Learning Input-agnostic Manipulation Directions in StyleGAN with Text Guidance

Yoonjeon Kim[1], Hyunsu Kim[2], Junho Kim[2], Yunjey Choi[2],Eunho Yang[1,3]*
Korea Advanced Institute of Science and Technology (KAIST)[1], NAVER AI LAB[2], AITRICS[3]
yoonkim313@kaist.ac.kr hyunsu1125.kim@navercorp.com
jhkim.ai@navercorp.com yunjey.choi@navercorp.com eunhoy@kaist.ac.kr

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 demonstrated 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 allows fast inference and 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)).

[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

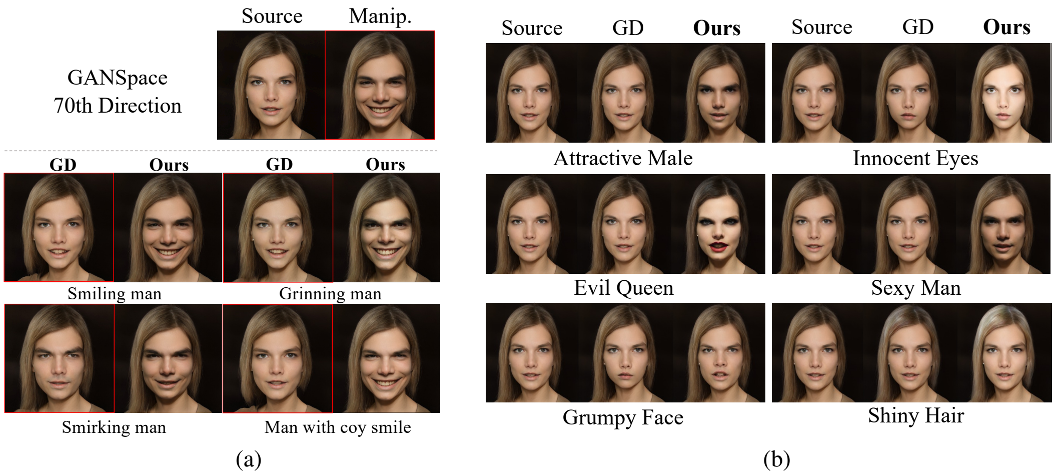


Figure 1: (a) Manipulation by the 70-th direction from GANspace generates 'a man with wide smile'. `GlobalDirection` (GD), highlighted in red, fails to reproduce similar result even when provided with various text guidances. (b) Manipulation results by *randomly selected text*, demostrating that GD has insufficient manipulation ability. Same number of channels are manipulated in both methods.

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 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 (Härkönen et al., 2020; Shen & Zhou, 2021) (see Fig. 1(a) for examples). In addition, we also show that this standard method does not properly perform manipulation on a large number of randomly selected texts (see Fig. 1(b) for examples). 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, named **Multi2One**, of *learning* a *Dictionary* that can manipulate multiple channels corresponding to a given text. However, here since there is no paired ground truth of text and manipulation direction corresponding to the text, we embed the directions found by existing unsupervised methods into the CLIP space and learn a dictionary to reproduce them in the CLIP space. Note that this has more meaning than simply reproducing the known directions derived by unsupervised methods. As the dictionary learns the relationship between channels in StyleSpace and CLIP space, we can find manipulations that could not be found with unsupervised methods using diverse text inputs.

Through extensive experiments, we confirm that 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 used for manipulation 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

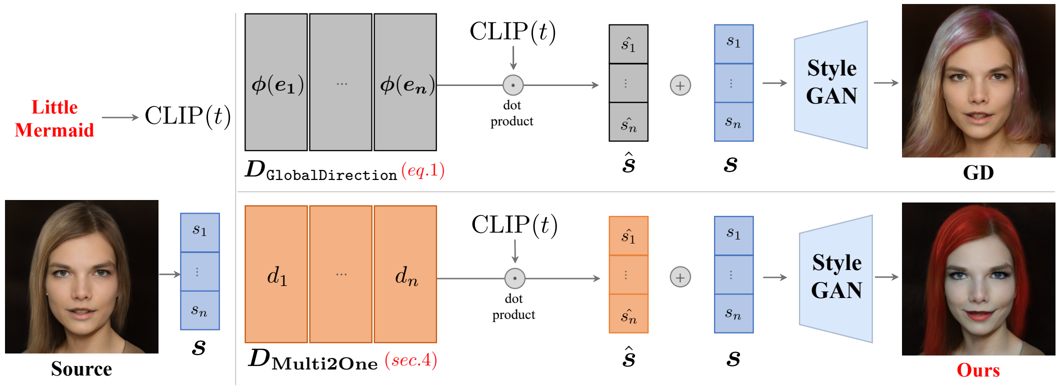


Figure 2: A diagram depicting the framework of dictionary-based image manipulation via text guidance. Our method, **Multi2One**, is differentiated from the previous methods in that it learns the dictionary for the text input. The proposed novel dictionary allows more flexible and expansive discovery of the manipulation direction $\hat{\boldsymbol{s}}$ with better result. $\phi(\cdot)$ in $D_{\texttt{GlobalDirection}}$ is an abbreviation of $\phi_{\text{CLIP}}(\cdot)$ of Eq. (1). The dictionary learning process of **Multi2One** to create $D_{\textbf{Multi2One}}$ is illustrated in Sec. 4.

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.

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 n to denote the total number of channels in $\mathcal{S}$ (that is, $\boldsymbol{s} = [s_1, s_2, \cdots, s_n]^T \in \mathcal{S}$ and s_i is a single parameter). 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. The goal of StyleGAN-based image manipulation via text is to find a direction $\hat{\boldsymbol{s}} = [\hat{s}_1, \hat{s}_2, ..., \cdots, \hat{s}_n]^T$ in StyleSpace $\mathcal{S}$ which generates an image $X_{\text{edited}} = G_s(\boldsymbol{s} + \hat{\boldsymbol{s}})$ suitable for the provided text guidance t. 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. The main benefit of considering StyleSpace $\mathcal{S}$ in image manipulation 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.

Table 1: Measurement of CLIP similarity score $cos(\cdot,\cdot)$ (↑) between the manipulated image and the CLIP representation $\boldsymbol{\phi}_{\text{CLIP}}(\cdot)$ of unsupervised direction $\boldsymbol{\alpha}$.

	$cos(\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}), X_{\text{unsup}})$	$cos(\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}), X_{\text{GD}})$
$\boldsymbol{\alpha}^{(10)}$	**0.3295**	0.2701
$\boldsymbol{\alpha}^{(30)}$	**0.3558**	0.3319
$\boldsymbol{\alpha}^{(50)}$	**0.3687**	0.2924

Text-guided Manipulation by StyleCLIP The `GlobalDirection` (Patashnik et al., 2021) is a representative method of text-driven image manipulation that provides an input-agnostic direction with a level of disentanglement. Intuitively, it computes the similarity between the input text and each style channel in the CLIP space (Radford et al., 2021) to find the channels that should be modified given the text. In order to compute the similarity between the text and style channel, both should be encoded into CLIP. While the text guidance t is trivially encoded via the text encoder of CLIP as CLIP$(t) \in \mathbb{R}^p$, *style* channels in StyleSpace $\mathcal{S}$ need additional pre-processing for the embedding. `GlobalDirection` proposes to embed the manipulation effect of i-th *style* channel using the following mapping from the StyleSpace $\mathcal{S}$ to CLIP space:

$$\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i) := \mathbb{E}_{\boldsymbol{s}\in\mathcal{S}}\Big[\text{CLIP}\big(G_s(\boldsymbol{s}+\boldsymbol{e}_i)\big) - \text{CLIP}\big(G_s(\boldsymbol{s}-\boldsymbol{e}_i)\big)\Big] \tag{1}$$

where the manipulation vector $\boldsymbol{e}_i \in \mathbb{R}^n$ is a zero-vector except for the i-th entry. Adding the manipulation vector $\boldsymbol{e}_i$ to the original style vector $\boldsymbol{s}$ indicates that only the i-th channel among n channels in StyleSpace is manipulated. Note that $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i)$ is also p-dimensional since the CLIP encoder maps images generated by $G_s(\cdot)$ into a p-dimensional CLIP space. The above mapping in Eq. (1) is enumerated across all n channels in StyleSpace $\mathcal{S}$ to create a dictionary $D_{\texttt{GlobalDirection}} := [\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_1), \boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_2), \cdots, \boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_n)] \in \mathbb{R}^{p\times n}$.

Finally, with this dictionary, the manipulation direction $\hat{\boldsymbol{s}} \in \mathbb{R}^n$ by `GlobalDirection` is given as the similarity score measured by the following equation:

$$\hat{\boldsymbol{s}} = D^T_{\texttt{GlobalDirection}}\text{CLIP}(t). \tag{2}$$

This overall manipulation procedure is visualized in Fig. 2.

In the following section (Sec. 3.2), we propose the evidence to the hypothesis that the single-channel encoding strategy with $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i)$ to create the dictionary D is the major bottleneck causing limited coverage issues that StyleCLIP suffers from.

3.2 Coverage Analysis of StyleCLIP `GlobalDirection` Method

In this section, we experimentally verify that `GlobalDirection` has limited ability to find manipulation directions, as described in Fig. 1 of the introduction. Toward this we show two empirical findings, which corresponds to Fig. 1(a) and Fig. 1(b), respectively.

First, we show that many edits using unsupervised methods (Härkönen et al., 2020; Shen & Zhou, 2021) cannot be recovered by `GlobalDirection`. For example, in Fig. 1(a), we can observe that applying the 70-th GANspace direction manipulates the source image to become a man with wide smile, showing that pre-trained StyleGAN itself is capable of such manipulation. However, `GlobalDirection` (GD) in Fig. 1(a) constantly fails to recover the same editing effect of the unsupervised direction, despite the variation on the text input such as ‘smiling man’, ‘grinning man’, ‘smirking man’, and ‘man with coy smile’.

