

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



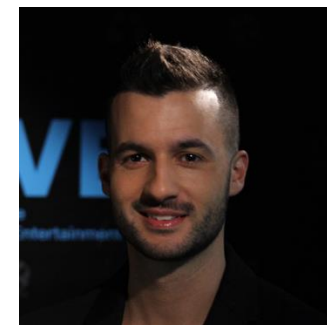
Gordon Wetzstein
Stanford



Alyosha Efros
UC Berkeley



Kfir Aberman
Snap

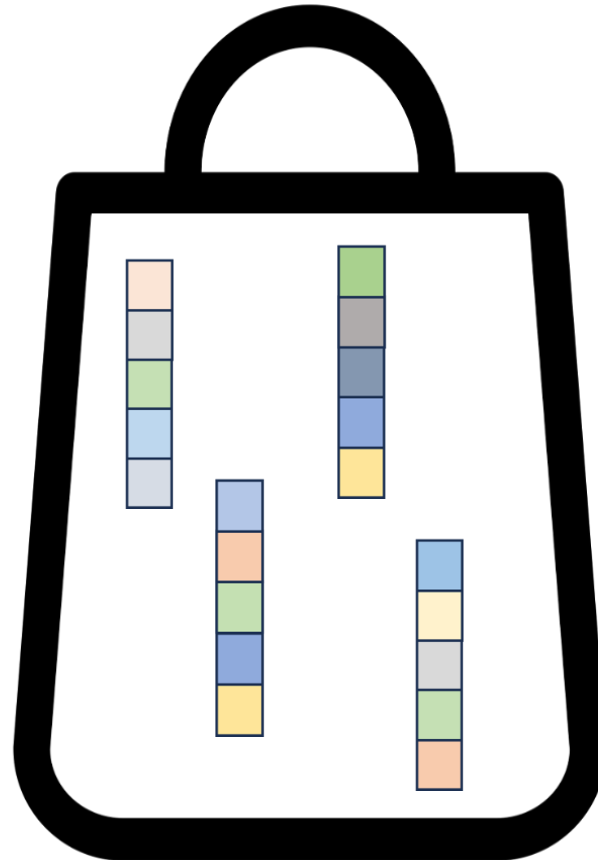


Linear Latent Spaces

Dataset of Images



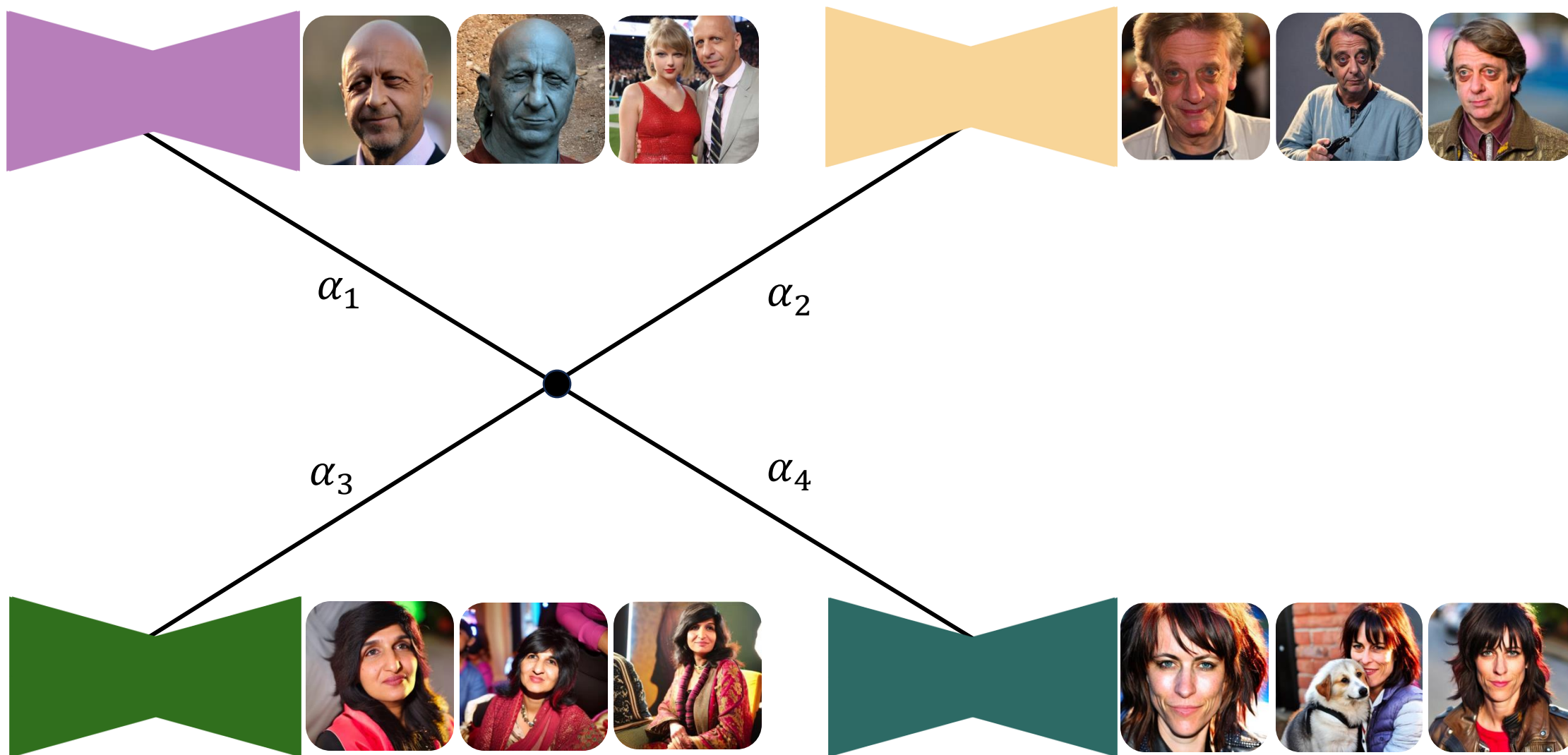
