HIGH RETRIEVAL ACCURACY DOES NOT IMPLY DISENTANGLEMENT: DIAGNOSING AND REPAIRING PSEUDO-DISENTANGLEMENT IN FROZEN VISUAL ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval accuracy is the default criterion for choosing a visual encoder, yet it is silent about whether content and style are actually separated: discriminable class boundaries suffice for high mAP even when the two factors are thoroughly mixed. We define this failure mode operationally—a representation is pseudo-disentangled when its retrieval mAP is high while its cross-factor leakage is not low—and make it measurable. Our metric, DScore, combines retrieval utility with two subspace-level leakage diagnostics, a kNN leakage proxy and a probe leakage rate; unlike the axis-aligned metrics (DCI, MIG, SAP), it does not assume each factor aligns with one latent dimension. Across nine frozen representations and three benchmarks, "high accuracy $\neq$ high disentanglement" holds throughout: the best style retriever falls to third once leakage is penalized, while a self-supervised encoder ranking eighth by style mAP ranks first. We then ask whether this entanglement can be repaired without fine-tuning the encoder, and show that it can. Cross-Nullspace Projection (CNP) is a closed-form, supervised post-processing step with two parts—a factor-specific LDA readout, and a bidirectional cross-projection that makes the two readouts approximately orthogonal. Most of its DScore gain (mean $+0.15$; probe leakage cut by 60–86%) comes from the readout, not the projection, but the projection is what makes the two factor vectors independently usable, lifting compositional retrieval from 0.19 to 0.45 mAP ($2.4\times$). Across the benchmarks studied, content-style entanglement in frozen encoders is thus substantially linear-remediable rather than a deep semantic fusion.

1 INTRODUCTION

Decomposing visual representations into content ("what is depicted") and style ("how it is depicted") is foundational for art analysis (Somepalli et al., 2024; Kumar et al., 2025), controllable generation (Huang & Belongie, 2017; Trinh et al., 2026; Ko et al., 2026), factorized retrieval (Gatys et al., 2016), and domain generalization (Peng et al., 2019; Li et al., 2017). Our experiments bear on the third of these, using compositional retrieval as its operational proxy. While training-based methods (β-VAE (Higgins et al., 2017), DICE (Zhang et al., 2026), SCAdapter (Trinh et al., 2026)) inject independence constraints during optimization, the dominant practice in visual applications is to reuse frozen pretrained encoders (CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), SigLIP2 (Tschannen et al., 2025), Qwen2.5-VL (Bai et al., 2025), DINOv3 (Siméoni et al., 2025)) as off-the-shelf backbones, selecting among them by downstream accuracy. This practice rests on an implicit assumption that has not been systematically tested, to our knowledge: that a representation which retrieves well also separates content from style. We show that this assumption can fail, and that the failure is quantifiable.

Retrieval accuracy is insensitive to entanglement. Representation quality is predominantly assessed by retrieval accuracy (mAP@K, Recall@K), whose implicit criterion is whether same-class samples rank at the top. Yet accuracy is achievable whenever class boundaries are linearly discriminable—a condition that holds even when content and style information are thoroughly interleaved within the embedding. We term this failure mode *pseudo-disentanglement*: a representation

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.

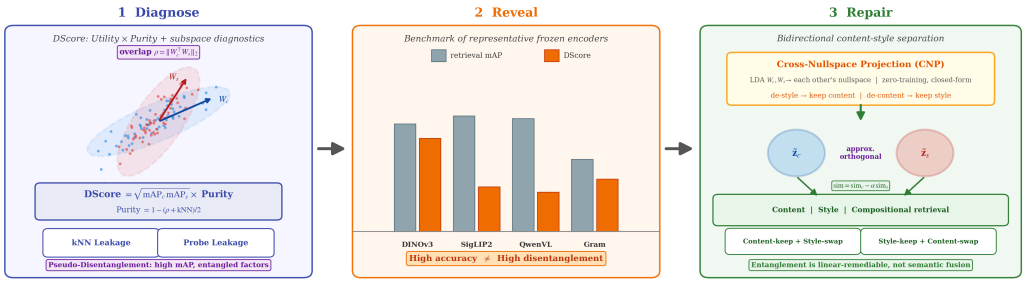

Figure 1: **Overview of our diagnose→reveal→repair chain. (1) Diagnose:** frozen features attain high retrieval mAP yet content/style subspaces overlap (overlap $\rho = \|W_c^\top W_s\|_2$); we quantify this with **DScore** (retrieval utility × disentanglement purity) and the kNN and probe leakage diagnostics. **(2) Reveal:** benchmarking 9 frozen representations shows high mAP does not imply low leakage—"high accuracy ≠ high disentanglement." **(3) Repair:** Cross-Nullspace Projection (CNP) cross-projects LDA subspaces onto each other's nullspace, yielding two vectors $\tilde{z}_c, \tilde{z}_s$ with approximately orthogonal estimated subspaces that enable content, style, and factorized compositional retrieval, without fine-tuning the encoder.

scores high on retrieval, while its content and style are not truly separated but merely rendered separable by discriminative boundaries. The consequence is practical and silent: selecting encoders by mAP alone can promote style descriptors that leak substantial content, causing downstream failures precisely in the tasks—factorized retrieval, controllable generation, spurious-correlation mitigation—that demand strict factor isolation.

Existing metrics do not expose this gap. Classical disentanglement metrics (DCI (Eastwood & Williams, 2018), MIG (Chen et al., 2018), SAP (Kumar et al., 2018)) presume axis-aligned latent factors in generative models, whereas content and style in encoder representations occupy multi-dimensional, non-axis-aligned subspaces, rendering per-dimension estimation unreliable. Retrieval benchmarks (VTAB (Zhai et al., 2019)) remain within the accuracy-as-quality paradigm and never interrogate cross-factor leakage. Concept-erasure methods from NLP (INLP (Ravfogel et al., 2020), LEACE (Belrose et al., 2023)) target the unidirectional removal of a single attribute for fairness, not the bidirectional separation and retention of two visual factors. A metric and protocol that directly quantify cross-factor leakage in frozen representations are therefore, to our knowledge, absent.

Contributions. We diagnose the phenomenon, measure how consistently it appears across training paradigms, and ask whether it can be repaired:

- **Diagnosis.** We propose DScore, a scalar that jointly measures retrieval utility and disentanglement purity for frozen representations, plus two subspace-level diagnostics—a kNN leakage proxy and a probe-based leakage rate—that together form an internally consistent suite (formal analysis in Appendix B).
- **Reveal.** We benchmark nine frozen representations under a unified protocol and examine pseudo-disentanglement across training paradigms: self-supervised encoders (DINOv2/v3) achieve the best content-style balance, semantic-alignment encoders (SigLIP2, Qwen2.5-VL) attain strong retrieval yet heavy entanglement, and style-specialized descriptors (Gram, IntroStyle) exhibit low leakage but content-blindness—rendering "high accuracy ≠ high disentanglement" a consistent, rankable phenomenon within this setting.
- **Repair.** We then ask whether this entanglement can be repaired without fine-tuning the encoder. Cross-Nullspace Projection (CNP) is a closed-form, supervised post-processing step with two parts—a factor-specific LDA readout, and a bidirectional cross-projection that makes the two readouts approximately orthogonal. The readout accounts for most of the accuracy gain; the projection is what makes the two factor vectors independently usable, which factorized compositional retrieval requires. That a linear operator suffices indicates that, in the benchmarks studied, the entanglement is substantially linear-remediable.

2 Related Work

2.1 Frozen Encoders and Style-Statistics Representations

CSD (Somepalli et al., 2024) fine-tunes CLIP for artistic style attribution; IntroStyle (Kumar et al., 2025) leverages channel statistics (μ/σ) of diffusion UNet intermediate features as a training-free style descriptor. Classical descriptors such as the Gram matrix (Gatys et al., 2016) and AdaIN (Huang & Belongie, 2017) characterize style through second-order feature statistics, while general-purpose encoders CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024) provide strong discriminative features but do not explicitly separate factors. StyleDecoupler (Jia et al., 2026) applies orthogonal projection to extract pure style vectors, but relies on large-scale image-text distillation and yields only a unidirectional style vector. Closest to our setting, Ngweta et al. (2023) recover separate content and style representations from frozen visual features by linear means, but their construction is unsupervised and transformation-driven, whereas CNP is label-supervised and needs no paired transformations (Appendix G). CNP differs from both: it requires no encoder fine-tuning, is bidirectional, and keeps content and style rather than discarding one.

