RETRIEVAL GEOMETRY SHAPES CACHE-BASED CLIP ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cache-based test-time adaptation improves CLIP predictions by storing and retrieving examples from the target stream while keeping the model frozen. However, existing methods largely treat the feature space used for image–image retrieval as fixed, leaving open how much adaptation depends on the retrieval space itself. We study this question by fixing the memory and changing only the retrieval encoder, finding that the same memory can yield very different gains: across sixteen retrieval spaces, ImageNet-A cache gain ranges from at most $+0.44$ points for CLIP and MAE to $+19.7 \pm 0.4$ for DINOv2-L, while label-free retrieval-space selection retains 98% of oracle gain on ImageNet-V2. These results show that memory quality depends not only on which examples are stored, but also on how they are retrieved. Motivated by this finding, we propose MARC (Memory-Augmented Retrieval for CLIP), a training-free system that uses frozen CLIP for prediction and DINOv2-B for retrieval with a single fusion weight. A single-view cache repairs $1{,}074 \pm 21$ baseline errors, compared with 878 ± 4 for a 64-view ensemble, at roughly one seventh of the cost. Across four ImageNet distribution shifts, MARC reaches a 67.91% OOD average and, at matched DINOv2-B scale and eight views, achieves $64.17 \pm 0.31\%$ versus $62.75 \pm 0.15\%$ for a graph-based cache system while running 2.6 times faster. Overall, our results establish retrieval space as a first-order design choice for robust cache-based adaptation in remote sensing, scientific imaging, and changing visual environments.

1 INTRODUCTION

Reliable recognition under distribution shift is essential for deploying open-vocabulary vision models. **C**ontrastive **L**anguage-**I**mage **P**re-training (CLIP) (Radford et al., 2021) can recognize a wide range of categories without task-specific training, but its performance often drops under distribution shifts (Hendrycks et al., 2021b;a; Wang et al., 2019). Cache-based test-time adaptation (TTA) provides a training-free remedy: it stores features and pseudo-labels from the test stream, retrieves similar examples for each new image, and uses them to refine the zero-shot prediction (Karmanov et al., 2024; Zhang et al., 2024; Han et al., 2025; Zhang et al., 2025; Huang et al., 2025). Recent work has mainly improved how the cache is built and used, with techniques such as regional bootstrapping, distributional estimation, and graph-based retrieval.

However, cache retrieval depends on another design choice that has received much less attention: **the feature space used to measure image-image similarity**. Most cache-based methods retrieve neighbors using CLIP's image representation (Karmanov et al., 2024; Zhang et al., 2024). This implicitly asks one representation to serve two different purposes. Zero-shot classification compares an image with C text prototypes, whereas cache retrieval compares it with other images in the memory. Under distribution shift, a representation that works well for image-text matching may not produce the most useful image-image neighborhoods (Fig. 1). This motivates a simple question: *how much of cache-based adaptation comes from the memory mechanism, and how much comes from the retrieval similarities induced by the encoder?*

We study this question through a controlled retrieval-encoder intervention. Self-supervised encoders already provide strong k-NN representations (Caron et al., 2021; Oquab et al., 2023); our contribution is to isolate their effect on cache retrieval. Frozen CLIP determines the predictions, pseudo-labels, admissions, and retained indices, while the stream order is shared across spaces and only

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.

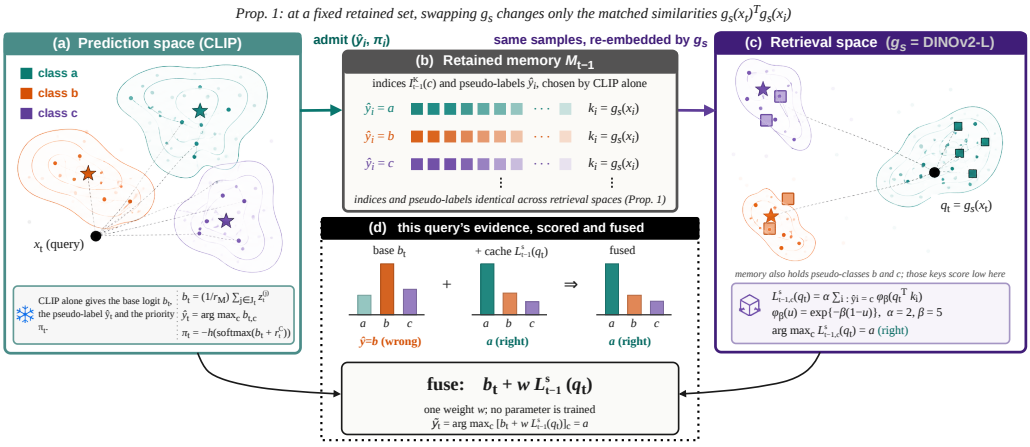


Figure 1: **Controlled retrieval-encoder intervention. (a)** Frozen CLIP computes the base logits b_t, pseudo-label $\hat{y}_t$, and admission priority π_t. **(b)** All retrieval spaces share the same retained samples and stored pseudo-labels, which are re-embedded as keys $k_i = g_s(x_i)$. **(c)** The retrieval encoder embeds the query as $q_t = g_s(x_t)$, so changing g_s modifies only the query–key similarities while keeping the cache trajectory fixed (Proposition 1). **(d)** DINOv2-L retrieves evidence for the correct class a, correcting CLIP's initial class-b prediction after fusion with $b_t + wL^s_{t-1}(q_t)$. The main MARC benchmarks use DINOv2-B (on 6 datasets: Caltech101, DTD, EuroSAT, Flowers102, Pets, and UCF101), while DINOv2-L is used for the controlled intervention and scale analyses.

the retrieval encoder changes the query–key similarities. This keeps the cache trajectory fixed and makes retrieval space the only changing variable. TDA (Karmanov et al., 2024) performs dynamic caching in CLIP space, while COSMIC (Huang et al., 2025) combines DINOv2 with dual semantic graphs and hyper-class querying. Our intervention instead directly measures the effect of changing retrieval space alone, with implications for adaptation across diverse application domains.

This intervention reveals that retrieval geometry affects not only neighbor quality, but also the gain and useful capacity of the cache. Existing methods use 1 to 64 augmented views (Karmanov et al., 2024; Huang et al., 2025; Shu et al., 2022; Feng et al., 2023; Zhang et al., 2024; Farina et al., 2024) and often tune the per-class capacity K by grid search. Across sixteen retrieval spaces and five streams, four different encoders achieve the highest cache gain. On ImageNet-A, changing only the retrieval encoder moves cache gain from at most $+0.44$ points (CLIP; MAE $+0.03$) to $+19.74 \pm 0.40$ for DINOv2-L, and from $+0.00$ to $+19.16$ once the fusion weight is selected on held-out ImageNet-V2. Neighborhood purity predicts the best space when candidate values are well separated, while pseudo-purity enables label-free selection from unlabeled target images.

Motivated by these findings, we introduce MARC (**M**emory-**A**ugmented **R**etrieval for **C**LIP), a training-free system with frozen CLIP and DINOv2-B encoders. MARC uses CLIP to build and admit samples into the memory, retrieves them with DINOv2-B, and fuses the retrieved evidence with the zero-shot prediction. This simple design decouples prediction from retrieval, allowing the retrieval space to be changed without altering the predictor or the cache trajectory, and provides a clean way to study their individual effects. Our core contributions are:

1. **We isolate retrieval similarity as a design variable.** Holding the predictor, prompts, admissions, retained indices, and test stream fixed, we vary only the retrieval encoder and its query–key similarities, directly isolating their effect on cache-based adaptation.
2. **We establish retrieval geometry as a key determinant of cache effectiveness.** Across retrieval spaces and streams, cache gain varies from nearly zero to about 20 points, with different encoders emerging as strongest across shifts. Neighborhood purity predicts gain, anti-hubness captures useful capacity, and stronger retrieval reduces the marginal benefit of additional views.
3. **We enable label-free retrieval-space selection.** Pseudo-purity closely tracks labeled rankings and selects effective retrieval spaces without target labels, providing a practical alternative to oracle selection under online constraints.

Our findings have substantial practical implications. Pseudo-purity selection retains 90% and 98% of oracle gain on ImageNet-A and ImageNet-V2, respectively. Across four ImageNet shifts, MARC reaches a 67.91% OOD average and, at matched DINOv2-B scale and eight views, achieves $64.17 \pm$

0.31% versus $62.75\pm0.15\%$ for a graph-based cache while running $2.6\times$ faster. These properties are particularly relevant to deployment settings with evolving visual distributions, from remote sensing and scientific imaging to continuously changing visual environments. Together, our results establish retrieval space as a first-order design choice and position MARC as a simple reference system for studying its interaction with prediction, memory, and fusion.

2 Related Work

Cache-based TTA keeps the predictor frozen and adapts through external memory: Tip-Adapter (Zhang et al., 2022b) introduced key–value caching for CLIP, while TDA (Karmanov et al., 2024) extended it online using entropy-prioritized pseudo-labels; subsequent methods modify cache construction, statistics, or fusion (Zhang et al., 2024; Han et al., 2025; Zhang et al., 2025; Hu et al., 2024; Zhang et al., 2026). COSMIC (Huang et al., 2025) combines CLIP and DINOv2 caches with semantic graphs and hyper-class queries, whereas we isolate retrieval geometry by varying only the retrieval encoder while fixing cache construction and admission. Self-supervised encoders such as DINO and DINOv2 (Caron et al., 2021; Oquab et al., 2023) provide strong image–image neighborhoods, unlike reconstruction-oriented MAE (He et al., 2022), making them useful retrieval spaces for studying cache gain and capacity. Prior work has examined neighborhood structure through k-NN accuracy, purity, hubness, and affinity graphs (Radovanovic et al., 2010; Hu et al., 2024; Boudiaf et al., 2022), but has not isolated the feature space defining those neighborhoods as a design variable; we test whether purity and anti-hubness predict gain and useful capacity. Finally, while augmentation-based methods such as MEMO, TPT, DiffTPT, and ZERO (Zhang et al., 2022a; Shu et al., 2022; Feng et al., 2023; Farina et al., 2024) use multiple views for test-time adaptation, we study how their value changes with retrieval effectiveness. Additional related work is discussed in detail in Appendix A.

