LEARNING INPUT-AGNOSTIC MANIPULATION DIRECTIONS IN STYLEGAN WITH TEXT GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

With the advantages of fast inference and human-friendly flexible manipulation, image-agnostic style manipulation via text guidance enables new applications that were not previously available. The state-of-the-art text-guided image-agnostic manipulation method embeds the representation of each channel of StyleGAN independently in the Contrastive Language-Image Pre-training (CLIP) space, and provides it in the form of a *Dictionary* to quickly find out the channel-wise manipulation direction during inference time. However, in this paper we argue that this dictionary which is constructed by controlling single channel individually is limited to accommodate the versatility of text guidance since the collective and interactive relation among multiple channels are not considered. Indeed, we show that it fails to discover a large portion of manipulation directions that can be found by existing methods, which manually manipulates latent space without texts. To alleviate this issue, we propose a novel method **Multi2One** that *learns* a *Dictionary*, whose entry corresponds to the representation of a single channel, by taking into account the manipulation effect coming from the interaction with multiple other channels. We demonstrate that our strategy resolves the inability of previous methods in finding diverse known directions from unsupervised methods and unknown directions from random text while maintaining the real-time inference speed and disentanglement ability.

1 INTRODUCTION

Wide range of generative models including adversarial networks (Goodfellow et al., 2014; Karras et al., 2018; 2019; 2020b; Kim et al., 2022; Kim & Ha, 2021; Karras et al., 2021), diffusion models (Dhariwal & Nichol, 2021), and auto-regressive models (Dosovitskiy et al., 2020; Chang et al., 2022) have been demonstrating notable ability to generate a high-resolution image that is hardly distinguishable from real images. Among these powerful models, style-based GAN models (Karras et al., 2019; 2020b) are equipped with a unique latent space which enables style and content mixing of given images, manipulation of local regions (Wu et al., 2021), and interpolation between different class of images (Sauer et al., 2022).

In this paper, we focus on the image manipulation based on the pre-trained StyleGAN, considering the unique advantages mentioned above and its popularity. Based on the steerability in the latent space of StyleGAN, researchers have put tremendous effort on finding a direction that causes semantically equivalent change to the entire samples of image. In this work, we refer to such latent direction as *global direction*. Unlike *local direction* which is a sample-wise traversal direction found by iterative optimization using a single image (Local Basis (Choi et al., 2021) and Latent Optimization of StyleCLIP (Patashnik et al., 2021)), global direction is applicable to any images once found using supervised (Jahanian et al., 2019), unsupervised (Shen & Zhou, 2021; Wang & Ponce, 2021; Härkönen et al., 2020; Voynov & Babenko, 2020), or text-guided methods (Global Mapper & `GlobalDirection`[1] of StyleCLIP (Patashnik et al., 2021)).

Among them, the text-guided methods have a unique advantage in that they can naturally provide the flexibility of manipulation through the diversity of the given driving text without human supervision

[1]In order to distinguish it from global direction, which means finding input agnostic directions, we express the method proposed in StyleCLIP in this way

to discover the direction in the latent space. However, in this paper, we argue that, contrary to this common belief on text-guidance, the standard method (Patashnik et al., 2021) for text-based StyleGAN manipulation surprisingly fails to even find the manipulation directions that are known to be found in unsupervised approaches. We hypothesize that the failure is due to the naïve approach that only considers a change of image caused by *a single channel* in StyleSpace, neglecting diverse directions that are visible only when manipulating multiple channels as a whole.

In order to address these issues, we propose a novel method **Multi2One**, which learns the representation of *one* channel considering the interactive effect of controlling *multiple* channels as a whole. More specifically, our method learns a novel channel-wise representation which are optimized to be able to reconstruct the direction with multiple channels when the CLIP representations of such directions are given as a guidance instead of the text. Contrary to the state-of-the-arts method (Patashnik et al., 2021) which explicitly encoded every single channel, our multi-channel based strategy not only excels in reconstruction of unsupervised directions but also in discovery of text-guided directions.

2 Related Work

Style-based Generators Generators of style-based models (Karras et al., 2019; 2020b;a; 2021) are built upon the progressive structure (Karras et al., 2018) that generates images of higher resolution in deeper blocks. The popularity of StyleGAN structure that has been employed in numerous number of researches comes from its ability to generate high-fidelity images, transfer styles to other images, and manipulate images in the latent spaces using inversion methods (Zhu et al., 2020; Roich et al., 2021; Tov et al., 2021; Collins et al., 2020). The latent spaces of StyleGAN are intermediate space $\mathcal{W}$ and StyleSpace $\mathcal{S}$ (Wu et al., 2021).

Unsupervised Global Directions Image-agnostic directions are latent vectors that create semantically equivalent shift when applied to the latent space of StyleGANs. In order to find such directions, SeFa (Shen & Zhou, 2021) performs PCA on the first weight that comes after intermediate space $\mathcal{W}$ in pre-trained StyleGAN, deducing the principal components as the global directions. On the other hand, GANspace (Härkönen et al., 2020) relies on the randomly sampled latent codes in $\mathcal{W}$ and the eigenvectors from the latent codes proved to be global directions that share an image-agnostic modification ability.

Text-guided Image Manipulations Most of the text-guided image manipulation methods aim to find a local direction (Kocasari et al., 2022; Xia et al., 2021; Patashnik et al., 2021) which is an image-specific direction that applies to a single sample of image. Methods that find local directions using text guidance could be found in various models including GAN, Diffusion (Kim & Ye, 2021; Nichol et al., 2021; Avrahami et al., 2022; Liu et al., 2021), and Vision transformers (Chang et al., 2022). Two unique approaches for finding a global direction using text guidance in StyleGAN are *Global Mapper* and `GlobalDirection` of StyleCLIP (Patashnik et al., 2021). Global Mapper finds an image-invariant direction for a single text by optimizing a fully connected layer. However, this method requires 10 hours of training time for every single text, making it less popular than `GlobalDirection`. On the other hand, `GlobalDirection` method offers a real-time manipulation in inference time using a dictionary-based framework that is applicable to any input image and text. Their methodology is to tweak a single channel of StyleSpace $\mathcal{S}$, then obtain the CLIP embedding variation by the change of semantics that happens in the manipulated image. The CLIP embeddings variations are pre-computed over all channels in StyleSpace, which forms a *dictionary*. Upon calculation of the cosine similarity between the CLIP-encoded target text and the dictionary, the set of most relevant latent channels with large similarities are manipulated by the cosine similarity score which is the global direction for the target text, while less relevant channels are unmodified leading to disentangled manipulation results.

3 Limited Coverage of StyleCLIP `GlobalDirection` Method

In this section, we briefly review the `GlobalDirection` of StyleCLIP (Patashnik et al., 2021), which performs StyleGAN-based image manipulation using text guidance (Sec. 3.1). Then, we provide our key motivation that this state-of-the-art text-guided manipulation method is surprisingly insufficient to fully utilize the manipulative power of StyleGAN (Sec. 3.2).

3.1 TEXT-GUIDED STYLEGAN IMAGE MANIPULATION

StyleSpace of StyleGAN The generators of StyleGAN family (Karras et al., 2019; 2020a;b) have a number of latent spaces: $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{W}+$, and $\mathcal{S}$. The original latent space is $\mathcal{Z}$, which is typically the standard normal distribution. The generator transforms an input noise $z \sim N(0, I)$ into intermediate latent spaces $\mathcal{W}$ (Karras et al., 2019), $\mathcal{W}+$ (Abdal et al., 2019), and $\mathcal{S}$ (Wu et al., 2021), sequentially. Recent study (Wu et al., 2021) shows that StyleSpace $\mathcal{S}$ is the most disentangled such that it can change a distinct visual attribute in a localized way. The number of style channels in $\mathcal{S}$ is 6048 excluding toRGB channels for StyleGAN-ADA (resolution 1024^2), and recent methods (Kocasari et al., 2022; Patashnik et al., 2021; Wu et al., 2021) modify the values of the channels to edit an image. Our method also adopts StyleSpace and we use k to denote the total number of channels in $\mathcal{S}$ (that is, $\boldsymbol{s} = [s_1, s_2, \cdots, s_k] \in \mathcal{S}$). For the convenience of explanation, the pre-trained generator G is re-defined with respect to StyleSpace as $X = G_s(\boldsymbol{s})$; X is a generated image. Note that $\boldsymbol{s}$ is the inverted style vector of image X, found via StyleGAN inversion methods such as (Alaluf et al., 2021; Roich et al., 2021; Tov et al., 2021) used for manipulation purpose in most cases.