2.2 Generative Content-Style Decomposition

Training-based disentanglement methods (DICE (Zhang et al., 2026), SCAdapter (Trinh et al., 2026), Chameleon (Ko et al., 2026), DRG-Font (Chakraborty et al., 2026), SplitFlux (Yang et al., 2026), CSD-VAR (Nguyen et al., 2025), USO (Wu et al., 2025), HyGDL (Fu et al., 2025)) advance style-content separation for generation and transfer, yet they are assessed by generation quality (FID, CLIP score) or task-specific accuracy rather than by the disentanglement quality of the representations themselves.

2.3 Disentanglement Metrics

Classical disentanglement metrics—DCI (Eastwood & Williams, 2018), MIG (Chen et al., 2018), SAP (Kumar et al., 2018)—target axis-aligned latent variables in generative models (β-VAE (Higgins et al., 2017), FactorVAE (Kim & Mnih, 2018)). Their governing assumption—that each generative factor aligns with a single latent dimension—breaks down on frozen encoder representations for three reasons: (i) content and style occupy multi-dimensional subspaces rather than individual axes; (ii) the metrics presume known, independently sampled generative factors, whereas encoder features are evaluated against real-image category labels; and (iii) per-dimension mutual-information estimation on 768–1024-dimensional embeddings is high-variance. Prior analyses (Mahon et al., 2023; Julka et al., 2024) likewise report systematic metric bias when axis-alignment is violated. Our kNN and probe leakage instead operate at the subspace level and are tailored to the "frozen encoder + category label" regime; benchmarks such as FED-Bench (Xue et al., 2026; Barami et al., 2025) evaluate generative latent spaces whose factors are sampled independently by construction rather than read off real images.

2.4 Linear Concept Erasure

INLP (Ravfogel et al., 2020) iteratively trains linear classifiers and projects weight directions onto their nullspace to erase a single sensitive attribute from text representations; Ravfogel et al. (2022) generalize this to a linear-adversarial formulation; LEACE (Belrose et al., 2023) derives the closed-form optimal least-squares erasure operator; and InstantForget (Yu, 2026) employs subspace projection for inference-time feature forgetting. A recent survey (Xie et al., 2025) identifies the simultaneous retention of two mutually entangled concepts as an open challenge.

These methods originate in NLP fairness and target the unidirectional removal of a single sensitive attribute (e.g., gender or race), and have not been systematically transferred to visual content-style separation. Three differences explain the gap: (i) fairness requires only one-way erasure, whereas content-style demands bidirectional separation *and* retention; (ii) NLP attributes are typically binary, whereas visual content and style span multi-class, subspace-level distributions; and (iii) fairness prioritizes preserving downstream accuracy, whereas disentanglement additionally requires quantifying that preservation. CNP inherits the nullspace-projection intuition but shifts the objective from "unidirectionally erase and discard" to "bidirectionally separate and retain."

3 METHOD

We first define a metric suite that makes content-style disentanglement measurable in frozen encoders, and then introduce Cross-Nullspace Projection (CNP), a closed-form operator that repairs the entanglement these metrics expose without fine-tuning the encoder. Throughout, $f : \mathcal{I} \to \mathbb{R}^D$ denotes a frozen encoder, and every image carries two labels—a content label y_c and a style label y_s.

3.1 SUBSPACE METRICS: KNN AND PROBE LEAKAGE

For each image, an encoder together with a disentanglement operator yields a content vector z_c and a style vector $z_s \in \mathbb{R}^d$ (for a raw single-vector representation, $z_c = z_s = x$; a decoupled pair instead lives in a reduced space of dimension $d \leq D$). We probe disentanglement from two complementary levels, so that no single blind spot can inflate the verdict: **kNN leakage** (neighbor level) asks whether a sample's content-nearest neighbors share its style label; and **probe leakage** (information level) asks whether style labels are linearly predictable from content vectors. A third, geometric quantity—the overlap between the two factors' LDA subspaces, $\rho = \|W_c^\top W_s\|_2$—is used alongside kNN leakage to form the DScore purity term (Sec. 3.2).

kNN leakage. Motivated by the KSG nonparametric estimator (Kraskov et al., 2004), we measure the deviation of the rate at which the content-retrieval top-K neighbors of sample i share its style label from the marginal $P(y_s)$:

$$L_{\text{knn}} = \left| \frac{1}{K} \sum_{j \in \text{kNN}_c(i)} \mathbb{1}[y_s^j = y_s^i] - P(y_s) \right|. \tag{1}$$

A value near zero (chance level) indicates that the content vector carries no residual style information. Here $\text{kNN}_c(i)$ is the set of K nearest neighbors of i in the content space under cosine similarity, excluding i itself; $P(y_s)$ is the global style marginal of the evaluation split, evaluated at y_s^i; and the quantity is averaged over queries. Appendix H gives the full protocol for every metric we report.

Probe leakage. The most direct test asks whether one factor is linearly decodable from the other. We fit a logistic-regression probe in each direction and read off its accuracy:

$$\begin{aligned} L_{\text{probe}}^{c \to s} &= \text{Acc}(\text{Probe}(z_c) \to y_s), \\ L_{\text{probe}}^{s \to c} &= \text{Acc}(\text{Probe}(z_s) \to y_c), \end{aligned} \tag{2}$$

and summarize the two directions by their mean, $L_{\text{probe}} = \frac{1}{2}(L_{\text{probe}}^{c \to s} + L_{\text{probe}}^{s \to c})$. Lower is better: a fully disentangled pair drives both probes to chance.

3.2 DSCORE: A JOINT ACCURACY-PURITY SCORE

Retrieval accuracy and disentanglement purity are individually insufficient—the former ignores leakage, the latter ignores usability. We combine them into a single rankable scalar,

$$\text{DScore} = \underbrace{\sqrt{\text{mAP}_c \cdot \text{mAP}_s}}_{\text{retrieval utility}} \times \underbrace{\big(1 - \text{leak}\big)}_{\text{disentanglement purity}}, \tag{3}$$

where the leakage term aggregates two complementary, scale-free measures of how confusable the two factors remain,

$$\text{leak} = \tfrac{1}{2}\big(L_{\text{sub}} + L_{\text{nn}}\big). \tag{4}$$

Here L_{nn} is the kNN leakage proxy, and L_{sub} is a subspace-level leakage term: for a raw single-vector encoder we use the overlap $\rho = \|W_c^\top W_s\|_2 \in [0, 1]$ between the two LDA factor subspaces, whereas for a decoupled pair (z_c, z_s)—where probe leakage provides a more direct measure of cross-factor predictability—we substitute the probe-based leakage. Both instantiations answer the same question: how much do the two factors remain linearly interchangeable. The switch is deliberate: a raw single-vector encoder makes self-probing meaningless, so the geometric ρ is the appropriate alternative; once decoupled, probe leakage becomes well-defined and more direct. We write DScore_ρ and $\text{DScore}_{\text{probe}}$ for the two instantiations. They are not on a common scale—ρ has a true zero, probe

leakage a positive chance floor—so the raw-to-decoupled ΔDScore that our post-processing tables report mixes both definitions: it is a within-row change, not a calibrated effect size (Appendix F).

Design rationale. DScore is high only when retrieval accuracy is high *and* both leakage terms are low. The geometric mean penalizes imbalance—content mAP $= 0.9$ with style mAP $= 0.1$ yields 0.30, versus 0.50 under the arithmetic mean—so a representation cannot compensate for one factor by over-performing on the other. Averaging a geometric term (L_{sub}) with an information-theoretic one (L_{nn}) makes the purity robust to the blind spot of either measure alone, and the linear discount $(1 - \text{leak})$ is monotonic and introduces no tuned weight of its own, beyond the settings of Sec. 4.1. A consistency analysis (DScore attains its maximum only when both factors are perfectly retrieved and fully separated) is given in Appendix B.

3.3 Cross-Nullspace Projection (CNP)

Given the pseudo-disentanglement exposed by our metrics, we ask whether the entanglement in a frozen encoder can be repaired *without fine-tuning the encoder*. CNP answers affirmatively. It is supervised linear post-processing—fitted on labeled estimation data, and never updating the encoder—and it has two components: a factor-specific LDA readout, and a bidirectional cross-projection that makes the two readouts approximately orthogonal (Figure 1). Sec. 4.5 shows that the readout accounts for most of the accuracy gain, while the projection is what makes the factors independently usable. The guiding intuition is symmetry: projecting a content vector away from the style-discriminative subspace suppresses its style leakage, and symmetrically, projecting a style vector away from the content-discriminative subspace suppresses its content leakage. The challenge is to perform both *jointly*, so each factor is cleaned of the other while being retained.