More quantitatively, we provide Tab. 1 to show that `GlobalDirection` (Patashnik et al., 2021) cannot effectively recover the directions found by unsupervised methods (Härkönen et al., 2020; Shen & Zhou, 2021). Scores in the table are the CLIP similarity between $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}) = \mathbb{E}_{\boldsymbol{s}\in\mathcal{S}}[\text{CLIP}\big(G_s(\boldsymbol{s}+\boldsymbol{\alpha})\big) - \text{CLIP}\big(G_s(\boldsymbol{s}-\boldsymbol{\alpha})\big)]$ and the modified image created by either `GlobalDirection` (X_{GD}; the construction of X_{GD} is explained below)[2] or unsupervised methods (X_{unsup}). Note that $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha})$ encodes the direction $\boldsymbol{\alpha}$ into the CLIP embedding space as `GlobalDirection` does for a single

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

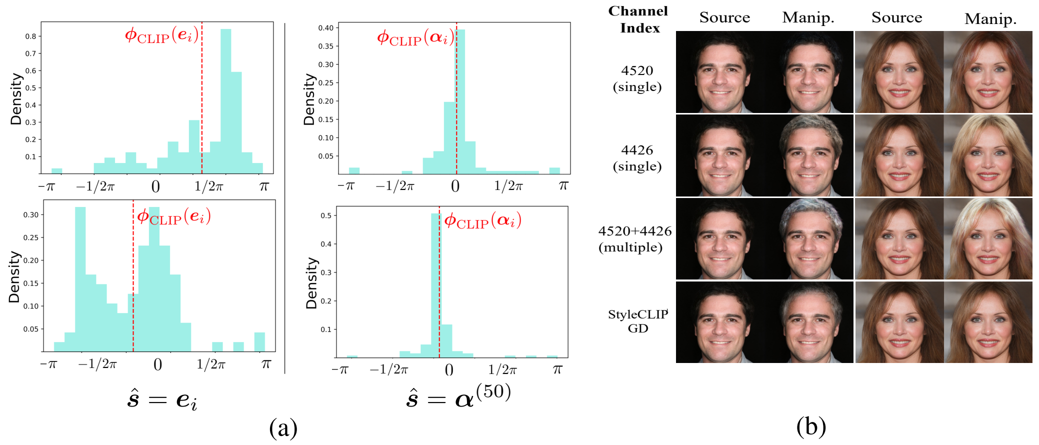


Figure 3: (a) Histogram of $\text{CLIP}\big(G_s(\boldsymbol{s} + \hat{\boldsymbol{s}})\big) - \text{CLIP}\big(G_s(\boldsymbol{s} - \hat{\boldsymbol{s}})\big)$ using single-channel manipulation $\boldsymbol{e}_i$ or multiple channel manipulation $\boldsymbol{\alpha}^{(50)}$. (b) Manipulation results using channel indices 4520, 4426. `GlobalDirection` (denoted as StyleCLIP GD) manipulated with text 'white hair'.

channel. From the table, we can confirm that `GlobalDirection` does not recover the similarity in CLIP space that the unsupervised methods have.

Here we illustrate the details on how we generate X_{unsup} and X_{GD} for Tab. 1. First, the source image $X = G_s(\boldsymbol{s})$ is manipulated by each unsupervised direction $\boldsymbol{\alpha} \in \mathbb{R}^n$ into $X_{\text{unsup}} = G_s(\boldsymbol{s} + \boldsymbol{\alpha})$. Instead of naïvely using the direction $\boldsymbol{\alpha} \in \mathbb{R}^n$, we only use the channels whose magnitude is the k-largest among n channels in $\boldsymbol{\alpha}$ to reduce the entanglement that naturally comes with directions found by unsupervised methods (please refer to Appendix C to find the effect of using limited number of channels). We denote such sparsified directions as $\boldsymbol{\alpha}^{(k)} \in \mathbb{R}^n$ and use them in the experiment of Tab. 1. Then, to generate $X_{\text{GD}} = G_s(\boldsymbol{s} + \hat{\boldsymbol{\alpha}})$, we compute $\hat{\boldsymbol{\alpha}} = D^T_{\texttt{GlobalDirection}}\text{CLIP}(t) \in \mathbb{R}^n$ as in Eq. (2). However, since there is no text corresponding to $\boldsymbol{\alpha}$ and our objective in this experiment is to find a direction $\hat{\boldsymbol{\alpha}}$ that may recover X_{unsup} by `GlobalDirection`, we replace $\text{CLIP}(t)$ with $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(k)}) = \mathbb{E}_{\boldsymbol{s}\in\mathcal{S}}[\text{CLIP}\big(G_s(\boldsymbol{s}+\boldsymbol{\alpha}^{(k)})\big) - \text{CLIP}\big(G_s(\boldsymbol{s}-\boldsymbol{\alpha}^{(k)})\big)]$. Since $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(k)})$ indicates the manipulation effect of $\boldsymbol{\alpha}^{(k)}$ encoded into CLIP space, it is used as a substitute for the CLIP-encoded text guidance, $\text{CLIP}(t)$. Finally, we compute the similarity score for 30 instances of source images and 1024 directions from SeFa and GANspace whose average is reported in Tab. 1.

Secondly in Fig. 1(b), we provide several instances where the `GlobalDirection` of StyleCLIP fails to manipulate an image given randomly selected text. For additional examples on the failure cases of the standard method, please refer to Appendix A where we provide the details on how we choose random texts and manipulated images using extensive list of random texts.

Regarding the phenomena found in the above experiments, we hypothesize the crude assumption of `GlobalDirection` - manipulation solely based on single channel could fully represent the manipulative ability of StyleGAN - leads to the limited coverage issue. To disprove this assumption of `GlobalDirection`, in Fig. 3(a), we demonstrate that single-channel manipulation framework of `GlobalDirection` (using $\boldsymbol{e}_i$) leads to improper construction of $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i)$ in the dictionary $D_{\texttt{GlobalDirection}}$. Then in Fig. 3(b), we show an example where editing two channels has a completely different effect from manipulating each of the channels individually, which proves the need for multi-channel manipulation.

In Fig. 3(a), we show that $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i) = \mathbb{E}\big[\text{CLIP}\big(G_s(\boldsymbol{s}+\boldsymbol{e}_i)\big) - \text{CLIP}\big(G_s(\boldsymbol{s}-\boldsymbol{e}_i)\big)\big]$, which forms the dictionary $D_{\texttt{GlobalDirection}}$, is not a robust representative of i-th style channel since there exists sample-wise inconsistency in the calculation of $\text{CLIP}\big(G_s(\boldsymbol{s}+\boldsymbol{e}_i)\big) - \text{CLIP}\big(G_s(\boldsymbol{s}-\boldsymbol{e}_i)\big)$. On the left side of Fig. 3(a), the distribution of $\text{CLIP}\big(G_s(\boldsymbol{s}+\boldsymbol{e}_i)\big) - \text{CLIP}\big(G_s(\boldsymbol{s}-\boldsymbol{e}_i)\big)$ and its average over the samples, $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i)$, are shown. As the samples show inconsistent angles over polar coordinate[3] when modified by a single-channel manipulation $\boldsymbol{e}_i$, we conclude that $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i)$ may not be a trust-worthy representative of i-th channel in StyleSpace. Thereby the basic assumption held by `GlobalDirection` which expects a single channel manipulation using $\boldsymbol{e}_i$ on any images to be semantically consistent in CLIP space, in fact, has numerous counter-examples showing inconsistency.

[3] A more detailed explanation on representing the samples as directions for an angle histogram is deferred to Appendix B due to space constraints.

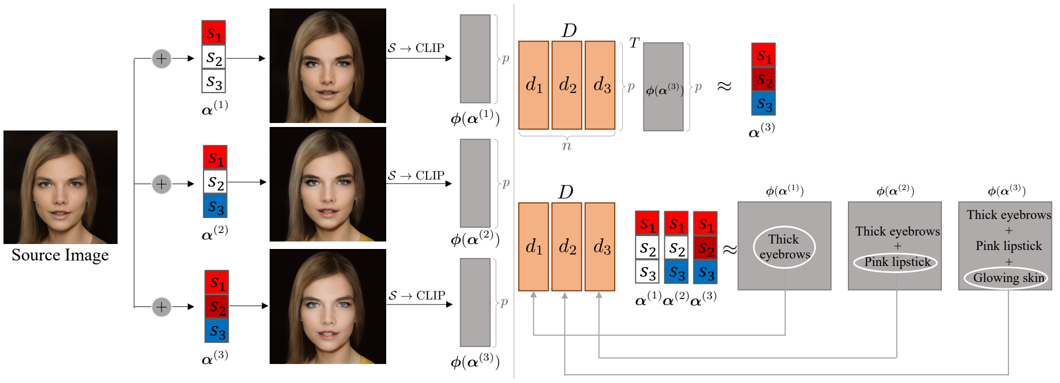


Figure 4: A concept diagram of our method in learning the *dictionary* D. $\phi(\cdot)$ is an abbreviation of $\phi_{\text{CLIP}}(\cdot)$. $\mathcal{S} \to$ CLIP represents a mapping from StyleSpace to CLIP.

On the other hand, it is observable on the right side of Fig. 3(a), that manipulation by unsupervised direction $\alpha^{(50)}$, shows much consistent results.

Moreover, in Fig. 3(b), we show an example where single channel alone has completely different meaning from the manipulation by the combination of multiple channels. This supports our claim that the interaction between multiple channels should be considered, rather than individually encoding single channel each. In Fig. 3(b), manipulation by 4520-th or 4426-th do not show "completely white hair" while the combination of the two shows white hair. Since `GlobalDirection` fails to learn that 4520-th also has a role for creating white hair, it fails to successfully modify an image when text guidance is given as ‘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* that could be possibly a set of channels in StyleSpace $\mathcal{S}$.

Ideally, given a text input t and its CLIP encoding $\text{CLIP}(t) \in \mathbb{R}^p$, we would like to construct a dictionary $D \in \mathbb{R}^{p\times n}$ which finds a manipulation direction $\hat{\boldsymbol{s}} = D^T\text{CLIP}(t) \in \mathbb{R}^n$ as in Eq. (2). The direction $\hat{\boldsymbol{s}}$ is expected to manipulate an image into $G_s(\boldsymbol{s} + \hat{\boldsymbol{s}})$ which faithfully represents t. Let us for now assume that the ‘ground truth’ manipulation direction $\hat{\boldsymbol{s}}_t \in \mathbb{R}^n$ corresponding to the driving text t is known. Then our estimated direction $\hat{\boldsymbol{s}} = D^T\text{CLIP}(t)$ should generate an image $X_{\text{edited}} = G_s(\boldsymbol{s} + D^T\text{CLIP}(t))$ close to $G_s(\boldsymbol{s} + \hat{\boldsymbol{s}}_t)$. This leads to the objective

$$\underset{D}{\text{minimize}} \left\| D^T\text{CLIP}(t) - \hat{\boldsymbol{s}}_t \right\|_2^2 \tag{3}$$

which we name *ideal* dictionary learning problem since the ground truth direction $\hat{\boldsymbol{s}}_t$ for a driving text t is unknown. To avoid this unrealistic assumption, we substitute $\hat{\boldsymbol{s}}_t$ with the known directions $\boldsymbol{\alpha} \in \mathbb{R}^n$ derived from unsupervised methods (Härkönen et al., 2020; Shen & Zhou, 2021). $\text{CLIP}(t)$ is also replaced by $\phi_{\text{CLIP}}(\boldsymbol{\alpha})$ (see Eq. (1) for the definition of $\phi_{\text{CLIP}}(\cdot)$) since we would like to use $\boldsymbol{\alpha}$ itself as a driving guidance rather than finding out corresponding text t that requires labor-intensive labeling. Then, Eq. (3) becomes

$$\underset{D}{\text{minimize}} \sum_{\forall \boldsymbol{\alpha}} \left\| \, D^T\phi_{\text{CLIP}}(\boldsymbol{\alpha}) - \boldsymbol{\alpha} \right\|_2^2. \tag{4}$$

We construct the set of known directions $\boldsymbol{\alpha}$ in Eq. (4) by using all 512 directions found by GANspace and directions with top 80 eigenvalues out of 512 from SeFa[4] (Härkönen et al., 2020; Shen & Zhou, 2021). Instead of naïvely using the unsupervised direction $\boldsymbol{\alpha} \in \mathbb{R}^n$, we prune the channels in directions to activate only k channels among n channels, zeroing out the other $n-k$ channels that have small magnitudes. Such pruning is represented by $\boldsymbol{\alpha}^{(k)} \in \mathbb{R}^n$ where we use $k = 10, 30, 50$ in this paper (the effect of k is described in Tab. 5 in the appendix). The reason for pruning the number of activated channels is that unsupervised methods show instability in terms of maintaining

[4] This is based on the observation that the change in image is not significant for the directions with lower eigenvalues in SeFa.