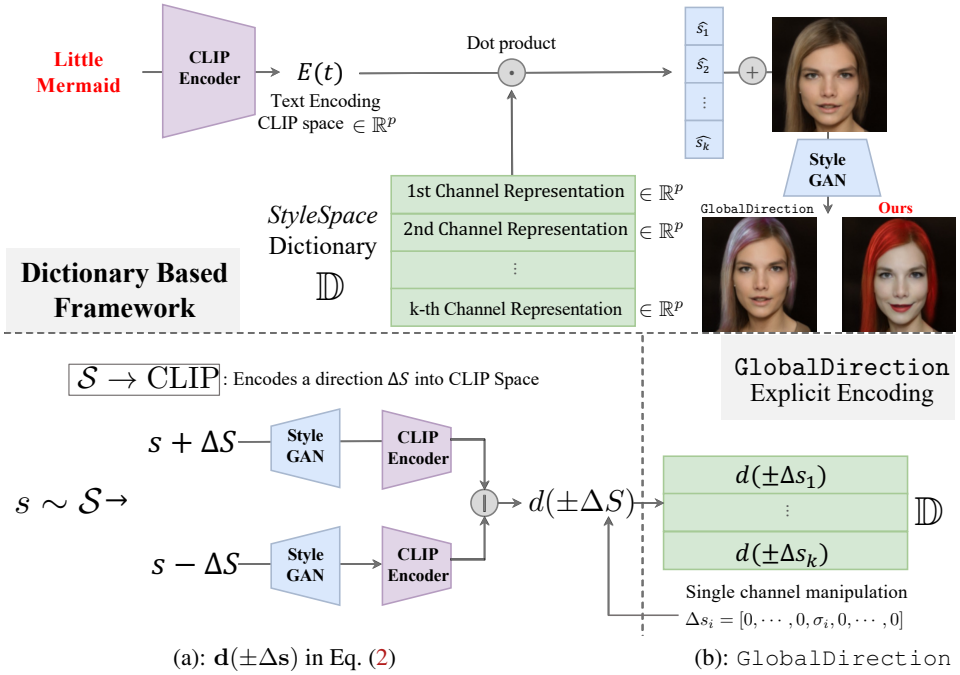


Figure 1: A diagram depicting the framework of dictionary-based image manipulation via text guidance

Text-guided Manipulation via StyleCLIP The goal of StyleGAN-based image manipulation is to find a direction $\hat{\mathbf{s}} = [\hat{s}_1, \hat{s}_2, ..., \cdots, \hat{s}_k]$ in StyleSpace $\mathcal{S}$ which generates an image $X_{\text{edited}} = G_s(\boldsymbol{s}+\hat{\mathbf{s}})$ suitable for the provided text guidance t. Note that the main benefit of considering $\mathcal{S}$ is that it does not change the undesired regions of the image when modifying a small number of style channels by its well disentangled property.

The `GlobalDirection` (Patashnik et al., 2021) is a representative method that provides an input-agnostic manipulation as well as some level of control on disentanglement by pre-computing set of CLIP embeddings corresponding to every style channel in StyleSpace $\mathcal{S}$. Specifically, StyleCLIP encodes the effect of manipulating i-th style channel using CLIP encoder $E(\cdot) \in \mathbb{R}^p$ (Radford et al., 2021) for all k channels to create the dictionary $\mathbb{D} \in \mathbb{R}^{p \times k}$ (Fig. 1(b)), then derive the direction $\hat{\mathbf{s}}$ by the similarity between CLIP encoded text $E(t)$ and $\mathbb{D}$ as follows:

$$\hat{\mathbf{s}} = E(t)^T \mathbb{D} \in \mathbb{R}^k. \tag{1}$$

StyleCLIP relies directly encodes the style channel individually to create the dictionary $\mathbb{D} := [\mathbf{d}(\pm\Delta \boldsymbol{s_1}), \mathbf{d}(\pm\Delta \boldsymbol{s_2}), \cdots, \mathbf{d}(\pm\Delta \boldsymbol{s_k})]$ which is described in the Fig. 1(a). More specifically, $\mathbf{d}$ is an

operation that encodes the change in image caused by the given direction $\Delta\mathbf{s} \in \mathbb{R}^k$ as follows:

$$\mathbf{d}(\pm\Delta\mathbf{s}) := \mathbb{E}_{\boldsymbol{s}\in\mathcal{S}}\Big[E\big(G_s(\boldsymbol{s}+\Delta\mathbf{s})\big) - E\big(G_s(\boldsymbol{s}-\Delta\mathbf{s})\big)\Big] \tag{2}$$

in which the expectation over style codes is approximately computed by training data in practice. Note that in `GlobalDirection`, the direction encoded by $\mathbf{d}$ is $\Delta\boldsymbol{s_i} = \sigma_i \boldsymbol{e}^{(i)}$, a standard basis vector whose i-th element σ_i is the standard deviation of the i-th channel, therefore $\mathbf{d}(\pm\Delta\boldsymbol{s_i})$ corresponds to the CLIP representation of the manipulation effect caused by i-th channel manipulation.

In the following section (Sec. 3.2), we propose the evidence to the hypothesis that the single-channel encoding strategy with $\Delta\boldsymbol{s_i}$ is the major bottleneck causing limited coverage issues that StyleCLIP suffers from.

3.2 Coverage Analysis of StyleCLIP `GlobalDirection` Method

Text-guided discovery of global direction is ideally preferred over unsupervised methods due to its ability to manipulate a given image with wide versatility (corresponding to various text inputs) which enables finding not only the directions found by unsupervised methods but also previously undiscovered directions as well. However, in this section, we argue that contrary to our expectation on the versatility of text guidance, the `GlobalDirection` of StyleCLIP surprisingly fails to discover a large portion of the instances from unsupervised directions. Moreover, it even fails to represent the new directions guided via text, to the fullest capacity of style-based GAN models (Karras et al., 2019; 2020a;b; Sauer et al., 2022).

To this end, we design two experiments that serves as an evidence to the limited coverage of StyleCLIP `GlobalDirection` (Patashnik et al., 2021) on unsupervised and text-guided directions respectively[2], using StyleGAN2-ADA (Karras et al., 2020a) trained on FFHQ dataset (Karras et al., 2019), image resolution of 1024.

Table 1: CLIP similarity score on manipulated images with unsupervised directions. r is number of non-zero entries in the unsupervised direction $\boldsymbol{\alpha}$.

r	Unsupervised	`Global Direction`
10	**0.3295**	0.2701
30	**0.3558**	0.3319
50	**0.3687**	0.2924

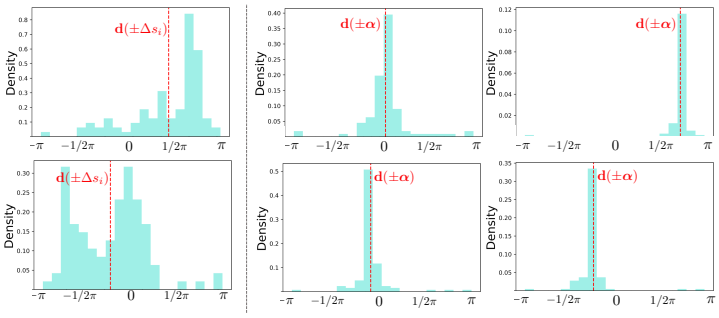
Figure 2: Manipulation results with randomly selected text. `GlobalDirection` is denoted as GD.

In Tab. 1, we measure the cosine similarity between $\mathbf{d}(\pm\boldsymbol{\alpha})$ and $E(X_{edited})$ where the manipulation of image is performed with either `GlobalDirection` or unsupervised approach (Shen & Zhou, 2021; Härkönen et al., 2020). Unlike the original framework where the guidance is given as a CLIP-encoded text t, we substitute $E(t)$ with $\mathbf{d}(\pm\boldsymbol{\alpha})$, which conceptually is a CLIP representation of image change induced by the direction $\boldsymbol{\alpha}$. Since $cos(\mathbf{d}(\pm\boldsymbol{\alpha}), E(X_{\text{edited}}))$ is constantly higher when using directions from unsupervised methods, it is evident that `GlobalDirection` is failing in recovery of unsupervised directions. Moreover, as shown in Fig. 2(b), despite the innate versatility of text, `GlobalDirection` shows inability in adapting to various text t. For additional results using randomly selected text prompts, please refer to Appendix A.

Regarding the phenomena found in the above two experiments, we hypothesize that the failure of `GlobalDirection` is due to the crude assumption on the StyleSpace in which the inconsistency of sample-wise manipulation result intensifies when modifying a single channel with $\Delta\boldsymbol{s_i}$. Such inconsistency harms the robust representation of i-th channel since $\mathbf{d}(\pm\Delta\boldsymbol{s_i})$ is the expectation of such inconsistent samples. Thereby we demonstrate that the basic assumption held by `GlobalDirection` which expects a single channel manipulation on any images to be semantically consistent in CLIP space, in fact, has numerous counter-examples showing inconsistency.

[2]We performed the experiments here using the `GlobalDirection` method and test examples provided on the StyleCLIP official site: https://github.com/orpatashnik/StyleCLIP

In Fig. 3(a), the distribution of $E\big(G_s(s+\Delta s_i)\big) - E\big(G_s(s-\Delta s_i)\big)$ over $s \sim \mathcal{S}$ is shown as an angle-histogram with red dashed line indicating the expectation $\mathbf{d}(\pm\Delta \boldsymbol{s_i})$ which is the i-th channel representation of `GlobalDirection`. As the samples show inconsistent angles over polar coordinate, the expectation is not a robust representative. Note that $\boldsymbol{\alpha}^{(r)}$ means a direction has r non-zero entries and $k-r$ zero entries.



(a) `GlobalDirection`: $\hat{s} = \Delta s$ (b) Unsupervised:$\hat{s} = \boldsymbol{\alpha}^{(50)}$ (c) Unsupervised:$\hat{s} = \boldsymbol{\alpha}^{(100)}$

Figure 3: Angle-histogram over directions $E\big(G_s(s+\hat{s})\big) - E\big(G_s(s-\hat{s})\big)$

On the other hand, it is observable that manipulation by unsupervised direction $\boldsymbol{\alpha}$, with Fig. 3(b)50 and Fig. 3(c)100 channels respectively, shows great improvement in consistency hence yields robust expectation, $\mathbf{d}(\pm\boldsymbol{\alpha})$. A more detailed explanation on representing the samples as directions for a angle histogram is deferred to Appendix B due to space constraints.

