Interpreting the Weight Space of Customized Diffusion Models

Amil Dravid[*1,2] **Yossi Gandelsman**[*1] **Kuan-Chieh Wang**[2]

Rameen Abdal[3] **Gordon Wetzstein**[3] **Alexei A. Efros**[1] **Kfir Aberman**[2]

[1]UC Berkeley [2]Snap Inc. [3]Stanford University

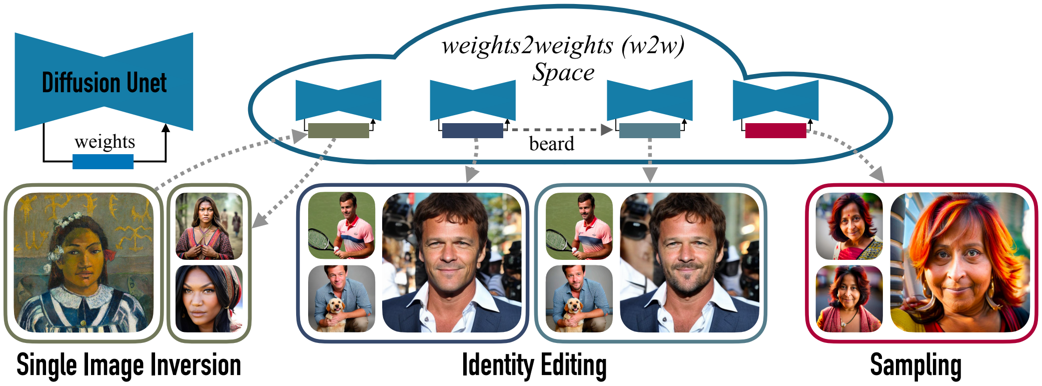


Figure 1: ***weights2weights* (*w2w*) space enables controllable creation of new customized diffusion models.** We model a manifold of customized diffusion models as a subspace of weights that encodes different instances of a broad visual concept (e.g., human identities, dog breeds, etc.). This forms a space that supports inverting the subject (e.g., identity) from a single image into a model, editing the subject encoded in the model, and sampling new models that encode new instances of the visual concept. Each of these operations results in a new model that can consistently generate the subject.

Abstract

We investigate the space of weights spanned by a large collection of customized diffusion models. We populate this space by creating a dataset of over 60,000 models, each of which is a base model fine-tuned to insert a different person's visual identity. We model the underlying manifold of these weights as a subspace, which we term *weights2weights*. We demonstrate three immediate applications of this space that result in new diffusion models – sampling, editing, and inversion. First, sampling a set of weights from this space results in a new model encoding a novel identity. Next, we find linear directions in this space corresponding to semantic edits of the identity (e.g., adding a beard), resulting in a new model with the original identity edited. Finally, we show that inverting a single image into this space encodes a realistic identity into a model, even if the input image is out of distribution (e.g., a painting). We further find that these linear properties of the diffusion model weight space extend to other visual concepts. Our results indicate that the weight space of fine-tuned diffusion models can behave as an interpretable *meta*-latent space producing new models.[1]

[*] Equal contribution

[1] Project page: https://snap-research.github.io/weights2weights
Code: https://github.com/snap-research/weights2weights

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-4363

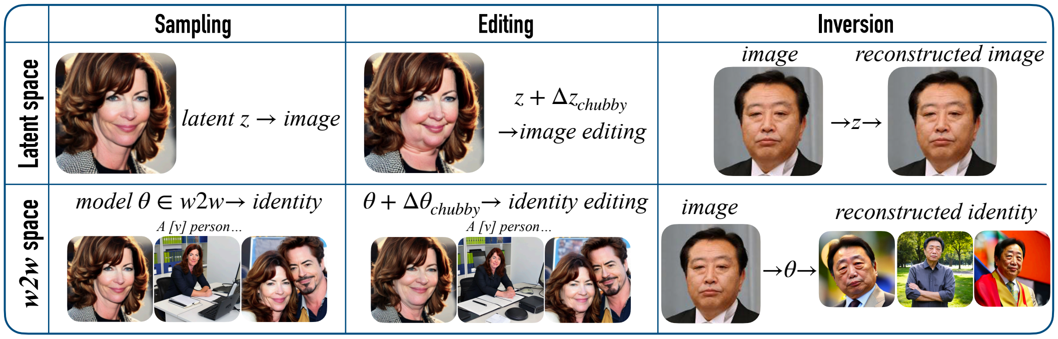


Figure 2: **The *weights2weights* space operates as a *meta*-latent space**. Unlike a traditional generative latent space, *w2w* space controls the model itself rather than single image instances. New identity-encoding models can be sampled from the space and edited by linearly traversing along semantic directions in weight space. Additionally, a single image can be inverted into the space to produce a model that consistently generates that identity.

1 Introduction

Generative models have emerged as a powerful tool to model our rich visual world. In particular, the latent space of single-step generative models, such as generative adversarial networks (GANs) [18, 28], has been shown to linearly encode meaningful concepts in the output images. For instance, datasets of latent vectors were used to discover linear directions in the GAN latent space encoding different attributes (e.g., gender or age of faces) [20, 60]. Even earlier, datasets of images and keypoints were leveraged to discover subspaces of facial shape and appearance [6, 53].

We aim to extend this even further, using datasets of model weights instead datasets of images or latents. Can we discover such interpretable subspaces in the model weights themselves? Recently introduced personalization approaches, such as Dreambooth [54] or Custom Diffusion [34], may hint that this is the case. These methods aim to learn an instance of a subject, such as a person's visual identity. Rather than searching for a latent code that represents an identity in the input noise space, these approaches customize diffusion models by fine-tuning on subject-specific images, which results in identity-specific model weights. We therefore hypothesize that a latent space can exist *in the weights themselves.*

To test our hypothesis, we fine-tune over 60,000 personalized models on individual identities to obtain points that lie on a manifold of customized diffusion model weights. To reduce the dimensionality of each data point, we use low-rank approximation (LoRA) [23] during fine-tuning and further apply Principal Components Analysis (PCA) to the set of data points. This forms our final space: *weights2weights* (*w2w*). Unlike traditional generative models like GANs, which model the pixel space of images, we model the *weight space* of these personalized models. Thus, each sample in our space corresponds to an identity-specific model which can consistently generate that subject. We provide a schematic in Fig. 2 that contrasts a typical latent space with our proposed *w2w* space, demonstrating the differences and analogies between these two representations. *w2w* space can be thought of as a *meta*-latent space, enabling controllable creation of new models instead of just images like a traditional latent space.

Creating this space unlocks a variety of applications that involve traversal in *w2w* (Fig. 1). First, we demonstrate that sampling model weights from *w2w* space corresponds to a new model encoding a novel subject. Second, we find linear directions in this space corresponding to semantic edits of the identity. Finally, we show that enforcing weights to live in this space enables a diffusion model to learn a subject given a single image, even if it is out of distribution.

We find that *w2w* space is highly expressive through quantitative evaluation on editing customized models and encoding new identities given a single image. Qualitatively, we observe this space supports sampling models that encode diverse and realistic identities, while also capturing the key characteristics of out-of-distribution identities. We finally demonstrate that similar weight subspaces exist for other visual concepts such as dog breeds and car types.

https://doi.org/10.52202/079017-4363

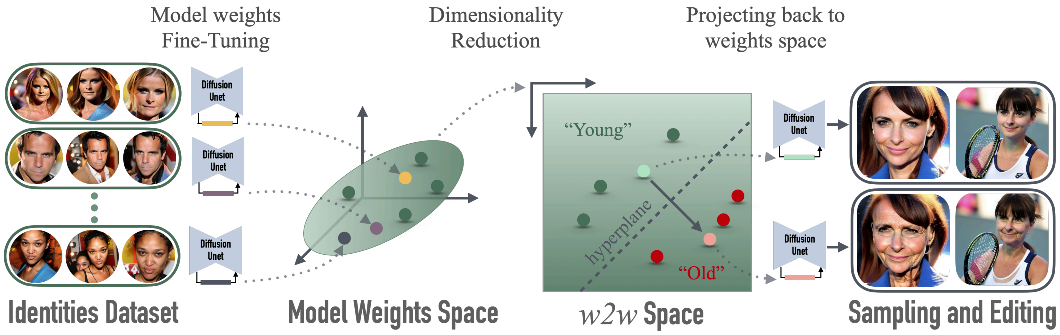


Figure 3: **Building *weights2weights* (*w2w*) space.** We create a dataset of model weights where each model is personalized to a specific identity using low-rank updates (LoRA). These model weights lie on a weights manifold that we further project into a lower-dimensional subspace spanned by its principal components. We train linear classifiers to find disentangled edit directions in this space.

2 Related Work

Image-based generative models. Various models have been proposed for image generation, including variational autoencoders (VAEs) [31], flow-based models [12, 32, 49], generative adversarial networks (GANs) [18], and diffusion models [22, 43, 62]. Within the realm of high-quality photorealistic image generation, GANs [25, 28, 29] and diffusion models [22, 43, 52, 63] have garnered significant attention due to their controllability and ability to produce high-quality images. Leveraging the compositionality of these models, methods for personalization and customization have been developed which aim to insert user-defined concepts via fine-tuning [16, 34, 40, 54]. Various works try to reduce the dimensionality of the optimized parameters for personalization either by operating in specific model layers [34] or in text-embedding space [16], by training hypernetworks [55], and by constructing a linear basis in text embedding space [73].

Latent space of generative models. Linear latent space models of facial shape and appearance were studied extensively in the 1990s, using PCA-based representations (e.g. Active Appearance Models [9], 3D Morphable Models [6]) as well as operating directly in pixel and keypoint space [53]. However, these techniques were restricted to aligned and cropped frontal faces. More recently, generative adversarial networks (GANs), particularly the StyleGAN series [26, 27, 28, 29], have showcased editing capabilities facilitated by their interpretable latent space. Furthermore, linear directions can be found in their latent space to conduct semantic edits by training linear classifiers or applying PCA [20, 60], among other methods for discovering semantic directions [8, 66]. Several methods aim to project real images into the GAN latent space in order to conduct this editing [1, 3, 51, 64, 77]. Beyond the latent space, works such as [4] found that directions could be discovered in the neuron activation space, suggesting the interpretability of weights.

Although diffusion models architecturally lack a GAN-like latent space, some works aim to discover similar spaces in these models. This has been explored in the UNet bottleneck layer [35, 41], noise space [10, 78], and text-embedding space [5]. Concept Sliders [17] explores the weight space for semantic image editing by conducting low-rank training with contrasting image or text pairs.

Weights as data. Past works have exploited the structure within weight space of deep networks for various applications. In particular, some have found linear properties of weights, enabling simple model ensembling and editing via arithmetic operations [24, 56, 59, 69]. Other works create datasets of neural network parameters for training hypernetworks [14, 19, 42, 55, 68], predicting properties of networks [58], and creating design spaces for models [46, 47].

3 Method

We start by demonstrating how we create a manifold of model weights as illustrated in Fig. 3. We explain how we obtain low-dimensional data points for this space, each of which represents an individual subject from a broad class (i.e., identity). We then use these points to model a weights manifold. Next, we find linear directions in this manifold that correspond to semantic attributes and use them for editing the identities. Finally, we demonstrate how this manifold can be utilized for constraining an ill-posed inversion task with a single image to reconstruct its identity.

https://doi.org/10.52202/079017-4363

3.1 Preliminaries

In this section, we first introduce latent diffusion models (LDM) [52], which we will use to create a dataset of weights. Then, we explain the approach for obtaining identity-specific models from LDM via Dreambooth [54] fine-tuning. We finally present a version of fine-tuning that uses low-dimensional weight updates (LoRA [23]). We will use the fine-tuned low-dimensional per-identity weights as data points to construct the weights manifold in Sec. 3.2.

Latent diffusion models [52]. We will extract weights from latent diffusion models to create *w2w* space. These models follow the standard diffusion objective [22] while operating on latents extracted from a pre-trained Variational Autoencoder [15, 31, 50]. With text, the conditioning signal is encoded by a text encoder (such as CLIP [44]), and the resulting embeddings are provided to the denoising UNet model. The loss of latent diffusion models is:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,t}[w_t||\epsilon - \epsilon_\theta(\mathbf{x}_t,\mathbf{c},t)||_2^2], \tag{1}$$

where ϵ_θ is the denoising UNet, $\mathbf{x}_t$ is the noised version of the latent for an image, $\mathbf{c}$ is the conditioning signal, t is the diffusion timestep, and w_t is a time-dependent weight on the loss.

To sample from the model , a random Gaussian latent x_T is deterministically denoised conditioned on a prompt for a fixed set of timesteps with a DDIM sampler [63]. The denoised latent is then fed through the VAE decoder to generate the final image.

