Grasping the Arrow of Time from the Singularity: Decoding Micromotion in Low-dimensional Latent Spaces from StyleGAN

Anonymous authors
Paper under double-blind review

Abstract

The disentanglement of StyleGAN latent space has paved the way for realistic and controllable image editing, but does StyleGAN know anything about temporal motion, as it was only trained on static images? To study the motion features in the latent space of StyleGAN, in this paper, we hypothesize and demonstrate that a series of meaningful, natural, and versatile small, local movements (referred to as "micromotion", such as expression, head movement, and aging effect) can be represented in low-rank spaces extracted from the latent space of a conventionally pre-trained StyleGAN-v2 model for face generation, with the guidance of proper "anchors" in the form of either short text or video clips. Starting from one target face image, with the editing direction decoded from the low-rank space, its micromotion features can be represented as simple as an affine transformation over its latent feature. Perhaps more surprisingly, such micromotion subspace, even learned from just single target face, can be painlessly transferred to other unseen face images, even those from vastly different domains (such as oil painting, cartoon, and sculpture faces). It demonstrates that the local feature geometry corresponding to one type of micromotion is aligned across different face subjects, and hence that StyleGAN-v2 is indeed "secretly" aware of the subject-disentangled feature variations caused by that micromotion. We present various successful examples of applying our low-dimensional micromotion subspace technique to directly and effortlessly manipulate faces. Compared with previous methods, our framework shows high robustness, low computational overhead, and impressive domain transferability. Our codes will be released upon acceptance.

1 Introduction

In recent years, the StyleGAN and its variants (Karras et al., 2017; 2021; 2019; 2020; Sauer et al., 2022) have achieved state-of-the-art performance in controllable image synthesis. These high qualities and fine-grained controls of the synthesized images are largely associated with the expressive latent space of StyleGAN. Previous researches have revealed that the learned latent space of StyleGAN can be smooth and interpretable (Abdal et al., 2019; 2020; Wu et al., 2021; Zhu et al., 2020). Furthermore, previous studies (Karras et al., 2019; 2020) have shown that by feature manipulations and interpolations in the latent space, the style-based GANs can generate a variety of intriguing images with desired changes. These findings have led to many downstream applications such as face manipulation (Wei et al., 2021; Alaluf et al., 2021a), style transfer (Abdal et al., 2019; Kwon & Ye, 2021), general image editing (Gu et al., 2020; Park et al., 2020; Suzuki et al., 2018), and even video generation (Chu et al., 2020; Fox et al., 2021; Skorokhodov et al., 2021; Zhang & Pollett, 2021).

Given this phenomenal result, many try to further understand the potential in the latent space of StyleGAN. Particularly, rather than per-image editing methods, people wonder whether it is possible to directly locate a series of latent codes that correspond to sample-agnostic semantically meaningful attributes (*e.g.* smiling, aging on human faces). These attempts to disentangle meaningful editing directions can be roughly categorized into supervised methods and unsupervised methods. The supervised approaches (Shen et al., 2020b;

Figure 1: **Representative examples created by the proposed method.** The original images are edited using a simple linear scaling with the discovered universal editing directions on various transformations. These three rows correspond to *eye-opening, aging, and head rotation*.

Yang et al., 2021; Goetschalckx et al., 2019) typically sample a series of latent codes, labeling the latent codes with pretrained attributes predictors, and finally learning classifiers for each desired attribute in the latent space. On the other hand, the unsupervised approach (Shen & Zhou, 2021; Härkönen et al., 2020) explores the principal components of the sampled latent codes and observes if these codes correspond to semantically meaningful editing directions. However, as will be shown later, the editing directions founded by these methods are shown to be still entangled with other attributes. When applying these discovered editing directions, the result images suffer from undesired changes in identity and other attributes. This leads to the following question: whether this sub-optimal entanglement is due to the intrinsic limits of the entangled latent space, or it is because previous methods do not fully reveal the potentials of the StyleGAN?

To answer the question, in this paper, we propose in-depth investigations on the StyleGAN-v2's latent space trained on face generations. In particular, we hypothesize that from the StyleGAN's high dimensional latent space, a low-rank feature space can be extracted where universal editing directions can be reconstructed for various facial style transformations including changes in expressions/emotions, heads movements, and aging effects, which we refer to as a series of **micromotion**. Thanks to the highly disentangled essence of the decoded editing directions, for any given input, linear scaling along the same found direction will make the image change its style smoothly. Furthermore, to find such a directional vector we leverage the guidance of proper "anchors" in the form of either short texts or a reference video clip and show the directional vector can be efficiently found via simple subtractions using a robustly learned linear subspace projection. Surprisingly, such latent subspace can be extracted using only a single query image, and then the resulting editing direction can be used for any unseen face image, even for those from vastly different domains including oil painting, cartoon, sculpture, *etc*. Figure 1 shows the generated images for multiple style transformations and face types. The contributions of our paper are three-fold:

- Leveraging the low-dimensional feature space hypothesis, we demonstrate the properties of StyleGAN's latent space from a global and universal viewpoint, using "micromotions" as the subject.
- We demonstrate that by using text/video-based anchors, low-dimensional micromotion subspace along with universal and highly disentangled editing directions can be consistently discovered using the same robust subspace projection technique for a large range of micromotion-style facial transformations.

- We show the editing direction can be found using a single query face input and then directly applied to other faces, even from vastly different domains (*e.g.*, oil painting, cartoon, and sculpture faces), in an easily controllable way as simple as linear scaling along the discovered subspace.

2 Related Works

2.1 StyleGAN: Models and Characteristics

The StyleGAN (Karras et al., 2021; 2019; 2020) is a style-based generator architecture targeting on image synthesis task. Leveraging a mapping network and affine transformation to render abstract style information, StyleGAN is able to control the image synthesis in a scale-specific fashion. Particularly, by augmenting the learned feature space and hierarchically feeding latent codes at each layer of the generator architecture, the StyleGAN has demonstrated surprising image synthesis performance with controls from coarse properties to fine-grained characteristics (Karras et al., 2019). Also, when trained on a high-resolution facial dataset (*e.g.,* FFHQ (Karras et al., 2019)), the StyleGAN is able to generate high-quality human faces with good fidelity.

2.2 StyleGAN-based Editing

Leveraging the expressive latent space by StyleGAN, recent studies consider interpolating and mixing the latent style codes to achieve specific attribute editing without impairing other attributes (e.g. person identity). (Hou et al., 2022; Shen et al., 2020a; Tewari et al., 2020a;b; Wu et al., 2021) focus on searching latent space to find latent codes corresponding to global meaningful manipulations, while (Chong et al., 2021) utilizes semantic segmentation maps to locate and mix certain positions of style codes to achieve editing goals.

