DIRECT LATENT EDITING FOR VFM-BASED IMAGE GENERATORS

Anonymous authors
Paper under double-blind review

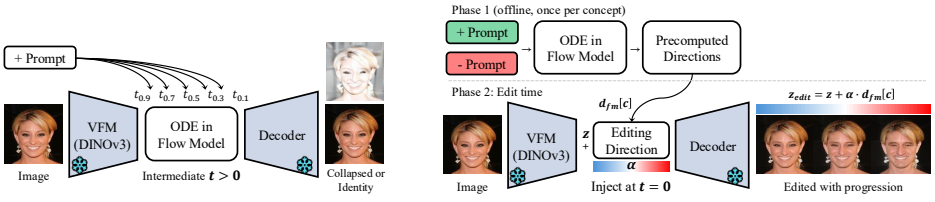
ABSTRACT

Modern text-to-image generators train a flow model in a compressed latent space defined by an encoder. Recent work replaces the standard VAE encoder with a frozen Vision Foundation Model (VFM) such as DINOv3 to exploit self-supervised semantic features. However, editing for these VFM-based generators has not kept pace with their generation quality. We apply four inference-time editors built for VAE-based generators to VFM-based generators, at their published or tuned settings, and find that they rarely reach the target attribute. We propose Direct Latent Editing (DLE), an inference-time editing method for fully VFM-based generators with a frozen encoder. DLE precomputes one editing direction per concept from flow-model samples. At edit time it adds this direction to the encoded latent and decodes; no flow-model call, inversion or weight update is needed. A single scalar gives continuous, signed control. On three VFM-based generators we trained (CelebA faces, MIMIC-CXR chest X-rays and ADNI-1 brain MRI), DLE reaches targets that the tested baselines do not (82.0% vs. at most 48.3% on CelebA; 0.927 vs. at most 0.569 AUROC on MIMIC-CXR), at $203\times$ lower per-edit flow-model and decoder time than FlowEdit at its official configuration.

1 INTRODUCTION

Modern text-to-image generators train a flow model (FM) that maps noise to data in the latent space of an encoder. The standard encoder, a VAE trained for pixel reconstruction, retains spatial structure but encodes low-level detail rather than semantics. Recent work replaces it with a frozen Vision Foundation Model (VFM) such as DINOv2 (Oquab et al., 2024) or DINOv3 (Siméoni et al., 2025) inside rectified-flow (RF) generators, first for class-conditional (Zheng et al., 2025a;b; Shi et al., 2026) and recently for text-to-image generation (Shi et al., 2025). Editing an attribute such as facial expression, or a finding such as cardiomegaly on a chest radiograph or Alzheimer's-associated appearance on a brain MRI slice, is useful for explanation, augmentation of rare concepts and counterfactual synthesis. For VFM-based generators, editing has not kept pace with generation quality (Figure 1).

We take four representative inference-time editors developed for VAE-based generators (Stable Diffusion (Rombach et al., 2022), FLUX (Black Forest Labs, 2024)): SDEdit (Meng et al., 2022),



(a) Baseline editing methods (b) DLE (ours)

Figure 1: **Conceptual comparison of DLE versus existing editing methods.** (a) Baselines inject the editing signal at intermediate steps of the sampling ODE, for every edit. (b) DLE precomputes per-concept editing directions from prompt pairs offline (Phase 1) and adds the scaled direction to the encoded latent before decoding (Phase 2), producing a progressive edit.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

SEGA (Brack et al., 2023), Concept Sliders (Gandikota et al., 2024) and FlowEdit (Kulikov et al., 2025). All four inject the edit at intermediate steps of the *sampling ODE* that carries noise z_1 to data z_0 in T steps (RF convention: $t{=}1$ noise, $t{=}0$ data). On our VFM-based generators for faces and chest X-rays (CXR) they rarely edit, whether at their published settings or after tuning on a development split. We attribute this to the latent, not the editors: in a VFM latent a concept is more linearly readable and better explained by one translation, and the CNN decoder keeps edits local (§3.4), so a single $t{=}0$ translation suffices; in a matched VAE latent SDEdit and FlowEdit work and DLE falls behind (§4.3).

Our method. We propose **Direct Latent Editing (DLE)**. For each concept it precomputes an editing direction from FM-generated samples. At edit time it encodes the source image (real or FM-generated), adds the direction at the sampling endpoint $t{=}0$, and decodes; no FM call, inversion or weight update is involved. Its flow-model and decoder time per edit is $203\times$ lower than FlowEdit's at its official configuration (Appendix F); the one-time cost per concept is recovered after about 282 edits.

Contributions.

- We introduce *Direct Latent Editing* (DLE), an inference-time editor for fully VFM-based generators, a frozen encoder with an RF flow model in its latent, where prior methods assume a VAE encoder.
- DLE precomputes one direction per concept from FM samples under a positive/negative prompt pair and adds it to the encoded latent at $t{=}0$. A single scalar gives continuous, signed control; edit time involves no flow-model evaluation, inversion or weight update.
- On three separately trained VFM-based generators (CelebA face attributes, MIMIC-CXR findings, ADNI-1 brain MRI), DLE reaches targets that the tested baselines do not (82.0% vs. at most 48.3% on held-out CelebA pairs; 0.927 vs. at most 0.569 AUROC on MIMIC-CXR), at $203\times$ lower per-edit flow-model and decoder time than FlowEdit at its official configuration. In a matched VAE-based generator the ordering reverses (§4.3), consistent with the account of §3.4.

2 Related Work

VFM-based generators. VFMs (Oquab et al., 2024; Siméoni et al., 2025) are self-supervised image encoders; recent work reuses their patch-token space as a generative latent: REPA (Yu et al., 2025) adds VFM alignment as an auxiliary loss to a VAE-based DiT (Peebles & Xie, 2023), while VFMTok (Zheng et al., 2025a) and RAE (Zheng et al., 2025b) replace the VAE with a frozen VFM encoder. Subsequent work returns to VAE alignment (iREPA (Singh et al., 2026)), moves alignment to pixel space (PixelREPA (Shin et al., 2026) on JiT (Li & He, 2025)), or drops the VFM (JiT); all evaluate *class-conditional* generation with image-level metrics.

Inference-time editing in diffusion/flow generators. SDEdit (Meng et al., 2022), SEGA (Brack et al., 2023), Concept Sliders (Gandikota et al., 2024) and FlowEdit (Kulikov et al., 2025) were developed on VAE-based backbones and modify the sampling ODE at intermediate $t > 0$ (Appendix A.4). Linear latent directions have been used in CAV/TCAV (Kim et al., 2018), InterFaceGAN (Shen et al., 2020) (an SVM normal) and CASteer (Gaintseva et al., 2026); DLE applies a paired mean difference at the sampling endpoint of a VFM latent. The closest concurrent work, PS-VAE (Zhang et al., 2025b), also targets DINOv2-based editing but unfreezes and retrains the encoder; DLE keeps it frozen. None targets a generator whose encoder is a frozen VFM and whose flow model lives in that latent; there, the four VAE-era baselines rarely edit even after tuning on a development split (Section 4, Appendix A.5).

3 Method: Direct Latent Editing (DLE)

DLE has two phases (Figure 2): a one-time extraction of a per-concept editing direction from a pretrained FM (§3.2), and a per-image edit that costs one encoder and one decoder forward (§3.3).

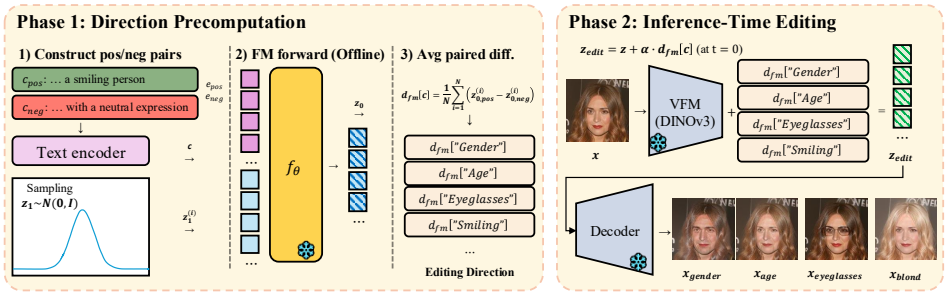


Figure 2: **DLE method overview.** Phase 1 precomputes a per-concept editing direction $d_{fm}[c]$ from positive and negative prompt FM samples. Phase 2 adds $\alpha\, d_{fm}[c]$ to the encoded source latent and decodes.

3.1 Preliminaries

Our pipeline is a text-to-image (T2I) generator trained with RF (Liu et al., 2023), following the VFM-as-encoder design (Shi et al., 2026; 2025) where the standard VAE is replaced by a frozen VFM encoder. RF trains a flow model f_θ to predict the velocity field that transports a noise sample z_1 to a data sample z_0 along a straight path in the VFM latent space; sampling integrates this velocity ODE from $t{=}1$ (noise) to $t{=}0$ (data). The pipeline has four components; for each we state whether it is shared or trained per domain (CelebA, CXR, brain MRI):

- **Encoder** E (shared, frozen): a self-supervised VFM that splits the image into an $h{\times}w$ grid of patches and outputs one D-dim semantic feature vector per patch.
- **Decoder** $\mathcal{D}$ (per-domain): a CNN trained to reconstruct x from $z{=}E(x)$.
- **Flow model** f_θ (per-domain): a velocity field $v{=}f_\theta(z_t, t, e)$ on the VFM latent space, where z_t is the latent at flow time $t \in [0, 1]$ (RF convention: z_1 noise, z_0 data) and e is the text embedding.
- **Text encoder** (per-domain, frozen): FaRL (Zheng et al., 2022) for CelebA, CXR-BERT (Boecking et al., 2022) for CXR, BiomedCLIP (Zhang et al., 2025a) for brain MRI.

At sampling we use classifier-free guidance (CFG) (Ho & Salimans, 2022) with scale $\omega{=}1.5$ (Appendix A.3).

3.2 Phase 1: Direction Precomputation

Given a pretrained FM, we precompute an editing direction $d_{fm}[c]$ for each desired *concept* c, where "concept" refers to a binary attribute on faces (e.g., Smiling), a clinical finding on chest radiographs (e.g., Cardiomegaly), or a diagnostic direction on brain MRI (e.g., Alzheimer's Disease, AD). The procedure (Algorithm 1, Phase 1) has three steps:

- **(i) Pick a positive and a negative prompt.** c_{pos} describes the concept (for *Smiling*, "a photo of a smiling person") and c_{neg} its absence or opposite ("a photo of a neutral face person"); the per-domain text encoder embeds them as e_{pos} and e_{neg}.
- **(ii) Generate N paired endpoint latents.** From each noise $z_1^{(i)} \sim \mathcal{N}(0, I)$, run the flow model with the same CFG scale ω once with e_{pos} and once with e_{neg}, giving $(z_{0,\text{pos}}^{(i)}, z_{0,\text{neg}}^{(i)})$, which differ only in the text condition.
- **(iii) Average the paired differences.**

$$d_{fm}[c] \;=\; \frac{1}{N}\sum_{i=1}^{N}\big(z_{0,\text{pos}}^{(i)} - z_{0,\text{neg}}^{(i)}\big). \tag{1}$$

Each pair starts from the same noise, so the difference cancels the noise contribution and approximately isolates the latent shift caused by the prompt change. Equation 1 is the least-squares single translation between the paired endpoints, $\arg\min_d \sum_i \|\Delta_i - d\|^2$ with $\Delta_i{=}z_{0,\text{pos}}^{(i)}{-}z_{0,\text{neg}}^{(i)}$, and averaging over N pairs scales its variance by $1/N$. A separating direction (an SVM normal, as in

Algorithm 1: Direct Latent Editing (DLE)

Data: N: samples per concept; K: ODE steps; ω: classifier-free guidance (CFG) scale

/* **Phase 1.** Direction Precomputation (once per concept, amortized) */

foreach *concept* $c \in \mathcal{C}$ **do**

  $e_{\text{pos}} \leftarrow \text{text_embed}(c_{\text{pos}})$

  $e_{\text{neg}} \leftarrow \text{text_embed}(c_{\text{neg}})$

  for $i \leftarrow 1$ **to** N **do**

    $z_1 \sim \mathcal{N}(0, I)$

    $z_{0,\text{pos},i} \leftarrow \text{ODE_solve}(f_\theta,\, z_1,\, e_{\text{pos}},\, \omega,\, K)$ // Euler, t:1→0

    $z_{0,\text{neg},i} \leftarrow \text{ODE_solve}(f_\theta,\, z_1,\, e_{\text{neg}},\, \omega,\, K)$

  $d_{fm}[c] \leftarrow \frac{1}{N}\sum_{i=1}^{N}\big(z_{0,\text{pos},i} - z_{0,\text{neg},i}\big)$ // tensor of shape $D{\times}h{\times}w$ (D=384, $h{=}w$=16)

/* **Phase 2.** Inference-Time Edit (per image, closed-form) */

Input: image x (real or FM-generated), concept c, strength α

$z \leftarrow E(x)$

$z_{\text{edit}} \leftarrow z + \alpha \cdot d_{fm}[c]$

$x_{\text{edit}} \leftarrow \mathcal{D}(z_{\text{edit}})$

return x_{edit} // FM forward passes: **0**; no inversion; no weight updates

InterFaceGAN) or a variance-maximizing one (PCA) estimates something else (§5, Appendix B.2). $d_{fm}[c]$ is the editing direction for concept c, and $\|d_{fm}[c]\|$ its L_2 norm.

