Automated Labeling of Semantic Directions in Diffusion Model Latent Spaces

Roman Mekashirskiy [1]

Abstract

Unsupervised methods such as PCA applied to the bottleneck activations (h-space) of diffusion models reveal interpretable semantic directions that control facial attributes like pose, gender, and hair style. However, these directions remain unlabeled: a human must visually inspect edited images to assign meaning. We propose a fully automated pipeline that discovers h-space directions via Incremental PCA and labels them using CLIP zero-shot classification and BLIP-2 visual question answering. On a DDPM trained on CelebA-HQ, our CLIP-based approach assigns semantically meaningful labels to the top 10 principal components with a mean confidence score of $|\Delta S| = 0.47$ (logit-scaled). We find that deterministic DDIM sampling (η=0) produces $2.1\times$ higher labeling confidence than stochastic DDPM (η=1), though at the cost of reduced label diversity due to attribute entanglement. We further characterize the sensitivity of labeling confidence to edit strength α and propose a specific-label heuristic that mitigates gender-attribute entanglement. Code is available at `https://github.com/ramzzes13/direction-in-latent`.

1. Introduction

Denoising Diffusion Probabilistic Models (Ho et al., 2020) generate high-quality images by iteratively denoising Gaussian noise through a learned reverse process. A growing body of work has shown that the internal representations of these models—particularly the bottleneck activations of the U-Net architecture—encode rich semantic structure (Kwon et al., 2023; Haas et al., 2024). Haas et al. (2024) demonstrated that applying Principal Component Analysis (PCA) to the so-called h-space (U-Net bottleneck activations collected across multiple generation runs) yields disentangled editing directions: adding a scaled principal component $\alpha \cdot v_k$ to the bottleneck activations during generation modifies a specific facial attribute such as pose, gender, or lighting.

A fundamental limitation of this unsupervised approach is that the discovered directions carry no semantic labels. As Haas et al. (2024) note, "their effects must be interpreted manually"—a human must generate positive and negative edits, visually compare them, and assign a textual description. This manual interpretation is subjective, time-consuming, and inherently unscalable: it becomes impractical when analyzing hundreds of directions or when applying the method to new datasets and models.

This problem parallels the earlier challenge in GAN latent space analysis, where methods like GANSpace (Härkönen et al., 2020) and SeFa (Shen & Zhou, 2021) similarly produce unlabeled directions. Solutions in the GAN setting have relied either on pre-trained attribute classifiers trained on labeled data (Shen et al., 2020) or on manual inspection (Voynov & Babenko, 2020). Neither approach transfers directly to diffusion models without adaptation.

We address this gap by proposing an automated labeling pipeline that takes a pre-trained diffusion model, discovers latent directions via Incremental PCA, generates positive/negative edit pairs, and automatically assigns textual labels using two complementary strategies:

1. **CLIP zero-shot classification**: We measure the change in CLIP (Radford et al., 2021) similarity between edited images and a predefined attribute vocabulary. The attribute with the largest absolute similarity delta is assigned as the label.
2. **VLM difference captioning**: We use BLIP-2 (Li et al., 2023) visual question answering to independently describe positive and negative images, then analyze the textual differences to produce an open-ended label.

Our contributions are:

- A complete, open-source pipeline that automates the

[1]Independent Researcher. Correspondence to: Roman Mekashirskiy <mekashirskiy@example.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

discovery-to-labeling workflow for h-space directions in diffusion models.

- Quantitative analysis of CLIP-based labeling across 10 principal components, including per-seed consistency, edit strength sensitivity, and the effect of sampling stochasticity (η).
- A specific-label heuristic that addresses the gender-attribute entanglement problem observed in CelebA-HQ directions.
- Qualitative evaluation of VLM-based open-ended labeling as a complement to fixed-vocabulary classification.

2. Related Work

Diffusion models. Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) define a forward process that gradually adds Gaussian noise to data and a learned reverse process that denoises it. Denoising Diffusion Implicit Models (DDIMs) (Song et al., 2021a) generalize this to a family of non-Markovian processes parameterized by η, where η=0 yields deterministic sampling and η=1 recovers stochastic DDPM. Song et al. (2021b) unify diffusion models under the framework of score-based stochastic differential equations. Dhariwal & Nichol (2021) demonstrate that diffusion models surpass GANs on image synthesis benchmarks. Latent Diffusion Models (Rombach et al., 2022) apply the diffusion process in a compressed latent space for computational efficiency.

