The Retrieval–Grounding Tradeoff in Cross-Modal Connectors, and How to Navigate It

Anonymous authors
Paper under double-blind review

Abstract

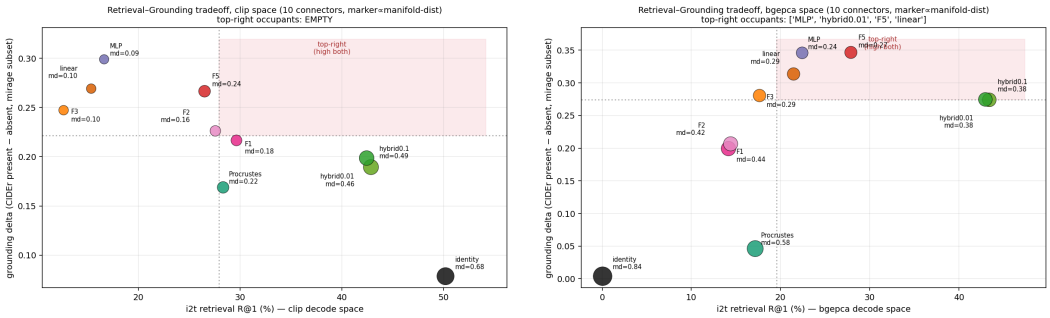
We ask whether a topology-preserving, invertible flow connector traces a better alignment-vs-structure frontier than projection heads for cross-modal (image↔text) translation, under the Aristotelian Representation Hypothesis (Gröger et al., 2026) that neural representations share local-neighborhood but not global-metric structure. Across ten connectors and two pre-registered testbeds—within-CLIP (pre-aligned) and cross-model (independent encoders)—the flow does not win: transport flows land off the target manifold; a guided conditional-generation flow buys single-shot retrieval but not manifold fidelity; and a one-line InfoNCE term on the same MLP beats every flow (42.9 vs. 27.9) at a fraction of the cost. Our central result is causal: a connector's grounding (how decodably image-specific its output is) is driven by proximity to the target manifold, which we establish with a projection/shuffle/intervention triad—so optimizing a connector for retrieval pushes its outputs off the manifold, a marginal manifold cost of discriminative pressure that is a hard retrieval–grounding exclusion when the spaces are pre-aligned and softens to a frontier otherwise. The mechanism is prescriptive: a train-time nearest-real-target pull on a contrastive objective navigates the tradeoff, filling the previously-empty top-right and Pareto-dominating every learned connector in both regimes (+11 R@1 over regression at equal grounding), and—the key control—beating post-hoc DeCap-style and subspace projections that target the same proximity. We scope all claims and release full pre-registration, code, and per-seed results.

1 Introduction

Cross-modal "connectors" that map one frozen encoder's embeddings into another's underlie retrieval, captioning, and multimodal LLMs. A natural hypothesis, motivated by the Aristotelian Representation Hypothesis (Gröger et al., 2026) and the diffeomorphism structure of flow matching (Lipman et al., 2023; Liu et al., 2023), is that an invertible, topology-preserving flow should trace a better alignment-vs-structure frontier than an unconstrained projection head, while providing invertibility as a free grounding certificate. We test this hypothesis adversarially, with kill conditions frozen before results.

Contributions. (i) A causal account of grounding: it is driven by proximity to the target manifold, which we establish with a projection/shuffle/intervention triad (§8)—so discriminative pressure carries a marginal manifold cost, a clean retrieval–grounding exclusion in the pre-aligned regime (Fig. 1) that softens to a frontier when the spaces are unaligned. (ii) The prescriptive fix, F6: a train-time nearest-real-target pull on a contrastive objective navigates the tradeoff, fills the previously-empty top-right, and—the key control—beats post-hoc DeCap-style and subspace projections that target the same proximity (§9). (iii) A pre-registered exclusion methodology and the negative results it yields: flows do not beat regression on the frontier, grounding, or selection, and a one-line contrastive term beats them all on retrieval. (iv) A positive: a single direction-conditioned flow field amortizes i2t/t2i (and captioning) tasks with flow-specific positive transfer. (v) A solvable model abstracting these findings: grounding follows an anisotropic decoder-metric distance (validated at

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.



(a) within-CLIP: top-right empty.

(b) cross-model: frontier with occupants.

Figure 1: The retrieval–grounding relationship (10 connectors, 5 seeds; marker size $\propto$ manifold distance). Grounding tracks proximity to the target manifold everywhere (Spearman $\rho = -0.89/-0.98$ vs. the k=1 nearest-target distance). (a) In the pre-aligned regime the top-right (high retrieval and high grounding) is empty: identity retrieves best ($R@1$ 50.2) yet grounds worst, and the contrastive hybrid retrieves at 42.9 but grounds poorly. (b) Unaligned, the exclusion softens—good learned maps are both discriminative and on-manifold, and only the extreme hybrid pays.

R^2=0.99, measured λ^2=2.31), the modality gap is its exclusion order parameter, and F6 is its Lagrangian tracing the empirical Pareto frontier (App. C).

What we do not claim. We never claim flows "strictly preserve" neighborhoods (a Lipschitz diffeomorphism can distort kNN graphs); the dataset coupling is a given coupling, not optimal transport, unless minibatch-OT is used; a post-hoc connector cannot fix Mirage-type decoder shortcut reliance (Asadi et al., 2026); and orthogonal Procrustes preserves all pairwise distances by construction and is unbeatable on structure metrics—our contribution is a frontier/mechanism claim.

2 Related work

Cross-modal priors and connectors. The DALL-E 2 prior (Ramesh et al., 2022) and ECLIPSE (Patel et al., 2024) transport CLIP text embeddings to CLIP image embeddings with diffusion / non-diffusion priors for text-to-image generation; connector modules similarly bridge a frozen vision encoder to a language model. Closest to our generative connector is Diffusion Bridge (Lee et al., 2025), which trains a diffusion model on text embeddings and injects image embeddings mid-reverse-process to pull them onto the text distribution for zero-shot captioning. These priors optimize one side of the tradeoff (generation/decodability); our contribution is to expose the retrieval–grounding tradeoff itself and navigate both sides with a single train-time objective (F6, §9), benchmarked against post-hoc projection. Our finding is cautionary for this line: on cross-modal retrieval a one-line contrastive projection dominates every flow, and a generative prior's outputs, while retrieval-strong, are less decodable than a regression map's—so the choice of objective matters more than the generative machinery.

Latent-space communication and translation. Relative representations (Moschella et al., 2023) and latent-space translation (Maiorca et al., 2023) align independently trained encoders via anchors or learned maps for zero-shot "communication." Our cross-model testbed is exactly this regime (CLIP image space $\rightarrow$ an independent text encoder's space, identity $\approx$ chance); we contribute a learned, invertible map and, more importantly, a grounding-side analysis of what such a translated embedding is good for once it lands.

Flow matching and conditional generation. We build directly on flow matching (Lipman et al., 2023), rectified flow (Liu et al., 2023), Riemannian flow matching on the sphere (Chen & Lipman, 2024), and stochastic interpolants (Albergo & Vanden-Eijnden, 2023), and instantiate the conditional-generation variant with classifier-free guidance (Ho & Salimans,

2022) and AdaLN-Zero conditioning (Peebles & Xie, 2023). Our selection analysis uses the continuous change-of-variables likelihood of CNFs (Grathwohl et al., 2019) (exact and Hutchinson-estimated divergence). We treat these as a family of connector classes and map where each sits on the retrieval/grounding plane, rather than proposing a new generative model.