Step 1: Discriminative subspace estimation. On the labeled estimation split we fit one linear discriminant analysis (LDA) per factor: with $S_W^{(f)}$ and $S_B^{(f)}$ the within- and between-class scatter matrices of factor f over the estimation samples, we keep the leading generalized eigenvectors

$$\begin{aligned} W_c &= \text{top-}k_c \text{ eigvec}\big((S_W^{(c)})^{-1} S_B^{(c)}\big), \\ W_s &= \text{top-}k_s \text{ eigvec}\big((S_W^{(s)})^{-1} S_B^{(s)}\big). \end{aligned} \tag{5}$$

The bases are Euclidean-orthonormalized and truncated to $k_f = \min(k,\, C_f - 1)$, the at most $C_f - 1$ discriminative directions a C_f-class LDA admits; with $k = 64$ the effective ranks k_c/k_s are $64/5$, $64/3$ and $59/5$ on DomainNet, OfficeHome and NICO++. Appendix E gives the regularization and rank-handling details.

Step 2: Cross-nullspace projection. Each factor's basis is projected onto the (approximate) nullspace of the other, with strength $\lambda \in [0, 1]$:

$$\begin{aligned} \tilde{W}_c &= (I - \lambda W_s W_s^\top)\, W_c, \\ \tilde{W}_s &= (I - \lambda W_c W_c^\top)\, W_s. \end{aligned} \tag{6}$$

Step 3: Factor readout. The disentangled vectors are the normalized projections $z_c = \text{normalize}(\tilde{W}_c^\top x)$ and $z_s = \text{normalize}(\tilde{W}_s^\top x)$.

Why it works. The operator $I - \lambda W_s W_s^\top$ scales any component lying in the style subspace by $(1 - \lambda)$, so the style energy carried by a content vector shrinks to a factor $(1 - \lambda)^2$. Meanwhile, for content-dominated vectors, the content energy is provably retained by at least $(1 - \lambda\rho)^2$, where $\rho = \|W_c^\top W_s\|_2 \in [0, 1]$ is the overlap between the two discriminative subspaces. Because the factors are largely non-collinear in practice ($\rho \ll 1$, typically < 0.3, with CSD and CLIP as mild exceptions at $\approx$0.40), the content loss is negligible while the style leakage is strongly attenuated. Both bounds are proved in Appendix B.

Distinction from unidirectional erasure. INLP (Ravfogel et al., 2020) iteratively erases a single attribute and *discards* it; LEACE (Belrose et al., 2023) gives the optimal least-squares erasure of one attribute. Both are one-way: they remove information rather than reorganize it. CNP instead *separates and retains both factors simultaneously*, producing two approximately orthogonal, independently usable vectors—precisely what enables factorized compositional retrieval, which single-projection baselines cannot provide. At $\lambda = 0$ the operator reduces to a factor-wise LDA readout (Sec. 4.5), and we set $\lambda = 0.7$ throughout (Sec. 4.8).

4 EXPERIMENTS

4.1 SETUP

Datasets. We evaluate on DomainNet (Peng et al., 2019) (345×6, $\sim$170K images), Office-Home (Venkateswara et al., 2017) (65×4, $\sim$15K), and NICO++ (Zhang et al., 2023) (60×6, $\sim$90K), reported as categories $\times$ domains/contexts. The object category is content; the domain or context is style.

Representations (9, by paradigm). *Semantic*: CLIP ViT-L/14 (Radford et al., 2021), SigLIP2 (Tschannen et al., 2025), Qwen2.5-VL (Bai et al., 2025). *Self-supervised*: DINOv2 ViT-L (Oquab et al., 2024), DINOv3 ViT-L (Siméoni et al., 2025). *Style/statistics*: CSD (Somepalli et al., 2024), IntroStyle (Kumar et al., 2025), Gram (Gatys et al., 2016). *Supervised baseline*: ResNet-50. These span the three paradigms of off-the-shelf visual backbones.

Metrics. Content/style mAP@10, kNN leakage ($K{=}10$), probe leakage, DScore, and compositional retrieval mAP@10, under a $70/30$ estimation/evaluation split averaged over 5 seeds. No network is trained; CNP fits only the linear subspaces of Sec. 3.3. Results are means over five seeds without dispersion, so smaller margins are indicative rather than significant (Appendix G).

4.2 LEADERBOARD AND RANK DIVERGENCE

Table 1: Content-Style Disentanglement Leaderboard on DomainNet, ranked by DScore.

Rank	Representation	C-mAP↑	S-mAP↑	Probe Leak↓	kNN Leak↓	**DScore↑**
1	DINOv2	0.687	0.789	0.776	**0.434**	**0.506**
2	DINOv3	**0.725**	0.796	0.815	0.435	0.483
3	Qwen2.5-VL	0.616	**0.890**	0.799	0.645	0.453
4	SigLIP2	0.716	0.878	0.855	0.591	0.446
5	CSD	0.668	0.841	0.801	0.518	0.404
6	CLIP	0.666	0.841	0.797	0.522	0.401
7	ResNet-50	0.443	0.822	0.653	0.550	0.400
8	IntroStyle	0.210	0.830	0.478	0.562	0.269
9	Gram	0.212	0.785	**0.468**	0.542	0.239

Subspace overlap ρ (row-wise): 0.192, 0.294, 0.132, 0.284, 0.404, 0.406, 0.122, 0.150, 0.284. DScore purity = $1 - (\rho + \text{kNN})/2$.

The leaderboard exposes systematic patterns tied to training paradigms. Semantic-alignment encoders (SigLIP2, Qwen2.5-VL, CLIP) attain the strongest style retrieval (mAP 0.841–0.890) but carry the highest neighborhood contamination—Qwen2.5-VL and SigLIP2 top the kNN leakage at 0.645 and 0.591—meaning their content neighborhoods are heavily populated by style-matched samples. Self-supervised encoders (DINOv2, DINOv3) instead pair balanced content/style mAP with the lowest kNN leakage (0.434/0.435), which is what lifts them to the top two DScore positions. Style-specialized descriptors (Gram, IntroStyle) achieve the lowest probe leakage (≈ 0.47) yet are content-blind (content mAP ≈ 0.21)—purity without usability, which DScore duly penalizes through the geometric-mean term. The consequence is a concrete rank reversal (Figure 2): Qwen2.5-VL, the top style retriever, falls to 3rd under DScore, whereas DINOv2—only 8th by style mAP—rises to 1st once purity is accounted for. An accuracy-only protocol therefore promotes an encoder whose factors remain entangled and overlooks the one that disentangles best.

4.3 CNP POST-PROCESSING EFFECTS

CNP improves DScore on all nine representations (mean $+0.15$): probe leakage drops by 60–86% across the board and the kNN leakage decreases for every encoder, confirming a genuine reduction in cross-factor dependence rather than a metric artifact, while style mAP is maintained or improved for all but one encoder. Two rank changes are telling. SigLIP2 rises from 4th to 1st ($+0.222$): the most entangled encoder becomes the strongest once its factors are separated—direct evidence that raw rankings misjudge disentanglement potential. DINOv3 overtakes DINOv2 ($+0.167$ vs. $+0.090$), as its stronger discriminative power surfaces once entanglement is removed. Gains are not uniform:

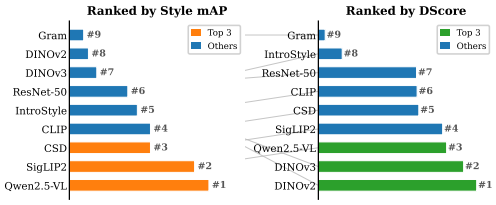

Figure 2: Rank reversal between style-mAP and DScore orderings. Encoders that rank highest by style retrieval (e.g., Qwen2.5-VL) drop under DScore once cross-factor leakage is penalized, while balanced self-supervised encoders (DINOv2) rise to the top.