Semantic latent spaces in diffusion models. Kwon et al. (2023) identify an asymmetric reverse process (Asyrp) that reveals a semantic latent space (h-space) in unconditional diffusion models. Haas et al. (2024) extend this by applying PCA to bottleneck activations collected over many generation runs, showing that the resulting principal components correspond to interpretable editing directions. Text-guided approaches such as DiffusionCLIP (Kim et al., 2022), Prompt-to-Prompt (Hertz et al., 2023), and Imagic (Kawar et al., 2023) achieve semantic editing through text conditioning, but require specifying the desired attribute in advance rather than discovering it.

GAN latent space analysis. GANSpace (Härkönen et al., 2020) applies PCA to intermediate GAN activations to find interpretable directions—the direct ancestor of the h-space PCA approach. SeFa (Shen & Zhou, 2021) obtains editing directions via closed-form factorization of the generator's weight matrix. InterFaceGAN (Shen et al., 2020) uses pre-trained binary attribute classifiers to find linear boundaries in GAN latent spaces that separate attribute values, enabling targeted editing but requiring labeled training data. Voynov & Babenko (2020) train a direction reconstructor network for unsupervised direction discovery. All of these methods either require labeled data for training classifiers or rely on manual interpretation of the discovered directions.

Vision-language models for image understanding. CLIP (Radford et al., 2021) learns a joint embedding space for images and text through contrastive pre-training on 400M image-text pairs, enabling zero-shot classification by comparing image embeddings to text embeddings of candidate labels. BLIP-2 (Li et al., 2023) bridges frozen image encoders and large language models through a lightweight Querying Transformer, enabling visual question answering without task-specific fine-tuning. We use both models as off-the-shelf tools for automated labeling—CLIP for fixed-vocabulary classification and BLIP-2 for open-ended captioning.

3. Method

Our pipeline consists of three stages: direction discovery, edit generation, and automated labeling. Figure 1 shows representative output from the full pipeline.

3.1. Stage 1: Direction Discovery

Following Haas et al. (2024), we extract semantic directions from the bottleneck activations of a pre-trained unconditional DDPM.

Let f_θ denote a U-Net denoiser with bottleneck (mid-block) activations $h_t \in \mathbb{R}^{C\times H\times W}$ at diffusion timestep t. For N random seeds, we run the full DDIM reverse process with T steps, collecting activations $\{h_t^{(i)}\}_{i=1}^N$ at each timestep.

For each timestep t, we flatten $h_t^{(i)}$ to a vector in $\mathbb{R}^{CHW}$ and fit Incremental PCA (Pedregosa et al., 2011) with batch size B to extract K principal components:

$$v_k^{(t)} = \mathrm{PCA}_k\big(\{h_t^{(1)}, \ldots, h_t^{(N)}\}\big), \quad k = 1, \ldots, K. \quad (1)$$

The full direction for component k is the sequence $v_k = (v_k^{(1)}, \ldots, v_k^{(T)}) \in \mathbb{R}^{T\times C\times H\times W}$.

In our experiments, the bottleneck has C=512, H=W=8 (for 256×256 generation), so each flattened activation is 32,768-dimensional. We use N=500 samples, K=20 components, T=50 DDIM steps, and batch size B=50.

3.2. Stage 2: Edit Generation

Given a direction v_k and edit strength $\alpha > 0$, we generate a triplet $(I_{\text{orig}}, I_{\text{pos}}, I_{\text{neg}})$ from a fixed seed s:

$$I_{\text{orig}} = \text{DDIM}(x_T^{(s)}), \tag{2}$$

$$I_{\text{pos}} = \text{DDIM}(x_T^{(s)};\ h_t \leftarrow h_t + \alpha \cdot v_k^{(t)}), \tag{3}$$

$$I_{\text{neg}} = \text{DDIM}(x_T^{(s)};\ h_t \leftarrow h_t - \alpha \cdot v_k^{(t)}), \tag{4}$$

where $x_T^{(s)}$ is the initial noise determined by seed s, and the notation $h_t \leftarrow h_t + \delta$ indicates adding δ to the bottleneck output at timestep t via a forward hook on the U-Net mid-block.

All three images share the same initial noise and diffusion trajectory up to the bottleneck modification, ensuring that differences between I_{pos} and I_{neg} reflect the semantic content of v_k rather than stochastic variation. We use deterministic DDIM (η=0) for this reason (see Section 4.6 for analysis of the effect of η).