To achieve zero-shot and open-vocabulary editing, recent works set their sights on using pretrained multi-modality models as guidance. With the aligned image-text representation learned by CLIP, a few works (Wei et al., 2021; Patashnik et al., 2021) use text to extract the latent edit directions with textual defined semantic meanings for separate input images. These works focus on extracting latent directions using contrastive CLIP loss to conduct image manipulation tasks such as face editing (Patashnik et al., 2021; Wei et al., 2021), cars editing (Abdal et al., 2021). Besides, a few recent works manipulate the images with visual guidance (Lewis et al., 2021; Kim et al., 2021). In these works, image editing is done by inverting the referential images into corresponding latent codes, and interpolating the latent codes to generate mixed-style images. However, these works focus on per-example image editing. In other words, for each individual image input, they have to compute corresponding manipulations in the latent space separately. With the help of disentangled latent space, it is interesting to ask whether we can decode universal latent manipulations and conduct sample-agnostic feature transformation.

2.3 Feature Disentanglement in Latent Space of StyleGAN

Feature disentanglement in StyleGAN latent space refers to decomposing latent vector components corresponding to interpretable attributes. Previous studies on StyleGAN latent space disentanglement can be roughly categorized into supervised and unsupervised methods. In supervised methods (Shen et al., 2020b; Yang et al., 2021; Goetschalckx et al., 2019), they typically leverage auxiliary classifiers or assessors to find the editing directions. To be more specific, Shen et al. (2020b) first sample a series of latent codes from the latent space and render corresponding images. Then, they train an SVM to learn the mapping between sampled latent codes and corresponding attributes, where the labeled attributes are supervised by the auxiliary classifiers. Finally, the normal direction of the hyperplane is the founded editing direction. Besides, Goetschalckx et al. (2019) directly optimize the editing direction based on an auxiliary classifier. On the other hand, the unsupervised methods typically explore the principal components of sampled latent codes, while they manually check if these components correspond to semantically meaningful attributes. However, as will be shown later, the editing directions founded by these methods are shown to be still entangled with other attributes. In this work, leveraging a stronger low-rank latent space hypothesis, we find highly-disentangled latent codes and show that sample-agnostic editing direction can be consistently found in StyleGAN's latent space.

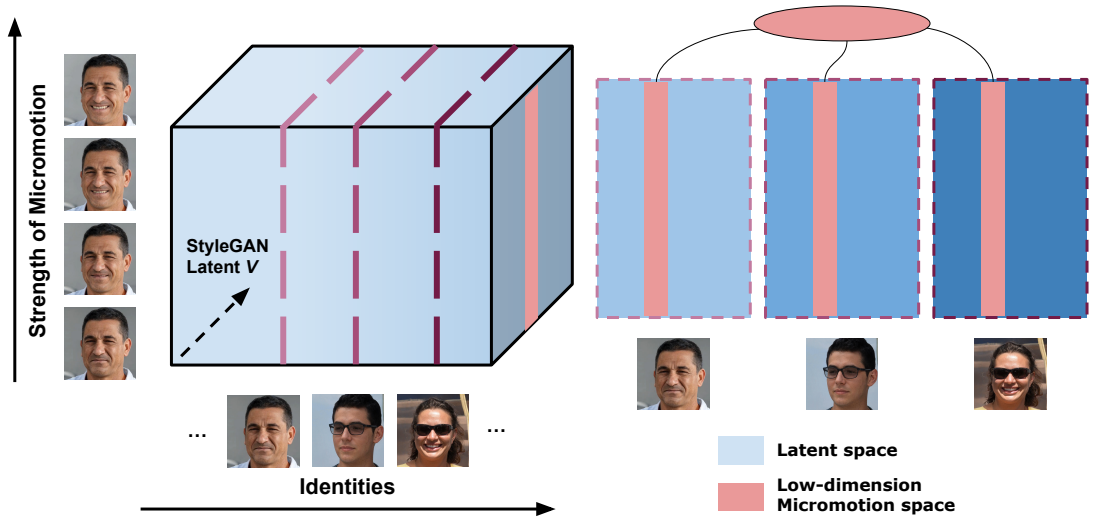


Figure 2: **A tensor illustration of our hypothesis.** In the StyleGAN latent space, we hypothesize the same type of micromotion, *at different quantitative levels but for the same identity*, can be approximated by a low-rank subspace. We further hypothesize that subspaces *for the same type of micromotion found at different identities* are extremely similar to each other, and can hence be transferred across identities.

3 Method

In this section, we first present the problem of decoding micromotion in a pre-trained StyleGAN latent space, and we define the notations involved in this paper. We then articulate the low-rank micromotion subspace hypothesis in Sec. 3.2, proposing that the locally low-dimensional geometry corresponding to one type of micromotion is consistently aligned across different face subjects, which serves as the key to decode universal micromotion from even a single identity. Finally, based on the hypothesis, we demonstrate a simple workflow to decode micromotions and seamlessly apply them to various in-domain and out-domain identities, incurring clear desired facial micromotions.

3.1 Problem Setting

Micromotions are reflected as smooth transitions in continuous video frames. In a general facial-style micromotion synthesis problem, given an arbitrary input image I_0 and a desired micromotion (*e.g.* smile), the goal is to design an identity-agnostic workflow to synthesize temporal frames $\{I_1, I_2, \ldots, I_t\}$, which constitute a consecutive video with the desired micromotion.

Synthesizing images with StyleGAN requires finding proper latent codes in its feature space. We use G and E to denote the pre-trained StyleGAN synthesis network and StyleGAN encoder respectively. Given a latent code $\mathbf{V} \in \mathcal{W}^+$, the pre-trained generator G maps it to the image space by $I = G(\mathbf{V})$. Inversely, the encoder maps the image I back to the latent space $\mathcal{W}^+$, or $\hat{\mathbf{V}} = E(I)$. Leveraging the StyleGAN latent space, finding consecutive video frames turns out to be a task of finding a series of latent codes $\{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_t\}$ corresponding to the micromotion.

