The Universal Normal Embedding - Supplementary Material

Chen Tasker* Roy Betser* Eyal Gofer* Meir Yossef Levi Guy Gilboa
Viterbi Faculty of Electrical and Computer Engineering
Technion - Israel Institute of Technology

Overview

In this supplementary material document, we provide additional implementation and experimental details to ensure the full reproducibility (Section A). We also provide additional analyses and qualitative examples of our linear editing approach, along with experiments on an additional dataset (Section B).

A. Reproducibility

Code and the NoiseZoo dataset are available here. Full implementation details are also available in the code repository.

A.1. NoiseZoo construction details

We constructed the NoiseZoo dataset by extracting latent representations for all 19,867 images in the CelebA [4] validation split, without any filtering. The dataset includes latents from three Stable Diffusion variants (SD 1.5, SD 2.1, and LCM [5, 10]), two CLIP variants (ViT-B/16 and ViT-L/14) [8], two OpenCLIP variants with the same architectures [6], and DINOv3 (ViT-L/16) [12]. Additionally, the NoiseZoo dataset was randomly split into 15,893 training samples and 3,974 test samples.

Stable Diffusion latents. Latent representations for diffusion models were obtained via DDIM inversion using the HuggingFace *diffusers* library. All images were center-cropped and bilinearly resized to 512×512 prior to inversion. Inversion was performed with an empty text prompt, classifier-free guidance enabled, a guidance scale of 3.5 and a fixed random seed (42). SD 1.5 and SD 2.1 were inverted with 50 DDIM steps, while LCM used 150 steps and a DDIMScheduler (as the default LCM scheduler is DDPM-based and does not support inversion). For all Stable Diffusion models, we saved only the initial latent obtained from the inversion procedure. All Stable Diffusion latents have shape (4, 64, 64) and are flattened before all the experiments.

*These authors contributed equally to this work.
Corresponding author: roybe@campus.technion.ac.il

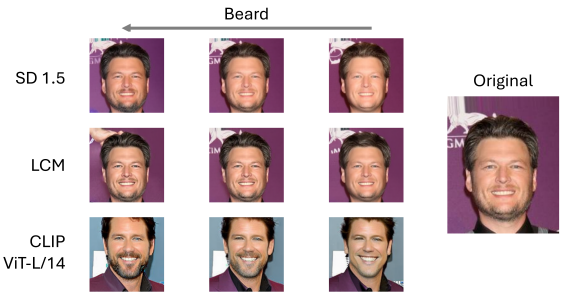

Figure 1. **Editing in different latent spaces.** The figure compares linear attribute editing across three latent spaces: two diffusion models and CLIP. The diffusion latents preserve the structure of the original image, so shifting along an attribute direction produces a modified version of the same image. In contrast, CLIP's latent space is not invertible to the pixel domain, so reconstruction yields a newly synthesized image that matches the target attribute but does not reconstruct the input. This highlights the trade-off: CLIP offers strong semantic control but poor original image faithfulness.

Encoder latents (CLIP, OpenCLIP, DINO). Encoder-based representations were obtained by passing each original CelebA image through the corresponding model using the model's default preprocessing pipeline. For DINOv3 (ViT-L/16), images were center-cropped and resized to 224x224 before encoding. No additional normalization was applied. The embedding dimensions for each model are:

- CLIP ViT-B/16: 512
- CLIP ViT-L/14: 768
- OpenCLIP ViT-B/16: 512
- OpenCLIP ViT-L/14: 768
- DINOv3 ViT-L/16: 768

Encoder embeddings were not normalized to unit norm.

A.2. Experimental details

Classification in latent space. For each feature set, the linear classifier consisted of a PCA projection, standard scaling, and an attribute-wise logistic regression stage. PCA was applied first (500 components for generative models