3.3. Stage 3A: CLIP Zero-Shot Classification

We define a vocabulary of M=20 candidate attributes $\{a_1, \ldots, a_M\}$ expressed as natural language phrases (e.g., "a smiling person", "a person with blonde hair"). For each attribute a_j, we compute the logit-scaled CLIP similarity delta:

$$\Delta S_{k,j} = \lambda \cdot \big[\cos(f_I(I_{\text{pos}}), f_T(a_j)) - \cos(f_I(I_{\text{neg}}), f_T(a_j))\big], \tag{5}$$

where f_I and f_T are the CLIP image and text encoders, $\cos(\cdot,\cdot)$ denotes cosine similarity, and $\lambda = \exp(\tau)$ is CLIP's learned logit scale ($\lambda \approx 100$ for ViT-B/32).

We average $\Delta S_{k,j}$ over S=5 seeds per direction. The top label for direction k is:

$$\hat{a}_k = \arg\max_j |\overline{\Delta S}_{k,j}|. \tag{6}$$

The sign of $\overline{\Delta S}_{k,\hat{a}_k}$ indicates the polarity: positive means the attribute increases from negative to positive edit.

Specific-label heuristic. We observe that "a female person" and "a male person" dominate the top label for many directions due to the strong correlation of gender with other facial attributes in CelebA-HQ. To address this, we define a *specific label* by excluding gender attributes from the ranking and selecting the next highest-scoring attribute.

3.4. Stage 3B: VLM Difference Captioning

As a complementary open-ended approach, we use BLIP-2 (Li et al., 2023) to independently describe I_{pos} and I_{neg} through targeted visual question answering prompts such as "Describe this person's appearance briefly." We then extract content words from each caption (removing stop words), compute the set difference between positive and negative captions across seeds, and report frequently gained/lost words as the consensus label.

4. Experiments

4.1. Setup

Model. We use the pre-trained `google/ddpm-celebahq-256` model from the Hugging Face `diffusers` library, an unconditional DDPM trained on CelebA-HQ (Karras et al., 2018) at 256×256 resolution. The U-Net has a mid-block with 512 channels and 8×8 spatial resolution.

Sampling. We use a DDIM scheduler (Song et al., 2021a) with T=50 steps. Our primary results use η=0 (deterministic DDIM); we also compare against η=1 (stochastic DDPM) in Section 4.6.

Direction discovery. We collect bottleneck activations from N=500 samples and fit Incremental PCA with K=20 components using batch size 50.

Edit generation. For the main experiments, we generate edit triplets for the top K'=10 directions using S=5 seeds (42, 123, 256, 789, 1024) at α=5.0.

CLIP labeling. We use OpenCLIP (Ilharco et al., 2021) ViT-B/32 pre-trained on OpenAI's CLIP data. The attribute vocabulary consists of 20 phrases covering rotation, expression, gender, accessories, age, hair properties, and eye/mouth state (full list in Appendix A).

VLM labeling. We use BLIP-2 with the OPT-2.7B language model backbone (Li et al., 2023), loaded from the Hugging Face Hub (`Salesforce/blip2-opt-2.7b`).

4.2. Experiment 1: PCA Direction Discovery

We verify that PCA on h-space activations recovers meaningful semantic directions. Figure 2 shows the explained variance ratio for the top 20 components. The first component explains 5.70% of variance (averaged over timesteps), and the top 10 components together explain 29.4%. The low per-component variance is expected: the h-space is 32,768-dimensional, and many directions encode subtle or localized variations. 21 components are needed to capture 90% of the variance.

Appendix C shows how PC0's explained variance varies across timesteps. The variance is concentrated in early-to-mid timesteps (indices 0–25), consistent with the finding

of Haas et al. (2024) that coarse semantic structure is determined early in the reverse process.

4.3. Experiment 2: CLIP-Based Labeling

Table 1 reports the CLIP labeling results for the top 10 directions. The mean absolute CLIP delta score across directions is $|\overline{\Delta S}| = 0.470$, with a standard deviation of 0.224. The top direction (PC0) achieves the strongest score ($|\Delta S| = 0.872$) and is labeled "a female person", reflecting a gender/appearance axis that is the dominant mode of variation in CelebA-HQ.

Label diversity. Out of 10 directions, only 5 unique labels are assigned. "A female person" appears 5 times, reflecting the strong entanglement of gender with other attributes (hair color, hair length, facial structure). The specific-label heuristic (which excludes gender from the ranking) assigns "a person with blonde hair" as the specific label for most gender-dominated directions, revealing the correlated attribute.

