NeurIPS 2026 Spotlight

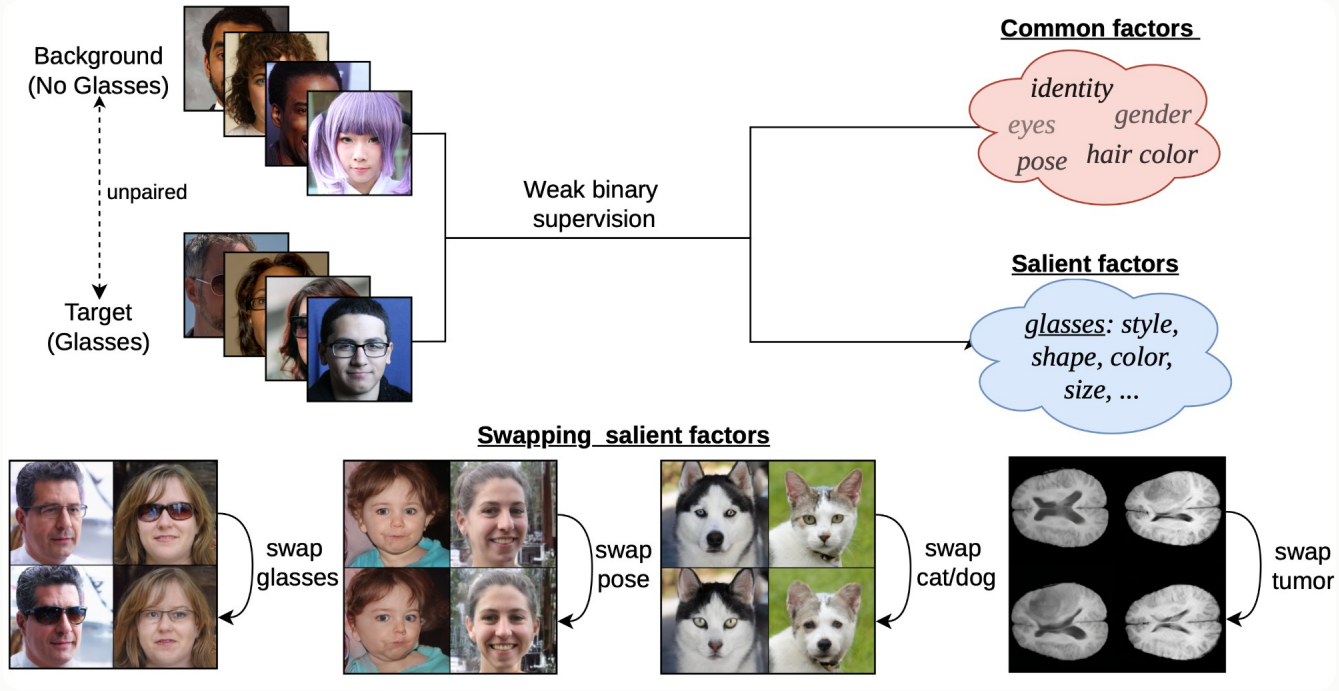
Diff-CA

Separating Common and Salient Factors with Diffusion Models

Michaël Soumm[1*] · Alexandre Fournier Montgieux[2*] · Yunlong He[1] · Pietro Gori[1] · Alasdair Newson[1,3]

[1] Télécom Paris, Institut Polytechnique de Paris [2] CEA List, Palaiseau [3] ISIR, Sorbonne Université * Equal contribution

1 Contrastive Analysis



Two unpaired datasets: a **background** (faces without glasses) and a **target** (faces with glasses).

Contrastive Analysis separates what is **common** to both (identity, pose, gender...) from what is **salient** to the target (glasses style, shape, color...), from the binary label alone.

1 Why Contrastive Analysis?

Discovery

A binary label (healthy / tumor) is cheap, while a description of what differs may not exist.

Control

Swap, remove or interpolate the salient factor, keeping the rest fixed.

Gap

VAE/GAN-based CA models are blurry, so what the salient code captures cannot be verified; text-guided editors need the factor to be named.

1 Contributions

1 **Theory**

Identifiability of the additive common/salient decomposition, and a bound on control.