Dreambooth [54]. To obtain an identity-specific model, we use the Dreambooth personalization method. Dreambooth fine-tuning introduces a novel subject into a pre-trained diffusion model given only a few images of it. During training, Dreambooth follows a two-part objective:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,t}[w_t||\epsilon - \epsilon_\theta(\mathbf{x}_t,\mathbf{c},t)||_2^2 + \lambda w_{t'}||\epsilon' - \epsilon_\theta(\mathbf{x}'_t,\mathbf{c}',t')||_2^2], \tag{2}$$

where the first term corresponds to the standard diffusion denoising objective using the subject-specific data $\mathbf{x}$ conditioned on the text prompt "[identifier] [class noun]" (e.g., "*[v]* person"), denoted $\mathbf{c}$. The second term, weighted by λ, corresponds to a prior preservation loss, which involves the standard denoising objective using the model's own generated samples $\mathbf{x}'$ for the broader class $\mathbf{c}'$ (e.g., "person"). This prevents the model from associating the class name with the specific instance, while also leveraging the semantic prior on the class.

Low Rank Adaptation (LoRA) [23]. Dreambooth requires fine-tuning all the weights of a model, which is a high–dimensional space. We turn to a more efficient fine-tuning scheme, LoRA, that modifies only a low-rank version of the weights. LoRA uses weight updates ΔW with a low intrinsic rank. For a base model layer $W \in \mathbb{R}^{m\times n}$, the LoRA update for that layer ΔW can be decomposed into $\Delta W = BA$, where $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$ are low-rank matrices with $r \ll min(m,n)$. During training, for each model layer, only the A and B are updated. This significantly reduces the number of trainable parameters. During inference, the low-rank weights are added residually to the weights of each layer in the base model and scaled by a coefficient $\alpha \in \mathbb{R}$: $W + \alpha\Delta W$.

3.2 Constructing the weights manifold

Creating a dataset of model weights. To construct the *weights2weights* (*w2w*) space, we begin by creating a dataset of model weights θ_i. We conduct Dreambooth fine-tuning on latent diffusion models in order to insert new subjects with the ability to control image instances using text prompts. This training is done with LoRA in order to reduce the space of model parameters. Each model is fine-tuned on a set of images corresponding to one human subject. After training, we flatten and concatenate all of the LoRA matrices, resulting in a data point $\theta_i \in \mathbb{R}^d$ which represents one identity. After training over N different instances, we have our final dataset of model weights $\mathcal{D} = \{\theta_1, \theta_2, ..., \theta_N\}$, representing a diverse array of subjects.

Modeling the weights manifold. We posit that our data $D \subseteq \mathbb{R}^d$ lies on a lower-dimensional manifold of weights that encode identities. A randomly sampled set of weights in $\mathbb{R}^d$, would not be guaranteed to produce a valid model encoding identity as the d degrees of freedom can be fine-tuned for any purpose. Therefore, we hypothesize that this manifold is a subset of the weight space. Inspired by findings that high-level concepts can be encoded as linear subspaces of representations [13, 37, 45, 48], we model this subset as a linear subspace $\mathbb{R}^m$ where $m < d$, and call it *weights2weights* (*w2w*) space. We represent points in this subspace as a linear combination of basis vectors $\mathbf{w} = \{w_1, ..., w_m\}$, $w_i \in \mathbb{R}^d$. In practice, we apply Principal Component Analysis (PCA) on the N models and keep the first m principal components for dimensional reduction and forming our basis of m vectors.

Sampling from the weights manifold. After modeling this weights manifold, we can sample a new model that lies on it, resulting in a new model that generates a novel identity. We sample a model represented with basis coefficients $\{\beta_1, ..., \beta_m\}$, where each coefficient β_k is sampled from a normal distribution with mean μ_k and standard deviation σ_k. The mean and standard deviation are calculated for each principal component k from the coefficients among all the training models.

3.3 Finding Interpretable Weight Space Directions

We seek a direction $\mathbf{n} \in \mathbb{R}^d$ defining a hyperplane that separates between binary identity properties embedded in the model weights (e.g., male/female), similarly to hyperplanes observed in the latent space of GANs [60]. We assume binary labels are given for attributes present in the identities encoded by the models. We then train linear classifiers using weights of the models as data based on these labels, imposing separating hyperplanes in weight space. Given an identity parameterized by weights θ, we can manipulate a single attribute by traversing in a direction $\mathbf{n}$, orthogonal to the separating hyperplane: $\theta_{\text{edit}} = \theta + \alpha\mathbf{n}$. An edit operation in *w2w* spaee produces a new model with the original subject edited, allowing the model to generate infinitely many new images of the edited subject.

3.4 Inversion into *w2w* Space

Traditionally, inversion of a generative model involves finding an input such as a latent code that best reconstructs a given image [38, 70]. This corresponds to finding a projection of the input onto the learned data manifold [77]. With *w2w* space, we model a manifold of model weights rather than images. Inspired by latent optimization methods [1, 77], we propose a gradient-based method of inverting a single identity from an image into our discovered space.

Given a single image $\mathbf{x}$, we follow a constrained denoising objective:

$$\max_{\theta} \mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,t}[w_t||\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)||_2^2] \quad \text{s.t. } \theta \in \textit{w2w} \tag{3}$$

Specifically, we constrain the model weights to lie in *w2w* space by optimizing a set of basis coefficients $\{\beta_1, ..., \beta_m\}$ rather than the original parameters. Unlike Dreambooth, we do not employ a prior preservation loss, since the optimized model lies in the subspace defined by our dataset of weights, and inherits their priors.

4 Experiments

We demonstrate *w2w* space on the visual concept of human identities for a variety of applications. We begin by detailing implementation details. Next, we use *w2w* space for 1) sampling new models encoding novel identities, 2) editing identity attributes in a consistent manner via linear traversal in *w2w* space, 3) embedding a new identity given a single image, and 4) projecting out-of-distribution identities into *w2w* space. Finally, we analyze how scaling the number of models in our dataset of model weights affects the disentanglement of attribute directions and preservation of identity.

4.1 Implementation Details

Creating an identity dataset. We generate a synthetic dataset of ∼65,000 identities using [67], where each identity is associated with multiple images of that person. Each identity is based on an image with labeled binary attributes (e.g., male/female) from CelebA [36]. Each set of images corresponding to an identity is then used as data to fine-tune a latent diffusion model with Dreambooth. Further details on this dataset and train/test splits are provided in Appendix E.

Encoding identities into model weights. We conduct Dreambooth fine-tuning using LoRA with rank 1 on the identities. Following [56], we only fine-tune the query and value projection matrices in the cross-attention layers. We utilize the RealisticVision-v51 checkpoint[2] based on Stable Diffusion

[2] https://huggingface.co/stablediffusionapi/realistic-vision-v51

https://doi.org/10.52202/079017-4363

1.5. Conducting Dreambooth fine-tuning on each identity training set results in a dataset of ∼65,000 weights θ where $\theta \in \mathbb{R}^{100,000}$.

Finding semantic attribute directions. We utilize binary attribute labels from CelebA to train linear classifiers on the dataset of model weights we curated. We run Principal Component Analysis (PCA) on the ∼65,000 training models and project to the first 1000 principal components in order to reduce the dimensionality. The orthogonal edit directions are calculated via the analytic least squares solution on the matrix of projected training models $\mathcal{D} \in \mathbb{R}^{65,000\times 1000}$, and then unprojected to the original dimensionality of the model weights: $\theta \in \mathbb{R}^{100,000}$.

4.2 Sampling from *w2w* Space

We present images generated from models that were sampled from the weights manifold (i.e., *w2w* Space) in Fig. 4. We follow the sampling procedure from Sec. 3.2, and generate images from the sampled model. As shown, each new model encodes a novel, realistic, and consistent identity. Additionally, we present the nearest neighbor model among the training dataset of model weights. We use cosine similarity on the models' principal component representations. Comparing with the nearest neighbors shows that these samples are not just copies from the dataset, but rather encode diverse identities with different attributes. Yet, the samples still demonstrate some similar features to the nearest neighbors. These include jawline and eye shape (top row), facial hair (middle row), and nose and eye shape (bottom row). Appendix A includes more such examples.

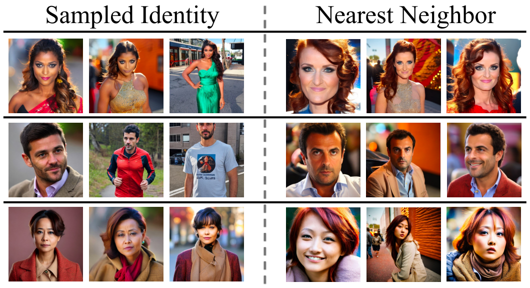


Figure 4: **Identity samples from *w2w* space.** We show the samples from *w2w* space do not overfit to nearest-neighbor identities, although they incorporate facial attributes from them. The identities are diverse and consistent across generations.

4.3 Editing Subjects

We demonstrate how directions found by the linear classifiers can be used to edit subjects encoded in the models. It is desired that these edits are 1) disentangled (i.e., do not interfere with other attributes of the embedded subject and preserve all other concepts such as context) 2) identity preserving (i.e., the person is still recognizable) 3) and semantically aligned with the intended edit.

Baselines. We compare against a naïve baseline of prompting with the desired attribute (e.g., "*[v]* person with small eyes"), and then Concept Sliders [17], an instance-specific editing method which we adapt to subject editing. In particular, we train their most accessible method, the text-based slider, which trains LoRAs to modulate attributes in a pretrained diffusion model based on contrasting text prompts. We then apply these sliders to the personalized identity models.

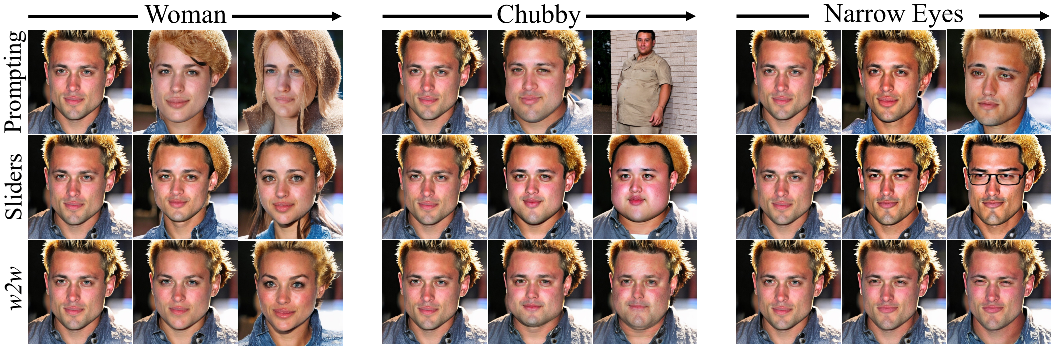


Figure 5: **Qualitative comparison.** *w2w* edits preserve identity while being disentangled and semantically aligned. Concept Sliders [17] tends to exaggerate effects which induces artifacts and degrades identity, while prompting the subject with the desired edit has unexpected effects.

https://doi.org/10.52202/079017-4363

Table 1: **Edits in *w2w* space preserve identity, are disentangled, and semantically aligned.**

	ID Score ↑			LPIPS ↓			CLIP Score ↑		
	Prompting	Sliders	*w2w*	Prompting	Sliders	*w2w*	Prompting	Sliders	*w2w*
Gender	$0.39_{\pm 0.08}$	$0.33_{\pm 0.09}$	$\mathbf{0.45}_{\pm 0.09}$	$\mathbf{0.30}_{\pm 0.05}$	$0.39_{\pm 0.09}$	$0.31_{\pm 0.03}$	$1.98_{\pm 0.78}$	$3.50_{\pm 0.68}$	$\mathbf{4.13}_{\pm 0.59}$
Chubby	$0.29_{\pm 0.14}$	$0.33_{\pm 0.09}$	$\mathbf{0.45}_{\pm 0.09}$	$0.41_{\pm 0.05}$	$0.38_{\pm 0.04}$	$\mathbf{0.36}_{\pm 0.04}$	$1.12_{\pm 0.61}$	$\mathbf{2.21}_{\pm 0.61}$	$2.16_{\pm 0.51}$
Eyes	$0.52_{\pm 0.06}$	$0.53_{\pm 0.04}$	$\mathbf{0.72}_{\pm 0.05}$	$0.32_{\pm 0.03}$	$0.30_{\pm 0.02}$	$\mathbf{0.19}_{\pm 0.02}$	$0.17_{\pm 0.17}$	$0.01_{\pm 0.22}$	$\mathbf{0.59}_{\pm 0.19}$

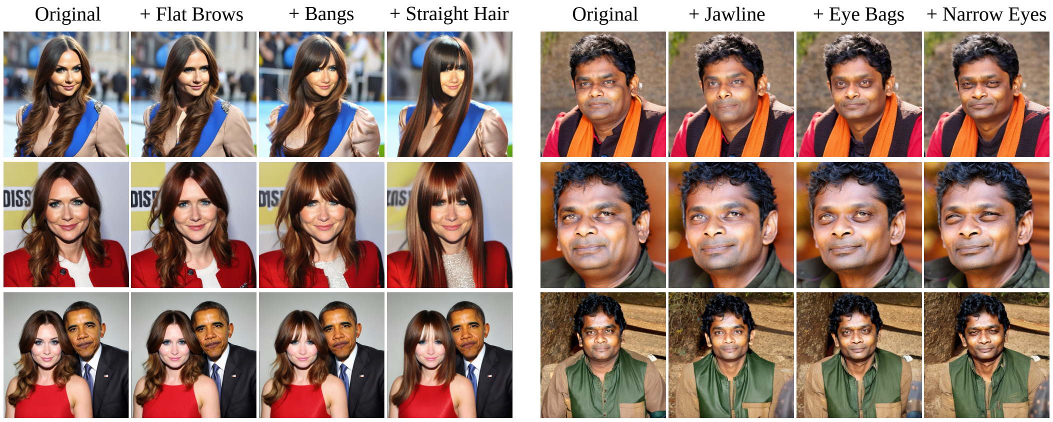


Figure 6: **Composing edits in *w2w* space.** Each column represents fixed seed samples from an edited model. Multiple edits in *w2w* space minimally degrade the original identity or interfere with other concepts, while maintaining edit appearance across different samples.

