presented in Table 3 and Fig. 8. Experiments show that our embeddings significantly enhance the quality of reconstruction for images within the same class.

5 Conclusion

We propose an image editing framework that optimizes semantic embeddings with a few image inputs, guided by a classifier, to deliver high-quality, disentangled edits and superior generalization compared to existing methods. However, since it builds upon the pre-trained Stable Diffusion model, its effectiveness is limited by Stable Diffusion. Additionally, the potential for malicious misuse, such as deepfake creation, is a concern [Korshunov and Marcel, 2018]. We suggest implementing strict access controls, maintaining usage logs, and promoting responsible use to ensure ethical deployment.

Acknowledgments

The work was supported by the National Key R&D Program of China (2023YFC3306100), National Natural Science Foundation of China (62172324, 62272379, 62403429), National Research Foundation, Key R&D in Shaanxi Province (2023-YBGY-269, 2022-QCY-LL33HZ), Xixian New Area Science and Technology Plan Project (RGZN-2023-002, 2022 ZDJS-001), Zhejiang Provincial Natural Science Foundation (LQN25F030008). This paper was also supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[Brack *et al.*, 2023] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

[Chen *et al.*, 2024] Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspaces in diffusion models for controllable image editing. *arXiv preprint arXiv:2409.02374*, 2024.

[Choi *et al.*, 2020] 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.

[Chung *et al.*, 2022] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[Dalva and Yanardag, 2024] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24209–24218, 2024.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Dinh *et al.*, 2014] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[Dong *et al.*, 2023] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7430–7440, 2023.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Han *et al.*, 2021] XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

[Hertz *et al.*, 2022a] 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.

[Hertz *et al.*, 2022b] 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.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[Karras, 2017] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[Kim *et al.*, 2022] 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.

[Kingma, 2013] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[Lee *et al.*, 2024] Hyunsoo Lee, Minsoo Kang, and Bohyung Han. Diffusion-based conditional image editing through optimized inference with guidance. *arXiv preprint arXiv:2412.15798*, 2024.

[Liu *et al.*, 2022] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[Lu *et al.*, 2024] Zeyu Lu, Chengyue Wu, Xinyuan Chen, Yaohui Wang, Lei Bai, Yu Qiao, and Xihui Liu. Hierarchical diffusion autoencoders and disentangled image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5374–5383, 2024.

[Luo *et al.*, 2023] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[Mokady *et al.*, 2022] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

[Papyan *et al.*, 2020] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[Preechakul *et al.*, 2022] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022.

[Radford *et al.*, 2021] 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.

[Rombach *et al.*, 2022] 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.

[Saharia *et al.*, 2022] 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.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[Tumanyan *et al.*, 2023] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[Valevski *et al.*, 2022] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2(3):5, 2022.

[Wang *et al.*, 2023] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023.

[Wu and De la Torre, 2023] 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.

[Wu *et al.*, 2023] 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.

[Yang *et al.*, 2022] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35:37991–38002, 2022.

[Yang *et al.*, 2023] Zhen Yang, Ganggui Ding, Wen Wang, Hao Chen, Bohan Zhuang, and Chunhua Shen. Object-aware inversion and reassembly for image editing. *arXiv preprint arXiv:2310.12149*, 2023.

[Zhang *et al.*, 2018] 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.