Cost and reuse. With the defaults $N{=}200$, $K{=}50$ Euler steps from $t{=}1$ to $t{=}0$ (denoted ODE_solve in Algorithm 1) and $\omega{=}1.5$, Phase 1 makes $2NK{=}20{,}000$ FM forward passes per concept, once: 16.7 min on one A5000, or 27.9 GPU-hours for 100 concepts (Appendix F). The direction $d_{fm}[c]$ (a $D{\times}h{\times}w$ tensor) is stored and reused for every later edit of concept c.

3.3 Phase 2: Inference-Time Edit Using Precomputed Direction

Given $d_{fm}[c]$ from Phase 1, editing a source image x (real or FM-generated) with strength $\alpha \in \mathbb{R}$ takes one pass (Algorithm 1, Phase 2): encode $z{=}E(x) \in \mathbb{R}^{D\times h\times w}$, add the scaled direction,

$$z_{\text{edit}} \;=\; z + \alpha \cdot d_{fm}[c],$$

and decode $x_{\text{edit}}{=}\mathcal{D}(z_{\text{edit}})$.

Strength control and cost. α gives continuous, signed control: positive values amplify the concept (e.g., add a smile), negative values suppress it, and $\alpha{=}0$ returns the reconstruction $\mathcal{D}(E(x))$; DLE has no other per-image hyperparameter. An edit costs one encoder forward, one addition and one decoder forward: about 18 ms for decoding, ≈31 ms including the encoder, on one NVIDIA RTX A5000 at 256×256 (Appendix F).

3.4 Why a VFM Latent Supports an Additive Edit

DLE relies on two properties of the latent. *The encoder.* The frozen DINOv3 produces semantic patch tokens (Caron et al., 2021; Oquab et al., 2024), whereas VAE latents carry low-level detail: five-fold linear probes on 1,000 labeled CelebA images, with a 64-dimensional PCA fitted inside each training fold and no decoder or flow model involved, read the five attributes from the VFM latent at macro AUROC 0.968 against 0.873 from an SD-VAE latent. A concept is therefore more linearly readable in the VFM latent, one translation explains more of its paired-difference energy there ($R_c{=}0.351$ vs. 0.243; Appendix B.4), and d_{fm} concentrates on the concept's tokens (Figure 7). *The decoder.* The CNN maps each token through a bounded receptive field (Araujo et al., 2019) whose effective response is concentrated well inside it (Luo et al., 2016), so token-local changes stay pixel-local, as observed for convolutional GAN generators (Bau et al., 2019; Kim et al., 2021); our qualitative

cross-talk observations agree (Appendix B.3). A single translation at $t{=}0$ is then enough to move the concept, whereas the baselines act on the sampling trajectory at $t{>}0$, where a VAE latent needs them. Section 4.3 tests this account with a matched generator trained in a VAE latent.

4 EXPERIMENTS

4.1 SETUP

Architecture and baselines. Our pipeline is a frozen DINOv3 ViT-S+/16 distilled encoder shared across domains, a per-domain NextDiT (Qin et al., 2025) flow model trained on the domain dataset (e.g., CelebA) and a per-domain CNN decoder at 256×256 (Appendix A.1). We compare against SDEdit (Meng et al., 2022), SEGA (Brack et al., 2023), Concept Sliders (Gandikota et al., 2024) and FlowEdit (Kulikov et al., 2025), adapted to our generator (Appendix A.4). FlowEdit runs at its official configuration; on CelebA and MIMIC-CXR, SDEdit, SEGA and Concept Sliders are tuned on a development split and scored once on the held-out test split, on ADNI-1 all baselines use published defaults, and Concept Sliders is not run on MIMIC-CXR; DLE uses $\alpha{=}3$ in every domain, the top of its CelebA development grid and a reporting point fixed before the CXR test run (grids and development-selected values in Appendix A.5). Real-edit denotes edits of real source images and Gen-edit edits of FM-generated sources.

CelebA face attributes (Liu et al., 2015). 5 binary attributes (Gender, Smiling, Age, Eyeglasses, Blond Hair). FaRL (Zheng et al., 2022) text encoder. We report Acc@$\tau{=}0.5$, the fraction of edits that CLIP-L-336 zero-shot classifies as the target at threshold 0.5 (Appendix A.6), on 350 real-image pairs held out from baseline tuning, 70 per attribute (Appendix A.5). Qualitative results: Figure 3 at $\alpha{=}{+}3$; Figure 4 at $\alpha{=}\pm\,1$ for Smiling and Eyeglasses and $\alpha{=}\pm\,3$ for the rest.

MIMIC-CXR chest radiographs (Johnson et al., 2019). 6 binary findings (Atelectasis, Cardiomegaly, Edema, Lung Opacity, Pleural Effusion, Pneumothorax) chosen for clinical relevance. CXR-BERT (Boecking et al., 2022) text conditioning. Each of 440 real sources (development 132, test 308; patient-disjoint) is edited toward and away from the finding, and CheXzero (Tiu et al., 2022) exact-pair AUROC ranks the toward-edit against the away-edit. Qualitative results: $\alpha \in \{-1, 0, +1\}$ in the main text (Figure 5); the $\alpha{=}{\pm}3$ edits and signed-difference maps in Appendix D.2.

ADNI-1 brain MRI (Jack Jr. et al., 2008). 3-class diagnosis: Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN); BiomedCLIP (Zhang et al., 2025a) text conditioning (FM training: 511 subjects, $36{,}540$ axial T1 slices; Appendix A.1). With no public slice-level AD/MCI/CN classifier, we train ADNI-DN121 (DenseNet-121 (Huang et al., 2017), ImageNet-pretrained; Appendix A.6); the editing pool comes from FM-validation subjects, disjoint from FM training and unseen by the evaluator. AUROC on $n{=}100$ Real-edit pairs per direction ($n{=}52$ for MCI); positive-α progressions in Figure 6, the full α grid in Appendix C.3 (Table 20).

4.2 CELEBA FACE ATTRIBUTE EDITING

Quantitative comparison. SDEdit, SEGA and FlowEdit stay at 14.0–18.0% and return near-copies of the source; Concept Sliders reaches 48.3% while replacing the face. DLE reaches $\mathbf{82.0}\%$ at $\alpha{=}{+}3$, with Smiling (100%) and Eyeglasses (98.6%) at the evaluator ceiling, and its edit needs no flow-model evaluation, $203\times$ less flow-model and decoder time than FlowEdit's 66 evaluations per edit, $115\times$ when the shared encoder forward is counted (Appendix F). A prompt-free ResNet-50 classifier trained on the CelebA training split rates DLE at 94.6% (Appendix A.7). Age is the weakest attribute (47.1% under CLIP, 81% under the classifier), the weak-direction case of Limitation 1. Weak text conditioning does not explain the baselines' failure: FlowEdit at its official target CFG of 13.5 stays at 18.0%, and the prompt-driven endpoint shifts are large for every CelebA and CXR concept other than Age and Pneumothorax ($\|d_{fm}\|$ 98–229; Table 6).

Qualitative comparison. In Figure 3, SDEdit, SEGA and FlowEdit return the source almost unchanged, and Concept Sliders loses identity without reaching the target. DLE edits all five attributes while pose and background appear unchanged. Figure 4 shows bidirectional DLE edits; full α-progressions are in Appendix D.1.

Table 1: CelebA Acc@τ=0.5 (%) on the held-out test split (70 pairs per attribute) in our VFM generator and in a matched VAE generator (Section 4.3). SDEdit, SEGA and Concept Sliders are tuned on a development split; FlowEdit runs at its official configuration (Appendix A.5). Concept Sliders uses our adapters and is run in the VFM generator only. Bold: best per column within each generator.

Method	Backbone	Gender	Smiling	Age	Eyegl.	Blond	Avg
VFM generator (ours)							
SDEdit	VFM+RF	1.4	21.4	30.0	8.6	15.7	15.4
SEGA	VFM+RF	1.4	20.0	27.1	5.7	15.7	14.0
Concept Sliders	VFM+RF	51.4	45.7	**47.1**	50.0	47.1	48.3
FlowEdit	VFM+RF	1.4	25.7	31.4	12.9	18.6	18.0
DLE (α=+1, ours)	**VFM+RF**	4.3	72.9	38.6	71.4	35.7	44.6
DLE (α=+3, ours)	**VFM+RF**	**71.4**	**100.0**	**47.1**	**98.6**	**92.9**	**82.0**
Matched VAE generator							
SDEdit	VAE+RF	**100.0**	88.6	50.0	**95.7**	80.0	82.9
SEGA	VAE+RF	30.0	38.6	30.0	67.1	14.3	36.0
FlowEdit	VAE+RF	**100.0**	**94.3**	**55.7**	92.9	87.1	**86.0**
DLE (α=+1, ours)	VAE+RF	1.4	41.4	32.9	30.0	52.9	31.7
DLE (α=+3, ours)	VAE+RF	11.4	81.4	34.3	71.4	**95.7**	58.9

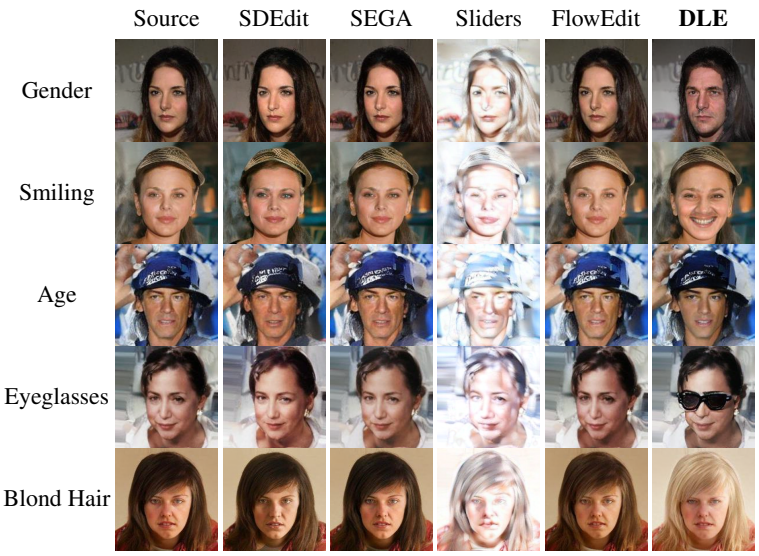


Figure 3: CelebA editing comparison on five attributes (FM-generated sources), one source per row. Baselines use the settings of Table 1 except Concept Sliders (published default); DLE at α=+3.