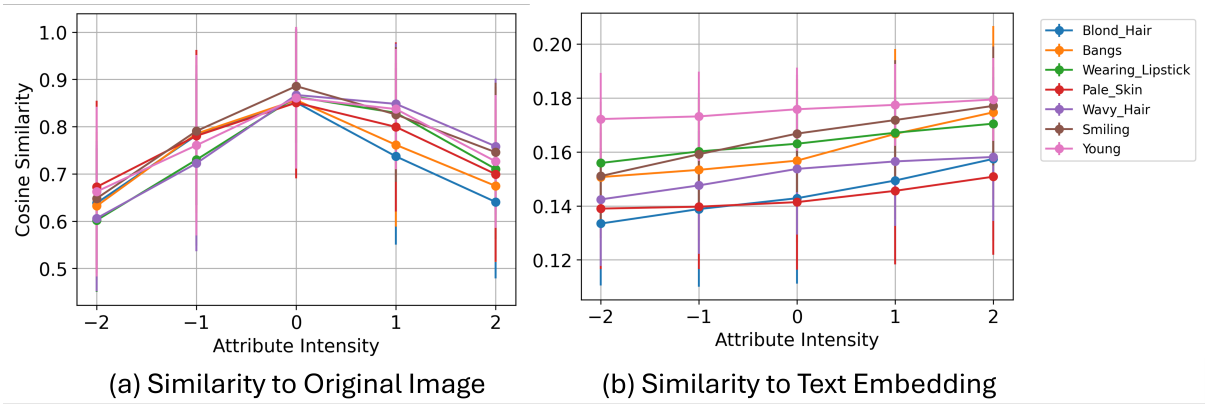

Figure 2. **Quantitative editing tests.** All measures are presented as a function of the edited *attribute intensity*, a normalized measure derived from the distance between the resulting latent and the appropriate classifier's decision plane. Edits are performed on SD 1.5 latents. Note that an x-axis value of 0 does not indicate no editing, but corresponds to editing the latent to the classifier's decision plane. Cosine similarities are measured between CLIP ViT-L/14 embeddings. (a) Cosine similarity between an edited image and the original image. (b) Cosine similarity between an edited image and CLIP text embeddings of the attribute's name.

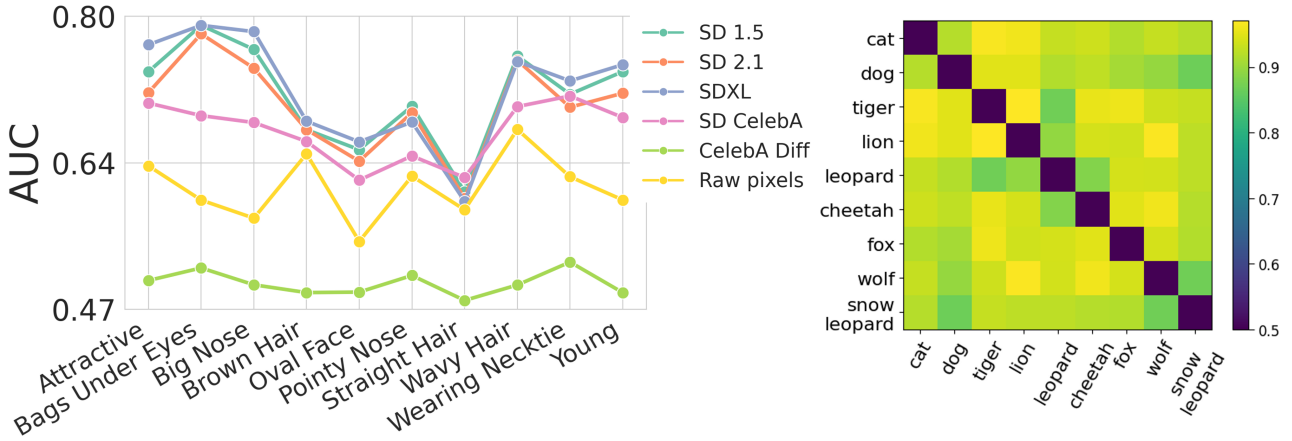

Figure 3. **Classification AUC.** Left: CelebA across different latent spaces. Right: AFHQ binary classification using SD 1.5 latents (AUC values).

and 310 for encoders), followed by standard scaling. Then, for each of the 40 attributes, a separate linear classifier was trained using scikit-learn's LogisticRegression with the saga solver, L2 regularization, a maximum of 25 iterations, and 30 parallel jobs. Each attribute's model forms one row of the overall weight matrix, with its corresponding bias term in the bias vector.

Cross-space transfer. We used ridge regression (scikit-learn's Ridge class) to learn a linear mapping between latent representations. The model was trained on paired samples in the training set, and evaluation (reported in Table 2 in the paper) was performed by applying a classifier trained in the target space to the translated test representations.