Table 2: CNP Post-Processing (DomainNet, $\lambda = 0.7$, $k = 64$, ranked by DScore descending)

Representation	C-mAP↑	S-mAP↑	Probe Leak↓	kNN Leak↓	**DScore↑**	ΔDS
SigLIP2	**0.740**	**0.904**	0.147	0.221	**0.668**	**+0.222**
DINOv3	0.677	0.857	**0.112**	**0.182**	0.650	+0.167
Qwen2.5-VL	0.657	0.901	0.130	0.189	0.647	+0.194
CSD	0.638	0.852	0.139	0.209	0.609	+0.205
CLIP	0.640	0.857	0.153	0.224	0.601	+0.200
DINOv2	0.599	0.818	0.116	**0.182**	0.596	+0.090
ResNet-50	0.323	0.816	0.141	0.213	0.423	+0.023
IntroStyle	0.325	0.862	0.189	0.271	0.408	+0.139
Gram	0.254	0.794	0.149	0.312	0.346	+0.107

ResNet-50 improves by only $+0.023$, reflecting the limited factor structure of a purely supervised CNN, whereas IntroStyle gains $+0.139$, showing that CNP can recover content-discriminative directions even from a style-specialized descriptor.

4.4 Metric Validity: Comparison with Classical Metrics

Table 3: Our Metrics vs. Classical Disentanglement Metrics (DomainNet, CLIP base, $\lambda = 0.7$, $k = 64$)

Representation	DCI↑	MIG↑	SAP↑	**Probe Leak↓**	**DScore↑**
CLIP (entangled)	0.371	0.007	0.004	0.797	0.401
IntroStyle (more disentangled)	0.697	0.002	0.004	0.478	0.269
CLIP+CNP (enhanced)	**0.729**	**0.010**	**0.015**	**0.153**	**0.601**
Dynamic Range (max−min)	0.358	0.008	0.011	0.644	0.332

Table 3 contrasts our metrics with classical ones. The classical metrics target a different object: DCI, MIG and SAP measure axis alignment between generative factors and individual latent coordinates, whereas our leakage terms measure cross-factor information between two representation subspaces. A rotation within an already-separated subspace can move DCI without changing content-style separation, so we read this table as a diagnostic sanity check rather than a like-for-like contest. On frozen encoders SAP and MIG are nearly flat (SAP spread 0.011; all MIG ≤ 0.010, against a typical VAE range of 0.1–0.5), consistent with their axis-aligned assumption failing once the factors occupy subspaces. DCI is more discriminative (0.358) but ranks IntroStyle (0.697) well above CLIP (0.371) despite IntroStyle's content mAP of 0.21: it rewards the per-dimension independence that channel statistics (μ/σ) satisfy by construction, regardless of retrieval utility. DScore instead couples usability with purity, demoting IntroStyle to last (0.269) and crediting the CLIP→CLIP+CNP repair ($0.401 \rightarrow 0.601$). Appendix H specifies how each of the classical metrics was computed.

Table 4: Comparison with Post-Processing Methods (DomainNet, CLIP base, $k = 64$, $\lambda = 0.7$). "—" = method does not produce independently usable factor vectors.

Post-Processing	C-mAP↑	S-mAP↑	Probe Leak↓	DScore↑	Comp. mAP↑
Raw	0.666	0.841	0.797	0.401	0.186
PCA Whitening	**0.681**	0.821	0.794	0.380	—
INLP (Ravfogel et al., 2020)	0.675	0.790	0.356	0.465	0.097
LDA Single Proj.	0.638	**0.873**	0.182	0.563	—
CNP (Ours)	0.640	0.857	**0.153**	**0.601**	**0.451**

4.5 Post-Processing Comparison

Table 4 compares CNP against the raw features and three post-processing baselines—PCA whitening, INLP, and LDA single-projection—under identical settings. PCA whitening even lowers DScore (-0.021): global decorrelation cannot isolate factor-specific directions. INLP reduces leakage to 0.356, but its style mAP drops ($0.841 \rightarrow 0.790$) and its compositional retrieval (0.097) falls *below* the raw baseline—the price of unidirectional erasure, which sacrifices the very factor it removes. The most instructive comparison is LDA single-projection (i.e., projecting the content vector onto the style nullspace only, without the reverse de-content step): its DScore (0.563) falls below CNP (0.601), and its content and style subspaces remain heavily overlapping, so the two vectors are not independently usable. Only CNP produces approximately orthogonal factor vectors that are simultaneously accurate and independent—a property confirmed by its strong compositional retrieval performance below, where no other post-processing method is viable. One attribution caveat belongs here: at $\lambda = 0$ the projection is off and CNP reduces to a factor-wise LDA readout that alone reaches DScore 0.574 against 0.601 at $\lambda = 0.7$ (Appendix A), so most of the gain is supervised readout and the projection's contribution is the independence of the two readouts (Appendix G).

4.6 Factorized Compositional Retrieval

A direct application of two independent factor vectors is factorized compositional retrieval: given a query, retrieve images that share its content but differ in style (content-keep + style-swap), or the converse. The task *intrinsically* requires separation—an entangled embedding cannot satisfy "same content" and "different style" simultaneously. For CNP-decoupled z_c, z_s we use the composite score $\text{sim}_{\text{compose}} = \text{sim}_c - \alpha\,\text{sim}_s$ ($\alpha = 0.5$), and compare against raw cosine retrieval on CLIP and an INLP-based variant.

Table 5: Factorized Compositional Retrieval mAP@10 (DomainNet, CLIP base, $\lambda = 0.7$, $k = 64$)

Method	Content-Keep + Style-Swap↑	Style-Keep + Content-Swap↑
Raw (joint embedding)	0.186	0.396
INLP Compositional	0.097	—
CNP (Ours)	**0.451**	**0.842**

Raw features reach only 0.186 on "content-keep + style-swap": an entangled embedding cannot match content while differentiating style. INLP is even worse (0.097, below raw), because erasing the style direction also removes the content-discriminative signal it was entangled with. CNP attains 0.451 ($2.4\times$ raw, $4.6\times$ INLP) and 0.842 on the converse task. The advantage follows directly from independence: because z_c and z_s lie in approximately orthogonal subspaces, they can be recombined with independent weights—a degree of freedom that neither a joint embedding nor a unidirectional erasure provides. This is the capability that distinguishes CNP from the unidirectional LDA single-projection.

4.7 Qualitative Analysis: t-SNE Visualization

Figure 3 (Appendix D) visualizes the content and style vectors from CLIP features before and after CNP via t-SNE. Before CNP, content vectors exhibit domain-conditioned sub-clusters and style vectors show pronounced category leakage (probe leakage 0.797); after CNP, content organizes by

category and style by domain, corroborating the leakage reduction to 0.153 (Table 2). Instance-level retrieval examples are also given there: a content query retrieves across all six domains and a style query retrieves across diverse categories, confirming the aggregate trends at the instance level.

4.8 Additional Experiments

Two further experiments are reported in full in the appendix. **Hyperparameters** (Appendix A): we ablate λ and k on DomainNet with CLIP as the base encoder; DScore peaks at $\lambda = 0.7$ and plateaus beyond $k = 64$, so we use those values throughout. **Cross-dataset** (Appendix C): replicating the full evaluation on OfficeHome (Venkateswara et al., 2017) and NICO++ (Zhang et al., 2023) reproduces the central findings—the "accuracy $\neq$ disentanglement" pattern, the self-supervised $>$ semantic $>$ style-specialized tier ordering, and consistent CNP gains on all nine encoders.

5 Discussion and Conclusion

We set out to test an assumption that encoder selection takes for granted—that a representation which retrieves well also separates content from style—and found that it fails. **DScore**, with its **kNN and probe leakage** diagnostics, makes the failure measurable on frozen encoders, and across nine representations on DomainNet, OfficeHome, and NICO++, "high retrieval accuracy $\neq$ high disentanglement" holds throughout this setting: semantic-alignment objectives yield strong retrievers with heavy entanglement, whereas self-supervised objectives strike a markedly better balance. CNP consistently repairs this entanglement and, uniquely among the baselines, enables factorized compositional retrieval.

Limitations and future work. CNP is deliberately linear: it assumes content and style occupy approximately low-overlap linear subspaces, and its content-preservation guarantee degrades as the overlap ρ grows, so representations whose factors are genuinely nonlinearly entangled may require nonlinear extensions. Our protocol also assumes two discrete labeled factors per sample; extending it to more than two factors, to continuous attributes (e.g., brightness or rotation angle), and to hierarchical structures (e.g., artist $\rightarrow$ period $\rightarrow$ genre) are natural directions. Because CNP operates on frozen features, it cannot raise the intrinsic ceiling of a representation—as Gram's persistent last-place ranking shows; combining CNP-style nullspace objectives with representation learning (e.g., as a training regularizer) may instead yield natively disentangled features. Finally, our study is confined to the visual domain, and extending DScore and CNP to text, audio, and video is a promising next step.

