RETRIEVAL GAINS, SIMILARITY LOSSES: CONTROLLABLE SPECIALISATION FOR ENTITY-SPECIFIC FEW-SHOT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large pretrained image encoders capture general visual similarity well, but this can be too coarse when retrieval must identify an operationally defined entity such as a specific character, logo, or product variant. We study target-specific few-shot retrieval adaptation: independently learning a lightweight head for each target from ten target and ten non-target images while keeping the encoder frozen. The adaptation improves target retrieval, but systematically degrades similarity judgement on general image pairs that were not used for adaptation. Across OpenCLIP ViT-L/14, DINOv3 ViT-B/16, and SigLIP base-patch16-224, with both a linear head and an MLP with one hidden layer, Recall improves and semantic AUROC declines in all six conditions; visual AUROC declines in five, with CLIP–MLP as the exception. On 100 entities, the linear head raises Recall@5 from 80.00% to 86.75% on CLIP, from 55.92% to 62.25% on DINOv3, and from 82.17% to 87.33% on SigLIP. The trade-off persists across head structures, while its magnitude varies: the MLP loses about twice as much semantic AUROC as the linear head on all three encoders, whereas retrieval ordering depends on the encoder and cutoff. Interpolating frozen and adapted scores with one per-target scalar α_c lets a user select an operating point on this specialisation–generality frontier. It does not remove the trade-off: CLIP admits useful intermediate settings, whereas DINOv3 has no non-zero setting that preserves both AUROCs at their frozen levels. Thus, the central design question is not whether to adapt, but how much generality to exchange for target-specific retrieval.

1 INTRODUCTION

Image encoders pretrained on web-scale corpora, such as CLIP (Radford et al., 2021), DINOv3 (Siméoni et al., 2025) and SigLIP (Zhai et al., 2023), provide strong representations for retrieving images that look or mean similar. In practical retrieval, however, similarity alone is sometimes insufficient: one may need to distinguish visually similar characters, find only a particular revision of a company logo, or retrieve a specific product variant. In each case, the semantic similarity captured by a general representation does not coincide with the entity identity that the operator needs.

We study the situation in which the entity of interest is known at retrieval time and a small number of images of it have been registered — the normal state of affairs in licensing, brand protection and catalogue management, where an entity is enrolled once together with a handful of reference images and a handful of deliberately chosen confusable images, and is then queried repeatedly. The adaptation therefore happens at *enrolment* time, per entity, and must not require class names, textual descriptions, joint training over all entities, or any modification of the pretrained encoder. We refer to this problem as entity-specific few-shot retrieval adaptation. Unlike few-shot classification with a fixed label set, entities here arrive asynchronously and must be added or removed without retraining previously enrolled entities.

A practical solution is to attach a small projection head to the frozen encoder output for each entity and to train it with a supervised contrastive objective (Khosla et al., 2020) that treats the registered

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.

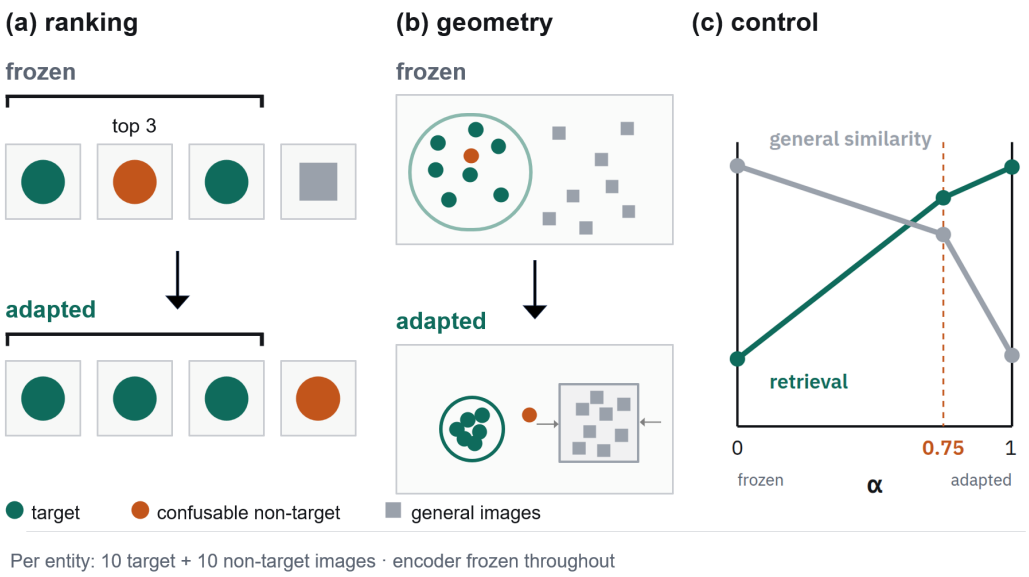


Figure 1: From practical motivation to representation effect. Left: generic similarity can rank a visually similar non-target beside the target. Middle: adaptation improves target separation but can alter general geometry. Right: score interpolation selects an operating point between frozen and adapted scores.

in-entity images as one group and the registered out-of-entity images as the other. It works: over 100 enrolled entities, mean Recall@5 rises from 80.00% to 86.75% (Section 6.1). A retrieval metric alone, however, cannot reveal whether the adapted representation preserves the broader similarity structure of the frozen encoder.

Recall alone cannot distinguish an adapted representation that learns a useful target boundary from one that merely distorts its general similarity structure in favour of that target.

What the task metric does not report. An entity-specific head is free to solve the retrieval task by collapsing the registered entity into a tight cluster, and a tight cluster is not the same thing as a better metric. The central question is therefore not whether retrieval can be improved, but whether an objective that supervises only entity discrimination preserves similarities it never supervises. To test this we evaluate the *same heads*, unchanged, on 304 human-labelled image pairs annotated for semantic and for visual similarity independently of our entities, with every pair that shares an image with a head's training set removed. Semantic AUROC falls from 0.860519 to 0.852533 (Section 6.2). Both effects are produced by the same weights, from the same twenty images, and nothing in the retrieval evaluation would have revealed the second. We do not assume that preserving general similarity is itself a direct retrieval objective; we use it as a diagnostic of how an adaptation trained only for entity discrimination changes similarity structure outside its training objective.

One knob, two things to protect. We further show that one scalar per entity can traverse this trade-off by interpolating frozen and adapted scores. Here $\alpha_c = 0$ favours generality and $\alpha_c = 1$ favours specialisation; the full definition and calibration protocols appear in Appendix G. We use this interpolation as a Pareto instrument, rather than as a universal remedy: a fixed intermediate setting is useful on CLIP, while the strict dual-AUROC rule selects no adaptation on DINOv3 (Appendix G, Section 6.4).

Contributions.

- **Specialisation–generality trade-off.** Over 100 targets and 304 human-labelled general similarity pairs, we show that the retrieval gain is accompanied by a loss in general-similarity discrimination: semantic AUROC declines in all six encoder × head conditions, while visual AUROC declines in five of six, with CLIP–MLP as the exception — Open-

CLIP ViT-L/14, DINOv3 ViT-B/16 and SigLIP base-patch16-224, linear and MLP head (Section 6.2, Section 6.3, Section 6.4). The trade-off persists across head structures and frozen backbones.

- **Head structure and adaptation strength.** Under paired bootstraps over identical heads, the MLP loses about twice as much semantic AUROC as the linear head on *all three* encoders, while the retrieval ordering between the two heads depends on the encoder and cutoff (Section 6.3). The nonlinear MLP incurs a larger generality cost than the linear head in our experiments. We therefore recommend the linear head for preserving general similarity at a fraction of the parameters, and explicitly do not claim it always retrieves better. We also show that score interpolation makes this trade-off observable and controllable, while its attainable operating points remain encoder-dependent (Appendix G, Section 6.4).
- **Target-specific retrieval adaptation and evaluation.** We formalise independent per-target adaptation from a small target/non-target support set and evaluate it jointly on retrieval and image-disjoint, task-external similarity pairs. This setting isolates whether an adaptation that improves target retrieval preserves the general similarity structure it was not trained to optimise.

2 Related Work

Lightweight adaptation of frozen encoders. CLIP-Adapter (Gao et al., 2024) and Tip-Adapter (Zhang et al., 2022) attach lightweight modules to a frozen CLIP backbone for few-shot classification, prompt-based methods (Zhou et al., 2022) adapt the text side instead, and parameter-efficient fine-tuning such as LoRA (Hu et al., 2022) modifies the backbone at low rank. These methods target a fixed label set, are trained jointly over all classes, and are evaluated on the task they were adapted for. Our unit of adaptation is a single enrolled entity, added and removed independently of every other, and a central part of our evaluation is deliberately *off* the adapted task.

