MagicMirror: Fast and High-Quality Avatar Generation with Constrained Search Space

Armand Comas[1,2*], Di Qiu[1], Menglei Chai[1], Marcel Bühler[1,3*], Amit Raj[4], Ruiqi Gao[4], Qiangeng Xu[1], Mark Matthews[1], Paulo Gotardo[1], Sergio Orts-Escolano[1], Thabo Beeler[1]
[1]Google AR/VR [2]Northeastern University COE [3]ETH Zürich [4]Google DeepMind *work done as a Google intern

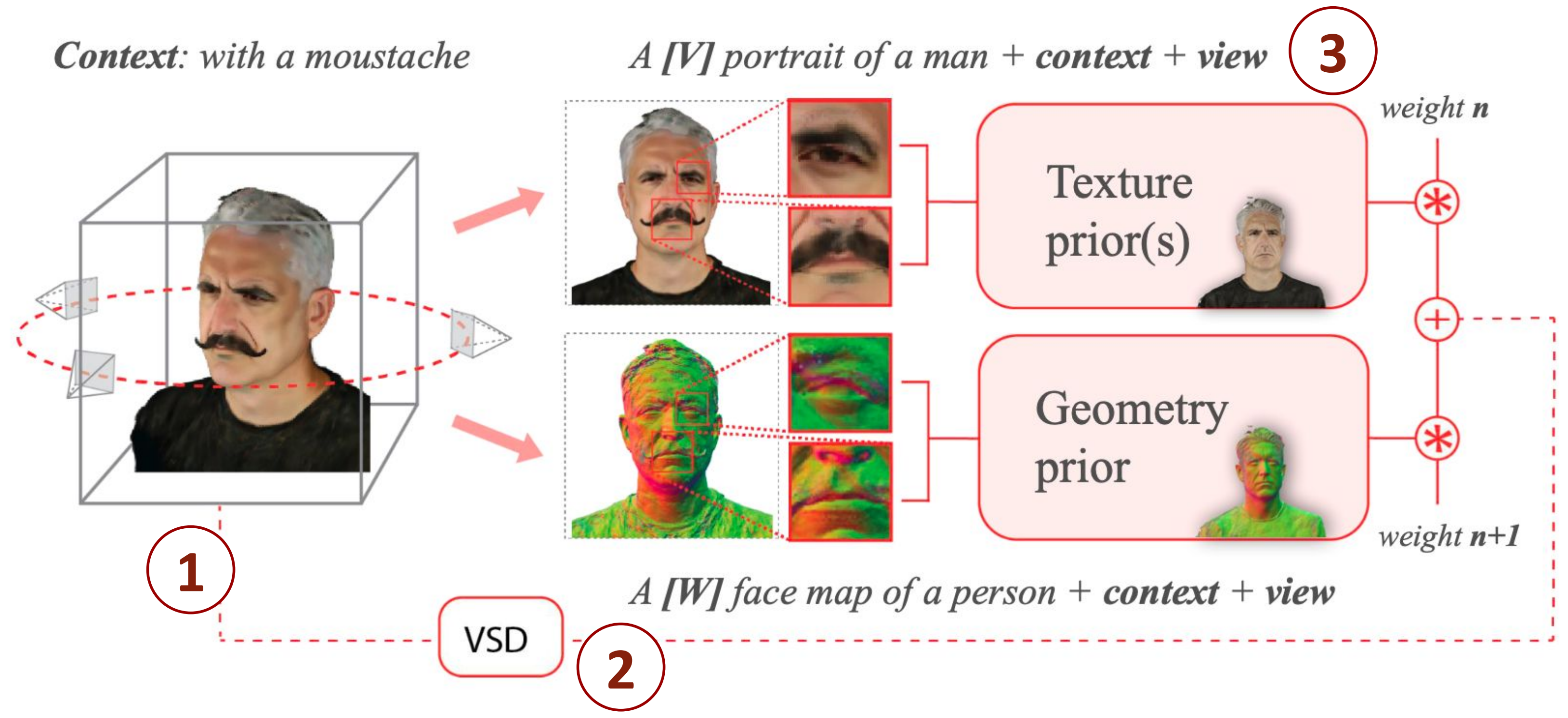
Summary

SCAN ME

Text-guided 3D head avatar generation and editing with high visual quality. Key components:

- The incorporation of **VSD [2]** to 3D human generation for improved texture.
- A set of **text-to-image diffusion priors** that capture:
  i) General human head distribution, **ii)** Identity of subjects **iii)** Geometric prior.
- A **constrained solution space**, learned as a conditional NeRF model trained on a dataset of human heads.