Scope. Two qualifications bound our claims: our leakage terms measure label dependence through the representation and so bound cross-factor information from above unless the two labels are independent; and the margins are small and undispersed. The attribution of CNP's gain between readout and projection is discussed in Sec. 4.5; Appendix G details each caveat.

Broader impact. A standardized disentanglement benchmark can guide model selection for domain generalization, fair representation learning, and controllable generation. By revealing that widely used frozen representations exhibit substantial pseudo-disentanglement, we hope to encourage the community to look beyond single-metric accuracy toward more interpretable, factorized representations. Moreover, that a simple linear operator can substantially reduce entanglement suggests this separation problem may be more tractable than previously assumed.

REPRODUCIBILITY STATEMENT

This work uses only frozen, publicly available pretrained encoders and requires no training. Sec. 4.1 specifies the three benchmarks (DomainNet, OfficeHome, NICO++) with their category$\times$domain/context counts, the nine representations and their paradigms, the $70/30$ estimation/evaluation split averaged over 5 seeds, and all evaluation metrics (content/style mAP@10, kNN leakage with $K{=}10$, probe leakage, DScore, and compositional retrieval mAP@10). CNP has two hyperparameters, the projection strength λ and the subspace dimension k; Appendix A reports the full ablations and we use $\lambda = 0.7$ and $k = 64$ throughout. Complete proofs of the CNP de-styling guarantee and the content-preservation guarantee are given in Appendix B, together with the DScore bounds; the consistency property is stated there with a proof sketch. All three datasets are established public benchmarks for domain generalization.

ETHICS STATEMENT

This study analyzes representations produced by existing pretrained encoders on three publicly available domain-generalization benchmarks (DomainNet, OfficeHome, NICO++). It does not collect new human-subject data, does not involve human participants, does not release a new dataset, and does not deploy any model in a downstream application. The content and style labels we use are the standard category and domain/context annotations shipped with those benchmarks. Potential risks of controllable generation and of factorized retrieval are discussed under Broader Impact in Sec. 5.

AI USE STATEMENT

We used generative AI tools for literature retrieval and discovery—that is, identifying relevant related work and searching for information, both of which the ICLR 2027 AI policy lists as recommended-disclosure tasks. Every reference surfaced by these tools was located and checked against the published record by the authors before it was cited, and all AI-assisted search results were reviewed manually. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tal Barami, Nimrod Berman, Ilan Naiman, Amos H. Hason, Rotem Ezra, and Omri Azencot. Disentanglement beyond static vs. dynamic: A benchmark and evaluation framework for multi-factor sequential representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 66044–66063, 2023.

Rejoy Chakraborty, Prasun Roy, Saumik Bhattacharya, and Umapada Pal. DRG-Font: Dynamic reference-guided few-shot font generation via contrastive style-content disentanglement. *arXiv preprint arXiv:2604.13797*, 2026.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Siming Fu, Sijun Dong, and Xiaoliang Meng. Disentangling content from style to overcome shortcut learning: A hybrid generative-discriminative learning framework. *arXiv preprint arXiv:2509.11598*, 2025.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017.

Zexi Jia, Jinchao Zhang, and Jie Zhou. StyleDecoupler: Generalizable artistic style disentanglement. *arXiv preprint arXiv:2601.17697*, 2026.

Sahib Julka, Yashu Wang, and Michael Granitzer. Towards an improved metric for evaluating disentangled representations. In *Proceedings of the IEEE International Conference on Knowledge Graph (ICKG)*, 2024.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Sukhun Ko, Soo Ye Kim, and Jihyong Oh. Chameleon: Style-content disentangled framework for cross-domain object compositing. *arXiv preprint arXiv:2606.01079*, 2026.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabelled observations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Anand Kumar, Jiteng Mu, and Nuno Vasconcelos. IntroStyle: Training-free introspective style attribution using diffusion features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14909–14918, 2025.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Louis Mahon, Lei Shah, and Thomas Lukasiewicz. Correcting flaws in common disentanglement metrics. *arXiv preprint arXiv:2304.02335*, 2023.

Quang-Binh Nguyen, Minh Luu, Quang Nguyen, Anh Tran, and Khoi Nguyen. CSD-VAR: Content-style decomposition in visual autoregressive models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Lilian Ngweta, Subha Maity, Alex Gittens, Yuekai Sun, and Mikhail Yurochkin. Simple disentanglement of style and content in visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pp. 26063–26086, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162, pp. 18400–18421, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

Luan Thanh Trinh, Kenji Doi, and Atsuki Osanai. SCAdapter: Content-style disentanglement for diffusion style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. USO: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025.

Yiwei Xie, Ping Liu, and Zheng Zhang. Erasing concepts, steering generations: A comprehensive survey of concept suppression. *arXiv preprint arXiv:2505.19398*, 2025.

Fengjian Xue, Xuecheng Wu, Heli Sun, Yunyun Shi, Shi Chen, Liangyu Fu, Jinheng Xie, Dingkang Yang, and Hao Wang. FED-Bench: A cross-granular benchmark for disentangled evaluation of facial expression editing. *arXiv preprint arXiv:2603.29697*, 2026.

Yitong Yang, Yinglin Wang, Changshuo Wang, Yongjun Zhang, Ziyang Chen, and Shuting He. SplitFlux: Learning to decouple content and style from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 572–582, 2026.

Zhenyu Yu. InstantForget: Update-free backdoor unlearning with inference-time feature reset. *arXiv preprint arXiv:2606.15730*, 2026.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Tong Zhang, Ru Zhang, and Jianyi Liu. DICE: Disentangling artist style from content via contrastive subspace decomposition in diffusion models. *arXiv preprint arXiv:2602.08059*, 2026.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. NICO++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

A Ablation Details

We ablate CNP's two hyperparameters (λ and k) on DomainNet with CLIP as the base encoder.

Projection strength λ. At $\lambda = 0$ no projection is applied and CNP reduces to a factor-wise LDA readout — the dual-LDA baseline that isolates supervised readout from cross-projection, and the control the reviewers of an earlier version asked for. DScore is non-monotonic in λ and peaks at $\lambda = 0.7$: it rises from 0.574 at $\lambda = 0$ through 0.595 at $\lambda = 0.5$ to 0.601 at $\lambda = 0.7$, then falls to 0.589 at $\lambda = 1.0$ — a 2.0% spread across $\lambda \in [0.5, 1.0]$. Probe leakage traces the same shape, reaching its minimum of 0.153 at $\lambda = 0.7$ before rising again to 0.166 at $\lambda = 1.0$.

Two readings of this curve matter. First, the trade-off is the expected one: a weak projection leaves style energy in the content vector, while a full projection ($\lambda = 1$) also removes content, so an intermediate strength is optimal. Second, and consequentially for attribution, most of the improvement over raw features is already present at $\lambda = 0$: the dual-LDA readout alone raises DScore from 0.401 to 0.574, so the cross-nullspace step contributes only the remaining 0.027 of the 0.200 total. We therefore do not claim that cross-projection drives the DScore gain; what it adds is the *independence* of the two readouts, which the dual-LDA readout does not provide and which is what makes factorized compositional retrieval possible (Appendix G). Because the two subspaces overlap only weakly to begin with, a mild projection ($\lambda = 0.5$) already recovers most of that benefit. We fix $\lambda = 0.7$ throughout.

Subspace dimension k. Content mAP rises from 0.564 at $k = 8$ to 0.640 at $k = 64$ and then plateaus; style mAP largely saturates earlier at $k = 32$, reflecting the narrower intrinsic dimensionality of 6 domains versus 345 categories. DScore plateaus beyond $k = 64$, so we adopt $k = 64$ as a balanced default. The saturation behavior shows that CNP does not demand precise dimensionality estimation: any reasonably large k suffices.

Table 6: Projection strength λ ablation (DomainNet, CLIP base, k=64). At λ=0 no projection is applied, giving the dual-LDA readout baseline.