3.2 Key Hypothesis: The Low-rank Micromotion Subspace

To generate semantically meaningful and correct micromotions using StyleGAN, the key objective is to find proper latent code series in its feature space. We hypothesize that those latent codes can be decoded by a low-rank micromotion subspace. Specifically, we articulate the key hypothesis in this work, stated as: *The versatile facial style micromotions can be represented as low-rank subspaces within the StyleGAN latent space, and such subspaces are subject-agnostic.*

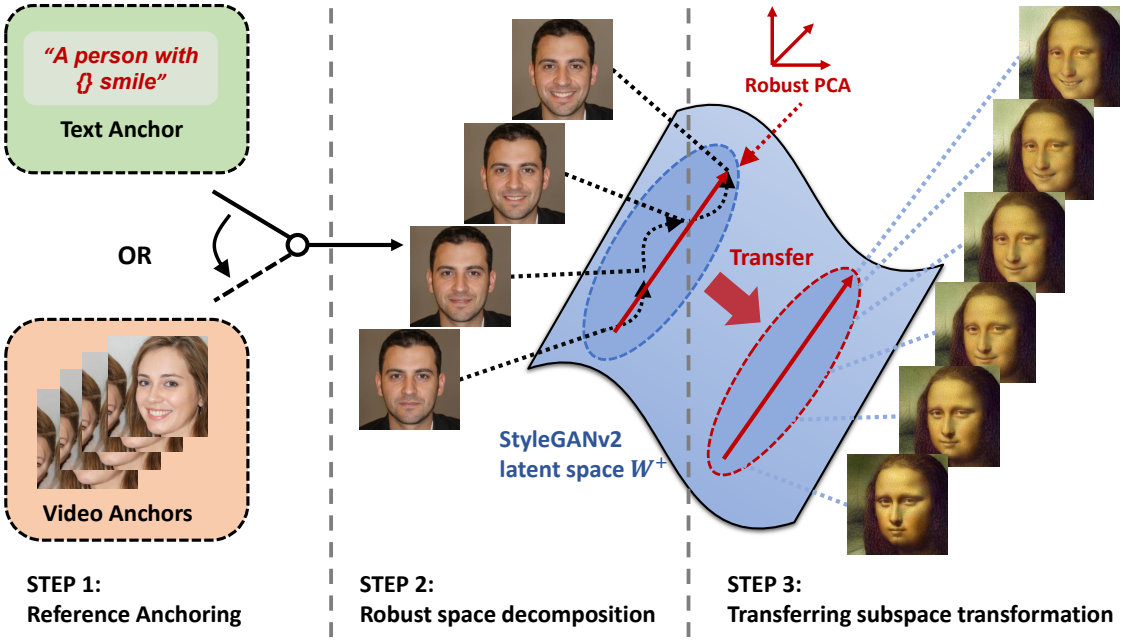


Figure 3: **Our workflow illustration.** In our workflow, we first extract a low-dimensional micromotion subspace from one identity, and then transfer it to a novel identity "Mona-lisa".

To give a concrete illustration of the hypothesis, we plot a tensor-view illustration of a micromotion subspace, smile, in Figure 2. The horizontal axis encodes the different face identities, and each perpendicular slice of the vertical plane represents all variations embedded in the StyleGAN latent space for a specific identity. We use the vertical axis to indicate the quantitative strength for a micromotion (*e.g.*, smile from mild to wild). Given a sampled set of images in which a subject face changes from the beginning (e.g., neutral) to the terminal state of a micromotion, each image can be synthesized using a latent code $\mathbf{V}$. Aligning these latent codes for one single subject formulates a *micromotion matrix* with dimension $V \times M$, where V is the dimension of the latent codes and M is the total number of images. Eventually, different subjects could all formulate their micromotion matrices in the same way, yielding a *micromotion tensor*, with dimension $P \times V \times M$ assuming a total of P identities. Our hypothesis is then stated in two folds:

- Each subject's micromotion matrix can be approximated by a "micromotion subspace" and it is inherently low-rank. The micromotion "strengths" can be reduced to linearly scaling along the subspace.
- The micromotion subspaces found at different subjects are substantially similar and even mutually transferable. In other words, different subjects (approximately) share the common micromotion subspace. That implies the existence of universal edit direction for specific micromotions regardless of identities.

If the hypothesis can be proven true, it would be immediately appealing for sample-agnostic image manipulations. First, micromotion can be represented in low-dimensional disentangled spaces, and the dynamic edit direction can be reconstructed once the space is anchored. Second, when the low-dimensional space is found, it can immediately be applied to multiple other identities with extremely low overhead, and is highly controllable through interpolation and extrapolation by as simple as linear scaling.

3.3 Our Workflow

With this hypothesis, we design a workflow to extract the edit direction from decomposed low-dimensional micromotion subspace, illustrated in Figure 3. Our complete workflow can be distilled down to three simple steps: (a) collecting anchor latent codes from a single identity; (b) enforcing robustness linear decomposition to obtain a noise-free low-dimensional space; (c) applying the extracted edit direction from low-dimensional space to arbitrary input identities.

Step 1: Reference Anchoring. To find the edit direction of a specific micromotion, we first acquire a set of latent codes corresponding to the desired action performed by the same person. Serving as anchors, these latent codes help to disentangle desired micromotions in later steps. Here, we consider two approaches, text-anchored and video-anchored methods, respectively.

Text-anchored Reference Generation: The StyleCLIP (Patashnik et al., 2021) has shown that expressive phrases can successfully manipulate attributes in synthesized images. We leverage the its latent optimization pipeline to generate the anchoring latent codes for desired micromotions. The main-idea to optimize these latent codes is to minimize the contrastive loss between the designed input texts and the images rendered by the codes with a few regularizations. Here, one major question is how to design the most appropriate text template to guide the optimization. To generate images with only variance in degrees of micromotions, a natural method is to specify the degrees in the text, where we concatenate a series of adjectives or percentages with the micromotion description text to indicate the various strength and the stage of the current micromotion. For example, for the micromotion "eyes closed", we use both percentages and adjectives to modify the micromotion by specifying "eyes *greatly/slightly* closed" and "eyes *10%/30%* closed". Here, we emphasize that this is just one of the possible text prompts design options. We compare various choices of text prompts, and the experiments of the text prompt choices will be covered in the ablation study.

Video-anchored Reference Generation: StyleCLIP relies on text guidance to optimize the latent codes, while for abstract and complicated motions, such as a gradual head movement with various head postures, the text might not be able to express the micromotion concisely. To overcome this issue, we leverage a reference video demonstration to anchor the micromotion subspace instead. In the reference video-based anchoring methods, we use frames of reference videos to decode the desired micromotions. Specifically, given a reference video that consists of continuous frames, we invert these frames with a pre-trained StyleGAN encoder to obtain the reference latent codes. We emphasize that different from the per-frame editing method, the goal of using reference video frames is to anchor the low-dimensional micromotion subspace. Thus, we use much fewer frames than per-frame editing methods, and no further video frames are used once we extract the space.

After applying either anchoring method, we obtain a set of t_n referential latent codes denoted as $\{\mathbf{V}_{t1}, \mathbf{V}_{t2}, \ldots, \mathbf{V}_{tn}\}$. We will use these codes to obtain a low-rank micromotion space in later steps.