4.3 Matched VAE Generator

To test the account of Section 3.4, we trained a matched CelebA generator in a VAE latent (stock SD-VAE; same NextDiT-0.8B, patch size, captions and text encoder) and ran every editor through the same protocol (Table 1, lower block). Which editor works depends on the latent. In the VAE generator FlowEdit and SDEdit reach more targets than DLE, but mostly by producing a different face (FaceNet similarity to the reconstruction 0.34 and 0.19 vs. 0.62 for DLE; Appendix A.7); in the VFM generator DLE edits and the three trajectory editors return near-copies (FaceNet similarity to the reconstruction 0.76–0.97; Appendix A.7). A single translation also explains a larger fraction of the paired-difference energy in the VFM latent (R_c=0.351 vs. 0.243; Appendix B.4). Encoder, decoder and flow checkpoint all differ between the two generators, so the comparison isolates the latent space and what is trained on it, not the encoder alone.

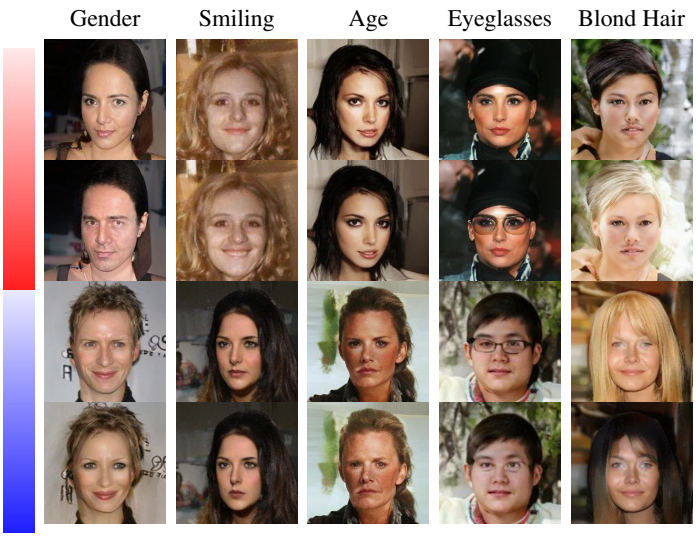


Figure 4: CelebA DLE bidirectional edits on 5 attributes. Rows 1–2: positive edits (row 1 source, row 2 at $\alpha{=}{+}3$, or $+1$ for Smiling/Eyeglasses). Rows 3–4: negative edits of a different source (row 3 source, row 4 at $\alpha{=}{-}3$, or -1 for Smiling/Eyeglasses). Red gradient marks α^{+}, blue gradient marks α^{-}. Each column: one $d_{fm}[c]$.

Table 2: Medical editing. **CXR**: exact-pair AUROC (CheXzero) on the held-out test split (308 images, 150 patients); SDEdit and SEGA tuned on a development split, FlowEdit at its official configuration, Concept Sliders not run. **Brain**: per-direction AUROC (ADNI-DN121), baselines at their published defaults. DLE at $\alpha{=}{+}3$ in both domains. Per-finding and per-transition values in Appendix C.2 and C.3.

	CXR (AUROC)			**Brain** (AUROC)		
Method	avg	best	worst	avg	best	worst
SDEdit (Meng et al., 2022)	0.529	0.565	0.501	0.646	0.806	0.482
SEGA (Brack et al., 2023)	0.505	0.515	0.500	0.552	0.806	0.366
Concept Sliders (Gandikota et al., 2024)	—	—	—	0.567	0.810	0.382
FlowEdit (Kulikov et al., 2025)	0.569	0.659	0.501	0.537	0.584	0.486
DLE (ours)	**0.927**	**1.000**	**0.579**	**0.878**	**0.920**	**0.836**

4.4 MIMIC-CXR Chest Radiograph Editing

Quantitative comparison. On the held-out test split (308 images, 150 patients) the baselines stay near chance (exact-pair AUROC 0.505–0.569; Table 2); DLE reaches $\mathbf{0.927}$ at $\alpha{=}{+}3$ and 0.848 at $\alpha{=}{+}1$; the development split selects $\alpha{=}8$, which reaches 0.945 at mean SSIM 0.50 to the reconstruction (Appendix A.5). Pneumothorax is DLE's weakest finding (0.579) and at chance for every baseline (0.501–0.502); it has the smallest direction ($\|d_{fm}\|{=}19.8$ vs. 104–229; Appendix E). Classifiers never trained on MIMIC-CXR also move with the edit: on the five findings all evaluators share, score-shift AUROC (the edit ranked against the source's autoencoder round-trip; Appendix C.2) is 0.758–0.764, against 0.855 for CheXzero (Appendix A.8).

Qualitative α-progressions. In Figure 5, five of six findings respond to α in the expected direction (basal opacity for Atelectasis and Pleural Effusion, a wider cardiac silhouette for Cardiomegaly, interstitial and patchy opacities for Edema and Lung Opacity); Pneumothorax is the weak-direction floor (Limitation 1; signed-difference maps in Appendix D.2).

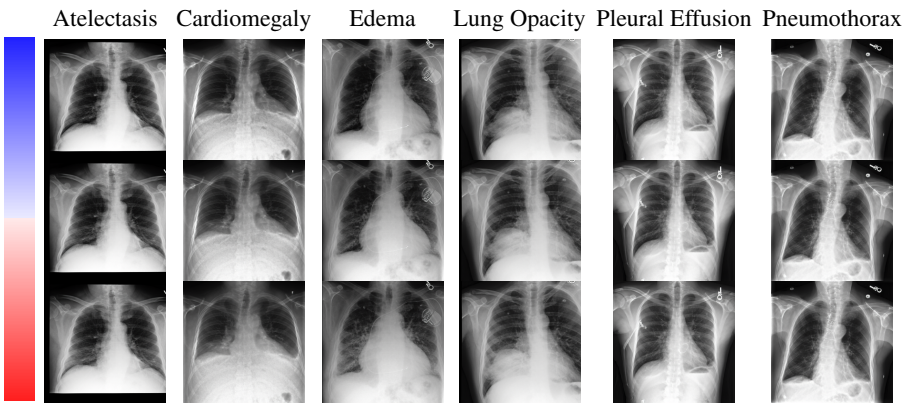


Figure 5: CXR DLE α-progression on six findings, $\alpha \in \{-1, 0, +1\}$ (blue→red). Each column: one $d_{fm}[c]$; the source (α=0) is FM-generated. Positive α amplifies, negative α suppresses; Pneumothorax barely changes (Limitation 1). $\alpha = \pm 3$ and signed-difference maps in Appendix D.2.

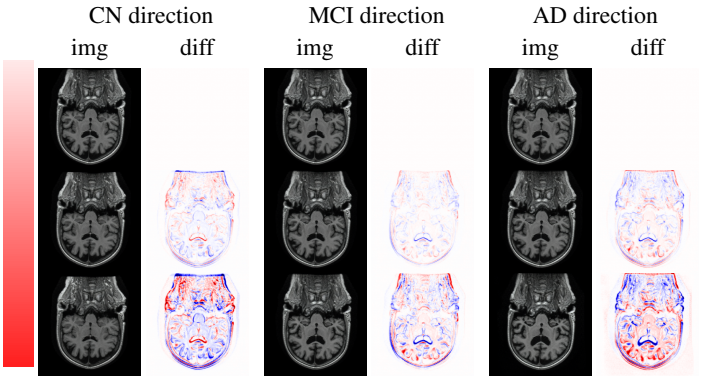


Figure 6: Brain DLE α^+-progression across 3 directions (CN, MCI, AD): image + signed-difference map (warm/cool: amplified/suppressed). Rows: $\alpha \in \{0, +1, +3\}$ (red gradient: light→dark). α=+3 pushes toward the corresponding state. Negative-α transitions in Table 20; per-direction values in Table 11.

4.5 ADNI-1 Brain MRI Editing

Quantitative comparison. Averaged over the CN, MCI and AD directions, DLE reaches AUROC **0.878** at α=+3 and the baselines 0.537–0.646 (Table 2); on the single reference pool of Table 11 the DLE average is 0.875 at α=+3 and 0.704 at α=+1. At α=+3 four of six transitions exceed 0.7 (CN→AD 0.886, MCI→AD 0.792, AD→CN 0.758, MCI→CN 0.743; Table 20), CN→MCI (0.665) and AD→MCI (0.557) are scored on the MCI axis, where the evaluator itself is near chance (Appendix A.6). Figure 6 shows the α^+-progression with signed-difference maps.

4.6 Ablations

The mean-difference direction beats random, PCA top-1 and SVM-normal directions at every positive α (Table 4; Appendix B.2). Accuracy rises with α for all five CelebA attributes (Table 17), and the Phase 1 pool size N=200 sits on the plateau in all three domains: N=50 changes the averaged metric by at most 0.021 (Appendix B.1). DLE is not tied to DINOv3. With other frozen DINO-family encoders, each with its own decoder and flow model, a DINOv2-B CelebA generator reaches 83.4% at α=+3 (DINOv3: 82.3% on the 500-pair pool), and RAD-DINO and DINOv2-S/B/L CXR generators reach score-shift AUROC 0.729–0.765 against 0.861 for DINOv3 (Table 3; Appendix B.6, Table 16). Sixteen further CelebA attributes average 61.0% at α=+3 on the same generator (Appendix B.5, Table 15).

Table 3: A second frozen encoder, DINOv2-B, with its own decoder and flow model in each domain. Top: CXR score-shift AUROC (CheXzero, 440 sources), per finding at $\alpha{=}{+}3$ and macro at $\alpha{=}{+}1/{+}3$. Bottom: CelebA Acc@$\tau{=}0.5$ (%) on the 500-pair pool, per attribute at $\alpha{=}{+}3$ and mean at $\alpha{=}{+}1/{+}3$. Three further CXR encoders in Appendix B.6.

CXR encoder	Atel.	Cardio.	Edema	Opacity	Effusion	PTX	macro $\alpha{=}1$	macro $\alpha{=}3$
DINOv3 ViT-S+/16 (ours)	0.958	0.982	0.921	0.890	0.864	0.550	0.765	**0.861**
DINOv2 ViT-B/14	0.665	0.869	0.948	0.829	0.604	0.511	0.584	0.738
CelebA encoder	Gender	Smiling	Age	Eyegl.	Blond		mean $\alpha{=}1$	mean $\alpha{=}3$
DINOv3 ViT-S+/16 (ours)	71.0	99.8	48.4	99.4	92.8		44.2	82.3
DINOv2 ViT-B/14	91.0	100.0	35.0	98.0	93.0		46.0	**83.4**

Table 4: Direction-extractor ablation on CelebA (Real-edit, 500-pair pool, CLIP-L Acc@$\tau{=}0.5$ in %, mean $\pm$ std over 5 bootstrap seeds). Mean-difference (DLE) is best at every positive α.

Direction	$\alpha{=}{-}3$	$\alpha{=}{-}2$	$\alpha{=}{-}1$	$\alpha{=}{+}1$	$\alpha{=}{+}2$	$\alpha{=}{+}3$
Random (control)	43.0±3.9	34.2±5.3	18.1±1.1	15.6±1.3	33.7±4.4	47.2±2.1
PCA top-1	8.5±0.6	8.9±1.1	10.9±0.5	28.0±0.8	41.9±3.0	57.3±3.7
SVM-normal	4.4±0.3	5.6±0.8	9.2±0.4	30.9±0.5	62.1±1.8	76.3±1.1
Mean-diff (DLE)	**3.9±0.3**	**6.0±0.4**	**8.4±0.1**	**44.1±0.5**	**70.6±1.5**	**82.3±0.3**

5 Discussion: Relation to Latent Arithmetic

Mean-difference latent arithmetic $z + \alpha\, n$ has been applied to GAN (Goodfellow et al., 2014) bottlenecks by InterFaceGAN (Shen et al., 2020), to classifier features by CAV/TCAV (Kim et al., 2018) and to diffusion activations by CASteer (Gaintseva et al., 2026); we do not claim it as new. $d_{fm}[c]$ is the least-squares translation between shared-noise endpoint pairs of the generator itself (Equation 1), not a classifier vector or a supervised boundary; with the same edit rule it reaches 82.3% on the 500-pair pool against 76.3% for an SVM normal and 57.3% for PCA (Appendix B.2). It is applied once at $t{=}0$, whereas CASteer and the baselines act during sampling at $t{>}0$. It is also a $D{\times}h{\times}w$ tensor concentrated on the concept's tokens (Figure 7), not a flat bottleneck vector. DLE's contribution is the paired-endpoint estimator and the endpoint operating point in a VFM latent.