Evaluation protocol. We evaluate these three methods for identity preservation, disentanglement, and edit coherence. To measure identity preservation, we first detect faces in the original generated images and the result of the edits using MTCNN [74]. We then calculate the similarity of the FaceNet [57] embeddings. We also use LPIPS [76] computed between the images before and after the edit to measure the degree of disentanglement with other visual elements, and CLIP score [21], to measure if the desired edit matches the text caption for the edit.

To generate samples, we fix a set of prompts and random seeds which are used as input to the held-out identity models. Then, we choose a set of identity-specific manipulations. For prompt-based editing, we augment the attribute description to the set of fixed prompts (e.g., “chubby *[v]* person"). For Concept Sliders and *w2w*, we apply the weight space edit directions to the personalized model with a fixed norm which determines the edit strength. The norm is calculated using the maximum projection component onto the edit direction among the training set of model weights.

***w2w* edits are identity preserving and disentangled.** We evaluate over a range of identity-specific attributes and present three (gender, chubby, narrow eyes) in Tab. 1. Edits in *w2w* preserve the identity of the original subject as measured by the ID score. These edits are semantically aligned with the desired effect as indicated by the CLIP score while minimally interfering with other visual concepts, as measured by LPIPS. We note that the CLIP score can be noisy in this setting as text captions can be too coarse to describe attributes as nuanced as those related to the human face. We supplement this with a user study presented in Appendix B.

Qualitatively, *w2w* edits make the minimal amount of changes to achieve semantic and identity-preserving edits (Fig. 5). For instance, changing the gender of the man does not significantly change the facial structure or hair, unlike Concept Sliders or prompting with text descriptions. Prompting has inconsistent results, either creating no effect or making drastic changes. Concept Sliders tends to make caricaturized effects, such as making the man cartoonishly chubby and baby-like.

Composing edits. Edit directions in *w2w* space can be composed linearly as shown in Fig. 6. The first column represents samples from the original model, and each subsequent column represents samples from the edited models. Each row shares the same fixed random generation seed. The composed edits persist in appearance across different generations, binding to the identity. Furthermore, the edited weights result in a new model, where the subject has different attributes while still maintaining as much of the prior identity. This is in contrast to editing in a traditional latent space, where an edit only corresponds to a single image. Additionally, as we operate on an personalized identity-specific weight manifold, minimal changes are made to other concepts, such as scene layout or other people. For instance, in Fig. 6, adding edits to the woman does not interfere with the person standing by her.

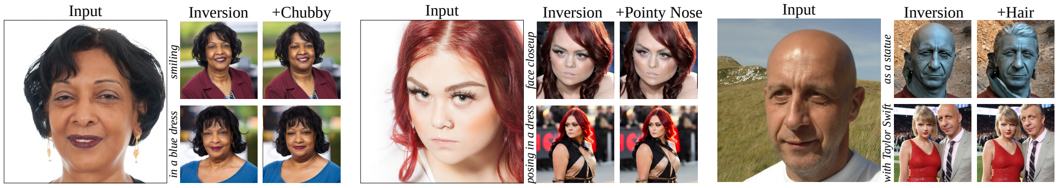


Figure 7: **Single image inversion reconstructs identity and enables editing in *w2w* space.** We present generated samples from the inverted models. These inverted identities can be composed in novel contexts and edited using our discovered semantic directions in weight space. These edits persist in appearance across generation seeds and prompts.

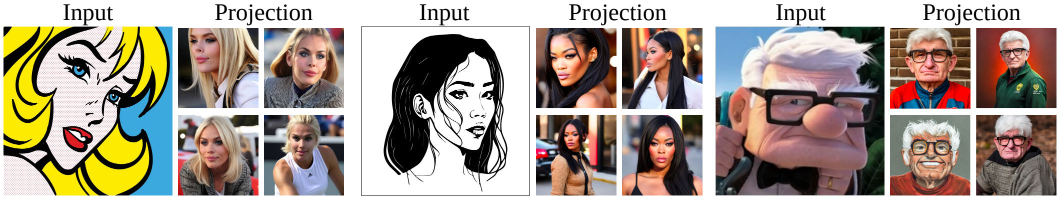


Figure 8: **Projecting out-of-distribution identities.** We show that our inversion method can convert unrealistic identities into realistic renderings with in-domain facial features. Each image represents a generated sample from the inverted model. The resulting identities can be composed in novel scenes, such as playing tennis or rendered into other artistic domains.

4.4 Inverting Subjects

Evaluation protocol. We measure *w2w* space's ability to represent novel identities by inverting a set of 100 random FFHQ [28] face images. We follow our inversion objective from eq. 3. We then provide a set of diverse prompts to generate multiple images and follow the identity preservation metric from Sec. 4.3 to measure subject fidelity. Implementation details are provided in Appendix C.

We compare our results to two approaches that use Dreambooth with rank-1 LoRA. The first is trained on a single image. The second is trained on *multiple images of each identity*. We generate such images by following our identity dataset construction from Sec. 4.1. This approach can be viewed as a pseudo-upper bound on modeling identity as it uses multiple images.

***w2w* space provides a strong identity prior.** Inverting a single image into *w2w* space improves on the single image Dreambooth baseline and closes the gap with the Dreambooth baseline that uses multiple identity images (Tab. 2). Conducting Dreambooth fine-tuning with a single image in the original weight space leads to image overfitting and poor subject reconstruction as indicated by a lower ID score. In contrast, by constraining the optimized weights to lie on a manifold of identity weights, *w2w* inversion inherits the rich priors of the models used to discover the space. As such, it can extract a high-fidelity identity that is consistent and compositional across generations. We present qualitative comparisons against Dreambooth and single-image Dreambooth in Appendix C. We additionally compare against other personalization methods in that section.

Inverted models are editable. Fig. 7 demonstrates that a diverse set of identities can be faithfully represented in *w2w* space. After inversion, the encoded identity can be composed in novel contexts and poses. For instance, the inverted man (rightmost example) can be seen posing with a celebrity or rendered as a statue. Moreover, semantic edits can be applied to the inverted models while maintaining appearance across generations.

Table 2: ***w2w* Inversion closes the gap with Dreambooth.**

Method	Single-Image	ID Score ↑
DB-LoRA	×	**0.69** ± 0.01
DB-LoRA	✓	0.43 ± 0.03
w2w	✓	0.64 ± 0.01

4.5 Out-of-Distribution Projection

***w2w* space captures out-of-distribution identities.** We follow the *w2w* inversion method from Sec. 4.4 to project images of unrealistic identities (e.g., paintings, cartoons, etc.) onto the weights manifold, and present these qualitative results in Fig. 8. By constraining the optimized model to live in *w2w* space, the inverted identities are converted into realistic renditions of the stylized identities, capturing prominent facial features. In Fig. 8, notice how the inverted identities generate a similar

https://doi.org/10.52202/079017-4363

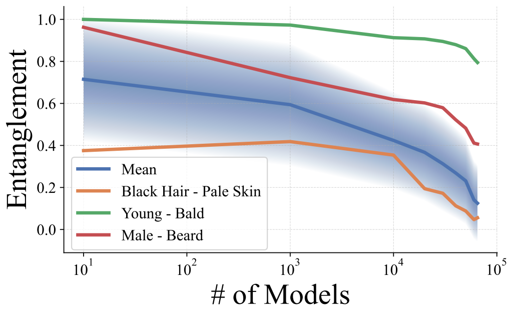


Figure 9: **Scaling dataset of models further disentangles classifier directions.** We highlight the trend in disentanglement of three examples where attributes may be strongly correlated among identities. As the number of models is increased, the features are less entangled.

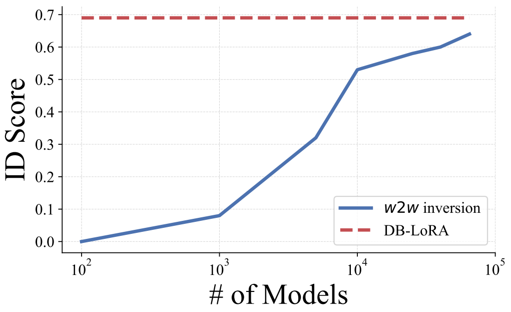


Figure 10: **Scaling the number of models improves identity preservation.** As the span of *w2w* space increases, inversion can reconstruct single-image identities more faithfully, approaching the pseudo-upper bound of multi-image Dreambooth (DB-LoRA).

blonde hairstyle and nose structure in the first example, defined jawline and lip shape in the second example, and head shape and big nose in the last example. As also shown in the figure, the inverted identities can also be translated to other artistic domains using text prompts. We present a variety of domains projected into *w2w* space in Appendix D.

4.6 Effect of Number of Models Spanning *w2w* Space

We ablate the number of models used to create *w2w* space and investigate the expressiveness of the resulting space. In particular, we measure the degree of entanglement among the edit direction and how well this space can capture identity.

Disentanglement vs. the number of models. We find that scaling the number of models in our dataset of weights leads to less entangled edit directions in *w2w* space (Fig. 9). We vary the number of models in our dataset of weights and reapply PCA to establish a basis. We then measure the absolute value of cosine similarity (lower is better) between all pairs of linear classifier directions found for CelebA labels. We repeat this as we scale the number of model weights used to train the classifiers. We report the mean and standard deviation for these scores, along with three notable semantic direction pairs. We observe a trend in decreasing cosine similarity. Notably, pairs such as "Black Hair - Pale Skin," "Young - Bald," and "Male - Beard" which may correlate in the distribution of identities, become less correlated as we scale our dataset of model weights.

Identity preservation vs. the number of models. We observe that as we scale the number of models in our dataset of weights, identities are more faithfully represented in *w2w* space (Fig. 10). We follow the same procedure as the disentanglement ablation, reapplying PCA to establish a basis based on the dataset of model weights. Next, following Sec. 4.4, we optimize coefficients for this basis and measure the average ID score over the 100 inverted FFHQ evaluation identities. As each model in our dataset encodes a different instance of an identity, growing this dataset increases the span of *w2w* space and its ability to capture more diverse identities. We plot the average multi-image Dreambooth LoRA (DB-LoRA) ID score from Sec. 4.4, which is agnostic to our dataset of models. This establishes a pseudo-upper bound on identity preservation. Scaling enables *w2w* to represent identities given a single image with performance approaching that of traditional Dreambooth with LoRA, which uses multiple images and trains in a higher dimensional space.

5 Extending to Other Domains

We extend our hypothesis of interpretable linear weight subspaces in diffusion models to other visual concepts beyond human identities. We apply the *weights2weights* framework to form a subspace for models encoding different dog breeds. To create a dataset for fine-tuning, we generate images with Stable Diffusion based on each of the 120 dog classes from ImageNet [11]. We then conduct Dreambooth fine-tuning on each set of dog breed images to create a dataset of 120 dog-encoding models, subsequently applying PCA. To find edit directions, we use GPT-4 [2] to create labels for

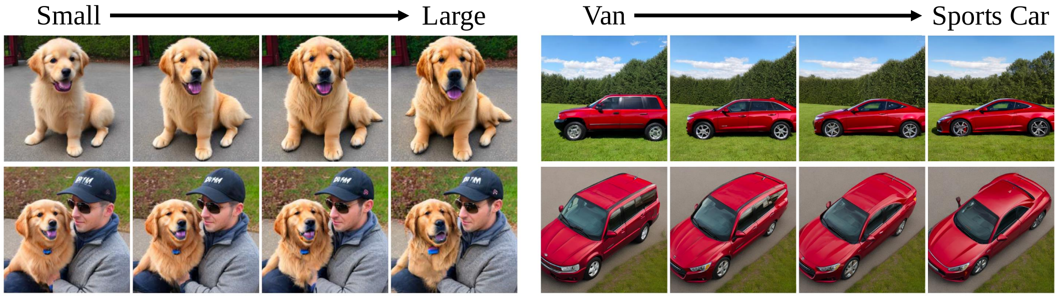


Figure 11: ***weights2weights* linear subspaces can be created for other visual concepts.** We follow the same procedure of applying PCA and finding edit directions with linear classifiers on datasets of models encoding dog breeds and models encoding car types.

each dog breed (e.g., wavy hair or not) and then train linear classifiers on the model weight principal component projections like Sec. 3.3. We apply the same framework to different car categories using models fine-tuned on images from a dataset of 197 different car types [33]. We present results for traversing edit directions in these two subspaces in Fig. 11. Each column represents samples from an edited model. Each row shares the same fixed random generation seed.