Few-shot retrieval and metric learning. Prototype-based few-shot methods (Snell et al., 2017; Tian et al., 2020) are the most direct learning-free counterpart to our setting; instance-level retrieval (Wang & Jiang, 2015) shares our granularity but not our enrolment protocol. UCDR-Adapter (Anonymous, 2025b) and MAIL (Anonymous, 2025a) adapt vision-language models so that a *shared* model generalises to unseen domains and categories; EGA (Anonymous, 2026) adapts a frozen encoder for vector search while bounding degradation on out-of-distribution regions. Both lines protect generalisation with one shared mechanism. We accept per-entity heads and make the protection a per-entity *decision*, which is possible precisely because enrolment supplies a small labelled support set for each entity.

Negative transfer and representation preservation. That transfer can hurt is an old observation (Rosenstein et al., 2005) and has been surveyed extensively (Zhang et al., 2023); in the fine-tuning literature it appears as feature distortion (Kumar et al., 2022), mitigated by interpolating with the zero-shot model (Wortsman et al., 2022). Our mixture is structurally that interpolation, with two operational differences: the coefficient is per-entity rather than global, and it is selected from evidence available at enrolment rather than tuned on a validation set of the target task. Methods that preserve the pretrained representation usually do so inside the objective; PCSF leaves the objective and the trained head untouched and intervenes only at the score level, which is what makes it applicable to heads that are already deployed. Reporting worst-group rather than average performance follows the group-robustness literature (Sagawa et al., 2020), with enrolled entities playing the role of groups.

3 Problem Setting

Enrolment and retrieval. Let $\mathcal{E}$ be a set of entities. For an entity $c \in \mathcal{E}$, enrolment provides a positive support set P_c of in-entity images with $|P_c| = 10$ and a negative support set N_c of out-of-entity images with $|N_c| = 10$, the latter chosen to include visually confusable material. Let $f : \mathcal{I} \to \mathbb{R}^D$ denote a frozen encoder whose parameters are never updated. At retrieval time the entity being searched for is known: a query image q is issued *against* entity c, and the system ranks a gallery $\mathcal{G}$ that contains images of c, images of other entities, hard negatives and unrelated images.

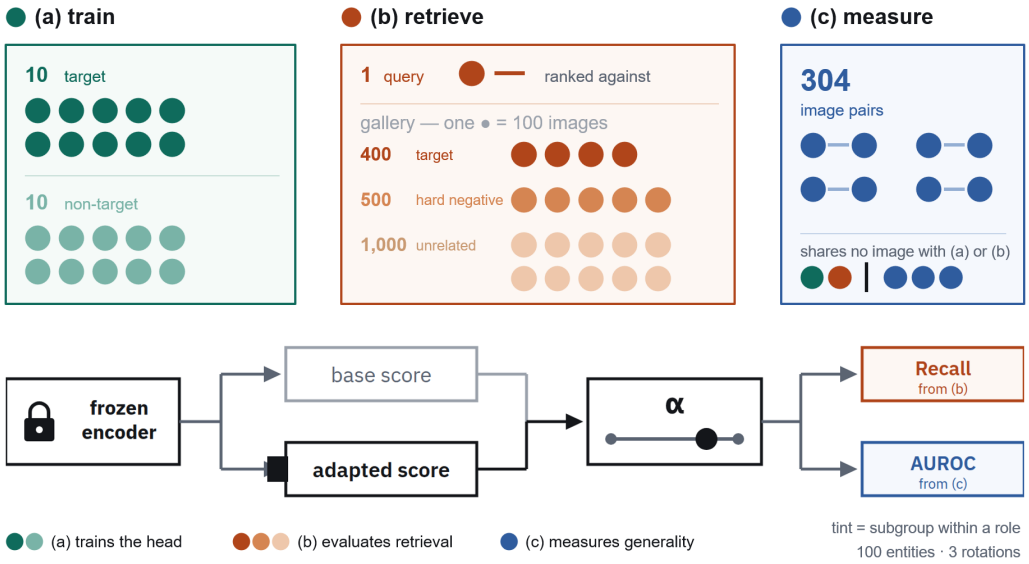


Figure 2: Evaluation protocol and data flow. The diagram makes explicit which data train each entity head, which data calibrate α_c, and which data evaluate retrieval, preventing train–calibration–evaluation leakage for each entity head by construction; a compact right panel recalls the scoring pipeline used to produce base/adapt/fused scores.

Performance is Recall@K for $K \in \{5, 10, 20\}$, averaged over entities and over the 3 rotated splits of Section 5.

The second objective. Retrieval accuracy alone cannot distinguish a head that has learned the operator's criterion from a head that has merely collapsed P_c. We therefore hold out an auxiliary set of general similarity pairs

$$\mathcal{B} = \{(x_i, y_i, \ell_i^{\text{sem}}, \ell_i^{\text{vis}})\}_{i=1}^{304}, \tag{1}$$

annotated by human raters for semantic and for visual similarity independently of any entity in $\mathcal{E}$. For each entity head h_c, every pair in $\mathcal{B}$ is scored by that head; this yields 100 head-specific AUROCs over the same pair set. We report AUROC for each label per head and the mean across heads, which requires no threshold. $\mathcal{B}$ is never used to train a head; it is used to *measure* every method and, for PCSF only, to calibrate one scalar per head under the protocol of Appendix G. A method is therefore evaluated on two axes at once,

$$\underbrace{\text{Recall@}K \text{ on } \mathcal{E}}_{\text{what adaptation targets}} \quad \text{and} \quad \underbrace{\text{AUROC on } \mathcal{B}}_{\text{what adaptation may damage}}, \tag{2}$$

and improvements on the first that are paid for on the second are reported as such.

Per-entity negative transfer and tail behavior are analyzed as an auxiliary result in Appendix B; the main paper focuses on mean retrieval and independent general-similarity preservation.

4 Method

4.1 Per-entity head

For every entity c we hold an independent head $h_c : \mathbb{R}^D \to \mathbb{R}^{128}$. Enrolling a new entity adds a head and updates nothing else; removing an entity deletes its head. The adapted representation is

$$u_c(x) = \frac{h_c(f(x))}{\|h_c(f(x))\|_2}, \tag{3}$$

and the adapted similarity is $s_c^{\text{adapt}}(x, y) = u_c(x)^\top u_c(y)$, to be contrasted with the frozen similarity $s^{\text{base}}(x, y) = \cos(f(x), f(y))$, which carries no entity index.

By default the head is a single linear map, $h_c(x) = W_c x + b_c$ with $W_c \in \mathbb{R}^{128\times D}$, followed by the ℓ_2 normalisation of Equation 3: $128(D{+}1)$ parameters, no hidden layer, no non-linearity. We also evaluate, as an ablation rather than as the proposed method, an MLP with one hidden layer, $D \to 256 \to 128$, with a ReLU and dropout of 0.1 between the two linear layers. Section 6.3 compares the two under an otherwise identical protocol, on the same 300 heads (100 targets $\times$ 3 repeats) across all three frozen encoders: the linear head loses about half as much general-similarity AUROC as the MLP across the three encoders, while which of the two retrieves better depends on the encoder. Since the extra capacity does not buy a consistent retrieval advantage and does cost generality, we do not carry it forward as the main method. Either head design absorbs the encoder's output width D without per-encoder tuning; only the first layer's input width depends on D across the encoders of Section 6.4, all of which emit $D = 768$ here.

4.2 Two-group supervised contrastive training

Each head is trained on $P_c \cup N_c$ with a supervised contrastive loss (Khosla et al., 2020) in which all of P_c carries one label and all of N_c carries the other:

$$\mathcal{L} = \sum_{i\in I} \frac{-1}{|A(i)|} \sum_{p\in A(i)} \log \frac{\exp\left(z_i^\top z_p/\tau\right)}{\sum_{a\neq i} \exp\left(z_i^\top z_a/\tau\right)}, \tag{4}$$

where z denotes the ℓ_2-normalised head output and $A(i)$ the set of samples sharing i's label. Treating the negatives as a group of their own, rather than only as negatives for the positives, is a deliberate choice: it makes the objective symmetric and stabilises training at this sample size. It is also a plausible mechanism for the degradation of Section 6.2, since the ten out-of-entity images are pulled together despite having nothing in common beyond not being c — and it is shared by every condition in which we observe the degradation, including a head with no hidden layer to exploit (Section 6.3) and a second, independently pretrained encoder (Section 6.4).

All entities and encoders use identical hyperparameters, with no per-entity tuning: AdamW (Loshchilov & Hutter, 2019) with learning rate 5×10^{-5}, weight decay 10^{-4}, $\beta = (0.9, 0.999)$, gradient clipping at 0.85, temperature $\tau = 0.07$, batch size 8, and 10 epochs. Only h_c receives gradients.

We additionally examine interpolation between frozen and adapted scores as an exploratory analysis; the score-fusion, PCSF, retrieval-gate, and target-level outer-CV details are provided in Appendix G and Appendix B.

