SAGE: SALIENT FACTOR DISCOVERY AND GENERATION WITH VISUAL FOUNDATION REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Given a target dataset, such as faces with eyeglasses, and a background dataset, such as faces without, contrastive analysis separates *salient* factors specific to the target from *common* content shared by both. We aim for salient representations that capture target-specific detail in each image, such as the shape, color, and position of the glasses, so that they reveal subtypes without subtype labels and guide the generation of new examples of a discovered subtype, even one with no name or text description. We introduce SAGE, which learns both factors directly in the high-dimensional spatial latent of a frozen representation autoencoder and conditions a diffusion transformer on the learned salient representation of a reference image. On Digits-ImageNet and FFHQ eyeglasses, SAGE combines high-fidelity *reconstruction* (rFID below 2) with unsupervised *subtype discovery*, recovering the digits better than baselines (probe accuracy 0.950 vs. at most 0.281) and revealing eyewear types, finer sunglasses styles, and mislabeled images; salient-conditioned *generation* raises Digits-ImageNet subtype accuracy over the unfactorized latent (90.5% vs. 27.7%) and diversity on both datasets. On retinal OCT, SAGE's salient space separates three diseases using only normal/disease labels.

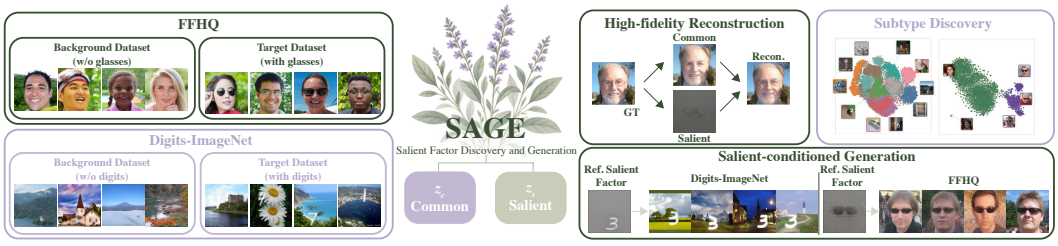


Figure 1: **Overview of SAGE.** Given background (BG) and target (TG) datasets, SAGE uses only dataset labels to decompose frozen self-supervised visual representations into *common* and *salient* factors, capturing shared content and target-specific variation, respectively. The decomposition supports high-fidelity reconstruction, subtype discovery in the salient space, and salient-conditioned generation that preserves target attributes while varying common content.

1 INTRODUCTION

Learning what two data distributions share and what sets one apart is central to multimodal learning (Dufumier et al., 2025), domain adaptation (Lee et al., 2021), and disentanglement (Sanchez et al., 2020). Contrastive analysis (CA) poses this problem with dataset labels alone: given a target and a background dataset, it separates *salient* factors z_s specific to the target from *common* factors z_c shared by both. For faces with and without eyeglasses as target and background (Fig. 1), z_s should capture eyewear-related variation (e.g., frame style and color) and z_c the content both groups share (e.g., identity, pose, and scene). The stakes are especially high in medicine, where clinical group labels may be available even when the corresponding imaging differences remain unknown. Autism, for instance, is diagnosed from behavior, which tells us who is affected but not how their brains differ; contrasting brain MRI of the autism and control groups can reveal these differences (Aglinskas et al., 2022). Such patterns are subtle, hard to describe in words, and not known in advance, yet

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.

finding them could provide objective imaging markers, deepen the understanding of a disease, and even uncover its subtypes. All of these settings raise the same question:

Can we learn salient representations that capture target-specific detail in each image, reveal subtypes without labels, and generate new examples of what they discover?

Answering it requires a representation that is semantically rich enough for discovery and decodable back to images, and existing CA methods offer complementary strengths toward these goals: generative CA models (Abid & Zou, 2019; Louiset et al., 2023; Carton et al., 2024) decode their factors but reconstruct complex natural images poorly, and contrastive CA methods (Louiset et al., 2024) learn semantically rich factors but have no decoder. In this work, we propose **SAGE** (**Sa**lient factor discovery and **Ge**neration), a contrastive analysis approach that factorizes the frozen spatial latent z of a representation autoencoder (RAE), RAEv2 (Singh et al., 2026), into z_c and z_s, whose sum the frozen RAE decoder maps back to images. Built on a pretrained foundation encoder, DINOv3 (Siméoni et al., 2025), this latent is semantically rich, keeps a high-dimensional spatial layout in which salient factors can retain instance-level detail such as the shape, color, and position of a particular pair of glasses, decodes with high fidelity, and supports state-of-the-art diffusion generation (Zheng et al., 2026; Singh et al., 2026).

This latent, however, mixes target-specific variation with all other image content (Li et al., 2023). Separating the two is nontrivial: reconstruction alone is already satisfied by an empty z_s, a high-dimensional z_s can easily absorb shared content, and it can capture only a generic template of the attribute while instance-specific detail stays in z_c. SAGE counters these failures with objectives that include swapping salient factors between background and target images, keeping them sparse, and requiring them to be recoverable after the swap. The learned salient factors cleanly separate target-specific content and reveal subtypes, and they can be swapped or removed to edit individual images. SAGE can further generate images. Because the salient variation is often not known in advance and has no name or text description, instruction- or prompt-conditioned generation (Liu et al., 2023; Wu et al., 2025) cannot even be applied; instead, in a second stage, SAGE conditions a diffusion transformer on the learned salient representation of a reference image to generate new images that keep its salient subtype while their common content varies. This in turn indicates that the salient condition carries little of the reference's shared content.