Per-seed consistency. The top label for PC0 is consistent across only 2 out of 5 seeds (40%), indicating that while the mean delta is large, individual seeds may show different dominant attributes depending on the face identity being edited. Lower-scoring directions (e.g., PC9 with $|\Delta S| = 0.147$) show even less consistency. Figure 6 visualizes the consistency across all directions. This low agreement rate (averaging $\sim$20% across directions) highlights a fundamental challenge: the same h-space perturbation produces different perceptual effects on different face identities, and CLIP's attribute ranking is sensitive to these identity-specific variations.

4.4. Experiment 3: VLM Difference Captioning

BLIP-2 captions for each direction are reported in the appendix. The VLM approach produces verbose, image-specific descriptions that capture fine-grained details (e.g., "a woman with blonde hair and green eyes") but struggle to distill the consistent semantic change across seeds. The consensus labeling (Section 3.4) extracts gained/lost words; for example, PC0's consensus label is "+necklace, hair, green, eye, blonde / −glow, way, hand, long, beard", which partially captures the gender-related transformation but mixes in irrelevant details.

The VLM approach is more expressive than CLIP's fixed vocabulary but less reliable for automated labeling: it requires aggregation heuristics to extract a single label from heterogeneous captions. We view it as a useful diagnostic complement rather than a standalone labeling method.

4.5. Experiment 4: Sensitivity to Edit Strength

We vary the edit strength $\alpha \in \{1, 2, 3, 5, 7, 10\}$ and measure the CLIP labeling confidence for the top 5 directions using 3 seeds per direction. Figure 4 shows that:

1. CLIP confidence increases monotonically with α: mean $|\Delta S|$ rises from 0.090 at α=1 to 0.450 at α=10.
2. The relationship is approximately logarithmic, with diminishing returns above α=5.
3. Individual directions exhibit different sensitivity curves, with some (e.g., PC3: "bald person") reaching saturation earlier.

At α=1, no direction achieves $|\Delta S| > 0.15$, making automated labeling unreliable. At α=10, the edits may introduce visual artifacts, though the CLIP score continues to increase. We recommend $\alpha \in [3, 7]$ as the operating range for reliable labeling.

4.6. Effect of Sampling Stochasticity (η)

Table 2 compares CLIP labeling results under deterministic (η=0) and stochastic (η=1) sampling.

Deterministic DDIM (η=0) produces:

- Higher mean confidence: $|\overline{\Delta S}| = 0.470$ vs. 0.227 ($2.1\times$).
- Lower label diversity: 5 unique labels vs. 9 unique labels.

Stochastic DDPM (η=1) assigns a wider variety of labels (including "smiling", "glasses", "bald", "hat") but with weaker confidence scores, because stochastic noise in the sampling process partially masks the effect of the h-space edit.

This creates a diversity-confidence trade-off: deterministic sampling gives clean, high-confidence labels but many directions collapse to the same dominant attribute (gender), while stochastic sampling reveals more diverse attributes at the cost of noisier scores. A practical strategy is to use η=0 with the specific-label heuristic to get both high confidence and attribute diversity.

5. Results

6. Discussion

Attribute entanglement. The most salient finding is the prevalence of gender-related labels: 5 out of 10 directions are labeled "a female person" under CLIP. This reflects genuine entanglement in the CelebA-HQ data distribution,

Table 1. CLIP labeling results for the top 10 h-space directions (DDIM, η=0, α=5.0). ΔS is the logit-scaled CLIP similarity delta averaged over 5 seeds. Specific label excludes gender attributes.

PC	Top Label	ΔS	Specific Label	ΔS
0	female	-0.872	blonde hair	-0.731
1	blonde hair	-0.169	blonde hair	-0.169
2	female	-0.778	blonde hair	-0.634
3	female	-0.587	blonde hair	-0.444
4	female	$+0.344$	blonde hair	$+0.216$
5	long hair	-0.359	long hair	-0.359
6	blonde hair	-0.553	blonde hair	-0.553
7	bangs	-0.438	bangs	-0.438
8	female	$+0.456$	blonde hair	$+0.419$
9	rotated left	-0.147	rotated left	-0.147
Mean $\lvert\Delta S\rvert$		0.470		0.411
Unique labels		5/10		4/10

Table 2. Effect of sampling stochasticity on CLIP labeling. η=0 (DDIM) vs. η=1 (DDPM), α=5.0.