Our results provide further evidence that diffusion models can encode visual concepts linearly. This enables the creation of new models in a controlled manner via simple interpolation. For instance, in Fig. 11, we rewrite the model's learned concept of a small golden retriever to make it bigger, or the model's encoding of a red van to make it a sports car. Additionally, unlike older PCA-based methods [6, 53, 61, 65] which rely on aligned pixels or keypoints of human faces, *weights2weights* can extend to other domains beyond human identities. We refer the reader to Appendix G for more results of applying *w2w* space to other visual concepts.

6 Limitations

As with any data-driven method, *w2w* space inherits the biases of the data used to discover it. For instance, co-occurring attributes in the identity-encoding models would cause linear classifier directions to entangle them (e.g. gender and facial hair). However, as we scale the number of models, spurious correlations will drop as evidenced by Fig. 9. These semantic directions are also limited by the labels present in CelebA. Additionally, the span of the *w2w* space is dictated by the models used to create it. Thus, *w2w* space can struggle to represent more complex identities as seen in Fig. 12. Inversion in these cases amounts to projecting onto the closest identity on the weights manifold. Despite this, our analysis on the size of the model dataset reveals that forming a space using a larger and more diverse set of identity-encoding models can mitigate this limitation.

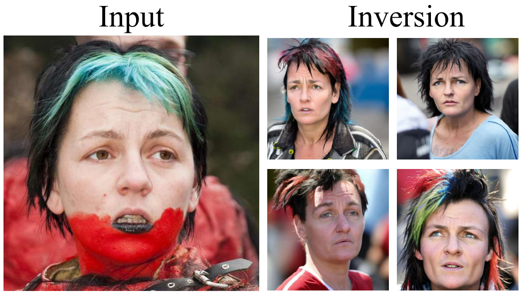


Figure 12: ***weights2weights* fails to capture identities with undersampled attributes.**

7 Discussion and Broader Impact

We presented a paradigm for representing diffusion model weights as a point in a subspace defined by other customized models – *weights2weights* (*w2w*) space. This enabled applications analogous to those of a generative latent space – inversion, editing, and sampling – but producing model weights rather than images, resulting in what we term a *meta*-latent space. We demonstrated these applications on model weights encoding human identities and extended this space to other visual concepts. Although these applications could enable malicious manipulation of real human identities and model weights, we hope the community uses the framework to explore visual creativity as well as utilize this interpretable space for controlling models for safety.

https://doi.org/10.52202/079017-4363

Acknowledgements

The authors would like to thank Grace Luo, Lisa Dunlap, Konpat Preechakul, Sheng-Yu Wang, Stephanie Fu, Or Patashnik, Daniel Cohen-Or, and Sergey Tulyakov for helpful discussions. AD is supported by the US Department of Energy Computational Science Graduate Fellowship. Part of the work was completed by AD as an intern with Snap Inc. YG is funded by the Google Fellowship. Additional funding came from ONR MURI.

References

[1] 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*, pages 4432–4441, 2019.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (TOG)*, 38(4):1–11, 2019.

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[5] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. *arXiv preprint arXiv:2403.17064*, 2024.

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[7] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

[8] Anton Cherepkov, Andrey Voynov, and Artem Babenko. Navigating the gan parameter space for semantic image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3671–3680, 2021.

[9] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*, pages 484–498. Springer, 1998.

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

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2016.

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

https://doi.org/10.52202/079017-4363

[14] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14300–14310, 2023.

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[16] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

[17] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023.

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

[19] David Ha, Andrew M Dai, and Quoc V Le. Hypernetworks. In *International Conference on Learning Representations*, 2016.

[20] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

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

[23] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[24] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2022.

[25] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023.

[26] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[27] 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*, 34:852–863, 2021.

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

[29] 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*, pages 8110–8119, 2020.

[30] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

https://doi.org/10.52202/079017-4363

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[32] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

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

[34] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[35] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2022.

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

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

[39] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[40] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022.

[41] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

[42] William Peebles, Ilija Radosavovic, Tim Brooks, Alexei A Efros, and Jitendra Malik. Learning to learn with generative models of neural network checkpoints. *arXiv preprint arXiv:2209.12892*, 2022.

[43] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit Bermano, Eric Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. In *Computer Graphics Forum*, volume 43, page e15063. Wiley Online Library, 2024.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[45] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[46] Ilija Radosavovic, Justin Johnson, Saining Xie, Wan-Yen Lo, and Piotr Dollár. On network design spaces for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1882–1890, 2019.

https://doi.org/10.52202/079017-4363

[47] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020.

[48] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

[49] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[50] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[51] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022.

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[53] Duncan A Rowland and David I Perrett. Manipulating facial appearance through shape and color. *IEEE computer graphics and applications*, 15(5):70–76, 1995.

[54] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[55] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6527–6536, 2024.

[56] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023.

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[58] Konstantin Schürholt, Dimche Kostadinov, and Damian Borth. Self-supervised representation learning on neural network weights for model characteristic prediction. *Advances in Neural Information Processing Systems*, 34:16481–16493, 2021.

[59] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023.

[60] 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*, pages 9243–9252, 2020.

[61] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. *Josa a*, 4(3):519–524, 1987.

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[63] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

https://doi.org/10.52202/079017-4363

[64] 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.

[65] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[66] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020.

[67] Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv e-prints*, pages arXiv–2404, 2024.

[68] Kuan-Chieh Wang, Paul Vicol, James Lucas, Li Gu, Roger Grosse, and Richard Zemel. Adversarial distillation of bayesian neural network posteriors. In *International conference on machine learning*, pages 5190–5199. PMLR, 2018.

[69] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[70] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3121–3138, 2022.

[71] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

[72] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[73] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *Advances in Neural Information Processing Systems*, 36, 2024.

[74] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

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

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[77] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 597–613. Springer, 2016.

[78] Ye Zhu, Yu Wu, Zhiwei Deng, Olga Russakovsky, and Yan Yan. Boundary guided learning-free semantic control with diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

https://doi.org/10.52202/079017-4363

A Sampling

We present additional examples of models sampled from *w2w* space in Fig. 13. The sampled models encode a diverse array of identities which are not copied from the dataset of model weights, as seen by comparing them to the nearest neighbor models from the training set. However, there are attributes borrowed from the nearest neighbors which visually appear in the sampled identity. For instance, the sampled man in the first row shares a similar jawline to the nearest neighbor identity. The sampled identities also demonstrate the same ability as the original training identities to be composed into novel contexts. A variety of prompts are used in Fig. 13 and the identities are still consistent.

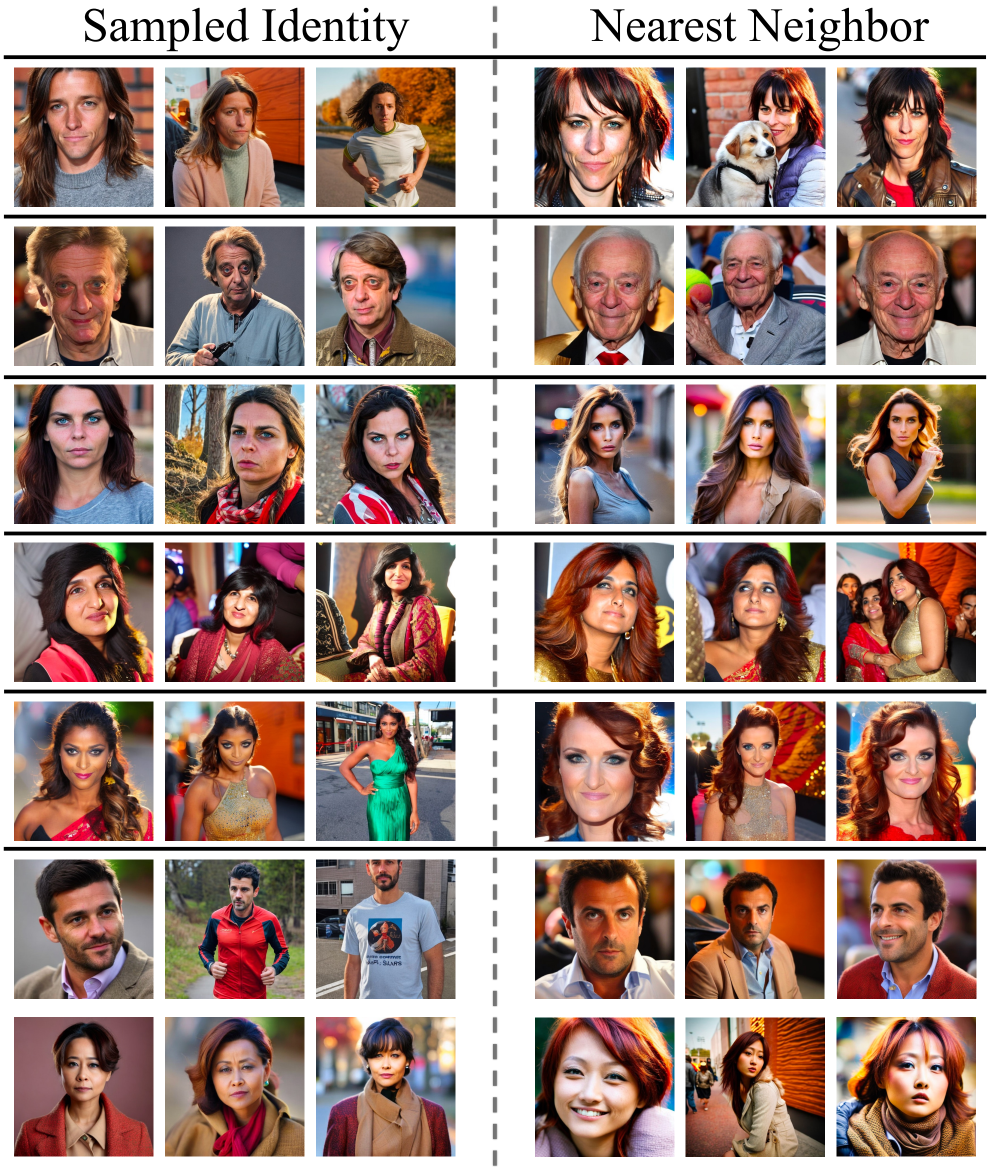


Figure 13: **Sampled identity-encoding models from *w2w* space and their nearest neighbor models.** The sampled identities share some characteristics with the nearest neighbors, but are still distinct. These identities can also be composed into novel contexts like a standard customized diffusion model.

https://doi.org/10.52202/079017-4363

B Model Editing in *w2w* Space

Qualitative Results. We display additional examples of applying edits in *w2w* space based on the directions discovered using linear classifiers and CelebA labels. In Fig. 14, we demonstrate how the strength of these edits can be modulated and combined with minimal interference. These edits are apparent even in more complex scenes beyond face images. Also, the edits do not degrade other present concepts, such as the dog near the man (top left example).

In Figs. 15 and 16, we demonstrate how multiple edits can be progressively added in a disentangled fashion with minimal degradation to the identity. Additionally, since we operate in a subspace of weight space, these edits persist with a consistent appearance across different generations. For instance, even the man exhibits the edits as a painting in Fig. 15.