Our contributions answer each part of the question above: (1) **Learning clean, high-fidelity salient factors**: salient factors decode to only the digit or eyeglasses, and together with the common factors they reconstruct Digits-ImageNet and FFHQ images with rFID below 2, versus above 120 for generative CA baselines. (2) **Revealing subtypes without labels**: on Digits-ImageNet, the salient space recovers the ten digits far better than a contrastive CA baseline on the same encoder (probe accuracy 0.950 vs. 0.148); on FFHQ, it separates eyewear types, finer sunglasses styles, and a mislabeled cluster; and on retinal OCT, it distinguishes three diseases. (3) **Generating what is discovered**: on Digits-ImageNet, conditioning on a reference's salient representation, without text prompts or subtype labels, preserves its subtype far more often and produces more varied samples than the unfactorized latent (subtype accuracy 90.5% vs. 27.7%; Vendi 12.50 vs. 2.65).

2 Related Work

Contrastive analysis. Contrastive analysis (CA) separates factors shared by background and target datasets from target-specific ones using only dataset labels, with applications in neuroimaging (Aglinskas et al., 2022; Zhu et al., 2026) and disease subgroup discovery (Louiset et al., 2026). cVAE (Abid & Zou, 2019), SepVAE (Louiset et al., 2023), and Double-InfoGAN (Carton et al., 2024) learn the factors with a generator trained from scratch; SepCLR (Louiset et al., 2024) learns strong salient representations under the InfoMax principle but has no decoder, and adding a reconstruction objective weakens its factor separation. CS-StyleGAN (He et al., 2025) separates factors in an inverted StyleGAN latent and refines reconstructions with input features, and the concurrent Diff-CA (Soumm et al., 2026) decomposes compact conditioning tokens of a diffusion generator. These methods represent the salient factor as a low-dimensional vector or global token; SAGE factorizes the full spatial latent of a frozen RAE and uses it for both discovery and generation.

Representation autoencoders and conditioned generation. Recent work replaces the VAE of latent diffusion with frozen pretrained encoders (Zheng et al., 2026; Tong et al., 2026; Shi et al.,

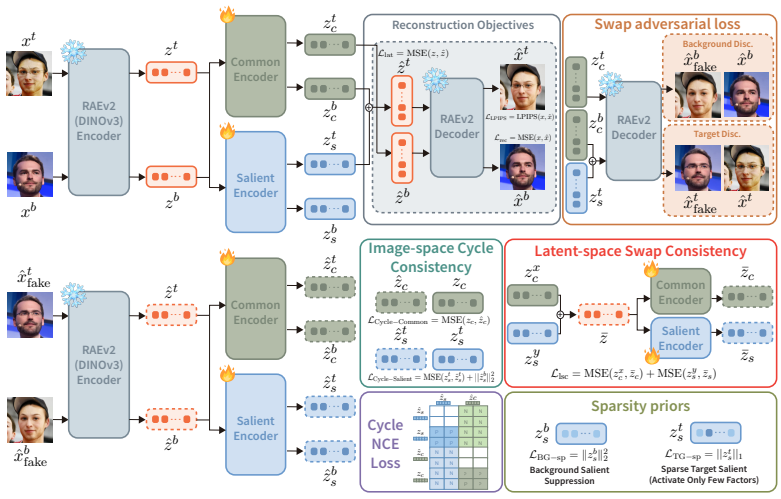


Figure 2: **Stage 1 of SAGE for salient factor discovery.** Trainable common and salient encoders split latents of the frozen RAEv2 encoder (DINOv3); the frozen RAEv2 decoder maps factors back to images. Top: reconstruction ($\mathcal{L}_{\text{rec}}$) and swap adversarial training ($\mathcal{L}^{G}_{\text{swap}}$). Bottom: swapped images are re-encoded for $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{\text{cNCE}}$, while $\mathcal{L}_{\text{lsc}}$ and the sparsity priors act directly on the factors. Snowflakes and flames mark frozen and trainable modules.

2026; Gao et al., 2026); RAEs pair such encoders with ViT decoders that match or exceed standard VAEs (Zheng et al., 2026), and RAEv2 adds REPA (Singh et al., 2026; Yu et al., 2025). SAGE factorizes this latent rather than only generating in it. Representation-conditioned generation (Li et al., 2024) conditions on a full self-supervised representation, and diffusion editing (Meng et al., 2022; Wu et al., 2025) changes prompt-specified attributes; SAGE instead conditions on a discovered salient representation, specifying a subtype by example.

3 Methodology

Given background and target datasets $\mathcal{D}_b$ and $\mathcal{D}_t$, SAGE uses only dataset labels $d \in \{b, t\}$, not subtype labels, to factorize the latent $z = E(x)$ of a frozen RAE with decoder D. Stage 1 learns the factorization; Stage 2 freezes it and trains a salient-conditioned diffusion transformer (Figs. 2, 3).

3.1 Salient Factor Discovery

Why a frozen RAE latent space. SAGE needs a latent that is semantically expressive for discovery and decodable for reconstruction, manipulation, and generation. Unlike the VAE latents of standard latent diffusion, which are trained for reconstruction and carry limited semantic structure (Zheng et al., 2026), RAEs pair frozen pretrained visual encoders with trained decoders, providing both properties and supporting high-quality latent diffusion (Zheng et al., 2026; Singh et al., 2026). We use the frozen RAEv2 encoder and decoder (Singh et al., 2026): E is DINOv3-L (Siméoni et al., 2025) with multi-layer summation (MLS), and D reconstructs images from z. We factorize this spatial latent directly, without a learned bottleneck, to retain localized salient content (Fig. 4); freezing E and D attributes decoded changes to the learned factors and lets Stage 2 reuse RAEv2's pretrained diffusion transformer.

Common and salient encoders. For $z = E(x) \in \mathbb{R}^{N \times C}$, trainable encoders produce same-shape common and salient factors $z_c = E_c(z)$ and $z_s = E_s(z)$, decoded as $D(z_c, z_s) := D(z_c + z_s)$; the sum stays in the frozen decoder's latent space, and $z_s = \mathbf{0}$ means no target-specific content. We require the factors to be decodable and composable, not statistically independent (Appendix E). Because dataset labels alone underdetermine this split, the objectives below reconstruct each image, push target-specific content into z_s, and preserve its image-specific appearance (Fig. 2).