Step 2: Robust space decomposition. Due to the randomness of the optimization and the complicacy of image contents (e.g., background distractors), the latent codes from the previous step could be viewed as "noisy samples" from the underlying low-dimensional space. Therefore, based on our low-rank hypothesis, we leverage further decomposition methods to robustify the latent codes and their shared micromotion subspace.

The first simple decomposition method we adopt is the principal component analysis (PCA), where each anchoring latent code serves as the row vector of the data matrix. Unfortunately, merely using PCA is insufficient for a noise-free micromotion subspace, since the outliers in latent codes degrade the quality of the extracted space. As such, we further turn to a classical technique called *robust PCA* (Wright et al., 2009), which can recover the underlining low-rank space from the latent codes with sparse gross corruptions. It can be formulated as a convex minimization of a nuclear norm plus an ℓ_1 norm and solved efficiently with alternating directions optimization (Candès et al., 2011). Through the principal component of the subspace, we get a robust micromotion edit direction $\Delta\mathbf{V}$.

Step 3: Applying the subspace transformation. Once the edit direction is obtained, we could edit any arbitrary input faces for the micromotion. Specifically, the editing is conducted simply through interpolation and extrapolation along this latent direction to obtain the intermediate frames. For an arbitrary input image I'_0, we find its latent code $\mathbf{V}'_0 = E(I'_0)$, and the videos can be synthesized through

$$I_t = G(\mathbf{V}_t) = G(\mathbf{V}_0 + \alpha t \Delta\mathbf{V}), \tag{1}$$

where α is a parameter controlling the degree of interpolation and extrapolation, t corresponds to the index of the frame, and the resulting set of frames $\{I_t\}$ collectively construct the desired micromotion such as "smiling", "eyes opening". Combining these synthesized frames, we obtain a complete video corresponding to the desired micromotion.

Figure 4: **Illustrations of versatile micromotions founded by text-anchored method.** We decode the micromotions across different identities, and apply them to in-domain identities. From Top to Bottom: (a) Smiling (b) Anger (c) Opening Eyes (d) Aging Face. Best view when zoomed in. Please refer to supplementary for complete video sequences.

4 Experiments

In the experiments, we focus on the following questions related to our hypothesis and workflow:

- Could we locate subspaces for various meaningful and highly disentangled micromotions? (Sec. 4.2)
- Could we transfer decoded micromotion to other subjects in various domains? (Sec. 4.3)
- Could we extend the micromotions to novel subjects with no computation overhead? (Sec. 4.3)
- Is this framework robust towards various choices of prompts and identities? (Sec. 4.4)

In short, we want to prove two concepts in following experiments: (a) **Universality**: Our pipeline can consistently find various micromotion, and the decoded micromotion can be extended to different subjects across domains; (b) **Lightweight**: Transferring the micromotion only requires a small computation overhead.

To validate these concepts, we now analyze our framework by synthesizing micromotions. We mainly consider five micromotions as examples: (a) smiling, (b) angry, (c) opening eyes, (d) turning head, and (e) aging face. We also consider more editings when comparing with other methods. Following the workflow, we obtain the edit directions for each micromotion, and then synthesize on other cross-domain images including characters in animations, sculptures, and paintings.

Figure 5: **Illustrations of the micromotion "turning head" founded by our video-anchored method.** Best view when zoomed in. Please refer to supplementary for complete video sequences.

4.1 Experiment Settings

The pre-trained models (StyleGAN-v2, StyleCLIP, and StyleGAN encoders) are all loaded from the public repositories (Abdal et al., 2020; Alaluf et al., 2021b; Patashnik et al., 2021; Radford et al., 2021). When optimizing the latent codes, the learning rate is 0.1 and we use Adam optimizer. For the text-anchored and video-anchored methods, the numbers of latent codes we generate are 16 and 7. In robust PCA, 4 principal dimensions are chosen. We also search the extrapolation scale hyperparameter α between 0.1 and 10.

For the text-anchored experiments, the original images are generated using random latent codes in StyleGAN-v2 feature space. The text prompts we construct is in the general form of (a) "A person with {} smile"; (b) "A person with {} angry face"; (c) "A person with eyes {} closed"; (d) "{} old person with gray hair", which correspond to the micromotions of smiling, angry, eyes opening and face aging. Here, the wildcard "{}" are replaced by a combination of both qualitative adjectives set including {"no", "a big", "big", "a slight", "slight", "a large", "large", " "} and quantitative percentages set including {10%, ..., 90%, 100%}. We will discuss the choice of various text templates and their outcomes in the ablation study. For the video-anchored experiments, we consider the micromotion of turning heads. The referential frames are collected from the Pointing04 DB dataset (Gourier et al., 2004), and the frames we used for anchoring include a single identity with different postures, which has the angle of $\{-45°, -30°, -15°, 0°, 15°, 30°, 45°\}$.

4.2 Micromotion Subspace Decoding

In this section, we use our anchoring methods to locate the micromotion subspace, discovering the editing direction, and apply it to the in-domain identities to generate desired changes. Figure 4 and Figure 5 show the generated five micromotions using text-anchored and video-anchored methods respectively. Within each row, the five demonstrated frames are sampled from our synthesized video with the desired micromotions. As we can see, these results illustrate continuous transitions of human faces performing micromotions smoothly, which indicates the edit direction decoded from the micromotion subspace is semantically meaningful and highly disentangled. Therefore, our framework successfully locates subspaces for various micromotions.

4.2.1 Quantitative Analysis

To validate if our framework can consistently produce high-quality micromotions, we further compare our decoded micromotions with results from other baselines. We consider the following two methods: InterfaceGAN (Shen et al., 2020b) and GANspace (Härkönen et al., 2020). InterfaceGAN is a supervised disentangle method obtaining edit directions from trained SVMs, while GANspace is an unsupervised disentangle method that discovers edit directions from the principal components of sampled latent codes. We obtain the editing directions from these baselines respectively, performing editings on 2,000 images, and comparing results quantitatively via the following two analyses.

(a) InterfaceGAN (b) GANspace (c) Our Framework

Figure 6: **Re-scoring Analysis** We conduct a re-scoring analysis similar to (Shen et al., 2020b), to quantitatively measure if our discovered editing directions can be successfully disentangled with other irrelevant attributes. We use the scores of a series of pre-trained attribute classifiers to measure how well different editing methods could depict the target attribute (e.g. smile, aging, glasses, etc.). We show a confusion matrix, where in each row we run latent-space editing with each of the attributes, and in each column shows the corresponding changes in the score by different pre-trained attribute classifiers, respectively. Higher scores indicate larger changes. We notice a clear diagonal pattern when editing latent space using our framework, indicating its good disentangle property.