Table 2: Mean Squared Error (Eq. (6); ↓) measured in StyleSpace. We manipulate total of 100 channels using both `GlobalDirection` and **Ours**. Note that $\boldsymbol{\alpha}$ is $\boldsymbol{\alpha}^{(100)}$ in this metric.

	`Global Direction`	**Ours**
$\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}})$	0.0315 ± 0.0037	**0.0057** ± 0.0034
$\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}\vert_{\neq 0}\big)$	0.0267 ± 0.00114	**0.0005** ± 0.00001
$\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}\vert_{=0}\big)$	0.006218 ± 0.00658	**0.00486** ± 0.00363

the original image due to the entanglement (please refer to Appendix C for examples of entanglement without pruning) and we circumvent such issue by constraining the direction $\boldsymbol{\alpha} \in \mathbb{R}^n$ to have limited effect on only k channels.

To help close the gap between Eq. (3) and Eq. (4), we introduce an additional loss term $\|D\boldsymbol{\alpha}^{(k)} - \boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(k)})\|_2^2$ to update the dictionary $D \in \mathbb{R}^{p \times n}$. This additional term comes from the observation in Sec. 3.2 that multi-channel manipulation should be considered. Unlike `GlobalDirection` which directly encodes $\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{e}_i)$ for all n channels in StyleSpace, we aim to learn the multi-channel manipulation by $\boldsymbol{\alpha}^{(k)}$ whose editing effect is mapped from StyleSpace to CLIP space by $\boldsymbol{\phi}_{\text{CLIP}}(\cdot)$. The main obstacle in this objective is identifying which channel is responsible for which part of the change induced by total of k channels in the given direction $\boldsymbol{\alpha}^{(k)}$. Here, we emphasize that augmenting the same direction $\boldsymbol{\alpha}$ into $\boldsymbol{\alpha}^{(10)}, \boldsymbol{\alpha}^{(30)}, \boldsymbol{\alpha}^{(50)}$ can encourage the dictionary learning process to identify a disentangled editing effect of each channel which will be explained further below.

Based on the concept figure in Fig. 4, we explain how the additional loss term encourages disentangled learning of dictionary D. We present a simplified version where $k = 1, 2, 3$, creating three augmentations $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \boldsymbol{\alpha}^{(3)}$ for the direction $\boldsymbol{\alpha}$. Suppose manipulation using $\boldsymbol{\alpha}^{(3)} = [s_1, s_2, s_3]^T$ modifies the source image to have 1) thick eyebrows, 2) pink lipstick and 3) glowing skin. Further suppose manipulation by $\boldsymbol{\alpha}^{(2)} = [s_1, 0, s_3]^T$ modifies 1) thick eyebrows and 2) pink lipstick. Finally, manipulation using $\boldsymbol{\alpha}^{(1)} = [s_1, 0, 0]^T$ modifies 1) thick eyebrows. If the loss term $\|D\boldsymbol{\alpha}^{(k)} - \boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(k)})\|_2^2$ is perfectly optimized to satisfy $D\boldsymbol{\alpha}^{(k)} = \boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(k)})$, then $s_1 d_1 \approx$ CLIP(thick eyebrows), $s_1 d_1 + s_3 d_3 \approx$ CLIP(thick eyebrows & pink lipsticks) and $s_1 d_1 + s_2 d_2 + s_3 d_3 \approx$ CLIP(thick eyebrows & pink lipsticks & glowing skin). Therefore, the linear combination on the sparsified direction helps to specifically link 'thick eyebrows' to d_1 'pink lipstick' to d_3, and 'glowing skin' to d_2. Hence our method may have greatly disentangled and diverse directions coming from the enriched coverage over manipulation in StyleGAN, thanks to the dictionary learning based on multi-channel manipulation. Moreover, even though our $\boldsymbol{\alpha}$ comes from the unsupervised methods, our method learns the individual role of s_1, s_2 and s_3 from the single unsupervised direction $\boldsymbol{\alpha} = [s_1, s_2, s_3]^T$. This leads to greatly diverse and disentangled results compared to unsupervised methods which will be proved in Tab. 3.

Finally, combining these two ingredients together yields our dictionary learning problem by considering all known directions $\boldsymbol{\alpha}^{(k)}$:

$$\underset{D}{\text{minimize}} \sum_{\forall \boldsymbol{\alpha}^{(k)}} \left\|D^T \boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(k)}) - \boldsymbol{\alpha}^{(k)}\right\|_2^2 + \lambda \cdot \left\|D\boldsymbol{\alpha}^{(k)} - \boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(k)})\right\|_2^2 \tag{5}$$

where λ is the tunable hyper-parameter and is set as 0.01 in all our experiments below.

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, 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

Table 3: Cosine similarity R_{CLIP} (Eq. (7); ↑) measured in CLIP Space. For unsupervised method, which does not depend on text guidance, we choose the text-related unsupervised direction $\boldsymbol{\alpha}$ that has the highest similarity value of $cos(\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}), \text{CLIP}(t))$ given text t. We also present the similarity value on our method where hyper-parameter λ is 0 for ablation study on the second term of Eq. (5).

	Methods	Unsupervised	`GlobalDirection`	**Ours** ($\lambda = 0$)	**Ours** ($\lambda = 0.01$)
Manip. Channels	**10**	0.219	0.224	0.235	**0.242**
	30	0.221	0.228	0.240	**0.249**
	50	0.223	0.232	0.243	**0.251**
	100	0.225	0.237	0.248	**0.254**

unsupervised directions and in CLIP space for evaluation over text guidance. **Multi2One** recover the unsupervised direction with maximum manipulative ability and minimum entanglement, for demonstration of which we define two metrics as follows.

- **StyleSpace metric for *unsupervised directions* (Tab. 2):** We employ the following metric:
$$\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) := \mathbb{E}_{\boldsymbol{\alpha}\sim\mathbb{A}}\left[\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2\right] \tag{6}$$
where $\boldsymbol{\alpha}$ is from a set of unsupervised directions $\mathbb{A}$. Note that the ground truth direction $\boldsymbol{\alpha}$ in this metric is $\boldsymbol{\alpha}^{(100)}$, where top-100 channels from unsupervised directions are modified while other $n - 100$ channels are zero entries. Therefore the reconstruction of such direction $\boldsymbol{\alpha}^{(100)}$, given by $\hat{\boldsymbol{\alpha}} = D^T\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}^{(100)})$ is expected to have 100 *style* channels whose values are similar to that of non-zero entries in $\boldsymbol{\alpha}^{(100)}$ while the other $n - 100$ *styles* are close to 0. $\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}|_{\neq 0}\big)$ denotes the difference between $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}$ measured between the 100 non-zero *style* channels that represents the ability of the given dictionary D to find the precise manipulation direction. On the other hand, $\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}|_{=0}\big)$ is the distance between $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}$ measured within $n - 100$ zero-entries. This represents the level of disentanglement since zero-entries are not expected to be modified in $\hat{\boldsymbol{\alpha}}$.
- **CLIP space metric for *text guidance* (Tab. 3):** We measure the similarity between edited image and the text guidance in CLIP space as follows:
$$\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}}) := \mathbb{E}_{t\sim\mathbb{T}}\left[cos\big(\text{CLIP}(t), \text{CLIP}(X_{\text{edited}})\big)\right]. \tag{7}$$
We denote the set of text prompts that describes the pre-trained domain of StyleGAN, namely FFHQ as $\mathbb{T}$ whose construction is described in Appendix A.

Tab. 2 shows that **Multi2One** consistently outperforms `GlobalDirection` in terms of MSE score between the unsupervised direction $\boldsymbol{\alpha}$ and the estimated direction $\hat{\boldsymbol{\alpha}}$. Especially, note that **Multi2One** have lower $\mathbf{MSE}\big(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}|_{=0}\big)$ than `GlobalDirection`, proving that our method shows less entanglement.

Moreover, Tab. 3 shows that the cosine similarity between the given text guidance and the manipulated image is always the highest using our method. Most importantly, we emphasize that our method could find manipulation directions that could not have been found by unsupervised methods. We prove this claim using 57 text prompts ($t \sim \mathbb{T}$) as a guidance to manipulate the images. Since we aim to find the unsupervised direction that could edit the image to become t, we select the $\boldsymbol{\alpha}$ whose similarity $cos(\boldsymbol{\phi}_{\text{CLIP}}(\boldsymbol{\alpha}), \text{CLIP}(t))$ is the largest then manipulate the image to produce $X_{\text{unsup}} = G_s(\boldsymbol{s} + \boldsymbol{\alpha})$. In Tab. 3, **Multi2One** scores better $\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$ compared to unsupervised methods. Therefore, we claim that our method successfully learns to find expansive editing directions which could not be found by unsupervised methods given diverse text inputs. Furthermore, we report $\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$ score with ablation on the second term of Eq. (5) in Tab. 3. For visual examples see Appendix G.

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 for StyleCLIP, 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[5] in StyleSpace is always fixed as 10.

[5] 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.