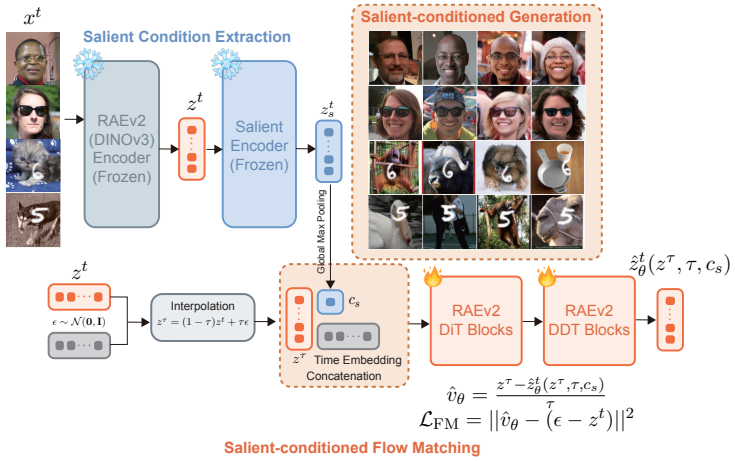


Figure 3: **Stage 2 of SAGE for salient-conditioned generation.** The frozen RAEv2 and salient encoders extract a reference salient representation, max-pooled to c_s. An RAEv2-initialized diffusion transformer is trained with conditional flow matching in this latent. Top right: generated samples; each row keeps the eyewear type or digit of the reference at left while faces and scenes vary.

Asymmetric reconstruction. Because the salient factor should carry only target-specific content, it is absent for background images but complements the common factor for targets: $\hat{x}^b = D(z_c^b, \mathbf{0})$ and $\hat{x}^t = D(z_c^t, z_s^t)$. With $\ell(x, \hat{x}) = \|x - \hat{x}\|_2^2 + \text{LPIPS}(x, \hat{x})$, we minimize

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x^b}\left[\ell(x^b, \hat{x}^b) + \left\|z_c^b - z^b\right\|_2^2\right] + \mathbb{E}_{x^t}\left[\ell(x^t, \hat{x}^t) + \left\|z_c^t + z_s^t - z^t\right\|_2^2\right]. \tag{1}$$

The latent terms tie decoder inputs to real latents, avoiding plausible but unreliable compositions.

Swap adversarial loss. Reconstruction leaves factor allocation underdetermined: for a target image, the trivial split $z_c^t = z^t$, $z_s^t = \mathbf{0}$ reconstructs perfectly. To move target-specific content to z_s, we zero-initialize E_s so that z_s starts empty and acquires content only when required, then swap target salient factors. Swapping exploits dataset labels: if z_s^t captures target-specific content, erasing it should yield a background-like image, while adding it to z_c^b should yield a target-like image: $\hat{x}_{\text{fake}}^b = D(z_c^t, \mathbf{0})$ and $\hat{x}_{\text{fake}}^t = D(z_c^b, z_s^t)$. Because swapped images have no paired ground truth, background and target discriminators Δ_b and Δ_t distinguish $\hat{x}_{\text{fake}}^b$ and $\hat{x}_{\text{fake}}^t$ from reconstructions $\hat{x}^b$ and $\hat{x}^t$, respectively (Goodfellow et al., 2014) (architecture in Appendix C.1); they minimize the hinge loss $\mathcal{L}_{\text{swap}}^D$ (Appendix C.3). Using reconstructions as real samples prevents the discriminators from exploiting decoder artifacts to distinguish swapped outputs from input images. The encoders minimize the following loss, where σ is the sigmoid function:

$$\mathcal{L}_{\text{swap}}^G = -\mathbb{E}\left[\log\sigma\left(\Delta_b(\hat{x}_{\text{fake}}^b)\right)\right] - \mathbb{E}\left[\log\sigma\left(\Delta_t(\hat{x}_{\text{fake}}^t)\right)\right], \tag{2}$$

Sparsity priors. Swap alone does not exclude common content from z_s. Since reconstruction discards z_s^b, we shrink background salient factors to zero with an ℓ_2^2 penalty; for targets, an ℓ_1 penalty favors sparse z_s^t to keep common content, such as the face wearing glasses, out of it:

$$\mathcal{L}_{\text{BG-sp}} = \mathbb{E}_{x^b}\left[\left\|z_s^b\right\|_2^2\right], \qquad \mathcal{L}_{\text{TG-sp}} = \lambda\,\mathbb{E}_{x^t}\left[\left\|z_s^t\right\|_1\right]. \tag{3}$$

The dataset-specific λ balances sparsity against retaining target detail (Appendix C.2).

Consistency objectives. Dataset-level adversarial matching cannot ensure z_s^t carries instance-specific target content: a generic eyeglass prototype could pass both discriminators while instance-specific details remain in z_c^t. We therefore require the factors to be recoverable after swapping, encouraging each salient representation to be instance-specific; the image-space cycle loss re-encodes swapped images, $\hat{z}^t = E(\hat{x}_{\text{fake}}^t)$ and $\hat{z}^b = E(\hat{x}_{\text{fake}}^b)$, to recover their factors (Zhu et al., 2017):

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}\left[\|E_c(\hat{z}^t) - z_c^b\|_2^2 + \|E_s(\hat{z}^t) - z_s^t\|_2^2 + \|E_c(\hat{z}^b) - z_c^t\|_2^2 + \|E_s(\hat{z}^b)\|_2^2\right], \tag{4}$$