Moreover, we show an example where single channel alone have completely different meaning from the manipulation by the combination of multiple channels in Fig. 4. This supports our claim that the interaction between multiple channels should be considered, rather than individually encoding single channel each. Due to the naïve strategy of `GlobalDirection`, it fails to find the two channels (4520, 4426) for white hair when the text guidance is given as 'white hair'.

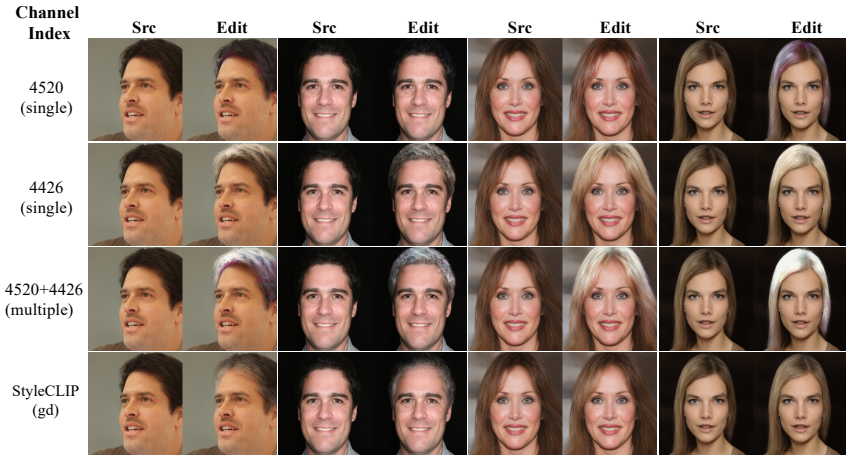


Figure 4: Manipulation result of controlling the channel index 4520, 4426, and both channels in latent space of pre-trained StyleGAN. `GlobalDirection` is manipulated by text 'white hair'.

4 Dictionary Learning for Sparse Representation of CLIP space

As discussed in Sec. 3.2, each style channel rather than the whole may not be meaningfully embedded in CLIP space. To address this issue, we propose a dictionary learning approach to find a robust representation of *styles*, in other words channels, in StyleSpace.

Ideally, we would like to construct a dictionary such that the manipulation direction $\hat{\boldsymbol{s}}$ obtained by $E(t)^T\mathbb{D}$ (Eq. (1)) with learned dictionary should correspond to the given driving text t. To see this, let us for now assume that the 'ground truth' manipulation direction $\boldsymbol{\alpha}_t$ corresponding to the driving text t is given. Then our estimated direction using learned dictionary as in Eq. (1), $\hat{\boldsymbol{\alpha}} = E(t)^T\mathbb{D}$, should be close to the ground truth $\boldsymbol{\alpha}_t$ which leads to the first term of the loss:

$$\underset{\mathbb{D}}{\text{minimize}} \left\|\boldsymbol{\alpha}_t - E(t)^T\mathbb{D}\right\|_2^2 \tag{3}$$

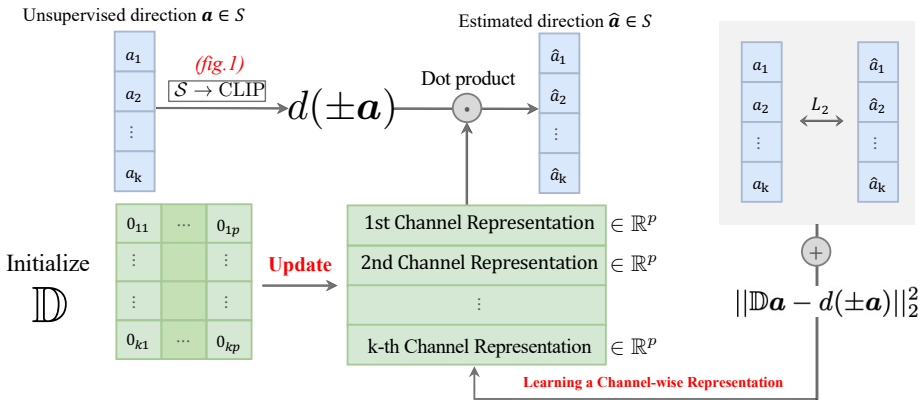


Figure 5: A concept diagram of our method in learning the *dictionary*.

which we name *ideal* dictionary learning problem since the ground truth direction $\boldsymbol{\alpha}_t$ for a driving text t is unknown. To avoid this, we substitute $\boldsymbol{\alpha}_t$ with the known directions $\boldsymbol{\alpha}$ derived from unsupervised methods (Härkönen et al., 2020; Shen & Zhou, 2021). Since $E(t)$ still does not exist in the second term, we use $\boldsymbol{\alpha}$ itself for driving manipulation in the second term rather than using t, to reconstruct $\boldsymbol{\alpha}$ given our dictionary (for ideal $\boldsymbol{\alpha}_t$, we have $E(t) = \mathbf{d}(\pm\boldsymbol{\alpha}_t)$ and hence $\|\boldsymbol{\alpha}_t - \mathbf{d}(\pm\boldsymbol{\alpha}_t)^T\mathbb{D}\|$).

The remaining question is how we can construct the set of known directions $\boldsymbol{\alpha}$ in Eq. (3). As mentioned above, we use all known directions found by GANSpace and SeFa (Härkönen et al., 2020; Shen & Zhou, 2021). However, it turns out that naïvely using them shows instability in terms of maintaining the original image, which is called entanglement issue. This is due to the innate properties of unsupervised methods that use principal component analysis or eigenvalue decomposition in the intermediate space of StyleGAN. The found style directions do not have zero entries, which indicates that all k channels are modified by the direction $\boldsymbol{\alpha}$ in $\mathcal{S}$. As can be seen in the experimental section, we empirically found that problems such as extreme entangling or image collapse occur when more than 300 coordinates of style direction $\boldsymbol{\alpha}$ are simultaneously activated.

To remedy this, we constraint $\boldsymbol{\alpha}$ to have impact on a limited number of *styles*. Specifically, we simply prune *styles* that have small magnitudes based on the ranking of coordinates, hence we have sparsified direction $\boldsymbol{\alpha} \otimes m$ where $m_j = 1$ if $\alpha_j \in \mathrm{Top-r}(\boldsymbol{\alpha})$ and 0 otherwise, $\mathrm{Top-r}(\boldsymbol{\alpha})$ is the set of top r coordinates of $\boldsymbol{\alpha}$ in magnitude and $\otimes$ indicates element-wise multiplication. This sparsified direction guarantees only TOP-r style channels to be modified while other channels relevant to unwanted local deformation stay intact.

Armed with the sparsified version of the directions $\boldsymbol{\alpha}^{(r)}$, where r is the number of non-zero entries, we introduce additional loss term $\|\mathbf{d}(\pm\boldsymbol{\alpha}_t) - \mathbb{D}\boldsymbol{\alpha}_t\|_2^2$ that encourages disentanglement in image manipulation to equation 3. Since we use 10, 30, ad 50 non-zero entries for each direction, each direction $\boldsymbol{\alpha}$ has three versions of itself, $\boldsymbol{\alpha}^{10}$, $\boldsymbol{\alpha}^{30}$ and $\boldsymbol{\alpha}^{50}$. Since the combination of overlapping and non-overlapping channels are simultaneously introduces in the dictionary learning process as a linear combination, the representation for each channel could be refined to have less entanglement. In Appendix C, we informally explain about the details of how our objective using sparsified $\boldsymbol{\alpha}$ set encourages to learn disentangled bases of $\mathbb{D}$. Hence our method may have greatly disentangled and diverse directions compared to the unsupervised directions (even though our $\boldsymbol{\alpha}$ comes from the unsupervised methods).

Combining these two ingredients together yields our dictionary learning problem by considering all known directions $\boldsymbol{\alpha}$:

$$\underset{\mathbb{D}}{\text{minimize}} \sum_{\boldsymbol{\alpha}} \left\|\boldsymbol{\alpha} - \mathbf{d}(\pm\boldsymbol{\alpha})^T\mathbb{D}\right\|_2^2 + \lambda \cdot \left\|\mathbb{D}\boldsymbol{\alpha} - \mathbf{d}(\pm\boldsymbol{\alpha})\right\|_2^2 . \tag{4}$$

where λ is the tunable hyper-parameter set as 0.01 and $\boldsymbol{\alpha}_t$ for *ideal* objective is substituted by unsupervised directions $\boldsymbol{\alpha}$.

5 Experiments

We evaluate the proposed method on StyleGAN2 (Karras et al., 2020a). We used StyleGAN2-ADA pretrained on multiple datasets, including FFHQ (Karras et al., 2019), LSUN (Yu et al., 2015) Car,

	Global Direction	Ours
$\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}})$	0.0315 $\pm$ 0.0037	**0.0057** $\pm$ 0.0034
$\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}} \cdot \mathbb{I}_{\{\mathbb{R}_{\neq 0}\}}(\boldsymbol{\alpha})\big)$	0.0267 $\pm$ 0.00114	**0.0005** $\pm$ 0.00001
$\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}} \cdot \mathbb{I}_{\{0\}}(\boldsymbol{\alpha})\big)$	0.006218 $\pm$ 0.00658	**0.00486** $\pm$ 0.00363

Table 2: MSE measured in StyleSpace.