3 Method

3.1 Preliminaries: Problem Formulation

We consider online test-time adaptation with a frozen CLIP classifier. Unlabeled images $\{x_t\}_{t=1}^{T}$ arrive sequentially, with each x_t predicted using the frozen model and memory from prior samples before possible insertion into the cache. Ground-truth labels are used only for evaluation; the predictor, prompts, and class prototypes remain fixed, making adaptation causal and entirely memory-based.

Let f denote the frozen CLIP image encoder and $\mathbf{t}_c$ the unit-normalized text prototype for class c. We use $\bar{f}(x)$ to denote the unit-normalized image feature and τ the fixed CLIP logit scale. Each test image x_t is represented by M views. Let J_t index the $r_M = \max(1, \lfloor 0.1M \rfloor)$ views with the lowest prediction entropy. For view j, the raw logit for class c and the averaged base logit are

$$z_{t,c}^{(j)} = \tau \left\langle \bar{f}(x_t^{(j)}), \mathbf{t}_c \right\rangle, \qquad b_t = \frac{1}{r_M} \sum_{j \in J_t} z_t^{(j)}, \tag{1}$$

The base pseudo-label is $\hat{y}_t = \arg\max_c b_{t,c}$. It is produced by the frozen CLIP predictor and provides the class assignment used by both online memories in Section 3.

3.2 Memory-Augmented Retrieval

Our key design choice is the retrieval space. Let g_s be a frozen, normalized retrieval encoder defining space s. For each test image x_t, its unaugmented feature $q_t = g_s(x_t)$ is used as the retrieval key if the image is admitted to memory.

Cache Construction. We also maintain a CLIP-indexed memory $\mathcal{C}_t$ with a fixed per-class capacity of K_C=16. Before updating the memory, we query it with the CLIP feature $k_t^C = \bar{f}(x_t)$ to obtain $r_t^C = L^{\mathcal{C}_{t-1}}(k_t^C)$ using the kernel in Eq. (3). For $p_t^{(j)} = \mathrm{softmax}(z_t^{(j)})$, we define

$$e_t = h\left(\frac{1}{M} \sum_{j=1}^{M} \mathrm{softmax}(p_t^{(j)}) \right), \; \pi_t = -h\Big(\mathrm{softmax}(b_t + r_t^C) \Big). \tag{2}$$

where $h(p) = -\sum_c p_c \log p_c$ is Shannon entropy with natural logarithms. Following the evaluated TDA implementation, e_t applies the additional softmax shown above; normalized entropy quantities are marked explicitly. Replacing it with entropy of the view posterior gives $60.10 \pm 0.41\%$ versus $60.14 \pm 0.38\%$ across three paired ImageNet-A seeds; further admission controls are reported in

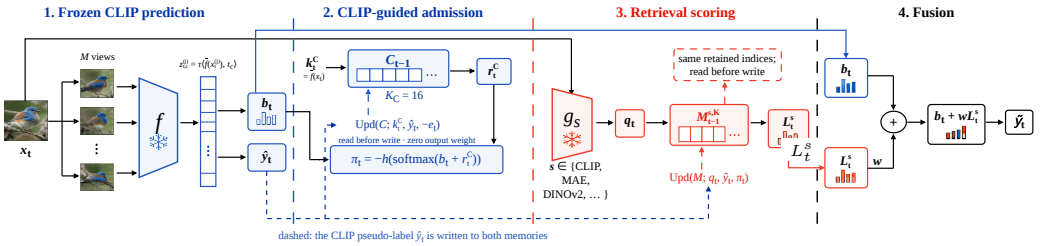


Figure 2: **The four-stage MARC workflow.** (1) Frozen CLIP aggregates M views into the base logit b_t and pseudo-label $\hat{y}_t$. (2) The blue, zero-output-weight CLIP memory is read before update to compute π_t and guide admission. (3) A frozen retrieval encoder g_s embeds the query and the same admitted samples; the red retrieval memory is read before update to produce L_t^s. Dashed arrows show the CLIP pseudo-label written to both memories. (4) Fusion returns $b_t + wL_t^s$. All encoder parameters remain frozen throughout.

App. H. The CLIP memory is updated with $(k_t^C, \hat{y}_t, -e_t)$, while π_t determines which samples enter the retrieval memory. The CLIP memory has no output weight, so it affects neither the final logit nor the stored pseudo-label. The strict one-memory setting uses $K_C{=}0$.

Retrieval and Prediction. For retrieval space s and per-class capacity K, the pre-query memory is $\mathcal{M}_{t-1}^{s,K} = \{(\mathbf{k}_i, \hat{y}_i, \pi_i)\}$, with at most K entries per pseudo-class. It produces class evidence

$$L_c^{\mathcal{M}}(q_t) = \alpha \sum_{i \in \mathcal{M}_{t-1}:\, \hat{y}_i = c} \phi_\beta(q_t^\top \mathbf{k}_i), \qquad \phi_\beta(u) = \exp\{-\beta(1-u)\}, \tag{3}$$

where $\alpha{=}2$ and $\beta{=}5$, following TDA (Karmanov et al., 2024). Both memories are read before they are updated. The prediction and memory updates are

$$\widetilde{y}_t = \arg\max_c \left[b_t + wL^{\mathcal{M}_{t-1}}(q_t)\right]_c, \; \mathcal{C}_t = \mathrm{Upd}_{K_C}(\mathcal{C}_{t-1}; k_t^C, \hat{y}_t, -e_t), \; \mathcal{M}_t^{s,K} = \mathrm{Upd}_K(\mathcal{M}_{t-1}^{s,K}; q_t, \hat{y}_t, \pi_t). \tag{4}$$

Each pseudo-class keeps its top-K arrivals by priority. Since stored labels and priorities do not depend on the fused prediction, both memory trajectories are independent of w and can be re-scored offline (Proposition 5). We vary the number of views M, memory capacity K, and retrieval space s, while keeping the predictor, prompts, kernel, admission rule, and stream fixed. The DINOv2 retrieval configuration in Eq. (4) is our MARC (Memory-Augmented Retrieval for CLIP). It removes output fusion from the CLIP cache, negative memory, adaptive fusion, and pseudo-label refinement; all variants are detailed in App. H.

3.3 What the Decoupling Identifies

To state the identification target, first consider a generic cache whose retained population $J_{t-1}^s = (J_{t-1}^s(c))_{c=1}^C$ may itself depend on encoder s. For fixed predictor, kernel, and fusion machinery, write $F_t(s, J) = b_t + wL_t^{s,J}$. Adding and subtracting the matched-population predictor gives the exact decomposition

$$F_t(s, J_{t-1}^s) - F_t(s', J_{t-1}^{s'}) = \underbrace{F_t(s, J_{t-1}^s) - F_t(s', J_{t-1}^s)}_{\text{encoder similarities on matched keys}} + \underbrace{F_t(s', J_{t-1}^s) - F_t(s', J_{t-1}^{s'})}_{\text{retained-population change}}. \tag{5}$$

A conventional composite-system comparison can contain both terms and may additionally change the aggregation machinery. MARC removes the second term by construction: $\hat{y}_t$ and π_t are CLIP functions, so admission is independent of s. Let $I_{t-1}^K(c)$ denote this common retained index set and write $u_{ti}^s = g_s(x_t)^\top g_s(x_i)$.

Proposition 1 (Exact space isolation and prediction stability)**.** *Fix the stream, its augmentation draws, K_C, K, $w \geq 0$, and all tie rules. For any two retrieval spaces s and s':* ***(i)*** *the retained index sets $I_{t-1}^K(c)$ are identical for every t, c;* ***(ii)*** *their cache logits satisfy*

$$L_{t,c}^s = \alpha e^{-\beta} \sum_{i \in I_{t-1}^K(c)} e^{\beta u_{ti}^s}, \quad \left|L_{t,c}^s - L_{t,c}^{s'}\right| \leq D_{t,c}^{s,s'} := \alpha\beta \sum_{i \in I_{t-1}^K(c)} \left|u_{ti}^s - u_{ti}^{s'}\right|. \tag{6}$$

Let $F_t^s = b_t + wL_t^s$, $a_t^s = \arg\max_c F_{t,c}^s$, and $\mathrm{gap}_t^s = F_{t,a_t^s}^s - \max_{c \neq a_t^s} F_{t,c}^s$. Then

$$\mathrm{gap}_t^s > w\left(D_{t,a_t^s}^{s,s'} + \max_{c \neq a_t^s} D_{t,c}^{s,s'}\right) \quad \Longrightarrow \quad a_t^{s'} = a_t^s. \tag{7}$$

Thus $J^s_{t-1} = J^{s'}_{t-1} = I^K_{t-1}$ and the population term in Eq. (5) vanishes exactly. The intervention replaces only the encoder-induced cosine similarities of matched query–key pairs; it cannot improve by retaining easier samples or receiving different pseudo-labels. Spaces inducing the same similarities on these pairs therefore produce identical logits and predictions, irrespective of model name. Equation (7) also localizes where an encoder swap can matter: only samples whose fused decision margin is smaller than the induced kernel perturbation can change prediction. This result does not order spaces or guarantee gain; it establishes that the controlled sweep estimates the effect of replacing the encoder-induced similarity matrix at a fixed cache population. The proof is in App. Q.