With the original factors as fixed targets (Appendix C.4), this loss checks rendered transfers and erasures; its final term enforces no salient content after erasure. The cheaper $\mathcal{L}_{\text{lsc}}$ applies the same recovery in latent space, without decoding: for a random permutation π of the minibatch,

$$\mathcal{L}_{\text{lsc}} = \mathbb{E}_i\Big[\|E_c(z_c^i + z_s^{\pi(i)}) - z_c^i\|_2^2 + \|E_s(z_c^i + z_s^{\pi(i)}) - z_s^{\pi(i)}\|_2^2\Big]. \tag{5}$$

Cycle NCE loss. Recovery losses do not preclude overlapping common and salient content, such as facial appearance present in both. Cycle NCE applies InfoNCE (van den Oord et al., 2018) to globally pooled, ℓ_2-normalized representations before and after swap–re-encoding, using the cycle to form positive pairs without augmentations. Because $\hat{x}^t_{\text{fake}} = D(z_c^b, z_s^t)$ combines factors from different images, its re-encoded factors should recover z_c^b and z_s^t despite the changed context. For each factor, minibatch representations of the other factor, original or re-encoded, serve as negatives; same-factor representations are excluded, so the loss separates the two factors without pushing apart images within one, preserving subtype grouping (Appendix C.5).

Full objective. The encoders minimize the unit-weighted sum,

$$\mathcal{L}_{\text{SAGE}} = \mathcal{L}_{\text{rec}} + \mathcal{L}^G_{\text{swap}} + \mathcal{L}_{\text{BG-sp}} + \mathcal{L}_{\text{TG-sp}} + \mathcal{L}_{\text{cyc}} + \mathcal{L}_{\text{lsc}} + \mathcal{L}_{\text{cNCE}}, \tag{6}$$

where λ enters only through $\mathcal{L}_{\text{TG-sp}}$. The discriminators minimize $\mathcal{L}^D_{\text{swap}}$, and the two updates alternate at each step; the RAEv2 encoder and decoder remain frozen throughout.

3.2 Salient-Conditioned Generation

After Stage 1, the frozen RAEv2 and salient encoders map a target image x^t to $z_s^t = E_s(E(x^t))$; the common encoder is unnecessary. This representation conditions Stage 2 generation (Fig. 3): samples retain the reference's salient characteristics while their common content varies. We condition on z_s rather than the full latent z, which also encodes common content such as the face or scene that generation would otherwise reproduce. Channel-wise max pooling gives $c_s = \text{GMP}(z_s^t) \in \mathbb{R}^C$, discarding location but keeping channels active in few tokens of the sparse map; c_s is projected to one token and concatenated with the noisy latent and time embeddings.

A diffusion transformer (DiT) with RAEv2's DDT head generates latents. Initialized from the pre-trained RAEv2 checkpoint, it is trained with flow matching (Lipman et al., 2023) in the same latent space on target images only. With $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\tau \in [0, 1]$, let $z^\tau = (1-\tau)z^t + \tau\epsilon$. The DiT F_θ predicts the clean latent, giving $\hat{v}_\theta = \big(z^\tau - F_\theta(z^\tau, \tau, c_s)\big)/\max(\tau, \tau_{\min})$, and is trained to match the flow velocity by minimizing $\mathcal{L}_{\text{fm}} = \mathbb{E}_{x^t,\epsilon,\tau}\|\hat{v}_\theta(z^\tau, \tau, c_s) - (\epsilon - z^t)\|_2^2$. During training, a learnable null token replaces c_s with probability p_{uncond}, enabling classifier-free guidance (Ho & Salimans, 2022) and unconditional sampling. At inference, we integrate from Gaussian noise at $\tau = 1$ to $\tau = 0$ conditioned on c_s and decode with frozen D: different noise samples vary common content, while c_s fixes salient characteristics (Appendix D).

4 Experiments

4.1 Experimental Settings

Datasets & metrics. We evaluate on ***Digits-ImageNet***, ImageNet images with and without an overlaid MNIST digit (ten digit subtypes) (LeCun et al., 1998; Deng et al., 2009); ***FFHQ*** faces with and without eyeglasses (reading glasses or sunglasses) (Karras et al., 2019; DCGM); and ***OCT-Kermany*** normal retinal scans versus scans of three diseases (CNV, DME, and DRUSEN) (Kermany et al., 2018). Subtype labels are used only for evaluation (Appendix B.1). We measure *reconstruction* by rFID (Heusel et al., 2017), PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018); *subtype discovery* by Salient LP and ARI (Hubert & Arabie, 1985)/NMI (Strehl & Ghosh, 2002) of k-means clusters; and *generation* by gFID (Heusel et al., 2017), generated-image subtype accuracy, and within-condition Vendi score (Friedman & Dieng, 2023) (Appendices B.4, B.5, and D.2).

Baselines and implementation. We compare with cVAE, SepVAE, Double-InfoGAN, and SepCLR (Abid & Zou, 2019; Louiset et al., 2023; Carton et al., 2024; Louiset et al., 2024), as well as DINOv3 + SepCLR and unfactorized DINOv3 features. Full implementation details appear in Appendices B.2, C, and D.

Table 1: **Benchmark comparison on Digits-ImageNet and FFHQ.** rFID: reconstruction Fréchet Inception Distance; N/A: no decoder; Double-InfoGAN uses its native 128×128 resolution; all other methods use 256×256. BG/TG Sil.: silhouette score of background versus target salient representations (Appendix Table 5). LP Acc.: linear-probe subtype accuracy, five-fold on Digits-ImageNet and balanced on FFHQ (Appendix B.4). ARI/NMI: k-means on the ℓ_2-normalized, unprojected salient representations, with k set to the subtype count (Appendix B.5). FFHQ labels are audit-corrected.