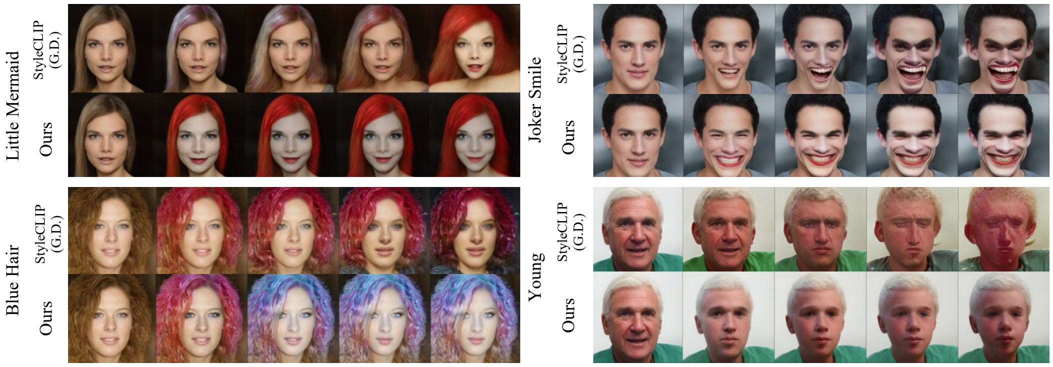


Figure 5: Manipulation results using FFHQ-pretrained StyleGAN2-ADA via text guidance on our method and StyleCLIP `GlobalDirection`. The samples are manipulated with increasing number of channels that are modified leading to larger changes.

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'.

Here we show that our method successfully discovers a direction that does not exist in unsupervised direction. To be more specific, we manipulate the image using an unsupervised direction $\boldsymbol{\alpha}$ whose CLIP representation, $\phi_{\text{CLIP}}(\boldsymbol{\alpha})$, is the most similar with the text 'Little Mermaid' and 'Joker smile'. Then from the manipulated image, we observe that the manipulation result is hardly plausible compared to the successful results in Fig. 5. Based on this examples, we emphasize that even though the dictionary learning process relies on the unsupervised methods, **Multi2One** demonstrate wider variety of manipulation results using the versatility of text.

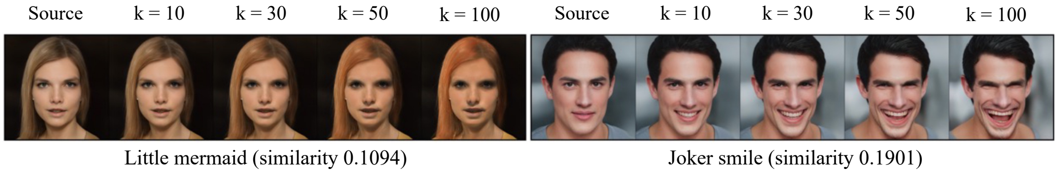

Source
k = 10
k = 30
k = 50
k = 100
Source
k = 10
k = 30
k = 50
k = 100
Little mermaid (similarity 0.1094)
Joker smile (similarity 0.1901)

Figure 6: **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 top-k channels from the found direction.

We show additional results of our method compared to `GlobalDirection` based on AFHQ and LSUN dataset in Appendix D due to space constraints.

6 Conclusion

Text-guided image manipulation has been largely preferred over unsupervised methods in finding manipulation directions due to its ability to find wide variety of directions corresponding to flexible text input. However, we have investigated that the state-of-the-art method of text-guided image manipulation has limited coverage over possible editing directions using StyleGAN and this problem rises from a simple assumption that a single-channel manipulation could fully represent the manipulation ability of StyleGAN. To overcome this issue, we have proposed a dictionary learning framework that embeds the interactive and collective editing effect based on modifying multiple style channels into CLIP space jointly. Our method have been proven to be superior over other text-guided image manipulation methods in terms of manipulative ability and disentanglement.

REPRODUCIBILITY STATEMENT

Our code is open sourced at here.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)). This project is also supported by KAIST-NAVER Hypercreative AI Center.

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.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

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. 1(b)

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'. The number of such adjective-noun pair amounts to 57 instances.

We manipulate the original image on the leftmost side using `GlobalDirection` and **Multi2One** using the same set of text description and equal configuration for both methods (manipulated channels 5, manipulation strength 5). The result in Fig. 7 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 **Multi2One** to overcome the coverage issue.

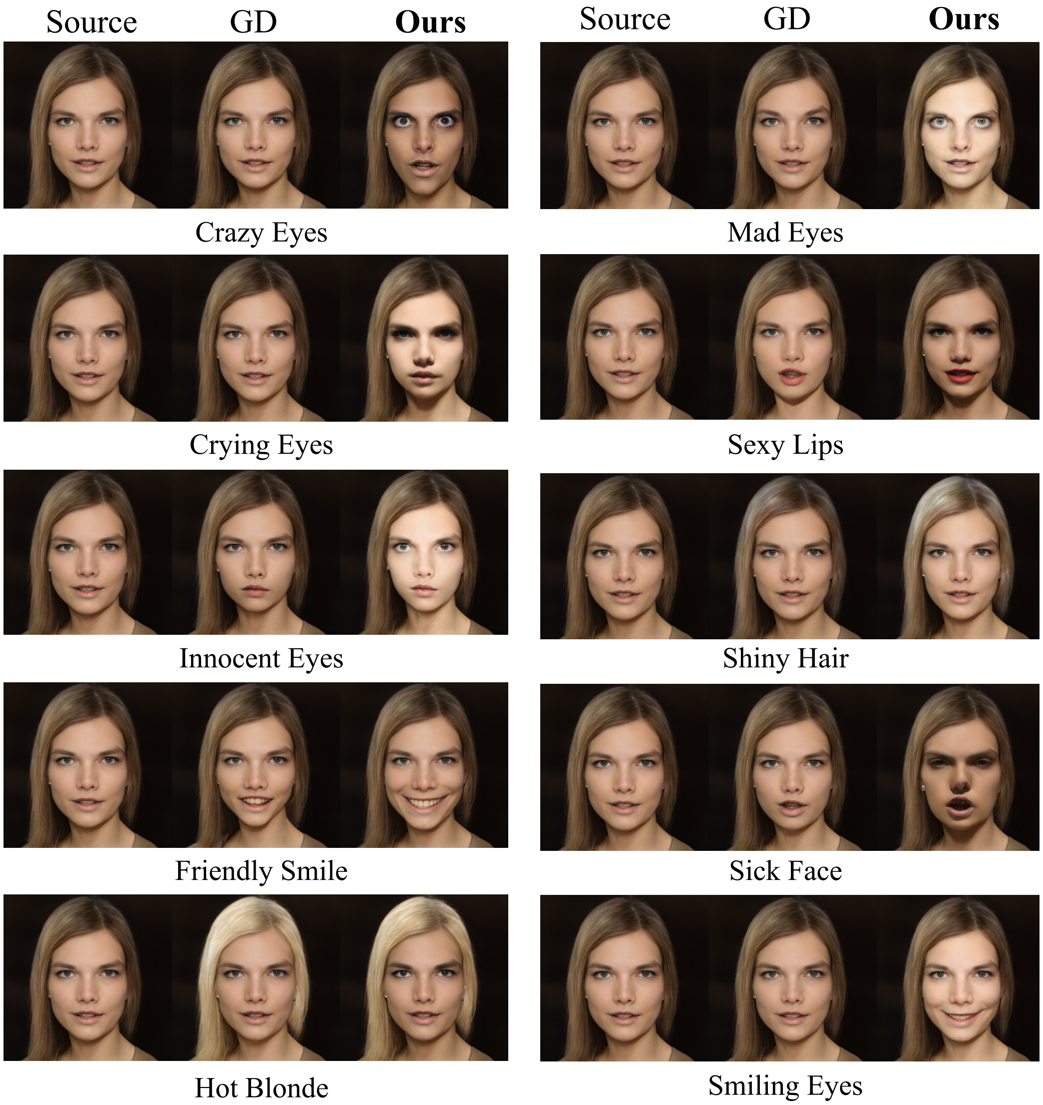


Figure 7: Manipulated images using randomly selected text from Visual Semantic Ontology (VSO). We show manipulated image via `GlobalDirection` (denoted as GD) and **Multi2One**.

B More Details on Fig. 3(a)

Both `GlobalDirection` and **Multi2One** 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, **Multi2One** 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 $E\big(G_s(s+\boldsymbol{s}_i)\big) - E\big(G_s(s-\boldsymbol{s}_i)\big)$ 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 $\mathbb{E}\Big[E\big(G_s(\boldsymbol{s}+\boldsymbol{s}_i)\big) - E\big(G_s(\boldsymbol{s}-\boldsymbol{s}_i)\big)\Big]$ in 2-dimensional space is $(\overline{a}, \overline{b})$. 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 $\arctan2(\overline{b}, \overline{a})$ is emphasized as a red line in Fig. 3(a).

C Examples of Entanglement using Unsupervised Methods

In Fig. 8, we demonstrate the effect of using limited number of channels among all non-zero channels in unsupervised direction $\boldsymbol{\alpha} \in \mathbb{R}^n$.

Since StyleGAN generator consists of progressiveGAN structure where the resolution of generated image progressively increases, it is possible to classify the generator structure into three blocks: coarse, medium and fine. We apply the unsupervised directions on each of the blocks. Moreover, by filtering out to only use k manipulated channels, we show that the source image shows less drastic change. By manipulating all channels, which is shown at the rightmost side of the figure, the source image is drastically changed showing entanglement.