To ensure consistent regularization across different latent representations, the ridge penalty was scaled by the energy of the source features. The effective ridge penalty was set to $\alpha_{\text{eff}} = \alpha \frac{||X_{\text{source}}||_F^2}{d}$, where α is the base regularization parameter, X_{source} is the source feature matrix and d is its dimensionality. We used $\alpha = 1.0$ in the reported results.

Shared latent spaces. The splits marked as $X1$-$X5$ in Figure 6 in the main paper are as follows:

- $X1$: SD 2.1, LCM, CLIP B/16, DINOv3
- $X2$: SD 1.5, LCM, OpenCLIP B/16, DINOv3
- $X3$: SD 1.5, SD 2.1, CLIP L/14, OpenCLIP B/16
- $X4$: SD 1.5, SD 2.1, CLIP L/14, DINOv3
- $X5$: SD 1.5, SD 2.1, LCM, CLIP L/14, OpenCLIP B/16, DINOv3

Figure 4. **Linear latent editing of animal faces.** We apply the method from Section 4.3 to the AFHQ dataset, which contains three categories: Cat, Dog, and Wild.

B. Editing Examples

B.1. Comparison of editing in different models

In Figure 1 we compare linear editing performed in the latent spaces of SD 1.5, LCM and CLIP ViT-L/14. As shown, diffusion latents allow faithful modification of the original image, whereas CLIP edits produce new images that satisfy the target attribute but do not preserve the input. The inversion of CLIP embeddings was done using the UnCLIP variant of Stable Diffusion [1, 9].

B.2. Quantitative analysis of editing

Figure 2 shows quantitative results of our editing method. The x-axis represents attribute intensity, a normalized measure derived from the distance to the classifier's decision plane (x = 0 corresponds to editing to the decision plane, not no editing). Edits are performed on SD 1.5 latents. Panel (a) shows cosine similarity between the edited and original images, while panel (b) shows similarity between the edited image and the CLIP text embedding of the attribute name.

As intensity increases, similarity to the target attribute text embedding increases, indicating successful controlled editing. The similarity to the original image peaks at zero intensity.

B.3. Effect of model scale, conditioning, and pixel space

In Figure 3 (left), we compare linear attribute classification across different representations. As a baseline, we evaluate pixel space, as well as latent spaces from several diffusion models: Stable Diffusion 1.5 (SD 1.5), a version fine-tuned on CelebA (SD CelebA [11]), a smaller unconditional model trained only on CelebA (CelebA Diff [3]), and a larger model (SDXL [7]).

Pixel-space representations yield substantially lower performance compared to generative latent spaces. The smaller model exhibits a clear degradation in linear separability, while increasing model scale (SDXL vs. SD 1.5/2.1) leads to only marginal improvements. Notably, fine-tuning on CelebA reduces linear separability even on the same dataset, highlighting the importance of broad and diverse training. Although CLIP influences training, it affects all samples uniformly at inference due to the use of empty-prompt DDIM inversion and generation.

B.4. Evaluation on additional datasets

To assess generalization beyond CelebA, we evaluate on AFHQ, which contains diverse animal faces [2]. The collection of animal-face images spans three categories: Cat, Dog, and Wild. We add more granular labels to the Wild category using CLIP score with dominant class labels. As shown in Figure 3 (right), semantic categories remain structured and linearly separable across species. We perform pairwise classification, with sub-categories defined via CLIP prompts.

To demonstrate that our latent editing procedure generalizes beyond human faces, we apply the method from Section 4.3 to this dataset. We shift the latents of images along the classifier's direction, following the procedure described in the main paper. This produces realistic edits that preserve the structure of the original image, indicating that the learned directions capture shared high-level semantic information despite the dataset's visual diversity. Figure 4 shows representative edits (towards the Dog class), illustrating attribute manipulation and confirming that the linearity assumption holds well in this domain.

References

[1] Stability AI. Stable diffusion 2.1 unclip (small). https://huggingface.co/stabilityai/stable-diffusion-2-1-unclip-small, 2022. 3

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 3

[3] Google. ddpm-celebahq-256. Hugging Face model. 3

[4] 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. 1

[5] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1

[6] mlfoundations. OpenCLIP. https://github.com/mlfoundations/open_clip, 2021. OpenCLIP: Open reproduction of CLIP training, Github page. 1

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[8] 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. 1

[9] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[10] 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. 1

[11] SG161222. Realistic vision v6.0 b1 novae. Hugging Face. 3

[12] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 1