λ	0.0	0.3	0.5	0.7	1.0
DScore↑	0.574	0.582	0.595	0.601	0.589
Probe Leak↓	0.198	0.175	0.161	0.153	0.166

Table 7: Subspace dimension k ablation (DomainNet, CLIP base, λ=0.7).

k	8	32	64	256
C-mAP↑	0.564	0.635	0.640	0.647
S-mAP↑	0.789	0.851	0.857	0.857

B Theoretical Results

B.1 CNP Guarantees

Proposition 1 (De-styling guarantee). Let $W_s \in \mathbb{R}^{D\times k}$ be an orthonormal basis for the style discriminative subspace. For any vector $v \in \mathbb{R}^D$, applying CNP projection $P_s = I - \lambda W_s W_s^\top$:

$$\|W_s^\top P_s v\|^2 = (1-\lambda)^2 \|W_s^\top v\|^2$$

Proof. $W_s^\top P_s = W_s^\top (I - \lambda W_s W_s^\top) = (1-\lambda) W_s^\top$, since $W_s^\top W_s = I_k$. Therefore $\|W_s^\top P_s v\|^2 = (1-\lambda)^2 \|W_s^\top v\|^2$.

Proposition 2 (Content preservation guarantee). Let $W_c \in \mathbb{R}^{D\times k}$ be orthonormal for the content subspace, and define the subspace overlap $\rho = \|W_c^\top W_s\|_2 \in [0, 1]$. For any content-dominated vector v (i.e., $\|W_s^\top v\| \le \|W_c^\top v\|$), applying P_s satisfies

$$\|W_c^\top P_s v\|^2 \ge (1 - \lambda\rho)^2 \|W_c^\top v\|^2.$$

Proof. $W_c^\top P_s = W_c^\top - \lambda W_c^\top W_s W_s^\top$. Let $M = W_c^\top W_s$. Then:

$$\begin{aligned}\|W_c^\top P_s v\| &= \|W_c^\top v - \lambda M W_s^\top v\| \\ &\geq \|W_c^\top v\| - \lambda \|M W_s^\top v\| \\ &\geq \|W_c^\top v\| - \lambda\rho \|W_s^\top v\|.\end{aligned}$$

(Step 2 uses the reverse triangle inequality $\|a-b\| \geq \|a\| - \|b\|$; step 3 uses the sub-multiplicativity of the spectral norm, $\|MW_s^\top v\| \leq \|M\|_2 \|W_s^\top v\|$, together with $\|M\|_2 = \rho$.) Invoking the content-dominated assumption $\|W_s^\top v\| \leq \|W_c^\top v\|$,

$$\|W_c^\top P_s v\| \geq (1 - \lambda\rho)\|W_c^\top v\|,$$

and squaring yields the stated bound.

Remark. In practice $\rho \ll 1$ (typically $<$0.3, with CSD and CLIP as mild exceptions at $\approx$0.40), so the content-preservation factor $(1-\lambda\rho)^2$ is close to 1. For $\lambda = 0.7$ and $\rho = 0.2$, the guaranteed content-energy retention is at least $(1 - 0.14)^2 \approx 74\%$, and is higher still in the common regime $\rho < 0.2$.

B.2 DScore Consistency

Proposition (consistency under factor independence). Let z_c, z_s be content and style vectors. If $I(z_c; y_s \mid y_c) = 0$ and $I(z_s; y_c \mid y_s) = 0$ (perfect disentanglement), the two leakage diagnostics fall to their floors—the kNN leakage to the label dependence $|P(y_s \mid y_c) - P(y_s)|$, and the probe leakage to the accuracy attainable from y_c alone. The subspace overlap ρ is geometric rather than distributional and is *not* implied by this condition; it is measured, not asserted, as discussed below.

Proof sketch. Under conditional independence the content neighbourhood of a sample carries no style information beyond y_c, so the same-style fraction among its neighbours matches $P(y_s \mid y_c)$ rather than the marginal $P(y_s)$; the residual kNN leakage is therefore the label dependence itself, and vanishes only when the two labels are independent, which our benchmarks do not guarantee (Appendix G). By the same argument, no linear classifier reading z_c can beat the accuracy attainable from y_c alone; that floor is below 0.14 on all three benchmarks and constant across the representations compared, so it does not affect relative rankings. **The geometric term is different in kind:** $\rho = \|W_c^\top W_s\|_2$ is computed from *fitted* bases, and label independence does not force two LDA subspaces to be orthogonal—each is built from its own class means—so their overlap is an empirical property that must be measured rather than assumed to vanish. DScore accordingly treats ρ (for raw encoders) or the probe leakage (for decoupled pairs) as a measured input. The converse of the proposition fails as well: these are sufficient, not necessary conditions, which is why DScore aggregates two complementary leakage terms rather than relying on any single one.

Proposition (DScore bounds). $0 \leq \text{DScore} < 1$. The subspace overlap ρ and the kNN leakage both have true zeros, so DScore approaches 1 as both mAPs approach 1 and both leakage terms approach zero. In the decoupled-vector case, L_{sub} is the probe leakage, whose minimum is the chance-level accuracy floor; this floor is below 0.14 in all our experiments and does not affect comparisons.

Proof. The geometric mean $\sqrt{\text{mAP}_c \cdot \text{mAP}_s} \in [0,1]$. For the raw-encoder case, $L_{\text{sub}} = \rho$ and L_{nn} both range in $[0,1]$ with attainable lower bounds of 0, so the purity $1 - \frac{1}{2}(\rho + L_{\text{nn}}) \in [0,1]$ and DScore can approach 1 arbitrarily closely. For decoupled pairs, L_{sub} is replaced by the probe leakage, whose floor is the chance accuracy; this introduces a small constant offset (< 0.07 in the purity term) that is identical across all compared representations.

C Cross-Dataset Validation

To test whether the observed patterns are dataset-specific or reflect intrinsic representation properties, we replicate the full evaluation on OfficeHome (Venkateswara et al., 2017) and NICO++ (Zhang et al., 2023). Tables 8–11 report the DScore leaderboards and CNP post-processing on both. Unlike Table 1, these tables do not list the raw-encoder subspace overlap ρ; since ρ enters the raw DScore by definition (Sec. 3.2), it is recoverable from the displayed cells as $\rho = 2\big(1 - \text{DScore}/\sqrt{\text{mAP}_c \cdot \text{mAP}_s}\big) - L_{\text{nn}}$, so the leaderboard entries remain mutually checkable at the reported precision.

OfficeHome. The "accuracy $\neq$ disentanglement" pattern replicates: strong retrievers carry the highest leakage (SigLIP2 probe leakage 0.948, Qwen2.5-VL 0.914), while style-specialized descriptors show low leakage but poor content mAP (Gram 0.474, IntroStyle 0.539). DINOv2 again claims the top rank (DScore 0.627; Table 8), with balanced factor mAP and the lowest kNN leakage among the leading encoders. The three paradigm tiers (self-supervised $>$ semantic $>$ style-specialized) are preserved.

NICO++. Here the factor definition shifts from "category-domain" to "object-context," testing DScore's generality under distribution shift. The ranking holds (Table 10)—DINOv2 (0.624), SigLIP2 (0.592), Qwen2.5-VL (0.591)—confirming that DScore captures an intrinsic property rather than a DomainNet artifact.

CNP transfer. CNP yields consistent DScore gains on all nine encoders across both datasets (Tables 9 and 11). On OfficeHome, gains span $+0.130$ (DINOv2) to $+0.288$ (IntroStyle), compressing leakage from 0.50–0.95 to 0.15–0.22, with Qwen2.5-VL reaching the best post-CNP DScore (0.844). On NICO++, gains span $+0.111$ to $+0.289$, with leakage compressed to 0.10–0.18. Gram remains last on both ($0.556/0.503$) despite gains, confirming that CNP repairs entanglement but cannot manufacture content structure that the encoder never captured. Across three datasets of differing scale, class count, and factor semantics, pseudo-disentanglement is present throughout and CNP repair transfers robustly, compressing probe leakage into the 0.10–0.22 band in every case.