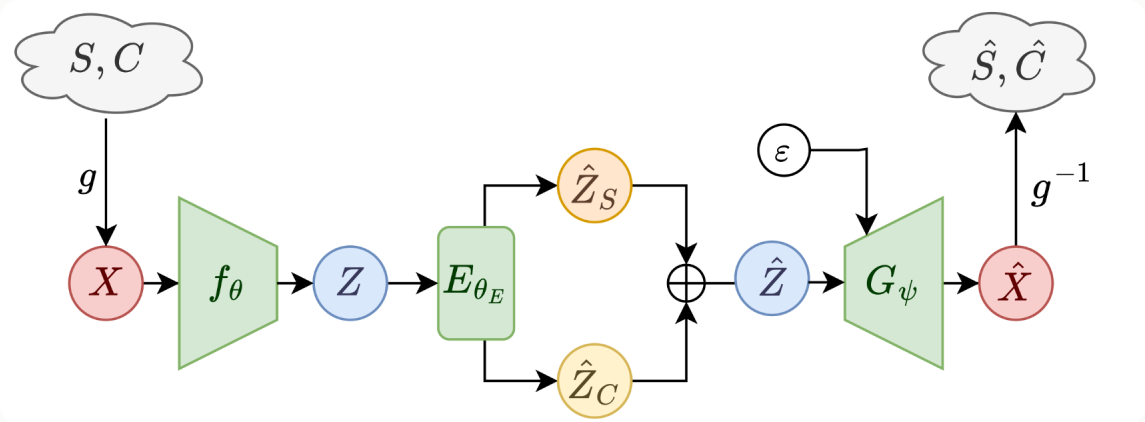
2 **Conditioning**

A high-fidelity, image-only conditioning space for diffusion (32 tokens).

3 **Diff-CA**

Contrastive Analysis on diffusion conditioning tokens, with 94.5% vs. 47.0% swap accuracy over the best CA baseline.

2 Identifiable CA



Factors (S, C) generate the image X. The separator E splits the conditioning Z into a salient and a common code, whose sum conditions the generator G.

Images come from independent factors: a salient S, zero on the background, and a common C. We assume an additive conditioning space: Z is the sum of a salient part and a common part, and the salient part is zero on the background.

Theorem 1 · Identifiability

If a learned split of Z into a salient and a common code satisfies

(i) **additive reconstruction:** the two codes sum to Z;

(ii) **pinning:** the salient code is zero exactly on the background;

(iii) **independence:** the two codes are independent,

then they equal the true salient and common parts: the factors are recovered.

2 From theory to design

Proposition 1 · Control

How well the salient code controls the generated salient factor is bounded below by

> **+ fidelity** of the generator
>
> **– noise dependence** of the output
>
> **– entanglement** of the two codes

Control needs a faithful generator that depends little on the noise, and disentangled codes.

Stage 1

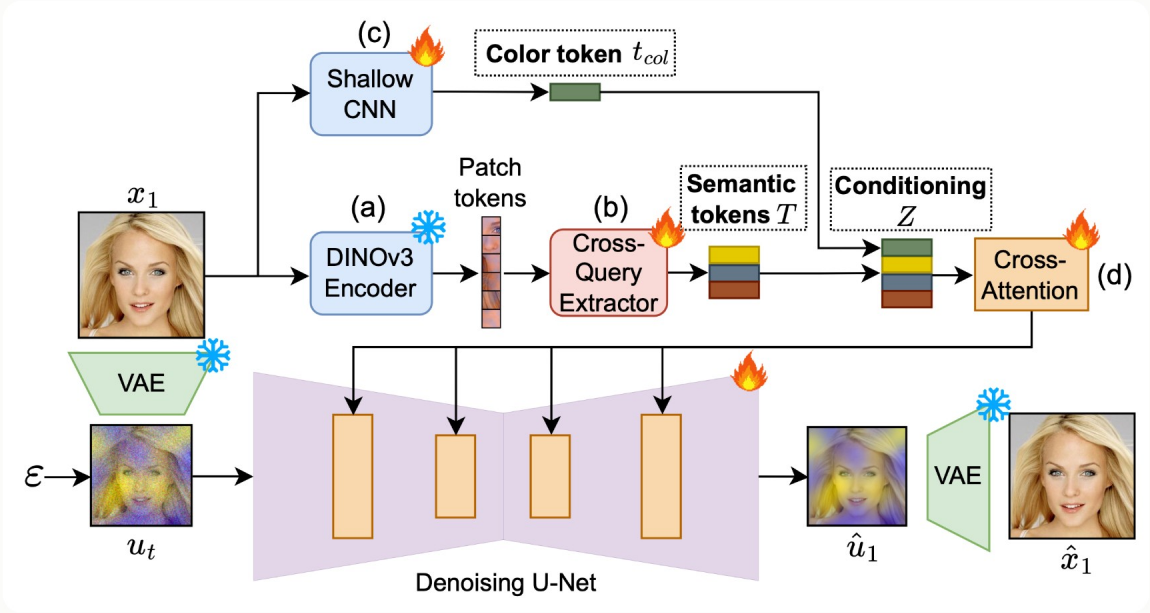
Builds a generator with high fidelity and low noise dependence.