6 Limitations

(1) **Weak-direction floor.** Directions with small $\|d_{fm}\|$ barely change the image (CelebA Age, CXR Pneumothorax with the plain prompt pair; Appendix E). (2) **Concept-dependent α.** Per-concept α chosen on a labeled development split matches or slightly exceeds fixed $\alpha{=}3$ on MIMIC-CXR (Appendix B.7) but needs labels. (3) **Architectural cross-talk.** Editing one concept can shift co-occurring concepts (e.g., Cardiomegaly $\rightarrow$ Edema); on CXR the off-target shift is 1.1–$3.1\times$ smaller than the on-target shift (Appendix B.3). (4) **Automatic evaluation.** CLIP, CheXzero, ADNI-DN121 and FaceNet are automatic proxies; the medical results make no clinical claim. (5) **Generality beyond our generators.** On the released SVG-T2I model (Shi et al., 2025), DLE reaches 25.7% against 64.0% for FlowEdit (Appendix B.8); transfer to generators we did not train is not established.

7 Conclusion

We introduced **Direct Latent Editing (DLE)**: a per-concept direction, estimated once from the generator's shared-noise prompt pairs and added to the encoded latent at $t{=}0$ without inversion, weight updates or per-edit FM sampling. On three VFM-based generators it reaches targets that the tested baselines do not, at $203\times$ lower per-edit flow-model and decoder time than FlowEdit at its official configuration.

AI USE STATEMENT

We used generative AI tools to edit and draft parts of the text, to check bibliographic entries against their sources, and to implement standard methods and evaluation and timing scripts. We did not use generative AI tools to generate research ideas or experimental results. All AI-assisted text, references and code were reviewed by the authors: reported numbers were traced to the stored outputs that produced them, and references were checked against publisher or arXiv records. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS AND REPRODUCIBILITY STATEMENTS

Data access, intended use and risks are discussed in Appendix **??**; implementation details, tuning grids and timing protocols in Appendices A, A.5 and F, and reproducibility in Appendix **??**.

REFERENCES

André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 2019. doi: 10.23915/distill.00021.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. GAN dissection: Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Black Forest Labs. FLUX.1: Open-source rectified-flow text-to-image model. `https://github.com/black-forest-labs/flux`, 2024.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria T. Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *European Conference on Computer Vision (ECCV)*, 2022.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In *Medical Imaging with Deep Learning (MIDL)*, 2022.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. CASteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations (ICLR)*, 2026.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

Clifford R. Jack Jr., Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008. doi: 10.1002/jmri.21049.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Asian Conference on Computer Vision (ACCV)*, 2022.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, 2018.

Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in GAN for real-time image editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Qi Qin, Le Zhuo, Yi Xin, et al. Lumina-Image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, Pengfei Wan, Kun Gai, Jie Zhou, and Jiwen Lu. SVG-T2I: Scaling up text-to-image latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2512.11749*, 2025.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, 2026.

Jaeyo Shin, Jiwook Kim, and Hyunjung Shim. Representation alignment for just image transformers is not easier than you think. *arXiv preprint arXiv:2603.14366*, 2026.

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. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Xingjian Leng, et al. What matters for representation alignment: Global information or spatial structure? In *International Conference on Learning Representations*, 2026. arXiv:2512.10794.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2025a. doi: 10.1056/AIoa2400640.

Shilong Zhang et al. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. *arXiv preprint arXiv:2512.17909*, 2025b.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive image generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025b.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 PER-DOMAIN GENERATORS

We train a separate generator on each of three domains. Per-domain components are the decoder, the flow model and the text encoder (Table 5); all other components are shared.

Table 5: Generator components. Five components (FM family, FM size, image encoder, latent dimension, input resolution) are shared across domains; the decoder is trained per-domain and the text encoder is selected per-domain.

Component	Value
FM family	NextDiT (Qin et al., 2025)
FM size	0.8B
Image encoder E	DINOv3 ViT-S+/16 distilled (`dinov3_vits16plus_pretrain_lvd1689m`)
Latent feature dim D	384
Decoder $\mathcal{D}$	CNN, trained per-domain on the target modality
Input resolution	256×256
Text encoder (CelebA)	FaRL (Zheng et al., 2022)
Text encoder (MIMIC-CXR)	CXR-BERT (Boecking et al., 2022)
Text encoder (Brain)	BiomedCLIP (Zhang et al., 2025a)

The decoder $\mathcal{D}$ uses the Taming-Transformer (Esser et al., 2021) decoder recipe with one modification: the input latent dimension is raised from the VAE default (4) to the DINOv3 patch-token width (D=384). Text encoder selection is described in §A.2.

Per-domain datasets.

- **CelebA** (Liu et al., 2015): original aligned set (178×218; *not* CelebA-HQ), 202,599 images, 40 binary attribute labels, official train/val/test split (162,770/19,867/19,962).
- **MIMIC-CXR-JPG v2.1.0** (Johnson et al., 2019): *PA (posteroanterior) view only* (the standard frontal chest X-ray, excluding lateral and AP views), CheXpert auto-labeler (Irvin et al., 2019). FM training pool: 237,966 entries from `p10`–`p19` (1,959 reserved for FM validation). The 440 Real-edit sources are 440 distinct PA images from these folders, each the source image of one real-image pair: 240 from incremental pairs (two patients whose labels differ in one finding) and 200 from longitudinal pairs (the same patient at two visits with a label change). Only the source image is edited and scored; the paired image fixes the target finding. Thirteen of the pair images fall in the FM-validation pool, so the flow model that yields d_{fm} and the decoder were trained on a pool containing most sources. The exact-pair metric compares the toward- and away-edits of the same source, so a memorised source enters both sides of each pair.
- **ADNI-1** (Jack Jr. et al., 2008): T1-weighted axial slices at 256×256, sampled at stride 2 in $z \in [110, 148]$ (20 slices per scan, multiple scans per subject across visits). The FM-training subset (§4.1; 511 subjects, 36,540 slices) has class balance MCI 48.1%, CN 30.6%, AD 21.3%, with 4,740 FM-validation slices from a subject-disjoint 64-subject set.

FM training. All three domain FMs share the same training framework: AdamW (learning rate 2×10^{-4}, weight decay 0, gradient clip 2.0) with bf16 precision, 1×1 patchify on the DINOv3 latent, lognormal SNR weighting, classifier-free-guidance training with 10% caption dropout, QK normalization, 2,000-step linear warmup. The reported checkpoint per domain is the one with the best FID on the domain's unseen validation pool among candidates saved during training, except for MIMIC-CXR, where we use the EMA weights at step 50,000 (global batch 288), fixed before evaluation. All quantitative MIMIC-CXR results in the main text and in Appendices A.5–A.8, B.5–B.7 and C use this checkpoint, except the N-scan (Section 4.6) and the cross-talk ratio (Limitation 3), which come from Appendices B.1 and B.3 on an earlier CXR checkpoint of the same design.

A.2 TEXT ENCODER SELECTION AND CONCEPT SEPARATION

The text encoders were chosen for domain fit (FaRL for faces, CXR-BERT for chest radiographs, BiomedCLIP for brain MRI); Table 6 lists, per concept, the cosine similarity $\cos(\bar{c}_+, \bar{c}_-)$ between the mean-pooled positive and negative prompt embeddings of the selected encoder (lower means better prompt separation) and the norm $\|d_{fm}\|$ of the resulting direction.

Table 6: Prompt-pair cosine similarity of the selected text encoder and d_{fm} direction norm per concept.

Domain	Attribute / Finding	$\cos(\bar{c}_+, \bar{c}_-)$	$\|d_{fm}\|$
Brain (BiomedCLIP)	AD	0.723	46.54
	MCI	0.703	31.58
	CN	0.714	46.65
CXR (CXR-BERT)	Atelectasis	0.778	176.48
	Cardiomegaly	0.720	229.46
	Edema	0.739	179.54
	Lung Opacity	0.872	227.39
	Pleural Effusion	0.877	103.64
	Pneumothorax	0.940	19.81
CelebA (FaRL)	Gender	0.984	155.41
	Smiling	0.943	115.97
	Age	0.947	73.43
	Eyeglasses	0.934	127.64
	Blond Hair	0.969	98.02

A.3 PHASE 1 FM SAMPLING

Phase 1 uses CFG scale ω=1.5, K=50 Euler steps and N=200 samples per prompt pool in every domain (Section 3).

Sign convention. By construction, the positive prompt pool encodes the *semantic presence* of the attribute and the negative pool its complement, so $d_{fm}[c] = \frac{1}{N}\sum_i (z_{0,\mathrm{pos},i} - z_{0,\mathrm{neg},i})$ points from absence to presence. Positive α amplifies the attribute, negative α suppresses it. The convention is fixed before evaluation. Near-chance accuracy under $+\alpha$ (e.g., CelebA Age at α=+3 reaches only 0.484 on the 500-pair pool, just below the 0.5 threshold; Table 17) reflects the weak-direction floor of d_{fm} (Limitation 1), not a sign issue.

Random seeds. Three independent seeds are used: an FM training seed (weight init, shuffle, dropout), a Phase 1 noise seed shared across each paired (positive, negative) ODE solve so the displacement isolates the prompt-driven shift, and a 5-seed bootstrap that resamples the cached Phase 1 pool with replacement to produce the per-cell std reported in Tables 4, 17 and 20 and the brain averages of Table 2. Seed values and exact RNG calls are released with the code.

A.4 BASELINE SETTINGS AND EXTENSION PROBES

Operand mechanisms. Each tested inference-time baseline modifies a flow operand at intermediate $t > 0$ via a different mechanism:

- **SDEdit** (Meng et al., 2022): replaces the latent at t=$t_{\mathrm{start}} > 0$ with a noise mixture, $z_t = (1 - t_{\mathrm{start}})z_{\mathrm{src}} + t_{\mathrm{start}}\,\varepsilon$.
- **SEGA** (Brack et al., 2023): adds a concept-direction velocity to the base, $f_\theta + \lambda\,[f_\theta(z, t, c_{\mathrm{tgt}}) - f_\theta(z, t, \varnothing)]$ (λ: concept scale, simplified form, omitting thresholding and momentum).
- **Concept Sliders** (Gandikota et al., 2024): adds a learned LoRA delta (Low-Rank Adaptation (Hu et al., 2022), two low-rank matrices $B_\theta A_\theta$ that perturb the velocity) to the FM, $f_\theta + (\alpha_L/r)\,B_\theta A_\theta z_t$ with α_L, r as scale and rank.

- **FlowEdit** (Kulikov et al., 2025): integrates the difference of source and target conditional flows, $z_0^{\mathrm{tgt}} = z_0^{\mathrm{src}} + \sum_\ell \Delta t \, [f_\theta(\cdot, c_{\mathrm{tgt}}) - f_\theta(\cdot, c_{\mathrm{src}})]$ (ℓ: ODE step index, $\Delta t = 1/T$: per-step size).

Default setting. In the main tables SDEdit, SEGA and Concept Sliders use settings selected on a development split (Appendix A.5, Table 7), FlowEdit its official configuration, and the brain baselines their published defaults. In the figures Concept Sliders uses $\alpha_{\mathrm{LoRA}}{=}{+}3$, its published default; Table 1 uses the development-selected setting of Table 7.

Concept Sliders training protocol. Concept Sliders is the only baseline requiring offline LoRA training per concept. All others are training-free. We train the LoRA from scratch on the VFM-NextDiT FM of CelebA and ADNI-1 (not run on MIMIC-CXR or in the VAE generator) using the published objective (Gandikota et al., 2024) on the same prompt pool as Phase 1 of DLE: rank 8, $\alpha_{\mathrm{LoRA-train}}{=}16$ (effective scale $\alpha/r{=}2$), AdamW at $\mathrm{lr}{=}1\times10^{-4}$ for 1,000 iterations, guidance scale $\eta{=}1.0$. Concept Sliders was developed on SD/SDXL; our from-scratch LoRA on the VFM+RF backbone may understate what it reaches on its original backbone.