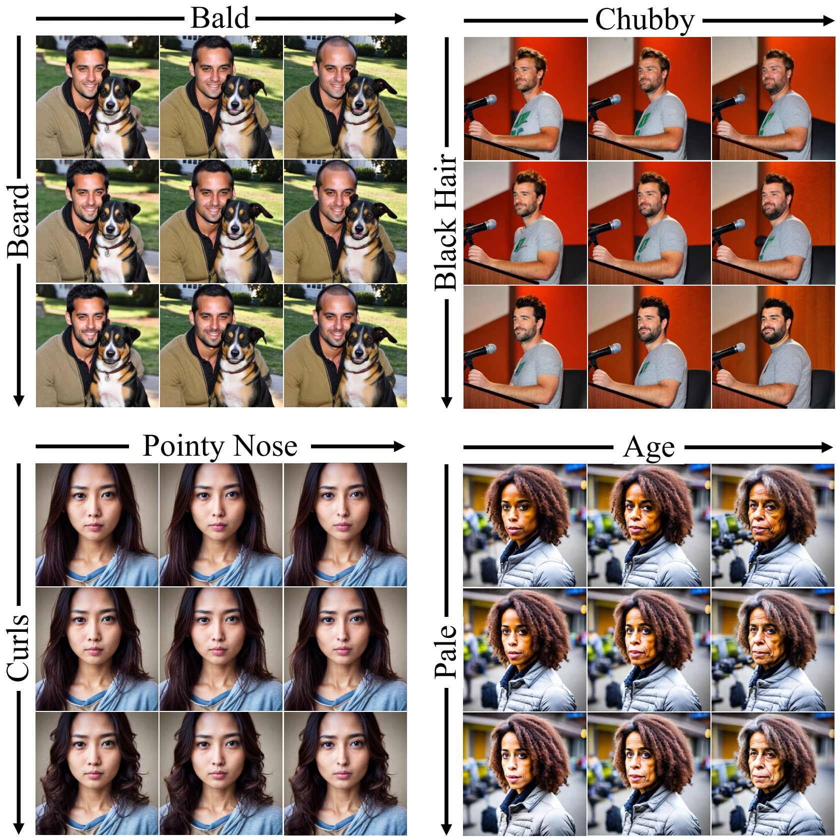


Figure 14: **Multiple edits can be controlled in a continuous manner.**

https://doi.org/10.52202/079017-4363

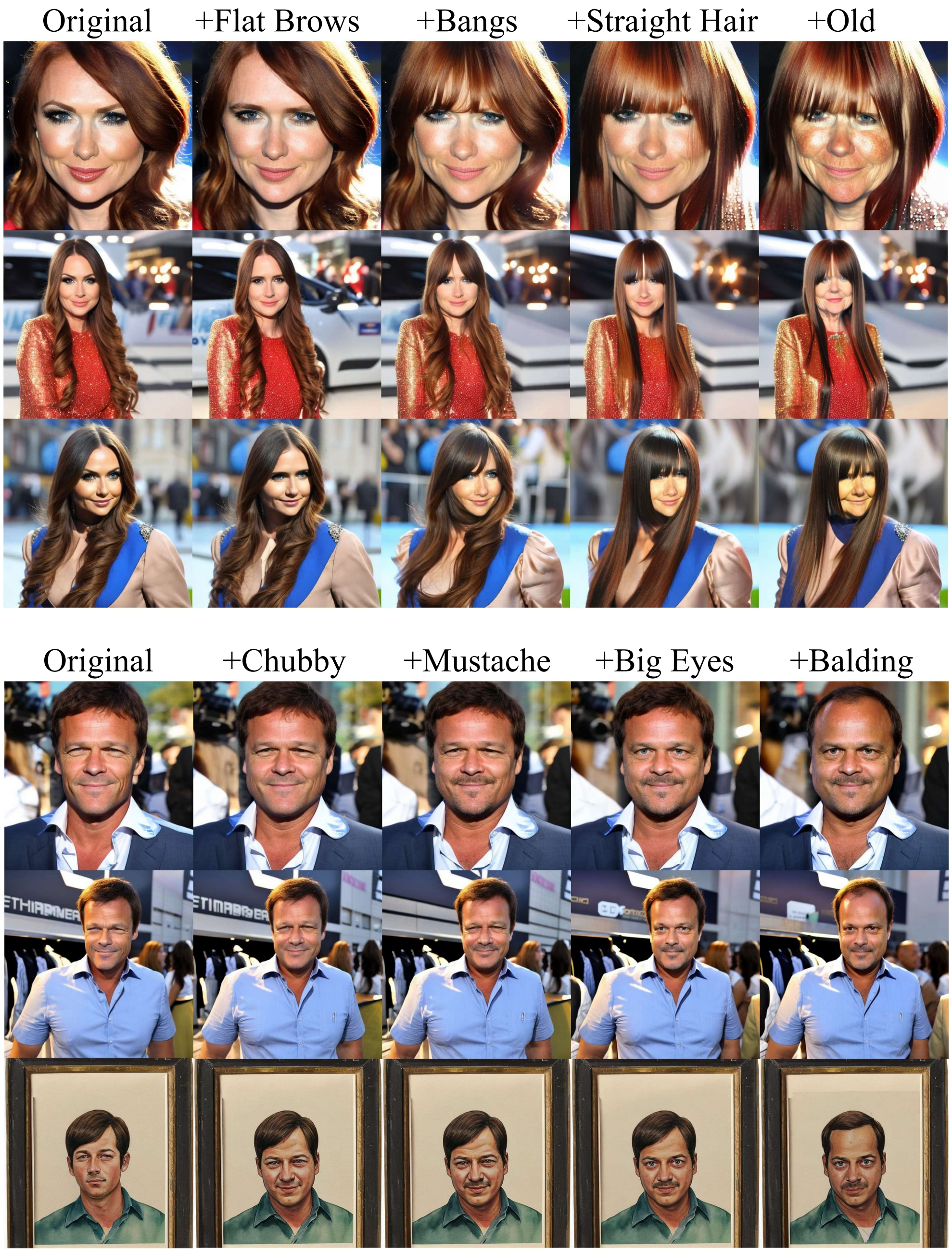


Figure 15: **Composing four different edits with minimal identity degradation**. These edits bind to the identity and persist in appearance across multiple generation seeds and prompts.

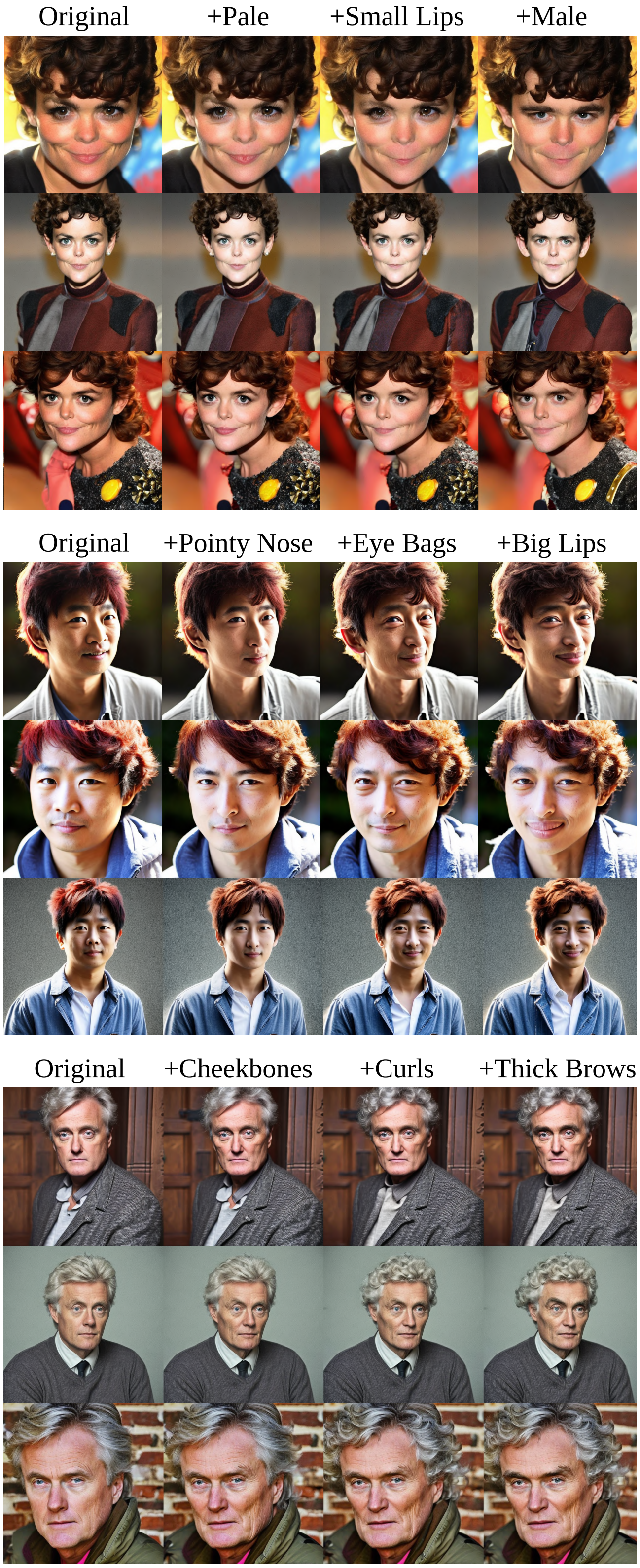
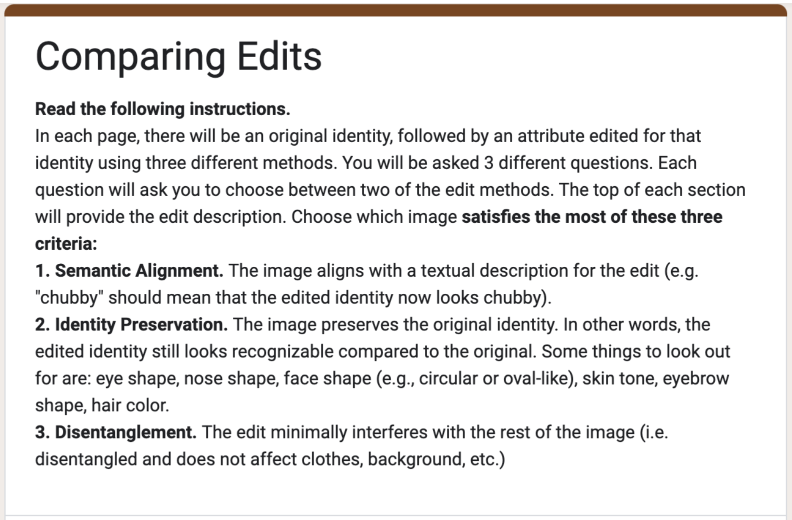


Figure 16: **Additional examples of composing multiple edits.** We provide more examples of semantic edits based on labels available from CelebA.

User Study. We present a two-alternative forced choice (2AFC) user study to evaluate the quality of identity edits in Tab. 3. Twenty-five users were given ten sets of images. Each set contained a randomly sampled original image of an identity, and then an image of that identity edited for an attribute using Concept Sliders [17], *w2w*, and text prompting. Users were then asked to choose between alternate pairs based on three criteria: identity preservation, alignment with the desired edit, and disentanglement. Our results in Tab. 3 show that users have a strong preference towards *w2w* edits. User instructions and an example question from the study are provided in Figs. 17, 18.

Table 3: **User study on identity editing.**

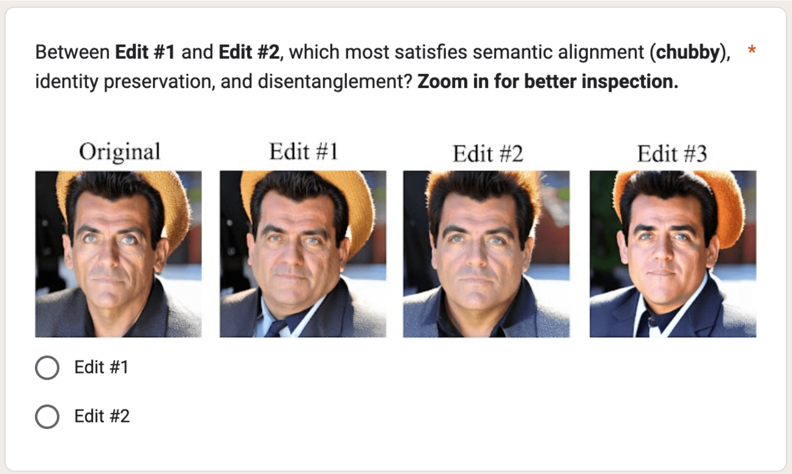
Method	Win Rate (%) ↑
Sliders	28.4
w2w	**71.6**
Prompting	12.8
w2w	**87.2**



Comparing Edits

Read the following instructions.
In each page, there will be an original identity, followed by an attribute edited for that identity using three different methods. You will be asked 3 different questions. Each question will ask you to choose between two of the edit methods. The top of each section will provide the edit description. Choose which image **satisfies the most of these three criteria:**
1. Semantic Alignment. The image aligns with a textual description for the edit (e.g. "chubby" should mean that the edited identity now looks chubby).
2. Identity Preservation. The image preserves the original identity. In other words, the edited identity still looks recognizable compared to the original. Some things to look out for are: eye shape, nose shape, face shape (e.g., circular or oval-like), skin tone, eyebrow shape, hair color.
3. Disentanglement. The edit minimally interferes with the rest of the image (i.e. disentangled and does not affect clothes, background, etc.)

Figure 17: **Instructions provided to users for the identity editing user study.**



Between **Edit #1** and **Edit #2**, which most satisfies semantic alignment (**chubby**), identity preservation, and disentanglement? **Zoom in for better inspection.** *

○ Edit #1

○ Edit #2

Figure 18: **Example question from the identity editing user study.**

C Inversion

We present additional details on *w2w* inversion and comparisons against training Dreambooth LoRA on a single image vs. multiple images.

Implementation Details: To conduct *w2w* inversion, we train on a single image following the objective from eq. 3. We qualitatively find that optimizing 10,000 principal component coefficients balances identity preservation with editability. This is discussed in Appendix F. We optimize for 400 epochs, using Adam [30] with learning rate 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and with weight decay factor $1e$-10. For conducting Dreambooth fine-tuning, we follow the implementation from Hugging Face [3] using LoRA with rank 1. To create a dataset of multiple images for an identity, we follow the procedure from Sec. 4.4.

***w2w* inversion is more efficient than previous methods.** Inversion into *w2w* space results in a significant speedup in optimization as seen in Tab. 4, where we measure the training time on a single NVIDIA A100 GPU. Standard Dreambooth fine-tuning operates on the full weight space and incorporates an additional prior preservation loss which typically requires hundreds of prior images. In contrast, we only optimize a standard denoising objective on a single image within a low-dimensional weight subspace. Despite operating with lower dimensionality, *w2w* inversion performs closely to standard Dreambooth fine-tuning on multiple images with LoRA.

Table 4: **Inversion into *w2w* space balances identity preservation and efficiency.**

Method	Single-Image	# Param	Opt. Time (s)	Identity Fidelity ↑
DB-LoRA	×	99,648	220	**0.69** ± 0.01
DB-LoRA	✓	99,648	200	0.43 ± 0.03
w2w Inversion	✓	10,000	55	0.64 ± 0.01

Qualitative Inversion Comparison. In Figs. 19 and 20, we present qualitative comparisons of *w2w* inversion against Dreambooth trained with multiple images and a single image. Although *mult-image* Dreambooth slightly outperforms *w2w* inversion in identity preservations, its samples tend to lack realism compared to *w2w* inversion. We hypothesize that this may be due to using generated images for prior preservation and training on synthetic identity images. Dreambooth trained on a single image either generates an artifacted version of the original image or random identities. Notice how inversion into *w2w* space is even able to capture key characteristics of the child although babies are nearly to completely absent in the identites based on CelebA used to fine-tune our dataset of models.