Method


Context: with a moustache
A [V] portrait of a man + context + view
3
weight n
Texture prior(s)
Geometry prior
weight n+1
A [W] face map of a person + context + view
1
VSD
2

2) Variational Score Distillation for 2D to 3D lifting

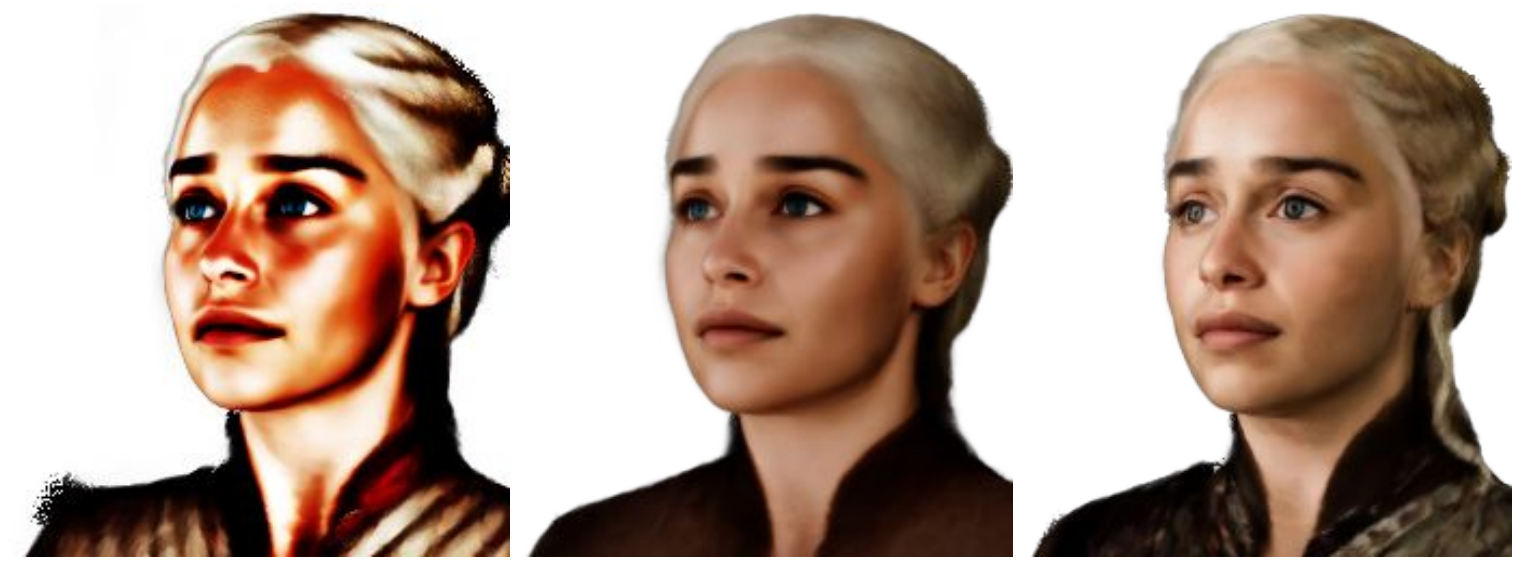
We use VSD [2] with multiple sets of generative priors which improves texture generation w.r.t SDS:

$$\mathcal{L}_{\text{VSD}}(\mathcal{D}', I) = \mathcal{L}_{\text{SDS}}(I) - \mathcal{L}_{\text{proxy}}(\text{sg}(\mathcal{D}'), I) + \mathcal{L}_{\text{proxy}}(\mathcal{D}', \text{sg}(I))$$

with: $\mathcal{L}_{\text{SDS}}(\text{sg}(\mathcal{D}), I, \epsilon, T, t) = \omega(t)\|\text{sg}(\mathcal{D}(I, \epsilon, T, t)) - I\|^2$

and $\mathcal{L}_{\text{proxy}}(\mathcal{D}', \text{sg}(I)) = \omega(t)\|\mathcal{D}'(I, \epsilon, T, t) - \text{sg}(I)\|^2$