Scope of the theoretical results. The decomposition and Proposition 1 are exact identities induced by the update rule and require no distributional assumptions. By contrast, the margin results in Eq. (9) provide query-level sufficient conditions under explicit separation assumptions, while retrieval-space selection based on purity or anti-hubness additionally relies on Assumption 3. These results characterize conditions under which retrieval geometry affects predictions; they do not constitute unconditional performance guarantees.

3.4 Retrieval-Geometry Measurements.

Before adaptation, let S be a common calibration sample of n target-domain images and let $N^s_\kappa(i)$ be the κ nearest neighbors of x_i in space s, excluding i. We measure

$$P_s(\kappa) = \frac{1}{n\kappa} \sum_{i \in S} \sum_{j \in N^s_\kappa(i)} \mathbf{1}[y_j = y_i], \qquad A_s(\kappa) = \frac{1}{n} \sum_{j \in S} \mathbf{1}[d^s_j(\kappa) = 0], \tag{8}$$

where $d^s_j(\kappa) = \sum_{i \in S} \mathbf{1}[j \in N^s_\kappa(i)]$ is k-NN in-degree. P_s is neighborhood purity and requires labels; A_s is the anti-hub fraction and is label-free. The probe radius κ is not the per-class cache capacity K; our tables use κ=10. Purity measures contaminating *edges*, whereas anti-hubness measures the fraction of keys with no incoming edge. They are diagnostics of different failure modes, not surrogate objectives optimized by the method. App. D gives their exact accounting identities and explains the calibration assumptions needed to connect them to an online, priority-filtered cache.

For the reported benchmark diagnostics, S is the complete target test split—for example, n=7,500 on ImageNet-A and n=10,000 on ImageNet-V2—and is the same split used for evaluation rather than a held-out set. Accordingly, labeled purity is an offline benchmarker's diagnostic. Pseudo-purity replaces y_i in Eq. (8) with frozen CLIP predictions and can instead be computed on an unlabeled target calibration batch.

Analysis Quantities. For a query of true class y, the base and cache margins are

$$\mu_t = b_{t,y} - \max_{c \neq y} b_{t,c}, \qquad \delta^{s,K}_t = L^{\mathcal{M}}_y(q_t) - \max_{c \neq y} L^{\mathcal{M}}_c(q_t). \tag{9}$$

For $w \geq 0$, fusion is correct whenever $\mu_t + w\delta^{s,K}_t > 0$; retrieval therefore helps only when its margin aligns with base-model errors. Under a separated-neighborhood condition, the cache-margin floor increases with true local purity and decreases with pseudo-label error and tail mass. Increasing capacity can instead expose a wrong-class key that is more similar than the newly exposed true-class key, lowering the pointwise margin. App. D states these conditional results, their calibration assumptions, and their proofs. We use purity to diagnose attainable signal and anti-hubness to diagnose useful capacity, not as unqualified guarantees.

Controlled Evaluation of Retrieval Spaces. Because cache-logit scale varies across spaces, controlled comparisons use the finite-grid envelope

$$\mathrm{Acc}^\star(s, K) = \max_{w \in \mathcal{W}} \mathrm{Acc}\big(\{b_t + wL^{\mathcal{M}^{s,K}_{t-1}}(q_t)\}^T_{t=1}\big), \tag{10}$$

over $\mathcal{W} = \{0, 0.1, 0.3, 1, 2, 3, 5, 8, 10, 15, 20, 30, 50, 75, 100\}$ and $\mathcal{K} = \{1, 2, 3, 4, 8, 16, 32, 64\}$. Since $0 \in \mathcal{W}$, zero gain is an oracle floor, not evidence that every weight is harmless. For the deployable check, we select w on ImageNet-V2 and apply it to ImageNet-A; this costs at most 0.95 points and preserves the purity ranking. App. G gives the complete transfer protocol.

4 Experiments

✓ **Evaluation Protocol.** We follow TDA (Karmanov et al., 2024) on four ImageNet shifts (A, V2, R, and Sketch) and six cross-domain datasets: Caltech101 (Fei-Fei et al., 2004), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flowers102 (Nilsback & Zisserman, 2008), Pets (Parkhi et al., 2012), and UCF101 (Soomro et al., 2012). Controlled runs use one frozen CLIP ViT-B/16 predictor and fixed prompts. The initial sweep compares CLIP, MAE (He et al., 2022), DINOv2-S/B/L (Oquab et al., 2023), DINO v1 (Caron et al., 2021), BEiT (Bao et al., 2022), and a supervised

ViT (Kolesnikov et al., 2021). We add eight retrieval spaces to cover a wider range of training objectives and architectures: OpenCLIP (Cherti et al., 2023), SigLIP (Zhai et al., 2023), EVA-02 CLIP (Sun et al., 2023), CLIP ViT-L/14, DINOv3-B (Siméoni et al., 2025), DeiT-III (Touvron et al., 2022), ConvNeXt-B (Liu et al., 2022), and a supervised ResNet-50 (He et al., 2016). App. O describes these comparisons. All encoders are frozen. We use three seeds for retrieval-space comparisons and all reported MARC benchmark rows. Runtimes are measured on one RTX 3060 after 100 warm-up samples; App. Q gives the full protocol.

♦ **Implementation Details.** All MARC configurations are training-free, with both prediction and retrieval encoders frozen throughout inference. Auxiliary-encoder groupings report inference resources separately from this designation. For every pair of retrieval spaces on ImageNet-A and ImageNet-V2, stream order, pseudo-labels, base logits, and base predictions are identical. Only the cache logit changes. Thus, the experiments vary only the encoder-induced query–key similarities, as specified in Proposition 1.

✓ **Final Design.** After fixing DINOv2 as the retrieval space, we test three additional components: pseudo-label refinement, a CLIP-indexed negative memory, and entropy-adaptive fusion, cumulatively. On ImageNet-A at $M{=}64$, accuracy drops from 65.25 ± 0.02 to 64.12 ± 0.26, 64.14 ± 0.25, and 62.12 ± 0.17, respectively, with the reduced configuration performing best in every seed. Across four OOD splits at $M{=}1$, the three additions change accuracy by -1.94, $+0.00$, and -0.09 points on average. We therefore retain the simpler read-before-write design: frozen CLIP prediction, a zero-output-weight CLIP admission memory, a frozen DINOv2 retrieval cache, and a single fusion weight. App. H details the tested components and reports all splits.

5 Experimental Findings

We organize our experimental findings into three tiers. *(i) To show why retrieval space matters,* controlled same-trajectory interventions isolate the effect of the retrieval encoder by keeping predictions, admissions, and retained indices identical (Tables 1 and A1). *(ii) To show how retrieval quality translates into practical gains,* we evaluate retrieval-space selection, transfer, robustness, and accuracy–efficiency trade-offs under controlled budgets (Table A30). *(iii) To place our results in a broader benchmark context,* we report comparisons under the protocols used by prior work.

5.1 Tier 1: Effect of Retrieval Geometry on Cache Gain

✓ **Retrieval Space Drives Cache Gain.** Our central experiment asks how much cache gain changes when we vary only the feature space used for retrieval. We use $M{=}1$ so that the cache is the only additional evidence. In the held-out comparison, all spaces use $K{=}8$, with w selected on ImageNet-V2 and evaluated on ImageNet-A. The gains range from $+0.00$ for MAE and CLIP to $+0.45$ for DINOv2-S, $+9.73$ for DINOv2-B, and $+19.16$ for DINOv2-L. Thus, the same cache construction and retained samples can produce very different gains solely because of the retrieval space. Because all sixteen spaces share one stream and one retained population, these differences are paired per image: at the diagnostic weight the DINOv2-L cache gains $+18.19$ $[+17.21, +19.17]$ points over the same no-cache predictor while the CLIP cache loses -10.28 $[-11.38, -9.18]$ (App. R).

✓ **Cache Gain Persists Across Capacities.** Table A1 reports three cache-gain measures: gain at a fixed weight, the best gain over the weight grid, and gain after selecting w on a held-out stream; App. G defines these measures and their use in each column. It also reports pre-adaptation retrieval diagnostics and, at each space's best capacity $K^\star$, both a fixed-w diagnostic and the oracle envelope selected on ImageNet-A. The oracle sweep preserves the same ordering, with attainable gain ranging from $+0.09$ to $+20.06$, while a CLIP cache reaches only $+0.62$ (App. B). DINOv2-S is smaller than CLIP ViT-B/16 yet has more than twice its gain ceiling, showing that model size alone does not explain the effect. Matched-population margin distributions are given in App. E.

✓ **Alternative Retrieval Spaces Remain Weaker.** Modifying CLIP's retrieval geometry closes only a small part of the gap: mean subtraction and whitening reduce hubness and raise CLIP's best gain from $+0.44$ to $+1.07$, still far below DINOv2-L's $+20.06$ (App. B). The pattern also extends beyond a single encoder family: OpenCLIP and SigLIP perform similarly to CLIP, CLIP ViT-L/14 reaches $+5.52$, and ConvNeXt and ResNet show that the effect is not specific to transformer architectures. Overall, cross-modal encoders rank low unless scaled up (Table 1).