	$\eta = 0$ (DDIM)		$\eta = 1$ (DDPM)	
PC	Label	$\lvert\Delta S\rvert$	Label	$\lvert\Delta S\rvert$
0	female	0.872	blonde hair	0.281
1	blonde hair	0.169	female	0.219
2	female	0.778	smiling	0.134
3	female	0.587	bangs	0.291
4	female	0.344	eyes closed	0.484
5	long hair	0.359	glasses	0.259
6	blonde hair	0.553	bald	0.231
7	bangs	0.438	long hair	0.156
8	female	0.456	glasses	0.072
9	rotated left	0.147	hat	0.137
Mean		0.470		0.227
Unique		5		9

where gender correlates strongly with hair length, hair color, skin texture, and facial structure. PCA extracts the direction of maximum variance, which in this dataset is dominated by gender-correlated visual features. This entanglement is not unique to our approach—InterFaceGAN (Shen et al., 2020) similarly observes that gender boundaries in GAN latent spaces are entangled with other attributes. The specific-label heuristic partially addresses this by revealing the correlated attribute (e.g., "blonde hair"), but it cannot fully disentangle the underlying factors. Independent Component Analysis (ICA) or supervised disentanglement methods may be needed for cleaner separation.

Comparison with GAN-based methods. Our pipeline is conceptually analogous to applying GANSpace (Härkönen et al., 2020) plus an automated labeling layer. The key differences are: (1) the h-space of diffusion models has different geometric properties than GAN intermediate layers—our directions are per-timestep, requiring aggregation across T=50 steps; (2) diffusion model edits are applied additively to the bottleneck activations rather than to the input latent code; and (3) deterministic DDIM sampling enables exact image-pair comparisons, which is impossible in unconditional stochastic generation. The $2.1\times$ confidence improvement from η=0 (Table 2) directly demonstrates the importance of deterministic sampling for reliable automated labeling.

Table 3. CLIP labeling confidence (mean $\lvert\Delta S\rvert$ over top 5 directions) as a function of edit strength α.

α	1.0	2.0	3.0	5.0	7.0	10.0
Mean $\lvert\Delta S\rvert$	0.090	0.211	0.237	0.364	0.406	0.450

CLIP vocabulary sensitivity. The fixed vocabulary constrains what CLIP can detect. If the vocabulary omits a relevant attribute (e.g., "background brightness"), the labeling will assign the closest available label, potentially introducing errors. Expanding the vocabulary helps but introduces a multiple-comparisons problem: with more candidates, chance correlations increase. The VLM approach avoids this issue but introduces its own challenges (caption aggregation, hallucination). A hybrid strategy—using CLIP for confident detections and falling back to VLM for low-confidence directions—could combine the strengths of both approaches.

Scalability. Stage 1 (PCA) takes approximately 25 minutes for 500 samples on a single GPU. Stage 2 (edit generation) takes approximately 30 minutes for 50 triplets. Stage 3 (CLIP + VLM labeling) completes in under 5 minutes. The entire pipeline runs in approximately 1 hour, making it practical for routine analysis of new models or datasets. In contrast, manual labeling of 10 directions requires generating and visually inspecting at least 30 image triplets (3 per direction with redundancy), typically taking 30–60 minutes of concentrated human effort that must be repeated for each model.

Limitations. Our evaluation is limited to CelebA-HQ faces, a domain where facial attributes are well-defined and relatively easy to detect. The method may be less effective on datasets with more abstract or less visually salient variations (e.g., LSUN Churches, where "style" or "era" may be harder for CLIP to distinguish). The per-seed consistency of CLIP labeling is low (often $\leq 40\%$ agreement; see Figure 6), suggesting that individual faces respond differently to the same h-space perturbation. This low consistency is partly an artifact of using absolute-maximum selection: if two attributes have similar $\lvert\Delta S\rvert$ values, the top label may flip between seeds. A soft labeling scheme that reports the top-k attributes with their confidence intervals

would be more informative. Additionally, the VLM-based approach produces noisy outputs that require significant post-processing to yield a single label, and our simple word-frequency heuristic for consensus extraction is limited compared to what a more sophisticated NLP pipeline could achieve.

7. Conclusion

We presented an automated pipeline for discovering and labeling semantic directions in the h-space of diffusion models. By combining PCA-based direction discovery with CLIP zero-shot classification and VLM-based captioning, we transform manually-interpreted editing directions into self-describing ones. Our experiments on CelebA-HQ demonstrate that CLIP can assign meaningful labels with reasonable confidence, though attribute entanglement remains a challenge. We characterized the effects of edit strength and sampling stochasticity on labeling quality, providing practical guidance for future applications of this pipeline. Future work should extend the evaluation to non-face datasets, explore disentanglement methods beyond PCA, and investigate whether larger CLIP models (e.g., ViT-L/14) improve labeling accuracy and consistency.