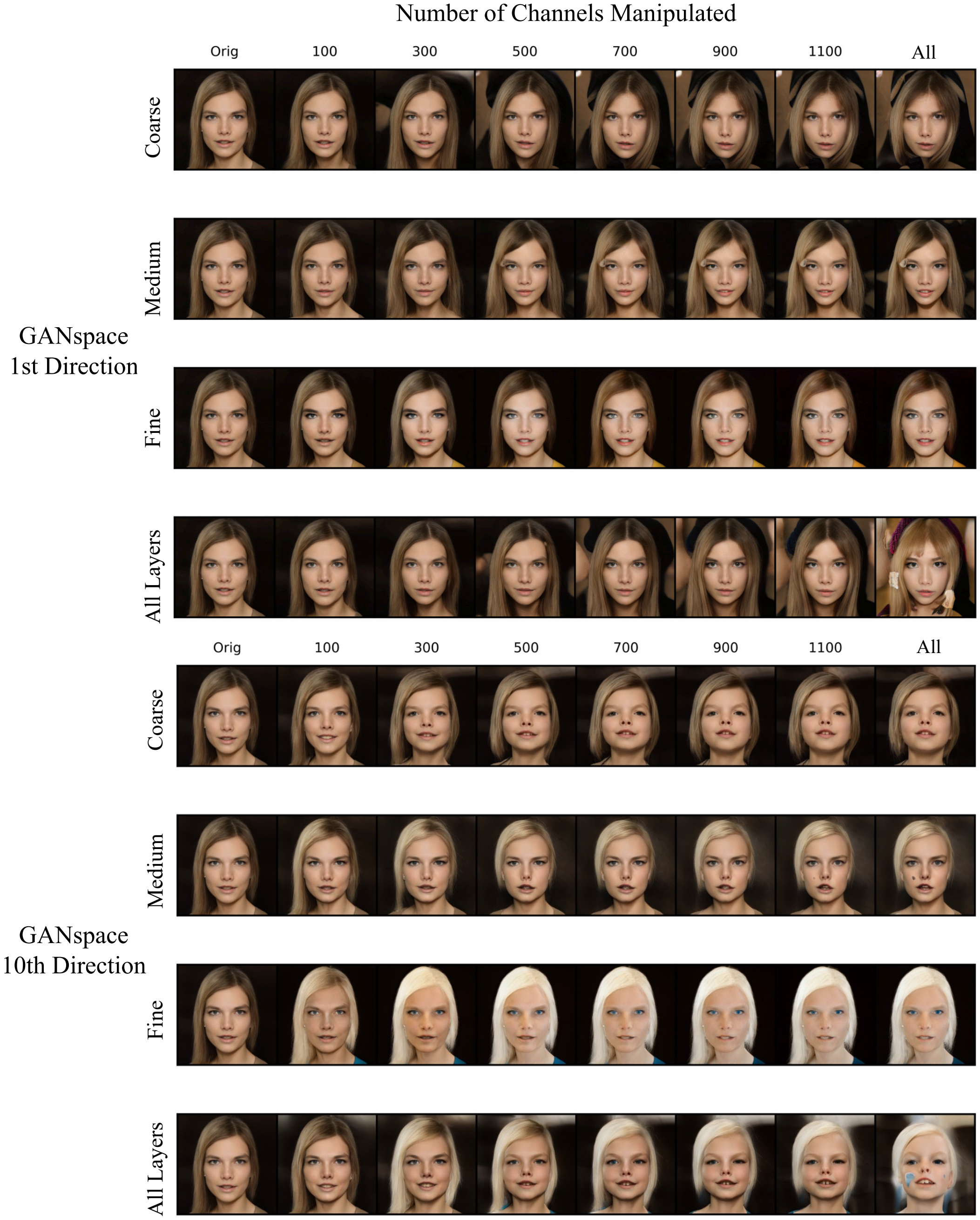


Figure 8: The manipulation results using two directions from GANSpace shows that the source identity is impaired when modifying all 18 layers of $\mathcal{W}$. Moreover, limiting the number of manipulated channels for each coarse, medium, and fine layers leads to significantly better results in terms of disentanglement.

D More Experimental Examples for all dataset

In this section, we provide more experimental results of **Multi2One** with FFHQ, AFHQ, and LSUN.

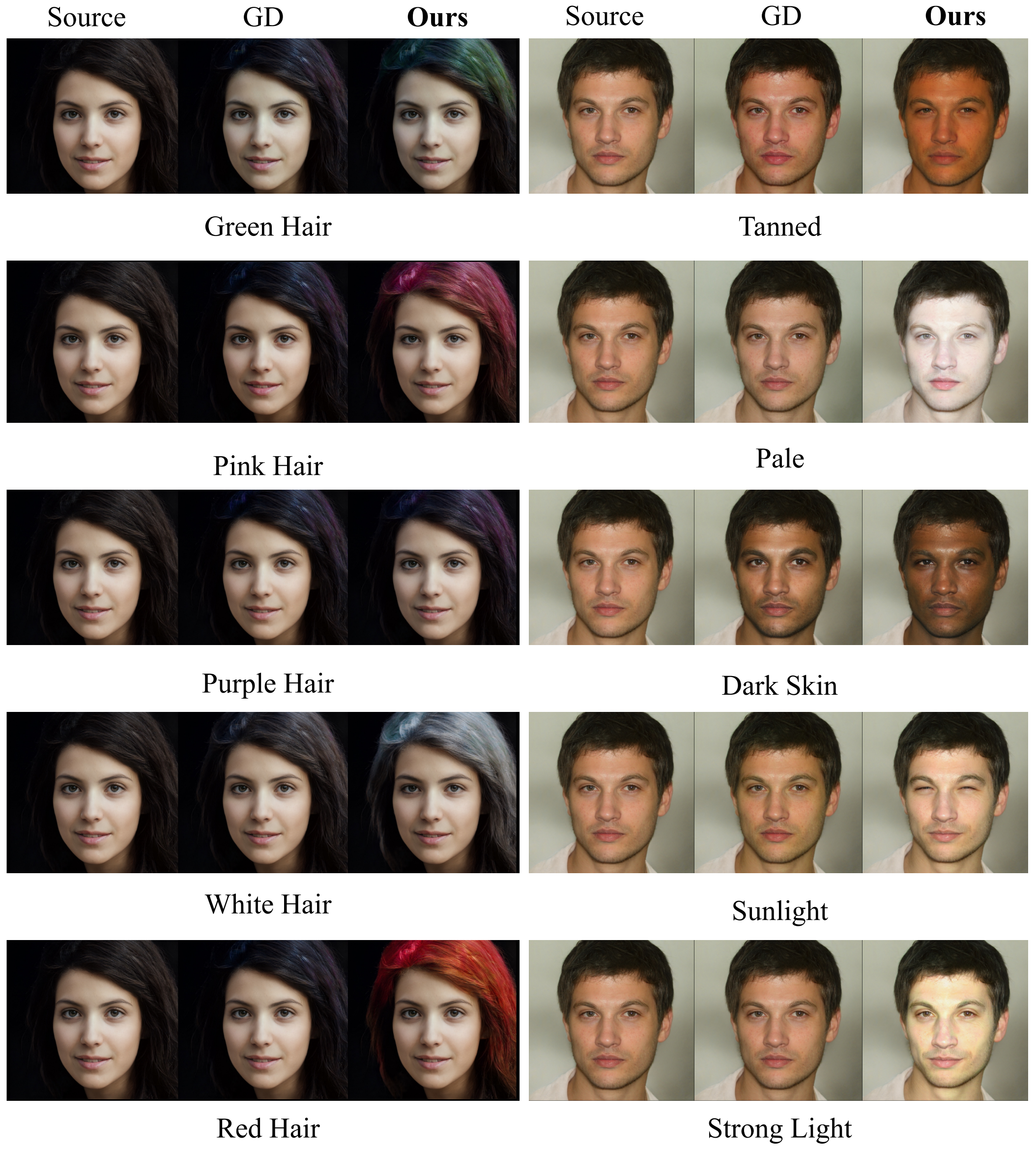


Figure 9: Additional manipulation results on FFHQ-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Multi2One**.

Fig. 10 shows the manipulation results on **Multi2One** 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, **Multi2One** shows consistently reasonable and disentangled images even when using all channels in StyleSpace.

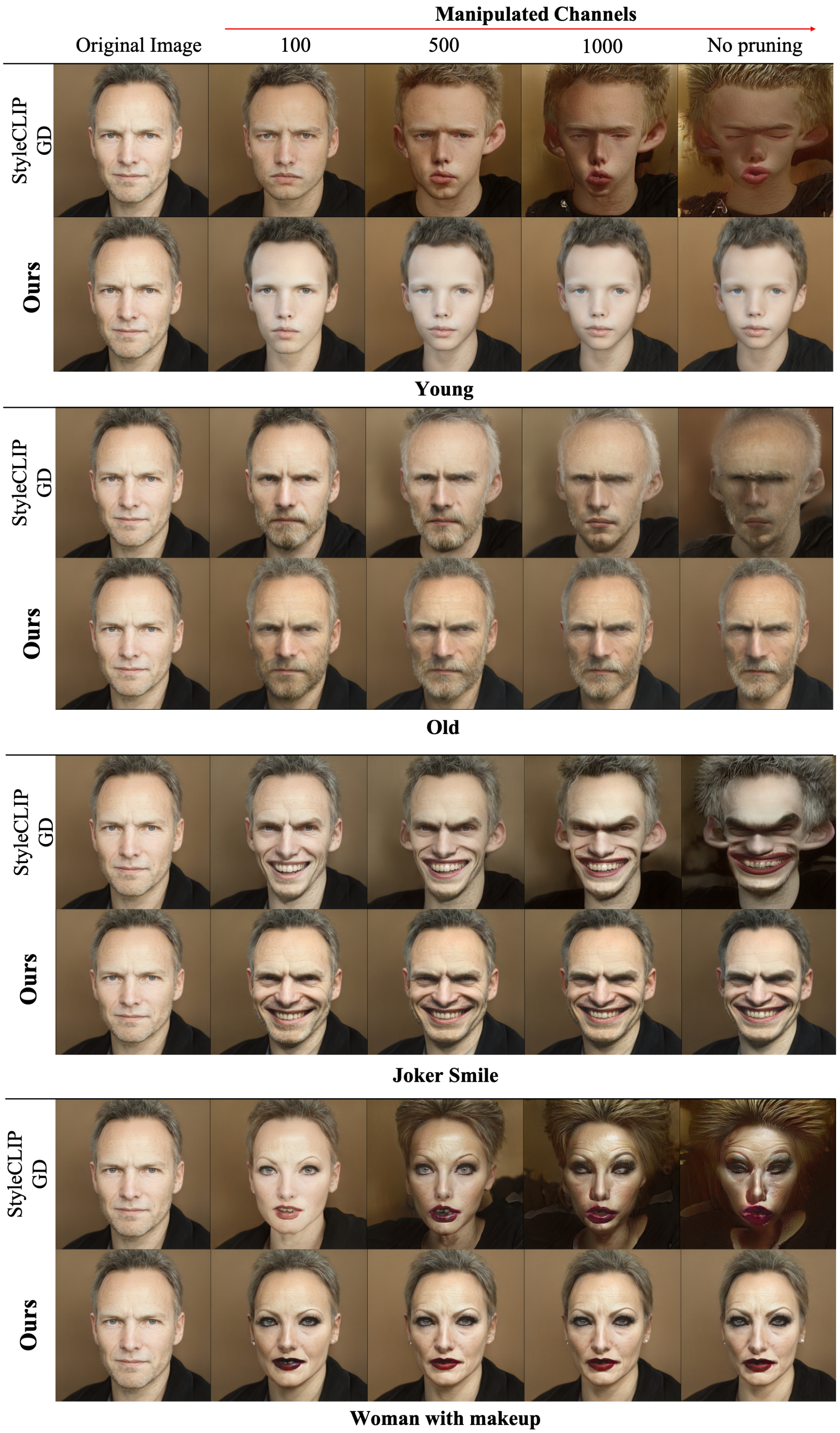


Figure 10: Manipulation results showing the disentanglement of **Multi2One** when using 100, 500, 1000 channels, and without pruning channels. StyleCLIP `GlobalDirection` manifests image collapse when using over 500 channels while **Multi2One** shows robustness against entanglement.