✓ **The Ranking Is Robust Across Seeds and Predictors.** Across three coupled seeds, Spearman ρ is 0.959, 0.961, and 0.967; DINOv2-L leads every seed with $G^\star = 19.74 \pm 0.40$ (Table cells show seed 0) and zero selection regret. The result is consistent with Table A1 (20.06 ± 0.34). With a CLIP ResNet-50 predictor, the same intervention gives $\rho{=}0.968$, again selecting DINOv2-L with

Table 1: **Results of controlled same-trajectory intervention, isolating the effect of retrieval space on cache gain (*Tier 1*).** Sixteen retrieval spaces compared under the **same retained samples** (K=8, M=1), with each gain reported as the evaluation-stream oracle envelope $G^\star$ maximized over the weight grid. *Purity* and *anti-hubness* are measured before adaptation. Paired 95% intervals, McNemar tests, and calibration-subsample intervals, are given in App. R.

Cache space	Purity @10 ↑ (A)	Purity @10 ↑ (V2)	AntiHub ↓ (A)	IN-A $G^\star$ ↑	IN-V2 $G^\star$ ↑
Self-supervised (self-distillation)					
DINOv2-L	67.3	48.7	1.67%	+19.36	+3.19
DINOv2-B	48.9	44.4	1.20%	+10.27	+2.12
DINOv3-B	44.6	35.6	0.79%	+6.49	+0.78
DINOv2-S	26.1	33.1	0.69%	+1.24	+0.35
DINO v1	16.8	29.1	1.95%	+0.05	+0.13
Label-supervised					
DeiT-III ViT-B/16	46.1	55.3	2.21%	+8.01	+5.06
Supervised ViT-B/16	35.0	50.8	1.37%	+4.43	+3.48
ConvNeXt-B	31.0	42.1	1.16%	+2.85	+1.63
ResNet-50 (sup)	12.3	35.8	2.73%	+0.04	+1.54
Cross-modal (language-supervised)					
CLIP ViT-L/14	42.8	28.3	5.59%	+5.52	+0.06
EVA-02 CLIP	38.9	33.9	2.28%	+4.15	+0.10
SigLIP ViT-B/16	29.5	29.3	5.43%	+0.99	+0.00
OpenCLIP ViT-B/16	24.3	24.4	3.35%	+0.76	+0.00
CLIP ViT-B/16	29.7	22.9	6.27%	+0.44	+0.00
Masked image modelling					
MAE ViT-B/16	3.2	1.5	3.44%	+0.03	+0.00
BEiT-B/16	19.1	29.0	11.00%	+0.00	+0.00
Spearman ρ(purity, gain)				+0.959	+0.943
Spearman ρ(− AntiHub, gain)				+0.541	+0.756
Selection regret of purity				0.00	0.00

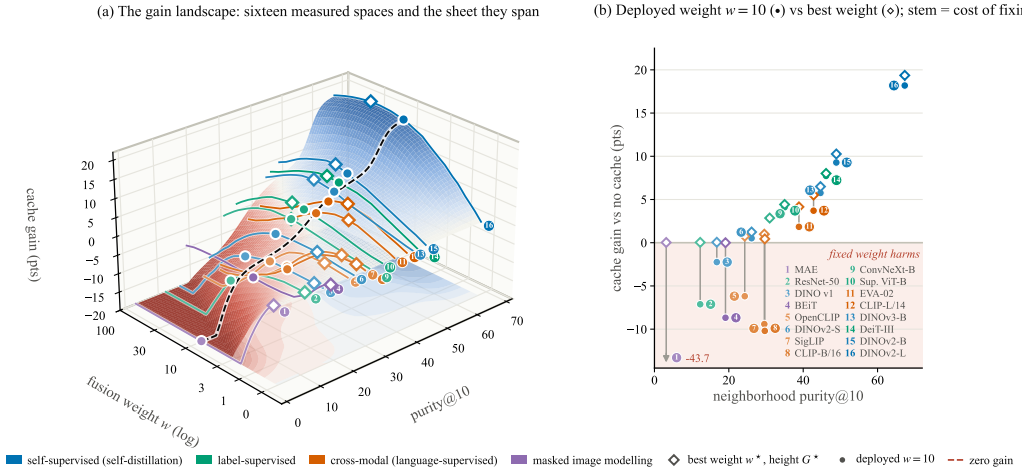


Figure 3: **Cache gain as a function of neighborhood purity and fusion weight on ImageNet-A (*Tier 2*).** High-purity spaces benefit from a larger cache contribution, whereas low-purity spaces degrade when the cache is weighted too strongly. Panel (b) compares the fixed choice w=10 with the best weight for each space. The surface interpolates between discrete encoder points and does not assert a continuous functional dependence.

zero regret and spanning 0.00–19.04 points of gain. Thus, the purity–gain ordering is robust across seeds and predictors.

5.2 Tier 2: Retrieval Quality, Transfer, and Efficiency

✓ **Purity predicts gain, while anti-hubness predicts capacity.** Useful capacity varies by more than an order of magnitude across retrieval spaces (Table A1). Purity strongly ranks attainable gain (ρ=0.90) but is less predictive of $K^\star$ (ρ=0.36), whereas negative anti-hubness ranks $K^\star$ at ρ=0.87 but gain at only ρ=0.50. On held-out ImageNet-V2, the corresponding correlations are 0.98 and 0.97. Calibration-subsample intervals separate both arms of this dissociation on ImageNet-A rather than leaving it as an observed ordering: purity ranks gain at 0.900 $[0.900, 0.900]$ against 0.547

$[0.500, 0.718]$ for anti-hubness, while anti-hubness ranks $K^\star$ at 0.817 $[0.718, 0.872]$ against 0.359 $[0.359, 0.359]$ for purity (App. R). Figure 3 shows the joint effect of purity and fusion weight, while Eq. (9) explains the process.

✓ **Purity predicts cache gain and enables label-free selection.** Across eight spaces, purity ranks cache gain at ρ=0.93 on ImageNet-A and 0.98 on ImageNet-V2, with $\rho \geq 0.80$ on five of six cross-domain streams; EuroSAT is the main exception due to similar candidate purity. Purity-based selection incurs only 0.35 points of mean regret, while pseudo-purity closely matches labeled rankings (ρ=0.98/1.00 on ImageNet-A/V2; 0.93 across eight streams), selecting the best encoder on $5/8$ streams with 1.48 points of mean regret. Jointly selecting s by pseudo-purity and transferring (K, w) from a disjoint labeled stream retains about 90% and 98% of oracle gain on ImageNet-A/V2.

✓ **The trend generalizes across representations, streams, and admission policies.** Projecting DINOv2-B to 32–512 dimensions preserves a ρ=0.94 purity–gain correlation. Across sixteen spaces, the correlation is 0.959 on ImageNet-A and 0.943 on ImageNet-V2; purity selects different winners across streams, confirming that the best retrieval space is stream-dependent. Among harder ImageNet-A choices spanning 4.4–10.0 points of gain, purity still gives ρ=0.943 with zero regret. Excluding zero-gain ties, the correlation rises to 0.971/0.973 on ImageNet-A/V2. Finally, across five admission policies, DINOv2-L maintains an approximately 9.6-point advantage over DINOv2-B, with gains of at least 18.3 and 8.7 points, respectively, under random/FIFO admission. Thus, the retrieval-space effect is robust to both representation scale and cache-admission policy; anti-hubness remains useful primarily for diagnosing capacity.

✓ **Diminishing Cache Gains with More Views.** We sweep $M \in 1, 2, 4, 8, 16, 32, 64$ with no cache, a CLIP cache, or a DINOv2 cache and compute the exact difference-in-differences from Eq. (11). Across three ImageNet-A seeds, DINOv2 cache gain drops from $+9.30 \pm 0.38$ points at one view to $+6.56 \pm 0.10$ at 64 views, giving an interaction contrast of -2.74 ± 0.33, with the same direction in every seed; the one-to-four-view contrast is likewise negative at -2.02 ± 0.28 across all seeds. The repair sets show the same pattern: a single-view cache repairs $1{,}074 \pm 21$ baseline errors versus 878 ± 4 for 64 views, at roughly one seventh of the cost, with 426 ± 14 shared repairs ($48.5 \pm 1.4\%$ overlap). The corresponding regression sets contain 377 ± 12 and 289 ± 3 images, with 80 ± 2 shared, while repair–regression cross-sets have cardinality zero by the disjoint base-error/base-correct partition. App. C provides the exact contrast, full sweeps, warm-up, and more information.

✓ **Accuracy–Efficiency Trade-off.** A single-view DINOv2-B cache exceeds the 64-view no-cache ensemble (60.14 ± 0.38 vs. 58.69 ± 0.09, both over three seeds) at $6.9\times$ lower measured cost. At matched M=8, auxiliary-encoder scale, and hardware, MARC reaches $64.17 \pm 0.31\%$ at 0.070 s/image versus $62.75 \pm 0.15\%$ at 0.181 s/image for COSMIC. This gives a 1.42-point accuracy advantage at $2.6\times$ $[2.52, 2.64]$ lower latency. A matched evaluation of both systems on three shared stream orders yields a per-image paired advantage of $+1.12$ $[+0.66, +1.58]$ points (App. R); App. S reports the complete Pareto table and three timing sessions (Tables A30 and A31).