Dataset of Latent Vectors



Dataset of Model Weights



Interpolation in Weight Space



Interpolation in Weight Space



**How Do We Get These
Models?**

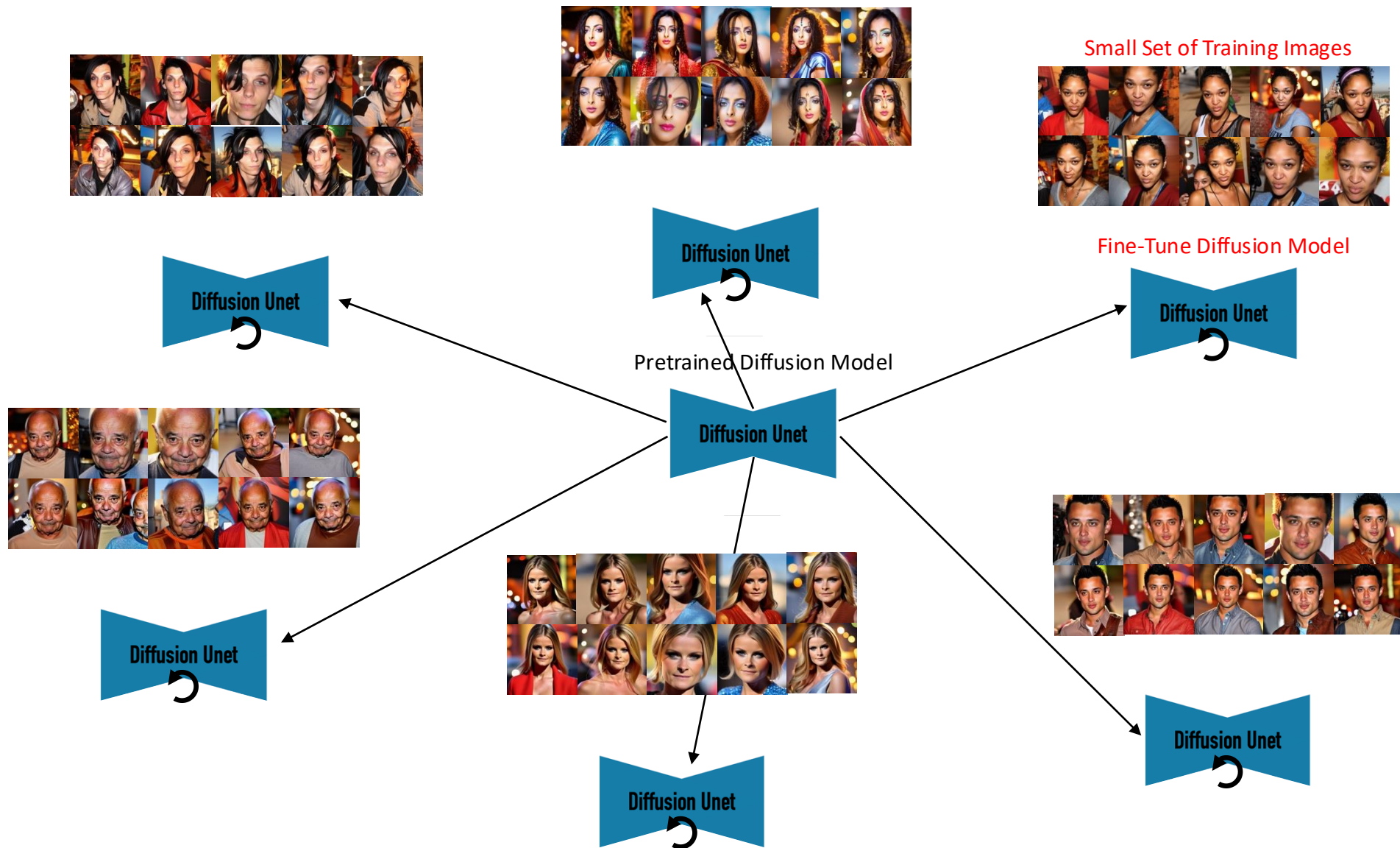
Fine-Tune for each Subject

Pretrained Diffusion Model



Diffusion Unet