References

Dhariwal, P. and Nichol, A. Q. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Haas, R., Huberman-Spiegelglas, I., Mulayoff, R., Grasshof, S., Brandt, S. S., and Michaeli, T. Discovering interpretable directions in the semantic latent space of diffusion models. In *Proceedings of the 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. GANSpace: Discovering interpretable GAN controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. OpenCLIP. *Zenodo*, 2021. doi: 10.5281/zenodo.5143773.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Kim, G., Kwon, T., and Ye, J. C. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2426–2435, 2022.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations (ICLR)*, 2023.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Shen, Y. and Zhou, B. Closed-form factorization of latent semantics in GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Shen, Y., Gu, J., Tang, X., and Zhou, B. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Voynov, A. and Babenko, A. Unsupervised discovery of interpretable directions in the GAN latent space. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing, 2020.

A. Attribute Vocabulary

The 20 candidate attributes used for CLIP zero-shot classification:

1. a face rotated to the left
2. a face rotated to the right
3. a smiling person
4. a frowning person
5. a male person
6. a female person
7. a person wearing glasses
8. a person without glasses
9. an older person
10. a younger person
11. a person with blonde hair
12. a person with dark hair
13. a bald person
14. a person with long hair
15. a person with bangs
16. a person with a hat
17. a person with eyes closed
18. a person with eyes open
19. a person with mouth open
20. a person with mouth closed

B. VLM Captioning Details

BLIP-2 is prompted with three targeted visual question answering prompts per image:

1. "Describe this person's appearance briefly."
2. "What stands out about this person's face?"
3. "Describe the hair, expression, and accessories."

The consensus label is computed by extracting content words (nouns, adjectives, verbs) from each caption, computing set differences between positive and negative images across all seeds, and reporting words that appear in at least $\lfloor S/3 \rfloor$ seed comparisons (where S is the number of seeds).

Representative VLM consensus labels:

- PC0: +necklace, hair, green, eye, blonde / −glow, way, hand, long, beard
- PC1: +short, woman, dress, blouse, wearing / −blonde, green, hair, think
- PC5: +woman, dark, hair / −blonde, man

C. PCA Variance by Timestep

Figure 7 shows the explained variance for the top 5 components as a function of the DDIM timestep index. Early timesteps (high noise level) show the most structured variation, consistent with the observation that coarse semantic structure is determined early in the reverse diffusion process (Haas et al., 2024; Kwon et al., 2023).

D. Edit Magnitude Analysis

Figure 8 shows the mean L2 pixel difference between edited and original images for each direction. PC0 shows the largest edit magnitude (mean L2 ≈ 2.80), consistent with it capturing the most variance. Positive and negative edits produce similar magnitudes, confirming the symmetry of linear h-space perturbations.

E. Implementation Details

The pipeline is implemented in Python using:

- **diffusers** (Wolf et al., 2020): DDPM model loading and DDIM scheduling
- **scikit-learn** (Pedregosa et al., 2011): Incremental PCA
- **OpenCLIP** (Ilharco et al., 2021): CLIP ViT-B/32
- **transformers** (Wolf et al., 2020): BLIP-2 model loading

The U-Net bottleneck hook intercepts the output of `unet.mid_block` via PyTorch's `register_forward_hook` API. In "capture" mode, it records activations; in "edit" mode, it adds $\alpha \cdot v_k^{(t)}$ to the output tensor at each timestep.

All experiments were run on NVIDIA H100 80GB GPUs. The full pipeline (Stage 1: PCA + Stage 2: edits + Stage 3: labeling) completes in approximately 60 minutes.

The complete codebase is available at `https://github.com/ramzzes13/direction-in-latent`.