	Digits-ImageNet							FFHQ						
	Reconstruction				Salient representation			Reconstruction				Salient representation		
Method	rFID↓	PSNR↑	SSIM↑	LPIPS↓	BG/TG Sil.↑	LP Acc.↑	ARI/NMI↑	rFID↓	PSNR↑	SSIM↑	LPIPS↓	BG/TG Sil.↑	LP Acc.↑	ARI/NMI↑
cVAE (Abid & Zou, 2019)	151.0	18.33	0.478	0.673	0.006	0.135	0.001/0.003	188.2	18.42	0.567	0.590	0.027	0.898	0.001/0.001
SepVAE (Louiset et al., 2023)	162.5	17.50	0.461	0.696	0.006	0.145	0.009/0.019	197.4	17.67	0.552	0.598	0.206	0.897	-0.001/0.000
Double-InfoGAN (Carton et al., 2024)	171.4	15.97	0.375	0.668	0.013	0.281	0.001/0.003	122.5	16.73	0.424	0.458	0.050	0.964	0.298/0.299
SepCLR (Louiset et al., 2024)	N/A	N/A	N/A	N/A	0.546	0.180	0.000/0.001	N/A	N/A	N/A	N/A	0.752	0.970	0.822/0.696
DINOv3 + SepCLR	N/A	N/A	N/A	N/A	0.642	0.148	0.000/0.001	N/A	N/A	N/A	N/A	0.551	0.976	0.915/0.825
SAGE (Ours)	**1.78**	**20.56**	**0.571**	**0.216**	**0.820**	**0.950**	**0.337/0.472**	**1.65**	**23.02**	**0.723**	**0.163**	**0.873**	**0.983**	**0.939/0.865**

4.2 Salient Factor Discovery

We first ask whether SAGE isolates a clean salient factor, one that contains the target content and little else, while the two factors together still reconstruct the image.

Qualitative results. Among decoder-based methods, only SAGE separates the target attribute cleanly (Fig. 4; additional randomly selected examples in Appendix Fig. 9): it reconstructs scenes and faces with their digits or eyeglasses, common-only decoding removes only the target attribute, and salient-only decoding renders the digit or eyeglasses alone at their original location and shape. Generative baselines blur reconstructions and alter the scene or face even in common-only decoding; their salient-only outputs show no recognizable digit on Digits-ImageNet and blurred eyeglasses with much of the face on FFHQ, indicating common-content leakage. At the dataset level, SAGE's salient t-SNE separates background from target images (Fig. 5a, d), as the sparsity prior drives background salient factors toward zero; those of cVAE and, on Digits-ImageNet, SepVAE mix the two (Appendix Fig. 10).

Quantitative results. SAGE achieves rFID 1.78 on Digits-ImageNet and 1.65 on FFHQ, whereas generative baselines trained from scratch exceed 120 (Table 1), because SAGE factorizes a frozen RAEv2 latent and decodes it with the pretrained decoder, adding little error over decoding z directly (1.78 versus 0.39 on Digits-ImageNet). Its salient space also separates background from target images best among all methods (silhouette 0.820 and 0.873). Linear probes (Appendix A.9) show that digit identity is far more accessible in z_s than in z (0.950 versus 0.328), while ImageNet category, a property of the shared scene, is readily decoded from z and z_c (0.765) but hardly from z_s (0.045). SepCLR and DINOv3 + SepCLR lack decoders; their comparison follows.

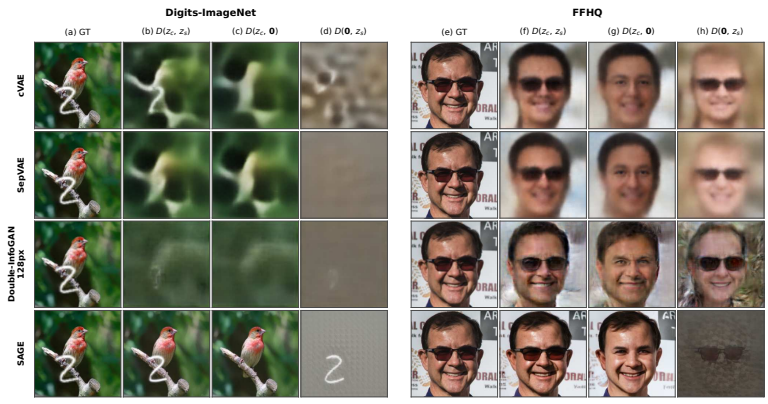


Figure 4: **Reconstruction, common-only, and salient-only decoding.** Columns show the input, the full reconstruction $D(z_c, z_s)$, common-only decoding $D(z_c, \mathbf{0})$, and salient-only decoding $D(\mathbf{0}, z_s)$ on Digits-ImageNet (left) and FFHQ (right); rows compare cVAE, SepVAE, Double-InfoGAN, and SAGE. A clean salient factor renders only the digit or eyeglasses.

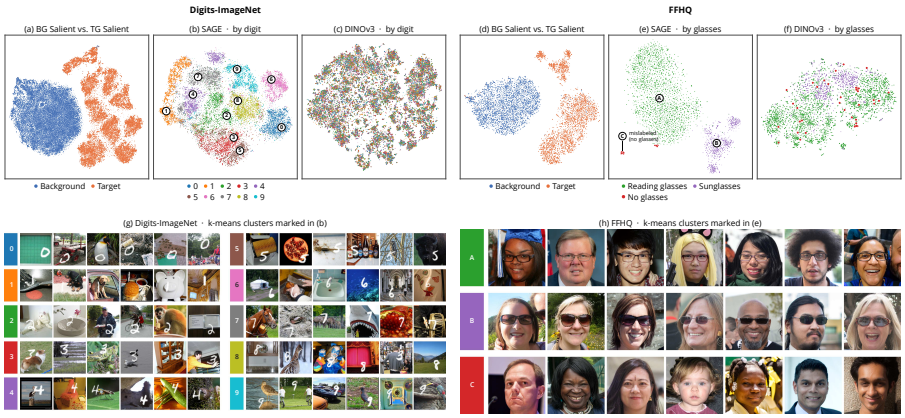


Figure 5: **Subtype discovery in the salient space.** Top: t-SNE of SAGE's salient representations separates background and target images (a, d) and organizes digit and eyewear subtypes (b, e), whereas raw DINOv3 features mix them (c, f). Bottom: examples from the k-means clusters marked in (b, e) show coherent digit and eyewear groups (g, h); FFHQ cluster C contains images mislabeled as wearing glasses, and red points are audited no-glasses images.