Stage 2

Trains a separator towards (i)–(iii), with reconstruction, pinning, adversarial and cycle losses.

3 Stage 1 · Image-only conditioning

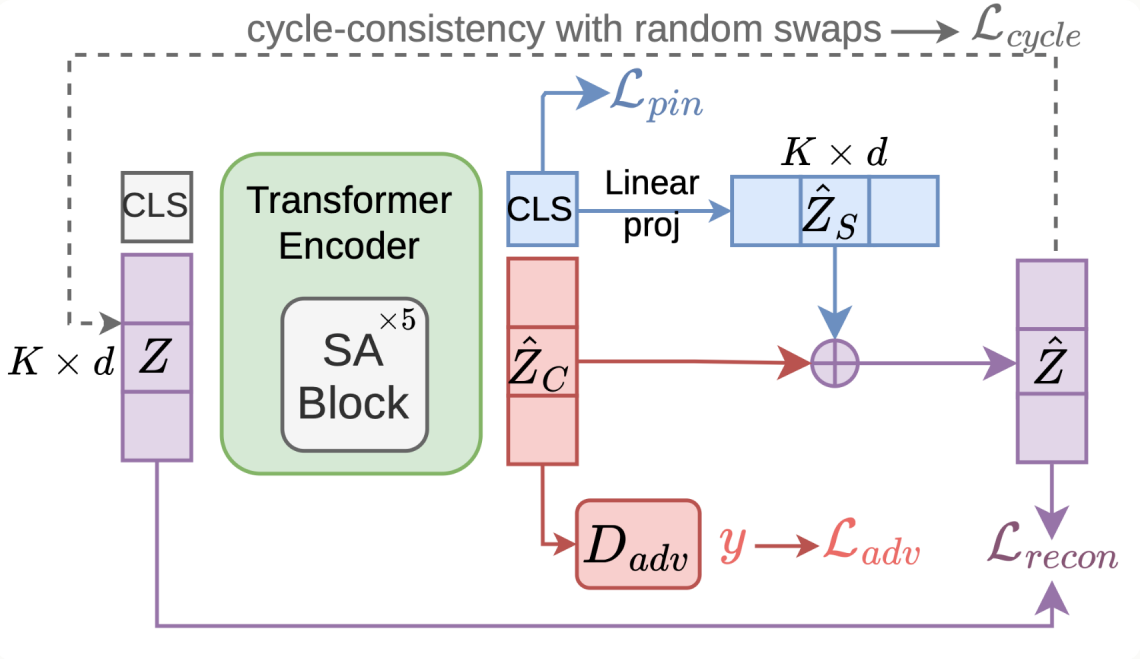
A latent flow-matching U-Net, trained from scratch and conditioned only on the image. A Cross-Query extractor compresses frozen DINOv3 features into 32 tokens, plus a color token. Deterministic sampling: the image is driven by *Z*, not by the noise.



Conditioning	Tokens	SSIM ↑	LPIPS ↓	FID ↓
DiffAE	1 × 512	.530	.209	14.28
EncDiff	32 × 128	.420	.323	18.43
DINOv3	261 × 128	.645	.099	5.66
Ours (CQC)	32 × 128	.611	.114	5.52

Reconstruction on CelebA-HQ (trained on FFHQ). With 8× fewer tokens than DINOv3, our conditioning stays close, and unlike DINOv3 it supports editing.

3 Stage 2 · Common/salient separator



A 5-layer Transformer splits Z into a **salient** code (CLS token) and a **common** code (32 tokens), summed back to Z.

Reconstruction (i): the two codes sum back to Z.

Pinning (ii): the salient code is zero on the background, non-zero on the target.

Adversarial: the common code carries no information on the label (gradient reversal).

Cycle (iii): swapped codes re-encode consistently.