Table 8: DScore Leaderboard on OfficeHome (65×4)

Rank	Representation	C-mAP↑	S-mAP↑	Probe Leak↓	kNN Leak↓	**DScore**↑
1	DINOv2	0.890	0.782	0.852	0.269	**0.627**
2	DINOv3	0.921	0.792	0.890	**0.267**	0.611
3	Qwen2.5-VL	0.877	0.893	0.914	0.499	0.609
4	ResNet-50	0.792	0.750	0.773	0.279	0.580
5	SigLIP2	**0.931**	**0.895**	0.948	0.478	0.575
6	CLIP	0.875	0.824	0.884	0.355	0.526
7	CSD	0.878	0.822	0.888	0.356	0.518
8	IntroStyle	0.539	0.787	0.552	0.393	0.476
9	Gram	0.474	0.714	**0.499**	0.270	0.425

Table 9: CNP Post-Processing on OfficeHome ($\lambda = 0.7$, $k = 64$, ranked by DScore descending)

Representation	C-mAP↑	S-mAP↑	Probe Leak↓	kNN Leak↓	**DScore**↑	ΔDS
Qwen2.5-VL	**0.980**	**0.977**	0.158	0.118	**0.844**	+0.235
SigLIP2	0.969	0.934	0.158	0.153	0.803	+0.228
DINOv3	0.962	0.866	**0.151**	0.125	0.787	+0.176
IntroStyle	0.931	0.922	0.208	0.143	0.764	**+0.288**
DINOv2	0.944	0.815	**0.151**	0.123	0.757	+0.130
CSD	0.928	0.857	0.187	0.136	0.748	+0.230
ResNet-50	0.928	0.805	0.174	**0.115**	0.740	+0.160
CLIP	0.921	0.834	0.197	0.150	0.725	+0.199
Gram	0.628	0.754	0.215	0.168	0.556	+0.131

D Qualitative Visualizations

Figures 3 and 4–5 give qualitative results for the CNP-decoupled CLIP representation on DomainNet: t-SNE embeddings of the content and style vectors before and after CNP, and instance-level retrieval examples. A content query retrieves across all six domains and a style query retrieves across diverse categories, confirming the aggregate trends at the instance level.

E LDA Estimation Details

Sec. 3.3 states the two LDA fits only in outline; we give the full specification here, since ρ, the attenuations and the CNP readouts all depend on it.

Table 10: DScore Leaderboard on NICO++ (60×6)

Rank	Representation	C-mAP↑	S-mAP↑	Probe Leak↓	kNN Leak↓	**DScore↑**
1	DINOv2	0.914	0.650	0.804	**0.179**	**0.624**
2	SigLIP2	**0.924**	0.717	0.864	0.270	0.592
3	Qwen2.5-VL	0.823	**0.726**	0.816	0.319	0.591
4	ResNet-50	0.823	0.647	0.765	0.239	0.584
5	DINOv3	0.904	0.672	0.827	0.213	0.578
6	CSD	0.902	0.699	0.824	0.253	0.528
7	CLIP	0.901	0.695	0.820	0.248	0.526
8	IntroStyle	0.450	0.595	0.439	0.212	0.420
9	Gram	0.411	0.571	**0.428**	0.235	0.357

Table 11: CNP Post-Processing on NICO++ ($\lambda = 0.7$, $k = 64$, ranked by DScore descending)

Representation	C-mAP↑	S-mAP↑	Probe Leak↓	kNN Leak↓	**DScore↑**	ΔDS
Qwen2.5-VL	**0.979**	**0.840**	**0.100**	**0.077**	**0.827**	+0.236
SigLIP2	0.967	0.749	0.103	0.110	0.760	+0.168
DINOv3	0.953	0.743	0.109	0.091	0.757	+0.179
ResNet-50	0.929	0.726	0.107	**0.077**	0.745	+0.161
DINOv2	0.952	0.692	0.101	0.088	0.735	+0.111
CSD	0.945	0.720	0.120	0.105	0.732	+0.204
CLIP	0.943	0.706	0.121	0.106	0.723	+0.197
IntroStyle	0.888	0.715	0.134	0.087	0.709	**+0.289**
Gram	0.597	0.595	0.183	0.131	0.503	+0.146

Scatter matrices. For factor $f \in \{c, s\}$ with C_f classes, class means μ_j and class sizes n_j, over the N samples of the labeled estimation split with mean $\bar{x}$:

$$S_W^{(f)} = \sum_{i=1}^{N} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^\top, \qquad S_B^{(f)} = \sum_{j=1}^{C_f} n_j (\mu_j - \bar{x})(\mu_j - \bar{x})^\top. \tag{7}$$

Regularization. $S_W^{(f)}$ is singular whenever D exceeds the pooled within-class degrees of freedom, which holds for every encoder we evaluate (D is 768–1024). We therefore solve against the ridge-regularized matrix

$$\tilde{S}_W^{(f)} = S_W^{(f)} + \tau \, \frac{\operatorname{tr}\big(S_W^{(f)}\big)}{D} \, I_D, \qquad \tau = 10^{-3}, \tag{8}$$

which leaves the leading discriminative directions essentially unchanged while making the solve well posed.

Orthonormalization. The generalized eigenvectors of $\big(\tilde{S}_W^{(f)}\big)^{-1} S_B^{(f)}$ are not mutually orthogonal in the Euclidean sense, but Eq. 6 treats W_f as an orthonormal projector. We therefore apply a QR factorization after truncation, so that $W_f^\top W_f = I_{k_f}$ exactly. Without this step the operators $I - \lambda W_f W_f^\top$ are not projectors and the attenuation statements of Appendix B do not apply.

Effective rank. A C_f-class LDA has at most $C_f - 1$ non-zero discriminative directions, because $S_B^{(f)}$ has rank at most $C_f - 1$. We set $k_f = \min(k, C_f - 1)$ accordingly; requesting more directions admits numerically arbitrary eigenvectors associated with near-zero generalized eigenvalues, which are not discriminative and inflate the measured overlap ρ. With $k = 64$ the effective ranks are:

Dataset	classes (content × style)	k_c	k_s
DomainNet	345×6	64	5
OfficeHome	65×4	64	3
NICO++	60×6	59	5

On OfficeHome content is not truncated ($C_c - 1 = 64 = k$); on NICO++ both factors are truncated. The style rank falls below k on all three datasets, so the style subspace is always the shorter of the

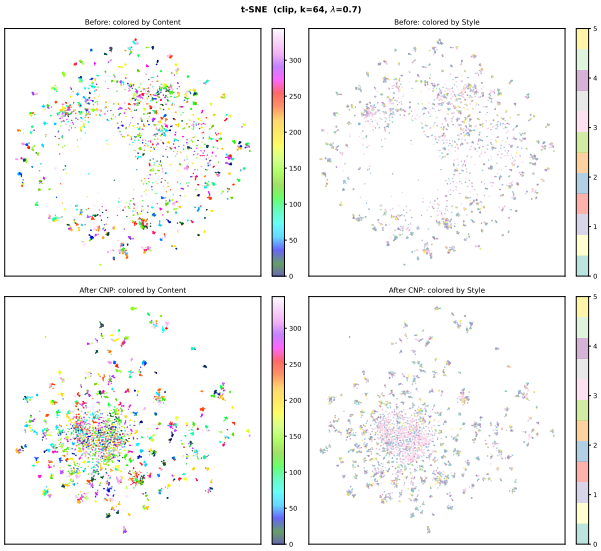

Figure 3: t-SNE visualization of content (left) and style (right) vectors before (top) and after (bottom) CNP. Content plot: colors = categories. Style plot: colors = domains.

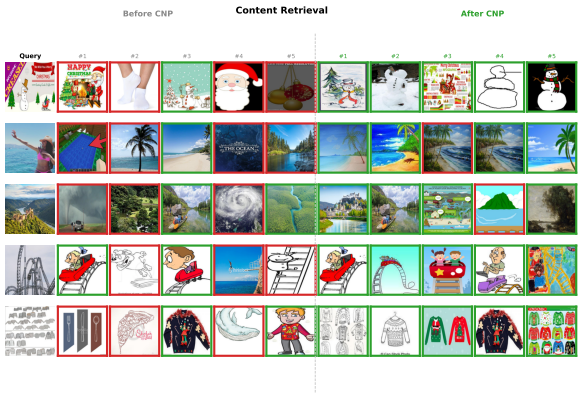

Figure 4: Content retrieval examples (DomainNet, CLIP base). Query (leftmost) and top-5 retrievals. CNP-processed vectors retrieve by category with substantially reduced domain interference.

two. The ρ values in Table 1 are computed from these truncated bases and would differ if k_s were left at 64.

F DSCORE SCALE AND THE PROBE-LEAKAGE FLOOR

$\rho = \|W_c^\top W_s\|_2$ has a true zero, attained when the two estimated subspaces are exactly orthogonal. Probe leakage is a classifier accuracy and therefore has a floor at the chance rate of the style label distribution, which is data-dependent and strictly positive. In our experiments that floor is below 0.14 on all three benchmarks, but it is not zero, so a decoupled pair cannot reach the same purity value as a perfectly disentangled raw encoder even under ideal separation.