Re-scoring Analysis We conduct a re-scoring analysis similar to (Shen et al., 2020b), to quantitatively measure if our discovered editing directions can be successfully disentangled with other irrelevant attributes. In this analysis, for a target attribute (*e.g.* smiling), we edit the input images using the editing direction discovered by our methods and baselines, and we use the scores of a series of pretrained classifiers (Na, 2021) to measure how the edits influence the target attribute as well as other non-target attributes. Ideally, a well-disentangled editing direction should result in a significant change for the target attributes, while it should have a neglectable influence on other ones.

Our results[1] can be found in Figure 6. In these three confusion matrices, for each row, we run latent-space editing with each of the attributes, and each column shows the corresponding changes measured by pretrained attribute classifiers. For example, in the matrix of InterfaceGAN, the first row means when applying the editing direction "smile" and measuring the changes using pretrained classifiers, the score of "smile" increases by 0.482 while the score of "age" also increases by 0.005, *etc.* Here, higher scores indicate larger changes. From these results, we observe that our framework produces a clear diagonal pattern, indicating a strong influence on desired attributes with little influence on non-target ones. On the other hand, the baseline methods lead to significant changes in both target and non-target attributes. Therefore, our framework demonstrates good disentangle performance compared with previous methods.

Identity-agnostic Analysis It is essential to preserve the identity in face editing. Following the previous work (Shen et al., 2020b), we also perform an identity-agnostic analysis. In this analysis, we use a pre-trained face identifier (Na, 2021) to quantitatively evaluate if the identity is changed after the image editing for different target attributes.

We compare our framework with other editing baselines and demonstrate the results in Table 1. Here, each row shows the changes in identity score when editing the target attributes, and a smaller score indicates better identity preservation. In this table, we

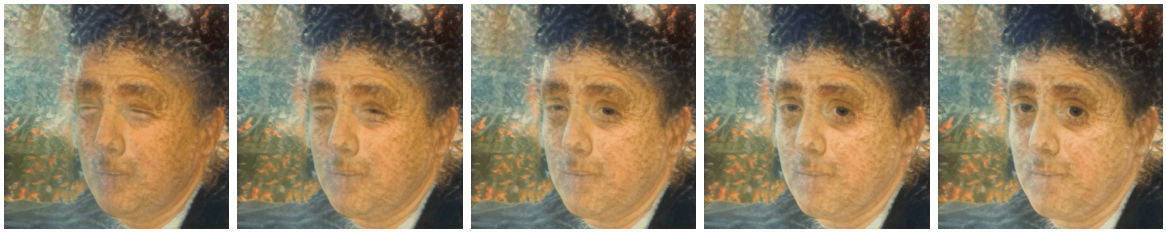
Table 1: Identity-agnostic Analysis. Each row shows the changes in identity score (the smaller the better) when editing the attributes in latent space. Here we want to make sure the attribute editing methods are identity-agnostic. We observed that our method yields the smallest changes in identity score.

	Smile	Age	Gender	Glass
InterfaceGAN	0.0515	0.1294	0.1225	0.0916
GANspace	0.1081	0.0975	0.0507	0.1420
Ours	**0.0047**	**0.0660**	**0.0491**	**0.0279**

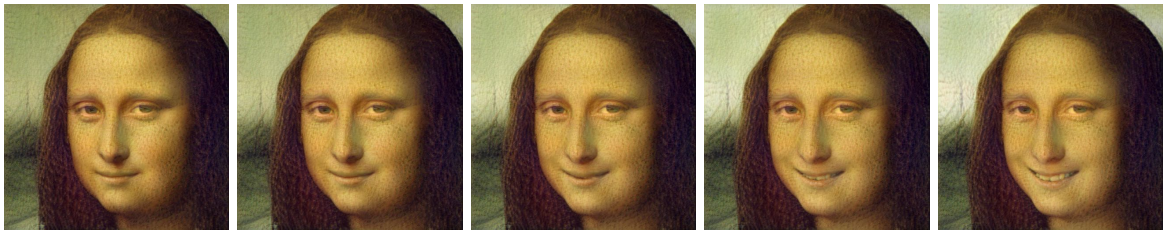
[1]Notice that InterfaceGAN does not release the classifiers they used. Therefore, we use different pretrained classifiers (Na, 2021) to conduct re-scoring analysis.

(a) Anime Character, Smiling

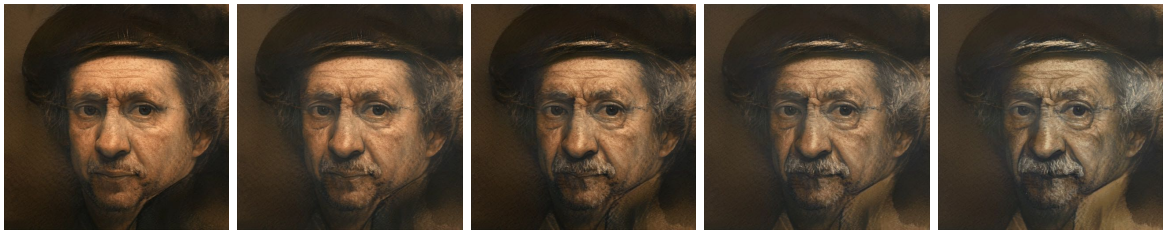
(b) Marble Sculpture, Opening Eyes



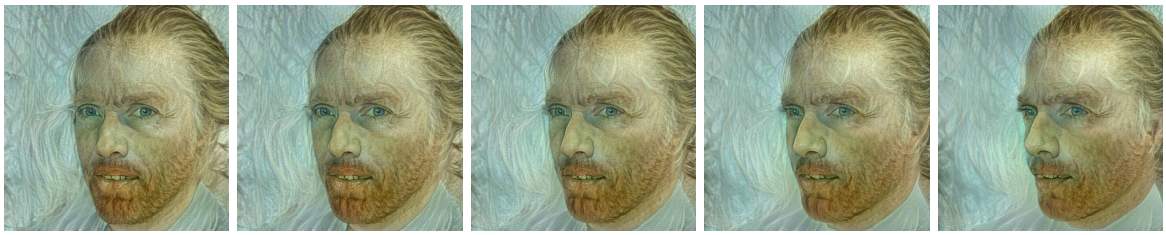
(c) Painting: Monet, Opening Eyes



(d) Painting: Da Vinci, Smiling



(e) Painting: Rembrandt, Aging face



(f) Painting: Van Gogh, Turning Head

Figure 7: **Micromotions on cross-domain identities.** Our micromotions generalize well when transferred to novel domains, including anime characters, sculptures, and various genres of paintings (Van Gogh, Monet, Da Vinci, Rembrandt). Best view when zoomed in. Please refer to supplementary for complete video.