Repeat over 60,000 Times



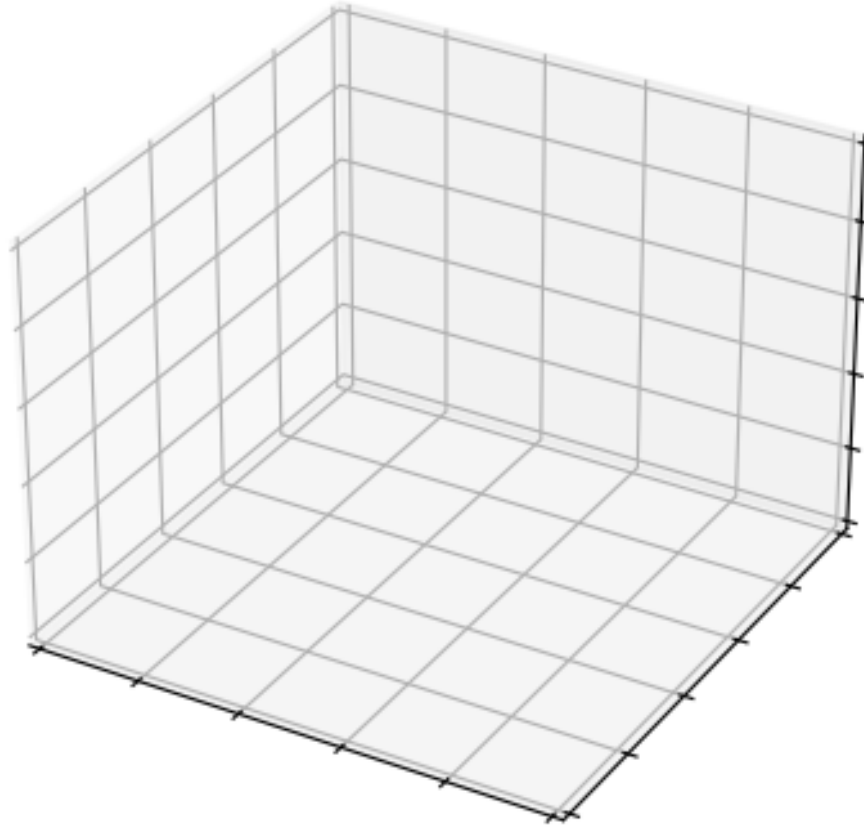
Modeling the Weights Manifold

weights



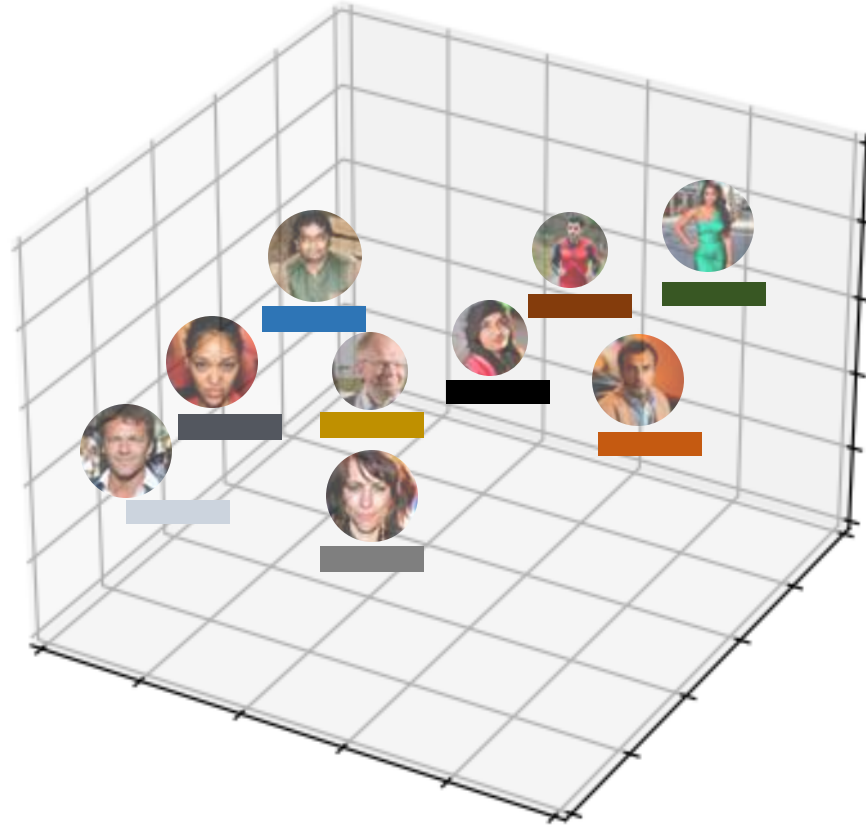
Modeling the Weight Manifold

Model Weight Space



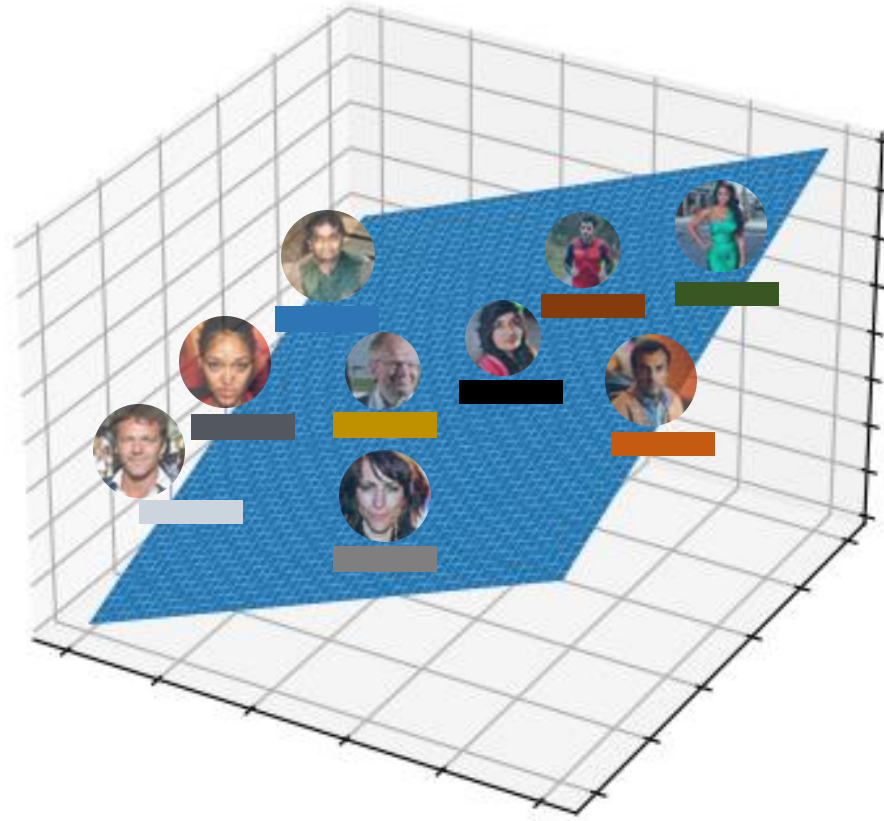
Modeling the Weight Manifold

Model Weight Space

