Interpreting the Weight Space of Customized Diffusion Models

NeurIPS 2024

Amil Dravid* UC Berkeley

Yossi Gandelsman* UC Berkeley

Kuan-Chieh Wang Snap

Rameen Abdal Stanford

Gordon Wetzstein Stanford

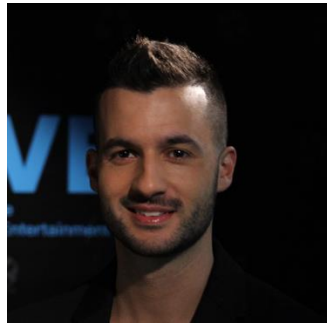
Alyosha Efros UC Berkeley

Kfir Aberman Snap







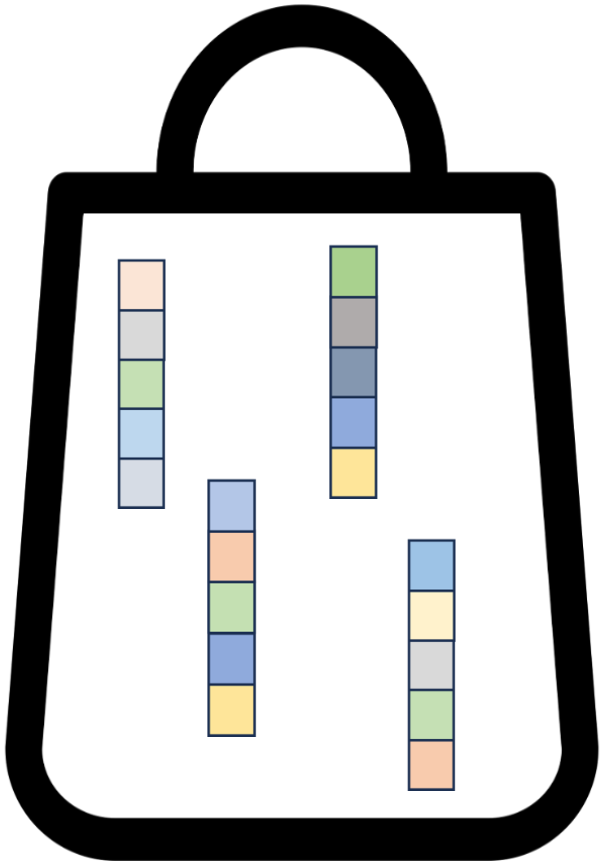


Linear Latent Spaces

Dataset of Images

Interpreting Face Images using Active Appearance Models, Edwards, Taylor, Cootes, 1998
A Morphable Model for the Synthesis of 3D Faces, Blanz and Vetter, 1999