Representational similarity and the modality gap. The Platonic Representation Hypothesis (Huh et al., 2024) posits converging representations; the Aristotelian view (Gröger et al., 2026) shows global-similarity scores are confounded by scale and, after a permutation-null calibration, only local neighborhood agreement survives—which motivates a topology-preserving map and supplies our calibrated mutual-kNN metric. The modality gap (Liang et al., 2022) (cone effect, contrastive temperature) explains why raw CLIP image embeddings, though retrieval-strong, sit off the text manifold—the mechanism behind "identity grounds worst" in our Fig. 1a. Schrodi et al. (2025) show the gap is driven by only a few embedding dimensions; our out-of-subspace-energy statistic (§8) is a complementary connector-side correlate—and we are explicit (via the intervention) that it is descriptive, not causal, and that the 95%-variance subspace is not the full text support.

Decoding embeddings, captioning, and evaluation. DeCap (Li et al., 2023) decodes CLIP latents into text with a text-only-trained decoder, and bridges the modality gap at inference by projecting a query embedding onto a support memory of training-text embeddings. Our grounding probe is a DeCap-style decoder; since this inference-time projection is itself a move toward the manifold, F6's train-time pull could be an expensive re-derivation of it, so we compare the two head-to-head (§9). We score with CIDEr (Vedantam et al., 2015). The Mirage effect (Asadi et al., 2026) shows multimodal benchmarks are inflated by textual shortcuts and contamination; we adopt its modality-ablation-delta protocol and mirage-mode filtering, and are explicit that a post-hoc connector cannot fix decoder-level shortcut reliance. Our contrastive control uses InfoNCE (van den Oord et al., 2018).

3 Setup

Backbones (frozen). CLIP ViT-B/32 (OpenAI weights) (Radford et al., 2021); SigLIP-base (Zhai et al., 2023) for the confirmatory replication. Data. MS-COCO (Lin et al., 2014) Karpathy split (Karpathy & Fei-Fei, 2015) (113,287/5,000/5,000); Flickr30k (Young et al., 2014) test for transfer. All training is on cached L2-normalized fp16 embeddings.

Two testbeds. (A) within-CLIP: source = CLIP image, target = CLIP text—already contrastively aligned, so identity is a strong baseline ($R@1$ 50.2). (B) cross-model: source = CLIP image, target = COCO captions embedded by an independent text encoder (bge-large (Xiao et al., 2023), PCA→512). The spaces are not jointly trained, so identity ≈ chance ($R@1$ 0.02) and a learned map is essential. (PCA is fit on the train split only; verified no leakage.)

Map classes. Identity; orthogonal Procrustes; linear ridge; MLP; and flows F1 (Euclidean rectified-flow transport (Liu et al., 2023)), F2 (spherical Riemannian flow matching (Chen & Lipman, 2024)), F3 (latent-noise / stochastic interpolant (Albergo & Vanden-Eijnden, 2023)), and F5 (conditional generation: pure-noise base, source injected as an AdaLN-Zero condition (Peebles & Xie, 2023) with classifier-free guidance (Ho & Salimans, 2022)). We also study a hybrid MLP trained with an added InfoNCE term (van den Oord et al., 2018). Non-generative maps share one regression objective ($\min \|g(x)-y\|^2$ on the paired coupling) so they lie on one comparable frontier.

Metrics. Retrieval $R@1/5/10$ (i2t/t2i); calibrated mutual-kNN via the permutation-null of Gröger et al. (2026) (our reimplementation matches their package to <1e-4); manifold distance = mean cosine distance to the k nearest held-out real targets ($k=1$ "nn1", $k=10$ "mdist"); grounding via a DeCap-style decoder (Li et al., 2023) (below); cycle consistency; and, for distributional maps, best-of-k and expected retrieval. For F3/F5 the manifold distance is reported for the single-sample draw (conservative), the expected map (mean of K; what the decoder decodes), and best-of-k (coverage); these differ and we label which is used.

Table 1: Baseline frontier (cross-model unless noted; R@1, 5 seeds).

	identity	Procrustes	linear	MLP
i2t R@1 (within-CLIP)	50.2	28.3	15.4	16.7
i2t R@1 (cross-model)	0.0	17.1	21.5	22.4
mKNN src→map	1.00	1.00	0.41	0.35
mKNN map→tgt (cross)	0.22	0.22	0.32	0.36
manifold-dist (cross)	0.84	0.58	0.29	0.24

Closed-form maps (identity/Procrustes/linear) are seed-invariant (std = 0); MLP i2t std = 0.25 (within), 0.34 (cross).

4 Flows do not trace a better frontier

Baselines. The two topology axes oppose (Table 1): Procrustes preserves source structure perfectly ($\text{mKNN}_{\text{src}\to\text{map}} = 1.0$) but cannot match the target's; regression maps do the reverse. No baseline wins retrieval and both axes.

Transport flows collapse. Deterministic F1 learns the conditional mean of the five captions: retrieval is flat across NFE 2–32 (14.3 → 14.06 cross-model) and off-manifold (mdist 0.44 vs. MLP 0.24); a 40× capacity sweep (0.5M→20M) leaves it at ≈ 15, so the collapse is objective-level, not capacity-level. F2 (spherical) ≈ F1, i.e. geometry does not help. F3 (noise) is distributional—best-of-10 covers the target (44.7) but single draws stay off-manifold and below the MLP on expected retrieval (17.6). KC1 (frontier) fails by construction: on the source→map axis Procrustes = 1.0 and no transporting map reaches 0.9; KC2 (transport vs. regression) fails for F1 and F3. Only invertibility survives (F1 image→text→image sem-cycle R@1 = 100%).

Directional asymmetry confirms the mechanism. One-to-many is image→5 captions, so a regression map collapsing to the conditional mean should hurt i2t far more than t2i (a caption still finds its image's collapsed point). Exactly this holds (Table S2): for the MLP/linear maps t2i exceeds i2t by +9 to +17 R@1, while orthogonal Procrustes—which cannot collapse—is roughly symmetric.

5 Conditional generation fixes retrieval, not manifold fidelity

F5 (guidance weight w selected on val only) is the one flow that beats the regression MLP on single-shot retrieval: cross-model **27.9** vs. MLP 22.4, within-CLIP 26.5 vs. 16.7 ($p < 10^{-4}$). We report F5 at the confirmatory-covered w=6 throughout; retrieval rises monotonically with w (a val-only sweep reaches 30.4 at w=10 but does not turn over, so we present w as a tradeoff parametrization, §8, not a higher headline number). Decoupling entropy (noise) from information (condition) is the mechanism. The gain replicates on SigLIP and Flickr30k and survives compute-matching (+2.7 R@1 vs. a 100-epoch MLP on SigLIP). But KC-F5 fails: F5's single-sample manifold distance exceeds the 1.1×MLP bar on both testbeds—a generative connector emits valid-but-novel points that retrieve well yet sit off the exact manifold (its distribution covers it: best-of-k mdist ≈ MLP).