5 Experimental Setup

Entities. The primary evaluation uses 100 enrolled entities in four groups: 40 character illustrations (CHAR), 20 trademarks and logos (LOGO), 20 fashion and product types (SKU), and 20 generic subjects (OTHER). The groups differ deliberately in granularity. CHAR and LOGO are instance-level: the entity is one specific character or one specific mark. SKU is category-level: the entity is a garment or accessory type such as a cardigan or a pair of high heels. OTHER mixes the two. This spread is intentional — an enrolled entity in deployment is whatever the operator declares it to be, and the granularity of that declaration is itself a variable we want the analysis to cover.

Membership is defined by an explicit inclusion rule written before collection — one character is admitted only in its default form and only when no visually adjacent character from the same work shares the frame; a logo rule decides affiliate marks and pre-rebrand versions explicitly. The ground truth therefore encodes an operator's criterion rather than a semantic category, which is precisely the gap that motivates adaptation. The complete entity list and the rules are in Appendix D.

Each entity contributes 15 images, split into 10 registered positives, 4 gallery positives and 1 query. A bank of 500 hard negatives and 1,000 unrelated images brings the collection to 3,000 images, and the retrieval gallery contains 1,900 images. Results are averaged over 3 rotations of the query/gallery/support assignment.

Auxiliary similarity pairs. $\mathcal{B}$ comes from an independently collected annotation exercise in which ten human raters (men and women in their 20s–50s) labelled each image pair on two axes through an

HTML interface displaying the two images side by side. Raters selected one of four categories: similar in meaning and in appearance (A), similar in meaning only (B), similar in appearance only (C), and neither (D). We retained a pair when at least seven of the ten raters selected the same category. The semantic binary label is $A/Bvs.\ C/D$ and the visual binary label is $A/Cvs.\ B/D$. The pairs were drawn without reference to our entity list, and the 312→304 filter of Appendix G removes the small number that nevertheless overlap a support set. Composition and annotation instructions are in Appendix E. Here, *general similarity* denotes similarity judgement on this independently collected, task-external pair benchmark, rather than similarity on the enrolled-entity retrieval task.

Encoders and configurations. The primary experiment uses a single frozen OpenCLIP ViT-L/14 encoder with plain cosine retrieval, over 300 heads (100 targets $\times$ 3 repeats), so that the effect of score interpolation in Appendix G and of head choice (Section 6.3) is not confounded with any fusion machinery. The OpenCLIP checkpoint uses the `datacomp_xl_s13b_b90k` pretrained weights. The whole protocol is then repeated on two additional frozen encoders, DINOv3 ViT-B/16 and SigLIP base-patch16-224, with the same 100 targets, the same 3 repeats, the same 304 auxiliary pairs, the same two head forms and the same target-level bootstrap, so that the cross-encoder comparison of Section 6.4 is like-for-like rather than suggestive. All three encoders emit $D = 768$, so the head shapes are identical across encoders as well. The SigLIP add-on evaluates only Base, Linear Adapt and MLP Adapt; its fusion, target-level outer-CV and seed-sensitivity analyses are outside the scope of this study. Appendix A reports the deployed system, which combines 3 encoders with element-wise colour, shape, texture and composition features by weighted linear fusion of z-normalised scores in the 4 configurations it exposes; there the fusion weights are held fixed between conditions so the comparison still isolates the heads.

6 Results

6.1 Entity-specific heads improve retrieval

On frozen CLIP with plain cosine retrieval, per-entity linear heads raise mean Recall@5 from 80.00% to 86.75%, Recall@10 from 88.50% to 94.00% and Recall@20 from 92.83% to 96.17% (Table 1, first two rows), over the same 300 heads used throughout Section 6. The gain is concentrated where the frozen encoder is weak, which is the instance-level groups: logos in particular are flat graphics that photographic pretraining has little reason to separate from other flat graphics. The category breakdown is in Appendix F.

6.2 The same heads degrade general similarity judgement

Table 1 also reports the two AUROCs. Semantic AUROC falls from 0.860519 to 0.852533 ($\Delta = -0.00802$, 95% CI $[-0.00902, -0.00699]$) and visual AUROC from 0.640658 to 0.636147. The retrieval gain and the similarity loss are produced by the same weights, from the same twenty images, and measured on pairs that share no image with the training set of any head. Section 6.3 shows the same pattern for the MLP head and Section 6.4 shows it again on additional frozen encoders. The trade-off persists across head structures and frozen backbones.

Two readings are available and we prefer the second. The first is that the loss is small in absolute terms, which it is. The second is that its sign is the informative part: twenty images of one entity, contrasted against ten images of nothing in particular, systematically move a general-purpose metric in the wrong direction, and no retrieval metric would have shown it. The observed per-entity changes could matter at catalogue scale, although we do not directly measure this accumulation.

We are explicit about what we do not claim. The visual AUROC of the frozen encoder is 0.640658 on CLIP, close enough to chance that we treat the visual axis on that encoder as a guardrail rather than as a result; on DINOv3 the frozen visual AUROC is 0.776662, high enough that the visual loss there (-0.00716, Section 6.4) does carry information. Our experiments do not distinguish between two possible mechanisms: collapse of the entity cluster, or the two-group objective binding the ten out-of-entity images into a group with no semantic content — though Section 6.3 rules out "it is a property of the MLP's hidden layer" as the explanation. Figure 3 shows that the trade-off varies by entity rather than appearing as a fixed exchange rate paid by every entity.

Table 1: Canonical CLIP results over 100 entities, 3 rotated splits, and 304 held-out human-labelled pairs. Recall is the objective the heads were trained for; AUROC is the task-external diagnostic.

	Entity retrieval (%)			General pairs (AUROC)	
Scoring rule	R@5	R@10	R@20	semantic	visual
Frozen CLIP	80.00	88.50	92.83	0.860519	0.640658
Linear Adapt	86.75	94.00	96.17	0.852533	0.636147
MLP Adapt	87.33	94.50	96.17	0.846592	0.651623

6.3 Is a hidden layer necessary? Linear versus MLP

The canonical absolute Linear/MLP values are shown in Table 1. The linear head incurs a smaller semantic-AUROC loss than the MLP across all three encoders, while the retrieval ranking between the two heads depends on the backbone. Thus a hidden layer is not necessary for the adaptation effect; the nonlinear MLP incurs a larger generality cost than the linear head in our experiments. Detailed paired-bootstrap statistics are in Appendix G; the full fusion-policy comparison is also in Appendix G.

The detailed score-fusion, PCSF, Global-CV, alpha-sweep, and target-level outer-CV analyses are provided in Appendix G.

The retrieval-gate analysis and its table are reported in Appendix B.

6.4 Does the trade-off hold for three frozen encoders?

If the phenomenon of Section 6.2 were a property of CLIP's particular geometry it would be a much narrower finding, so we repeated the entire protocol on DINOv3 ViT-B/16 and SigLIP base-patch16-224: the same 100 targets, the same 3 repeats, the same 304 auxiliary pairs, the same two head forms and the same target-level paired bootstrap (5,000 resamples over targets). Table 2 reports DINOv3, Table 3 reports SigLIP, and Table 4 places all three encoders beside each other.

The trade-off replicates. On DINOv3 the linear head raises Recall@5 from 55.92% to 62.25% ($\Delta = +6.33$ points, 95% CI $[+4.50, +8.25]$), Recall@10 by +6.92 (95% CI $[+4.67, +9.42]$) and Recall@20 by +5.92 (95% CI $[+3.33, +8.75]$), while semantic AUROC falls by -0.00520 (95% CI $[-0.00594, -0.00443]$) and visual AUROC by -0.00716 (95% CI $[-0.00810, -0.00630]$). The MLP head behaves the same way with larger numbers on both axes. On SigLIP, Linear Adapt raises Recall@5 from 82.17% to 87.33% ($\Delta = +5.17$ points, 95% CI $[+2.83, +7.50]$) and MLP Adapt reaches 89.25%; their semantic AUROC values are 0.8827 and 0.8754, below the frozen value 0.8909. Across all six encoder $\times$ head conditions, all Recall intervals lie strictly above zero and all semantic-AUROC intervals strictly below it: not one condition gains retrieval for free.

Two things generalise across encoders and one does not. What generalises is the sign of the semantic effect, and the fact that the semantic-AUROC loss is about twice as large for the MLP as for the linear head on every backbone (Table 4). Visual AUROC declines in five of the six conditions; CLIP MLP is the sole exception. What does not generalise is which head retrieves better: the ordering varies across encoders and retrieval cutoffs (Section 6.3) — and whether a strictly non-degrading fusion point exists: on CLIP the intermediate α of Appendix G give useful operating points, whereas on DINOv3 the strict dual-AUROC rule selects $\alpha = 0.00$ because visual AUROC decreases monotonically in α. The generalisable claim of this paper is therefore the trade-off and its dependence on head structure, not any particular operating point. SigLIP does not include fusion, target-level outer-CV, or seed-sensitivity experiments, and is not used for claims about those analyses.