A.5 Tuning protocol

CelebA: 500 directed real-image pairs in 250 reciprocal clusters, split by cluster into a development split (150 pairs, 30 per attribute) and a test split (350 pairs, 70 per attribute). MIMIC-CXR: 440 images from 213 patients, split by patient into development (132 images, 63 patients) and test (308 images, 150 patients). Noise seeds are shared across methods and settings. SDEdit and SEGA evaluate the same 12 settings on the development split, and Concept Sliders 8 settings on CelebA with its trained adapters fixed; each selects the best and is scored once on the test split. The SDEdit and SEGA grids are identical across domains. Selection maximizes development target accuracy (CelebA) or

Table 7: Settings (12 for SDEdit and SEGA, 8 for Concept Sliders; FlowEdit's official configuration is fixed; DLE grids as listed) and development-selected choices.

Method	Grid	Selected (CelebA VFM / CelebA VAE / MIMIC-CXR)
SDEdit	$t_{\mathrm{start}} \in \{0.2, 0.3, 0.5, 0.7\}\times$ CFG $\in \{1.5, 3, 5\}$	(0.7, 3) / (0.7, 5) / (0.7, 5)
SEGA	$\lambda \in \{3, 5, 8\} \times t_{\mathrm{start}} \in \{0.2, 0.3\}\times$ CFG $\in \{1.5, 3\}$	(5, 0.2, 3) / (8, 0.3, 1.5) / (8, 0.3, 1.5)
Concept Sliders	$\alpha_{\mathrm{LoRA}} \in \{1, 3\} \times t_{\mathrm{start}} \in \{0, 0.3\}\times$ CFG $\in \{1.5, 3\}$ (CelebA VFM only; rank-8 adapters fixed)	(3, 0.3, 1.5) / — / —
FlowEdit	official SD3 configuration (fixed); search within the published window: $n_{\max} \in \{10, 20, 33\}\times$ target CFG $\in \{1.5, 3, 5\}$; $T{=}50$, $n_{\mathrm{avg}}{=}1$, source CFG 1.5	official: $(33, 3.5/13.5)$ in all three; search: (33, 5) in all three
DLE	α (6 values in the CelebA VFM generator; 10 values up to 32 in the VAE generator; $\{1, 2, 3, 4, 6, 8\}$ for MIMIC-CXR)	$\alpha{=}3$ / $\alpha{=}32$ / $\alpha{=}8$ (see text)

exact-pair AUROC (CXR), with ties broken by mean SSIM to the reconstruction. FlowEdit's primary setting is its official SD3 configuration ($T{=}50$, $n_{\max}{=}33$, $n_{\mathrm{avg}}{=}1$, source/target CFG $3.5/13.5$; the FLUX configuration uses $n_{\max}{=}24$ of $T{=}28$), fixed before its evaluation: CelebA test 18.0% [14.1, 22.1] in the VFM and 86.0% [83.4, 88.5] in the VAE generator. We also searched, on the development split, the nine settings inside FlowEdit's published editing window, $n_{\max} \in \{10, 20, 33\}\times$ target CFG $\in \{1.5, 3, 5\}$ (SD3 uses $n_{\max}{=}33$ of $T{=}50$, FLUX 24 of 28). The search selects $(33, 5)$ (test: VFM 17.1%, VAE 65.1%). Concept Sliders' selected setting reaches 48.3% [45.1, 51.1] on the CelebA test split with mean SSIM 0.35 to the reconstruction (Table 1). On the CelebA development split, the VFM cells stay at 14.7–15.3%, whereas $n_{\max}{=}33$ already reaches 68.7% in the VAE generator; the corresponding MIMIC-CXR test value is in Table 19. DLE's selected $\alpha{=}3$ is the largest value of its VFM grid. On MIMIC-CXR, DLE's development AUROC rises with α up to the grid maximum

while mean SSIM to the reconstruction falls, so the rule selects α=8, which reaches 0.945 [0.939, 0.952] exact-pair AUROC at mean SSIM 0.50 on the test split, whereas the main text reports the unified α=3, a reporting point fixed before the test run (Table 19). In the VAE generator, DLE was first run at α=3 without selection, and a follow-up protocol fixed before its outcome selected α=32 from a grid up to 32, whose outputs ignore the source (Section 4.3); we report both.

A.6 EVALUATORS

Each domain uses a zero-shot or pre-trained evaluator independent of the FM's text encoder, to avoid evaluator–encoder leakage (the evaluator using the same text encoder as Phase 1, which would inflate scores).

CelebA: CLIP-L-336. Zero-shot CLIP-L (`openai/clip-vit-large-patch14-336`) on bidirectional cross-attribute pairs. Each attribute uses a single (positive, negative) prompt pair (Table 8). The same pair is reused for Phase 1 $d_{fm}[c]$ extraction so that the evaluator and direction extraction share semantic anchors by design. At inference, CLIP-L produces a softmax over $\{p_{\text{pos}}(x),\ p_{\text{neg}}(x)\}$. We report Acc@$\tau$=0.5, the fraction of edited pairs for which $\mathbf{1}[p_{\text{pos}}(x) > 0.5]$ matches the target-attribute direction. For positive ($+\alpha$) edits the metric scores $\mathbf{1}[p_{\text{pos}}(x) > 0.5]$; for negative ($-\alpha$) edits it scores $\mathbf{1}[p_{\text{pos}}(x) < 0.5]$. Acc@$\tau$=0.5 averages the two halves of the bidirectional pair pool. CLIP-L is a different model from the FaRL text encoder used in Phase 1, but its prompts overlap in wording with the extraction prompts; we therefore also report a prompt-free classifier (Appendix A.7) and held-out wording (Appendix B.5). The appendix per-attribute analysis (Table 17) uses the full 500-pair pool (100 per attribute); Table 1 reports its 350-pair test split.

Table 8: CelebA CLIP-L-336 zero-shot prompt template (same pair reused for Phase 1 $d_{fm}[c]$ extraction). CXR (CheXzero) ships with built-in prompts and Brain (ADNI-DN121) takes no text input, so prompts are listed for CelebA only.

Attribute	Positive prompt	Negative prompt
Gender	"a photo of a man"	"a photo of a woman"
Smiling	"a photo of a smiling person"	"a photo of a neutral face person"
Age	"a photo of a young person"	"a photo of an elderly person"
Eyeglasses	"a photo of a person wearing glasses"	"a photo of a person without glasses"
Blond Hair	"a photo of a person with blond hair"	"a photo of a person with dark hair"

CXR: CheXzero. CheXzero (Tiu et al., 2022) is a zero-shot chest-radiograph classifier pre-trained on MIMIC-CXR text–image pairs, independent of the CXR-BERT used in Phase 1. CheXzero outputs a per-finding probability. AUROC is computed on the (positive-target, negative-target) probability pair after editing each source.

Brain: ADNI-DN121. No public zero-shot AD/MCI/CN slice-level classifier exists, so we trained *ADNI-DN121*, a DenseNet-121 (Huang et al., 2017) (ImageNet-pretrained, 3-class CN/MCI/AD head; 7.04M parameters). The FM-training slice pool (511 subjects, §A.1) is split 80/10/10 at the *subject* level (random shuffle, seed 0): $\sim$408 training subjects, $\sim$51 validation, $\sim$52 test (the 64-subject FM-validation pool is reserved for the editing evaluation and is unseen by the evaluator). Training: AdamW, lr=1×10^{-4}, weight decay 10^{-4}, batch size 64, 6 epochs, image 224×224 with ImageNet normalization. Best epoch is selected by AD-vs-CN AUROC on validation. The chosen checkpoint (epoch 3) reaches val AD-vs-CN AUROC 0.723 (held-out test 0.710); its 3-class accuracy is 43.3% on validation and 42.5% on test, below the 48.1% majority-class (MCI) rate of the training pool, so the evaluator is informative mainly on the AD-vs-CN axis. ADNI-DN121 weights will be released alongside the code.

Pairwise binary AUROC on val. MCI-pair validation AUROCs of the evaluator are near chance (0.49–0.54 with the margin score $p(\text{class}_a) - p(\text{class}_b)$), so the two transitions scored on $p(\text{MCI})$ in Table 20, CN→MCI and AD→MCI, are scored by an evaluator near chance on that axis and should be read as exploratory; AD→MCI (0.557) is the clearest case.

Editing AUROC vs. evaluator AUROC. Brain editing AUROCs (e.g., CN→AD 0.886 at $\alpha{=}{+}3$, Table 20) are *within-subject probability shifts*: for each source image, $p_{\mathrm{orig}}(\mathrm{target})$ vs. $p_{\mathrm{edit}}(\mathrm{target})$ are ranked across the source pool labeled $\{0, 1\}$ as a paired test. The evaluator's val AD-vs-CN AUROC (0.723 with $p(\mathrm{AD})$ score; 0.802 with margin score) is a *between-subject* test ranking two independent images. The two AUROCs answer different questions: the editing AUROC ranks each image's own before and after scores, whereas the evaluator's AUROC ranks different subjects, so the two numbers are not comparable.

A.7 IDENTITY

We measure identity with FaceNet (Schroff et al., 2015) (InceptionResnetV1 trained on VGGFace2 (Cao et al., 2018)) embeddings of the edit and of the autoencoder reconstruction $D(E(x))$ of the source, on $160{\times}160$ resizes of the aligned crops without face detection. As in face-editing work that compares against the inverted image (Yao et al., 2021), the reference is the reconstruction, so no editor is charged for the autoencoder's loss; the reconstructions themselves keep the identity (mean cosine to the input 0.909 in the VFM and 0.984 in the VAE generator, all ≥ 0.5). We report the flip rate (FR, target accuracy), face similarity (FS, mean cosine) (Jeanneret et al., 2023), and FR×FVA, where FVA is the fraction of successful edits with cosine $\geq \tau$ (computed on successful edits only, as in (Jeanneret et al., 2022)); FR×FVA is thus the fraction of all edits that reach the target and keep the identity. On 6,654 images of 326 identities from the official CelebA test split (77,571 same-identity and 200,000 different-identity pairs, same preprocessing), $\tau{=}0.5$ gives a false-accept rate of 0.31% and a true-accept rate of 83.7% (equal error rate 4.4% at 0.33). We added this axis after observing that target accuracy rewarded outputs that ignore or replace the source. All edits were regenerated with the selected settings and seeds; the baseline and FlowEdit edits reproduce the test accuracies exactly, and the DLE edits differ from Table 1 on one pair per generator (82.3% vs. 82.0%; 59.1% vs. 58.9%). Paired differences DLE ($\alpha{=}3$) minus FlowEdit at its official configuration in FR×FVA at

Table 9: Editors in the VFM and the matched VAE generator (CelebA held-out test split, 350 pairs per cell; Section 4.3): flip rate (FR, %), face similarity (FS) and FR×FVA at $\tau{=}0.5$ (%). FlowEdit: official configuration; "search": development-selected setting. For DLE in the VAE generator, the α selected on accuracy is 32, whose outputs ignore the source.

Generator	Method	FR	FS	FR×FVA
VFM	SDEdit	15.4	0.762	15.1
VFM	SEGA	14.0	0.919	14.0
VFM	FlowEdit	18.0	0.971	18.0
VFM	FlowEdit (search)	17.1	0.995	17.1
VFM	DLE ($\alpha{=}3$)	82.3	0.640	65.4
VAE	SDEdit	82.9	0.186	3.7
VAE	SEGA	36.0	0.663	30.9
VAE	FlowEdit	86.0	0.339	15.4
VAE	FlowEdit (search)	65.1	0.665	49.4
VAE	DLE ($\alpha{=}3$)	59.1	0.617	36.3
VAE	DLE ($\alpha{=}32$)	80.6	0.037	0.0

$\tau{=}0.5$ (95% cluster bootstrap) are $+47.4$ [41.4, 53.4] in the VFM and $+20.9$ [15.7, 26.0] in the VAE generator. An independent ResNet-50 attribute classifier (trained on the CelebA training split, macro AUROC 0.948 over 40 attributes) scores DLE at $\alpha{=}3$ at 94.6% [92.3, 96.6] in the VFM generator and 60.3% in the VAE generator, at or above CLIP in both.