Coverage without selection (KC-SEL). F5's oracle best-of-10 is ≈ 2× its single-shot, but no target-blind selector recovers this gap: exact-trace likelihood selection (Grathwohl et al., 2019) recovers −1.3% and medoid +4% (a norm-matched MLP+noise+medoid control recovers 1–3% of its own gap). Invertibility is a property, not a deployable selection mechanism.

6 What actually fixes retrieval: a contrastive term

Adding InfoNCE at λ=0.01 to the same MLP lifts cross-model i2t from 22.4 to **42.9**—+15 over F5's guidance, at one loss term and no ODE (Fig. S1a). Even $\lambda = 3 \times 10^{-4}$ already beats F5 (32.3); the discriminative signal the regression objective lacks is tiny. Conversely, F1+InfoNCE reaches only 24–28 despite more training compute (a differentiable-ODE forward per step): flow machinery underperforms a contrastive MLP under equal-or-greater

budget. On transfer the hybrid also dominates (Flickr30k i2t 65.3 vs. F5 39.2, MLP 23.8): F5's better transfer is a claim relative to regression only—the contrastive crown not only survives OOD but widens. (Flickr30k uses a 1k-image gallery vs. COCO's 5k, so absolute $R@1$ is not comparable across datasets; only within-dataset method contrasts are.) The retrieval collapse that motivates the flow program is an objective problem, not an architecture one. (Scope: in the pre-aligned regime even the best hybrid, 42.9, does not beat raw-CLIP identity, 50.2; the recommendation is for the unaligned regime.)

7 A positive: one field, many paths (RQ2)

A single direction-conditioned generative field, selected by a discrete direction/task token, matches or beats dedicated per-direction F5 heads on i2t and t2i on both testbeds (Fig. S1b; KC3 pass, within-CLIP $+4.7/+2.5$ $R@1$, $p < 0.01$), and absorbs a third captioning task without hurting retrieval. The required control shows the transfer is flow-specific: conditioning helps the flow on all four cells but hurts the MLP and hybrid (Table S3). Honest bound: the conditioned hybrid still retrieves at ≈ 42, so this is an amortization property of the flow, not a claim that it is the best connector.

8 The retrieval–grounding tradeoff and grounding (KC4)

A frozen DeCap-style decoder (GPT-2 + trained prefix, unimodal text-embedding training) captions each connector's output; grounding = CIDEr (Vedantam et al., 2015) (image-present)−CIDEr(image-absent), on a mirage-filtered subset. Across all ten connectors (Table 3, Fig. 1a), grounding tracks proximity to the target manifold (nn1 = cosine distance to the nearest real target): Spearman(nn1, grounding) = $\mathbf{-0.89}$ within-CLIP and $\mathbf{-0.98}$ cross-model (n=10; exploratory). Three controls (Fig. 2, 3 seeds × 2 testbeds) establish that this proximity is causal, not merely correlational. These controls (and Table 4) retrain the decoder for the head-to-head comparison, so their absolute grounding shifts by $\sim\pm 0.05$ from Table 3 per the decoder-seed caveat (e.g. MLP $0.30 \to 0.37$); orderings and every conclusion are invariant.

(a) Correspondence, not distribution (shuffle). Permuting which image each output is scored against—holding the output set fixed, so nn1 and every distributional statistic are unchanged—collapses grounding to ≈ 0 (within-CLIP MLP $0.37 \to 0.01$, F6 $0.30 \to 0.003$; cross-model both $\to 0.01$). Grounding is image-specific; it is not a property of the output distribution.

(b) Proximity, not subspace membership (intervention). Take the contrastive map and progressively project its outputs onto the 95%-variance text subspace. This drives out-of-subspace energy essentially to zero (gsr $0.27 \to 0.01$ within-CLIP, $0.18 \to 0.00$ cross-model) but does not reduce nn1 (the points stay far from real targets, just inside the subspace), and grounding does not move ($0.203 \to 0.208$; $0.295 \to 0.294$). Being in the subspace is not sufficient. Conversely, methods that actually cut nn1—a DeCap-style projection onto the text memory, and F6's nearest-real pull (§9)—raise grounding monotonically. Proximity to real target points is the operative causal quantity. All three controls are decoder-robust: they replicate under a decoder retrained with embedding-noise augmentation (App. B).

The one anomaly, reported as a descriptive correlate. At matched nn1≈ 0.20, orthogonal Procrustes grounds below the nn1 trend (0.169 vs. F5's 0.267). Two natural explanations fail—wrong semantic neighborhood and off the local sheet both predict no better than nn1 (Table 2). What best describes the anomaly is out-of-subspace energy (gsr): Procrustes, a rigid rotation, keeps 39% of its energy in image-space directions the text manifold underpopulates (vs. F5's 17%), and gsr correlates with grounding slightly better than nn1 ($\rho = -0.96$ vs. -0.89). But control (b) shows gsr is not causal—projecting that energy away restores no grounding—so we report it as a descriptive correlate, not the mechanism. The causal statement stands: Procrustes lands near a real point but along a direction the text-trained decoder never sees. This is captured by a single anisotropic decoder-metric distance $\Delta = dV^2 + \lambda^2\,\mathrm{perp}^2$ (in-subspace proximity error plus λ-weighted out-of-subspace energy): fit to the ten measured groundings it gives λ^2=2.31 (R^2=0.99), subsumes nn1 (λ=1) and

Table 2: The Procrustes anomaly (within-CLIP, 3 seeds). At matched nn1 and local residual (mrec), Procrustes' global subspace residual (gsr) is 2.3× F5's and correlates with its low grounding. gsr describes the anomaly better than nn1/mrec across all ten connectors, but the intervention (Fig. 2b) shows it is a correlate, not a causal lever.

	nn1 (k=1)	mrec (local sheet)	gsr (subspace)	grounding
MLP	0.074	0.029	0.120	0.299
F5	0.198	0.144	0.171	0.267
Procrustes	0.205	0.156	0.388	0.169
identity	0.663	0.627	0.636	0.079
Spearman vs. grounding	−0.89	−0.89	−0.96	—

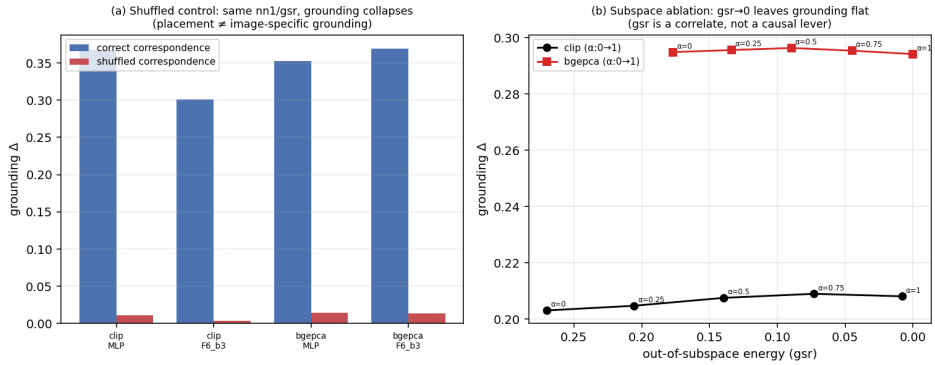


Figure 2: Grounding is caused by proximity to real target points, not by subspace membership. (a, left) Shuffling the image↔output correspondence (identical output set, hence identical nn1/gsr) collapses grounding to ≈0: manifold placement alone does not produce image-specific grounding. (b, right) Projecting a contrastive map's outputs onto the 95%-variance text subspace drives out-of-subspace energy (gsr) to ≈0 yet leaves grounding flat on both testbeds—gsr is a correlate, not a causal lever.

gsr (perp-only), and absorbs the Procrustes residual—the anomaly is exactly the decoder's off-subspace anisotropy, not a new variable. A solvable model built on Δ, its bounds, and their validation on these embeddings are in App. C.