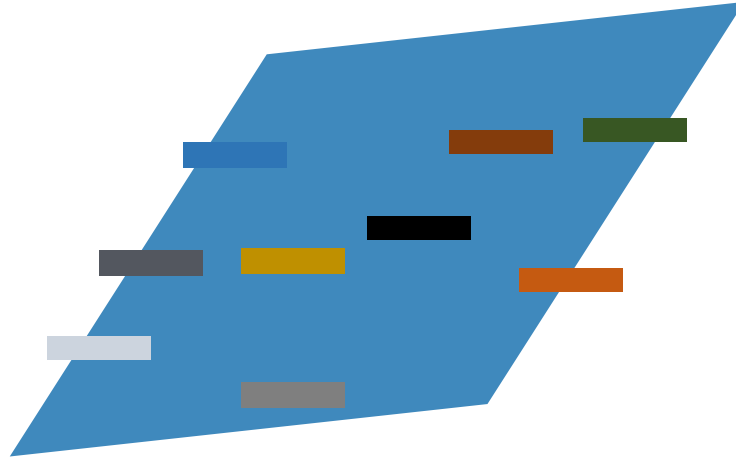


weights2weights (w2w) Space

Model Weight Space

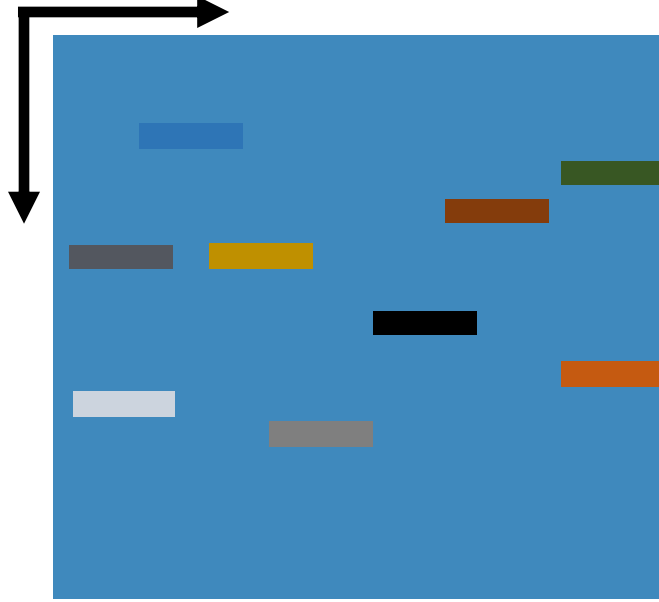


weights2weights (w2w) Space



weights2weights (w2w) Space

PCA Basis: $\mathbf{w} = \{w_1, \dots, w_m\}$



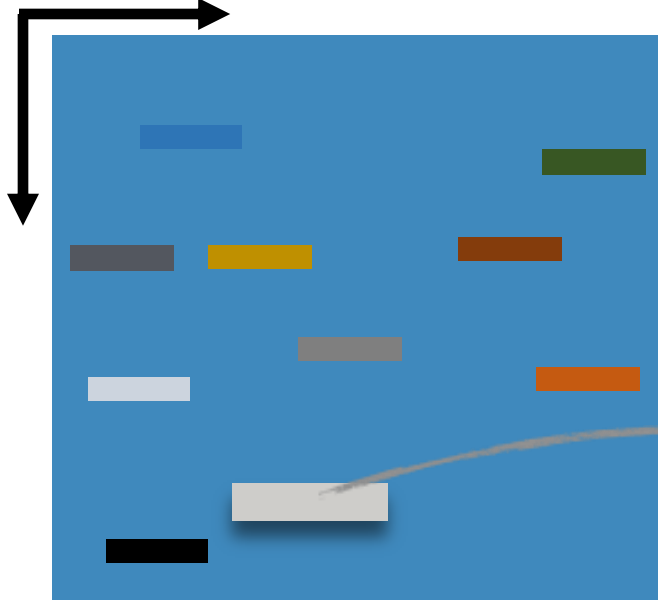
“Meta”-Latent Space