(a) Original

(b) w/ Robust PCA

(c) w/o Robust PCA

Figure 8: **Comparison between with and without Robust PCA.** For each column, from left to right, the micromotions are "closing eyes" (for the first and third columns), "angry", "smiling", and "aging face". For conciseness, we only show the original and last frame. Best view when zoomed in.

find our method preserves identities well in different edits. Other baseline methods, however, lead to more severe identity changes. Therefore, leveraging the strong latent space hypothesis, we find that our framework is able to discover more precise editing directions and better disentangles the target attributes from the human intrinsic identities.

4.3 Micromotion Applications on Cross-domain Identities

Sec.4.2 decodes the micromotion from low-dimensional micromotion subspace, which verifies the first part of the hypothesis. In this section, we further verify the second part of the hypothesis, exploring if the decoded micromotion can be applied to arbitrary and cross-domain identities.

Figure 7 shows the result of transferring the decoded micromotions on novel identities. Within each row, we exert the decoded micromotions on the novel identities, synthesize the desired movements, and demonstrate sampled frames from the generated continuous videos. From these results, we observe that the sampled frames on each new identity also depict the continuous transitions of desired micromotions. This verifies that the decoded micromotions extracted from our workflow can be successfully transited to the out-domain identities, generating smooth and natural transformations. Also, this shows the low-dimensional micromotion subspace in StyleGAN is indeed not isolated nor tied to certain identities. On the contrary, leveraging the low-rank micromotion hypothesis in StyleGAN latent space, the *identity-agnostic* micromotions can be found using our framework and can be ubiquitously applied to those even out-of-domain identities.

Moreover, we emphasize that to generate dynamic micromotion on a novel identity, the entire computational cost boils down to inverting the identity into latent space and then extrapolating along the obtained edit direction, without the requirement of retraining the model or conducting identity-specific computations. This leads to effortless editing of new identity images using the found direction, with little extra cost.

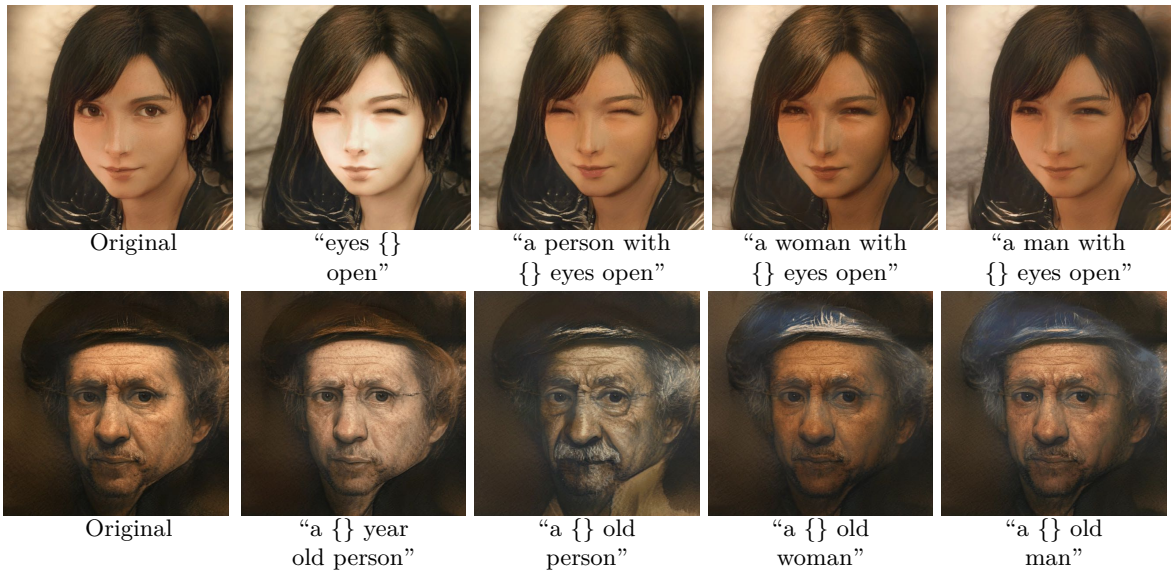


Figure 9: **Ablation on the choice of text template for micromotion "opening eyes" and "aging face".** For each template, we fill the wildcard "{}" using descriptive text, including {10%, 20%, ..., 100%}, {10, 20, ..., 60}, and {small, big, ...}. For conciseness, we only show the last frame of each group. Best view when zoomed in.

4.4 Design Choices and Ablation Study

Ablation on component decomposition in micromotion subspace To verify the effectiveness of the robust decomposition in our workflow, instead of doing robust PCA to decompose the low-rank micromotion space, we sample two anchoring latent codes and adopt its interpolated linear space as the low-rank space. Then, we compare the qualitative results of the decoded micromotions. Results in Figure 8 show that synthesized videos without robust space decomposition step incur many undesired artifacts, often entangling many noisy attributes not belonging to the original and presumably mixed from other identities. Adding a robustness aware subspace decomposition, however, effectively extracts more stable and clearly disentangled linear subspace dimensions in the presence of feature fluctuations and outliers.

Ablation on text templates To explore the sensitivity of the micromotion subspace w.r.t the text templates, we study various text templates that describe the same micromotion. In Figure 9 top row, we can see that the micromotion "closing eyes" is agnostic to the choice of different text templates and generate similar visual results. On the other hand, In Figure 9 bottom row, we observe the opposite where the micromotion "face aging" is sensitive to different text templates, which results in diverse visual patterns. This suggests the choice of text template may influence the performance of some micromotions, and a high-quality text guidance based on prompts engineering or prompts learning could be interesting future work.

Ablations on number of anchors and identities In our framework, we rely on a series of latent codes to anchor the low-rank space, and these codes are obtained from one identity performing micromotion. Therefore, we further ask two questions for our design: (a) How would the number of latent codes influence the editing performance? (b) Could we obtain better editing results from multiple identities? For the first question, we hypothesize the number of latent codes would influence the quality of the discovered low-dimensional space, therefore influencing the editing performance. For the second question, although we have obtained high-quality micromotion editing direction from a single identity, we explore if multiple identities decrease the correlation between micromotion and identity and lead to better disentanglement.