KC4 fails: the MLP's grounding delta exceeds every flow's ($p < 0.05$, opposite the hypothesis). On-manifold regression outputs decode to more image-faithful captions; flows retrieve better but reconstruct worse.

The exclusion is regime-dependent (reported against our own headline). Within-CLIP the top-right (high retrieval and high grounding) is empty—identity is the sharpest case (retrieves best, grounds worst, nn1= 0.66 via the modality gap; that raw CLIP image embeddings decode poorly under a text-trained decoder is the known gap DeCap (Li et al., 2023) is built to bridge, which we here reproduce as one point on the curve). Cross-model, with a bge-space decoder, the top-right is occupied (MLP, F5, linear): good learned maps are both discriminative and on-manifold, and only the extreme hybrid pays (Fig. 1b). Thus the clean exclusion is a pre-aligned-regime phenomenon; the manifold-fidelity regularity is what holds across both.

A unifying mechanism: the marginal manifold cost of discriminative pressure. The two panels of Fig. 1 are one mechanism, not a claim and a counterexample. Grounding requires the output to be a typical point of the target manifold (a text-trained decoder decodes typical text embeddings); retrieval requires the output to be discriminative (well-separated across images). These conflict only to the extent that increasing discriminability moves an output off the target manifold—a marginal manifold cost of discriminative pressure that the geometry sets. In the pre-aligned regime CLIP's contrastive training has already made the

Table 3: Within-CLIP grounding (10 connectors, 5 seeds; mean±std). nn1 = k=1 nearest-target distance (the manifold metric grounding tracks); mdist = k=10. Closed-form maps (identity/Procrustes/linear) are seed-invariant.

connector	i2t R@1	grounding-Δ	nn1	mdist
identity	50.2±0.0	0.079 (worst)	0.663	0.682
hybrid λ=0.1	42.9±0.3	0.190±.002	0.447	0.493
hybrid λ=0.01	42.4±0.4	0.199±.003	0.413	0.457
F1	29.6±1.4	0.217±.004	0.157	0.178
Procrustes	28.3±0.0	0.169±.000	0.205	0.224
F2	27.6±1.1	0.227±.003	0.142	0.162
F5	26.5±1.1	0.267±.005	0.197	0.235
MLP	16.7±0.3	0.299 (best)±.001	0.074	0.089
linear	15.4±0.0	0.270±.000	0.087	0.102
F3	12.7±0.5	0.248±.006	0.090	0.104

Protocol: 5 connector seeds × a single decoder training (the decoder is the fixed measuring instrument); absolute CIDEr Δ shifts $\sim\pm0.05$ across decoder-retraining seeds, connector ordering is stable.

image cone the discriminatively optimal region, and the modality gap holds that cone off the text manifold; so among the standard connectors any step toward discriminability (up to identity, the raw cone) pays immediately, and the top-right is empty. In the unaligned regime a useful map must move to the target space regardless, so moderate discriminability is compatible with on-manifold placement—the marginal cost is near zero until the contrastive extreme forces outputs into low-density directions. The cost is not a hard wall, though: an objective with an explicit manifold-typicality term pays it down and navigates the tradeoff in both regimes (§9); what stays gap-limited within-CLIP is only the extreme corner (identity-level retrieval and grounding). The retrieval–grounding tradeoff is the general statement of this cost; the empty top-right is its special case under pre-aligned geometry. (Here "typical point" means near real target points, the causally-established quantity of §8; the Procrustes anomaly—a map near real points that still grounds poorly—is flagged descriptively by out-of-subspace energy but, per the intervention, is not a separate causal mechanism.)

9 Navigating the tradeoff: the mechanism is prescriptive (F6)

If grounding fails because discriminative pressure pushes outputs off the manifold, the mechanism prescribes a fix: keep the contrastive objective (for retrieval) but add an explicit manifold-typicality term that pulls each output toward the nearest actual target embedding—a soft nearest-real snap, not an MSE to the conditional mean (which would collapse retrieval). We call this F6: $L = \text{InfoNCE}(\text{map}(x), y_{\text{paired}}) + \beta \cdot \text{pull}$, $\text{pull} = \mathbb{E}[1 - \cos(\text{map}(x), \text{soft-NN}_{\text{real}})]$.

F6 navigates the tradeoff and Pareto-dominates every learned connector, in both regimes (Fig. S2, Table S1; the only undominated point is the raw-embedding retrieval corner—within-CLIP identity at R@1 50.2, grounding 0.079—which F6 does not reach on retrieval). Sweeping β traces a curve from the contrastive corner into the previously-empty top-right: at β=3, cross-model F6 reaches R@1 33.1 at grounding 0.347—higher retrieval than F5 (27.9) and the MLP (22.4) at their grounding (≈ 0.35); within-CLIP F6 reaches R@1 32.1 at grounding 0.301, dominating the MLP $(17, 0.30)$ and F5 $(27, 0.27)$. The MMD ablation (match the target marginal instead of a point-wise pull) does not navigate: cross-model it leaves grounding and nn1 essentially unchanged ($0.273 \to 0.278$, nn1 0.325 both), and though within-CLIP it does move nn1 ($0.449 \to 0.353$) it converts far less of that into grounding than the point-wise pull ($0.204 \to 0.230$ vs. the pull's $0.204 \to 0.301$ at the same β=3) and reaches no top-right point—confirming that grounding is a point-wise manifold property, and that the pull, not mere distribution-matching, is what works.

The train-time pull beats post-hoc projection (the key control). Since grounding is caused by proximity (§8), a skeptic could ask whether F6's training-time objective is necessary at all: why not take the retrieval-optimal contrastive map and simply project its outputs toward the

Table 4: F6 (train-time pull) vs. post-hoc projection of the same contrastive map, same decoder, 3 seeds (R@1 / grounding). This decoder is retrained for the head-to-head, so absolute grounding differs $\sim \pm 0.05$ from Table 3/S1 (decoder-seed caveat); orderings are unchanged. DeCap- and subspace-projection are inference-time. Top-right = above both 10-connector medians (within-CLIP 27.96/0.222; cross-model 19.53/0.275); $\star$ marks top-right points. No post-hoc projection reaches it.

	within-CLIP (R@1/grnd)	cross-model (R@1/grnd)
contrastive (β=0)	42.9/0.203	43.5/0.295*
F6 β=1	40.0/0.248*	40.5/0.319*
F6 β=3	**32.5/0.301***	**33.0/0.369***
F6 β=6	23.4/0.336	24.5/0.395*
DeCap-proj (best τ)	21.2/0.413	24.0/0.390*
subspace-proj (α=1)	42.3/0.208	42.9/0.294*
MLP (regression)	16.5/0.367	22.3/0.353*

Cross-model's top-right is broadly reachable (soft frontier), so the discriminating claim there is Pareto-domination of the projection curves; within-CLIP the top-right is empty for every method but F6.