4 Results on FFHQ glasses

94.5%

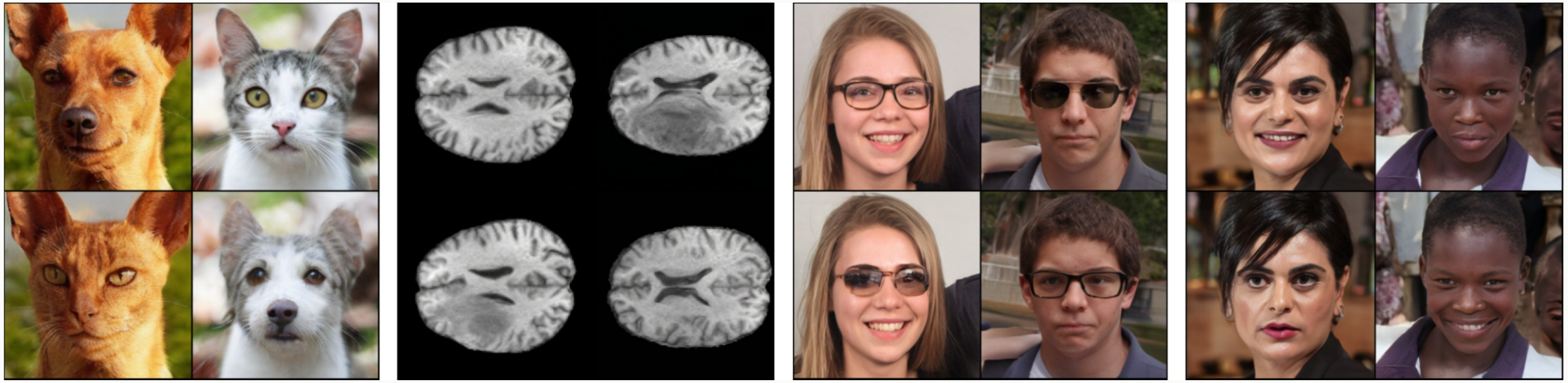
swap accuracy on fine-grained glasses classes, from binary labels only

Best CA baseline: 47.0%

Method	SSIM ↑	LPIPS ↓	Acc ↑	ID-Sim ↑	Gender ↑	Smile ↑	Pose ↓
MM-cVAE	.370	.643	11.8	.302	62.3	57.7	4.62
SepVAE	.358	.642	19.6	.301	57.7	56.4	8.29
DoubleInfoGAN	.326	.430	47.0	.314	68.5	79.2	7.35
Diff-CA (ours)	.610	.115	94.5	.474	92.5	92.9	1.87

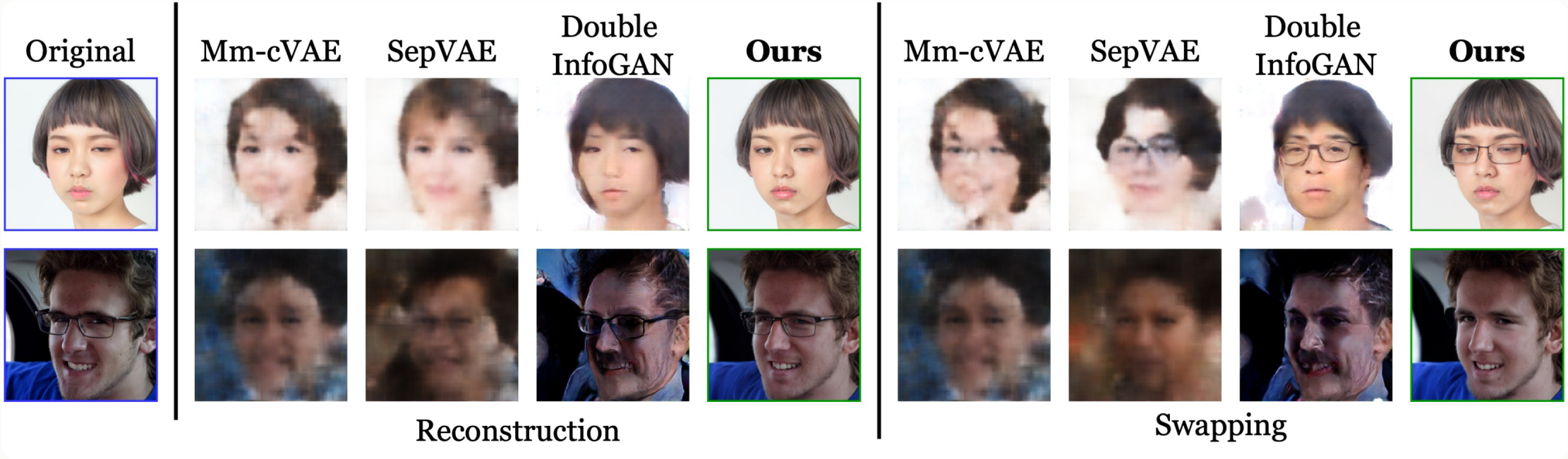
Swap accuracy (%) on fine-grained classes (none, reading, sunglasses), learned from binary labels only. ID-Sim, Gender, Smile (%) and Pose (MAE, °) measure how well common attributes are preserved by the swap.