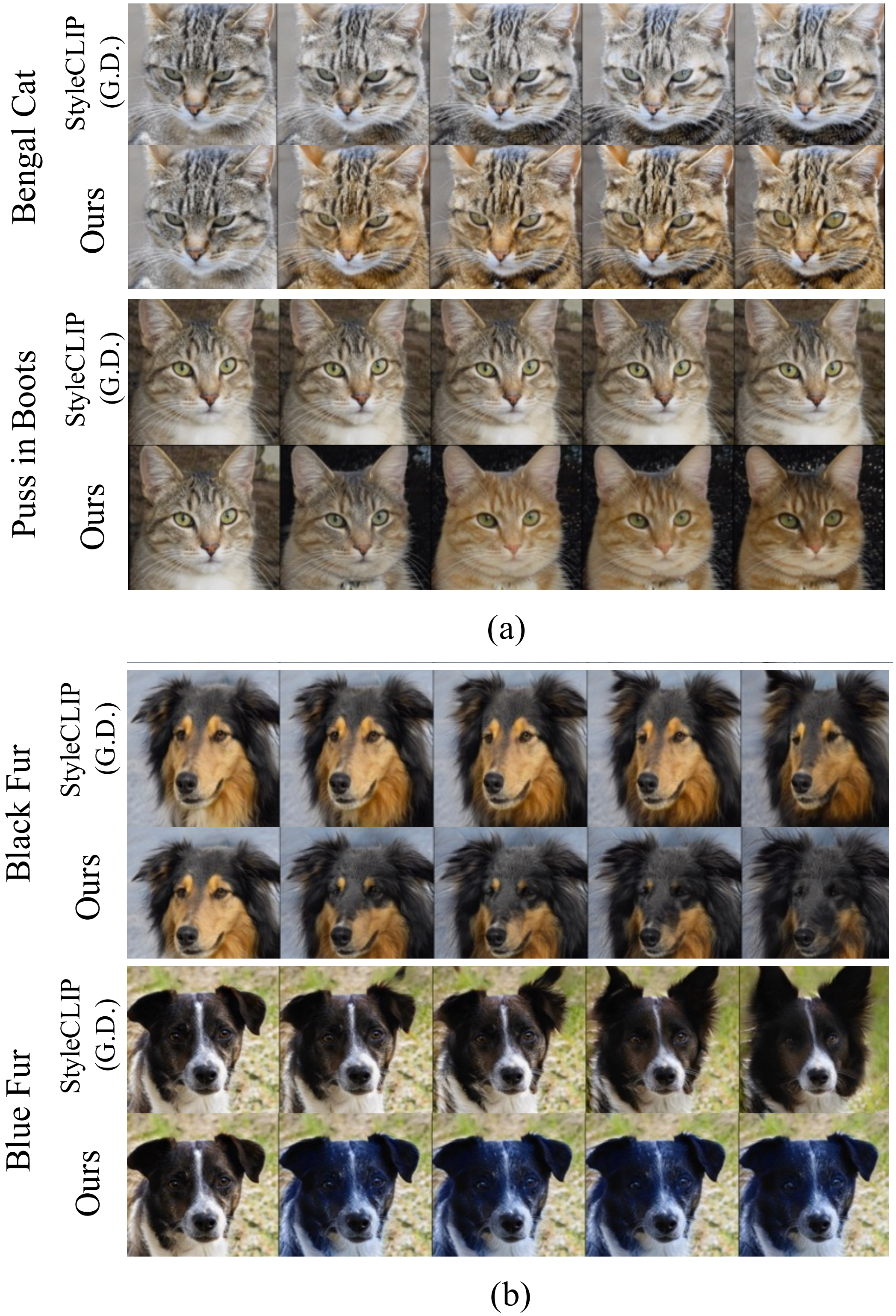
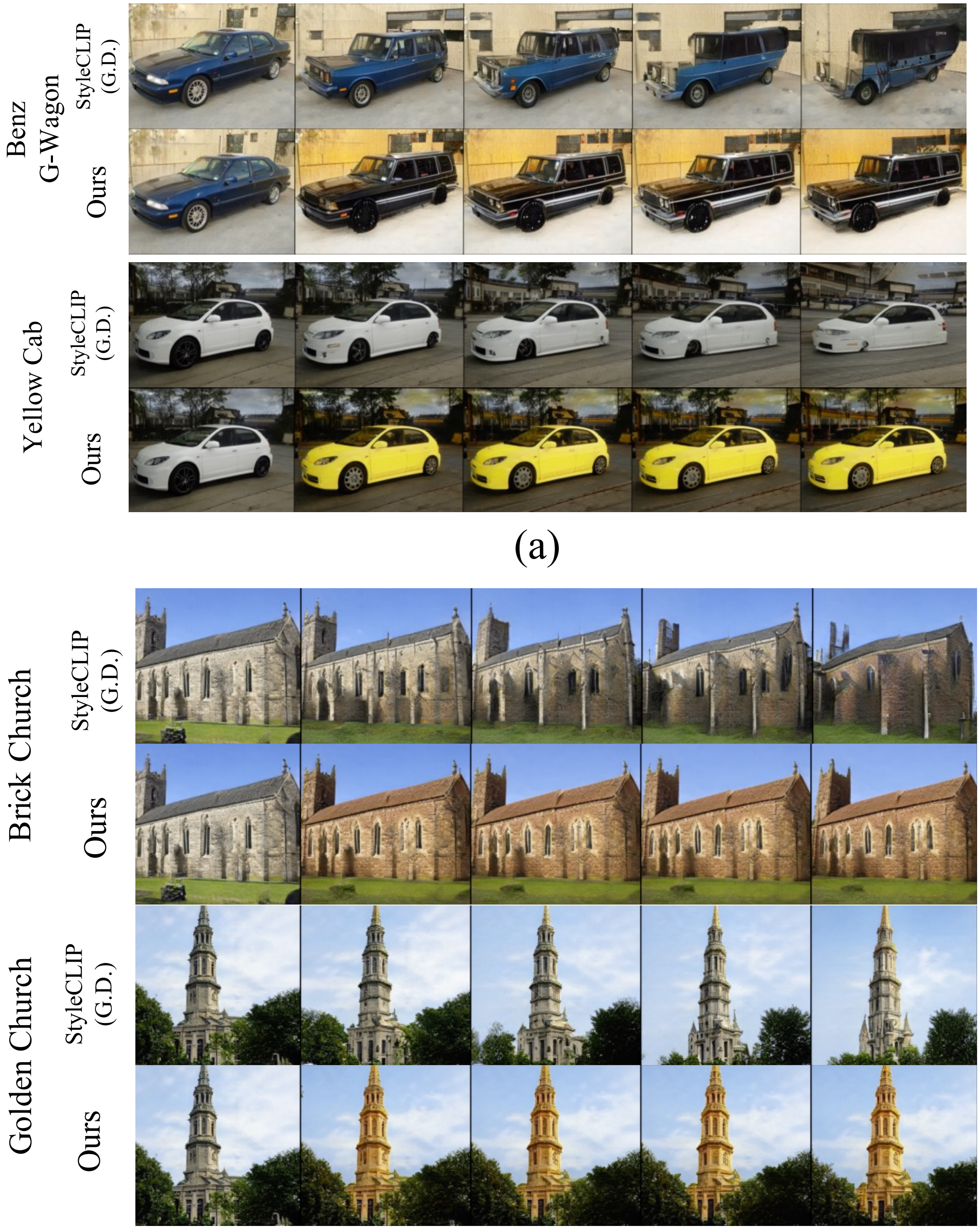


Figure 11: Manipulation results using AFHQ cat and AFHQ dog pretrained StyleGAN2-ADA via text guidance on our method **Multi2One** and StyleCLIP `GlobalDirection`. The samples are manipulated with increasing number of channels that are modified leading to larger changes.



(a)

(b)

Figure 12: Manipulation results using LSUN car and LSUN church pretrained StyleGAN2-ADA via text guidance on our method **Multi2One** and StyleCLIP `GlobalDirection`. The samples are manipulated with increasing number of channels that are modified leading to larger changes.

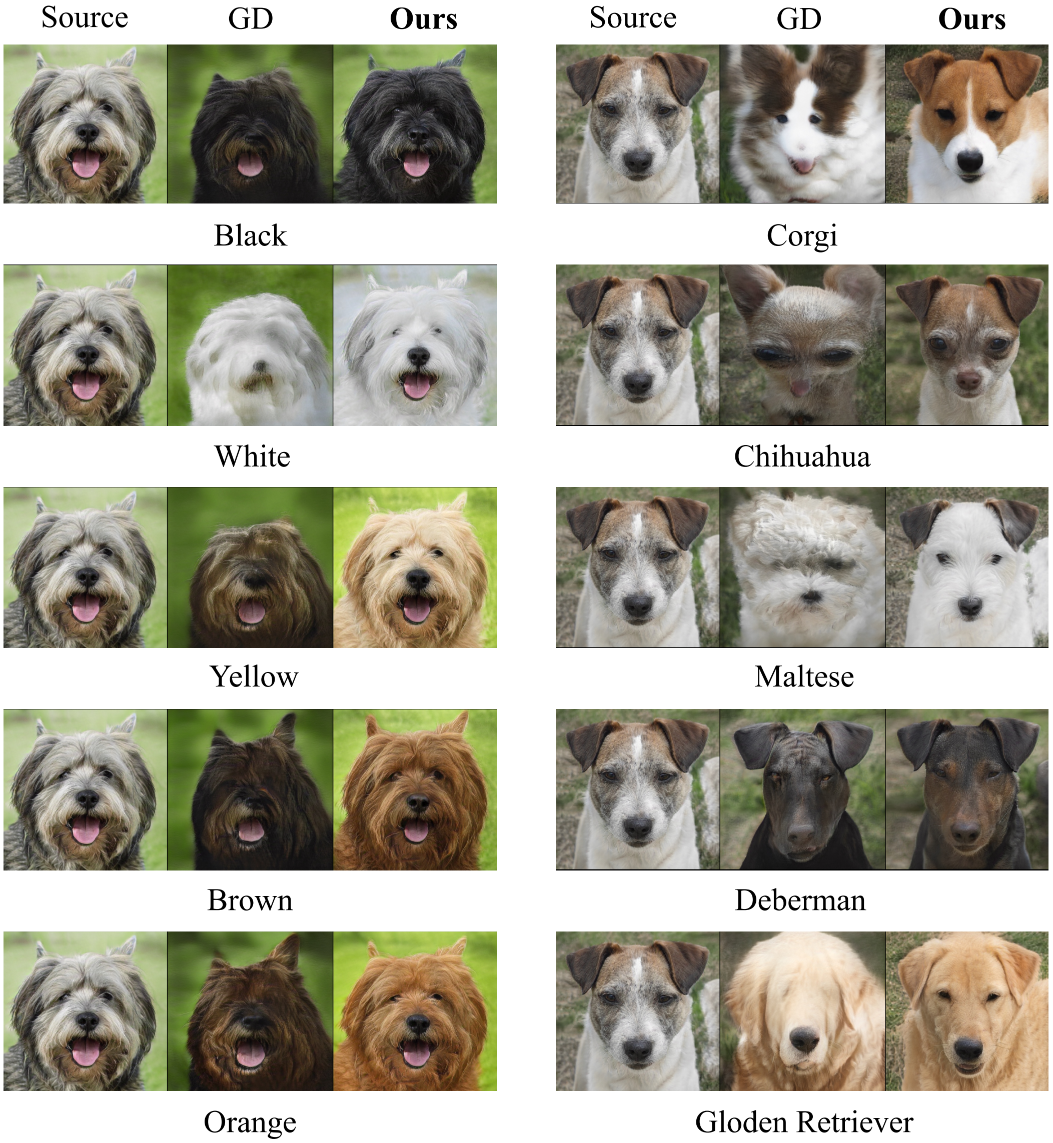


Figure 13: Additional manipulation results on AFHQdog-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Multi2One**.