A.8 INDEPENDENT CHEST-RADIOGRAPH EVALUATORS

CheXzero was trained on MIMIC-CXR. We rescored DLE edits (440 sources, the five findings shared by all evaluators) with the released TorchXRayVision DenseNet-121 classifiers (Cohen et al., 2022) (weights `all`, `chex` and `nih`), each with its own preprocessing; the five findings shared by all evaluators are all except Lung Opacity (a different set from the '5 strong' set of Table 10, which excludes Pneumothorax). Score-shift AUROC at $\alpha{=}3$ (five-finding macro, patient-cluster bootstrap): CheXzero 0.855 [0.840, 0.870], pooled data including MIMIC 0.793 [0.766, 0.819], CheXpert only

0.758 [0.724, 0.789], NIH ChestX-ray14 only 0.764 [0.742, 0.788]; at α=1, 0.739, 0.679, 0.670 and 0.650. Pneumothorax stays near chance for all (0.53–0.57 at α=3).

B Ablations

B.1 Phase 1 sample size N

We validate the default Phase 1 sample size $N{=}200$ (§3.2) by sub-sampling the cached Phase 1 pool with replacement to $N \in \{50, 100\}$ (5 bootstrap seeds at $N{=}50$, 3 at $N{=}100$; $N{=}200$ is the single-seed reference) and re-measuring Real-edit accuracy/AUROC over the full $\alpha \in \{-3, \ldots, +3\}$ grid on each domain. Table 10 summarizes the scan at $\alpha{=}{+}3$ together with the largest sub-sample deviation $|\Delta(200{-}50)| = |(\text{reference at } N{=}200) - (\text{mean over the } N{=}50 \text{ subsamples})|$ observed anywhere on the grid. The plateau begins at $N \leq 50$ across all three domains; the default $N{=}200$ is retained as a conservative choice.

Table 10: Phase 1 sample-size scan. Columns $N{=}50/100/200$: Real-edit metric at $\alpha{=}{+}3$, averaged over the domain's concepts. $\max|\Delta|$: largest $|\Delta(200{-}50)|$ over the full α grid, for the concept-averaged rows and for any single (concept, α) cell. CelebA: CLIP-L-336 Acc@$\tau{=}0.5$, 5 attributes ($n{=}100$ pairs each). CXR: CheXzero score-shift AUROC, $n{=}440$ sources, over the five findings other than Pneumothorax ('5 strong') and over all six. Brain: ADNI-DN121 per-direction AUROC, CN/MCI/AD.

Domain (metric)	$N{=}50$	$N{=}100$	$N{=}200$	$\max\|\Delta\|$ (avg)	$\max\|\Delta\|$ (cell)
CelebA (Acc@$\tau{=}0.5$)	0.815	0.827	0.822	0.021	0.068
CXR (AUROC, 5 strong)	0.972	0.970	0.967	0.005	0.028
CXR (AUROC, all 6)	0.893	0.889	0.889	0.004	0.028
Brain (AUROC, 3 directions)	0.874	0.866	0.875	0.014	0.018

CelebA. The five-attribute mean stays within 0.021 of the $N{=}200$ reference at every α; the largest single-cell deviation is Blond Hair at $\alpha{=}{+}2$ (0.068).

CXR. This scan uses an earlier CXR flow model trained on the same data with the same encoder, decoder and text encoder, and one Phase 1 reference seed, so absolute AUROCs differ from Table 18; the finding-averaged deviation stays at or below 0.005, the largest single-cell deviation is Pleural Effusion (0.028), and Pneumothorax stays at the weak-direction floor (≈ 0.5) regardless of N or α (Appendix E).

Brain. Deviations stay at or below 0.018 across all 18 (direction, α) cells. Table 11 lists the per-direction values at $\alpha{=}{+}1$ and $\alpha{=}{+}3$. This per-direction AUROC scores each direction's edit against ADNI-DN121 $p(\text{target})$ on a fixed pool ($n_{\text{eval}}{=}100$ for CN/AD, 52 for MCI) without splitting into source-conditioned transitions (Table 20); absolute values differ slightly from Table 2, which averages 5 bootstrap resamples of the Phase 1 pool (Appendix A.3), whereas this scan uses the single reference pool.

Table 11: Brain per-direction N-scan (Real-edit, ADNI-DN121 AUROC; mean $\pm$ std over subsample seeds at $N{=}50, 100$; single reference seed at $N{=}200$).

	$\alpha{=}{+}1$			$\alpha{=}{+}3$		
Direction	$N{=}50$	$N{=}100$	$N{=}200$	$N{=}50$	$N{=}100$	$N{=}200$
CN	0.737 ± 0.026	0.729 ± 0.019	0.720	0.881 ± 0.034	0.869 ± 0.038	0.869
MCI	0.647 ± 0.043	0.629 ± 0.041	0.631	0.819 ± 0.062	0.806 ± 0.076	0.832
AD	0.754 ± 0.023	0.752 ± 0.021	0.761	0.922 ± 0.014	0.923 ± 0.017	0.925
avg	**0.713**	**0.703**	**0.704**	**0.874**	**0.866**	**0.875**

B.2 Direction extractor choice

We compare four ways of constructing the latent direction d on CelebA (5 attrs, 5-seed bootstrap, CLIP-L-336 evaluator). All four extractors share the same encode-add-decode protocol and the same Phase 1 sample pool ($N{=}200$ paired positive/negative latents, §A.3). They differ only in how $d \in \mathbb{R}^{D \times h \times w}$ is constructed.

Variants.

- **Random (control)**, $d \sim \mathcal{N}(0, I)$ rescaled to match the median $\|d_{fm}\|$, gives a chance baseline at matched magnitude.
- **PCA top-1**, the first principal component of the $2N$ positive+negative latents (variance-maximizing direction, agnostic to label).
- **SVM-normal**, the weight vector (hyperplane normal) of a linear SVM fit on the $2N$ latents with positive/negative labels (label-aware separator).
- **Mean-difference (DLE, ours)**, $d = \frac{1}{N}\sum_{i=1}^{N}(z_{0,\mathrm{pos},i} - z_{0,\mathrm{neg},i})$ (Algorithm 1, Phase 1), a label-aware direction that exploits the paired-noise displacement.

Mean-difference (DLE) has the highest accuracy at every positive α (Table 4): at $\alpha{=}{+}3$ Real-edit it reaches 82.3 vs. SVM-normal 76.3, PCA top-1 57.3 and Random 47.2, and the gap to SVM-normal/PCA is $+13.2/{+}16.1$ pp at $\alpha{=}{+}1$. It also matches the lowest seed-std (tied with SVM-normal at $\alpha{=}{+}1$, strictly lowest at $\alpha{=}{+}2, {+}3$); the DLE–PCA gap at $\alpha{=}{+}3$ exceeds PCA's own std ($\sim$3.7 pp) by $\sim 7\times$. At $\alpha{=}{-}3$ (wrong-direction control), DLE drops to 3.9% while Random stays near 43%, confirming the extracted direction is sign-correct.

Direction-attention visualization. Figure 7 overlays the magnitude of each extractor's direction $|d|$ onto a representative source. Mean-diff (DLE) concentrates on the concept's spatial extent (eye region for Eyeglasses, mouth for Smiling, hair for Blond Hair), while Random spreads uniformly and PCA/SVM mix concept-relevant and concept-irrelevant tokens.

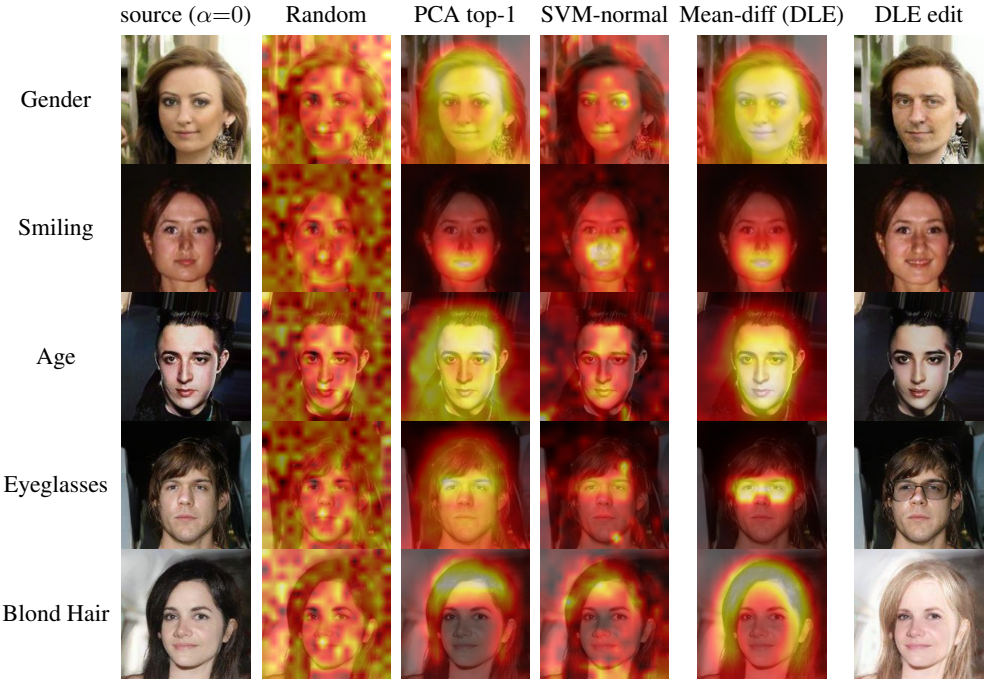


Figure 7: Extractor direction-attention overlays $|d|$ on a representative CelebA source per attribute (5 attrs $\times$ 4 extractors). Source (α=0) and the resulting DLE edit at the per-attribute saturation α are shown in the leftmost and rightmost columns (Smiling/Eyeglasses at $\alpha{=}{+}1$; Gender/Age/Blond Hair at $\alpha{=}{+}3$). Mean-diff (DLE) concentrates on the concept-relevant region; Random is diffuse, PCA/SVM are mixed.

B.3 Direction cross-talk

The CXR measurements in this subsection use the CXR flow model of Appendix B.1. Editing for one concept can shift co-occurring concepts (e.g., Cardiomegaly with Edema in CXR) and intensities outside the target region. We quantify two axes:

- **Semantic axis.** Editing for one finding shifts co-occurring findings on the same source. We measure this by applying $d_{fm}[B]$ ($B \neq A$) to sources carrying finding A, then reporting $\Delta p(A)$= shift in the source finding's probability after an off-target edit (positive = A amplified) and $\Delta p(B)$ (shift in target-finding probability: how much B is induced), averaged over the 5 off-target B-directions per source row, n=140 source images per row, α=+1.
- **Spatial axis.** The edit shifts intensity statistics in regions outside the target's spatial extent (e.g., mediastinum, scapula on CXR; cortical luminance on Brain). Qualitative observation across all six CXR findings and all three Brain directions.

Table 12: CXR semantic cross-talk at α=+1. Protocol: source x carries finding A; we apply $d_{fm}[B]$ for $B \neq A$. $\Delta p(A)$= shift in source-finding probability (positive = A amplified); $\Delta p(B)$= shift in target-finding probability (how much B is induced); averaged over 5 off-target B-directions per row, n=140 sources per finding A. SSIM compares the source x_A to its edit $x_{A\to B}$.

Finding A	$\Delta p(A)$	$\Delta p(B)$	SSIM$(x_A, x_{A\to B})$
Atelectasis	+0.101	+0.112	0.777
Cardiomegaly	+0.178	+0.412	0.688
Edema	+0.182	+0.530	0.707
Lung Opacity	+0.204	+0.329	0.717
Pleural Effusion	+0.064	+0.200	0.800
Pneumothorax (weak-direction floor)	−0.050	−0.012	0.830

Applying an off-target direction (Table 12) raises the source finding's probability by +0.06 to +0.20 on the five strong findings, the on-target shift $\Delta p(B)$ is 1.1–3.1× larger, so the intended finding moves more than the co-occurring one but the co-occurring one is not left unchanged, and mean SSIM to the source stays at or above 0.688. By inspection, the off-target change is a luminance and contrast shift rather than a deformation of the rib cage, ventricles, gyri or face shape at $|\alpha| \leq 3$; we did not measure structural preservation.