Applications

Sampling New Models

weights2weights Space

PCA Basis: $\mathbf{w} = \{w_1, \dots, w_m\}$



Diffusion
Unet

Generate Identity



Sample model
weights

Sampling New Models

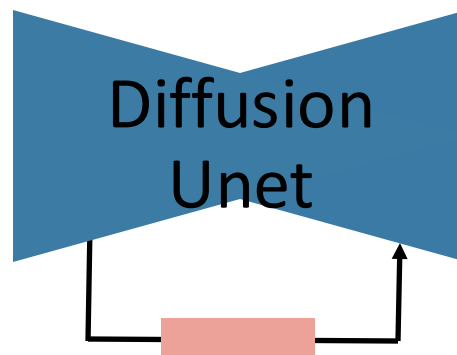
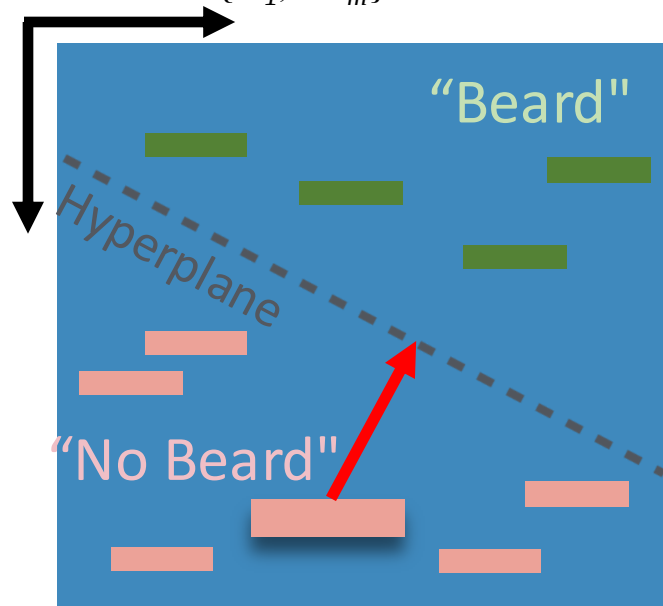
Sampled Identity