		Source Image (Real)		Source Image (Generated)	
		Global Direction	Ours	Global Direction	Ours
$Top - r$	**10**	0.224	**0.242**	0.221	**0.234**
	30	0.228	**0.249**	0.226	**0.241**
	50	0.232	**0.251**	0.229	**0.243**
	100	0.237	**0.254**	0.235	**0.246**

Table 3: Cosine similarity R_{CLIP} measured in CLIP Space.

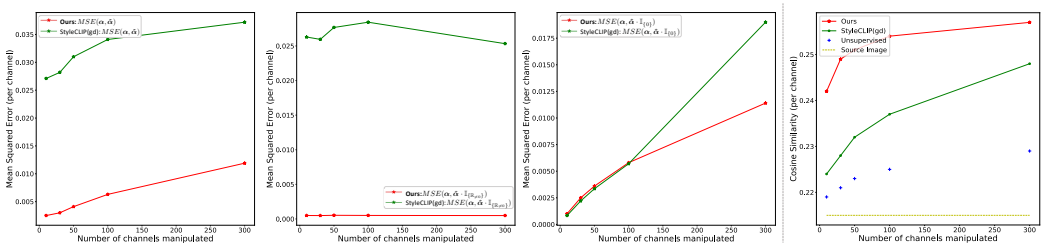
Church, AFHQ (Choi et al., 2020) Dog, Cat. In this section, we show the widened StyleGAN manipulation coverage of our method compared to a text-guided method, `GlobalDirection` of StyleCLIP (Patashnik et al., 2021) mainly on FFHQ pre-trained model. For additional details in experimental design, please refer to Appendix E.

Quantitative Results We measure the superiority of our method in terms of recovering the unsupervised directions $\boldsymbol{\alpha}$ and text guidance t. The metrics are defined in StyleSpace $\mathcal{S}$ to evaluate over unsupervised directions and CLIP space for evaluation over text guidance. In order to assess is **Multi2One** recover the unsupervised direction with maximum manipulative ability and minimum entanglement, we define three metrics as follows.

- **StyleSpace metric for *unsupervised directions*:** As a measure for discrepancy existent between `GlobalDirection` and unsupervised directions, we employ the following metric:
$$\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) := \mathbb{E}_{\boldsymbol{\alpha} \sim \mathbb{A}}\big[\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2\big] \tag{5}$$
where $\boldsymbol{\alpha}$ is from a set of unsupervised directions $\mathbb{A}$ and the reconstruction of such direction is given by $\hat{\boldsymbol{\alpha}} = \mathbf{d}(\pm\boldsymbol{\alpha})^T\mathbb{D}$ in which the text guidance $E(t)$ is replaced with $\mathbf{d}(\pm\boldsymbol{\alpha})$, following the standard framework described in Sec. 3.1. Conceptually, $\mathbf{d}(\pm\boldsymbol{\alpha})$ is a CLIP representation of change in images caused by manipulation with the direction $\boldsymbol{\alpha}$, computed in the same exact manner of Eq. (2). Tab. 2 shows that **Multi2One** consistently outperforms `GlobalDirection` in terms of cosine similarity between the unsupervised direction and the estimated direction. $\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}} \cdot \mathbb{I}_{\{\mathbb{R}_{\neq 0}\}}(\boldsymbol{\alpha})\big)$ indicate the non-zero entries of ground truth direction $\boldsymbol{\alpha}$ which could be used for measuring the pure manipulation ability by computing the error in channels that should be modified. On the other hand, $\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}} \cdot \mathbb{I}_{\{0\}}(\boldsymbol{\alpha})\big)$ is a metric used for calculating the entanglement that happened in channels that should not be modified.
- **CLIP space metric for *unknown directions*:** We measure the similarity between edited image and the text guidance in CLIP space as follows:
$$\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}}) := \mathbb{E}_{t \sim \mathbb{T}}\big[cos\big(E(t), E(X_{\text{edited}})\big)\big]. \tag{6}$$
We denote $\mathbb{T}$, the set of text prompts that describes the pre-trained domain of StyleGAN, namely FFHQ. Tab. 3 shows that the cosine similarity between given text and the edited image is consistently higher using our method, **Multi2One**. $Top - r$ indicates the number of channels that are manipulated in StyleSpace.

Qualitative Results We demonstrate the effectiveness of our method, compared to the state-of-the-arts method, `GlobalDirection`. We conducted all experiments using the pretrained models and the precomputed set of CLIP embeddings, $\mathbf{d}(\Delta \boldsymbol{s}_i)$, that are readily provided for FFHQ, AFHQ Cat and AFHQ Dog. We computed the CLIP embeddings of the StyleSpace of LSUN Church and Car using the same configuration with `GlobalDirection`. All the results are compared under same condition, where manipulation strength and disentanglement levels are set to be equal. More specifically, we modify same number of channels for both methods and the magnitude of change[3] in StyleSpace is always fixed as 10.

[3]The original paper StyleCLIP refers to the magnitude of change in StyleSpace as α and the disentanglement level as β which we substitute by number of channel that are changed.



(a) $\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}})$ (b) $\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}} \cdot \mathbb{I}_{\{\mathbb{R}_{\neq 0}\}}(\boldsymbol{\alpha})\big)$ (c) $\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}} \cdot \mathbb{I}_{\{0\}}(\boldsymbol{\alpha})\big)$ (d) $\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$

Figure 6: Evaluation on the ability to recover unsupervised directions. (a) Error between ground truth $\boldsymbol{\alpha}$ and the estimate $\hat{\boldsymbol{\alpha}}$ (b) Manipulation ability: Error between the non-zero entries(c) Disentanglement: Error between the zero entries (d) CLIP similarity between text and the edited image by **Multi2One**, `GlobalDirection`, and unsupervised method.

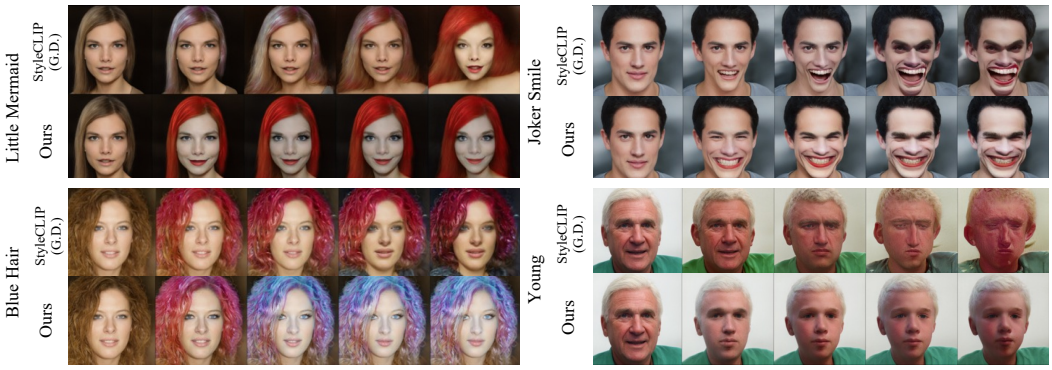
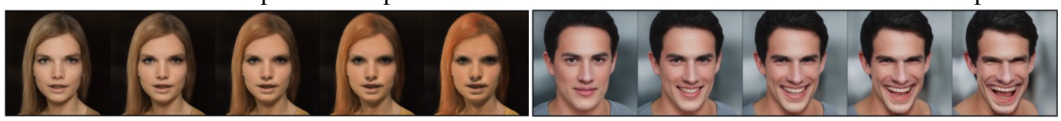


Figure 7: Manipulation results using FFHQ-pretrained StyleGAN2-ADA via text guidance on our method and StyleCLIP `GlobalDirection`. The samples are manipulated with increasing number of channels that are modified leading to larger changes.

The manipulated images show that `GlobalDirection` fails to represent some of the most basic text such as 'Young' for FFHQ. On the other hand, ours successfully manipulates the images not only on the basic text guidances but also on advanced semantics such as 'Little mermaid' and 'Joker smile' for FFHQ. We emphasize that despite the dictionary of our method is learned from the known directions in unsupervised approaches (Shen & Zhou, 2021; Härkönen et al., 2020), the manipulation results show that our learned dictionary could adapt to previously unseen combination of semantics such as red hair, pale skin, and big eyes to represent 'Little Mermaid' and unnatural smiles with red lipstick and pale face to represent 'Joker smile'.