We use different numbers of anchoring latent codes and identities to discover the editing direction and apply it to novel images. In the first study, multiple latent codes are used to determine the low-rank space. Notice that when using only one anchoring latent code, the framework reduces to using StyleCLIP to find editing

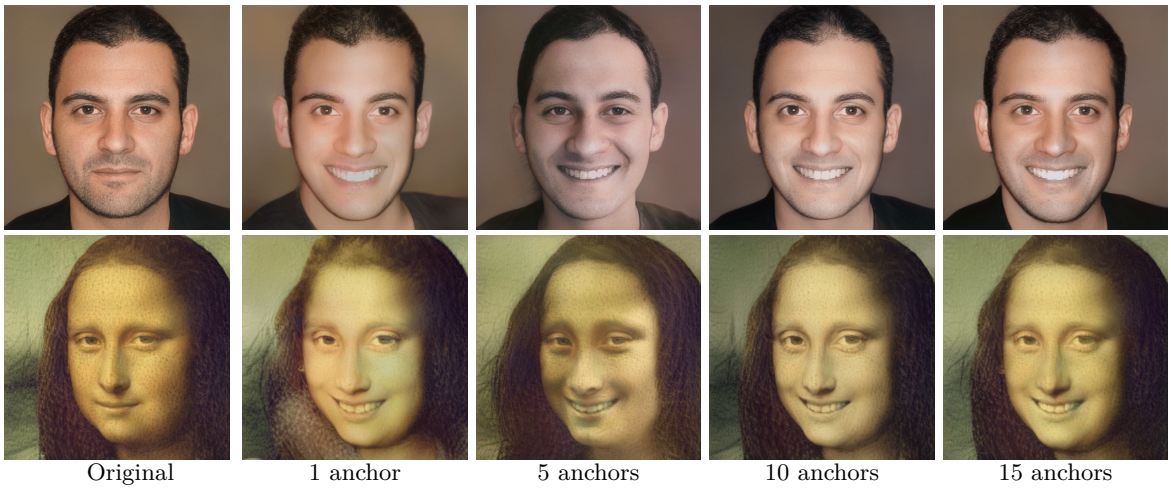


Figure 10: **Ablation on the number of anchoring latent codes used to find the low-rank latent space.** We use the micromotion "smiling" as an example, while applying to both in-domain (human face) and out-of-domain (painting) images, we notice the quality of latent space editing improves proportionally w.r.t the number of anchors until it finally saturated at around 10 anchors.

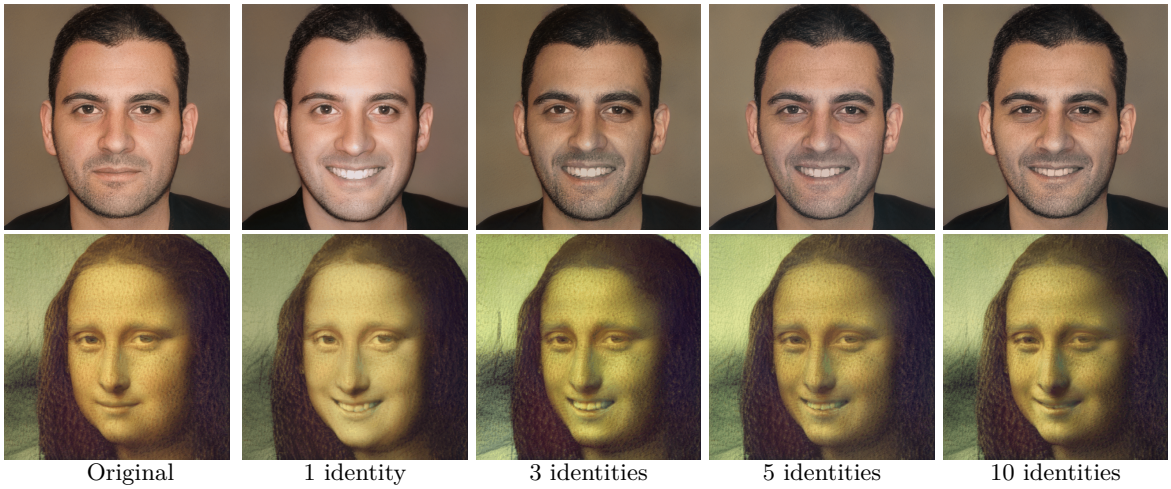


Figure 11: **Ablation on the number of identities used in latent-space optimization.** We use the micromotion "smiling" as an example while applying to both in-domain (human face) and out-of-domain (painting) images. We notice there is no clear visual improvement in the quality of micromotion as the number of identities grows, however it would result in different styles of "smile".

direction and directly apply it to novel images. In the second study, with multiple identities, we optimize the latent code on each identity separately and use the average latent code as the final editing direction.

Our results can be found in Fig 10 and Fig 11. For the effects of anchoring latent codes, we observe that fewer anchors result in noises and artifacts, indicating insufficient disentanglement. Meanwhile, we observe the quality of latent space editing improves gradually with respect to the number of anchors. This motivates us to use a series of anchoring latent codes for a better low-rank latent space. For the effects of identities, we observe that although using more identities has some weak benefits (*e.g.* better background preservation), there is no clear visual improvement compared with using one identity. This motivates us to stick with one identity with better efficiency in our framework.

5 Conclusions

In this work, we analyze the latent space of StyleGAN-v2, demonstrating that although trained with static images, the StyleGAN still captures temporal micromotion representation in its feature space. We find versatile micromotions can be represented by low-dimensional subspaces of the original StyleGAN latent space, and such representations are disentangled and agnostic to the choice of identities. Based on this finding, we explore and successfully decode representative micromotion subspace by two methods: text-anchored and video-anchored reference generation, and these micromotions can be applied to arbitrary cross-domain subjects, even for the virtual figures including oil paintings, sculptures, and anime characters. We also explore the various design choices and their corresponding effects on our framework. Future works may study more complex motion subspace and further explore if latent space corresponding to larger-scale motion is also ubiquitous and can be disentangled from other attributes and identities, which could potentially lead to many interesting applications.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8296–8305, 2020.

Rameen Abdal, Peihao Zhu, John Femiani, Niloy J Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. *arXiv preprint arXiv:2112.05219*, 2021.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021a.

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, 2021b.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Min Jin Chong, Wen-Sheng Chu, Abhishek Kumar, and David Forsyth. Retrieve in style: Unsupervised facial feature transfer and retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3887–3896, 2021.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020.

Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. *arXiv preprint arXiv:2107.07224*, 2021.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.

Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial features. In *ICPR International Workshop on Visual Observation of Deictic Gestures*. Citeseer, 2004.

Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3012–3021, 2020.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

Xianxu Hou, Xiaokang Zhang, Hanbang Liang, Linlin Shen, Zhihui Lai, and Jun Wan. Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. *Neural Networks*, 145:209–220, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in gan for real-time image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 852–861, 2021.

Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021.

Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Vogue: Try-on by stylegan interpolation optimization. *arXiv e-prints*, pp. arXiv–2101, 2021.

Dongbin Na. Celeba hq face identity and attributes recognition. https://github.com/ndb796/CelebA-HQ-Face-Identity-and-Attributes-Recognition-PyTorch, 2021.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532–1540, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020a.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020b.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. *arXiv preprint arXiv:2112.14683*, 2021.