The deployed multi-feature system results are reported in Appendix A.

7 Discussion and Limitations

What the result is and is not. The finding we would defend is that entity-specific few-shot adaptation is not free, that the cost appears on an axis not captured by the retrieval metric, and that the

Table 2: Protocol-matched second encoder: DINOv3 ViT-B/16, 100 entities, 3 rotated splits (300 heads per head type) and the same 304 held-out pairs used for CLIP. Base carries no head and is therefore common to both head types. The Adapt − Base differences and their 95% target-level bootstrap intervals are in Table 4 and Table 12; every Recall interval lies above zero and every AUROC interval below it.

	Entity retrieval (%)			General pairs (AUROC)	
Scoring rule	R@5	R@10	R@20	semantic	visual
Frozen DINOv3 ($\alpha = 0$)	55.92	66.00	73.08	0.816104	0.776662
Linear head ($\alpha = 1$)	62.25	72.92	79.00	0.810900	0.769503
MLP head ($\alpha = 1$)	63.33	73.33	78.92	0.803935	0.762303

Table 3: Protocol-matched third encoder: SigLIP base-patch16-224, 100 entities, 3 rotated splits and the same 304 held-out pairs. SigLIP uses the same head shapes but a separately specified frozen-feature extraction and optimisation path.

Scoring rule	R@5	R@10	R@20	semantic AUROC	visual AUROC
Frozen SigLIP	82.17	90.42	93.75	0.8909	0.6799
Linear head	87.33	95.33	97.83	0.8827	0.6766
MLP head	89.25	96.25	97.58	0.8754	0.6724

trade-off persists across head structures and frozen backbones (Section 6.3, Section 6.4). A single scalar per entity lets an operator choose a point on the trade-off without retraining the head. The finding we would not defend is that the price is large: semantic AUROC moves by -0.00802 on CLIP, -0.00520 on DINOv3, and -0.82 points on SigLIP for the linear head. A reader who weights absolute magnitudes will read this as a null result with a small correction. Our argument for the alternative is that the semantic direction is consistent across three encoders and two head architectures — six bootstrap intervals, all strictly negative — while the observed per-entity changes could matter at catalogue scale, we do not directly measure this accumulation.

What does not transfer between encoders. Two claims are specific to CLIP in our experiments. The first is not that the linear head universally retrieves better than the MLP: in our data the retrieval ordering varies with the backbone and cutoff, so the head recommendation is stated on the similarity-preservation axis rather than on an unconditional retrieval win. The second is that an intermediate α can keep the retrieval gain while holding both AUROCs at the frozen encoder's level; on DINOv3 no such α exists, because visual AUROC decreases monotonically in α there. We consequently present fusion as a Pareto instrument and report the operating points as observations on a given encoder, not as defaults.

Additional analysis of adaptation strength. We additionally investigated interpolation between frozen and adapted representations. The detailed α sweep, per-entity calibration, and target-level outer-CV analyses are provided in Appendix G. These analyses confirm that adaptation strength controls the specialisation–generality trade-off, but do not provide a universally optimal fusion strength across encoders.

Why PCSF is not the proposed method. PCSF remains an Appendix ablation rather than the proposed method because its extra per-entity calibration complexity does not yield a universally better operating point.

Other limitations. The visual AUROC of the frozen CLIP encoder is 0.640658, too close to chance for the visual axis to carry weight on that encoder, so we use it there as a guardrail only, for both the fixed and the calibrated fusion rules; on DINOv3 it is 0.776662 and the visual axis is informative, which is one reason the strict fusion rule behaves differently there. Storage and head count grow linearly in entities times encoders, and the setting assumes the target entity is known at query time. The retrieval gate needs a small labelled probe per entity, and the support-only alternative of Appendix B is described but not validated. Per-entity retrieval resolution is also coarse:

Table 4: Adapt − Base across all frozen encoders and head types. Values are from the single final cross-encoder summary, with every delta expressed in percentage points. Recall improves and semantic AUROC declines in all six conditions. Visual AUROC declines in five; CLIP MLP is the exception.

Encoder	Head	ΔR@5	ΔR@10	ΔR@20	Δsem. AUROC	Δvis. AUROC
OpenCLIP ViT-L/14	Linear	+6.75	+5.50	+3.33	-0.80	-0.45
OpenCLIP ViT-L/14	MLP	+7.33	+6.00	+3.33	-1.39	+1.10
DINOv3 ViT-B/16	Linear	+6.33	+6.92	+5.92	-0.52	-0.72
DINOv3 ViT-B/16	MLP	+7.42	+7.33	+5.83	-1.22	-1.44
SigLIP base-patch16-224	Linear	+5.17	+4.92	+4.08	-0.82	-0.34
SigLIP base-patch16-224	MLP	+7.08	+5.83	+3.83	-1.55	-0.75

with 1 query per entity per rotation over 3 rotations, per-entity Recall@5 moves in steps of 8.3 points. The submitted supplementary package provides the split manifests and provenance needed for the canonical Base/Adapt comparison, but it does not package every exploratory artifact used by the older fusion and gate analyses. The DINOv3 and SigLIP tracks are protocol-matched to CLIP for their Base/Adapt comparison. All three report one training seed; seed sensitivity is not a completed result. SigLIP does not include score fusion or target-level outer-CV, so it supports the cross-encoder adaptation–generality result but not fusion or unseen-target generalisation claims. The α sweep of Appendix G on DINOv3 selects over all targets rather than under the target-level outer cross-validation in Appendix G, so it is a Pareto description of this evaluation set and not a generalisation claim. The non-inferiority margins of Appendix G were registered before the DINOv3 results were inspected, but the Recall@20 comparison sits close enough to its margin to be sensitive to the bootstrap seed, and we do not build any conclusion on it.

8 Conclusion

Entity-specific few-shot adaptation improves target retrieval, while semantic AUROC declines in all six encoder–head conditions and visual AUROC declines in five of six. This specialisation–generality trade-off persists across CLIP, DINOv3, and SigLIP, and across both linear and nonlinear heads. The magnitude of the generality cost depends on head structure, with the linear head consistently incurring a smaller semantic-AUROC loss, while retrieval gains remain encoder-dependent. These results show that evaluating few-shot adaptation requires both task-specific retrieval metrics and an independent measure of broader similarity preservation.

Reproducibility Statement

The anonymous supplementary package provides the code, encoder configurations, anonymised support/retrieval manifests, 304-pair labels, and evaluation scripts for the canonical Frozen Base, Linear Adapt, and MLP Adapt comparison. A machine-readable summary of the reported aggregate results is included for cross-checking. The package provides an algorithmic and protocol-level reproduction path, not a self-contained reconstruction of every third-party image: reviewers must lawfully obtain the remaining images from the documented sources, whose bytes and availability can change. It does not contain the score-fusion, PCSF, Global-CV, target-level outer-CV, or retrieval-gate code; the corresponding analyses are therefore presented only as exploratory auxiliary material, not as reproducible primary results of the package.

Ethics Statement

The evaluation data consists of publicly available images of characters, trademarks and products, collected for the purpose of measuring retrieval accuracy and not redistributed. The application domain — brand protection and licensing — involves decisions about third-party content, and a retrieval system of this kind should support rather than replace human adjudication; the per-entity failure modes we document in Appendix B are directly relevant to that point. The auxiliary similarity pairs were annotated by ten human raters under the instructions reproduced in Appendix E; each rater evaluated image pairs for semantic and visual similarity. Trademarks, logos and character

names referenced in this paper belong to their respective owners; they are used solely for descriptive research purposes and do not imply endorsement, affiliation or sponsorship. Raw per-rater annotation records are not released. The supplementary material provides anonymised consensus labels and anonymous image identifiers for the 304 evaluation pairs, but does not release per-rater responses or private entity-name mappings.

AI USE STATEMENT

Our empirical claim concerns the supervised-contrastive entity-specific adaptation studied here; we do not claim that every possible adaptation objective must exhibit the same trade-off.

The research hypotheses, experimental design, and initial drafting of the manuscript were conducted by the authors. GitHub Copilot was used in a limited capacity to assist with method implementation, data processing, language editing, and translation. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Anonymous. Multi-modal interactive agent layer for few-shot universal cross-domain retrieval and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Anonymous. UCDR-Adapter: Exploring adaptation of pre-trained vision-language models for universal cross-domain retrieval. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025b.