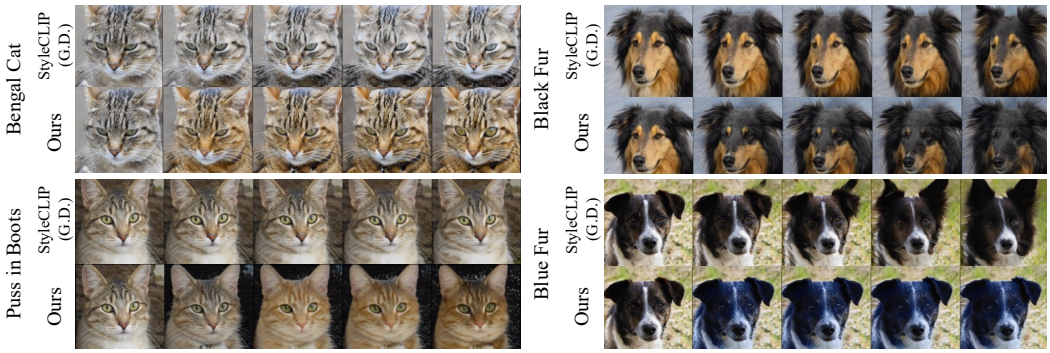
We show in Fig. 8 that the unsupervised direction $\boldsymbol{\alpha}$ whose CLIP representation $\mathbf{d}(\pm\boldsymbol{\alpha})$ has the maximum cosine similarity with the CLIP encoding of text 'Little Mermaid' and 'Joker smile' have slightly different direction compared to the manipulation result in Sec. 5. For $t =$'Little mermaid', among the 1440 directions that we used for dictionary learning, the direction that scored the highest cosine similarity of 0.1094 with $E(t)$ was the 1475th direction. As shown in Fig. 8, the direction has little relevance with the given text. Also, given a text 'Joker smile', the direction that scored highest cosine similarity of 0.1901 was 316th whose application to StyleSpace shows exaggerated smiles but not with red lipsticks and pale skin. We show additional results of our method compared to



Little mermaid (similarity 0.1094) Joker smile (similarity 0.1901)

Figure 8: **Manipulation by the direction with highest cosine similarity**. We find the unsupervised direction with highest similarity given text 'Little mermaid' and 'Joker smile'. Then we manipulate the image using the found direction. This shows that our method successfully discovers directions that has not been encountered during the course of dictionary learning.

`GlobalDirection` based on AFHQ(Sec. 5) and LSUN(Sec. 5) dataset on which StyleGAN-ADA is pretrained. Similar to FFHQ, we observed that styleCLIP fails to manipulated given images with simple text like 'Black fur' for AFHQ-dog and 'Brick church' for LSUN-church. We demonstrate the superiority of ours which discovers 'Blue fur' from AFHQ-dog pretrained StyleGAN and 'Golden church' from LSUN-church pretrained StyleGAN.



(a) (b)

Figure 9: Manipulation results using AFHQ cat and dog pretrained StyleGAN2-ADA via text guidance on our method and StyleCLIP `GlobalDirection`. The samples are manipulated with increasing number of channels that are modified leading to larger changes.

(a) (b)

Figure 10: Manipulation results using LSUN car and church pretrained StyleGAN2-ADA via text guidance on our method and StyleCLIP `GlobalDirection`. The samples are manipulated with increasing number of channels that are modified leading to larger changes.

REPRODUCIBILITY STATEMENT

We will release our code and the implementation details to be publicly available for reproducibility.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019.

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

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. *arXiv preprint arXiv:2202.04200*, 2022.

Jaewoong Choi, Junho Lee, Changyeon Yoon, Jung Ho Park, Geonho Hwang, and Myungjoo Kang. Do not escape from the manifold: Discovering the local coordinates on the latent space of gans. *arXiv preprint arXiv:2106.06959*, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *CVPR*, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020.

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

Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 159–168, 2015.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020b.

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

Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*, 2021.

Junho Kim, Yunjey Choi, and Youngjung Uh. Feature statistics mixing regularization for generative adversarial networks. In *CVPR*, 2022.

Yunji Kim and Jung-Woo Ha. Contrastive fine-grained class clustering via generative adversarial networks. 2021.

Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: Multi-channel based fast text-guided image generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 895–904, 2022.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *ArXiv*, abs/2112.05744, 2021.

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

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 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. *arXiv preprint arXiv:2103.00020*, 2021.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv preprint arXiv:2202.00273*, 2022.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, 2021.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020.

A MORE EXAMPLES ON FIG. 2

In this section, we show the extensive results on the limited coverage of `GlobalDirection` on wide variety of text guidance. The experiment is conducted on FFHQ-pretrained StyleGAN-ADA, and the text guidance are the adjective-noun pairs extracted from Visual Semantic Ontology (Jou et al., 2015) from which extract the nouns that describe human face. For example, the adjective-noun pairs contain human face descriptions such as 'Handsome smile', 'Bad hair', and 'Stupid face'. We manipulate the original image on the leftmost side using `GlobalDirection` and **Ours** using the same set of text description and equal configuration for both methods (manipulated channels 5, manipulation strength 5). The result in Fig. 11 shows that StyleCLIP `GlobalDirection` constantly fails to represent the given text guidance faithfully, manifesting its limited capacity in exploiting the flexibility of text hence the need for **Ours** to overcome the coverage issue.

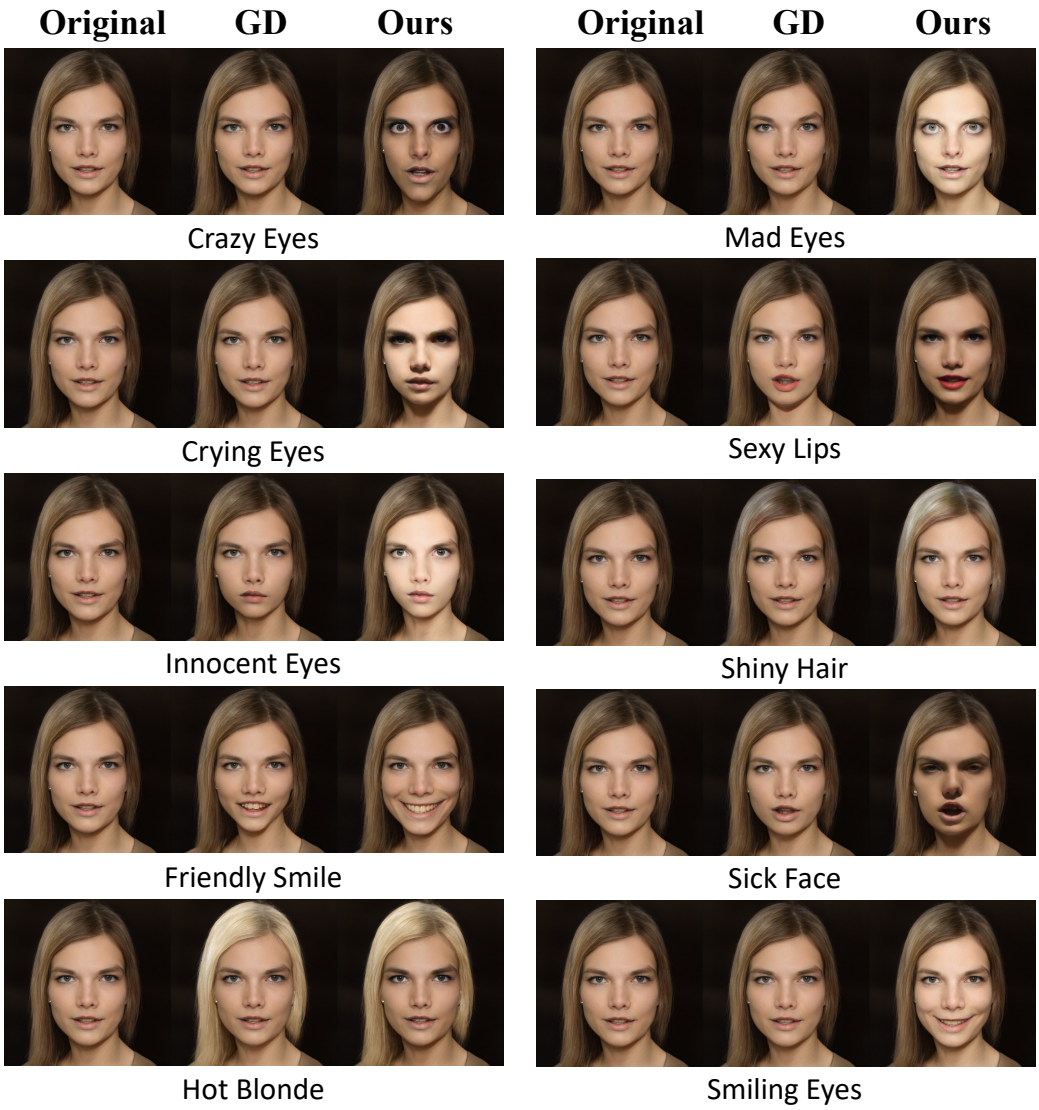


Figure 11: Manipulated images using randomly selected text from Visual Semantic Ontology (VSO). We show manipulated image via `GlobalDirection` (denoted as GD) and **Ours**.

B MORE DETAILS ON FIG. 3

Both `GlobalDirection` and **Ours** rely on multiple samples of images to encode the averaged difference caused by manipulating a sample into positive and negative direction. The difference is that unlike `GlobalDirection` which relies on a single channel manipulation, **Ours** encode change of image caused by image-agnostic direction found by unsupervised methods (Härkönen et al., 2020; Shen & Zhou, 2021). Single channel manipulation is not image-agnostic since the change caused by moving a single channel may not be applicable to certain source images used as a sample. For unconditional FFHQ pretrained StyleGAN-ADA, we have observed that a style channel which is known to make a person's hair grey fails to modify half of the samples. On the other hand, the image-agnostic directions found by unsupervised methods hardly fails to manipulate any given image to become older with grey hair. As a consequence, the CLIP embeddings of difference caused by single channel manipulation may greatly vary sample by sample while image-agnostic directions show consistent direction of CLIP embeddings due to the successful manipulation on any given sample of images.

We visualize the direction of each sample when a single channel is modified, and when 100 or 200 channels from unsupervised directions are modified. More specifically, for a single channel manipulation of `GlobalDirection` the samples are $E\big(G_s(s + \Delta \boldsymbol{s_i})\big) - E\big(G_s(s - \Delta \boldsymbol{s_i})\big)$ and for multiple channel manipulation, the samples are $E\big(G_s(s + \boldsymbol{\alpha})\big) - E\big(G_s(s - \boldsymbol{\alpha})\big)$ where $s \sim \mathcal{S}$.