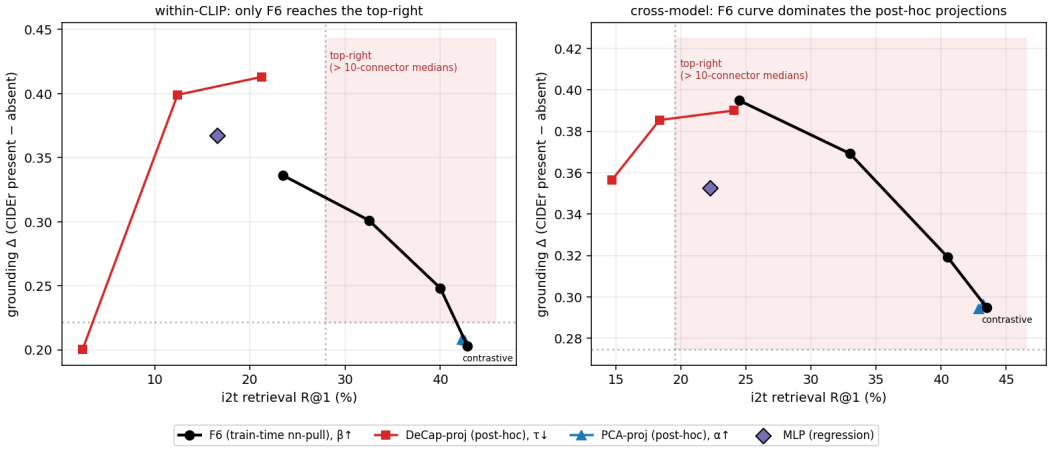


Figure 3: F6's train-time pull vs. post-hoc projection of the contrastive map. The DeCap projection (red) reaches high grounding only at low retrieval (top-left); the subspace projection (blue) sits on the contrastive point (no grounding gain). Only F6 (black) traces into the top-right (within-CLIP) and dominates the projection curves (cross-model). Shaded = above both 10-connector medians.

manifold at inference—exactly what DeCap (Li et al., 2023) does with its support memory? We test this head-to-head on the same decoder (Table 4, Fig. 3), applying to the contrastive outputs: (i) a DeCap-style projection (similarity-weighted snap onto a 100k text memory, temperature τ swept) and (ii) a subspace projection (fraction α onto the text subspace). Neither reaches the top-right. The DeCap projection buys grounding only by destroying retrieval—at its best it grounds well (0.41/0.39) but at R@1 $\approx$ 21/24, because snapping onto real points discards the discriminative structure retrieval needs, so it lands top-left; the subspace projection is inert (grounding 0.20 $\rightarrow$ 0.21; it does not cut nn1). F6's pull instead keeps discriminability while gaining typicality: within-CLIP it is the only method above both 10-connector medians, and cross-model its curve Pareto-dominates the DeCap-projection curve (e.g. F6 R@1 24.5/grounding 0.395 dominates DeCap 24.0/0.390). Jointly optimizing retrieval and proximity is not reducible to retrieving-then-projecting.

Scope and the pre-registered verdict. This refutes our own P-NAV2 (we predicted the within-CLIP exclusion was geometrically absolute; it is not). What stays gap-limited within-CLIP is only the extreme corner: no connector reaches identity-level retrieval (50) and good grounding. KC-NAV's strict secondary clause (F6 must also match the hybrid's own

$R@1 \geq 42$) fails—a mis-calibrated bar (Table S4)—but its substantive primary holds on both testbeds: F6 fills the top-right and Pareto-dominates the prior frontier ($p < 0.05$ vs. MLP/F5 at matched grounding). We read the tradeoff as navigable by design, not fundamental.

Generality: it holds off CLIP. To rule out a CLIP-specific artifact we replicate on a fully non-CLIP backbone pair—DINOv2 image (self-supervised, no language alignment) → bge text—with the same bge-space decoder (Fig. S3). Everything transfers: grounding tracks nn1 with Spearman $\rho = \mathbf{-1.00}$ ($p < 0.001$, 9 connectors); the contrastive corner is again retrieval-maximal but off-manifold (nn1 0.34); F6 again fills the top-right and Pareto-dominates the learned connectors (β=3: $R@1$ 27.4/grounding 0.324 strictly dominates F5 21.5/0.324 and trades a hair of grounding for +9.7 $R@1$ over the MLP 17.7/0.332; β=6: $R@1$ 20.2/grounding 0.349 then strictly dominates that MLP too); and the MMD ablation is again flat (0.261 → 0.269 grounding, nn1 unchanged). Absolute retrieval is lower (DINOv2 is not language-aligned), but the mechanism, the tradeoff, and its navigation are backbone-independent.

10 Diagnosis, limitations, and methodological lessons

Diagnosis. Retrieval collapse is an objective problem: every map that improves retrieval injects a retrieval-aligned signal the regression objective lacks, and a one-line InfoNCE term does this better and cheaper than a conditioned generative ODE. Grounding is caused by proximity to the target manifold, which regression has and transport/guidance trades away. The flow's distinctive properties—coverage and invertibility—are real but neither target-blind-recoverable nor a retrieval/grounding mechanism; its one genuine upside is flow-specific multitask amortization.

Limitations. (i) Data scale (10^5 pairs vs. $\sim 10^8$ for the DALL-E 2 prior (Ramesh et al., 2022)) bounds the negative manifold/selection claims. (ii) Results replicate on a non-CLIP backbone (DINOv2→bge, §9), but all targets are embedding text spaces on one dataset (COCO/Flickr). (iii) The retrieval–grounding exclusion is within-CLIP; cross-model it is a soft frontier—the universal claim is the narrower manifold-fidelity regularity. (iv) Grounding uses one DeCap-style decoder and CIDEr; quadrant/rank tests use n=10 connectors (exploratory). (v) A post-hoc connector cannot fix Mirage-type decoder shortcut reliance (Asadi et al., 2026).

Methodological lesson. KC1 was unsatisfiable by construction—it asked a transporting map to preserve source neighborhoods within 90% of an orthogonal map, which only identity/orthogonal maps can meet; a frontier kill condition must compare against the achievable frontier, not a ceiling. Separately, three results were nearly derailed by base/target scale mismatch (Gaussian base vs. unit targets; input-noise σ), each caught by a sanity check before a verdict. Reproducibility: frozen backbones, 5 seeds, logged FLOPs/NFE; the preregistration and every deviation are released with per-seed tables and figure code.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In ICLR, 2023.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. MIRAGE: The illusion of visual understanding. arXiv preprint arXiv:2603.21687, 2026.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In ICLR, 2024.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In ICLR, 2019.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In Proceedings of the 43rd International Conference on Machine Learning (ICML), 2026. arXiv:2602.14486.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. NeurIPS Workshop on Deep Generative Models, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. International Conference on Machine Learning (ICML), 2024.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.

Jeongryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. In ICLR, 2023.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In NeurIPS, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. In ECCV, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In ICLR, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In ICLR, 2023.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In NeurIPS, 2023.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In ICLR, 2023.

Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, and Yezhou Yang. Eclipse: A resource-efficient text-to-image prior for image generations. In CVPR, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In ICCV, 2023.

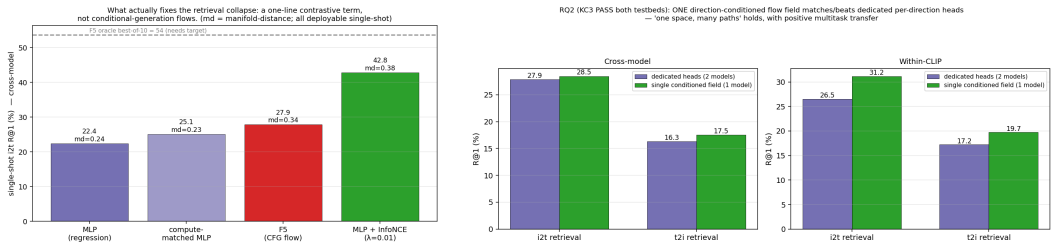
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In International Conference on Learning Representations (ICLR), 2025. Oral.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In CVPR, 2015.



(a) Cross-model single-shot i2t. (b) RQ2: one conditioned field vs. dedicated heads.

Figure S1: Left: a one-line contrastive term beats every flow on retrieval; the regression MLP grounds best but retrieves modestly. Right: a single direction-conditioned field matches/beats dedicated per-direction heads (KC3 pass), via flow-specific positive transfer.

Table S1: F6 (β sweep, nearest-real pull) on retrieval $\times$ grounding, 5 seeds. Points in the top-right (above the connector medians of Table 3) marked $\star$. The MMD ablation reaches no top-right point (cross-model flat; within-CLIP it moves nn1 but yields little grounding).

	cross-model			within-CLIP		
β	R@1	grounding	nn1	R@1	grounding	nn1
0 (contrastive)	43.9	0.273	0.325	43.1	0.204	0.449
1	40.3*	0.303	0.286	39.6*	0.249	0.312
3	33.1*	0.347	0.227	32.1*	0.301	0.188
6	24.6*	0.375	0.179	22.7	0.321	0.127
MMD (β=3)	43.7	0.278	0.325	43.1	0.230	0.353

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packed resources for general chinese embeddings (bge). arXiv preprint arXiv:2309.07597, 2023.

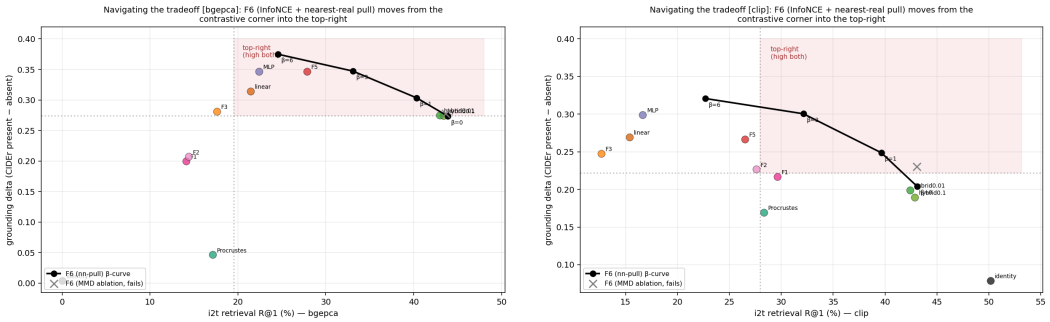
Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the ACL, 2014.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In ICCV, 2023.

A Kill-condition scorecard and additional tables

B Decoder robustness: a perturbed decoder

A skeptic may ask whether grounding measures image-specific manifold fidelity or merely familiarity with our particular decoder. Beyond the shuffle control (§8), we retrain the decoder with DeCap-style embedding-noise augmentation—each training embedding is perturbed by a random vector of 30% its norm (renormalized, fresh per batch), making the decoder tolerant of off-manifold inputs—and re-run the full battery (Table S6, 3 seeds $\times$ 2 testbeds). All three conclusions replicate: grounding still tracks proximity (Spearman(nn1, grounding) $= -0.93$ within-CLIP, -0.92 cross-model); the subspace intervention stays null (gsr $\to 0$ leaves grounding flat, $\Delta \leq 0.006$); the shuffle still collapses grounding ($\to 0.01$); and no post-hoc projection matches F6. The noise-robust decoder slightly compresses the grounding range (low-grounding maps rise, high-grounding maps fall, by ≤ 0.03) but leaves connector ordering unchanged—so the gains reflect image-specific fidelity, not decoder-specific adaptivity.



(a) cross-model (b) within-CLIP

Figure S2: F6 navigates the tradeoff. The β-curve (black) sweeps from the contrastive corner into the top-right, passing above/right of the standard connectors (Pareto-domination); the MMD ablation ($\times$) stays stuck at the corner.

Table S2: i2t vs. t2i R@1 (5 seeds). Regression maps show a large t2i>i2t gap (conditional-mean collapse of the image→text direction); Procrustes is near-symmetric.

testbed	direction	MLP	linear	Procrustes
within-CLIP	i2t	16.7	15.4	28.3
within-CLIP	t2i	34.0	32.8	21.9
cross-model	i2t	22.4	21.5	17.1
cross-model	t2i	33.7	30.3	22.7

C A model of the retrieval–grounding tradeoff: bounds and validation

We abstract the empirical findings into a solvable model and test its predictions on the same embeddings. Code and full derivations accompany the release.

Model. An output $\hat{y} = [u; p]$ (in-subspace u, off-manifold p) is summarised by three numbers: separation (drives retrieval R), in-subspace proximity error $dV = \|u - s\|$, and off-manifold energy perp $= \|p\|$. The decoder is trained only on the text subspace, so its decode error is the anisotropic decoder metric and grounding is a monotone function of it:

$$\mathrm{err}^2 \;=\; dV^2 + \lambda^2\,\mathrm{perp}^2, \qquad G \;=\; G_{\max}\,\exp\!\big(-\mathrm{err}^2/2\tau^2\big),$$

with λ the decoder's off-subspace sensitivity. Two order parameters govern the regime: δ (modality gap—how much off-manifold energy the standard family must spend to sharpen) and γ (image/text information imbalance). Every empirical result is one line of this model (Procrustes anomaly = metric anisotropy; the flat intervention = removing p barely changes err^2 when dV dominates; F6 navigation; post-hoc failure; soft-vs-hard regime).

Bounds. B1 (causal law): $G \le G_{\max}\exp(-(dV^2+\lambda^2\mathrm{perp}^2)/2\tau^2)$; nn1 is the isotropic ($\lambda$=1) proxy and gsr the perp-only proxy—neither is a causal lever (§8b). B2 (exclusion, order parameter δ): $G(R) \le G_{\max}\exp(-(\lambda\kappa\delta)^2(\mathrm{sep}(R)-\mathrm{sep}_{\mathrm{reg}})_+^2/2\tau^2)$; δ=0 is flat (soft frontier), δ>0 bends (hard exclusion), severity $\propto \delta^2$. B3: F6 solves $\max_{\hat{y}} R - \beta\,\mathrm{err}^2$, so its β-sweep is the Lagrangian/Pareto frontier and post-hoc projections are interior. B$_\gamma$: grounded retrieval $\le R_{\mathrm{cap}}(\gamma)$, a ceiling binding every connector.