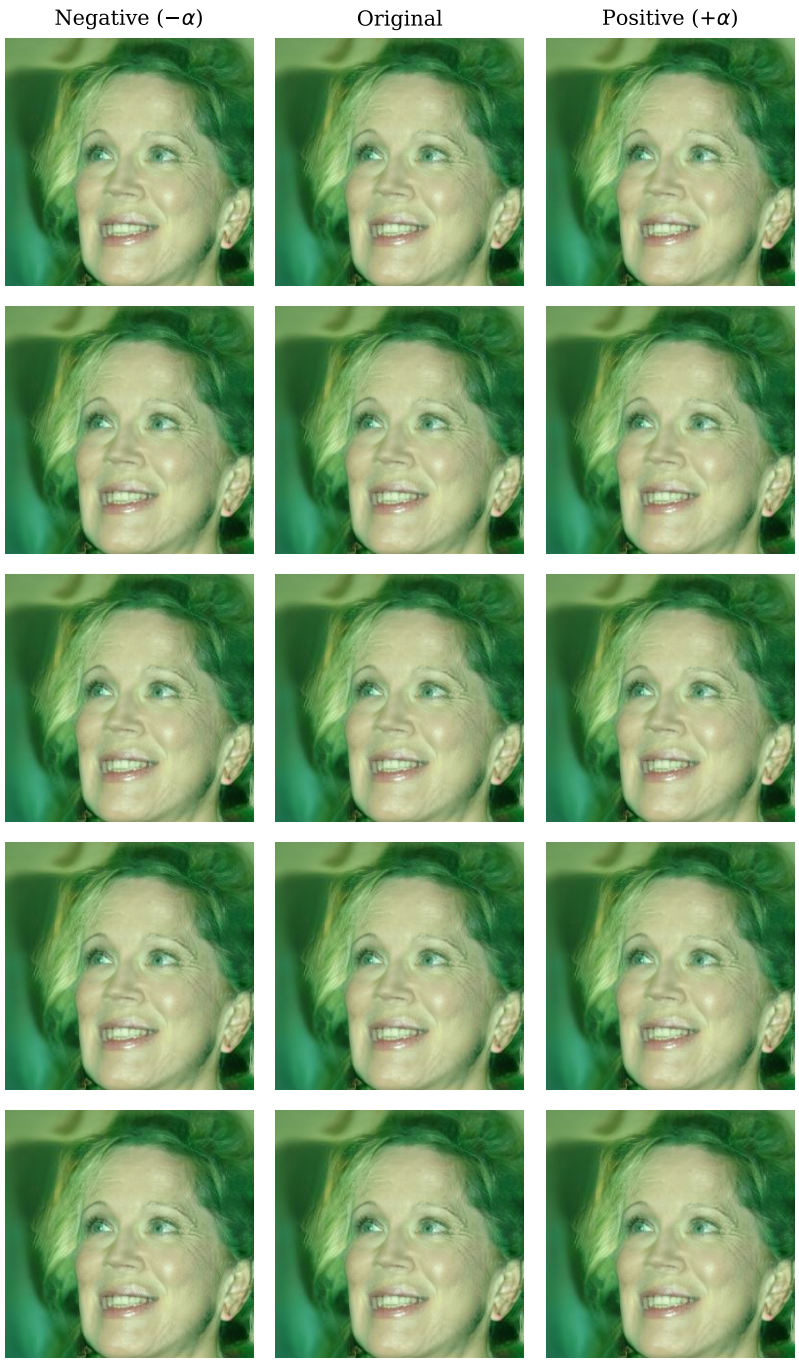

Figure 1. Semantic edits via h-space PCA directions on CelebA-HQ faces. Each row shows one PCA direction (PC0–PC4) applied to a single seed at α=5.0 with DDIM (η=0). Left: negative edit ($-\alpha$), center: original, right: positive edit ($+\alpha$). The CLIP-assigned label is shown on the left.

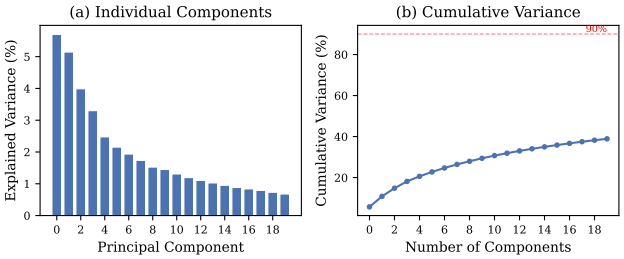

Figure 2. PCA explained variance ratio for the top 20 h-space components, averaged over 50 DDIM timesteps. (a) Individual component variance. (b) Cumulative variance with 90% threshold.

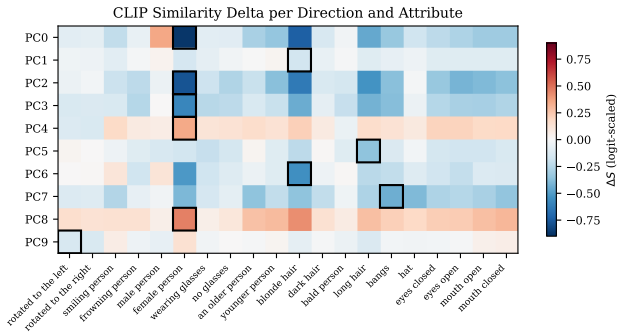

Figure 3. CLIP similarity delta heatmap (ΔS, logit-scaled) for 10 directions $\times$ 20 attributes. Black rectangles mark the top label per direction. Blue indicates positive ΔS (attribute increases with positive edit), red indicates negative.