Anonymous. EGA: Adapting frozen encoders for vector search with bounded out-of-distribution degradation. *arXiv preprint arXiv:2605.05674*, 2026.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 2024.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 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 *International Conference on Machine Learning (ICML)*, 2021.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. To transfer or not to transfer. In *NeurIPS Workshop on Inductive Transfer*, 2005.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Oriane Siméoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, 2020.

Shuang Wang and Shuqiang Jiang. INSTRE: A new benchmark for instance-level object retrieval and recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(3), 2015.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2022.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2), 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *International Journal of Computer Vision (IJCV)*, 2022.

A Deployed Multi-Feature System

The experiments above deliberately strip away everything except one encoder and a cosine. Inside the production system, which fuses 3 encoders with element-wise visual features in 4 configurations, mean Recall improves in every configuration and at every cut-off, by 3.25–7.92 points at $K{=}5$, with bootstrap intervals over entities excluding zero everywhere though only marginally for the high-precision ensemble ($[0.50, 5.83]$). The gain is largest for the shape-oriented configuration, whose unadapted performance is weakest, and smallest for the high-precision ensemble, whose unadapted performance is already high. The adaptation being analysed is therefore the one that is actually deployed, not a laboratory reduction of it.

B Retrieval-Gate Analysis

B.1 The other calibration signal: negative transfer inside retrieval

For a method $\mathcal{M}$ against the unadapted baseline, the negative-transfer rate is

$$\mathrm{NTR@}K(\mathcal{M}) = \frac{1}{|\mathcal{E}|} \sum_{c \in \mathcal{E}} \mathbf{1}\left[R@K_c^{\mathcal{M}} < R@K_c^{\mathrm{base}} \right]. \tag{5}$$

We also report mean degradation over the degraded subset and two tail summaries: the low-performance tail, defined by the bottom quartile of unadapted Recall, and the negative-transfer tail, defined by the bottom quartile of $\Delta R@K_c$.

Section 6.2 concerns damage outside the task. There is also damage inside it, and the same score interpolation described in Appendix G addresses it once α_c is calibrated differently. Measured over the deployed system's 4 configurations, so that the finding is not an artefact of one score definition, the per-entity change in Recall@5 is strongly bimodal: 53 of the 400 entity–configuration pairs *lose* accuracy, by 10.00 to 15.35 points on average, with a worst single case of -50.00 points (Figure 4a in Appendix G).

The tail is not random. Figure 4b in Appendix G plots each entity's improvement against its unadapted Recall@5; the relationship is negative in every configuration (pooled $r = -0.44$, Spearman $\rho = -0.52$). Binning by unadapted accuracy makes it plain: entities the frozen system already

Table 5: Retrieval-gated fusion at K=5, over the deployed system's 4 configurations. The gate is a threshold on the entity's unadapted Recall@5, fitted leave-one-entity-out. NTR@5 is Equation 5; the oracle is the per-entity upper bound and has NTR $= 0$ by construction.

	Semantic		Shape		Balanced		High-prec.		Mean	
Policy	R@5	NTR	R@5	NTR	R@5	NTR	R@5	NTR	R@5	NTR
No adaptation	80.00	—	55.92	—	82.17	—	82.08	—	75.04	—
Adapt all	86.83	10.0	63.83	10.0	89.17	14.0	85.33	19.0	81.29	13.3
Retrieval gate	87.00	7.0	63.92	8.0	88.92	11.0	86.58	13.0	81.61	9.8
Oracle	87.83	0.0	65.17	0.0	90.75	0.0	88.25	0.0	83.00	0.0

retrieves perfectly ($n = 106$) lose -1.81 points on average and degrade 11.3% of the time, whereas entities below 50% ($n = 55$) gain +13.03 points and degrade 7.3% of the time. Where unadapted retrieval is already perfect the head has nothing to gain and can only distort a working geometry, consistent with the fine-tuning distortion of Kumar et al. (2022). The tail is also configuration-specific — 35 of 100 entities degrade in at least one configuration but 0 degrade in all four — which is itself evidence that the failure is a property of the entity–geometry interaction rather than of the entity alone.

Table 5 applies the retrieval gate of Appendix B. Averaged over configurations it raises mean Recall@5 from 81.29% to 81.61% — it does not pay for its safety with accuracy — while cutting the negative transfer rate from 13.3% to 9.8%, and lifts mean Recall@5 over the worst quartile of entities from 55.50% to 55.84%. The honest framing is the oracle row: an entity-level oracle that knows the true per-entity outcome is worth only 1.71 points of mean Recall@5 over adapting everything. No gating rule can move the mean far. What gating buys is the tail, and the tail is what an individual operator experiences.

C Implementation Details

C.1 Head

The default head is a single linear layer $D \to 128$ followed by ℓ_2 normalisation: $128D+128$ parameters, and no per-encoder redesign needed since only the first (only) layer's input width depends on D. The MLP ablation of Section 6.3 instead uses a five-stage structure regardless of which encoder it sits on: a linear layer $D \to 256$, a ReLU, dropout at rate 0.1, a linear layer $256 \to 128$, and ℓ_2 normalisation, for $256D + 33{,}152$ parameters, of which $33{,}152$ is independent of the encoder. The shapes are taken from the shipped checkpoints rather than from the design document: across all three encoders the first weight is $(256, D)$ and the second is $(128, 256)$, so all three encoders use identical head forms and the MLP has exactly one hidden layer. Both heads carry biases. Dropout is placed after the activation rather than before it, so that the units suppressed are the ones the non-linearity has just amplified. With $10{+}10$ training images this is the MLP's main regulariser; the linear head has no dropout because it has no hidden activation to regularise. We did not find early stopping necessary at 10 epochs for either head.

C.2 Training

Only h_c receives gradients; the encoder runs in inference mode throughout. Optimisation uses AdamW (Loshchilov & Hutter, 2019) with learning rate 5×10^{-5}, weight decay 10^{-4}, $\beta = (0.9, 0.999)$ and gradient clipping at 0.85; the contrastive temperature is $\tau = 0.07$, the batch size 8 and the schedule 10 epochs with shuffling enabled. These values are identical for every entity and every encoder. No per-entity hyperparameter search is performed, which is a requirement of the setting rather than a simplification: enrolment must complete without a held-out set to tune against.

C.3 Score normalisation and fusion

Cosine similarities from different feature modules occupy different ranges, so each module's score is z-normalised before fusion. The reference statistics $(\mu, sigma)$ are computed once, at the end of enrolment, over the pairwise cosine similarities of all registered images of the entity (positives and negatives together), and the normalised score is mapped to a bounded display range by $S = 50 + 10\,s$. In the deployed system of Appendix A the final score is the weighted linear combination $S_{\text{final}} = \sum_i w_i S_i$ with $\sum_i w_i = 1$, and an adapted module replaces its unadapted counterpart in that sum without any change to w; this is what allows Table 6 to isolate the effect of the heads. The single-encoder experiments that carry the main claims use no fusion weights at all.

D Data Construction

D.1 Entities

Characters (CHAR, 40 entities, instance-level). Entity IDs: CHAR_001 to CHAR_040.

Logos (LOGO, 20 entities, instance-level). Entity IDs: LOGO_001 to LOGO_020.

Fashion and products (SKU, 20 entities, category-level). Entity IDs: SKU_001 to SKU_020.

Generic subjects (OTHER, 20 entities, mixed granularity). Entity IDs: OTHER_001 to OTHER_020. The first few IDs in this group correspond to instance-level landmarks, while the remainder correspond to object categories.

The real-name to ID mapping is maintained privately and is not required to interpret the aggregate findings reported in this paper.

D.2 Inclusion rules

Each entity carries a written rule fixed before collection, stating which renditions count as the entity and which do not. Representative examples:

- A character is admitted only in its default form. Transformed or power-up states, chibi or deformed renditions and live-action or costumed appearances are excluded unless the rule says otherwise, and images in which a visually similar character from the same work shares the frame are excluded.
- A character defined by a distinguishing accessory is admitted only when that accessory is visible and no substitute accessory is present.
- A logo rule states explicitly whether affiliate and subsidiary marks, pre-rebrand designs and co-branded lockups are inside or outside the entity.
- A category-level entity is defined by the garment or object type, with no constraint on brand, colourway or viewpoint.

These rules are the operational definition of the ground truth. They are also the reason the frozen encoder is not sufficient: several of them draw a boundary that is finer, or coarser, than any semantic category the encoder was trained to represent.

E Auxiliary Similarity Pairs

The auxiliary set $\mathcal{B}$ of Equation 1 originates in an annotation exercise carried out for a different purpose, in which raters judged each image pair on two independent axes and assigned it to one of four categories:

Label	Meaning	Used as
A	similar in meaning *and* in appearance	semantic 1, visual 1
B	similar in meaning, different in appearance	semantic 1, visual 0
C	similar in appearance, different in meaning	semantic 0, visual 1
D	different in both	semantic 0, visual 0

so that the semantic binary label is $A/Bvs.\ C/D$ and the visual binary label is $A/Cvs.\ B/D$. AUROC rather than accuracy or F_1 is used throughout because it requires no decision threshold, which matters here since the fused score of the score interpolation changes scale with α.