Editing Models

weights2weights Space

PCA Basis: $\mathbf{w} = \{w_1, \dots, w_m\}$



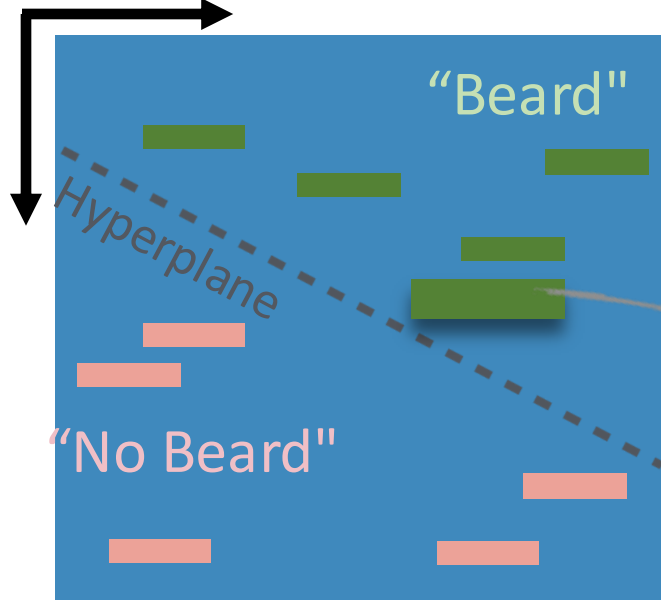
Generate Identity



Editing Models

weights2weights Space

PCA Basis: $\mathbf{w} = \{w_1, \dots, w_m\}$



Diffusion
Unet

Generate Identity



Editing Models

Original



Editing Models

Original

+ Flat Brows



Editing Models

Original

+ Flat Brows

+ Bangs



Editing Models

Original

+ Flat Brows

+ Bangs

+ Straight Hair



Editing Models

Original

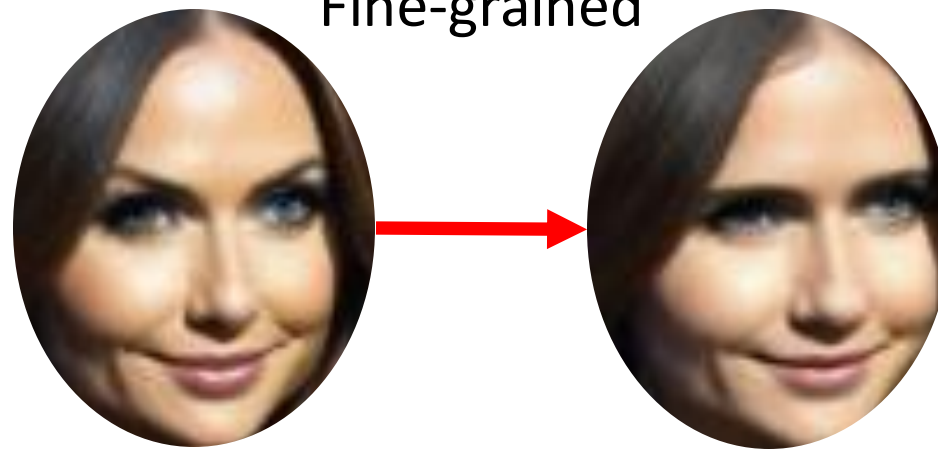
+ Flat Brows

+ Bangs

+ Straight Hair



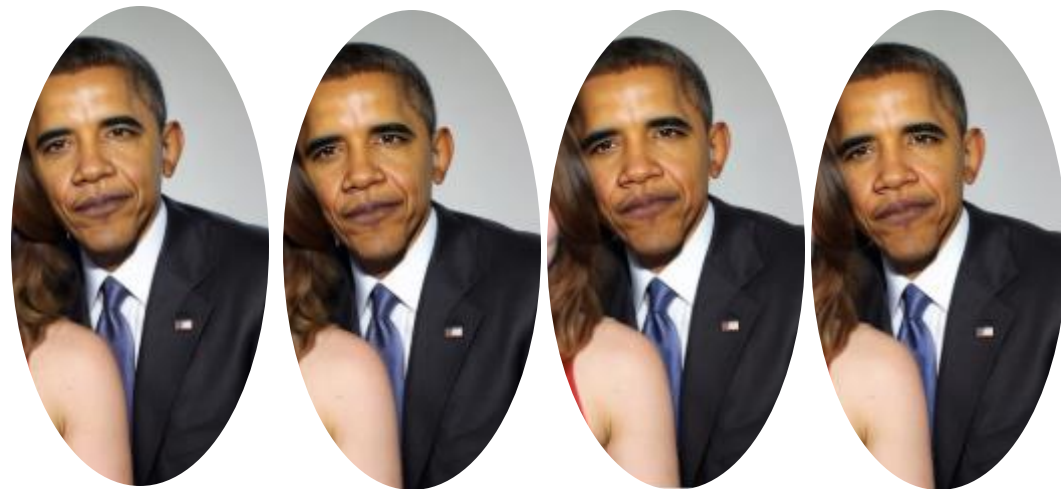
Fine-grained



Editing Models