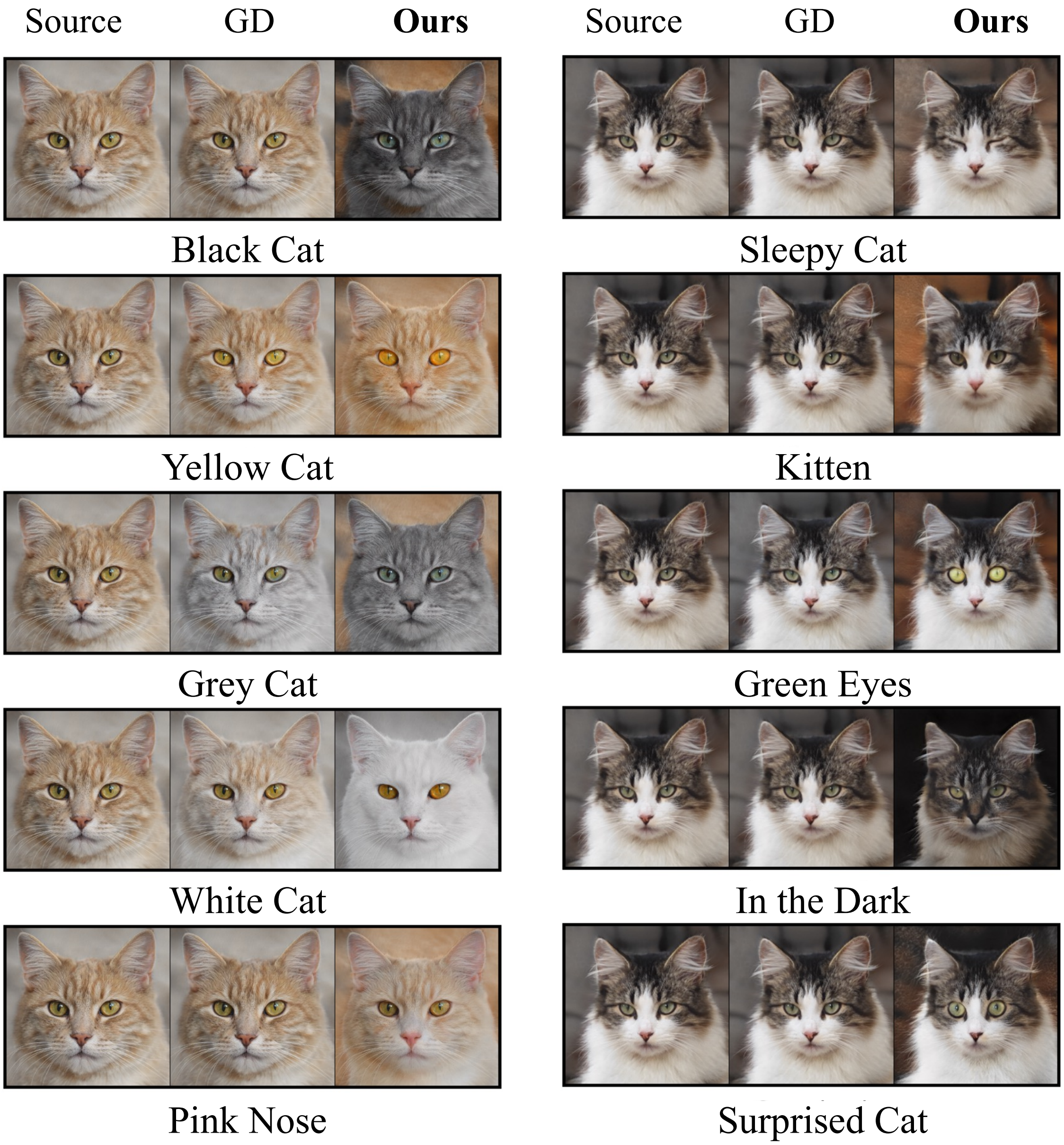
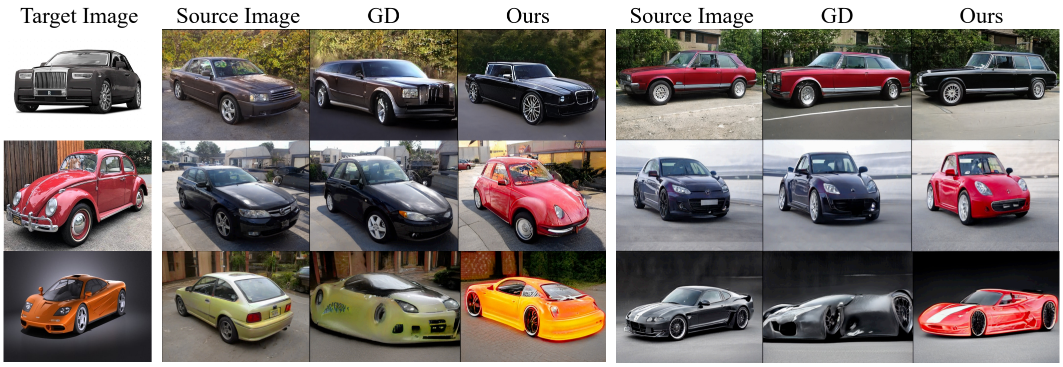


Figure 14: Additional manipulation results on AFHQcat-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Multi2One**.



1

Figure 15: Additional manipulation results on LSUNcar-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD)and **Multi2One**. Used image instead of text as a guidance encoded into CLIP with encoder $E(\cdot)$.

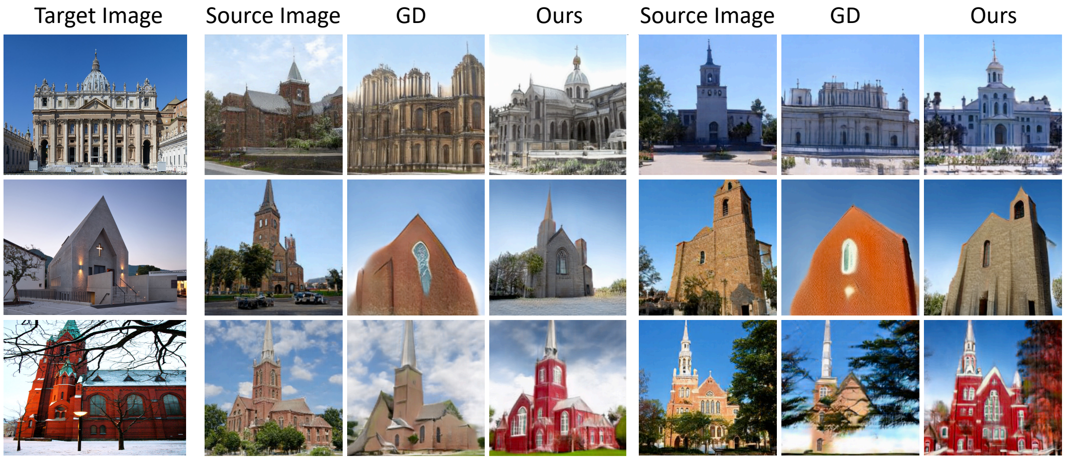


Figure 16: Additional manipulation results on LSUNchuch-pretrained StyleGAN-ADA. Equal configuration for both `GlobalDirection` (denoted as GD) and **Multi2One**. Used image instead of text as a guidance encoded into CLIP with encoder $E(\cdot)$.

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. We use 512 directions from GANSpace and 80 directions from SeFa. 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$\sim$32), medium(resolution 64$\sim$128) and fine(resolution 128$\sim$) (Karras et al., 2020a), which amounts to total of 1776 unsupervised directions.

Since the text-guided manipulation directions are found in StyleSpace $\mathcal{S}$, we map the 1776 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. Sparsifying the known manipulation vectors $\boldsymbol{\alpha}$ into $\boldsymbol{\alpha}^{(k)}$ where $k = 10, 30, 50$ boosts the disentangled property of the learned dictionary.

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 Comparison with Other SoTA methods

Table 4: Evaluation on other text-guided image manipulation methods.

Metrics	Methods			
	Multi2One	StyleCLIP GD	StyleCLIP LO	StyleMC
$\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$ (↑)	$\underline{0.2617}$	0.2482	$\underline{\mathbf{0.2698}}$	0.2574
$\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}})$ (↓)	$\underline{0.0054}$	0.0244	$\underline{\mathbf{0.0047}}$	0.0683
$\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}\|_{=0})$ (↓)	$\underline{\mathbf{0.0012}}$	0.0157	$\underline{0.0018}$	0.0651
$\mathbf{MSE}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}\|_{\neq 0})$ (↓)	$\underline{0.0041}$	0.0086	$\underline{\mathbf{0.0028}}$	0.0032
Inference Time (per image-text)	✗		100 secs	

In Tab. 4, we provide additional results from other image manipulation methods, StyleCLIP GD (`GlobalDirection` from Patashnik et al. (2021)), StyleCLIP LO (local optimization from Patashnik et al. (2021)) and StyleMC from (Kocasari et al., 2022). **Multi2One** and StyleCLIP `GlobalDirection` are two unique methods that provides instant manipulation on provided text and image without additional computation for optimization. Local optimization (Patashnik et al., 2021) and StyleMC (Kocasari et al., 2022) are optimization based methods which require about 100 seconds of inference time using a single NVIDIA 2080Ti. Moreover, such optimization based methods do not have disentanglement property unlike **Multi2One** and `GlobalDirection` leading to significantly worse performance in source identity preservation.

The text-image similarity is measured using $\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$ from Eq. (7), which measures the similarity between text t and the manipulated image. We used total of 57 text descriptions on human face[6] to evaluate the performance based on FFHQ-pretrained generator. We manipulate all 6048 channels in StyleSpace accordingly with optimization based methods which do not prune certain channels as a remedy for entanglement. Despite being an input-agnostic manipulation method, **Multi2One** is on par with the input-dependent StyleCLIP Local Optimization method (Patashnik et al., 2021) in the CLIP similarity score, $\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$

The measurement on the ability to find the unsupervised directions $\boldsymbol{\alpha}^{(100)}$ (only the top 100 channels from unsupervised direction $\boldsymbol{\alpha}$ are modified) is measured using mean squared error in Eq. (6). Given

[6] The construction of 57 random texts for manipulation on FFHQ pre-trained StyleGAN-ada is described in Appendix A

an $\boldsymbol{\phi}_{\mathrm{CLIP}}(\cdot)$ operation which maps the manipulation effect of $\boldsymbol{\alpha}^{(100)}$, it is desirable to find a direction that only modifies the 100 channels that were non-zero entries among n channels in $\boldsymbol{\alpha}^{(100)}$ while the other $n - 100$ channels are not modified. Therefore, $\mathbf{MSE}\big(\boldsymbol{\alpha},\, \hat{\boldsymbol{\alpha}}|_{=0}\big)$ measure the level of disentanglement which measures the distance between the $n - 100$ channels in $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^{(100)}$. The result in Tab. 4 shows that ours, **Multi2One**, scored the lowest $\mathbf{MSE}\big(\boldsymbol{\alpha},\, \hat{\boldsymbol{\alpha}}|_{=0}\big)$ proving to be the most disentangled among the four text-guided manipulation methods.