Daenerys Targaryen

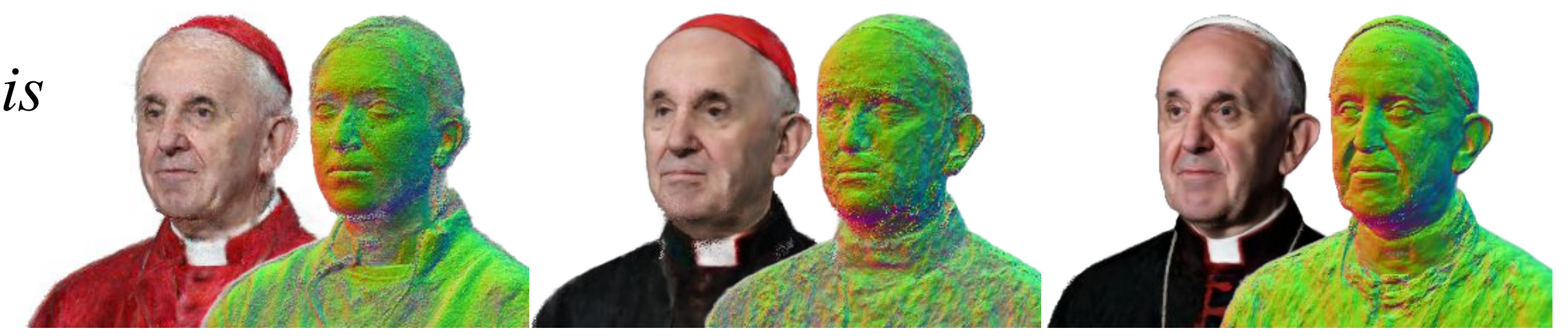


SDS, CFG=100 SDS, CFG=20 VSD, CFG=3

1) Avatar Prior (constrained solution space):

We use [3], a Mip-NeRF360 conditioned by an identity code. It is pretrained with 1450 human faces with natural expression. The number of faces is key to performance:

Pope Francis



1 subject 350 subjects 1450 subjects

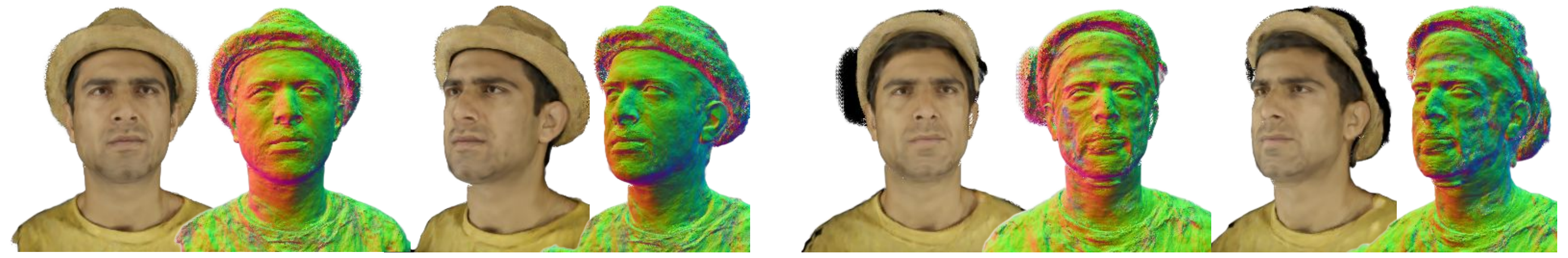
3) A mixture of Generative priors

We mix linearly the contributions of several priors:

- A generic prior based on a pre-trained text-to-image diffusion model [4].

We leverage a DreamBooth scheeme [1] to capture:

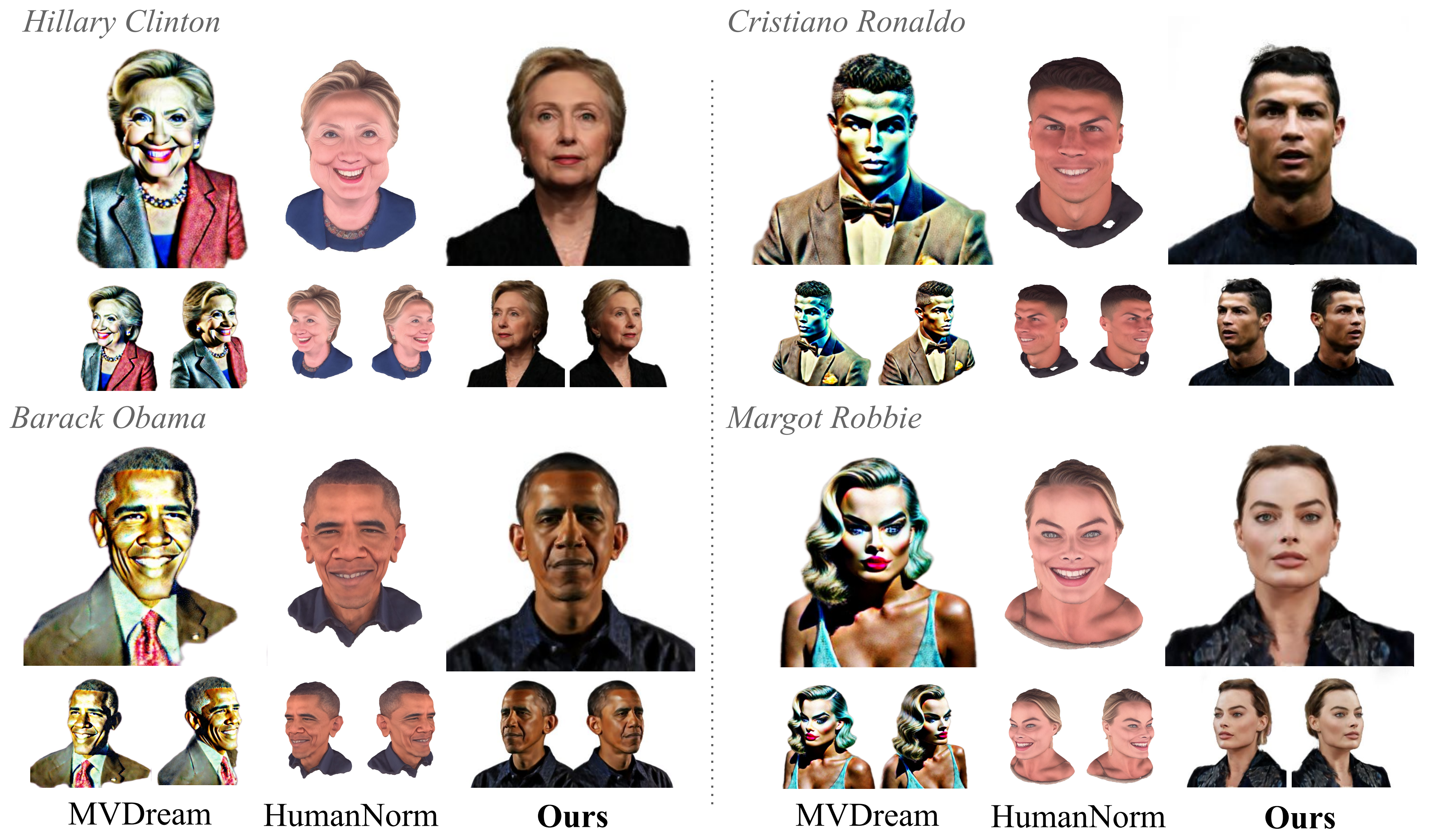
- The identity of a person for editing: *A [V] portrait…*
- A geometric prior (novel): *A [W] face map…*