Minimally interferes with other concepts

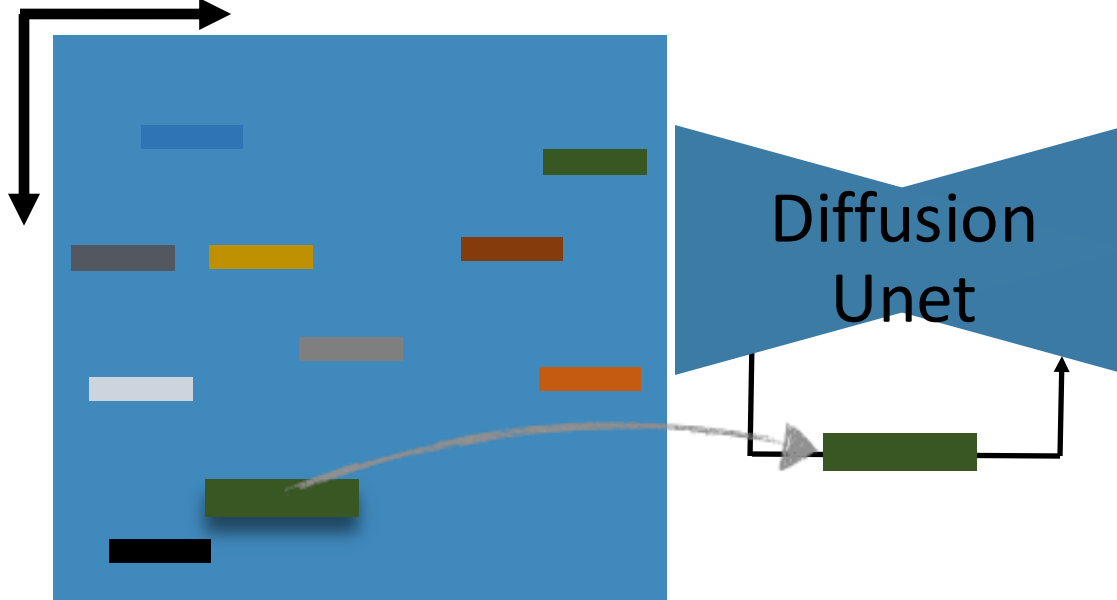


Inversion into $w2w$ Space

weights2weights Space

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

Input



Invert into weights

Generate Identity

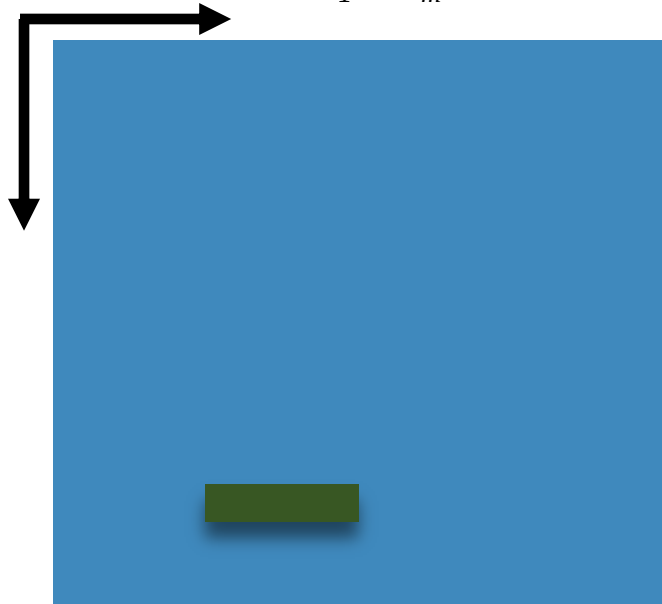


Inversion into $w2w$ Space

weights2weights Space

PCA Basis: $\mathbf{w} = \{w_1, \dots, w_m\}$

Input



Invert into weights

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

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

Inversion into w2w Space

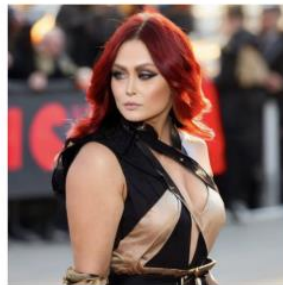
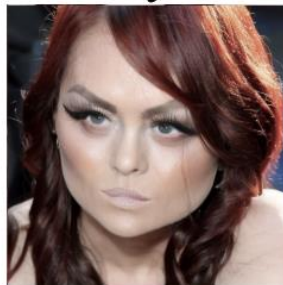
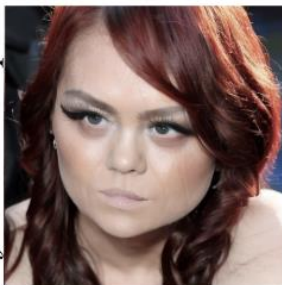
Input