Of the 312 annotated pairs, those containing an image that also appears in the support set of any head are removed, leaving 304. The removal is by exact image identity.

F Per-Entity Results

Table 7 gives the category breakdown for each retrieval configuration of the deployed system, behind the summary in Appendix A, and Table 6 gives the aggregate the summary quotes. Table 8 lists every entity. The last two are generated directly from the evaluation output by `tools/make_tables.py`; no value is transcribed by hand.

Table 6: Deployed system: mean Recall@K (%) over 100 entities and 3 rotated splits, before and after entity-specific adaptation.

	Recall@5			Recall@10			Recall@20		
Configuration	base	adapt	Δ	base	adapt	Δ	base	adapt	Δ
Semantic-oriented	80.00	86.83	+6.83	88.58	95.08	+6.50	92.83	96.75	+3.92
Shape-oriented	55.92	63.83	+7.92	66.00	73.58	+7.58	73.08	79.00	+5.92
Balanced	82.17	89.17	+7.00	90.42	96.25	+5.83	93.75	97.42	+3.67
High-precision	82.08	85.33	+3.25	90.25	93.50	+3.25	93.83	96.83	+3.00

Table 7: Recall@5 by category and retrieval configuration (deployed system).

Configuration	Category	n	base	adapt	NTR@5
Semantic	CHAR	40	77.92	86.67	10.0%
	LOGO	20	59.58	72.92	5.0%
	SKU	20	90.42	92.92	15.0%
	OTHER	20	94.17	95.00	10.0%
Shape	CHAR	40	41.88	50.42	12.5%
	LOGO	20	21.67	32.92	5.0%
	SKU	20	84.58	91.67	15.0%
	OTHER	20	89.58	93.75	5.0%
Balanced	CHAR	40	80.83	87.29	17.5%
	LOGO	20	67.08	82.50	10.0%
	SKU	20	90.83	95.83	15.0%
	OTHER	20	91.25	92.92	10.0%
High-prec.	CHAR	40	78.54	81.67	25.0%
	LOGO	20	65.83	72.92	25.0%
	SKU	20	95.00	96.25	10.0%
	OTHER	20	92.50	94.17	10.0%

Table 8: Per-target Recall@5 before adaptation and the change produced by adaptation, for each retrieval configuration. Values are percentages and points, averaged over the 3 split rotations.

Entity ID	Cat.	Semantic		Shape		Balanced		High-prec.	
		base	Δ	base	Δ	base	Δ	base	Δ
CHAR_001	CHAR	91.7	-8.3	33.3	+16.7	100.0	+0.0	91.7	+0.0
CHAR_002	CHAR	58.3	+41.7	50.0	+0.0	66.7	+8.3	58.3	+25.0
CHAR_003	CHAR	83.3	+8.3	41.7	-16.7	91.7	+8.3	83.3	-16.7
CHAR_004	CHAR	58.3	+25.0	25.0	+8.3	75.0	+16.7	66.7	+25.0
CHAR_005	CHAR	66.7	+16.7	66.7	+33.3	91.7	+0.0	91.7	+0.0
CHAR_006	CHAR	66.7	+8.3	0.0	+25.0	75.0	+25.0	83.3	-41.7
CHAR_007	CHAR	75.0	-8.3	25.0	+8.3	91.7	-8.3	91.7	+8.3
CHAR_008	CHAR	66.7	+8.3	41.7	+0.0	100.0	+0.0	91.7	-8.3
CHAR_009	CHAR	91.7	+0.0	91.7	+0.0	100.0	+0.0	100.0	+0.0
CHAR_010	CHAR	83.3	+16.7	33.3	+16.7	66.7	+0.0	66.7	+0.0
CHAR_011	CHAR	66.7	+8.3	33.3	+16.7	83.3	-16.7	75.0	-8.3
CHAR_012	CHAR	75.0	+25.0	58.3	+0.0	83.3	+16.7	75.0	+25.0
CHAR_013	CHAR	83.3	+16.7	33.3	+16.7	91.7	+0.0	100.0	-8.3
CHAR_014	CHAR	83.3	-8.3	25.0	+16.7	66.7	+16.7	58.3	+25.0
CHAR_015	CHAR	83.3	+16.7	58.3	-16.7	91.7	-8.3	91.7	+8.3
CHAR_016	CHAR	75.0	+8.3	50.0	+16.7	75.0	+16.7	75.0	+16.7
CHAR_017	CHAR	100.0	+0.0	41.7	+16.7	100.0	+0.0	100.0	-25.0
CHAR_018	CHAR	100.0	+0.0	25.0	+50.0	91.7	+8.3	91.7	+8.3
CHAR_019	CHAR	66.7	+16.7	50.0	+0.0	83.3	-8.3	75.0	+8.3
CHAR_020	CHAR	91.7	+8.3	83.3	+16.7	91.7	+8.3	91.7	+8.3
CHAR_021	CHAR	83.3	+0.0	41.7	+0.0	83.3	+8.3	83.3	+0.0
CHAR_022	CHAR	66.7	+16.7	16.7	+8.3	58.3	+16.7	58.3	+8.3
CHAR_023	CHAR	66.7	+8.3	16.7	+0.0	58.3	-8.3	50.0	-16.7
CHAR_024	CHAR	83.3	+0.0	66.7	+16.7	83.3	+16.7	83.3	+8.3
CHAR_025	CHAR	66.7	+25.0	58.3	+8.3	75.0	+25.0	75.0	+25.0
CHAR_026	CHAR	100.0	+0.0	66.7	+8.3	100.0	+0.0	100.0	+0.0
CHAR_027	CHAR	58.3	+16.7	8.3	+16.7	58.3	+16.7	50.0	+25.0
CHAR_028	CHAR	75.0	+8.3	25.0	+25.0	58.3	+25.0	58.3	+8.3
CHAR_029	CHAR	75.0	+0.0	58.3	+8.3	75.0	+0.0	75.0	+0.0
CHAR_030	CHAR	83.3	+8.3	50.0	+0.0	75.0	+16.7	83.3	+0.0
CHAR_031	CHAR	75.0	+0.0	50.0	+16.7	91.7	+8.3	83.3	+8.3
CHAR_032	CHAR	66.7	+25.0	41.7	+0.0	66.7	+25.0	66.7	+16.7
CHAR_033	CHAR	83.3	+0.0	25.0	+16.7	66.7	+16.7	66.7	+16.7
CHAR_034	CHAR	100.0	+0.0	25.0	-25.0	100.0	+0.0	100.0	-50.0
CHAR_035	CHAR	83.3	+0.0	66.7	-16.7	75.0	+8.3	75.0	+8.3
CHAR_036	CHAR	75.0	+8.3	41.7	-8.3	58.3	-25.0	50.0	-8.3
CHAR_037	CHAR	50.0	+33.3	25.0	+16.7	75.0	-8.3	75.0	-8.3
CHAR_038	CHAR	100.0	+0.0	91.7	+0.0	100.0	+0.0	100.0	+0.0
CHAR_039	CHAR	75.0	-8.3	16.7	+8.3	91.7	+8.3	91.7	+0.0
CHAR_040	CHAR	83.3	+8.3	16.7	+16.7	66.7	+25.0	58.3	+33.3
LOGO_001	LOGO	33.3	+50.0	16.7	+16.7	33.3	+25.0	33.3	+25.0
LOGO_002	LOGO	41.7	+8.3	8.3	+8.3	75.0	+8.3	75.0	+8.3
LOGO_003	LOGO	75.0	+8.3	16.7	-8.3	50.0	-8.3	50.0	-8.3
LOGO_004	LOGO	83.3	+0.0	50.0	+8.3	58.3	+16.7	58.3	+0.0
LOGO_005	LOGO	100.0	+0.0	16.7	+8.3	83.3	+16.7	83.3	+0.0
LOGO_006	LOGO	66.7	+0.0	33.3	+0.0	66.7	+8.3	75.0	+0.0
LOGO_007	LOGO	25.0	+25.0	16.7	+0.0	50.0	+25.0	41.7	+0.0
LOGO_008	LOGO	58.3	+25.0	16.7	+16.7	100.0	+0.0	100.0	-8.3
LOGO_009	LOGO	41.7	+0.0	33.3	+8.3	58.3	+16.7	58.3	-16.7
LOGO_010	LOGO	75.0	+16.7	8.3	+0.0	91.7	+8.3	83.3	+16.7
LOGO_011	LOGO	83.3	+0.0	25.0	+8.3	75.0	+16.7	75.0	+8.3
LOGO_012	LOGO	58.3	-8.3	25.0	+8.3	83.3	+0.0	83.3	-8.3
LOGO_013	LOGO	75.0	+16.7	16.7	+8.3	100.0	+0.0	91.7	-8.3
LOGO_014	LOGO	58.3	+25.0	33.3	+25.0	66.7	+33.3	75.0	+16.7
LOGO_015	LOGO	25.0	+25.0	16.7	+8.3	50.0	+41.7	41.7	+25.0
LOGO_016	LOGO	66.7	+16.7	16.7	+8.3	50.0	+41.7	50.0	+25.0