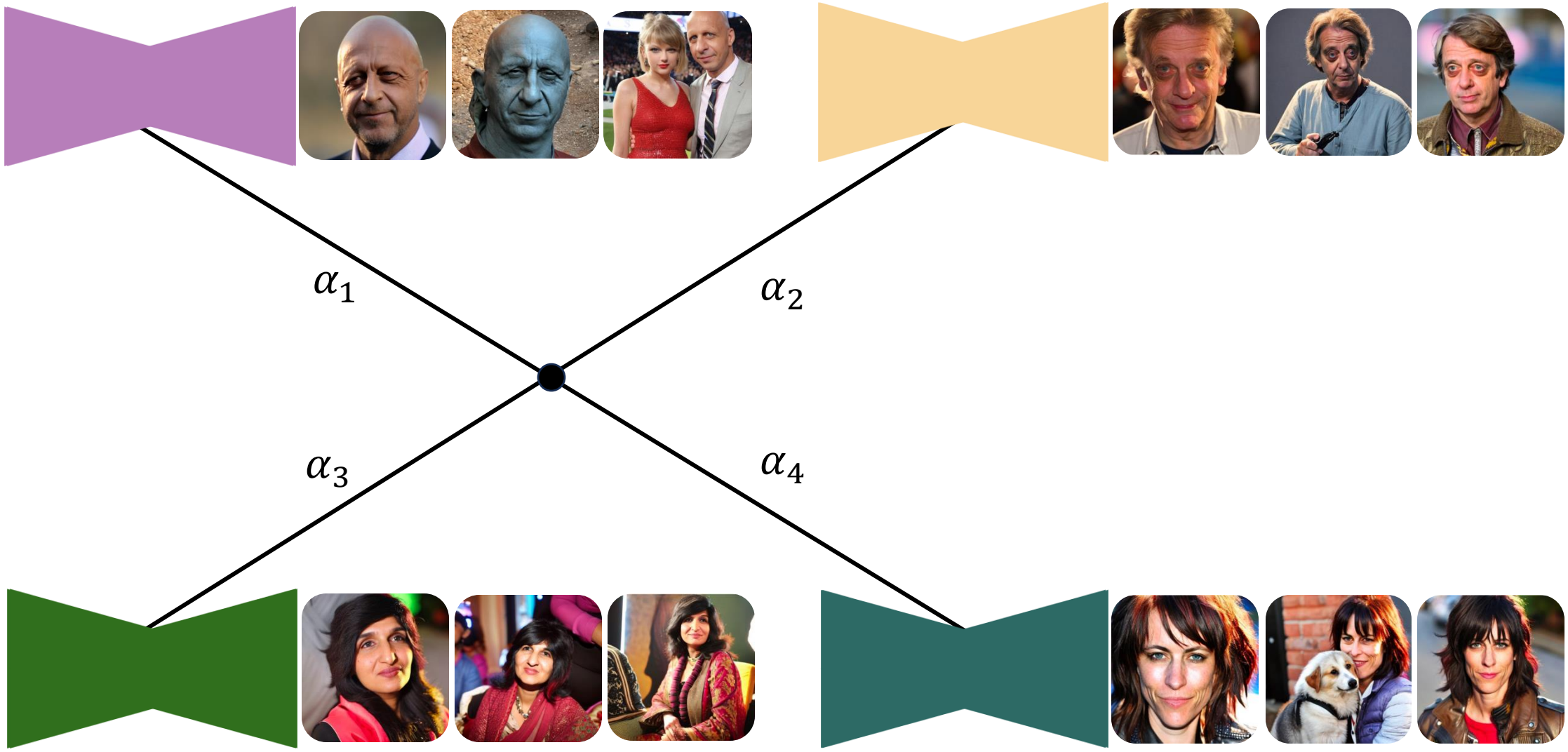
Dataset of Latent Vectors



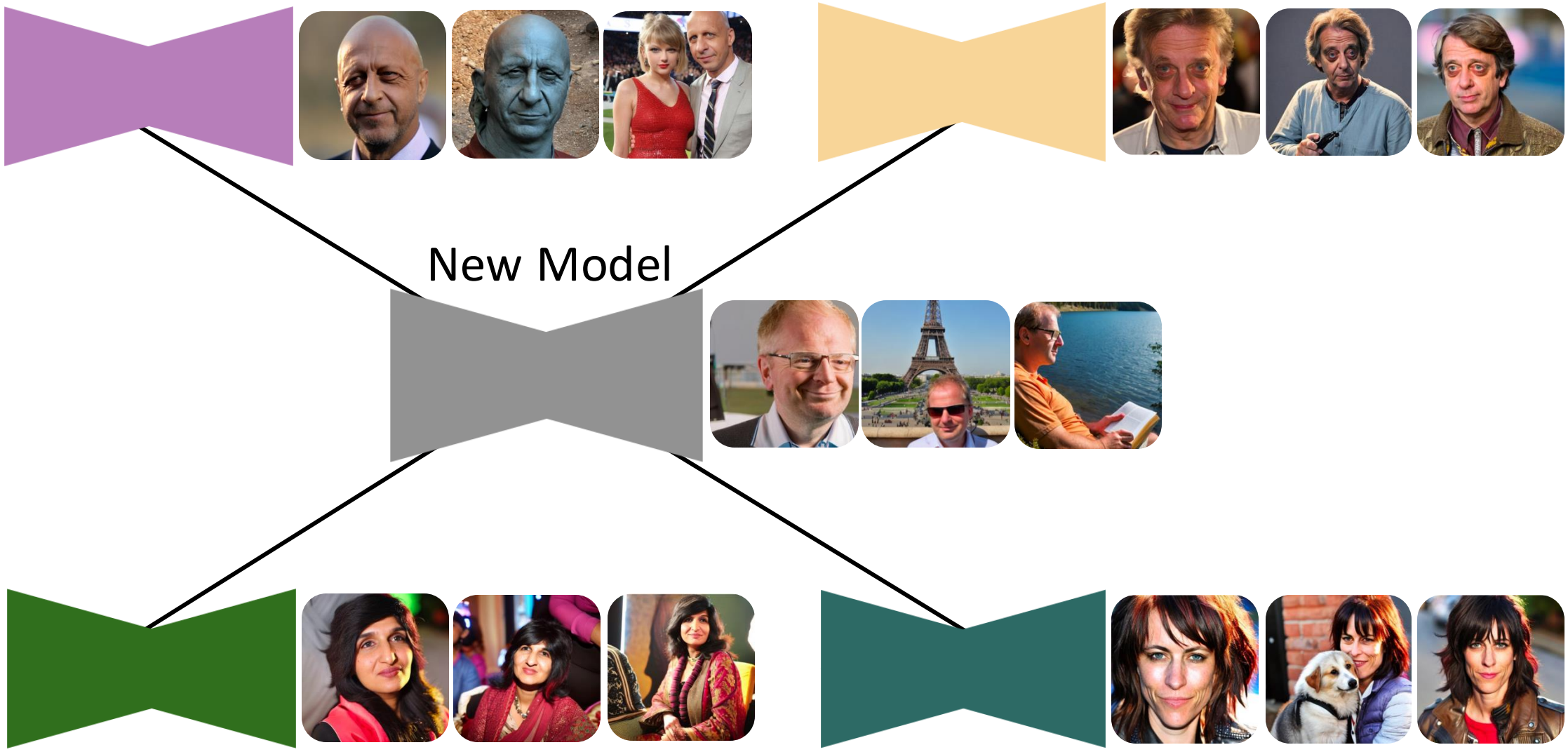
GANSpace: Discovering Interpretable GAN Controls, Härkönen et al., 2020
Interpreting the Latent Space of GANs for Semantic for Semantic Face Editing, Shen et al., 2020

Dataset of Model Weights

Interpolation in Weight Space



Interpolation in Weight Space



How Do We Get These Models?