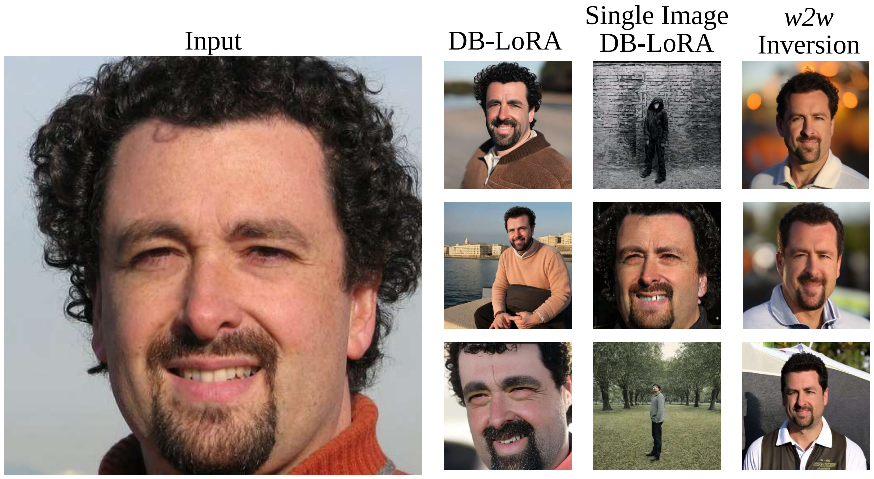


Figure 19: **Inversion into *w2w* space preserves identity and realism.** We compare against Dreambooth fine-tuning with LoRA on multiple images and a single image.

[3] https://github.com/huggingface/peft

https://doi.org/10.52202/079017-4363

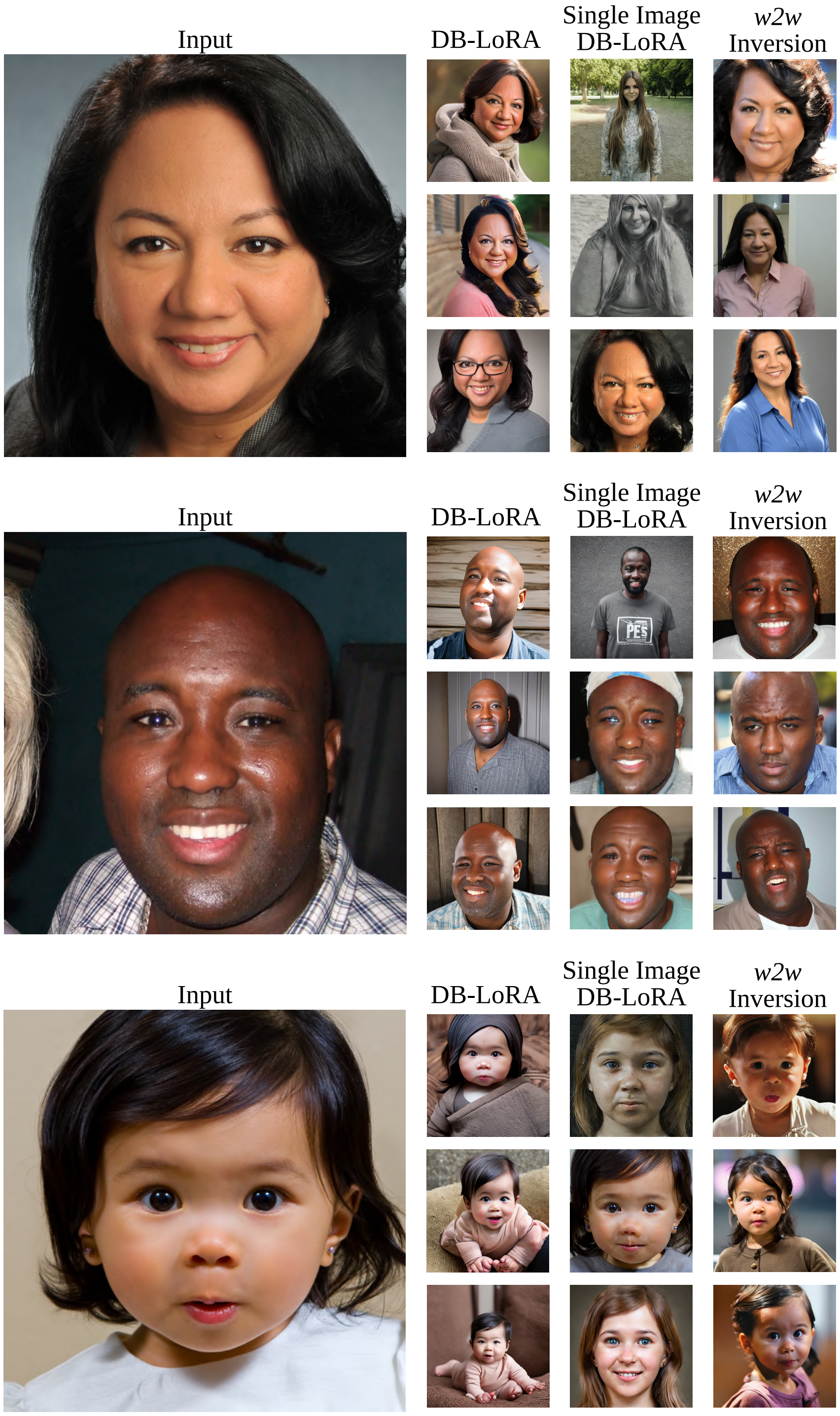


Figure 20: **Inversion into *w2w* space preserves identity and realism (cont.).**

Comparison against Single Image Personalization Methods. We compare *w2w* inversion to single shot personalization methods Celeb-Basis [73] and IP-Adapter FaceID[4] [72], following the same evaluation protocol from Sec. 4.4. These quantitative results are presented in Tab. 5. While Celeb-Basis optimizes in the input text embedding space for a given face image, IP-Adapter trains light-weight adapters [39, 75] to condition generation on any input face image.

We further conducted a two-alternative forced choice (2AFC) user study on the perceptual quality of *w2w* identity preservation. Twenty users were given ten sets of images. Each set contained a randomly sampled original image of the identity, and then three random images generated using IP-Adapter FaceID, Celeb-Basis, and *w2w* with the same random prompt. Users were then asked to choose between alternate pairs based on three criteria: identity preservation, prompt alignment, and diversity of generated images. Our results in Tab. 6 show that users found generations from *w2w* models capture identity better while also generating more diverse images that better align with the prompts.

Across both these metrics, *w2w* performs stronger than Celeb-Basis and IP-Adapter FaceID. Our results indicate that operating in our weight subspace is highly expressive and flexible as it is able to faithfully capture nuanced identity without overfitting to the input image. For instance, in Fig. 21, *w2w* inversion enables diverse generation with various poses, facial expressions, and clothing while maintaining identity.

Table 5: **Single-shot personalization comparison.**

Method	ID Score ↑
Celeb-Basis	0.60 ± 0.02
IP-Adapter FaceID	0.62 ± 0.02
w2w	$\mathbf{0.64} \pm 0.01$

Table 6: **User study on identity inversion.**

Method	Win Rate (%) ↑
Celeb-Basis	13.4
w2w	**86.6**
IP-Adapter FaceID	22.8
w2w	**77.2**

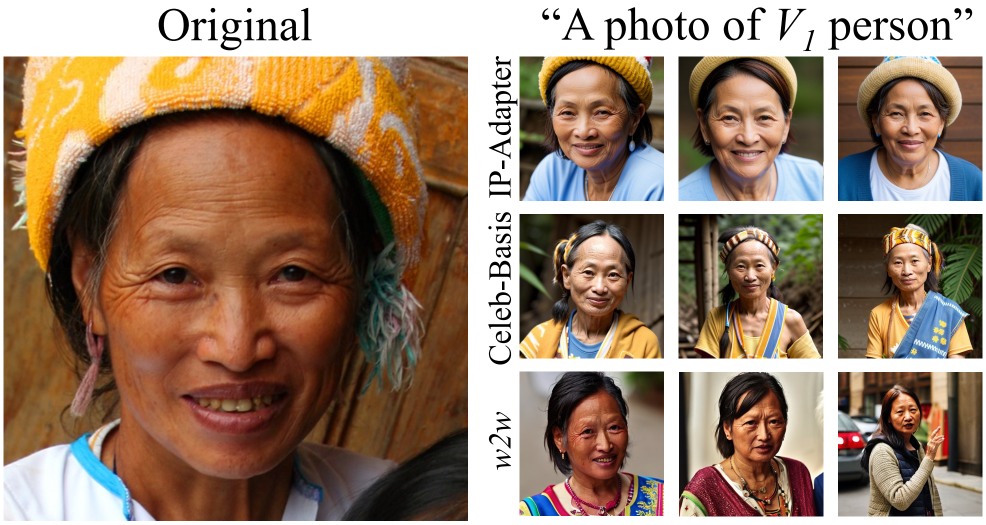
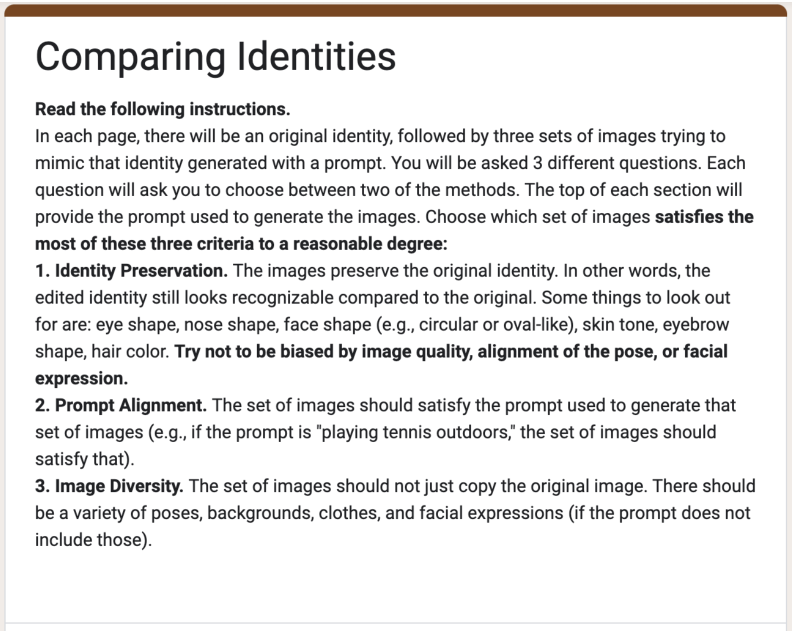


Figure 21: **Qualitative comparison of single-shot personalization methods.**

[4] https://huggingface.co/h94/IP-Adapter-FaceID

https://doi.org/10.52202/079017-4363

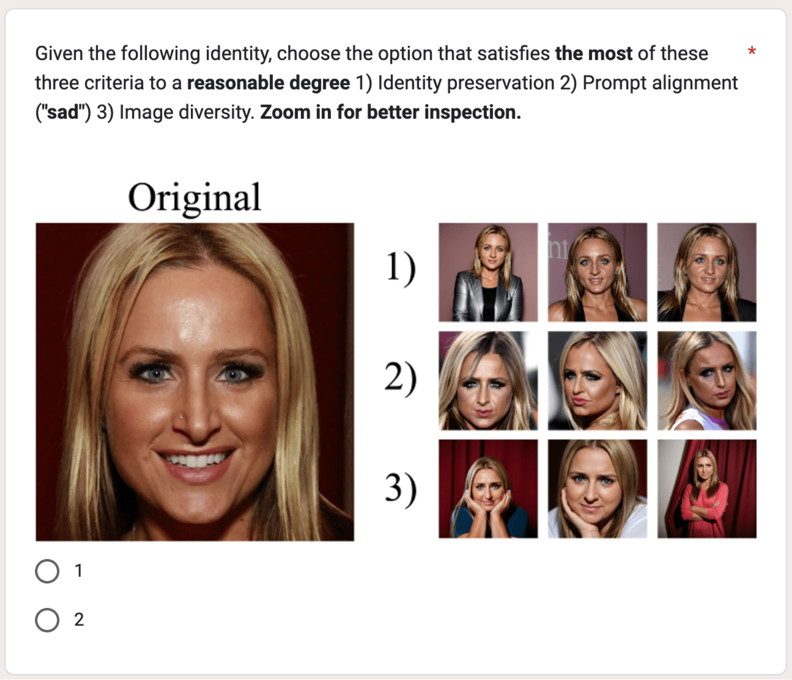
Below in Figs. 22 and 23, we provide the instructions provided to users for the identity inversion user study in addition to an example question.