B.4 Translation consistency

Table 13: Translation consistency R_c (Section 4.3; N=200 paired endpoints) and DLE accuracy at α=3 under the independent classifier (test split, 70 pairs per attribute; per-attribute values rounded to the nearest percent, mean of unrounded values 94.6%).

	R_c		Accuracy (%)	
Attribute	VFM	VAE	VFM	VAE
Gender	0.362	0.073	97.0	48.6
Smiling	0.439	0.116	100.0	68.6
Age	0.255	0.074	81.0	15.7
Eyeglasses	0.415	0.325	100.0	84.3
Blond Hair	0.286	0.627	94.0	84.3
Mean	0.351	0.243	94.4	60.3

R_c is the fraction of the paired-difference energy explained by the single translation $d_{fm}[c]$, $R_c = 1-\sum_i \|\Delta_i - d_{fm}[c]\|^2 / \sum_i \|\Delta_i\|^2$ over the N=200 pairs (Table 13); its mean over the five attributes is 0.351 in the VFM generator and 0.243 in the VAE generator. The prediction that R_c is higher in the VFM generator was stated before measurement and holds for four attributes. The linear probe of Section 3.4 uses 1,000 labeled CelebA images: five-fold linear probes on each frozen latent, with a 64-dimensional PCA fitted inside each training fold and no decoder or flow model involved, give macro AUROC 0.968 for the VFM latent and 0.873 for the SD-VAE latent.

B.5 Prompt wording and additional CelebA attributes

Sixteen further CelebA attributes, with direction prompts written in the grammar of the generator's training captions and fixed before testing, average 61.0% (Table 15) at α=3 under CLIP-L with held-out wording on the test split (70 pairs per attribute) and 70.2% under the independent ResNet-50

classifier, and caption-grammar prompts score higher than generic prompts that do not follow the caption grammar, so direction prompts should match how the generator was captioned. We also rebuilt d_{fm} under several wording families (CXR values in Table 14; CelebA values in the text), changing the positive and negative prompts together, and rescored on the same evaluation sets at $\alpha{=}3$. On CelebA (Acc@$\tau{=}0.5$ on the test split), the four families stay within 6 points of each other in mean accuracy, 75.1–81.1%. On CXR five findings vary by at most 0.08 across families. Pneumothorax is

Table 14: Prompt-wording families on CXR: CheXzero score-shift AUROC on 440 sources; family directions use $N{=}50$ samples, so the plain row differs slightly from Table 18 ($N{=}200$). Ensemble: mean of the family directions.

CXR family	Cardio.	Edema	Effusion	Atel.	Opacity	PTX
plain (mean 0.857)	0.978	0.920	0.857	0.940	0.896	0.548
descriptive (0.910)	0.985	0.923	0.896	0.882	0.930	0.846
clinical (0.931)	0.982	0.914	0.900	0.959	0.896	0.936
ensemble (0.913)	0.982	0.926	0.910	0.905	0.921	0.835

the most wording-sensitive finding: near chance with the plain pair (0.548), it reaches 0.936 with clinical wording. The main results keep the plain pair, since choosing the wording on these numbers would select on the evaluation sources.

Table 15: Sixteen further CelebA attributes: CLIP-L Acc@$\tau{=}0.5$ (%) at $\alpha{=}{+}3$ on the test split (70 pairs per attribute), held-out wording.

Attribute	Acc	Attribute	Acc
Goatee	100.0	Pointy Nose	54.3
Mustache	87.1	Brown Hair	51.4
Black Hair	82.9	Bushy Eyebrows	50.0
Heavy Makeup	78.6	Arched Eyebrows	48.6
Narrow Eyes	75.7	Rosy Cheeks	44.3
No Beard	70.0	Bags Under Eyes	42.9
Sideburns	64.3	Gray Hair	41.4
High Cheekbones	55.7	Bald	28.6
Macro (16)			61.0

B.6 Other VFM encoders

DLE is not tied to DINOv3. Replacing it with other frozen DINO-family VFMs, each with its own decoder and flow model trained on the same data, gives the CheXzero score-shift AUROC of Table 16 on the same 440 CXR sources; the CelebA comparison with DINOv2-B is in Table 3 (main text). The five encoders differ in size and training budget. With held-out CLIP wording the DINOv2-B

Table 16: CXR score-shift AUROC (CheXzero, 440 sources) by encoder, per finding at $\alpha{=}{+}3$ and macro at $\alpha{=}{+}1/{+}3$. Each encoder has its own decoder and flow model.

Encoder	Atel.	Cardio.	Edema	Opacity	Effusion	PTX	macro $\alpha{=}1$	macro $\alpha{=}3$
DINOv3 ViT-S+/16 (ours)	0.958	0.982	0.921	0.890	0.864	0.550	0.765	**0.861**
RAD-DINO ViT-B/14	0.639	0.841	0.925	0.873	0.619	0.478	0.597	0.729
DINOv2 ViT-S/14	0.824	0.830	0.929	0.877	0.634	0.498	0.636	0.765
DINOv2 ViT-B/14	0.665	0.869	0.948	0.829	0.604	0.511	0.584	0.738
DINOv2 ViT-L/14	0.712	0.824	0.936	0.875	0.625	0.511	0.598	0.747

generator reaches 73.8% at $\alpha{=}3$.

B.7 Choosing α

A unified α does not match every concept's optimum: on CelebA, Smiling and Eyeglasses exceed 70% already at $\alpha{=}{+}1$ while Gender needs $\alpha{=}{+}3$ (Table 17). On MIMIC-CXR we choose, per

finding, from a separate grid $\alpha \in \{0.5, 1, \ldots, 3\}$ (not the grid of Appendix A.5) the smallest α whose development score-shift AUROC is within one bootstrap standard error of the best (patient-disjoint split, 132/308 images). The rule (2,000 bootstrap replicates for the standard error) selects α=2.5 for Atelectasis and Cardiomegaly, 1.5 for Edema and Lung Opacity, and 3 for Pleural Effusion and Pneumothorax, and reaches a test macro score-shift AUROC of 0.872 [0.850, 0.892] against 0.863 [0.835, 0.888] for a fixed α=3 (6-finding macro on the 308 test images; Table 18 reports all 440 sources; exact-pair values for α=3 are in Table 19).

B.8 A RELEASED VFM-BASED GENERATOR

We repeated the CelebA comparison on the released SVG-T2I model (Shi et al., 2025) (stage 1, 256 px, 2.6B parameters, frozen DINOv3 encoder, trained by its authors on 76M text–image pairs), with the same pairs, prompts, evaluator and tuning, under a protocol fixed before the outcome that counted a gap of at most 5 points to FlowEdit as transfer. SDEdit and SEGA were not run. DLE selected α=8 (the largest value of the grid $\{1, 2, 3, 4, 6, 8\}$) and reached 25.7% on the test split, FlowEdit 64.0%; only 9.1% and 6.9% of their edits also kept the identity (FR$\times$FVA), so neither method edits reliably in this model on our prompts. The Eyeglasses and Blond Hair directions are almost zero ($\|d_{fm}\|$=6.3 and 3.7 against 111–162 for the others). We report this as an exploratory negative result.

C PER-DOMAIN QUANTITATIVE RESULTS

This appendix expands main Tables 1–2 into per-attribute / per-finding / per-direction breakdowns for each domain. Evaluators are defined in §A.6; pair pools in §A.1 and Appendix A.5.

C.1 CELEBA: PER-ATTRIBUTE REAL-EDIT ACCURACY

Table 17: CelebA per-attribute Real-edit Acc@τ=0.5 across the full α grid on all 500 pairs (main Table 1 uses the 350-pair test split) (n=100/attr, 5-seed bootstrap, mean $\pm$ std). Rows sorted by α=+3 accuracy, descending. $\|d_{fm}\|$ is the extracted direction norm per attribute.

Attribute	α=−3	α=−2	α=−1	α=+1	α=+2	α=+3	$\|d_{fm}\|$
Smiling	0.002±0.004	0.000±0.000	0.020±0.000	0.730±0.023	0.990±0.012	**0.998±0.004**	115.97
Eyeglasses	0.024±0.011	0.026±0.005	0.030±0.000	0.730±0.007	0.960±0.012	**0.994±0.005**	127.64
Blond Hair	0.070±0.007	0.080±0.007	0.140±0.000	0.354±0.018	0.778±0.035	**0.928±0.008**	98.02
Gender	0.000±0.000	0.000±0.000	0.010±0.000	0.026±0.005	0.376±0.036	**0.710±0.014**	155.41
Age	0.100±0.007	0.192±0.018	0.222±0.004	0.368±0.004	0.426±0.011	0.484±0.021	73.43
avg	**0.039**	**0.060**	**0.084**	**0.442**	**0.706**	**0.823**	

C.2 CXR: PER-FINDING REAL-EDIT AUROC

Table 18: CXR per-finding score-shift AUROC (CheXzero; autoencoder round-trip of the source, label 0, against the decoded DLE edit, label 1) across the α grid on all 440 Real-edit sources (213 patients). Macro at α=+3: 0.861 [0.840, 0.881] (patient-cluster bootstrap). Findings rise from α=−3 to α=+2; Edema and Lung Opacity dip slightly at α=+3 as the edit spreads over the lung fields, and Pneumothorax stays near chance (weak-direction floor, Appendix E).

Finding	α=−3	α=−2	α=−1	α=+1	α=+2	α=+3
Atelectasis	0.137	0.208	0.335	0.725	0.909	0.958
Cardiomegaly	0.084	0.107	0.223	0.881	0.965	0.982
Edema	0.115	0.153	0.252	0.927	0.956	0.921
Lung Opacity	0.131	0.192	0.301	0.892	0.922	0.890
Pleural Effusion	0.245	0.316	0.397	0.649	0.806	0.864
Pneumothorax	0.473	0.481	0.489	0.514	0.530	0.550
avg (5 strong)	0.142	0.195	0.302	0.815	0.911	**0.923**
avg (6 all)	0.197	0.243	0.333	0.765	0.848	**0.861**

Table 19: CXR per-finding exact-pair AUROC on the held-out test split (308 images, 150 patients; Table 2). FlowEdit: official SD3 configuration; search: development-selected setting $(33, 5)$ (Appendix A.5). SDEdit and SEGA: development-selected settings. SSIM: mean between each edit (at the listed α for DLE) and the source reconstruction $D(E(x))$, over the six findings and both edit directions.

Method	Atel.	Cardio.	Edema	Opacity	Effusion	PTX	Macro	SSIM
SDEdit	0.519	0.541	0.565	0.536	0.511	0.501	0.529	0.79
SEGA	0.502	0.508	0.515	0.502	0.500	0.502	0.505	0.92
FlowEdit	0.546	0.596	0.659	0.587	0.525	0.501	0.569	0.99
FlowEdit (search)	0.517	0.536	0.560	0.529	0.511	0.500	0.526	1.00
DLE (α=1)	0.856	0.990	0.989	0.979	0.751	0.525	0.848	0.86
DLE (α=3)	1.000	1.000	1.000	1.000	0.981	0.579	0.927	0.64
DLE (α=8)	1.000	1.000	1.000	1.000	1.000	0.671	0.945	0.50

C.3 BRAIN: BIDIRECTIONAL TRANSITION MATRIX

Table 20: Brain MRI transition matrix across the full α grid, Real-edit (5-seed bootstrap, mean $\pm$ std; n per cell: CN 60, AD 64, MCI 132 slices from the 64-subject FM-validation pool; the per-direction pool of Table 2 uses 100 pairs for CN/AD and 52 for MCI). AUROC scored by ADNI-DN121 $p(\text{target} \mid x)$. CN→MCI (0.665) and AD→MCI (0.557) are scored on the MCI axis, where ADNI-DN121 is near chance (Appendix A.6).