Validation on the real embeddings. B1 (Table S7, Fig. S5): regressing $-\log G$ on measured dV^2 and perp^2 over the ten connectors, the anisotropic metric fits best (R^2=0.99) with λ^2=2.31, and the Procrustes residual collapses from +0.32 (nn1) to −0.04—the outlier is exactly the λ-weighting. B2: inflating the gap by scaling the image's out-of-subspace component holds retrieval fixed (≈ 50) while grounding falls $0.111 \to 0.054$ within-CLIP (fit by the B1 form with the measured λ, R^2=0.72); on the cross-model testbed the same manipulation is far flatter and the natural operating point is on-manifold (gsr 0.19 vs. within-CLIP 0.64)—i.e. cross-model sits at the $\delta\approx$0 end (Fig. S6). B3 (Fig. S7): on one shared

Table S3: RQ2 control (E1): Δ_{transfer} = conditioned − dedicated R@1, per (task, testbed), 5 seeds. Transfer is flow-specific: positive for F5 on all 4 cells, ≤ 0 for MLP/hybrid on 7/8. Param columns: conditioned model = one dedicated head's size (standard amortization); the dedicated set is 2× (one head per direction).

testbed	task	Δ(MLP)	Δ(hybrid)	Δ(F5)
cross-model	i2t	−0.45	−1.28	**+0.59**
cross-model	t2i	−0.83	−0.73	**+1.24**
within-CLIP	i2t	+0.08	−0.40	**+4.68**
within-CLIP	t2i	−0.13	−0.64	**+2.51**
conditioned R@1 (cross i2t)		21.4	41.7	28.5
dedicated R@1 (cross i2t)		21.8	43.0	27.9

KC	claim	verdict
KC1	flow Pareto-dominates frontier (src→map)	FAIL (by construction)
KC2	flow transports on-manifold better than regression	FAIL (F1, F3)
KC-F5	cond-gen > MLP retrieval AND on-manifold	FAIL (retrieval, not manifold)
KC-SEL	target-blind selector recovers best-of-k coverage	FAIL (exact-trace −1.3%)
KC3	one conditioned field ≈ dedicated heads	PASS (flow-specific)
KC4	flow grounds better than regression	FAIL (regression grounds best)
confirmatory	F5>MLP retrieval (SigLIP, Flickr)	holds; compute-matched too
KC-NAV	F6 fills top-right and beats hybrid at R@1 ≥ 42 (primary: fill top-right + dominate frontier)	FAIL on strict clause; primary holds, both regimes

Table S4: Pre-registered kill conditions, reported at equal weight. KC-NAV's strict secondary clause (match the retrieval king's R@1) was mis-calibrated; its primary—F6 fills the top-right and Pareto-dominates every learned connector (all but the raw-retrieval corner)—holds on both testbeds ($p < 0.05$).

decoder the F6 β-sweep is the empirical Pareto frontier—it dominates every standard learned connector (only the raw-retrieval identity corner within-CLIP escapes), the frontier is convex with β (Lagrangian signature), and 8/11 (within-CLIP) / 9/12 (cross-model) frontier points are F6; post-hoc projections are dominated except the hard DeCap-snap, which co-holds only the extreme high-grounding/low-retrieval corner. B_γ: within-CLIP the on-manifold projection already retrieves at 49.0 ≈ identity's 50.2 (off-manifold-only retrieval 1.2 R@1), so the unfillable corner is set by proximity (B2/B1), not information imbalance—B_γ is a valid mechanism but subdominant to the gap here.

λ	10^{-3}	10^{-2}	10^{-1}
F1+InfoNCE cross-model i2t R@1	25.3 ± 0.2	24.0 ± 0.5	28.4 ± 1.0

Table S5: E7: F1+InfoNCE, full λ grid, 5 seeds. All $\ll$ the contrastive MLP (42.9) despite greater training compute (differentiable-ODE forward per step)—flow machinery does not help under the right objective.

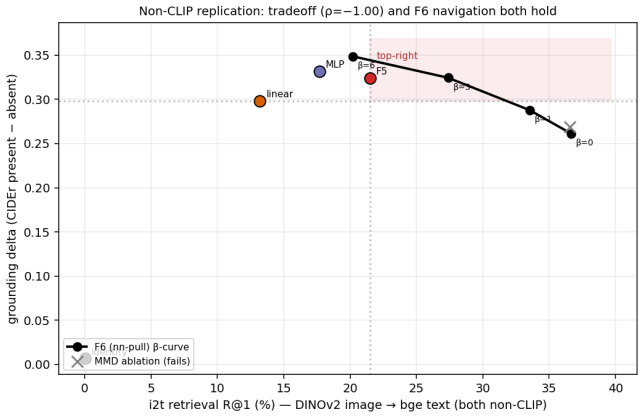


Figure S3: Non-CLIP replication (DINOv2 image → bge text, 3 seeds): the manifold-fidelity tradeoff ($\rho = -1.00$) and F6 navigation both hold; the MMD ablation ($\times$) again fails.

Table S6: Grounding under the standard vs. noise-augmented decoder (3 seeds). Ordering and the proximity trend are unchanged; the range compresses by ≤ 0.03.

	within-CLIP		cross-model	
connector (nn1↓)	standard	perturbed	standard	perturbed
contrastive	0.203	0.223	0.295	0.299
pca-proj (α=1)	0.208	0.229	0.294	0.299
F6 β=3	0.301	0.307	0.369	0.370
F6 β=6	0.336	0.324	0.395	0.389
MLP	0.367	0.339	0.353	0.371
DeCap-proj	0.413	0.382	0.390	0.381

Table S7: B1 on real CLIP: fit of $-\log G$ over the ten connectors, and the Procrustes residual. The anisotropic decoder metric fits best and absorbs the outlier (λ^2=2.31).

predictor of $-\log G$	R^2	Procrustes residual
nn1^2 (isotropic proxy)	0.82	+0.32 (outlier)
gsr only (perp^2)	0.97	−0.12
$dV^2 + \text{perp}^2$ (λ=1)	0.98	—
$dV^2 + \lambda^2\text{perp}^2$ (fit)	0.99	**−0.04**