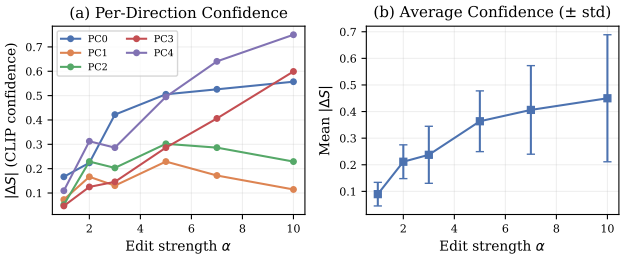

Figure 4. CLIP labeling confidence vs. edit strength α. (a) Per-direction curves for the top 5 components. (b) Mean $|\Delta S|$ with standard deviation. Confidence increases monotonically but saturates above $\alpha \approx 5$.

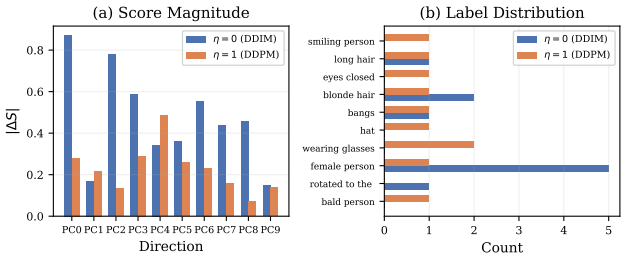

Figure 5. Comparison of DDIM (η=0) vs. DDPM (η=1) sampling. (a) Score magnitude per direction. (b) Label distribution. DDIM gives higher scores but lower label diversity.

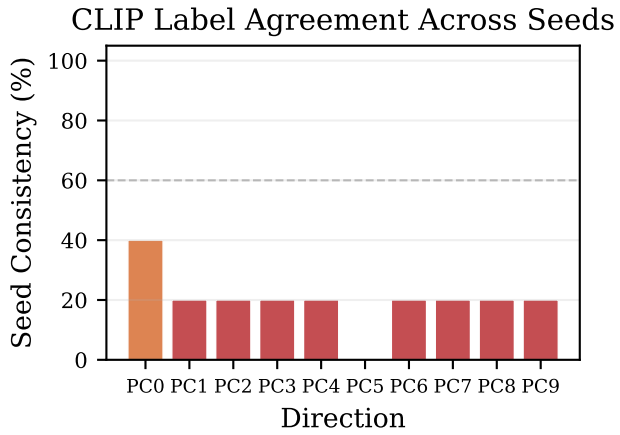

Figure 6. CLIP label agreement across 5 seeds for each direction. Green: $\geq$60% agreement, orange: $\geq$40%, red: $<$40%. Most directions show low consistency, reflecting identity-dependent effects of h-space perturbations.

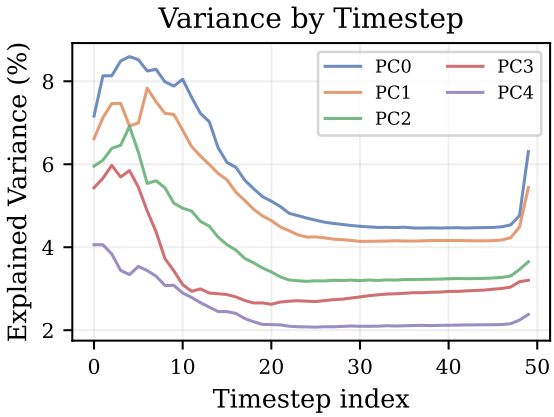

Figure 7. Explained variance ratio by DDIM timestep for the top 5 principal components.

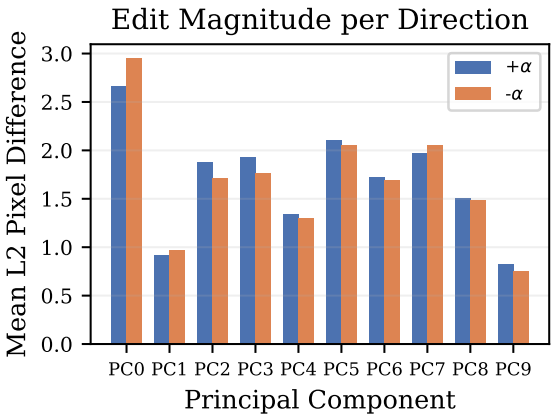

Figure 8. Mean L2 pixel difference between edited and original images for each direction.