Transition	α=−3	α=−2	α=−1	α=+1	α=+2	α=+3
CN → MCI	0.288±0.017	0.351±0.016	0.420±0.010	0.574±0.006	0.635±0.017	0.665±0.016
CN → AD	0.257±0.032	0.293±0.033	0.361±0.024	0.667±0.021	0.790±0.027	**0.886±0.022**
MCI → CN	0.291±0.024	0.286±0.017	0.376±0.007	0.619±0.011	0.700±0.020	**0.743±0.020**
MCI → AD	0.252±0.021	0.293±0.020	0.367±0.016	0.639±0.016	0.739±0.022	**0.792±0.024**
AD → CN	0.404±0.020	0.364±0.014	0.419±0.009	0.600±0.013	0.686±0.021	**0.758±0.024**
AD → MCI	0.399±0.004	0.435±0.006	0.462±0.003	0.532±0.008	0.553±0.018	0.557±0.029

D PER-DOMAIN QUALITATIVE RESULTS

D.1 CELEBA: α-PROGRESSION

All CelebA qualitative figures in this paper use FM-generated sources to avoid displaying real-identity images. Figure 8 shows the full bidirectional $\alpha \in \{-3, \ldots, +3\}$ grid. Signed-difference maps are omitted for CelebA: face edits manifest as distributed attribute-aligned shifts, and skin-texture variability dominates the diff signal.

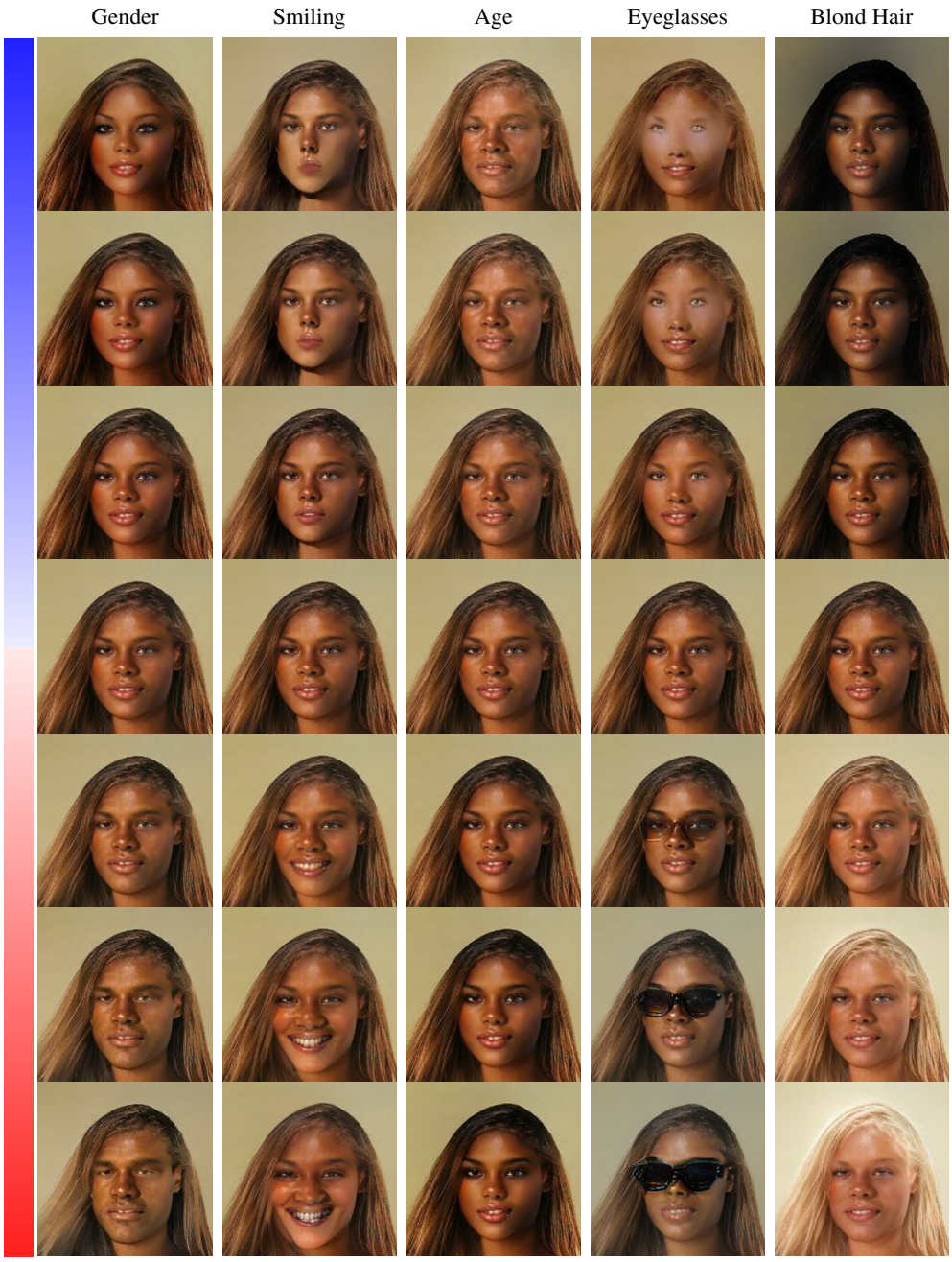


Figure 8: CelebA DLE full bidirectional α-progression on all 5 attributes (Gen-edit, CFG 1.5). Rows top-to-bottom: $\alpha \in \{-3, -2, -1, 0, +1, +2, +3\}$ (left bar: blue top $\rightarrow$ red bottom encodes negative $\rightarrow$ positive α). Each column: one $d_{fm}[c]$.

D.2 CXR: $\alpha{=}\pm3$ EDITS AND SIGNED-DIFFERENCE MAPS

Figure 9 shows DLE on all 6 CXR findings at $\alpha{=}\pm3$ (the $\alpha \in \{-1, 0, +1\}$ rows on the same sources are in Figure 5) and the signed-difference maps ($x_{\text{edit}} - x_{\text{orig}}$, clipped) at $\alpha \in \{-3, -1, +1, +3\}$ on the same source pairs.

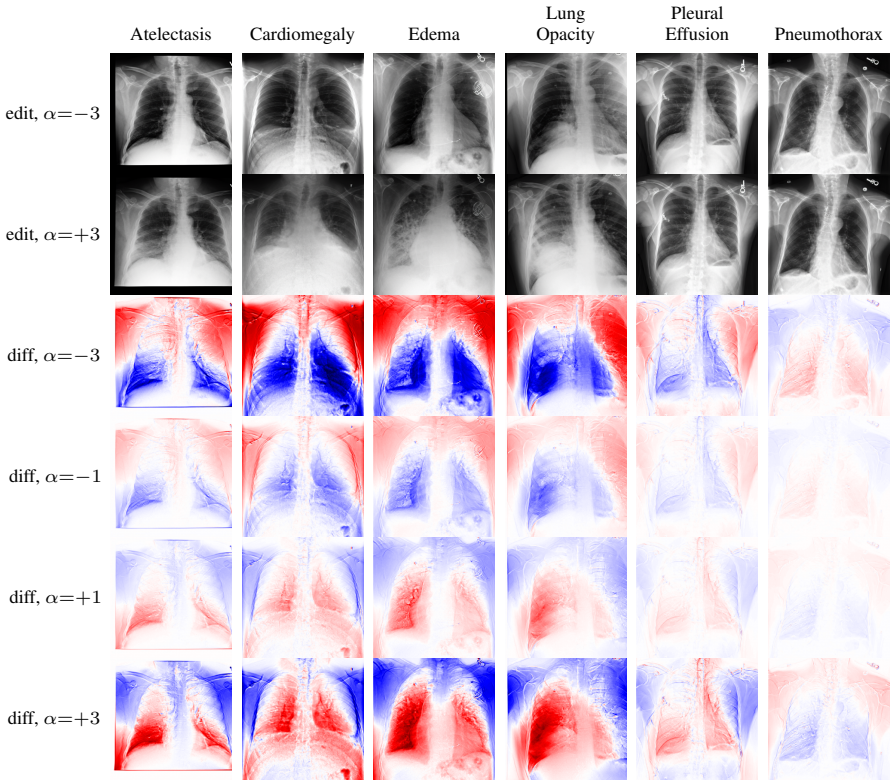


Figure 9: CXR DLE on 6 findings, same generated sources as Figure 5. Top two rows: edits at $\alpha{=}-3$ and $\alpha{=}+3$, where the edit spreads over the lung fields. Bottom four rows: signed-difference maps ($x_{\text{edit}} - x_{\text{orig}}$, per-image clipped; warm = amplified, cool = suppressed) at $\alpha \in \{-3, -1, +1, +3\}$. Pneumothorax shows near-uniform low-amplitude residuals at every α (weak-direction floor, Appendix E).

E FAILURE CASES: DIRECTION-EXTRACTION FLOOR

Directions much smaller than the domain's median $\|d_{fm}\|$ barely move the image (Limitation 1). Two concepts exhibit this regime in our experiments: CelebA *Age* and CXR *Pneumothorax*; the brain AD→MCI transition (0.557) shares a small $\|d_{fm}\|$ but is scored on an axis where the evaluator is near chance (Appendix A.6). Per-concept $\|d_{fm}\|$ values are listed in Appendix A.2, Table 6. At $\alpha{=}{+}3$, both remain near chance: Age 0.484 Acc (Table 17), Pneumothorax 0.550 AUROC (0.47–0.55 across α; Table 18). The signed-difference map for Pneumothorax (Figure 9, Pneumothorax column) shows near-uniform low-amplitude residuals; for Pneumothorax the floor depends on the prompt wording (Appendix B.5).

Failure modes at large $|\alpha|$. Figure 10 shows the two CelebA failure modes that appear at $|\alpha|{=}3$ on some sources: collapse under aggressive suppression and overshoot past the target, the reason a per-concept α can beat the fixed value (Limitation 2).

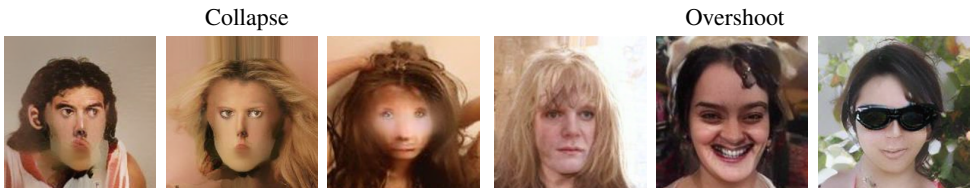


Figure 10: Two CelebA failure modes at $|\alpha|{=}3$. **Collapse** ($\alpha{=}{-}3$, left): aggressive suppression degrades the image (Smiling: face distortion; Eyeglasses on a no-glasses source: eye-region distortion). **Overshoot** ($\alpha{=}{+}3$, right): the edit goes past the target (Blond Hair: background luminance shift; Smiling: exaggerated expression; Eyeglasses: turns into sunglasses). (Limitation 2)

F COMPUTATION COST

Hardware. Per-domain FM training on 1×H200 141GB. Phase 1 direction extraction and Phase 2 editing on 1×A5000 24GB.

Measurement. All times in Table 21 were measured in one session on one RTX A5000 (BF16, batch 1, PyTorch 2.8, standard attention) with the CelebA VFM generator of Table 1. Each FM call is timed with CUDA events; the encoder (13.5 ms per source) is excluded because every method needs it, and decoding is included. After warm-up, each method runs five times.

Table 21: Compute cost on one A5000 (same session). Edit time is per image; extraction is once per concept. FlowEdit is timed at the official SD3 configuration used in Table 1 (T=50, $n_{\max}$=33, n_{avg}=1).

	FM calls	Time	Relative
DLE, per edit (ours)	**0**	$\mathbf{17.6 \pm 0.5}$ ms	**1**×
FlowEdit, official configuration, per edit	66	$3{,}578 \pm 42$ ms	203×
DLE direction extraction, per concept	20,000	1,004 s (16.7 min)	—

Break-even. A DLE edit saves 3.6 s over FlowEdit at its official configuration, so the one-time extraction pays for itself after $1{,}004/(3.578 - 0.018) \approx 282$ edits of the same concept; 100 concepts cost 27.9 GPU-hours. Extraction here generates samples one at a time; batching would shorten it. Counting the shared encoder forward (13.5 ms) in both methods gives $3{,}591.5/31.1 \approx 115\times$.