Fine-Tune for each Subject

Pretrained Diffusion Model

Repeat over 60,000 Times

Small Set of Training Images

Diffusion Unet

Diffusion Unet

Fine-Tune Diffusion Model

Diffusion Unet

Pretrained Diffusion Model

Diffusion Unet

Diffusion Unet

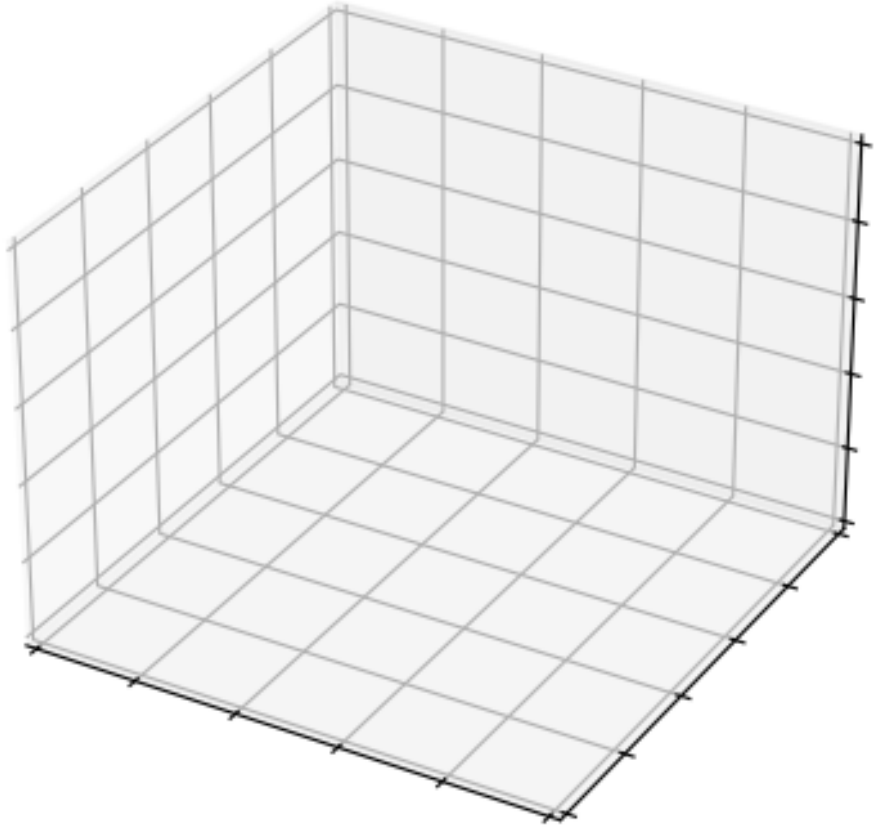
Diffusion Unet

Diffusion Unet

Modeling the Weights Manifold

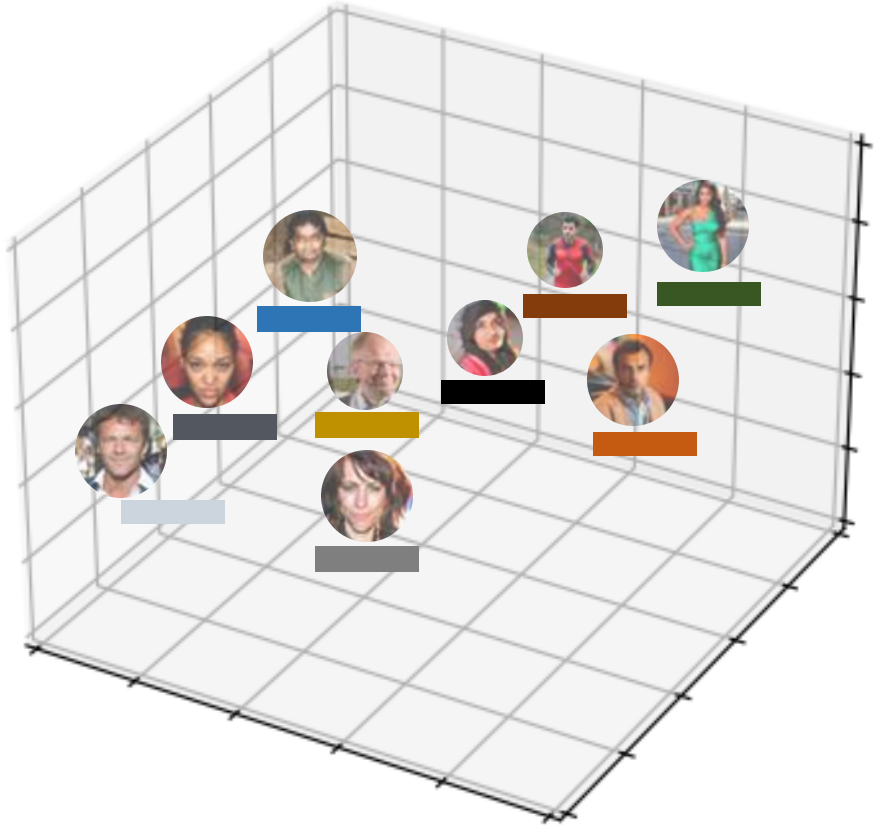
Modeling the Weight Manifold

Model Weight Space



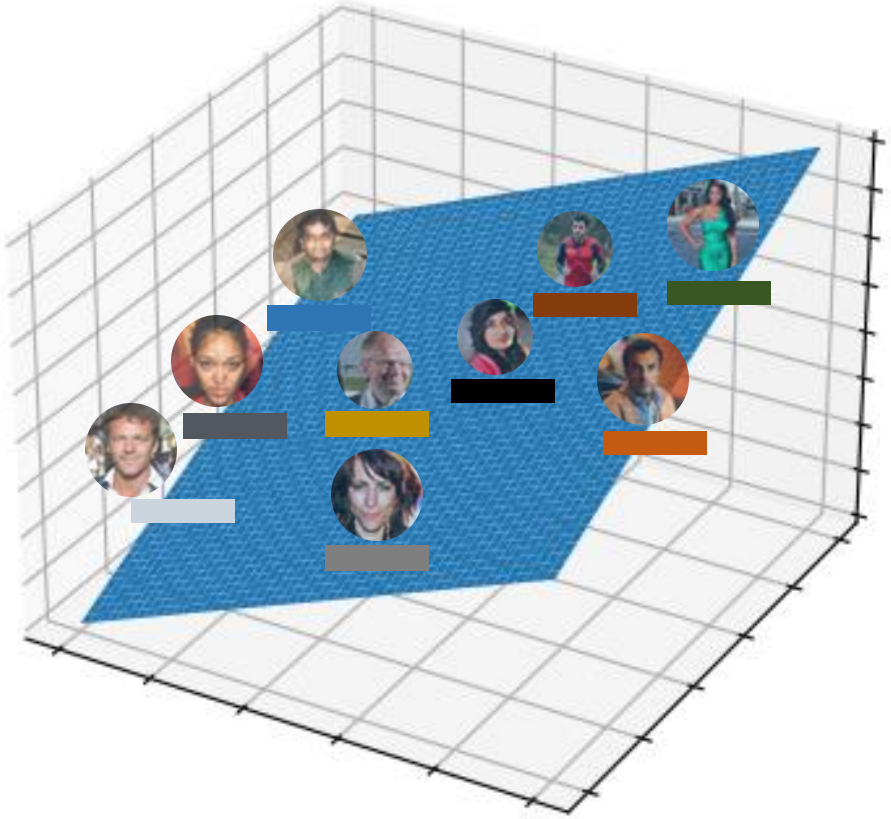