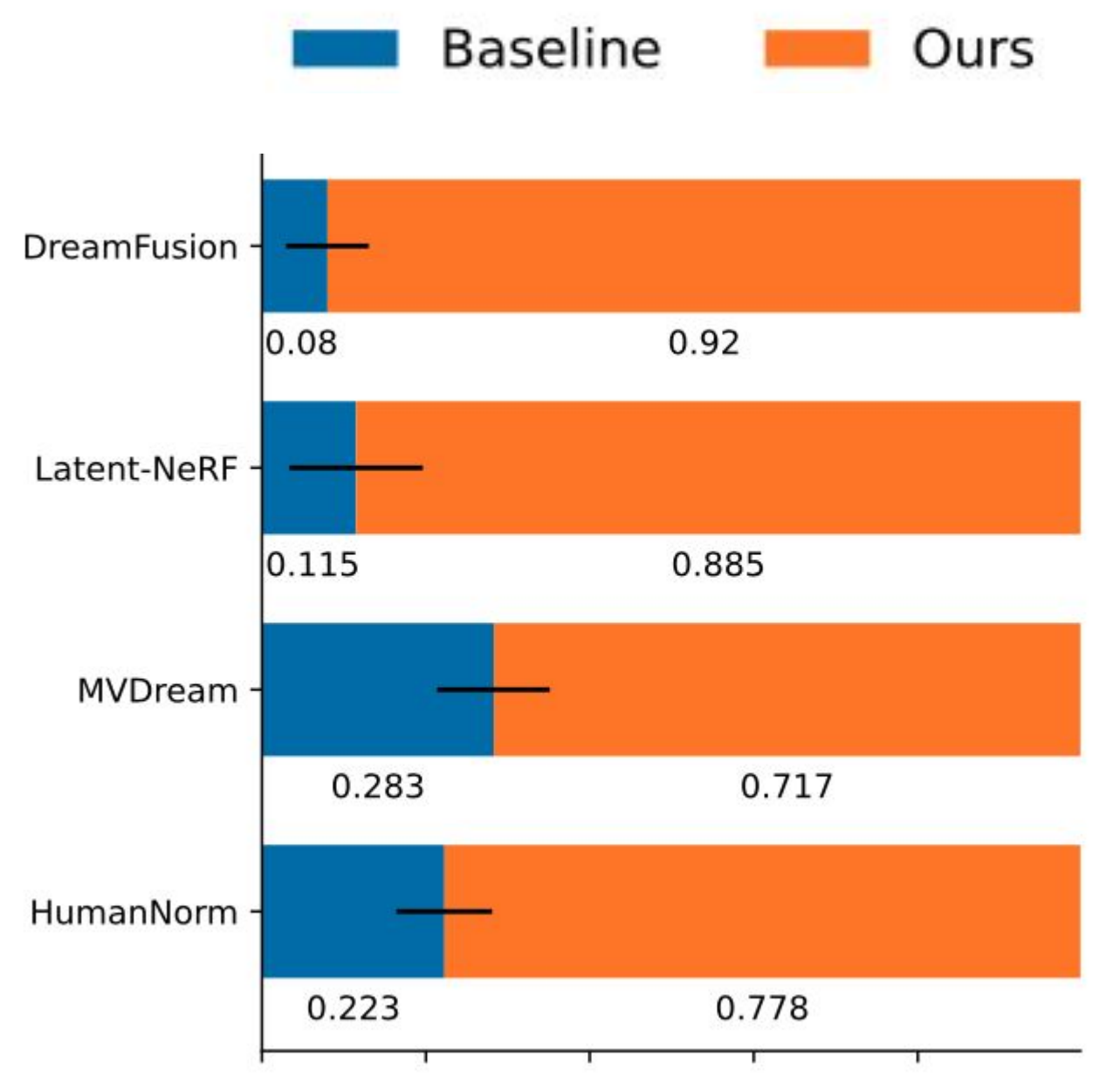


w/ geo. prior w/o geo. prior

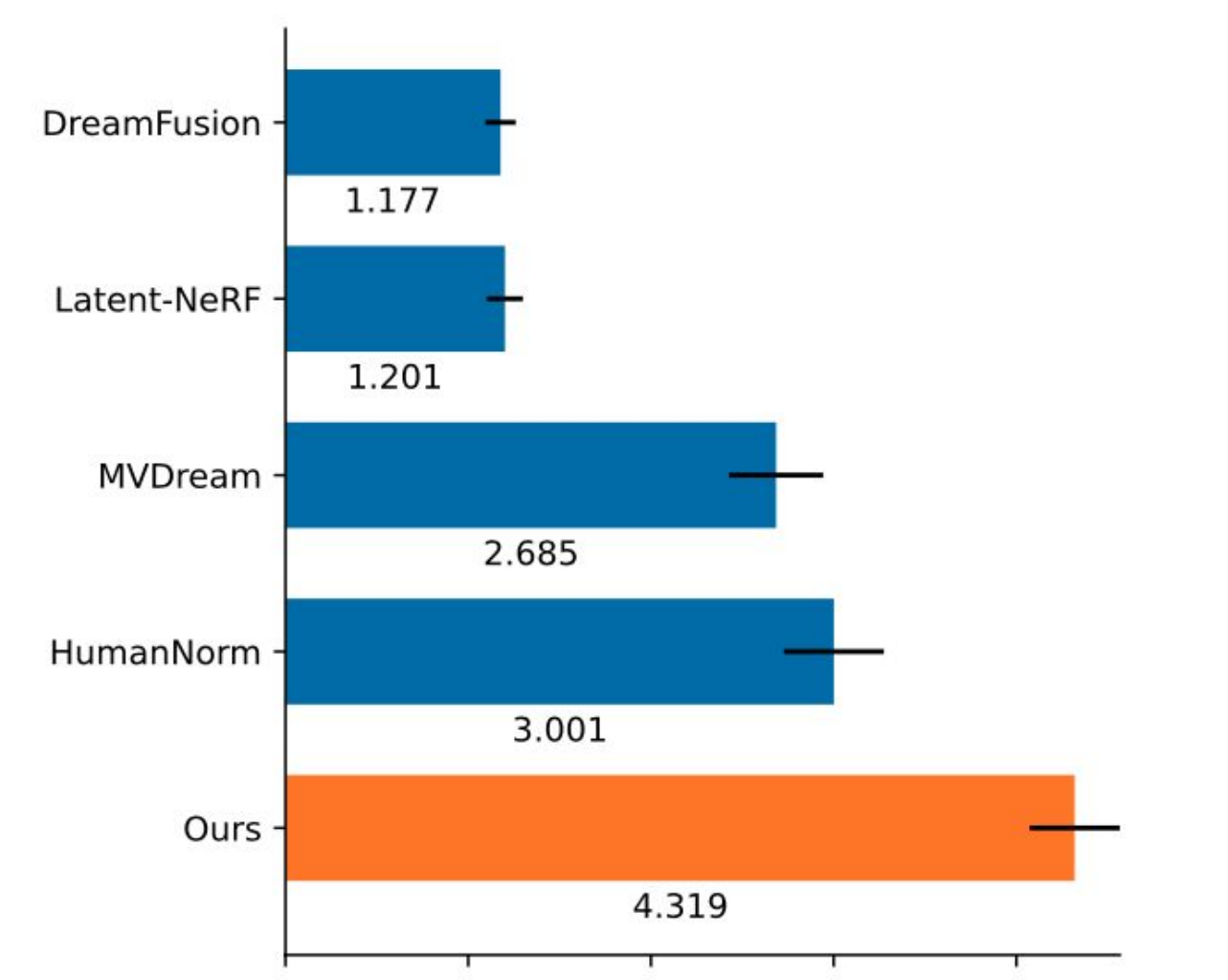
[1] Ruiz, N.et al (2023). DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
[2] Wang, Z., et al (2023). ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)
[3] Buehler, M. C., et al(2023). Preface: A Data-driven Volumetric Prior for Few-shot Ultra High-resolution Face Synthesis. Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)
[4] Saharia, C.et al (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Advances in Neural Information Processing Systems (NeurIPS)