4.3 Subtype Discovery

We next ask whether the salient factor discovers target subtypes without subtype labels, both the expected ten digits and two eyewear types and groups we did not expect.

Qualitative results. Figure 5 shows t-SNE (van der Maaten & Hinton, 2008) of salient representations, colored by dataset label and, for targets, subtype labels used only for visualization. Background–target separation (Fig. 5a, d; Section 4.2) does not imply subtype separation: SepCLR and DINOv3 + SepCLR also separate the datasets, yet their target digits remain mixed (Appendix A.3). SAGE recovers the expected subtypes: the ten digit identities form locally coherent, partly overlapping groups, and reading glasses and sunglasses form two separate groups (Fig. 5b, e); raw z separates neither (Fig. 5c, f). Samples from each group (Fig. 5g, h) share the digit or eyewear type while their scenes and faces differ, so the grouping follows the target attribute rather than shared content.

Discovering finer eyewear styles. Within the sunglasses group (Fig. 5e), SAGE separates three styles: thick, angular wayfarer-like frames; thin-rimmed aviator-like lenses; and a broader mixed group (Appendix Fig. 11). The names describe the samples; the data have no style labels.

Discovering mislabeled images. Beyond reading glasses and sunglasses, the FFHQ t-SNE contains cluster C (Fig. 5e, h), whose images lack eyeglasses but whose automatic annotations label them as wearing glasses, often because of face paint or hat-brim shadows (Appendix Fig. 13c). This cluster led us to audit all 3,437 target validation images: annotators labeled the visible eyewear without seeing SAGE's assignments and corrected 125 labels (3.64%; all corrections are shown in Appendix Fig. 12), including 39 no-glasses images; SAGE places 34 of these 39 in cluster C, while the other five show face paint or masks around the eyes (Appendix Fig. 13a, b). With $k = 3$, including no-glasses images, SAGE reaches ARI/NMI $0.937/0.865$, whereas every baseline remains below $0.46/0.56$; e.g., DINOv3 + SepCLR drops from $0.915/0.825$ to $0.453/0.550$ (Appendix Table 4). SAGE thus also exposes label noise; Table 1 uses the audit-corrected labels for all methods.

Quantitative results. On Digits-ImageNet, SAGE reaches digit probe accuracy 0.950, versus at most 0.281 for baselines (Table 1) and 0.328 for the unfactorized RAEv2 latent z (Appendix Table 7); because SAGE and z share an encoder, this gain comes from factorization. Its ARI/NMI is $0.337/0.472$, versus at most $0.009/0.019$ for baselines. On FFHQ, SAGE achieves the best probe accuracy and clustering (0.983; $0.939/0.865$), ahead of DINOv3 + SepCLR (0.976; $0.915/0.825$). High probe accuracy alone does not ensure well-separated clusters: cVAE and SepVAE reach FFHQ probe accuracies near 0.90 yet ARI near zero. SAGE helps most when z obscures subtypes.

4.4 Factor Manipulation

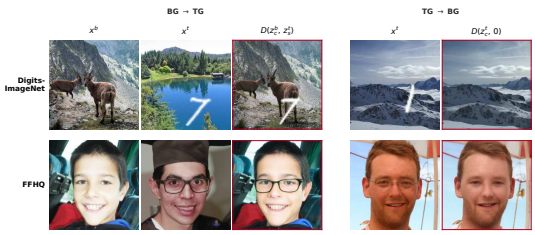


Figure 6: **Salient swapping and erasure.**

Because both factors live in the latent space of the frozen decoder, the salient factor can be moved between images, removed, or interpolated, and the result decoded directly. We test whether such edits change only the target attribute. **Swapping and erasure.** Adding a target image's salient factor to a background image transfers its particular attribute (Fig. 6, left): a digit 7 retains its location and stroke shape, and thick black frames, rather than generic glasses, appear on a new face while the scene and identity remain. Conversely, zeroing a target's salient factor removes the digit or reading glasses while preserving the rest (Fig. 6, right; more in Fig. 14). These edits illustrate that each salient factor carries its own image's target attribute and little else. **Interpolation.** Interpolating salient factors of two target images while retaining the first common factor changes mainly the target attribute (Fig. 7, salient-only rows): a digit turns from 1 to 7 on a fixed photograph, and clear reading glasses darken into sunglasses on a fixed face. Interpolating both factors cross-fades the whole image (full-latent rows). Salient interpolation thus traverses subtypes while preserving common content.

4.5 Discovering Disease Subtypes in Retinal OCT

Unlike digits and eyeglasses, retinal disease subtypes are clinically subtle and vary across patients. On OCT-Kermany, SAGE is trained only with normal/disease labels. **Disease subtypes.** Its salient space separates the three diseases, with t-SNE groups that overlap mainly at their boundaries (Fig. 8, left), a linear-probe accuracy of 0.960, and ARI/NMI $0.415/0.415$ (Appendix A.7). **Disease-specific edits.** Adding a disease scan's salient factor to a normal scan introduces fluid-filled cavities, while removing it yields a retina closer to normal (Fig. 8, middle and right), suggesting disease-related features rather than patient-specific anatomy. These edits are qualitative, not clinical evidence.