continued on next page

Entity ID	Cat.	Semantic		Shape		Balanced		High-prec.	
		base	Δ	base	Δ	base	Δ	base	Δ
LOGO_017	LOGO	66.7	+16.7	41.7	+16.7	66.7	+16.7	66.7	+8.3
LOGO_018	LOGO	41.7	+0.0	0.0	+0.0	16.7	+33.3	16.7	+33.3
LOGO_019	LOGO	50.0	+41.7	0.0	+58.3	83.3	+16.7	83.3	+16.7
LOGO_020	LOGO	66.7	+0.0	41.7	+25.0	83.3	-8.3	75.0	+8.3
SKU_001	SKU	100.0	+0.0	91.7	-8.3	100.0	+0.0	100.0	+0.0
SKU_002	SKU	100.0	+0.0	91.7	+0.0	100.0	+0.0	100.0	+0.0
SKU_003	SKU	100.0	+0.0	50.0	+33.3	100.0	+0.0	100.0	+0.0
SKU_004	SKU	83.3	+16.7	75.0	+16.7	83.3	+16.7	91.7	+8.3
SKU_005	SKU	100.0	+0.0	75.0	+16.7	100.0	+0.0	100.0	+0.0
SKU_006	SKU	83.3	+16.7	83.3	-16.7	91.7	+0.0	91.7	-8.3
SKU_007	SKU	100.0	-16.7	91.7	+0.0	100.0	+0.0	100.0	+0.0
SKU_008	SKU	75.0	+8.3	83.3	+16.7	75.0	+25.0	83.3	+8.3
SKU_009	SKU	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
SKU_010	SKU	83.3	+8.3	75.0	+25.0	75.0	+25.0	83.3	+8.3
SKU_011	SKU	100.0	-16.7	83.3	+8.3	100.0	-8.3	100.0	+0.0
SKU_012	SKU	66.7	+8.3	75.0	-8.3	75.0	+8.3	83.3	+0.0
SKU_013	SKU	83.3	+0.0	100.0	+0.0	75.0	+16.7	83.3	+16.7
SKU_014	SKU	91.7	+8.3	75.0	+16.7	100.0	+0.0	100.0	+0.0
SKU_015	SKU	83.3	+8.3	83.3	+16.7	100.0	-8.3	100.0	+0.0
SKU_016	SKU	91.7	-8.3	100.0	+0.0	66.7	+33.3	83.3	+16.7
SKU_017	SKU	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
SKU_018	SKU	75.0	+8.3	83.3	+8.3	91.7	-8.3	100.0	-25.0
SKU_019	SKU	100.0	+0.0	83.3	+16.7	100.0	+0.0	100.0	+0.0
SKU_020	SKU	91.7	+8.3	91.7	+0.0	83.3	+0.0	100.0	+0.0
OTHER_001	OTHER	66.7	+16.7	66.7	+16.7	58.3	+25.0	50.0	+25.0
OTHER_002	OTHER	100.0	-8.3	91.7	+8.3	100.0	-8.3	91.7	+0.0
OTHER_003	OTHER	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
OTHER_004	OTHER	91.7	+0.0	91.7	+0.0	91.7	+8.3	91.7	+8.3
OTHER_005	OTHER	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
OTHER_006	OTHER	100.0	+0.0	66.7	+25.0	100.0	+0.0	100.0	+0.0
OTHER_007	OTHER	83.3	+0.0	83.3	+0.0	83.3	+8.3	83.3	+0.0
OTHER_008	OTHER	100.0	+0.0	91.7	-8.3	100.0	+0.0	100.0	+0.0
OTHER_009	OTHER	91.7	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
OTHER_010	OTHER	91.7	+0.0	100.0	+0.0	91.7	-25.0	100.0	-8.3
OTHER_011	OTHER	100.0	+0.0	100.0	+0.0	83.3	+16.7	100.0	+0.0
OTHER_012	OTHER	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
OTHER_013	OTHER	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
OTHER_014	OTHER	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
OTHER_015	OTHER	100.0	+0.0	91.7	+0.0	100.0	+0.0	100.0	+0.0
OTHER_016	OTHER	75.0	+16.7	58.3	+16.7	58.3	+0.0	75.0	+16.7
OTHER_017	OTHER	91.7	-8.3	66.7	+16.7	75.0	+0.0	75.0	-8.3
OTHER_018	OTHER	91.7	+0.0	83.3	+8.3	83.3	+8.3	83.3	+0.0
OTHER_019	OTHER	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0
OTHER_020	OTHER	100.0	+0.0	100.0	+0.0	100.0	+0.0	100.0	+0.0

G Additional Analysis of Adaptation Strength

G.1 Score fusion and per-entity calibration

Let both similarities be z-normalised with per-entity statistics computed once at enrolment from pairwise support scores (positives and negatives together):

$$\tilde{s}_c^m(x,y) = \frac{s_c^m(x,y) - \mu_c^m}{\sigma_c^m}, \qquad m \in \{\text{base}, \text{adapt}\}. \tag{6}$$

Fusion is applied to the normalised scores:

$$s_c^\alpha(q,g) = (1-\alpha_c)\,\tilde{s}_c^{\text{base}}(q,g) + \alpha_c\,\tilde{s}_c^{\text{adapt}}(q,g), \qquad \alpha_c \in [0,1]. \tag{7}$$

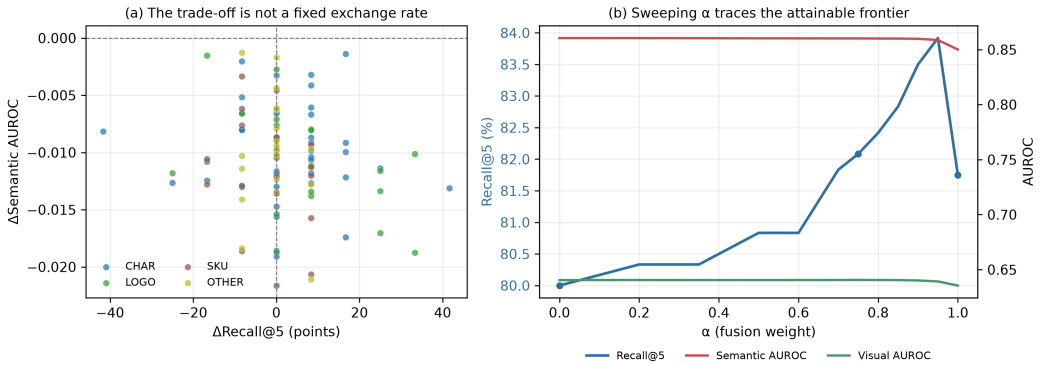


Figure 3: The retrieval gain and semantic-similarity loss vary by entity, while sweeping α traces the attainable specialisation–generality frontier.

Thus $\alpha_c = 1$ is full adaptation and $\alpha_c = 0$ is the frozen encoder under the same normalisation convention. No retraining is involved in evaluating a new α from stored base and adapted scores.

For PCSF, on the grid $\alpha \in \{0, 0.05, \ldots, 1\}$, select

$$\alpha_c^\star = \max\left\{\alpha : \mathrm{AUC}_c^{\mathrm{sem}}(\alpha) \geq \mathrm{AUC}_c^{\mathrm{sem}}(0) - \epsilon \ \wedge \ \mathrm{AUC}_c^{\mathrm{vis}}(\alpha) \geq \mathrm{AUC}_c^{\mathrm{vis}}(0) - \epsilon\right\}. \tag{8}$$

This is the strongest per-entity adaptation that does not exceed the permitted general-similarity degradation. Pair-level calibration uses image-disjoint folds; target-level outer-CV holds out entities entirely before selecting α.

The following score-fusion analysis uses a separate exploratory result series and is not reproduced by the submitted supplementary package. Table 9 compares, from the same stored base and adapted scores with no retraining, three fixed values of α, a single globally calibrated α (Global-CV), and PCSF, for both head types. For the linear head, fixed $\alpha = 0.75$ reaches 85.67% Recall@5 (versus 86.75% at $\alpha = 1$ and 80.00% at $\alpha = 0$) while giving up only -0.0065 of semantic AUROC relative to the frozen encoder. PCSF, calibrated per entity on $\mathcal{B}$, reaches 81.33% Recall@5 (95% CI $[+0.57, +2.30]$ pt) while giving up almost nothing on semantic AUROC. Global-CV lands close to the frozen encoder (80.10% Recall@5, 0.856914 semantic AUROC).