Generation Results



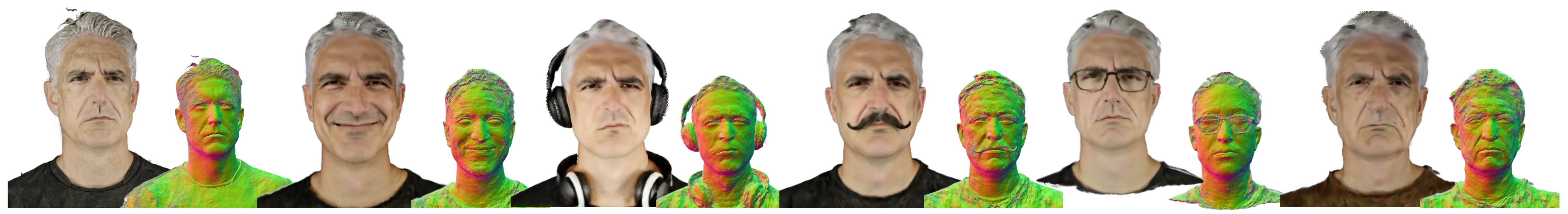


PickScore



Human Study: Quality

Editing/Stylizing Results



Original Avatar who is happy wearing headphones with a mustache wearing glasses old person

Smooth Transitions

A [V] man who is a child → A [V] man who is an old person

Smooth optimization across concepts

Mixture of Concepts

1 Sad / 0 Happy 1 Sad / 0.5 Happy 1 Sad / 1 Happy 0.5 Sad / 1 Happy 0 Sad / 1 Happy

We mix concepts by linear interp. of updates