4.6 Salient-Conditioned Generation

Representation-conditioned generation (Li et al., 2024) conditions on a full self-supervised representation; we instead condition on SAGE's pooled salient representation c_s. Raw DINOv3 GMP pools the unfactorized latent z identically, testing whether factorization removes reference face or scene information. **Qualitative results.** Salient conditioning generates new faces and scenes that keep the reference digit or eyewear type, whereas Raw DINOv3 GMP reproduces the reference person or scene almost exactly and often changes the digit (Appendix Fig. 15); Fig. 3 (top right) shows several samples per reference. **Quantitative results.** On Digits-ImageNet, salient conditioning raises subtype accuracy from 27.7% to 90.5% and Vendi from 2.65 to 12.50 (real: 15.79), and lowers Cos-to-ref from 0.794 to 0.116 (Table 2). On FFHQ, both conditions preserve eyewear type

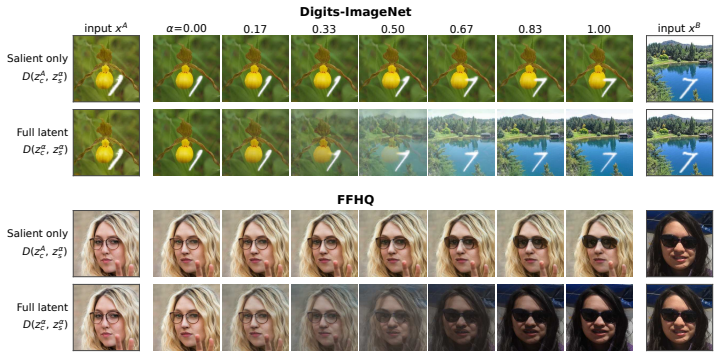


Figure 7: **Salient interpolation.** For two held-out target images x^A and x^B, salient-only rows decode $D(z_c^A, z_s^\alpha)$ with $z_s^\alpha = (1-\alpha)z_s^A + \alpha z_s^B$; full-latent rows interpolate both factors.

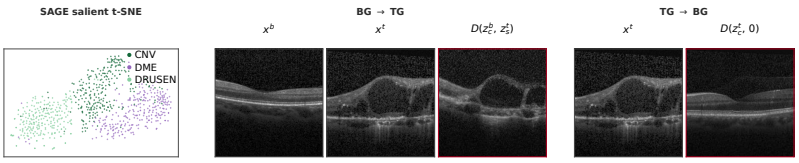


Figure 8: **OCT disease subtypes and salient edits.** Left: salient t-SNE of disease scans, colored by disease subtype labels unused in training. BG→TG adds a disease scan's salient factor to a normal scan; TG→BG removes a disease scan's salient factor.

Table 2: **Salient-conditioned generation.** Raw DINOv3 GMP conditions on the unfactorized latent. Subtype Acc.: share of samples assigned the reference subtype by a DINOv3-L classifier trained on real targets. Vendi (diversity) and Cos-to-ref (reference similarity) use 16 samples per reference and DINOv2-L features. *Real images*: classifier accuracy and within-subtype Vendi (Appendix D.2).

Dataset	Conditioning	Uncond. gFID ↓	Cond. gFID ↓	Subtype Acc. ↑	Vendi ↑	Cos-to-ref ↓
Digits-ImageNet	*Real images*	–	–	*99.3*	*15.79*	–
	Raw DINOv3 GMP	8.86	**2.40**	27.70	2.65	0.794
	SAGE Salient GMP	**4.76**	4.09	**90.48**	**12.50**	**0.116**
FFHQ	*Real images*	–	–	*99.1*	*10.09*	–
	Raw DINOv3 GMP	16.07	**11.17**	**98.5**	2.32	0.779
	SAGE Salient GMP	**15.80**	16.03	96.6	**6.96**	**0.470**

(96.6% versus 98.5%), but only salient conditioning produces varied samples (Vendi 6.96 versus 2.32; Cos-to-ref 0.470 versus 0.779). Since the two conditions differ only in the salient encoder, reduced copying supports z_s as a more selective condition. Raw DINOv3 GMP attains lower conditional gFID, consistent with closer reference copying rather than subtype control, while SAGE has lower unconditional gFID on both datasets.

4.7 Ablation Studies

Table 3: **Loss ablations on Digits-ImageNet.** Common-only: rFID and SSIM of $D(z_c, \mathbf{0})$ for target images against their digit-free ImageNet backgrounds. Salient probes measure digit identity (target content, higher is better) and ImageNet category (shared content, lower is better). ARI/NMI are computed as in Table 1; their standard deviations over three k-means seeds are at most 0.0007.

	Full Reconstruction				Common-only		Salient Linear Probe		Clustering
Configuration	rFID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	rFID ↓	SSIM ↑	Digit ↑	ImageNet ↓	ARI/NMI ↑
SAGE (all objectives)	1.78	20.56	0.571	0.216	1.89	0.569	0.950	0.045	0.337/0.472
w/o swap adversarial ($\mathcal{L}^G_{\text{swap}}$)	1.01	20.67	0.566	0.210	4.88	0.544	0.123	0.169	0.000/0.001
w/o target sparsity ($\mathcal{L}_{\text{TG-sp}}$)	1.15	21.06	0.598	0.193	1.36	0.585	0.923	0.302	0.274/0.409
w/o Cycle NCE ($\mathcal{L}_{\text{cNCE}}$)	1.29	20.78	0.576	0.205	1.58	0.573	0.950	0.100	0.348/0.479
w/o image-space cycle consistency ($\mathcal{L}_{\text{cyc}}$)	2.84	19.88	0.562	0.236	3.46	0.559	0.951	0.086	0.360/0.496
w/o latent-space swap consistency ($\mathcal{L}_{\text{lsc}}$)	1.82	20.30	0.568	0.218	2.25	0.567	0.953	0.107	0.361/0.494

Table 3 removes each objective in turn. The **swap adversarial loss** is what moves target content into the salient factor: without it, the digit probe drops from 0.950 to 0.123 and ARI/NMI to nearly zero, consistent with z_s staying near its zero initialization. The other objectives mainly keep shared content out: removing **target sparsity**, **Cycle NCE**, or either **consistency** loss raises the salient ImageNet probe from 0.045 to between 0.086 and 0.302, and removing the image-space cycle loss also degrades reconstruction (rFID 1.78 → 2.84). We keep all objectives, which together give the lowest salient ImageNet probe accuracy (Appendix A.9).