Comparing Identities

Read the following instructions.
In each page, there will be an original identity, followed by three sets of images trying to mimic that identity generated with a prompt. You will be asked 3 different questions. Each question will ask you to choose between two of the methods. The top of each section will provide the prompt used to generate the images. Choose which set of images **satisfies the most of these three criteria to a reasonable degree:**
1. Identity Preservation. The images preserve the original identity. In other words, the edited identity still looks recognizable compared to the original. Some things to look out for are: eye shape, nose shape, face shape (e.g., circular or oval-like), skin tone, eyebrow shape, hair color. **Try not to be biased by image quality, alignment of the pose, or facial expression.**
2. Prompt Alignment. The set of images should satisfy the prompt used to generate that set of images (e.g., if the prompt is "playing tennis outdoors," the set of images should satisfy that).
3. Image Diversity. The set of images should not just copy the original image. There should be a variety of poses, backgrounds, clothes, and facial expressions (if the prompt does not include those).

Figure 22: **Instructions provided to users for the identity inversion user study.**



Given the following identity, choose the option that satisfies **the most** of these three criteria to a **reasonable degree** 1) Identity preservation 2) Prompt alignment (**"sad"**) 3) Image diversity. **Zoom in for better inspection.** *

○ 1

○ 2

Figure 23: **Example question from the identity inversion user study.**

https://doi.org/10.52202/079017-4363

D Out of Distribution Projection

Additional examples of out-of-distribution projections are displayed in Fig. 24. A diverse array of styles and subjects (e.g. paintings, sketches, non-humans) can be distilled into a model in *w2w* space. After embedding an identity into this space, the model still retains the compositionality and rich priors of a standard personalized model. For instance, we can generate images using prompts such as “*[v]* person writing at a desk” (top example), “*[v]* person with a dog” (middle example), or “a painting of *[v]* person painting on a canvas” (bottom example).

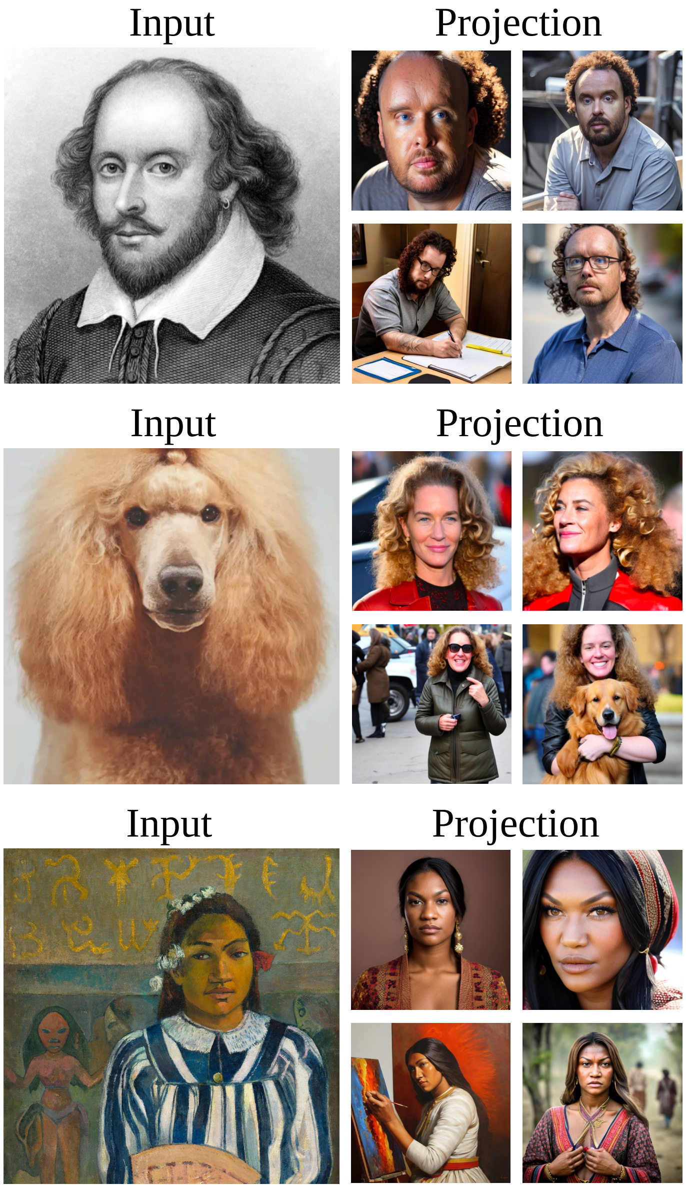


Figure 24: **Projection into *w2w* space generalizes to a variety of inputs.** A range of styles and entities can be inverted into a realistic identity in this space. Once a model is obtained, it can be prompted to generate the identity in a variety of contexts.

E Identity Datasets

In Fig. 25, we present examples of synthetic identity datasets used to conduct our Dreambooth fine-tuning as discussed in Sec 4. Each dataset is a set of ten images generated with [67] conditioned on single CelebA [36] images associated with binary attribute labels. Note that we only display a subset of images per identity in the figure. The same identity can occur multiple times in different images in CelebA but have different appearances. So, creating these synthetic datasets reduces intra-dataset diversity and creates a more consistent appearance for each subject. For instance, the first two rows in the figure are the same identity, but look drastically different. So we instead treat them as different identities associated with a different set of images.

For evaluating identity edits from Sec. 4.3, we hold out 100 identities, which results in leaving out ∼1000 models since multiple models may encode different instances of the same identity. For instance, if we left out the model encoding the identity in the first row of Fig. 25 for evaluation, the model encoding the second row identity would also be left out since it encodes the same identity but a different instance.

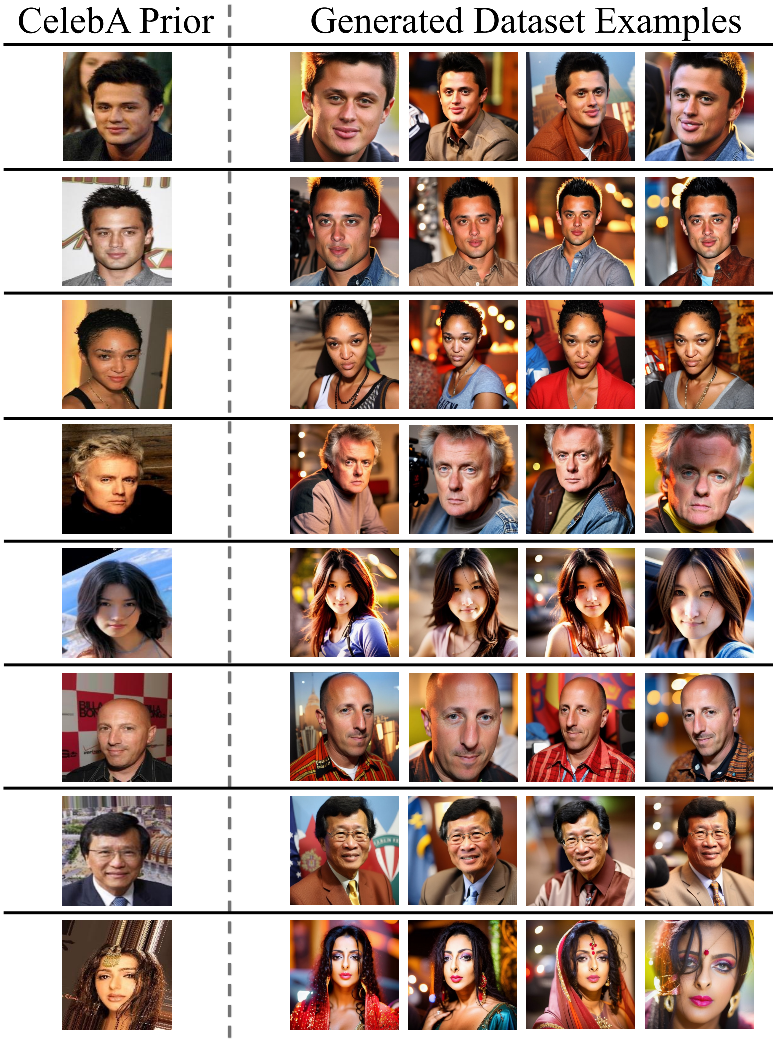


Figure 25: **Fine-tuning on synthetic examples allows Dreambooth fine-tuning to distill a consistent identity.** The left column shows a CelebA image used to condition generation of a set of identity-consistent images in the right column associated with that identity using [67]. The consistent appearance of the identity enables a more consistent identity encoding.

https://doi.org/10.52202/079017-4363

F Principal Component Basis

In this section, we analyze various properties of the Principal Component (PC) basis used to define *w2w* Space. We investigate the distribution of PC coefficients and the effect of the number of PCs on identity editing and inversion.

Distribution of PC Coefficients. We plot the histogram of the coefficient values for the first three Principal Components in Fig. 26. They appear roughly Gaussian. Next, we rescale the coefficients for these three components to unit variance for visualization purposes. We then plot the pairwise joint distributions for them in Fig. 27. The circular shapes indicates roughly diagonal covariances. Although the joint over other combinations of Principal Components may exhibit different properties, these results motivate us to model the PCs as independent Gaussians, leading to the *w2w* sampling strategy from Sec. 3.2.

Number of Principal Components for Identity Editing We empirically observe that training classifiers based on the 1000 dimensional PC representations (first 1000 PCs) of the model weights results in the most semantically aligned and disentangled edits directions. We visualize a comparison for the "goatee" direction in Fig. 28. After finding the direction, we calculate the maximum projection component onto the edit direction among the training set of model weights. This determines the edit strength. As seen in the figure, restricting to the first 100 Principal Components may be too coarse to achieve the fine-grained edit, instead relying on spurious cues such as skin color. Training with the first 10,000 Principal Components suffers from the curse of dimensionality and the discovered direction may edit other concepts such as eye color or clothes. Finding the direction using the first 1000 Principal Components achieves the desired edit with minimal entanglement with other concepts.

Number of Principal Components for Identity Inversion We qualitatively observe that inversion using the first 10,000 Principal Components balances identity preservation while not overfitting to the source image. We visualize a comparison in Fig. 29, where each column has a fixed seed and prompt. Optimizing with the first 1000 PCs underfits the identity and does not generate a consistent identity. Inversion with the first 20,000 Principal Components overfits to the source image of a face shot, which results in artifacted face images despite different generation seeds and prompts. Optimizing with the first 10,000 Principal Components enjoys the benefits of a lower dimensional representation than the original LoRA parameter space ($\sim$100,000 trainable parameters), while still preserving identity and compositionality. This is supported quantitatively by Fig. 30, which shows the average ID score for 100 inverted FFHQ identities optimized over a varying number of principal components.

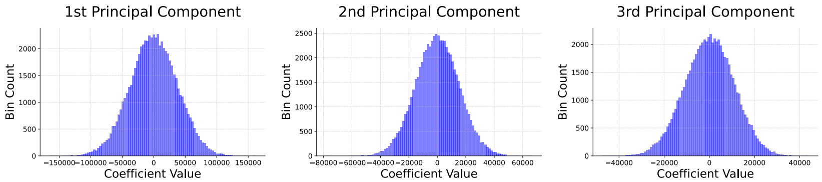


Figure 26: **Histogram of principal component coefficients.** The first three principal component coefficients appear approximately Gaussian.

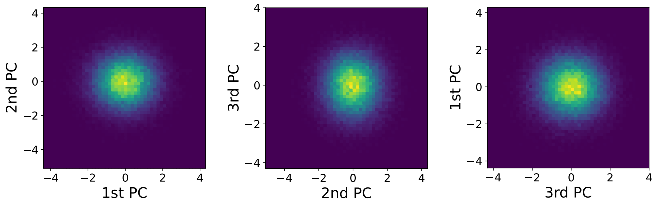


Figure 27: **Pairwise joint histogram of principal component coefficients.** We rescale the first three principal component coefficients and plot the pairwise joint distributions for visualization purposes. Given that the marginals are roughly Gaussian, the circular appearance of the joint suggests pairwise independence for the first three components.

https://doi.org/10.52202/079017-4363

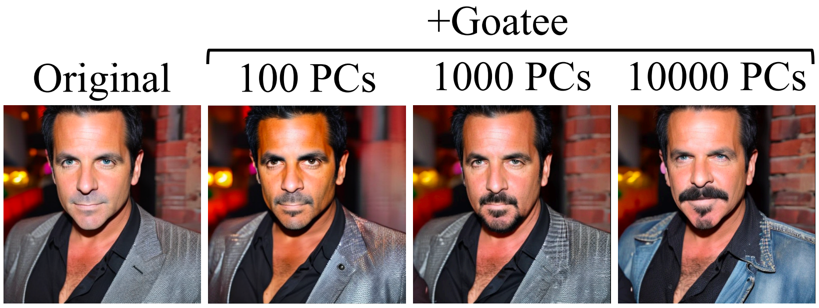


Figure 28: **Edit results with varying number of Principal Components.** Training classifiers to find semantic weight space directions with the first 1000 Principal Components achieves the most semantically aligned and disentangled results.