4 Swapping the salient factor



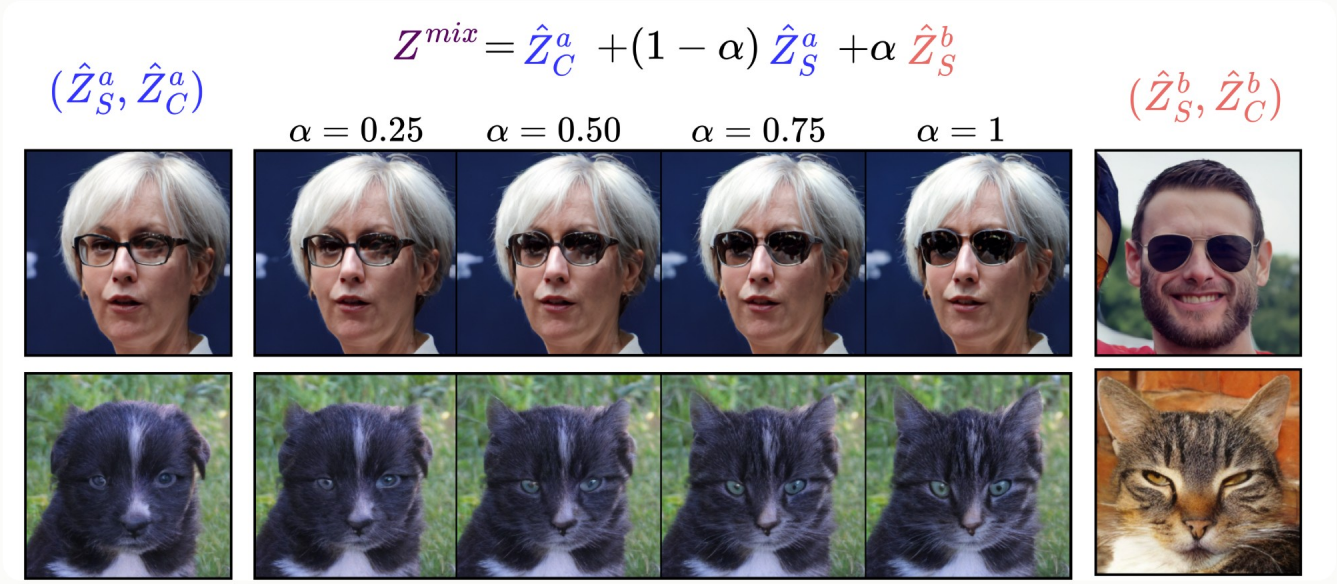
AFHQ (cat), BraTS (tumor) and FFHQ (glasses, smile). Top: real background and target images; bottom: after swapping the salient factor, with the common factor kept.

4 Comparison with CA baselines



VAE- and GAN-based CA baselines swap the attribute, but produce blurry images and lose identity.

4 Interpolation and ablations



Salient interpolation with the common code fixed: the glasses style (top) and cat features (bottom) change progressively.

Ablations

Variant	Acc ↑
Without adversarial loss	≤ 2.1
+ adversarial loss	91.4
+ cycle loss (removes color shifts)	94.5
Our separator on EncDiff tokens	58.1

Swap accuracy (%) on FFHQ glasses. On EncDiff tokens of the same size, our separator still beats all CA baselines: both stages matter.

Take-home

Contrastive Analysis on diffusion conditioning tokens gives high-fidelity images and a clean separation of common and salient factors, from a binary label alone.

Project page

msoumm.github.io/Diff-CA