5 Conclusion

We presented SAGE, which learns salient factors in the high-dimensional spatial latent of a frozen representation autoencoder using only background/target labels. Its factors faithfully reconstruct and edit individual images. Without subtype labels, its salient representation recovers digits, eyewear types, and retinal diseases, reveals finer sunglasses styles and mislabeled images, and conditions the generation of new images that keep a discovered subtype while common content varies.

REPRODUCIBILITY STATEMENT

We will release code, trained models, and the audited FFHQ labels upon publication. All experiments use public data: MNIST and ImageNet for Digits-ImageNet, FFHQ with the FFHQ Features eyewear annotations, and OCT-Kermany. Appendix B.1 describes how each dataset is built, including the Digits-ImageNet compositing, the audited FFHQ evaluation labels, and the OCT preprocessing and splits. Appendices C and D give the Stage 1 and Stage 2 architectures, objectives, optimization settings, and checkpoints, and Appendices B.2 and B.3 give the baseline configurations. Appendices B.4, B.5, and D.2 specify the linear-probe, clustering, and generation evaluation protocols. SAGE uses the pretrained RAEv2 encoder and decoder of Singh et al. (2026), which remain frozen throughout.

AI USE STATEMENT

In this work, we used generative AI to assist with the implementation of experimental methods and code. We have not used generative AI tools for other tasks requiring disclosure, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to polish the writing of the manuscript. All AI-assisted experimental code was reviewed and tested by the authors, and the resulting experimental outputs were also verified by the authors. All AI-assisted writing was also reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code and results produced with the aid of generative AI.

REFERENCES

Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019. URL https://arxiv.org/abs/1902.04601.

Aidas Aglinskas, Joshua K Hartshorne, and Stefano Anzellotti. Contrastive machine learning reveals the structure of neuroanatomical variation within autism. *Science*, 376(6597):1070–1074, 2022. doi: 10.1126/science.abm2461. URL https://www.science.org/doi/10.1126/science.abm2461.

Florence Carton, Robin Louiset, and Pietro Gori. Double InfoGAN for contrastive analysis. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 172–180, 02–04 May 2024.

DCGM. FFHQ Features Dataset: Gender, age, and emotion for Flickr-Faces-HQ dataset. https://github.com/DCGM/ffhq-features-dataset. Accessed: 2026-09-23.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023, 2023. ISSN 2835-8856.

Yuan Gao, Chen Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4688–4697, June 2026.

Ian J. 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*, volume 27, 2014.

Yunlong He, Gwilherm Lesné, Ziqian Liu, Michaël Soumm, and Pietro Gori. Learning common and salient generative factors between two image datasets. *arXiv preprint arXiv:2512.12800*, 2025. URL https://arxiv.org/abs/2512.12800.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985. ISSN 1432-1343. doi: 10.1007/BF01908075. URL https://doi.org/10.1007/BF01908075.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A. N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131.e9, February 2018. ISSN 1097-4172 0092-8674. doi: 10.1016/j.cell.2018.02.010.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Seunghun Lee, Sunghyun Cho, and Sunghoon Im. DRANet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15252–15261, 2021.

Tianhong Li, Lijie Fan, Yuan Yuan, Hao He, Yonglong Tian, Rogerio Feris, Piotr Indyk, and Dina Katabi. Addressing feature suppression in unsupervised visual representations. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1411–1420. IEEE, 2023.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: a self-supervised representation generation method. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, 2024. ISBN 9798331314385.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Robin Louiset, Edouard Duchesnay, Antoine Grigis, Benoit Dufumier, and Pietro Gori. Sep-vae: a contrastive vae to separate pathological patterns from healthy ones. *arXiv preprint arXiv:2307.06206*, 2023. URL https://arxiv.org/abs/2307.06206.

Robin Louiset, Edouard Duchesnay, Antoine Grigis, and Pietro Gori. Separating common from salient patterns with contrastive representation learning. In *International Conference on Learning Representations*, volume 2024, pp. 26887–26918, 2024.

Robin Louiset, Edouard Duchesnay, Benoit Dufumier, Antoine Grigis, and Pietro Gori. Automatic discovery of disease subgroups by contrasting with healthy controls. *arXiv preprint arXiv:2605.21301*, 2026. doi: 10.48550/arXiv.2605.21301. URL https://arxiv.org/abs/2605.21301.

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. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *European Conference on Computer Vision*, 2020.

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*, volume 2026, pp. 154506–154537, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fa5ddd6bac0d665c72969d79221b680a-Paper-Conference.pdf.

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. URL https://arxiv.org/abs/2508.10104.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026. URL https://arxiv.org/abs/2605.18324.

Michaël Soumm, Alexandre Fournier Montgieux, Yunlong He, Pietro Gori, and Alasdair Newson. Diff-CA: Separating common and salient factors with diffusion models. *arXiv preprint arXiv:2606.06120*, 2026. doi: 10.48550/arXiv.2606.06120. URL https://arxiv.org/abs/2606.06120.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026. URL https://arxiv.org/abs/2601.16208.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

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*, volume 2025, pp. 87400–87442, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d9e42b4d7163931f3689d6d6fbaa11d0-Paper-Conference.pdf.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Yu Zhu, Aidas Aglinskas, and Stefano Anzellotti. DeepCor: denoising fMRI data with contrastive autoencoders. *Nature Methods*, 23(2):334–337, 2026. doi: 10.1038/s41592-025-02967-x. URL https://www.nature.com/articles/s41592-025-02967-x.