5.3 Tier 3: Benchmark Comparison

✓ **MARC Improves OOD Accuracy and Transfers Across Domains.** Table 2 compares MARC with CoOp, CoCoOp, TPT, DiffTPT, and other training-free methods under their reported protocols. Under reported protocols, MARC reaches a 53.00 ± 0.08 OOD average with a ResNet-50 predictor and one view, versus 48.20 for the strongest listed training-free baseline; with ViT-B/16, four views reach 66.79. Across six cross-domain datasets, it averages 74.96 at four views and 75.45 at 64 views, with cache gains ranging from -0.48 to $+11.20$ points (App. J, Table A16); on Caltech101 and DTD the paired interval includes zero at both view budgets, so the cache has no detectable effect there (App. R). On two additional streams, the average cache gain falls to $+0.84$ versus $+3.32$ on the original six (Table A17).

✓ **Gains Depend on View Budget and Evaluation Protocol.** For ResNet-50, increasing from one to 64 views adds only 1.29 points, while four views reach 71.48 on ImageNet-val. The dense 50-per-class stream favors cache methods with K=16, so we treat this in-distribution result as protocol-dependent rather than evidence of robustness. Likewise, SOTA comparisons follow each method's reported protocol; notably, the 2.28-point comparison with SCAP uses auxiliary DINOv2-B retrieval for MARC versus a CLIP-only baseline, with an auxiliary-matched comparison reported separately.

6 Discussion

Our findings suggest a broader view of cache-based adaptation: the value of a memory depends not only on what it stores, but also on the geometry through which its contents become accessible. This separation is practically important because retrieval can adapt the evidence available to a frozen

Table 2: **Comparison with SOTA frameworks under their reported protocols.** Top-1 accuracy (%) on ImageNet and its OOD variants, evaluating the overall effectiveness of MARC (*Tier 3*). Prompt sets, view budgets, and implementations follow the cited methods unless marked †. **Bold** denotes the best training-free result overall; underlining denotes the best training-free result using CLIP as the sole visual encoder. Green marks the best result among methods using frozen DINOv2-B. Blocks indicate the visual-encoder resources used by each method. Shaded rows denote MARC, † denotes our runs, and subscripts report standard deviations over three seeds.

Method	ImageNet ↑	IN-A ↑	IN-V2 ↑	IN-R ↑	IN-Sketch ↑	Average ↑	OOD Average ↑
CLIP ResNet-50 backbone							
CLIP-RN50	58.16	21.83	51.41	56.15	33.37	44.18	40.69
Methods with parameter adaptation							
CoOp (Zhou et al., 2022b)	63.33	23.06	55.40	56.60	34.67	46.61	42.43
CoCoOp (Zhou et al., 2022a)	62.81	23.32	55.72	57.74	34.48	46.81	42.82
TPT (Shu et al., 2022)	60.74	26.67	54.70	59.11	35.09	47.26	43.89
DiffTPT (Feng et al., 2023)	60.80	31.06	55.80	58.80	37.10	48.71	45.69
Training-free methods: CLIP visual encoder							
TDA (Karmanov et al., 2024)	61.35	30.29	55.54	62.58	38.12	49.58	46.63
ETTA (Dastmalchi et al., 2025)	61.80	31.17	55.87	62.67	38.33	49.97	47.01
BoostAdapter (Zhang et al., 2024)	—	35.12	56.14	62.66	38.87	—	48.20
Training-free methods: CLIP + frozen DINOv2-B							
MARC(Ours) (DINOv2-B), M=1†	71.48$_{\pm 0.18}$	35.52$_{\pm 0.48}$	58.43$_{\pm 0.17}$	70.16$_{\pm 0.04}$	47.88$_{\pm 0.15}$	56.69$_{\pm 0.03}$	53.00$_{\pm 0.08}$
MARC(Ours) (DINOv2-B), M=64†	**71.92**$_{\pm 0.18}$	**39.07**$_{\pm 0.44}$	**58.98**$_{\pm 0.21}$	**70.45**$_{\pm 0.17}$	**48.65**$_{\pm 0.21}$	**57.82**$_{\pm 0.16}$	**54.29**$_{\pm 0.23}$
CLIP ViT-B/16 backbone							
CLIP-ViT-B/16	66.73	47.87	60.86	73.98	46.09	59.11	57.20
Methods with parameter adaptation							
CoOp (Zhou et al., 2022b)	71.51	49.71	64.20	75.21	47.99	61.72	59.28
CoCoOp (Zhou et al., 2022a)	71.02	50.63	64.07	76.18	48.75	62.13	59.91
TPT (Shu et al., 2022)	68.98	54.77	63.45	77.06	47.94	62.44	60.81
DiffTPT (Feng et al., 2023)	70.30	55.68	65.10	75.00	46.80	62.58	60.65
Training-free methods: CLIP visual encoder							
TDA (Karmanov et al., 2024)	69.51	60.11	64.67	80.24	50.54	65.01	63.89
ZERO+Ens (Farina et al., 2024)	71.17	62.75	65.23	80.75	50.59	66.10	64.83
BoostAdapter (Zhang et al., 2024)	—	64.53	65.51	80.95	51.28	—	65.57
SCAP (Zhang et al., 2025)	—	64.52	64.65	81.68	51.65	—	65.63
TaTa (Zhang et al., 2026)	70.63	61.87	65.37	81.78	52.39	66.41	65.35
Training-free methods: CLIP + frozen DINOv2-B							
COSMIC (Huang et al., 2025) (DINOv2-B), M=8†	**76.39**	62.75$_{\pm 0.15}$	**67.54**	82.39	**57.95**	**69.40**	67.66
MARC(Ours) (DINOv2-B), M=4†	74.53$_{\pm 0.04}$	63.00$_{\pm 0.42}$	65.73$_{\pm 0.08}$	82.18$_{\pm 0.09}$	56.26$_{\pm 0.04}$	68.34$_{\pm 0.08}$	66.79$_{\pm 0.10}$
MARC(Ours) (DINOv2-B), M=64†	74.84$_{\pm 0.07}$	**65.25**$_{\pm 0.02}$	66.50$_{\pm 0.09}$	**83.15**$_{\pm 0.02}$	56.76$_{\pm 0.15}$	69.30$_{\pm 0.04}$	**67.91**$_{\pm 0.06}$

foundation model without parameter updates, making it attractive for changing visual environments where repeated retraining is costly or unavailable. Across domains, cache contributions range from -0.48 to $+11.20$ points, indicating greater utility when the target shift leaves correction headroom and incoming data provide complementary neighborhood structure. This setting is relevant to remote sensing, scientific imaging, robotics, and other streaming applications where recent observations can provide useful local evidence. Retrieval also changes compute allocation: its gain decreases from $+9.30$ points with one view to $+6.56$ with 64 views, suggesting partial overlap between retrieval and augmentation and favoring retrieval under constrained inference budgets.

Purity and anti-hubness provide complementary diagnostics of gain and useful capacity, while closely matched spaces motivate richer label-free selection criteria. Our evaluation spans two predictor backbones, diverse retrieval spaces, four ImageNet shifts, and multiple cross-domain datasets; broader model families, modalities, and non-stationary streams offer natural extensions. These findings suggest establishing retrieval quality before allocating additional compute to large view ensembles or complex adaptation machinery. Ultimately, our results position retrieval geometry as a practical interface between a fixed foundation model and a changing world, providing a path toward adaptation that is lightweight, modular, and responsive to the structure of incoming data.

7 Conclusion

Cache-based CLIP adaptation typically uses the same representation for prediction and retrieval, despite their different geometric requirements. We show that changing only the retrieval space can shift cache gain from nearly zero to about 20 points, with consistent effects across seeds, architectures, scales, and admission policies. Stronger retrieval also reduces the value of additional views, suggesting that better memory geometry can outperform added inference complexity. Future work should study label-free retrieval selection across shifts, modalities, and applications such as remote sensing and scientific imaging. More broadly, retrieval space should be treated as a first-order design dimension for memory-based adaptation, particularly in changing visual environments where retraining is impractical, opening a new direction for building memory-based adaptation systems.

LLM USAGE DISCLOSURE

The authors used large language model (LLM)-assisted tools for non-experimental support tasks during research and manuscript preparation. Specifically, these tools were used for grammar checking, language editing, code debugging, figure ideation, and visualization assistance. In addition, agentic review tools, including the ICLR-provided PAT framework and Stanford Agentic AI tools, were used to support manuscript review and revision by identifying potential issues, suggesting improvements, and providing additional critical feedback on the presentation and clarity of the work. All AI-assisted suggestions, reviews, and outputs were manually evaluated by the authors and, where relevant, independently tested or verified before incorporation. The authors retain full responsibility for the integrity and correctness of the work and for the interpretation of all reported results and analyses.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8334–8343, 2022. doi: 10.1109/CVPR52688.2022.00816.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Hamidreza Dastmalchi, Aijun An, and Ali Cheraghian. Etta: Efficient test-time adaptation for vision-language models through dynamic embedding updates. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025.

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

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, 2004. doi: 10.1109/CVPR.2004.383.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2704–2714, 2023.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 142772–142798. Curran Associates, Inc., 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Xuefeng Hu, Ke Zhang, Min Sun, Albert Chen, Cheng-Hao Kuo, and Ram Nevatia. Bafta: Backprop-free test-time adaptation for zero-shot vision-language models. *arXiv preprint arXiv:2406.11309*, 2024.

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9772–9781, 2025.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient Test-Time Adaptation of Vision-Language Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171. IEEE Computer Society, 2024. doi: 10.1109/CVPR52733.2024.01343.

Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 6028–6039, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of machine learning research*, 11(86):2487–2531, 2010.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 2022. ISBN 9781713871088.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Oriane Siméoni, Huy V Vo, Vasil Khalidov, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, pp. 516–533. Springer, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100.