Ryohei Suzuki, Masanori Koyama, Takeru Miyato, Taizan Yonetsuji, and Huachun Zhu. Spatially controllable image synthesis with internal representation collaging. *arXiv preprint arXiv:1811.10153*, 2018.

Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020a.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6142–6151, 2020b.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. *arXiv preprint arXiv:2112.05142*, 2021.

John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, 22, 2009.

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.

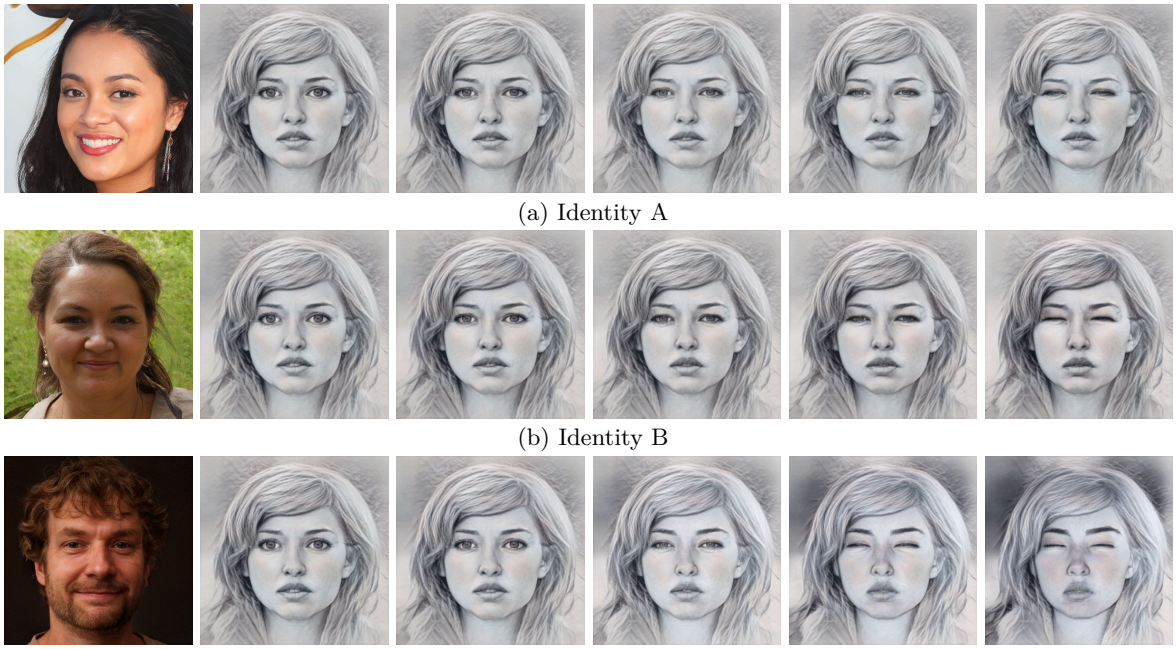
Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, 129(5):1451–1466, 2021.

Lei Zhang and Chris Pollett. Facial expression video synthesis from the stylegan latent space. In *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*, volume 11878, pp. 118781M. International Society for Optics and Photonics, 2021.

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 Ablation on the identities of anchoring latent codes

In this ablation, we explore if the choice of identity influences the micromotion quality. We use photos of different people (denoted as Identity A, B, and C) to discover editing directions, and we generalize to the same sketch painting. The result is shown in Figure 12. We observe that latent codes decoded from various identities generate visually similar micromotions. Therefore, the micromotion can be decoded using different identities and still result in semantically correct edits.



(a) Identity A

(b) Identity B

(c) Identity C

Figure 12: **Ablations on the identities of anchoring latent codes.** The first column shows synthesized images of three different identities generated by three latent codes. From the second column, we show the micromotion subspace (i.e. "closing eyes") decoded from three identities in the first column, exhibiting visually similar results, when generalized to a sketch painting in the novel domain.

B Additional examples of micromotions transferred to novel domains

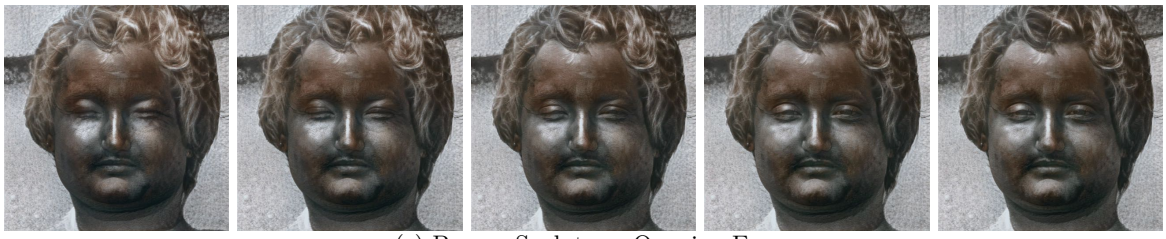
In Figure 13, we include additional visual examples to demonstrate that our micromotions generalize well when transferred to novel domains. The additional novel domains include bronze sculptures, oil/sketch painting, and more anime characters.

C Supplementary Video

We present a series of in-domain and out-of-domain synthesized videos to fully demonstrate the visual effects of our method. Those videos can be found in the "video" folder of the supplementary materials.

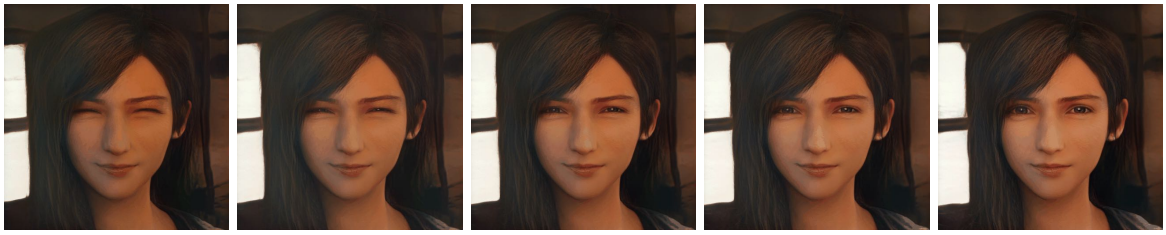
(a) Painting: Pope, Smiling

(b) Anime Character, Opening Eyes



(c) Bronze Sculpture, Opening Eyes

(d) Sketch, Smiling



(e) Anime Character, Opening Eyes

(f) Sketch, Smiling

Figure 13: **Additional examples of micromotions transferred to novel domains.** Our micromotions generalize well when transferred to novel domains, which include anime characters, sculptures, and various genres of paintings (oil painting, sketch). Best view when zoomed in. Please refer to supplementary for complete video sequences.