It follows that a ΔDScore computed across the two instantiations—comparing a raw encoder to its own CNP-processed version—mixes a quantity with a true zero against one with a positive floor, and part of the apparent gain is the change of definition rather than a change in separation. The per-representation improvements in Table 2 and in the cross-dataset tables are of exactly this form. We report them because both the raw and the decoupled scores are internally consistent across the

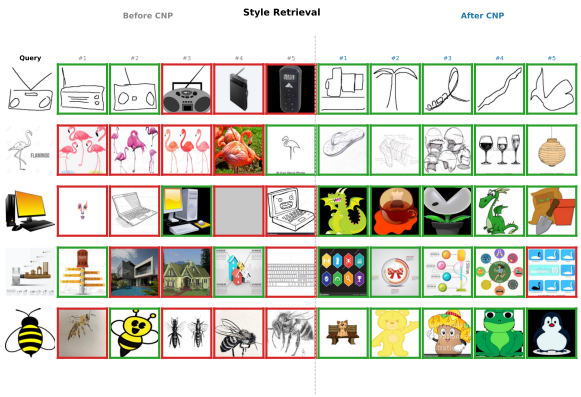

Figure 5: Style retrieval examples (DomainNet, CLIP base). Query (leftmost) and top-5 retrievals. CNP-processed vectors retrieve by domain without being distracted by content similarity.

nine representations, which is what the ranking comparison uses, but we do not read the raw-to-decoupled ΔDScore as a calibrated effect size. A single leakage definition valid for both regimes, and a calibrated comparison under it, remain open.

Scope of the consistency claim. The consistency proposition of Appendix B states *sufficient* conditions: it shows that perfect conditional independence drives the two leakage diagnostics to their floors. The geometric term is not covered by it: statistical independence of the two labels is *not* equivalent to Euclidean orthogonality of the estimated LDA subspaces, since ρ is a property of the fitted bases rather than of the underlying distributions. The proposition should therefore be read as describing an idealised regime in which the diagnostics agree, not as a guarantee that a small ρ certifies independence.

G Additional Limitations

Attribution between readout and projection. At $\lambda = 0$ Eq. 6 leaves the factor basis unchanged and CNP reduces to a factor-wise LDA readout fitted on the same labeled estimation split. That is the dual-LDA baseline, and Table 6 includes it, so the two contributions can be separated rather than merely acknowledged. The readout alone raises DScore from 0.401 (raw) to 0.574, against 0.601 at $\lambda = 0.7$: of the 0.200 total gain, 0.173 (86%) is attributable to supervised factor-specific readout and only 0.027 (14%) to the cross-nullspace step. We therefore make no claim that cross-projection drives the DScore improvement. What the projection adds is a property the readout does not supply: W_c and W_s are estimated independently, so the $\lambda = 0$ vectors are not mutually orthogonal and are not separately usable, whereas the cross-projected pair is — which is what the compositional-retrieval result depends on. One gap remains: Table 4 still has no full dual-LDA row, because that configuration is scored for DScore and probe leakage only, in Table 6, and not for content or style mAP or compositional retrieval.

Label dependence versus representational leakage. Eq. 1 compares neighbor agreement against the marginal $P(y_s)$ and Eq. 2 measures cross-label prediction accuracy. Both quantify dependence between the content and style *labels* as mediated by the representation. When y_c and y_s are correlated in the data—which they are in every benchmark we use, since domains and categories co-occur non-uniformly—a content vector carrying no style-specific information can still predict y_s through that correlation. A positive leakage value is therefore an upper bound on genuinely extra cross-factor information, and the appropriate null is not the marginal $P(y_s)$ but a label-only oracle. We do not report label-permutation, class-balanced or oracle controls, so the absolute leakage values should be read with that caveat; the ranking comparison is less affected, since the same bias applies to every representation scored on a given benchmark.

Uncertainty. All results are means over five seeds and we report no dispersion. Several adjacent pairs are close enough that this matters: the gap between the top two DScore entries in Table 2 is 0.003, and the leading margin in Table 1 is 0.023. Without paired variability estimates we cannot

establish that these orderings are stable, and we therefore present them as indicative of the reported setting rather than as a settled ranking.

Scope of the empirical claims. Three datasets with nine encoders support an observation about those datasets and encoders. In addition, both factors are proxied by dataset annotations—the object category stands in for content and the domain or context for style—so the results speak to that operationalisation rather than to artistic, authorial or texture style. We do not claim that pseudo-disentanglement is pervasive in general, nor that the underlying entanglement is linear-remediable in general, only that both hold throughout this setting.

Related post-hoc linear separation. Ngweta et al. (2023) also recover separate content and style representations from frozen visual features by linear means. The methodological difference is supervision: their construction is unsupervised and driven by controlled transformations, whereas CNP fits factor-specific discriminative subspaces on labeled data and needs no paired transformations. We do not claim that post-hoc linear factor separation is unaddressed in general, only that the specific combination of bidirectional retention, approximate orthogonality of the two readouts, and the absence of encoder fine-tuning distinguishes CNP within the setting we evaluate.

Computational cost. We report no wall-clock time or peak memory. CNP adds one pass over the labeled estimation split to accumulate the two scatter matrices, then two $D \times D$ generalized eigenproblems and two QR factorizations; it involves no gradient computation and no encoder update. We do not provide measured timings, so the "lightweight" characterisation should be read as an architectural statement about what the method computes rather than as a measured efficiency claim.

H Evaluation Protocol

We specify every metric reported, so that the numbers can be reproduced without the code.

Retrieval mAP. Features are L2-normalized and ranked by cosine similarity. Each query is scored against the full gallery with the query itself excluded (leave-one-out). For tractability we subsample at most 2000 queries per split, seeded; the gallery is never subsampled. mAP@K averages, over queries, the average precision of same-label entries among the top K retrievals.

Probe leakage. A linear classifier is fit on a $70/30$ split and scored on the held-out 30%. The split is stratified by the target label whenever the class count does not exceed half the sample count, and unstratified otherwise. We use logistic regression (max. 500 iterations) for at most 20000 samples and 50 classes, and an SGD logistic classifier (max. 1000 iterations) beyond that. L_{probe} is the mean of the two directions. Reported values are accuracies, so the chance floor is $1/C_s$ for style and $1/C_c$ for content rather than zero.

kNN leakage. Content vectors are L2-normalized; for each query we take its K nearest neighbors under cosine similarity, excluding itself, and measure the absolute deviation of the same-style fraction among them from the global style marginal evaluated at the query's own style label. The reported value is the mean deviation over queries.

Compositional retrieval. The task is: given a query, retrieve images sharing its content label but with a *different* style label. No paired examples are constructed—the relevant set for a query is defined by the label predicate $(y_c = y_c^{\text{query}}) \wedge (y_s \neq y_s^{\text{query}})$ over the gallery, and average precision is computed against it as above. At most 1000 queries are sampled, seeded, with the query excluded from its own candidate list. For decoupled representations the ranking score is $\text{sim}_c - \alpha\,\text{sim}_s$ with $\alpha = 0.5$ fixed, not tuned; the reverse task (same style, different content) uses $\text{sim}_s - \alpha\,\text{sim}_c$. We flag an asymmetry we do not correct: the raw baseline ranks by a single joint cosine similarity, because a joint embedding provides no separate factors to reweight, whereas the CNP and INLP rows do subtract a style term. The raw number should therefore be read as the no-factorization reference rather than a tuned competitor, and the content-only row of Table 5 is the upper bound for the content component alone.

Classical disentanglement metrics. We record the configuration behind Table 3, since these metrics are sensitive to it. DCI is the weighted per-dimension disentanglement score: a random forest (50 trees, max. depth 8) is fit per factor, per-dimension importances are normalized across the two factors, and each dimension contributes $1 - H$ (normalized importance entropy) weighted by its total importance. SAP fits a single-dimension logistic classifier per factor and takes the accuracy

gap between the best and second-best dimension. MIG estimates per-dimension mutual information from a 20-bin discretization. Two properties of this configuration bear on how the comparison should be read: the metrics are computed on the same split used to estimate the factors, without a held-out partition, and the CLIP+CNP row is computed on the concatenation $[z_c; z_s]$ rather than on a joint embedding. Both follow from these metrics assuming a single latent vector, and both are further reasons we treat the table as a diagnostic sanity check rather than a like-for-like contest.