Chenyu Zhang, Kunlun Xu, Zichen Liu, Yuxin Peng, and Jiahuan Zhou. Scap: Transductive test-time adaptation via supportive clique-based attribute prompting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30032–30041, 2025.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: test time robustness via adaptation and augmentation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022a.

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 *Computer Vision – ECCV 2022*, pp. 493–510. Springer Nature Switzerland, 2022b.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *Advances in Neural Information Processing Systems*, 37:67795–67825, 2024.

Yi Zhang, Chun-Wun Cheng, Angelica I. Aviles-Rivero, Zhihai He, and Liang-Jie Zhang. Training-free test-time adaptation with brownian distance covariance in vision-language models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12782–12786, 2026. doi: 10.1109/ICASSP55912.2026.11462803.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vision*, 130(9):2337–2348, 2022b. ISSN 0920-5691. doi: 10.1007/s11263-022-01653-1. URL https://doi.org/10.1007/s11263-022-01653-1.

Appendix

Table of Contents

A More Details on Related Work **16**

A.1 Cache based Adaptation 16

A.2 Self supervised Representations as Retrieval Spaces 16

A.3 Geometry of Nearest neighbor Retrieval 16

A.4 Prompt based and Training free CLIP Adaptation 17

B Retrieval-Space Comparison: Additional Controls **17**

B.1 Single versus Parallel Caches 17

B.2 Why DINOv2 Is Not a Zero-Shot Predictor 17

C Diminishing Marginal Cache Gain: Additional Results **18**

C.1 Results Across OOD Splits 19

C.2 Joint View-Budget Sweeps 19

C.3 Cache Warm-Up over the Stream 19

C.4 Interaction with a ResNet-50 Predictor 20

C.5 Oracle Coverage Analysis 20

D Geometry and Capacity: Additional Results **20**

D.1 Estimators and conditional interpretation 21

E Encoder-Induced Retrieval Similarities at Query Level **22**

F Limits of a Composite Geometry Diagnostic **24**

F.1 EuroSAT Failure Case 24

F.2 Limits of Cross-Stream Prediction 25

G Held-Out Fusion-Weight Transfer **26**

H Component Ablations Across OOD Splits **26**

I Full Rank-Correlation Results **27**

J Additional Benchmark Baselines **28**

K Cross-Domain Retrieval-Space Rankings **30**

L Qualitative Examples **31**

M Cache Capacity at a Large View Budget **32**

N Additional ViT-B/16 Pretraining Regimes **32**

O The Full Sixteen-Space Candidate Set **34**

P Hyperparameter Sensitivity **35**

Q Full Experimental Protocol and Isolation Proof **36**

R Confidence Intervals for the Reported Gains **37**

S Timing Reproducibility **40**

T Open Questions and Future Research Directions **41**

Table conventions. The following apply to every table in the paper and appendix; where a caption gives a table-specific meaning, the caption takes precedence.

Green	Best entry in its column or block.
Green / Red	Favourable / unfavourable value, e.g. a gain versus a loss, or a strong versus a weak correlation.
$XX.XX_{\pm Y.YY}$	Mean with standard deviation across seeds.
↑ / ↓	Higher / lower is better.
†	Our runs of the method.
—	Not applicable or not reported.

A More Details on Related Work

A.1 Cache based Adaptation

Test time adaptation (TTA) traditionally adapts a model to the target stream by updating its parameters, for example through entropy minimization (Wang et al., 2021) or source free hypothesis transfer (Liang et al., 2020). Cache based methods offer a different paradigm: they keep the pretrained model frozen and perform adaptation through an external memory constructed from target examples. Tip Adapter (Zhang et al., 2022b) establishes a representative formulation of this approach. It stores image features and their labels in a key value cache, retrieves examples similar to the current query, and combines the retrieved evidence with the zero shot CLIP logits. TDA (Karmanov et al., 2024) extends this formulation to the online, unlabeled setting. Instead of relying on labeled support examples, it populates the cache with entropy prioritized pseudo labels from the incoming test stream and maintains separate positive and negative memories for each class. Our method adopts TDA's cache construction and update rule, allowing us to focus specifically on how the choice of retrieval space affects adaptation.

Several subsequent methods improve other components of the cache adaptation pipeline. These include regional bootstrapping (Zhang et al., 2024), online per class Gaussian modeling (Han et al., 2025), clique based attribute prompting (Zhang et al., 2025), SVD based clustering (Hu et al., 2024), and Brownian distance covariance (Zhang et al., 2026). COSMIC (Huang et al., 2025), in particular, combines CLIP and DINOv2 caches through dual semantic graphs and clique guided hyper class queries. While these approaches modify cache construction, representation, or aggregation jointly, our goal is more controlled: we isolate retrieval geometry by varying only the encoder that defines similarity, while keeping cache construction and admission fixed.

A.2 Self supervised Representations as Retrieval Spaces

The representation used for retrieval directly determines which cached examples are considered neighbors. Self supervised vision models therefore provide a useful family of alternative retrieval spaces. DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2023) learn representations through cross view consistency and exhibit strong transfer to tasks that depend directly on image to image similarity, including nearest neighbor classification and retrieval. Their embeddings can consequently induce neighborhood structures that differ substantially from those produced by CLIP's vision encoder.

MAE (He et al., 2022), in contrast, learns representations through masked image reconstruction and generally exhibits weaker k NN structure. This difference gives us a useful testbed for separating representation quality from the mechanics of cache adaptation. Rather than treating the retrieval encoder as an interchangeable implementation detail, we study how encoder induced similarity structures affect cache gain, the number of useful augmented views, and the effective memory capacity.

A.3 Geometry of Nearest neighbor Retrieval

The effectiveness of a cache depends not only on what it stores, but also on the geometry that determines which entries are retrieved. Neighborhood purity and hubness (Radovanovic et al., 2010) are established measures for characterizing k NN behavior in high dimensional representation spaces. High purity indicates that nearby samples tend to share semantic labels. Hubness captures an orthogonal phenomenon in which a small subset of points appears in many neighbor lists, while anti hubs are rarely or never retrieved.

Prior adaptation work touches on these geometric properties without explicitly treating retrieval space as an experimental variable. BaFTA (Hu et al., 2024), for example, uses k NN accuracy to assess the quality of a CLIP projection. Cache based TTA methods also show that increasing memory capacity introduces a tradeoff between greater sample diversity and increased pseudo label contamination (Han et al., 2025; Karmanov et al., 2024). LAME (Boudiaf et al., 2022) further demonstrates that neighborhood structure itself can support adaptation by constructing a k NN affinity graph over test features without updating model parameters. However, these works largely operate within a fixed feature space. We instead compare retrieval spaces explicitly, characterize them through neighborhood purity and anti hubness, and test whether these geometric properties predict both cache gain and useful memory capacity.

A.4 Prompt based and Training free CLIP Adaptation

Test time augmentation provides another mechanism for extracting useful information from a single target image. Augmentation has long been used to improve robustness under distribution shift (Hendrycks et al., 2020), and recent TTA methods use multiple augmented views as an explicit adaptation signal. MEMO (Zhang et al., 2022a), for instance, adapts predictions by minimizing the marginal entropy across augmented views. In the vision language setting, TPT (Shu et al., 2022) and DiffTPT (Feng et al., 2023) use view ensembles for test time prompt optimization, while ZERO (Farina et al., 2024) explores training free adaptation from multiple views.
Across these methods, the number of augmented views is typically treated primarily as a protocol or computational choice. In a cache based system, however, additional views can also influence the reliability of pseudo labels and the evidence entering the memory. We therefore study view budget as part of the retrieval problem and examine whether the benefit of additional views changes systematically with the effectiveness of the underlying retrieval space.

B Retrieval-Space Comparison: Additional Controls

Table A1: Tier (i), controlled same-trajectory comparison on ImageNet-A with a CLIP ViT-B/16 predictor, $M{=}1$, and a 50.84% no-cache baseline; only encoder-induced query–key similarities change. Purity and anti-hubness are measured at $K{=}10$ before adaptation, and fixed-w entries are separate diagnostic runs. $G^\star$ is the evaluation-stream oracle envelope, not a deployable result. Oracle subscripts denote standard deviations over three seeds; transfer uses $K{=}8$ and weights selected on ImageNet-V2, averaged over three seeds. Table A14 reports a second, independent three-seed sweep of the same protocol, differing only in that the stream permutation is drawn from an explicit generator rather than the global RNG; the two means agree within seed dispersion ($+20.06 \pm 0.34$ here versus $+19.74 \pm 0.40$ there for DINOv2-L; $+9.88 \pm 0.38$ versus $+9.97 \pm 0.29$ for DINOv2-B).

Cache space	Purity @10 ↑	AntiHub ↓	$K^\star$	w	accuracy (G_{10})	$G^\star$ ↑	G^{tr} ↑
Control spaces							
MAE ViT-B/16	3.2	3.4	1	0.1	32.93 (-17.91)	$+0.09_{\pm 0.07}$	$+0.00$
CLIP ViT-B/16	29.7	6.3	2	1	46.84 (-4.00)	$+0.62_{\pm 0.10}$	$+0.00$
Self-supervised spaces							
DINOv2-S	26.1	0.7	16	3	50.40 (-0.44)	$+1.33_{\pm 0.09}$	$+0.45$
DINOv2-B	48.9	1.2	8	15	60.01 ($+9.17$)	$+9.88_{\pm 0.38}$	$+9.73$
DINOv2-L	67.3	1.7	8	20	68.99 ($+18.15$)	$+20.06_{\pm 0.34}$	$+19.16$
Span, weakest to strongest						19.97	19.16

B.1 Single versus Parallel Caches

Every row of Tab. A1 uses the named representation as the *only* cache. A parallel-cache pipeline instead gives the CLIP arm 51.59 at $K{=}2$; because this is a different system, the controlled comparisons use the single-cache result.