The visualization is conducted in a similar manner with the work of Wang & Isola (2020). First we embed the $p = 512$ dimensional CLIP vectors into 2-dimension using t-SNE (Van der Maaten & Hinton, 2008). We denote each sample as (a_i, b_i) for $i = 1, 2, \cdots, 100$ and the averaged direction of interest $\mathbb{E}\Big[E\big(G_s(\boldsymbol{s} + \Delta \boldsymbol{s_i})\big) - E\big(G_s(\boldsymbol{s} - \Delta \boldsymbol{s_i})\big)\Big]$ in 2-dimensional space is (a, b). The angles are derived as $\theta_i = \arctan2(b_i, a_i)$ for all samples which range between $[-\pi, +\pi]$. The histogram of 100 angles are plotted with the angle of averaged difference in $\mathbb{R}^2$ with the angle $\arctan2(b, a)$ is emphasized as a red line in Fig. 3.

C Discussion on How Our Method Achieve Disentangled Bases

Here we informally explain how our objective function

$$\underset{\mathbb{D}}{\text{minimize}} \sum_{\boldsymbol{\alpha}} \left\|\boldsymbol{\alpha} - \mathbf{d}(\pm\boldsymbol{\alpha})^T\mathbb{D}\right\|_2^2 + \lambda \cdot \left\|\mathbb{D}\boldsymbol{\alpha} - \mathbf{d}(\pm\boldsymbol{\alpha})\right\|_2^2 . \tag{7}$$

finds a disentangled estimate $\hat{\boldsymbol{\alpha}} = E(t)^T\mathbb{D}$ of a ground truth direction $\boldsymbol{\alpha}$. For clarification, the disentangled estimate means that the element α_i for irrelevant coordinate s_i in StyleSpace $\mathcal{S}$ has a small value for the estimate $\hat{\alpha}_i$ which is a cosine similarity between CLIP encoded text and i-th basis of dictionary $\mathbb{D}$. Therefore, for all irrelevant coordinates in StyleSpace which might cause entanglement, disentangled estimate should satisfy $E(t)^T d_i \approx 0$

Suppose that the second term of the objective function $\left\|\mathbb{D}\boldsymbol{\alpha} - \mathbf{d}(\pm\boldsymbol{\alpha})\right\|_2^2$ is perfectly optimized that the dictionary $\mathbb{D}$ could represent any given $\mathbf{d}(\pm\boldsymbol{\alpha})$ as the linear combination of bases in dictionary as $\mathbf{d}(\pm\boldsymbol{\alpha}) = \mathbb{D}\boldsymbol{\alpha} = \alpha_1 \cdot d_1 + \alpha_2 \cdot d_2 + \cdots + \alpha_k \cdot d_k$. Since the given $\boldsymbol{\alpha}$ is pruned to have only a limited number (300 elements) of non-zero elements among k coordinates in StyleSpace, the other $k - 300$ elements for inactivated coordinates in StyleSpace $\mathcal{S}$ are zeros creating a sparse direction $\boldsymbol{\alpha}$ (FFHQ pretrained StyleGAN2-ADA has $k = 6048$ style coordinates). Based on the sparse direction, we explain how the estimate $\hat{\boldsymbol{\alpha}} = E(t)^T\mathbb{D}$ enforces to not modify the irrelevant *styles*, where the cosine similarity becomes nearly zero for any irrelevant elements $\hat{\alpha} \in \hat{\boldsymbol{\alpha}}$.

Here we describe a carefully designed environment for intuitive explanation. For simplification of the problem statement, we reduce the size of StyleSpace into $k = 4$. Moreover, there are four known directions from unsupervised approach with only two non-zero elements each. The set of directions is denoted as $\mathcal{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4] \in \mathbb{R}^{k\times 4}$ under the assumption that $\mathbf{d}(\pm\boldsymbol{\alpha}) = \mathbb{D}\mathcal{A}$ since the second term is perfectly optimized which transforms $\mathbf{d}(\pm\boldsymbol{\alpha})$ as the following matrix

$$\mathbf{d}(\pm\boldsymbol{\alpha}) = \begin{bmatrix} d_1 & d_2 & d_3 & d_4 \end{bmatrix} \begin{bmatrix} 0.5 & 0.2 & 0 & 0 \\ -0.5 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & -0.7 \\ 0 & 0.4 & -0.3 & 0.7 \end{bmatrix} = \begin{bmatrix} 0.5 \cdot d_1 - 0.5 \cdot d_2 \\ 0.2 \cdot d_1 + 0.4 \cdot d_4 \\ 0.3 \cdot d_2 - 0.3 \cdot d_4 \\ -0.7 \cdot d_3 + 0.7 \cdot d_4 \end{bmatrix} .$$

Then, proceeding to the first term of the objective function, $\arg\min_D \left\|\boldsymbol{\alpha} - \mathbf{d}(\pm\boldsymbol{\alpha})^T\mathbb{D}\right\|_2^2$, we may substitute $\mathbf{d}(\pm\boldsymbol{\alpha})$ as a linear combination of basis in dictionary $\mathbb{D} = [d_1, d_2, d_3, d_4]$ and focus on the first element α_1 of direction $\boldsymbol{\alpha}$ and its estimate $\hat{\alpha}_1 \in \hat{\boldsymbol{\alpha}}$.

Leaving only the first basis d_1 in $\mathbb{D}$, the optimization objective of the first term is to find d_1 such that satisfies $\alpha_1 = \mathbf{d}(\pm\boldsymbol{\alpha})d_1$ for all directions $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3$, and $\boldsymbol{\alpha}_4$ as the following equation

$$\begin{bmatrix} 0.5 \cdot d_1 - 0.5 \cdot d_2 \\ 0.2 \cdot d_1 + 0.4 \cdot d_4 \\ 0.3 \cdot d_2 - 0.3 \cdot d_4 \\ -0.7 \cdot d_3 + 0.7 \cdot d_4 \end{bmatrix} \cdot d_1 = \begin{bmatrix} 0.5 \\ 0.2 \\ 0 \\ 0 \end{bmatrix} .$$

We may further simplify the equation above into

$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} \cdot d_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

which constrains d_1 to be orthogonal to d_2, d_3, d_4. Likewise, for all bases in $\mathbb{D}$, there exist a set of other bases that are orthogonal.

By the second term of objective function, CLIP embedding of text $E(t)$ may be expressed as a linear combination of bases in $\mathbb{D}$ as $\mathbb{D}\boldsymbol{\alpha}_t$. Though the true direction $\boldsymbol{\alpha}_t$ given a text t is unknown, we assume that the equivalence of $\mathbf{d}(\pm\boldsymbol{\alpha})$ and $E(t)$ in CLIP space naturally satisfies follow: $\arg\min_{\mathbb{D}} \left\|\mathbb{D}\boldsymbol{\alpha} - \mathbf{d}(\pm\boldsymbol{\alpha})\right\|_2^2 \approx \arg\min_{\mathbb{D}} \left\|\mathbb{D}\boldsymbol{\alpha} - E(t)\right\|_2^2$. Based on this assumption, now suppose that the text encoding $E(t)$ is expressed as $0.1 \cdot d_2 + 0.4 \cdot d_3$. Then the estimate for the first coordinate of StyleSpace will be $\alpha_1 = (0.1 \cdot d_2 + 0.4 \cdot d_3)^T d_1 = 0$ and the fourth coordinate $\alpha_4 = (0.1 \cdot d_2 + 0.4 \cdot d_3)^T d_4 = 0$ since d_1 is orthogonal to d_2, d_3, d_4 and d_4 is orthogonal to d_1, d_2, d_3. Thus we may derive an estimate $\hat{\boldsymbol{\alpha}} = E(t)^T\mathbb{D}$ that has zeros for the StyleSpace coordinates s_1, s_4 that are irrelevant to the given text.

D MORE EXPERIMENTAL EXAMPLES FOR ALL DATASET

In this section, we provide more experimental results of **Ours** with FFHQ, AFHQ, and LSUN.