Modeling the Weight Manifold

Model Weight Space

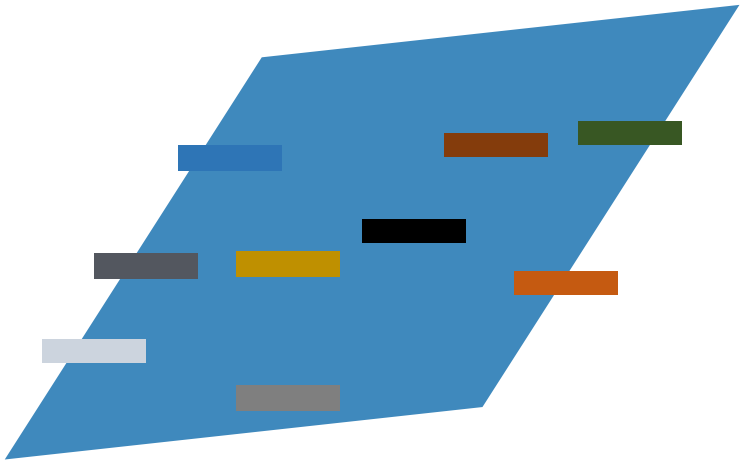


weights2weights (w2w) Space

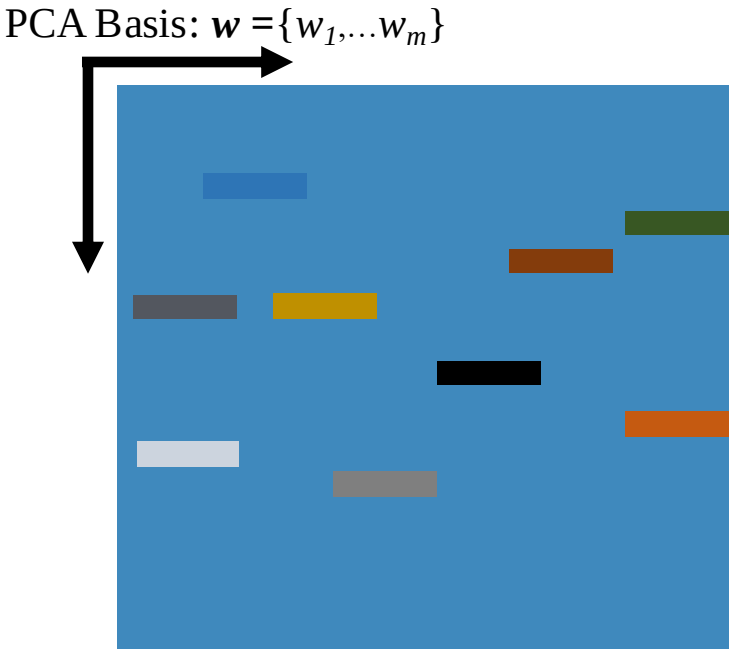
Model Weight Space



weights2weights (w2w) Space



weights2weights (w2w) Space



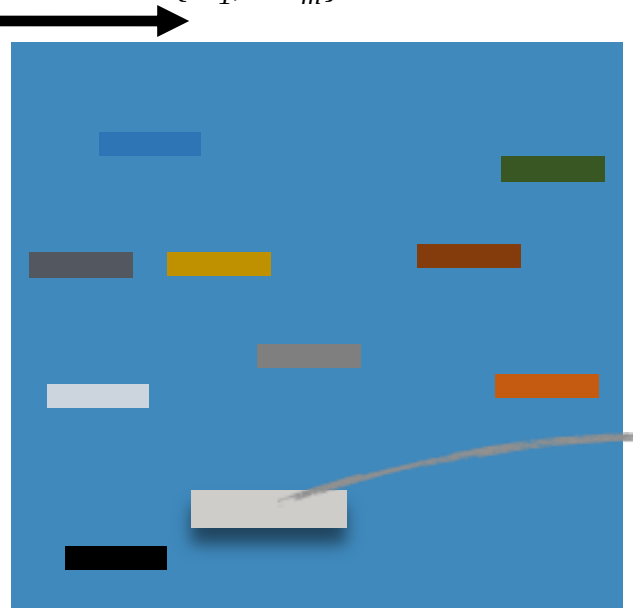
"Meta"-Latent Space

Applications

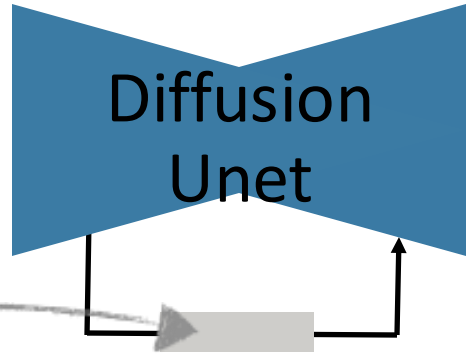
Sampling New Models

weights2weights Space

PCA Basis: $\boldsymbol{w} = \{w_1, \ldots w_m\}$