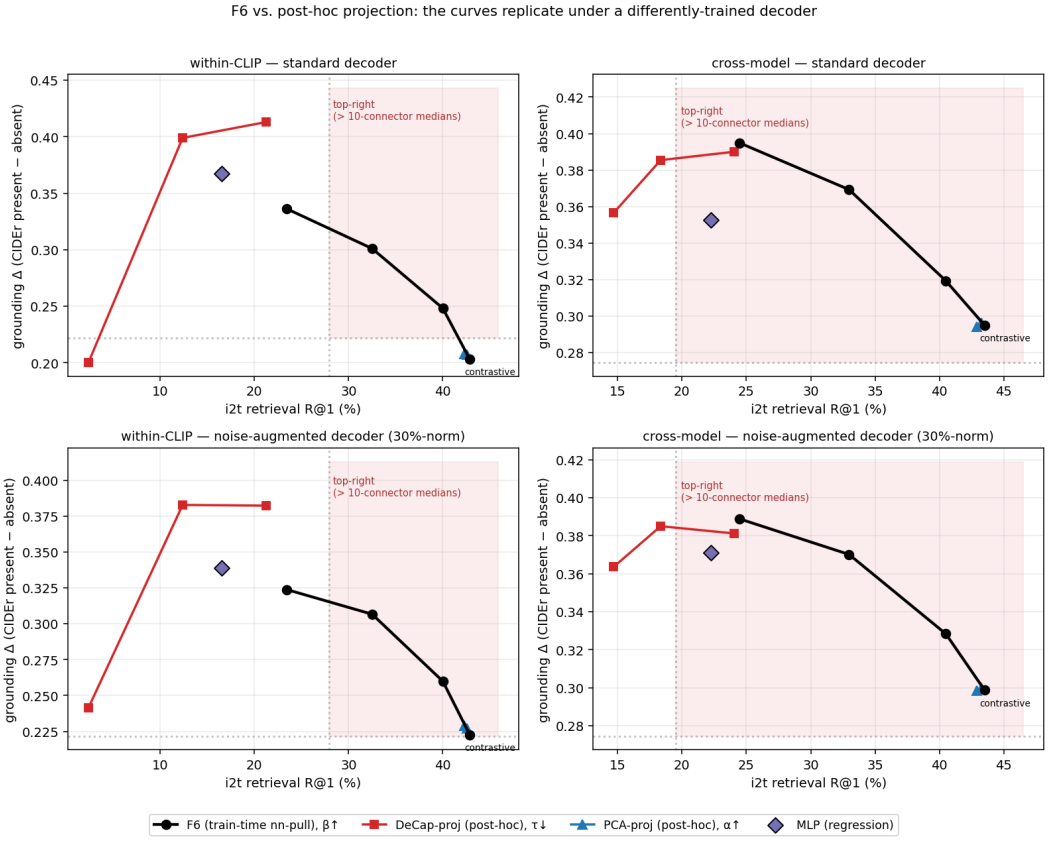


Figure S4: F6 vs. post-hoc projection under the standard (top) and noise-augmented (bottom) decoder. The retrieval×grounding curves are near-identical between decoders: F6 (black) reaches the top-right / dominates the projection curves in both rows; DeCap-projection (red) stays top-left; PCA-projection (blue) stays at the contrastive point. The noise-augmented decoder only compresses the vertical range slightly. Cf. the main-body Fig. 3 (standard decoder).

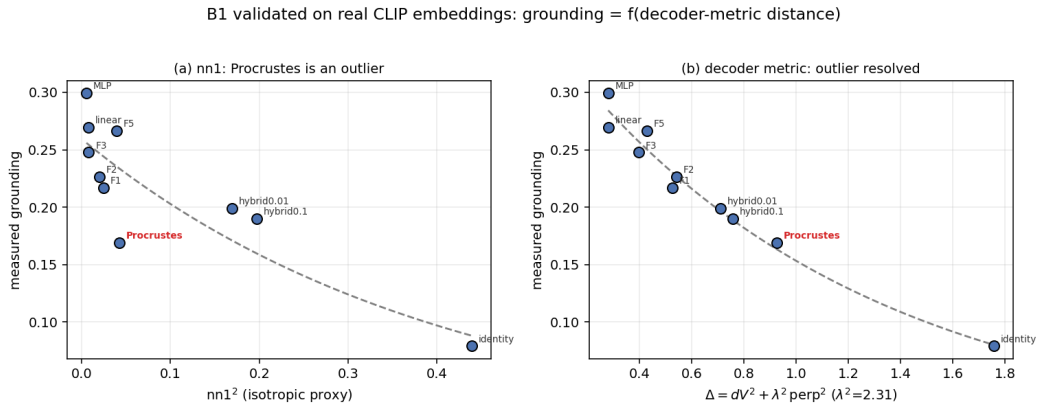


Figure S5: B1 on real CLIP. Under nn1 (left) Procrustes is an outlier below the curve; under the decoder metric $dV^2 + \lambda^2 \text{perp}^2$ (λ^2=2.31, right) it lies on the curve.

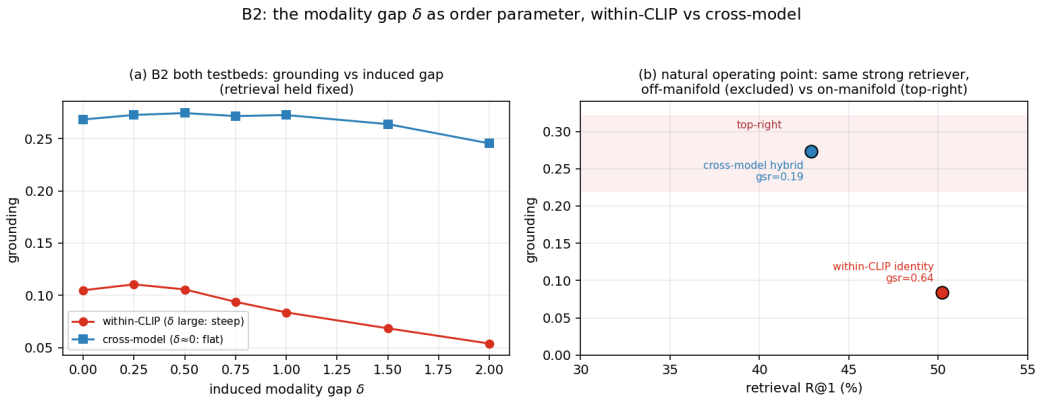


Figure S6: B2: the modality gap δ is the order parameter. (a) Inflating the gap lowers grounding at fixed retrieval—steeply within-CLIP, flat cross-model. (b) The same strong retriever is off-manifold/excluded within-CLIP but on-manifold/top-right cross-model.

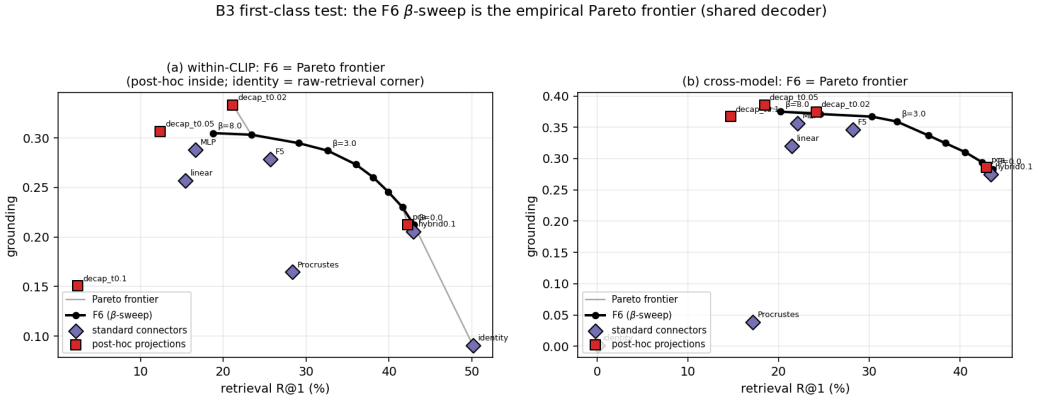


Figure S7: B3: the F6 β-sweep is the empirical Pareto frontier (shared decoder). Standard connectors sit inside; post-hoc projections are dominated (except the hard DeCap-snap top-left corner); only the raw-retrieval identity corner escapes F6 within-CLIP.