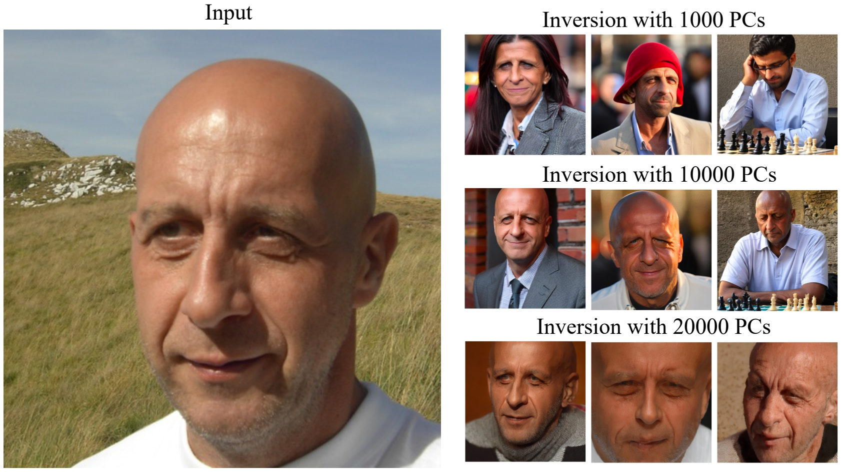


Figure 29: **Identity inversion results with varying number of principal components.** We optimize the coefficients for the first 1000, $10,000$, and $20,000$ Principal Component. Each column indicates a fixed generation seed and prompt. Inversion with the first $10,000$ components balances parameter efficiency, realism, and identity preservation without overfitting to the single image.

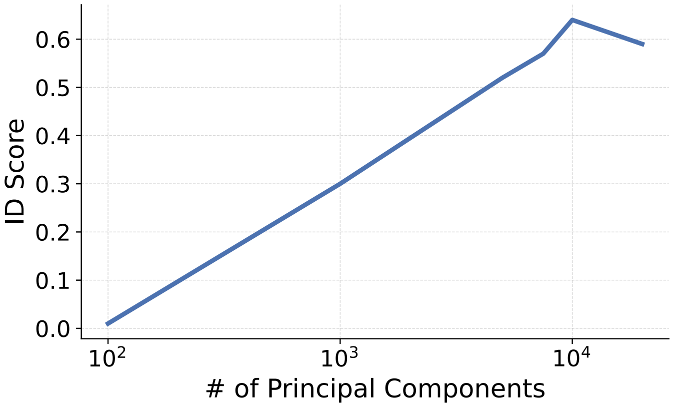


Figure 30: **Identity preservation vs. number of principal components used for *w2w* inversion.** We optimize the coefficients for the first N principal components up to 20,000 and measure the average ID score for 100 inverted FFHQ identities.

https://doi.org/10.52202/079017-4363

Visualizing Principal Components. We provide a visualization of traversals along a set of principal components in Fig. 31. The principal components change attributes of the identity, although various semantic attributes are entangled. For instance, the first PC appears to change age, hair color, and hair style. The second PC appears to change gender and skin complexion. The third PC seems to change age, skin complexion, and facial hair. This motivates our use of linear classifiers to find separating hyperplanes in weight space and disentangle these attributes.

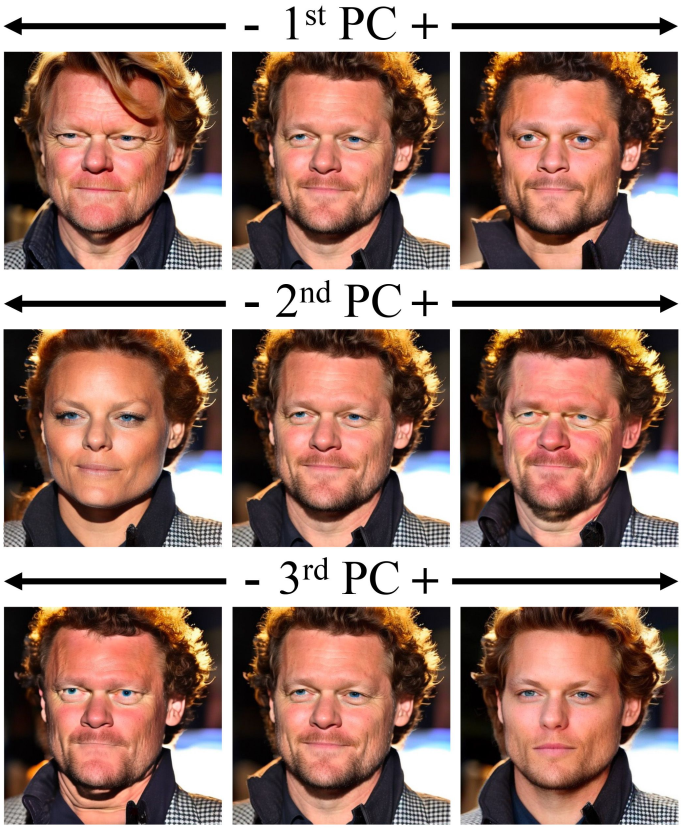


Figure 31: **Traversal along the first three principal components in *w2w* space.** The directions encode various entangled identity attributes such as age, gender, and facial hair.

https://doi.org/10.52202/079017-4363

G *weights2weights* for Other Visual Concepts

We find that similar subspaces can be created for other visual concepts beyond human identites. For instance, we apply the *weights2weights* framework to create two subspaces for models encoding different dog breeds and models encoding car types. We present examples of editing these models in Fig. 32. This suggest the generality of *weights2weights* and linear subspaces within diffusion model weights.

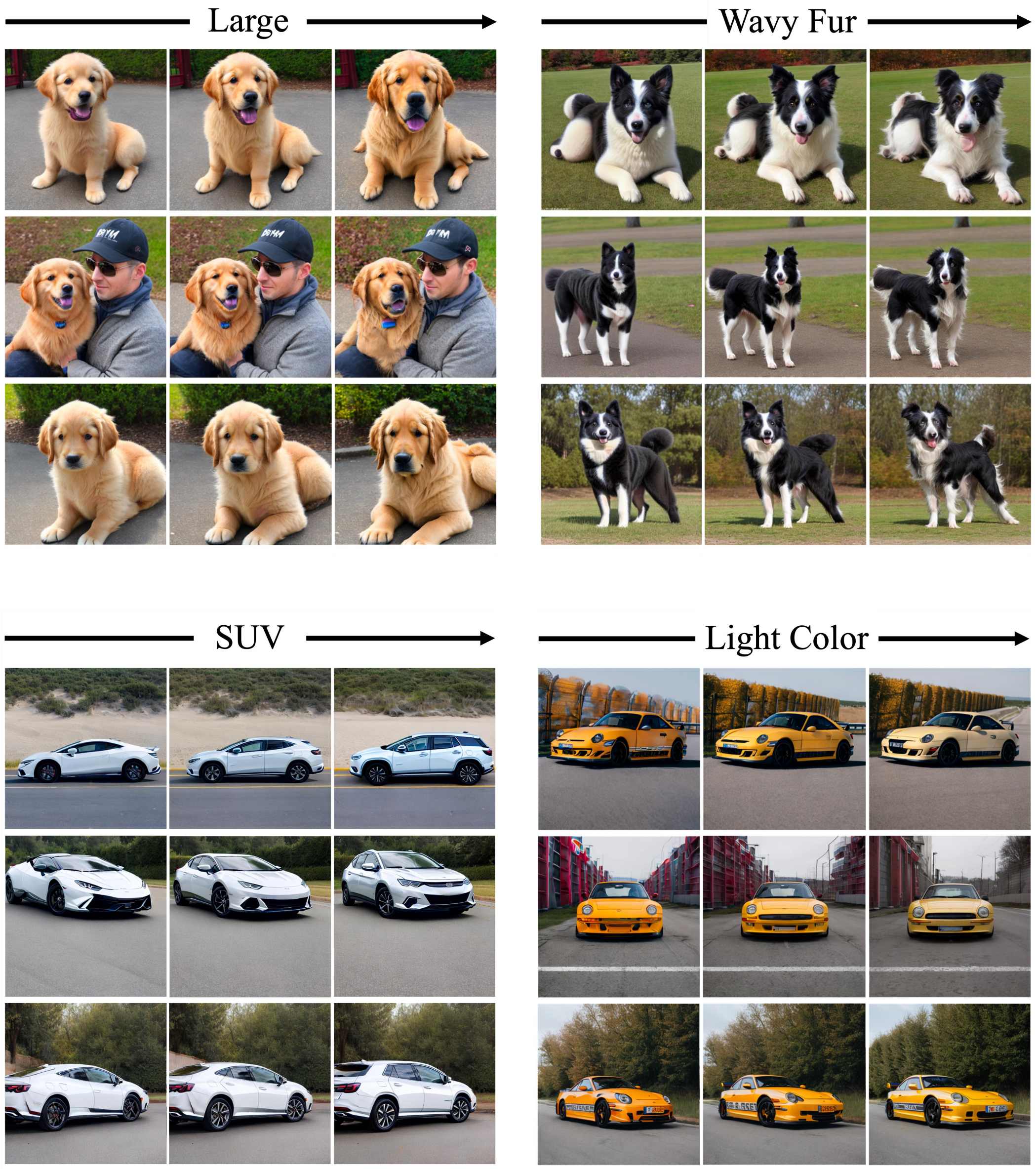


Figure 32: **Applying *weights2weights* edits on dog-encoding models and car-encoding models.** We create two datasets of model weights and apply PCA to define two separate weight subspaces. We then train linear classifiers to find semantic edit directions.

H Multi-Concept Merging

Multiple models living in *w2w* space cannot be merged since they live in the same weight subspace. So, merging will lead to interpolation of the identities. However, *w2w* models can be merged with models lying approximately orthogonal to the subspace. This merging can be done adding the weights. We present an example of merging a model living in *w2w* space with a model fine-tuned to encode “Pixar" style in Fig. 33.

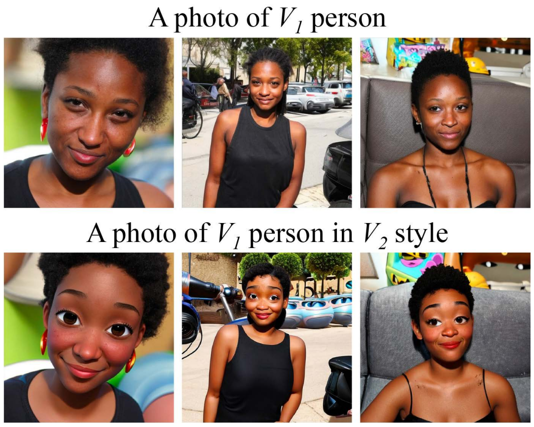


Figure 33: **Merging *w2w* models with non-identity models.** Here, another model is fine-tuned to map V_2 to “Pixar” style. The two models are merged with simple addition.

I Timestep Analysis

Edits in *w2w* space correspond to identity edits with minimal interference with other visual concepts. Although not a focus, image editing is achieved as a byproduct. For further context preservation, edits in *w2w* Space can be integrated with delayed injection [7, 17, 35, 71], where after T timesteps, the edited weights are used instead of the original ones. We visualize this in Fig. 34. Larger T in the range $[700, 1000]$ are helpful for more global attribute changes, while smaller $[400, 700]$ can be used for more fine-grained edits. However, by decreasing the timestep T, the strength of the edit is lost in favor of better context preservation. For instance, the dog’s face is better preserved in the second row at $T = 600$, although the man is not as chubby compared to other T.

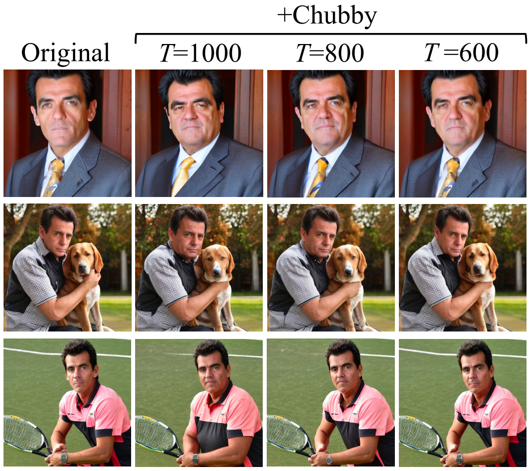


Figure 34: **Injecting edited weights at varying timesteps.** Using the edited weights at a smaller timestep T better preserves context at the expense of edit strength and fidelity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Experiments in the paper back up the claims made in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a dedicated limitations section in Sec. 6 detailing failure cases. Computational efficiency is addressed in Sec. C and scale is addressed in Sec. 4.6. Privacy and fairness is addressed in Sec. 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-4363

Answer: [NA]

Justification: There are no theoretical results in the submission.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Secs. 3, 4 and the Appendix detail all necessary steps to reproduce the results. We plan to release the model weights we trained.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-4363

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code, data, and weights can be found through the project page at: https://snap-research.github.io/weights2weights/

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We present data splits, hyperparameters, etc. in Secs. 3, 4, and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Confidence intervals are reported for the results in the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-4363

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix details the compute resources used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have read the NeurIPS Code of Ethics and confirm that the paper conforms with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts has a dedicated section in Sec. 7.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-4363

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work utilizes pretrained models and well-established datasets that already have proper safeguards in place.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original creators of assets are either cited or linked.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No assets are being submitted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Screenshots and instructions are provided for both identity editing and identity inversion user studies.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Potential risks are not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/079017-4363

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-4363