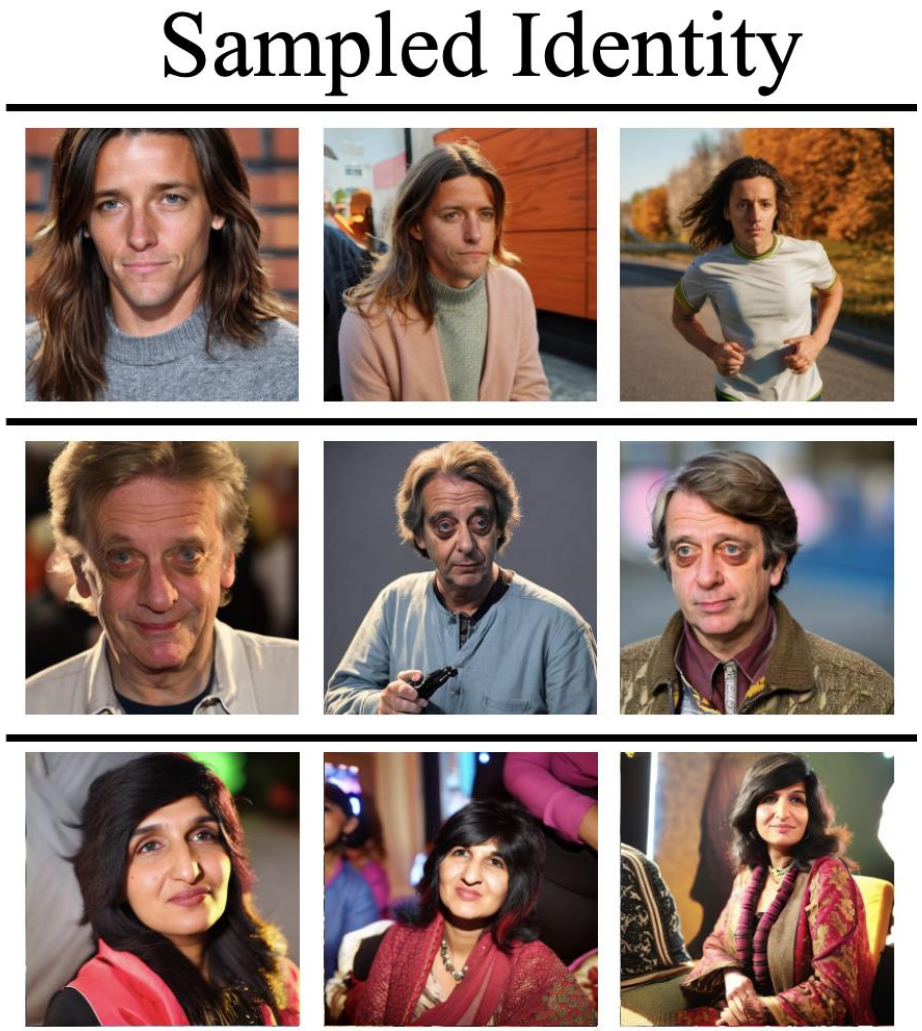


Sample model weights



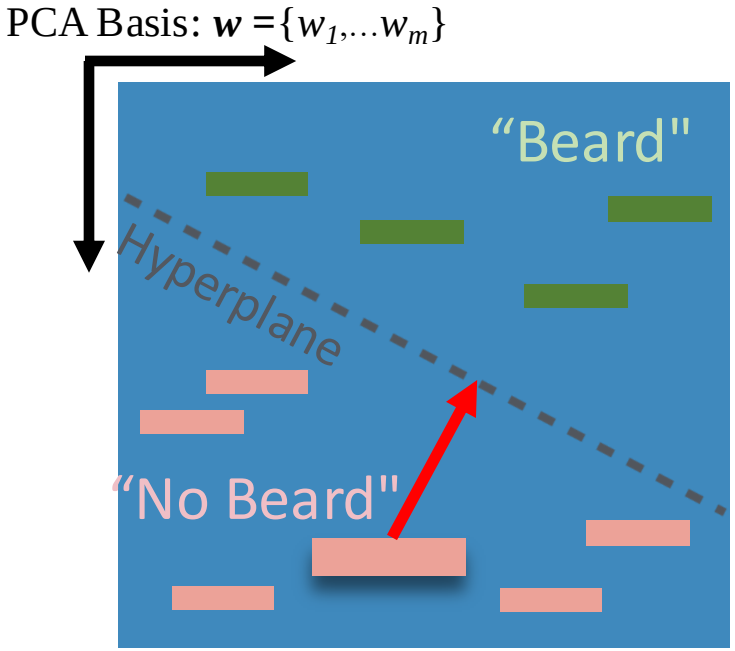
Generate Identity

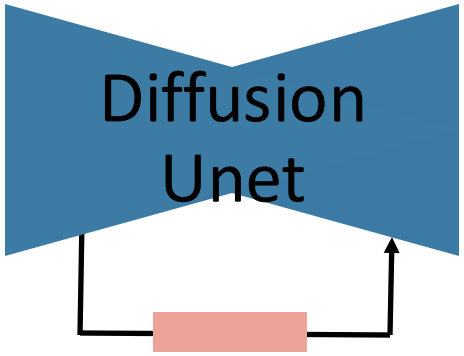
Sampling New Models



Editing Models

weights2weights Space

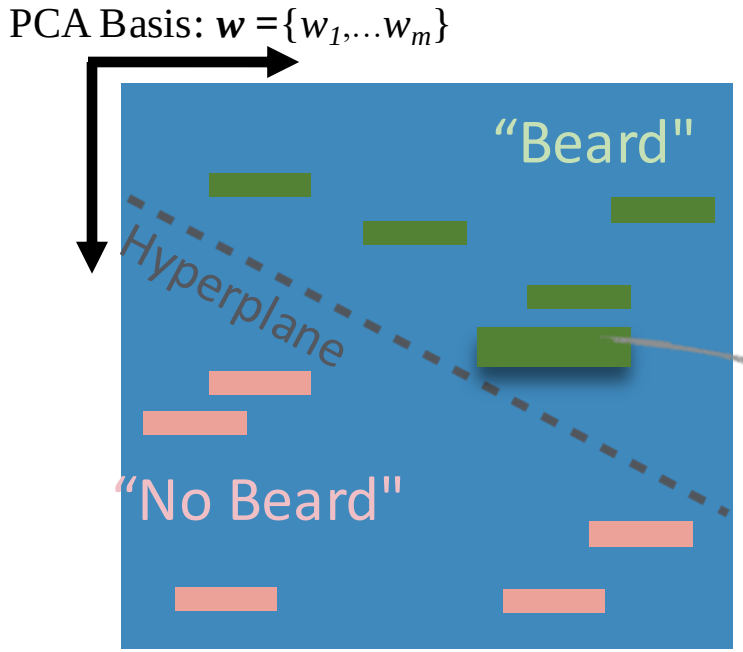


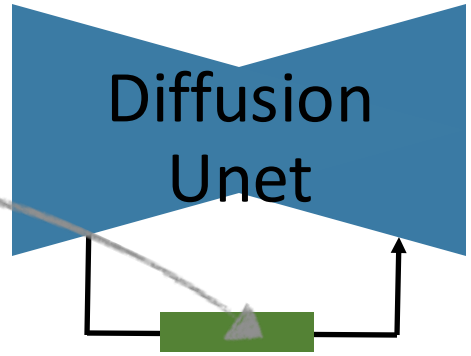


Generate Identity

Editing Models

weights2weights Space





Generate Identity

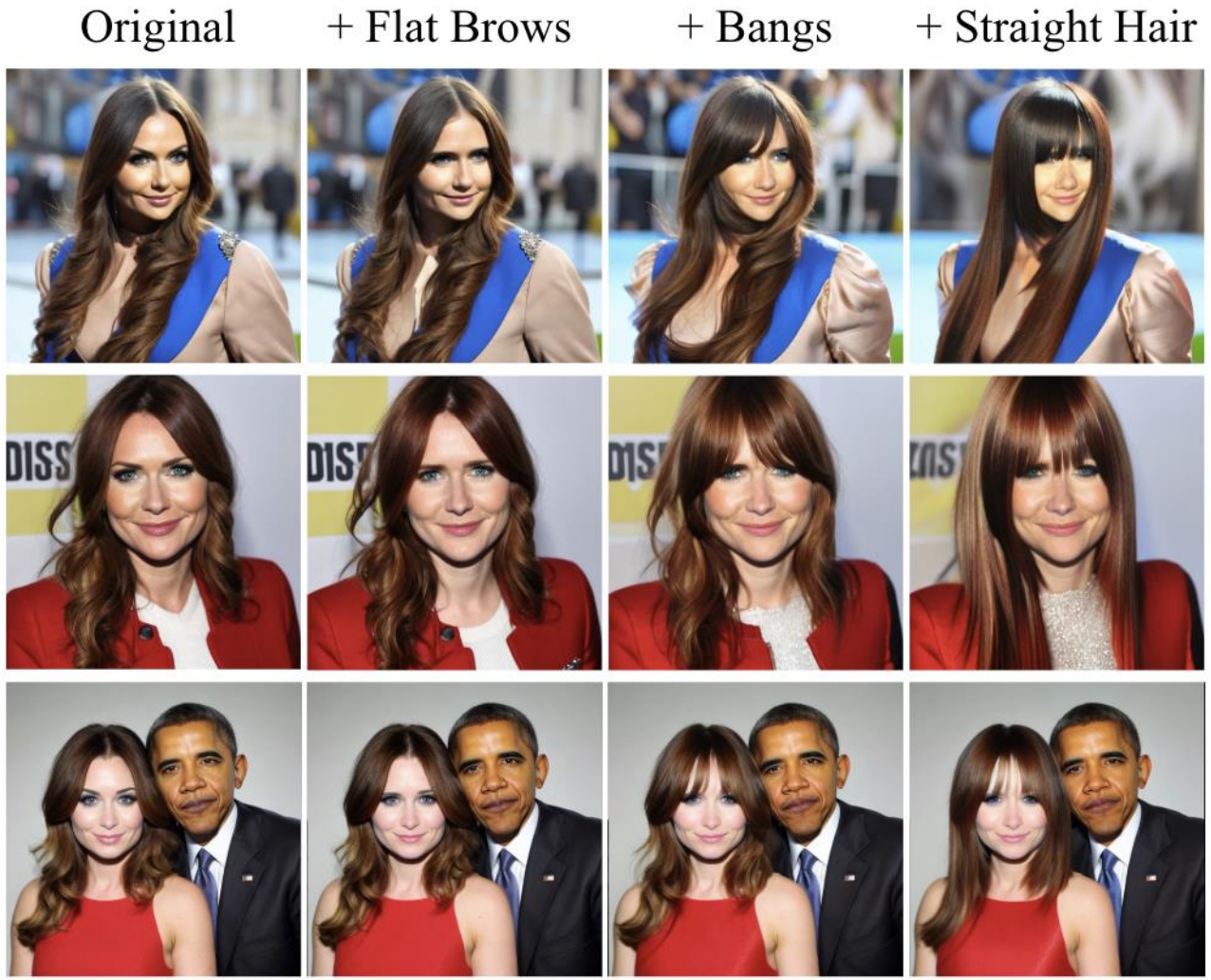
Editing Models