B.2 Why DINOv2 Is Not a Zero-Shot Predictor

The converse experiment DINOv2 for zero-shot prediction is not directly defined because DINOv2 has no language supervision or text-aligned head. An unlearned projection into CLIP text space would evaluate the map rather than the encoder; a fixed random projection obtains 0.44%, approximately chance.

A better-defined test discards CLIP logits at query time and scores only the cache. Table A2 shows strong dependence on encoder scale.

Table A2: Cache-only prediction on ImageNet-A at $M{=}1$, with CLIP logits discarded at query time. CLIP still supplies stored pseudo-labels, so this is not an independent classifier.

Cache scored alone	Top-1 ↑	vs. no-cache prompt ensemble
No-cache, prompt ensemble	50.84	—
CLIP cache	27.00	−23.84
DINOv2-B cache	49.40	−1.44
DINOv2-L cache	64.20	+13.36

CLIP still supplies the pseudo-labels, while DINOv2 supplies the image neighborhoods; the result supports their complementary roles rather than a universally stronger encoder.

C Diminishing Marginal Cache Gain: Additional Results

Let $A_{M,r}$ be accuracy at view budget M with retrieval indicator $r \in \{0,1\}$. For $M_1 < M_2$, our exact interaction contrast is

$$\begin{aligned}\mathrm{Sub}(M_1, M_2) &= [A_{M_2,1} - A_{M_2,0}] - [A_{M_1,1} - A_{M_1,0}] \\ &= [A_{M_2,1} - A_{M_1,1}] - [A_{M_2,0} - A_{M_1,0}].\end{aligned} \tag{11}$$

Values below zero mean that retrieval contributes less after increasing the view budget. The contrast is an algebraic difference in differences that summarizes the measured interaction.

Measurement. For fixed stream, retrieval space, capacity, and weight, let $\Delta_M = \mathrm{Acc}_{\mathrm{cache}}(M) - \mathrm{Acc}_0(M)$. Then $\mathrm{Sub}(M_1, M_2) = \Delta_{M_2} - \Delta_{M_1}$: values below, at, and above zero represent decreasing, unchanged, and increasing marginal cache gain, respectively.

Table A3: ImageNet-A accuracy across view budgets and cache spaces (CLIP ViT-B/16). Δ is the fixed-weight gain G_w of the DINOv2-B cache over the same no-cache ensemble; the shaded row uses one view. The bottom-right entry is $\mathrm{Sub}(1, 64)$ from Eq. (11). This table shows the complete seed-0 budget sweep; Table A4 reports three-seed replication at the anchor budgets. Costs are reported in Table A30.

M	no cache ↑	+CLIP cache ↑	+DINOv2 cache ↑	$G_w{=}\Delta$ ↑
1	50.84	49.59	59.73	+8.89
2	53.76	52.85	61.52	+7.76
4	55.77	55.23	62.93	+7.16
8	57.20	56.20	63.96	+6.76
16	59.03	57.85	64.40	+5.37
32	58.21	57.69	64.72	+6.51
64	58.77	58.21	65.23	+6.45
Δ, $M{=}1 \to 64$	+7.93	+8.62	+5.50	−2.44

Table A4: Three-seed replication of the ImageNet-A augmentation–retrieval measurements. Cache gains are $G_w(M)$ and $\mathrm{Sub}(1, M) = G_w(M) - G_w(1)$. Repair counts compare the 64-view ensemble and the single-view DINOv2-B cache with the single-view predictor.

Quantity	seed 0	seed 1	seed 2	mean ± sd
$G_w(1)$	+8.89	+9.65	+9.35	$+9.30 \pm 0.38$
$G_w(4)$	+7.16	+7.60	+7.06	$+7.27 \pm 0.29$
$G_w(64)$	+6.45	+6.56	+6.66	$+6.56 \pm 0.10$
$\mathrm{Sub}(1, 4)$	−1.73	−2.05	−2.29	-2.02 ± 0.28
$\mathrm{Sub}(1, 64)$	−2.44	−3.09	−2.69	-2.74 ± 0.33
Augmentation repairs	881	880	873	878 ± 4
Retrieval repairs	1,057	1,097	1,069	$1,074 \pm 21$
Shared repairs	425	440	412	426 ± 14
Shared / augmentation	48.2%	50.0%	47.2%	$48.5 \pm 1.4\%$

Repair-set accounting. Relative to the one-view no-cache predictor f_0, define

$$E_0 = \{t : f_0(x_t) \neq y_t\}, \quad R_u = \{t \in E_0 : f_u(x_t) = y_t\}, \quad B_u = \{t \notin E_0 : f_u(x_t) \neq y_t\}.$$

Then $\mathrm{Acc}(f_u) - \mathrm{Acc}(f_0) = (|R_u| - |B_u|)/T$. Across three seeds, retrieval repairs $1{,}074 \pm 21$ baseline errors versus 878 ± 4 for augmentation, while their intersection contains 426 ± 14 images ($48.5 \pm 1.4\%$ of augmentation repairs). Together with the gain contrasts, these counts provide directional evidence that the two sources exploit partially shared errors.

Table A5: Joint repair and regression accounting on ImageNet-A over three seeds. The reference is the one-view cache-free predictor; augmentation uses 64 views, and retrieval uses a one-view DINOv2-B cache. Values are mean ± standard deviation.

Set	Augmentation	Retrieval	CLIP cache	Aug. ∩ retrieval
Repairs $\lvert R_u \rvert$	878 ± 4	$1{,}074 \pm 21$	378 ± 30	426 ± 14
Regressions $\lvert B_u \rvert$	289 ± 3	377 ± 12	429 ± 16	80 ± 2

The mean repair partition contains 452 augmentation-only, 648 retrieval-only, and 426 shared images. By definition, $R_u \subseteq E_0$ and $B_u \subseteq E_0^c$, so $|R_{\text{aug}} \cap B_{\text{ret}}| = |R_{\text{ret}} \cap B_{\text{aug}}| = 0$. The identity $(|R_u| - |B_u|)/T$ recovers the measured accuracy change for each source.

C.1 Results Across OOD Splits

Table A6 repeats the comparison on all four OOD splits; cache gain decreases from one to four views on each.

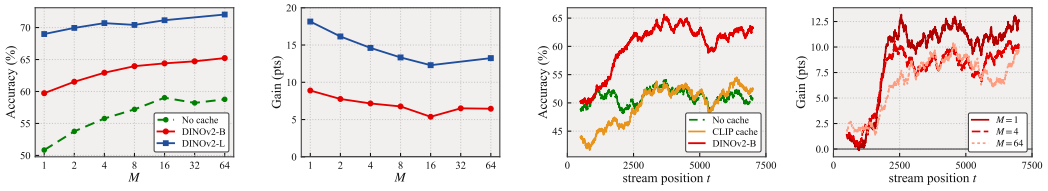
Table A6: Single-run DINOv2-B cache gain at one and four views across OOD splits (CLIP ViT-B/16). Cells show no cache → +DINOv2; $\Delta_M = G_w(M)$ is fixed-weight gain, and *shrink* is $\Delta_4 - \Delta_1$, which is negative on every split.

Split	$M{=}1$	$M{=}4$	$G_w(1)\uparrow$	$G_w(4)\uparrow$	shrink
IN-A	50.84 → 59.73	55.77 → 62.93	+8.89	+7.16	−1.73
IN-Sketch	49.27 → 55.49	50.63 → 56.23	+6.22	+5.60	−0.62
IN-R	77.37 → 81.34	79.07 → 82.12	+3.97	+3.05	−0.92
IN-V2	63.66 → 65.25	64.49 → 65.64	+1.59	+1.15	−0.44

Gain follows shift severity more closely than the CLIP-to-DINOv2 purity gap. ImageNet-V2 has the largest gap but smallest gain, plausibly because its predictor already reaches 63.66%.

C.2 Joint View-Budget Sweeps

Figure A1 sweeps the view budget for two cache spaces. Both gains decline overall, giving the same directional pattern, with the stronger space starting higher and losing more as views increase.



(a) Accuracy vs view budget. (b) Cache gain vs view budget. (c) Warm-up over the stream ($M = 1$). (d) Warm-up per view budget.

Figure A1: View budget and cache warm-up on ImageNet-A, at the fixed weights of Tab. A1. (a) One-view DINOv2-L reaches 68.99, above 64-view DINOv2-B (65.25) and no cache (58.77). (b) Three-seed mean cache gain changes $9.30 \to 6.56$ for DINOv2-B and $18.15 \to 13.24$ for DINOv2-L. (c) At $M{=}1$ the DINOv2-B arm climbs 13.9 points from the first to the last decile of the stream while the no-cache control moves $+0.9$; the climb is the cache being built. (d) The same climb shrinks as views are added, consistent with augmentation–retrieval overlap inside the stream.

C.3 Cache Warm-Up over the Stream

Because aggregate accuracy mixes retrieval quality with cache maturity, we also examine gain over the stream (Fig. A1, panels c and d).

This warm-up also clarifies the ImageNet-val caveat in Sec. 5.3. With 50 images per class and K=16, the cache is populated for most of the stream; all cache methods benefit from this density, and high-purity retrieval uses it more effectively.