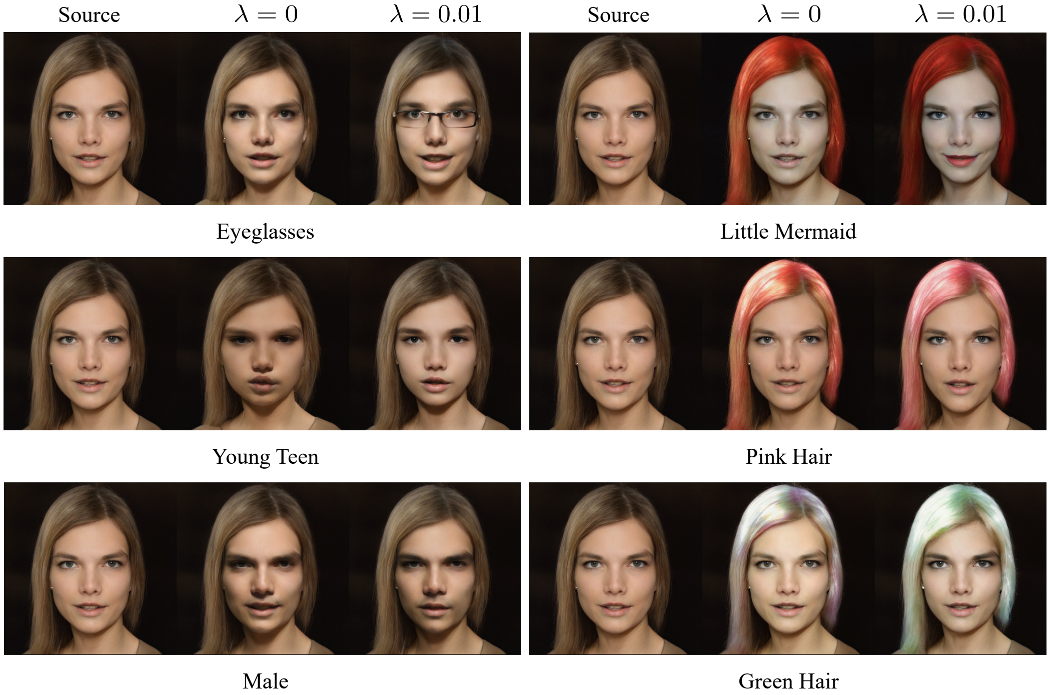
G Ablation Study on Loss Term

We report the similarity score as $\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}}) - \mathbf{R}_{\text{CLIP}}(t, X)$, to show the increase in similarity value when modified using the text guidance t. Note that **ID** loss tends to increase as the image is manipulated properly, therefore it is not a robust metric for evaluating the entanglement level.

Table 5: We report the effect of the hyper-parameter λ and the candidates for top-k for creating the sparsified unsupervised directions. $\mathbf{ID}(.,.)$ indicates ArcFace Loss, which measures a change in identity, between the original image and edited image.

		Similarity		$\mathbf{ID}(X, X_{edit})$	
	λ	0.1	0.01	0.1	0.01
Top-k	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

In Fig. 17, we visualize the manipulation results to show the effect of the second loss term when removed by $\lambda = 0$ compared to when the second loss term is included by $\lambda = 0.01$. We observe that using the additional loss term does help finding novel directions which are not present in the unsupervised directions, leading to better manipulative ability.



Eyeglasses

Little Mermaid

Young Teen

Pink Hair

Male

Green Hair

Figure 17: Manipulation results with ablation of the second loss term using $\lambda = 0$ and $\lambda = 0.01$.

H Ablation Study on the effect of Unsupervised Methods

The dictionary learning process of **Multi2One** employs the directions $\boldsymbol{\alpha} \in \mathbb{R}^n$ from unsupervised methods (Shen & Zhou, 2021; Härkönen et al., 2020). Therefore, we conduct an ablation study on the effect of using unsupervised directions by comparing the two cases where directions $\boldsymbol{\alpha}$ come from supervised method (Shen et al., 2020) and unsupervised methods (Shen & Zhou, 2021; Härkönen et al., 2020). Fig. 18 shows that using directions from supervised methods are less capable of manipulating images to have desired change. This could also be observed by quantitative measure $\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$ in Tab. 6 since the similarity score between manipulated image and the text guidance is significantly lower using directions from supervised method compared to using unsupervised method. This is due to the limited number of directions that could be found by supervised method, which amounts to only 14 using InterFaceGAN (Shen et al., 2020).

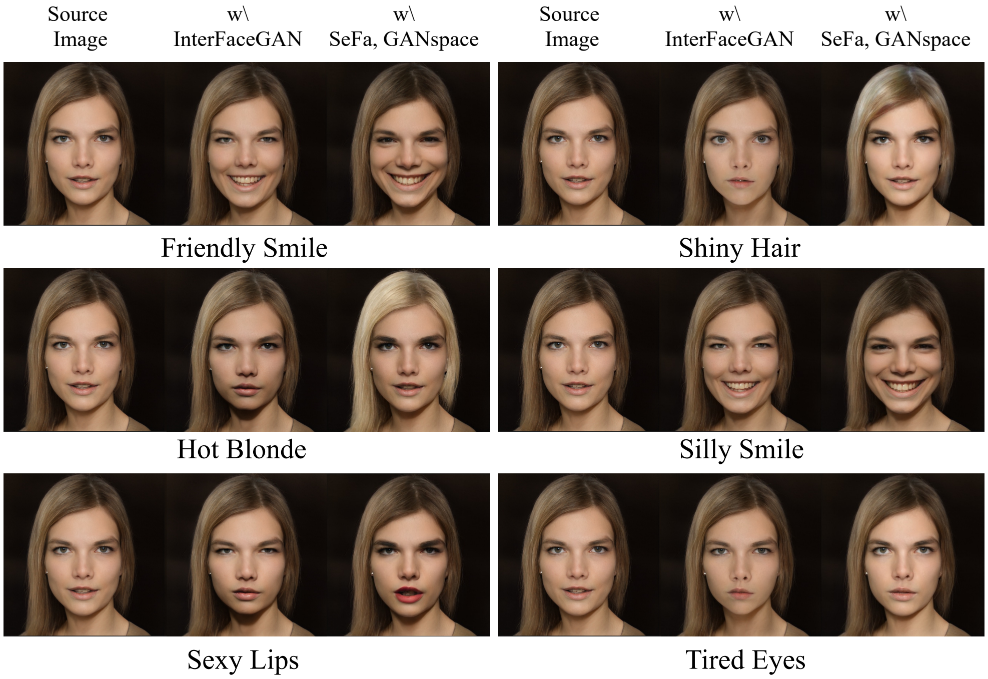


Figure 18: Ablation study on the effect of using unsupervised directions from Härkönen et al. (2020); Shen & Zhou (2021). The substitute the ground truth directions in the dictionary learning process with directions from InterFaceGAN (Shen et al., 2020). We manipulate same number of channels with equal manipulation strength for both.

Table 6: Evaluation on other text-guided image manipulation methods. 100 channels manipulated with manipulation strength of 5. Supervised method indicates using directions from Shen et al. (2020) as $direction$. Unsupervised method indicates that dictionary learning process uses ground truth directions $\boldsymbol{\alpha}$ from Härkönen et al. (2020); Shen & Zhou (2021).

$\boldsymbol{\alpha}$	Supervised Method	Unsupervised Method
$\mathbf{R}_{\text{CLIP}}(t, X_{\text{edited}})$	0.228	**0.254**

I Computational Overhead

StyleCLIP `GlobalDirection` and **Multi2One** require preprocessing time to compute $\phi_{\text{CLIP}}(\cdot)$ operation. Such operation is computed using either $\boldsymbol{e}_i$ over n channels (StyleCLIP) or $\boldsymbol{\alpha}$ for total of q directions. More specifically, StyleGAN-ADA generator has a fixed StyleSpace dimension consisting of n channels which varies by the resolution of output image. 1024 resolution has 6048 channels, while 512 resolution has 5952 and 256 has 5760. Therefore the total computational time is measured by the number of $\phi_{\text{CLIP}}(\cdot)$ operations calculated over n or q directions using 100 instances of source images to compute the expectation. In Tab. 7 we measure the average computation seconds per iteration using NVIDIA 2080Ti which is equal for both methods. Since one iteration means a single $\phi_{\text{CLIP}}(\cdot)$ operation, comparing the number of iterations shows that our method requires less computation time. Moreover, our method introduces an additional dictionary learning phase which only takes about 15 minutes. Therefore we conclude that our method is slightly more efficient in terms of time to construct the dictionary D.

Table 7: Measured time required for a forward pass with batch size 1 in NVIDIA 2080Ti-seconds. SG2 stands for StyleGAN-ADA. The first row show the duration of single iteration in seconds.

	Preprocessing Time			Dictionary Learning
Dataset	SG2-**res1024$^{\mathbf{2}}$**	SG2-**res512$^{\mathbf{2}}$**	SG2-**res256$^{\mathbf{2}}$**	All **resolution**
Sec/it	14.98s/it	13.11s/it	8.19s/it	
GD (iters)	6048	5952	5760	✗
Ours (iters)	**5328**	**5328**	**5328**	15 mins

J Limitations

The flexibility and diversity of text is not fully exerted due to the limited encoding ability and the deterministic representation of CLIP (Radford et al., 2021). For example, it is possible to observe change in pose when using the eigenvectors (vectors that correspond to large eigenvalues often shows large change in image) from unsupervised methods. However, since CLIP is known to have limited ability in terms of understanding the relative position, it is impossible to modify an image using text guidance such as ‘move right’ and ‘turn left’. Moreover, we observe that CLIP encoder fails to represent some of the text instance such as ’Smiling eyes’ (Fig. 7) while it is shown in Fig. 1(a) that change in eyes happens using text ’Smile’. Since it is proven that the editing direction of ’smiling eyes’ could be discovered when provided with appropriate text, we conjecture that the CLIP encoder fails to represent some text input faithfully leading to unsuccessful manipulation results.