Original

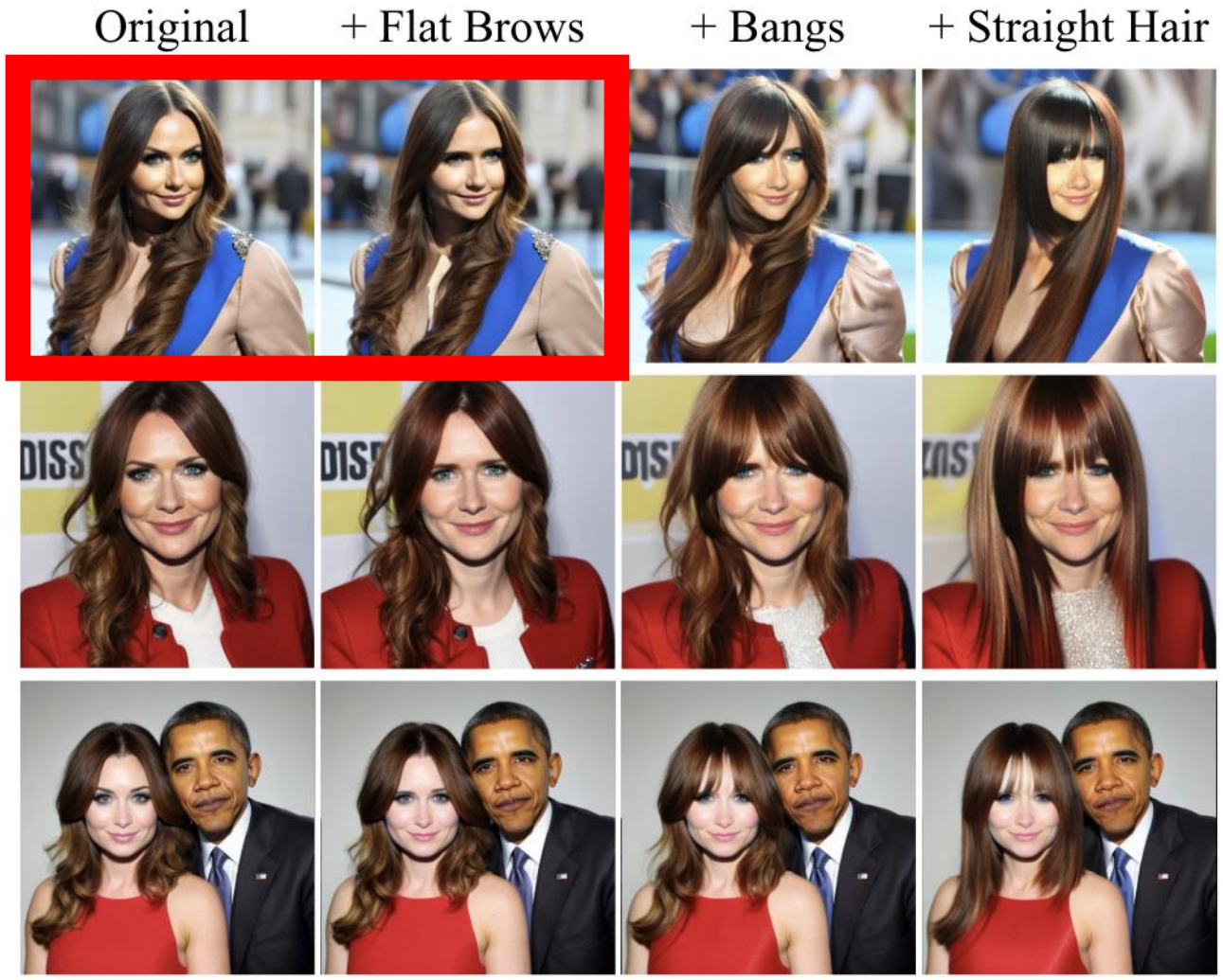
Editing Models

Editing Models

Editing Models

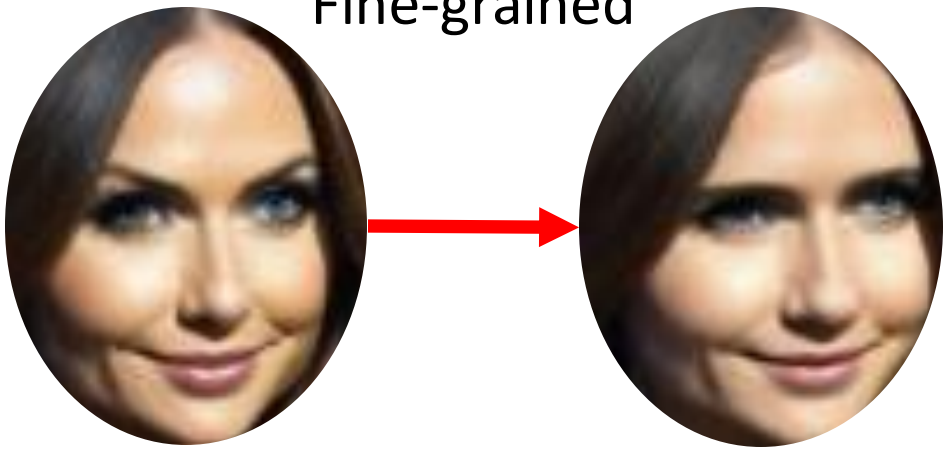


Editing Models

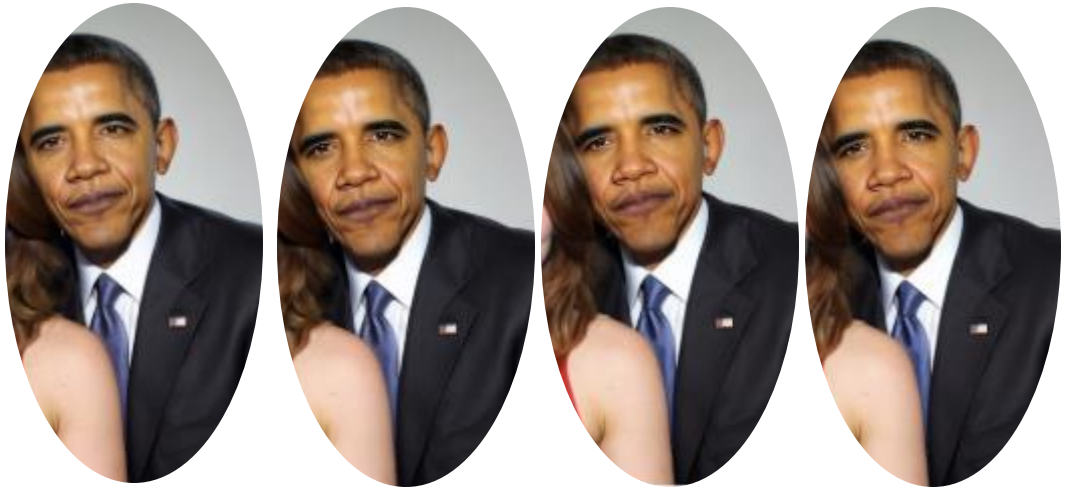


Editing Models

Fine-grained



Minimally interferes with other concepts

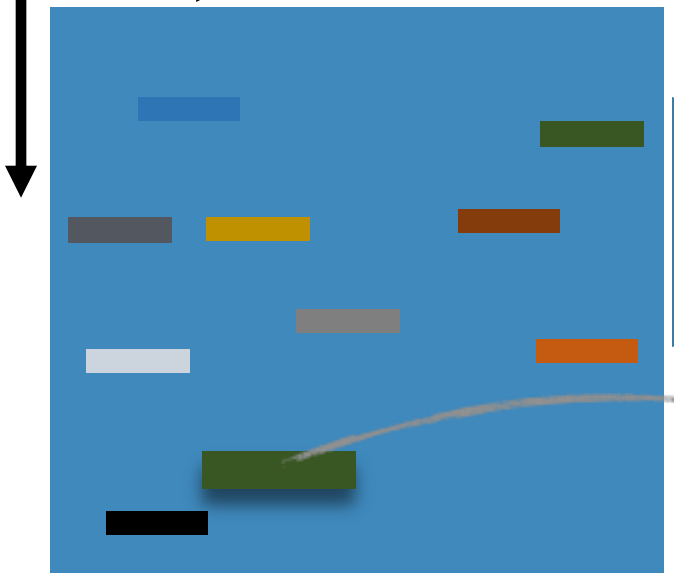


Inversion into *w2w* Space

weights2weights Space

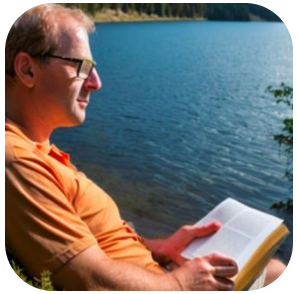
PCA Basis: $\boldsymbol{w} = \{w_1, \ldots w_m\}$