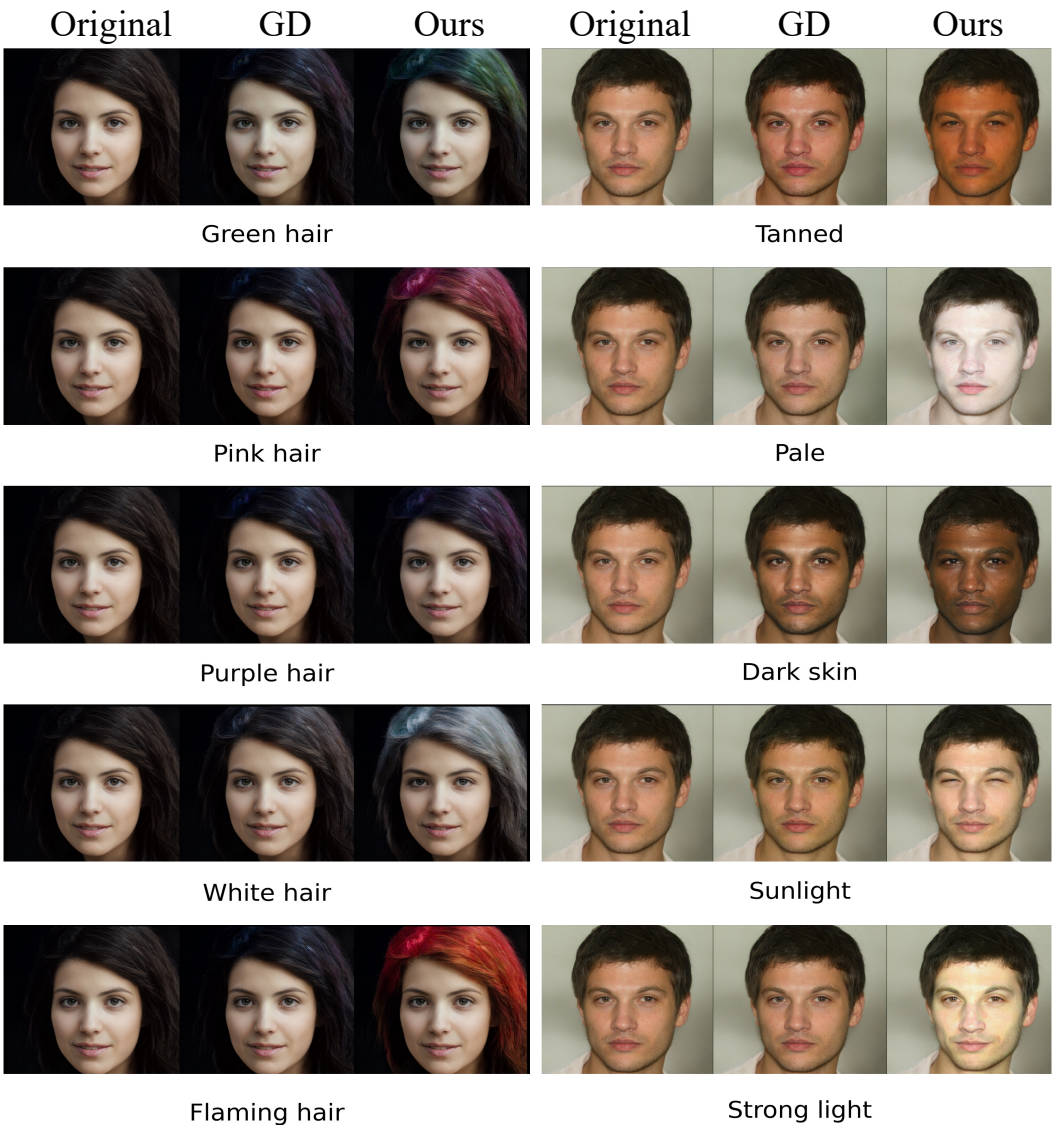


Figure 12: Additional manipulation results on FFHQ-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Ours**.

Fig. 13 shows the manipulation results on **Ours** and StyleCLIP `GlobalDirection` using 100, 500, 1000 channels. In addition, the manipulation results without channel pruning strategy is shown as well. Using over 500 channels is required for large structural change of the object in source image. However, StyleCLIP `GlobalDirection` shows signs of severe image collapse before creating an image that perfectly satisfies the given text guidance. For example, an image at the first row with 100 channels manipulated do not seem 'young' enough and the extent of manipulation is not improved with using more channels since the image collapse and entanglement becomes a dominating factor. On the other hand, **Ours** shows consistently reasonable and disentangled images even when using all channels in STYLESPACE.

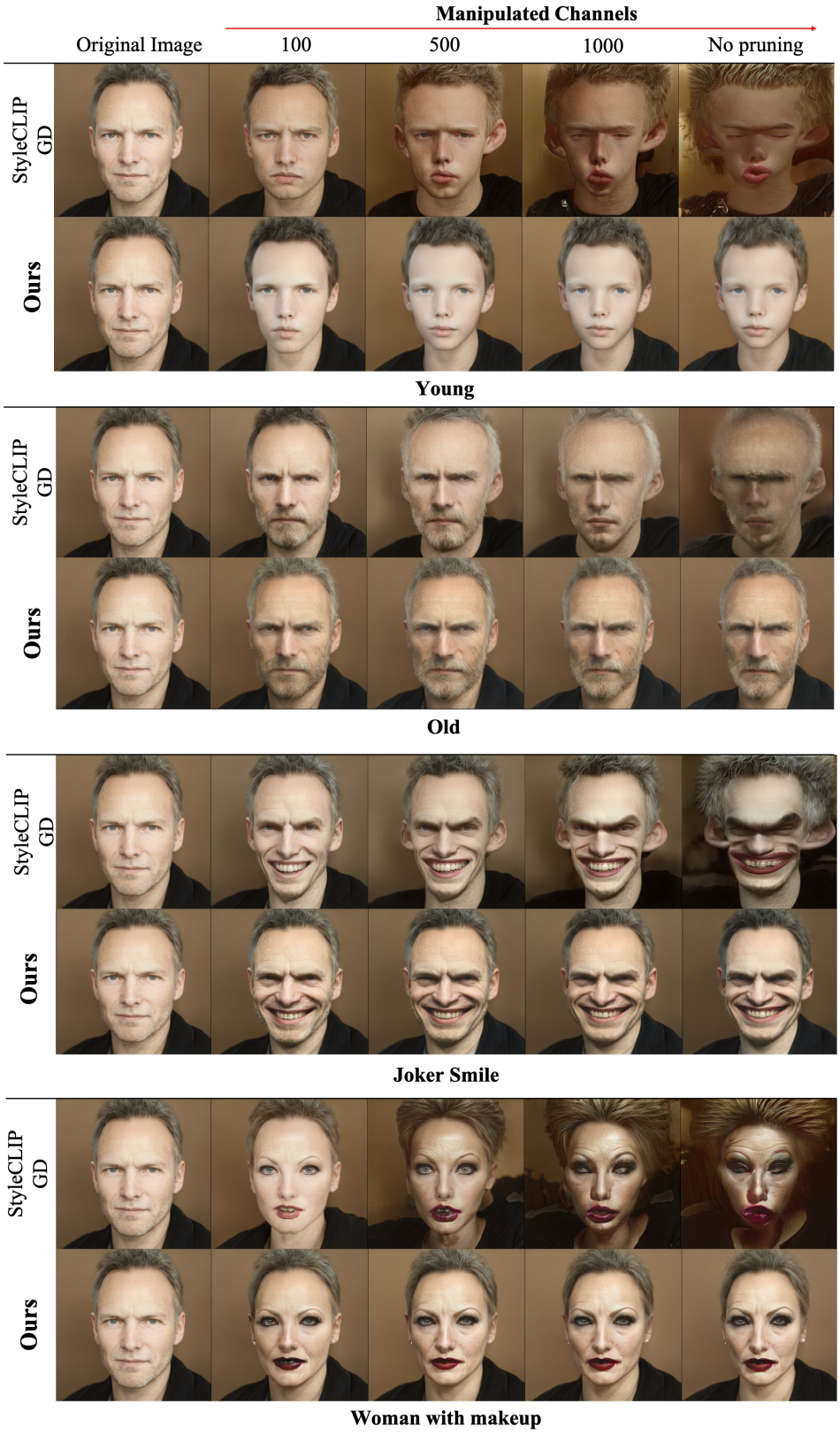


Figure 13: Manipulation results showing the disentanglement of **Ours** when using 100, 500, 1000 channels, and without pruning channels. StyleCLIP `GlobalDirection` manifests image collapse when using over 500 channels while **Ours** shows robustness against entanglement.