Repeating the sweep on DINOv3 shows the encoder dependence: Recall@5 rises monotonically in α while visual AUROC falls monotonically, so the strict rule admits no non-zero α and selects $\alpha = 0.00$ for both head types. Fusion therefore makes the trade-off explicit rather than dissolving it.

G.2 Target-level outer cross-validation

To ask what α a fusion rule would choose for *unseen* entities, target-level outer cross-validation partitions entities into five outer folds, selects α on four folds, and evaluates it on the held-out fold. The selected values are uniformly small and conservative ($0.05, 0.05, 0.05, 0.05, 0.00$, median 0.05, mean 0.04), far below the in-distribution $\alpha = 0.75$. Held-out Recall gains are close to zero, while semantic and visual AUROC remain essentially unchanged from the frozen encoder (+0.0000046 and +0.0000320).

Fusion, calibration, and encoder ablations. Table 12, Table 13 and Table 14 accompany Section 6.4; Table 9 accompanies Appendix G and Section 6.3 and is the full 300-head condition behind Table 1. Table 10 and Table 11 accompany Appendix G and Appendix G.

Distribution of the per-entity retrieval effect. Figure 4 is the evidence behind Appendix B: the effect of adaptation on retrieval, per entity, and its dependence on unadapted accuracy.

Worst-quartile accuracy. Mean Recall@5 over the low-performance tail (bottom quartile by unadapted Recall@5), which is the quantity most aligned with deployment complaints:

Table 9: Fusion ablation on frozen CLIP, 300 heads (both head types). The rows of Table 1 are the Base and Linear-Adapt rows here, repeated for reference. All rows use the interpolation in Appendix G over identical stored scores and differ only in how α is chosen; Global-CV uses one α pooled over all heads, selected under the protocol in Appendix G.

Head	Rule (α)	R@5	R@10	R@20	semantic AUROC	visual AUROC
Common						
—	Base ($\alpha = 0$)	80.00	88.50	92.83	0.860519	0.640658
Linear	Adapt ($\alpha = 1$)	86.75	94.00	96.17	0.852533	0.636147
Linear	Fixed $\alpha = 0.25$	82.08	90.33	94.17	0.856747	0.635872
Linear	Fixed $\alpha = 0.50$	84.33	92.17	94.58	0.856103	0.635703
Linear	Fixed $\alpha = 0.75$	85.67	92.58	95.58	0.854026	0.634559
Linear	Global-CV	80.10	88.70	92.90	0.856914	0.635055
Linear	PCSF	81.33	89.70	93.35	0.856737	0.635205
MLP	Adapt ($\alpha = 1$)	87.33	94.50	96.17	0.846592	0.651623
MLP	Fixed $\alpha = 0.25$	80.33	89.17	93.00	0.856975	0.635459
MLP	Fixed $\alpha = 0.50$	80.83	89.83	93.42	0.857015	0.635561
MLP	Fixed $\alpha = 0.75$	82.08	90.92	94.42	0.857063	0.636040
MLP	Global-CV	80.18	88.93	93.02	0.857146	0.635188
MLP	PCSF	81.15	89.80	93.17	0.856650	0.635384

Table 10: Mean per-head out-of-fold AUROC (Repeat0, 100 heads), the definition in Appendix G: the OOF AUROC of each head on its own held-out pair fold, averaged across heads, as distinct from pooling every head's out-of-fold predictions into one stacked AUROC.

Method	semantic AUROC	visual AUROC
Base	0.860476	0.640442
Linear Adapt	0.853385	0.636372
Linear Global-CV	0.860286	0.640175
MLP Adapt	0.850162	0.635033
MLP Global-CV	0.853014	0.635108

Policy	Semantic	Shape	Balanced	High-precision	Mean
No adaptation	55.67	16.00	57.67	56.00	46.34
Adapt all	69.00	23.67	68.67	60.67	55.50
Retrieval gate	69.00	23.67	68.67	62.00	55.84
Oracle	70.33	26.00	72.67	67.67	59.17

Per-split versus per-entity gating. A gate that could choose per split rather than per entity would be worth 3.02 points on average instead of 1.71. The difference is the part of the variance that is not a property of the entity at all, and no deployable rule can capture it, since the split is an artefact of evaluation. We report it only to bound how much of the remaining headroom is addressable.

Table 11: Target-level outer cross-validation (Appendix G): the linear fusion α selected on 4 of 5 entity folds and evaluated, held out, on the fifth.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Selected α	$0.05, 0.05, 0.05, 0.05, 0.00$				

Summary	median α	mean α	ΔR@5	ΔR@10	ΔR@20
Held-out	0.05	0.04	0.000000	+0.000833	+0.000833

Δsemantic AUROC $= +0.0000046$, Δvisual AUROC $= +0.0000320$.

Table 12: DINOv3 Adapt − Base with 95% target-level bootstrap intervals (100 targets, 5,000 resamples), behind Table 2. Recall deltas are in points. No interval crosses zero.

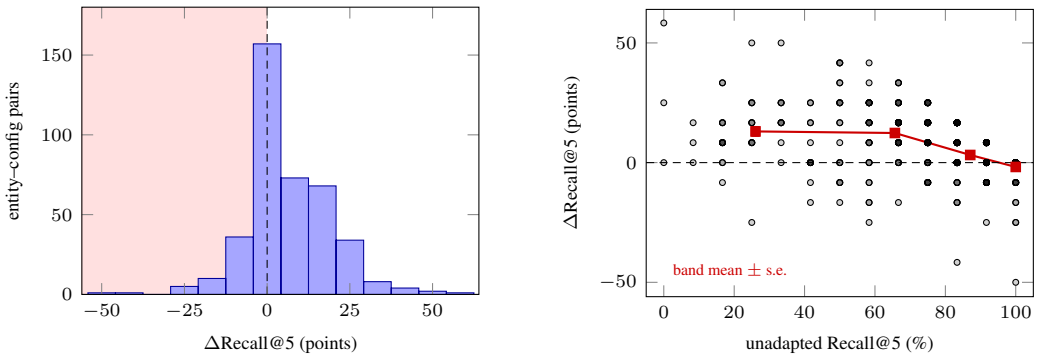
	Linear head		MLP head	
Metric	Δ	95% CI	Δ	95% CI
R@5	+6.33	$[+4.50, +8.25]$	+7.42	$[+4.92, +10.00]$
R@10	+6.92	$[+4.67, +9.42]$	+7.33	$[+4.67, +10.17]$
R@20	+5.92	$[+3.33, +8.75]$	+5.83	$[+3.08, +8.67]$
semantic AUROC	-0.00520	$[-0.00594, -0.00443]$	-0.01217	$[-0.01324, -0.01116]$
visual AUROC	-0.00716	$[-0.00810, -0.00630]$	-0.01436	$[-0.01560, -0.01310]$

Table 13: DINOv3 α sweep for the linear head (Appendix G), grid step 0.05. Recall@5 increases monotonically in α while visual AUROC decreases monotonically, so the strict rule in Appendix G, which requires both AUROCs to stay at or above the $\alpha = 0$ value, admits no non-zero α and selects $\alpha = 0.00$. The MLP head behaves the same way.

α	R@5 (%)	semantic AUROC	visual AUROC
0.00	55.92	0.81610	0.77666
0.05	56.42	0.81626	0.77644
0.25	57.75	0.81600	0.77582
0.50	60.00	0.81511	0.77452
1.00	62.25	0.810900	0.769503

Table 14: Non-inferiority check on DINOv3 (Section 6.4), as a sensitivity analysis rather than a main result. Margins were registered before the results were inspected: 1.0, 1.5 and 2.0 points for Recall@5, @10 and @20, and 0.010 for either AUROC. A condition is non-inferior when the lower bound of the 95% target-level bootstrap interval exceeds −margin.

Comparison	Metric	95% CI lower bound	Non-inferior
Linear Adapt − Base	semantic AUROC	-0.0059	yes
Linear Adapt − Base	visual AUROC	-0.0081	yes
MLP Adapt − Base	semantic AUROC	-0.0132	no
MLP Adapt − Base	visual AUROC	-0.0156	no
Linear − MLP (Adapt)	R@5	-3.08	no
Linear − MLP (Adapt)	R@10	-2.42	no
Linear − MLP (Adapt)	R@20	-1.92	marginal



(a) Distribution of the per-entity effect. Shaded: negative transfer.

(b) Effect against unadapted accuracy. No headroom at the right edge.

Figure 4: Negative transfer *inside* the retrieval task, over the 400 entity–configuration pairs at $K=5$. This is the signal the retrieval gate in Appendix B calibrates on.