Input



Diffusion Unet

Generate Identity



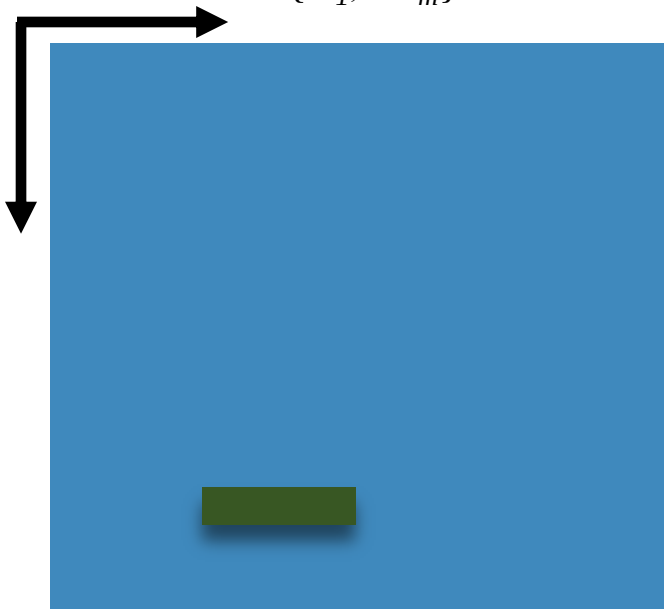
Invert into weights

Inversion into *w2w* Space

Input

weights2weights Space

PCA Basis: $\boldsymbol{w} = \{w_1, \ldots w_m\}$



Invert into weights

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

- Standard diffusion denoising objective with a subspace constraint
- Enforce constraint by operating directly on principal component basis