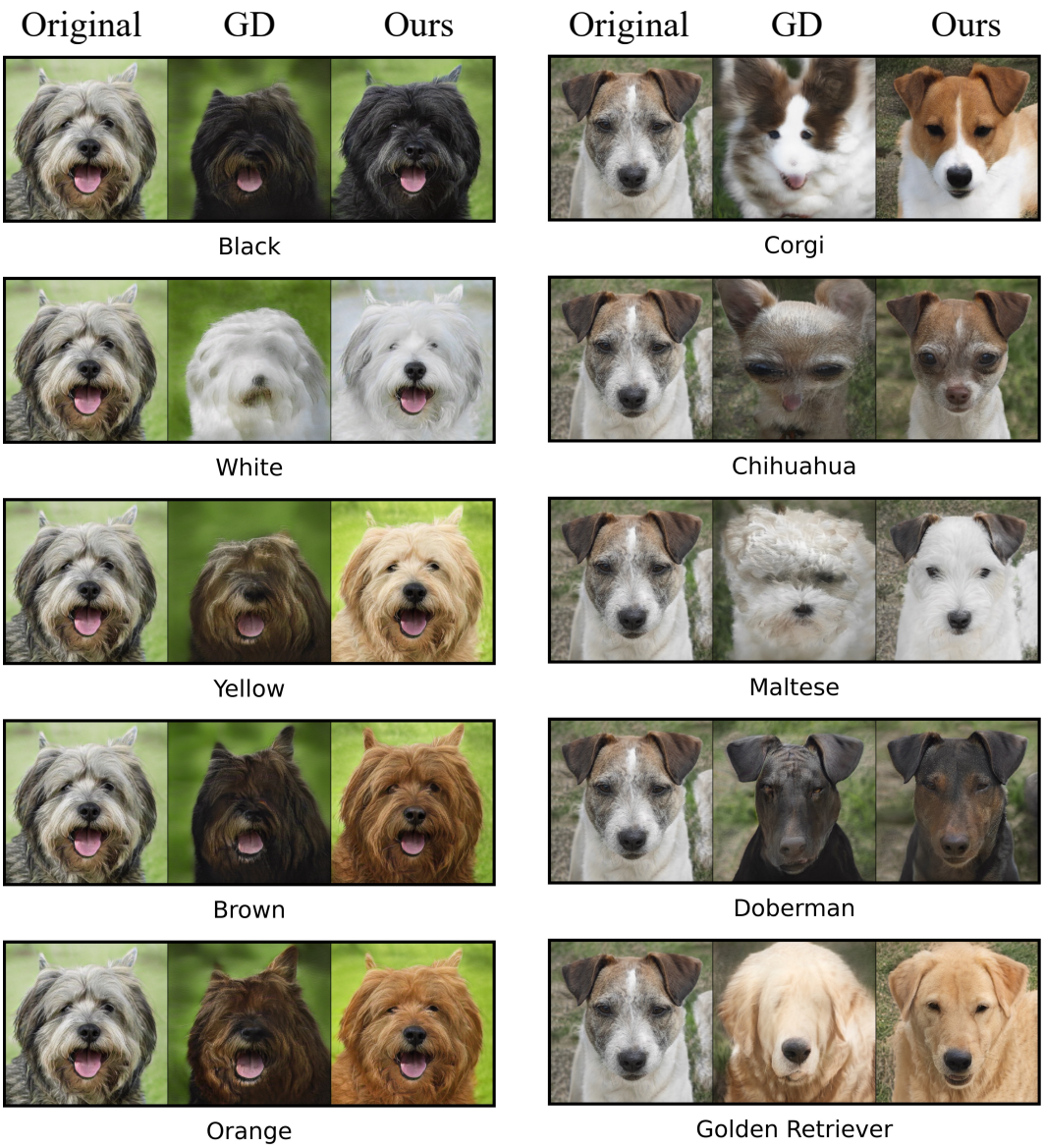


Figure 14: Additional manipulation results on AFHQdog-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Ours**.

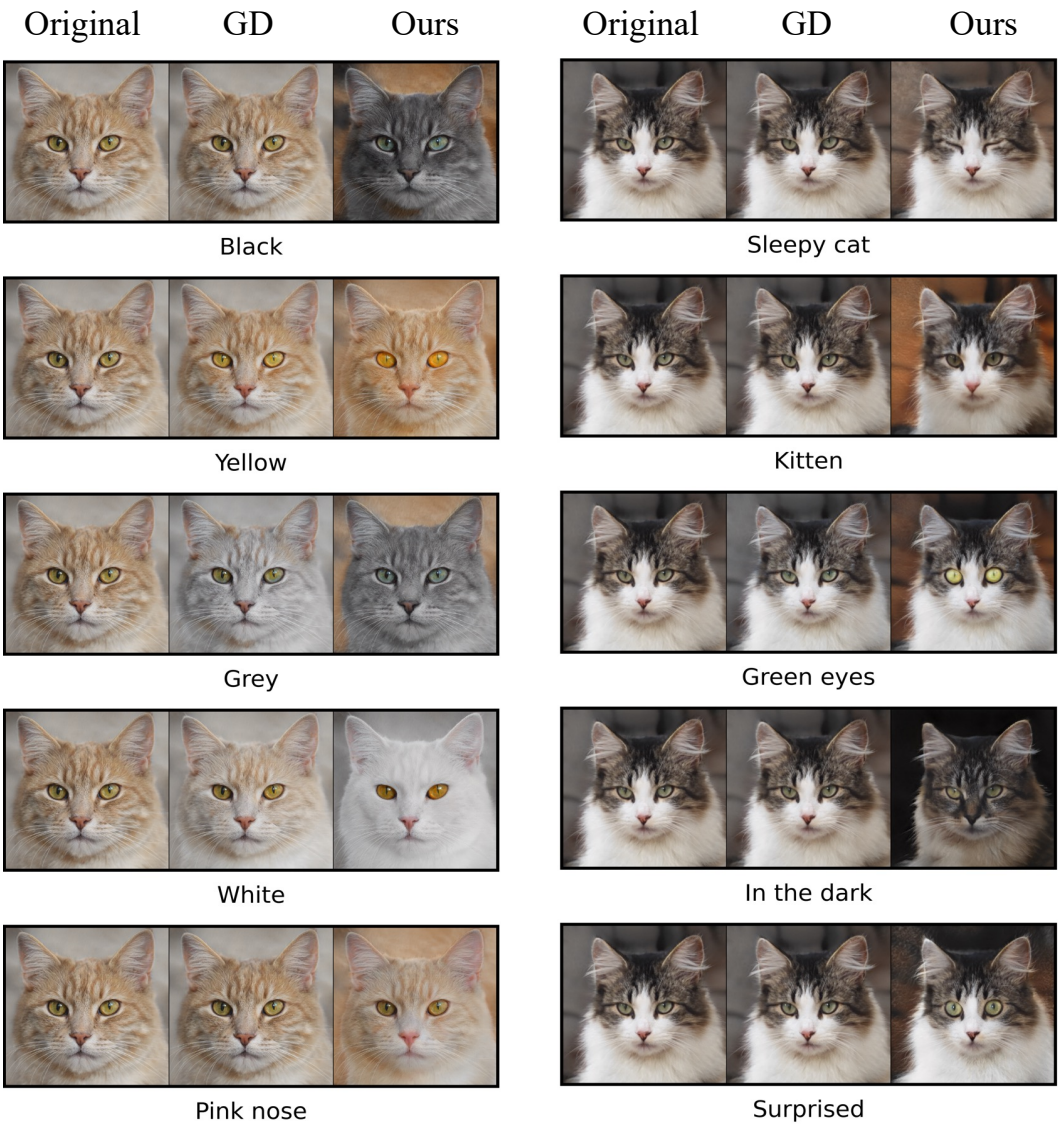


Figure 15: Additional manipulation results on AFHQcat-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Ours**.

Figure 16: Additional manipulation results on LSUNcar-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD)and **Ours**. Used image instead of text as a guidance encoded into CLIP with encoder $E(\cdot)$.

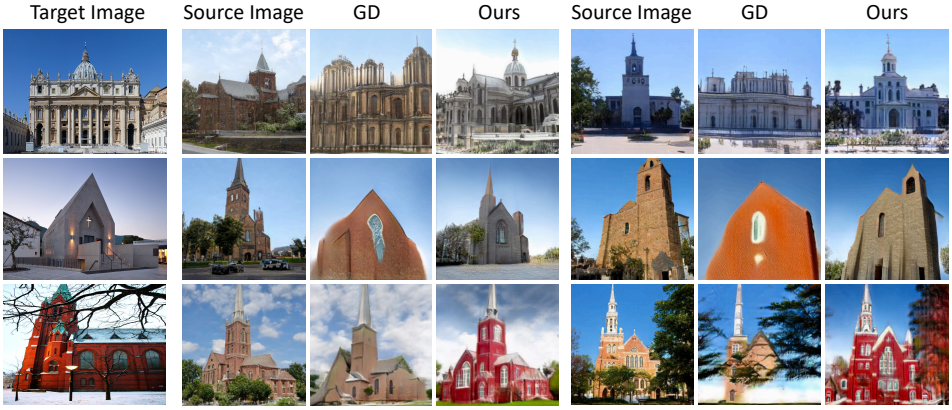


Figure 17: Additional manipulation results on LSUNchuch-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Ours**. Used image instead of text as a guidance encoded into CLIP with encoder $E(\cdot)$.

E EXPERIMENTAL DETAILS

We use unsupervised directions from SeFa and GANspace, both of which are found in intermediate space $\mathcal{W}$ limiting the maximum number of directions to 512, which is the dimension of the intermediate latent space. Therefore, we use two variants of 512 vectors. Since the earlier layers of StyleGAN mainly shows large structural changes while the deep layers are related to a small detailed changes, we apply the 1024 vectors on three groups of layers: coarse(resolution 4∼32), medium(resolution 64∼128) and fine(resolution 128∼) (Karras et al., 2020a), which amounts to total of 3072 unsupervised directions.

Since the text-guided manipulation directions are found in StyleSpace $\mathcal{S}$, we map the 3072 directions into StyleSpace S, where the maximum value of the style channel parameter in each direction is 15. The unsupervised directions are *dense* vectors, where the elements are non-zeros affecting every channel in StyleSpace. *Dense* manipulation vectors are prone to modify multiple regions of images, causing an image to lose the original identity. Therefore, we select top 300 channel parameters whose absolute value is large. Sparsifying the known manipulation vectors boosts the disentangled property of the learned dictionary.

Given 3072 instances of $\boldsymbol{\alpha}$ and the CLIP encoding $\mathbf{d}(\pm\boldsymbol{\alpha})$, we train the dictionary from zero matrix with 30000 epochs. The hyper-parameter λ is set as 0.01, and the learning rate is 3.0 with Adadelta (Zeiler, 2012), an optimization method with adaptive learning rate.

F ABLATION STUDY

Table 4: We report the effect of the hyper-parameter λ and the candidates for $top - r$ for creating the sparsified unsupervised directions. $\mathbf{ID}(.,.)$ indicates ArcFace Loss, which measures a change in identity, between the original image and edited image. Note that **ID** loss tends to increase as the image is manipulated properly.

		$\mathbf{R}_{CLIP}(t, X_{edit})$		$\mathbf{ID}(X, X_{edit})$	
	λ	0.1	0.01	0.1	0.01
Top-r	10	0.125	0.106	0.375	0.175
	10, 30	0.098	0.140	0.075	0.390
	10, 30, 50	0.115	**0.141**	0.417	0.284
	10, 30, 50, 100	0.120	0.121	0.225	0.007