Inversion

+Pointy Nose

face closeup

posing in a dress



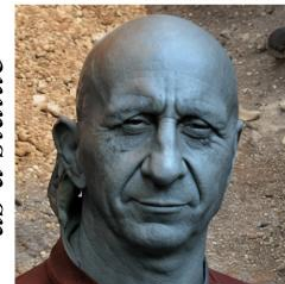
Input

Inversion

+Hair

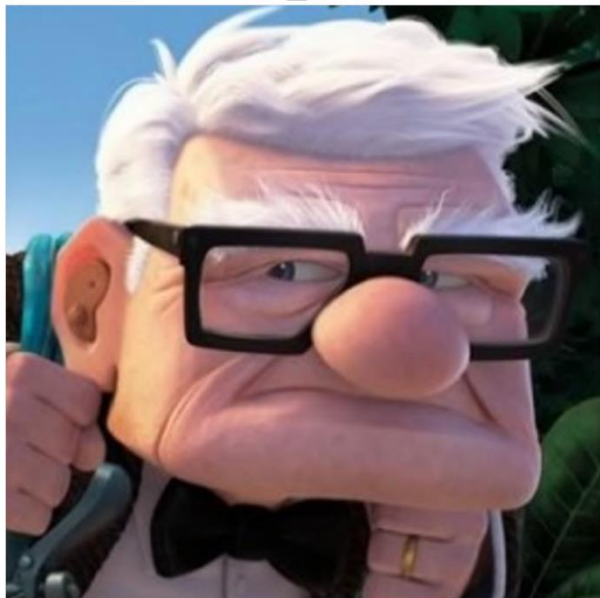
as a statue

with Taylor Swift



Out-of-Distribution Inversion

Input



Projection



Input



Projection



Out-of-Distribution Inversion

Input

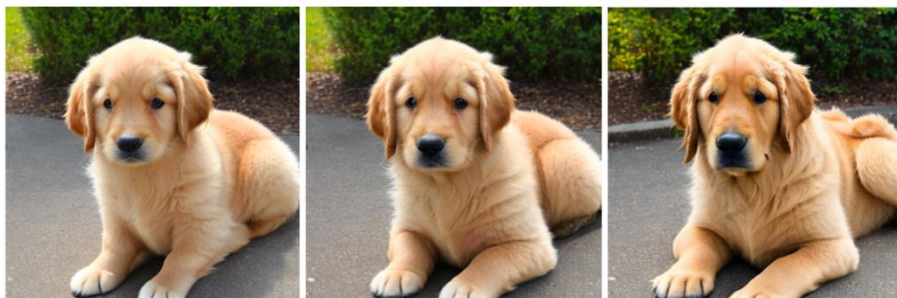
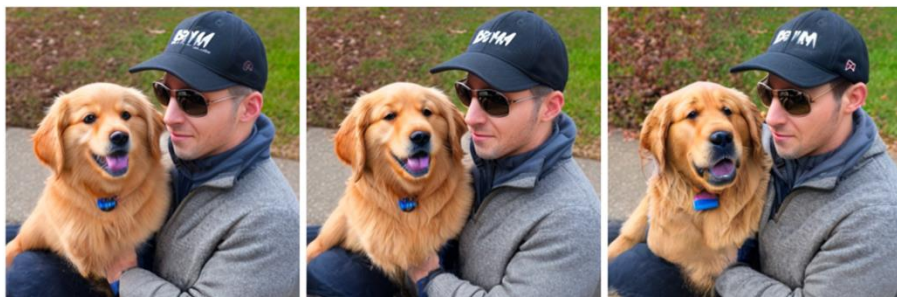
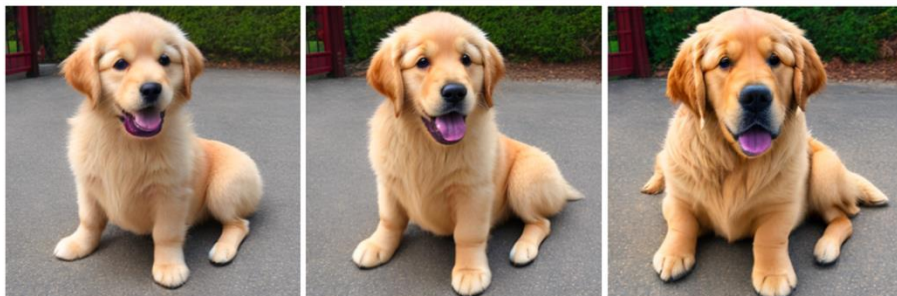


Projection



Extending to Other Tasks

———— Large —————→



———— Wavy Fur —————→



Extending to Other Tasks

Van  Sports Car



Thanks!

Project Page:

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

Code:

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