Inversion into *w2w* Space

Out-of-Distribution Inversion

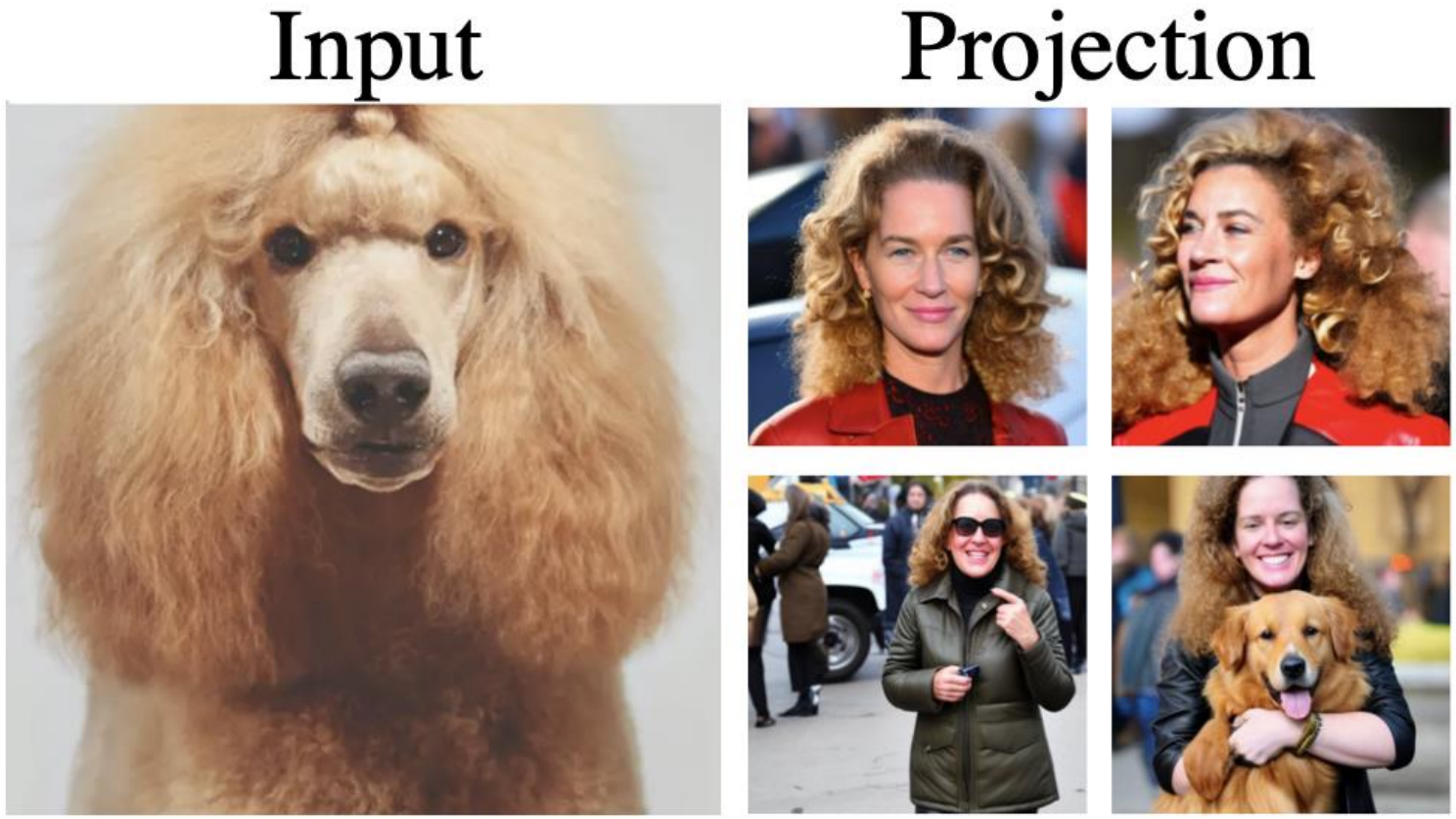
Input

Projection

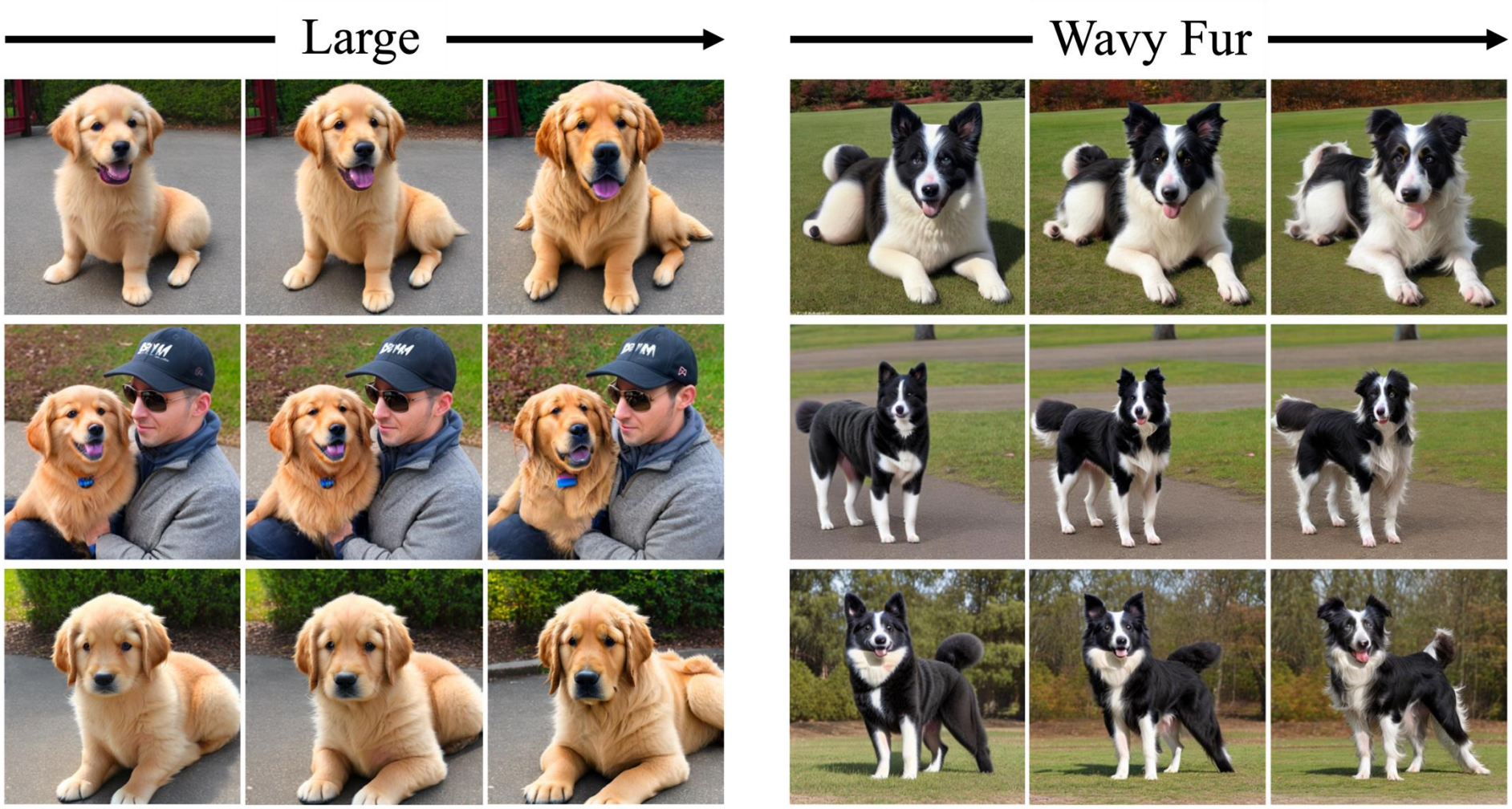
Input

Projection

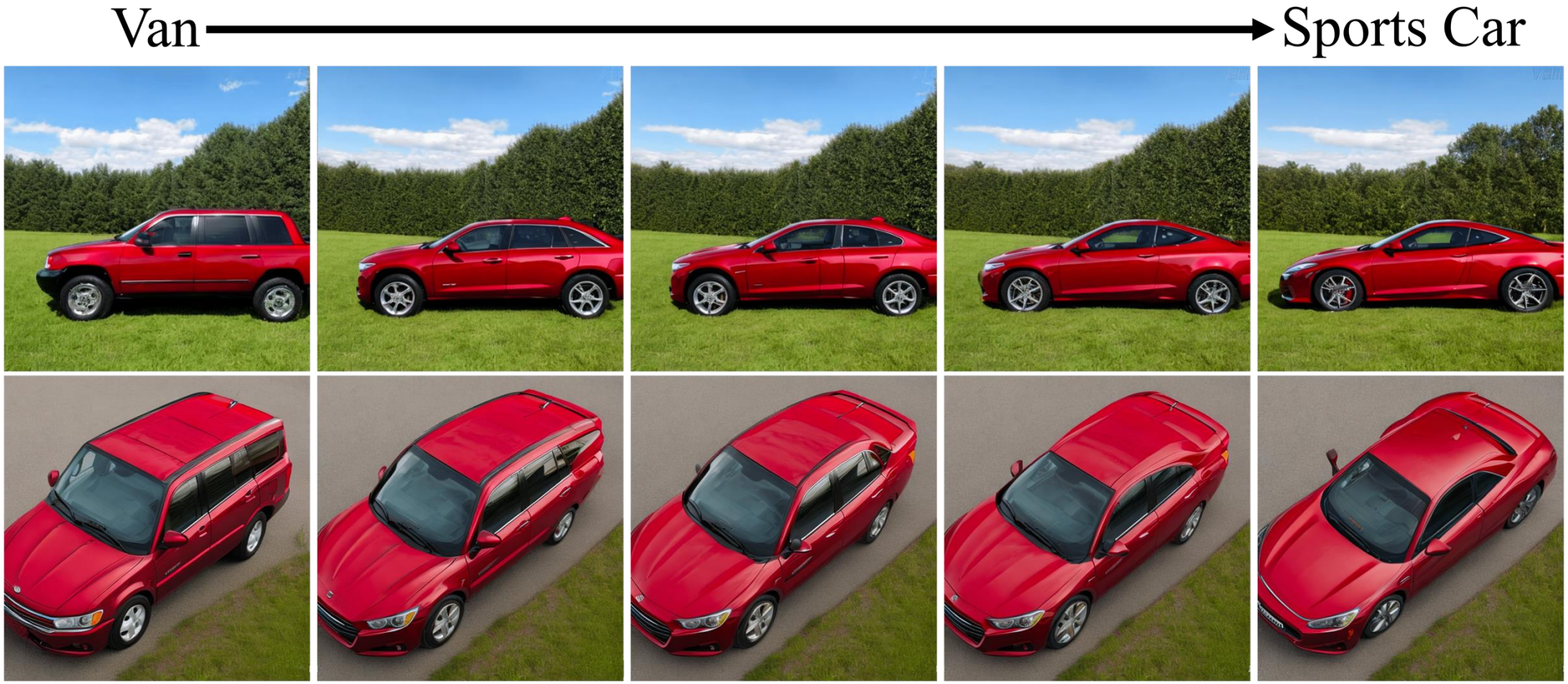
Out-of-Distribution Inversion



Extending to Other Tasks



Extending to Other Tasks



Thanks!

Project Page:

https://snap-research.github.io/weights2weights/

Code:

https://github.com/snap-research/weights2weights