C.4 Interaction with a ResNet-50 Predictor

With a CLIP ResNet-50 predictor on ImageNet-A, 64 views improve accuracy from 24.12 to 28.97 (+4.85), whereas a single-view DINOv2-B cache reaches 34.99 (+10.87); the full configuration reaches 39.07. The sweep uses one run; the replicated single-view benchmark result is 35.52 ± 0.48.

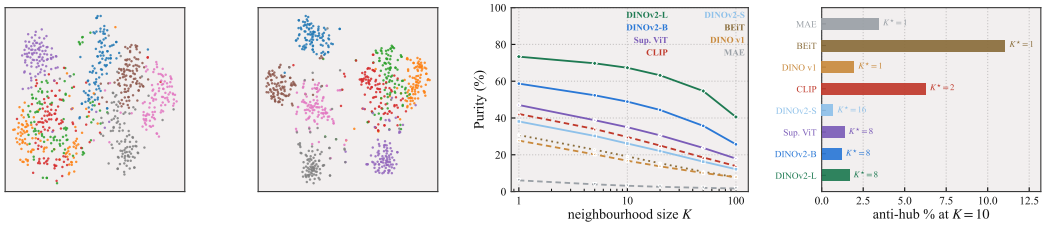
Scope of the ± 0.4 band. The dispersion quoted throughout is measured at M=64, where the cache-only configuration on ImageNet-A spans 0.04 points across three seeds. It does not apply uniformly at M=1. On ImageNet-A, which has 7,500 images, the cache-only configuration spans 0.76 points across seeds, and the configurations that add pseudo-label refinement span 1.08–1.61. Refinement is therefore itself a source of run-to-run variance, consistent with its negative mean effect, and comparisons at M=1 on this split should be read against the wider band. This does not affect the conclusion of Sec. 4, because the cache-only configuration is the most accurate in every seed.

C.5 Oracle Coverage Analysis

An oracle retaining every prediction that either source gets right would reach 71.01%, versus 65.23% for their actual combination. For the CLIP cache, the corresponding values are 64.99% and 58.21%. Because these gaps mix source interactions with regressions on previously correct images, the main paper uses the directly observed 48% repair-set overlap.

D Geometry and Capacity: Additional Results

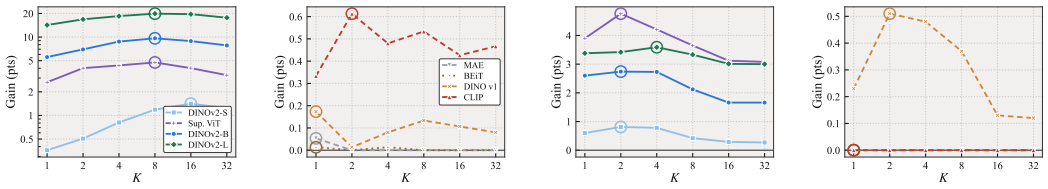
Figure A2 visualizes the measurements from Sec. 3.4; Fig. A3 gives the associated capacity-response curves. The t-SNE panels use the same 800 images from eight classes, with perplexity 30, PCA initialization, and seed 0. Their quoted purity is within those classes and is not comparable to the full 200-class measurement.



(a) CLIP: classes interleave. (b) DINOv2-B: classes separate. (c) Purity, eight spaces, 200 classes. (d) Dead capacity vs usable $K^\star$.

Figure A2: Retrieval geometry on ImageNet-A before adaptation. The t-SNE projections illustrate the cleaner DINOv2-B neighborhoods; the unprojected full-feature result is the purity curve over all eight retrieval spaces, including BEiT, DINO v1, and the supervised ViT. CLIP has 6.3% anti-hubs and $K^\star$=2, versus 0.7–1.7% and $K^\star$=8–16 for DINOv2; BEiT is the extreme case, 11.1% anti-hubs and $K^\star$=1, while DINO v1 (2.0%) and the supervised ViT (1.4%) sit between CLIP and the DINOv2 family.

Tab. A1 reports each optimum $K^\star$; Fig. A3 shows the full curves.



(a) ImageNet-A, strong spaces. (b) ImageNet-A, weak spaces. (c) ImageNet-V2, strong spaces. (d) ImageNet-V2, weak spaces.

Figure A3: Full capacity response: all eight retrieval spaces on both ImageNet streams. Rings mark each space's optimum $K^\star$, the summary statistic reported for each curve in Tab. A1. Curves are weight-swept envelopes, so a space that cannot be scaled usefully sits at $+0.00$ rather than going negative; this isolates capacity behaviour from the scaling failure that dominates a common-weight view. Each stream is split into its four strong and four weak spaces so no curve hides at the bottom of another's scale; the strong ImageNet-A panel uses a log ordinate because DINOv2-L spans twenty points. The optimum shifts with the stream, not only with the space: every DINOv2 space peaks lower on ImageNet-V2 than on ImageNet-A.

Beyond the optimum, more memory can reduce accuracy. At $K{=}64$, the CLIP cache's best weight is 0, so the oracle ignores it.

D.1 Estimators and conditional interpretation

On a shared calibration set S of n images, let $N^s_\kappa(i)$ be the κ nearest neighbors of i in space s, excluding i, and let $d^s_j = |\{i : j \in N^s_\kappa(i)\}|$. We compute

$$P_s(\kappa) = \frac{1}{n\kappa} \sum_{i \in S} \sum_{j \in N^s_\kappa(i)} \mathbf{1}[y_j = y_i], \qquad A_s(\kappa) = \frac{1}{n} \sum_{j \in S} \mathbf{1}[d^s_j = 0]. \tag{12}$$

The probe radius κ is distinct from cache capacity K. Purity requires labels; anti-hubness does not. The graph has exactly $n\kappa$ edges and $n\kappa(1 - P_s)$ cross-class edges, while $n(1 - A_s)$ points receive any edge. Thus purity measures contamination and anti-hubness measures how broadly retrieval mass is distributed.

For query t, let μ_t and $\delta^{s,K}_t$ be the base and cache margins from Eq. (9). If an influential band contains m_t entries, has true-label purity p°_t, pseudo-label error η_t, minimum within-band weight ρ_t, and tail mass bounded by $N_t\varepsilon_t$, then

$$\delta^{s,K}_t \geq \alpha\big\{m_t\big[(1+\rho_t)[p^\circ_t - \eta_t]_+ - 1\big] - N_t\varepsilon_t\big\}. \tag{13}$$

Corollary 2 (Sufficient local-purity threshold)**.** *A sufficient condition for a positive cache margin is*

$$p^\circ_t - \eta_t > \frac{1}{1+\rho_t} + \frac{N_t\varepsilon_t}{m_t(1+\rho_t)}. \tag{14}$$

These are conditional, query-level statements, not performance guarantees.

Assumption 3 (Calibration representativeness)**.** For each space s, the mean true-label purity $\bar{p}_s$ of influential online neighborhoods satisfies $|\bar{p}_s - P_s(\kappa)| \leq \xi_s$, where ξ_s absorbs sampling, online-prefix, per-class truncation, priority-selection, and probe-versus-kernel discrepancies.

This is the empirical link tested by rank correlation, held-out transfer, and the EuroSAT failure case; it is not implied by the kernel.

Capacity is non-monotone. If increasing capacity exposes weights a^{K+1}_c for class c, and $c^\star_K$ is the strongest wrong class at capacity K, then

$$\delta^{s,K+1}_t - \delta^{s,K}_t \leq \alpha\big(a^{K+1}_y - a^{K+1}_{c^\star_K}\big). \tag{15}$$

The margin can fall when the newly exposed confuser is more similar than the new true-class key. Anti-hubness therefore diagnoses potentially wasted capacity; it does not directly estimate $K^\star$ or select a retrieval space.

E ENCODER-INDUCED RETRIEVAL SIMILARITIES AT QUERY LEVEL

The aggregate measurements above are reflected in individual margins, retrieved keys, and kernel weights. Statistics use the full embedding; projections are visual only, and matched panels use the same retained indices.

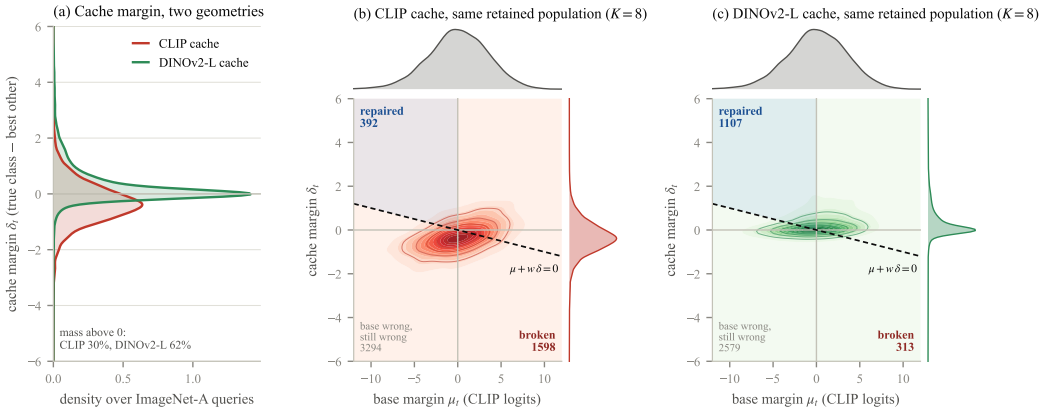


Figure A4: Base and cache margins on ImageNet-A for matched CLIP and DINOv2-L caches at K=8. Fusion at w=10 is correct above $\mu + w\delta = 0$; DINOv2-L places many more base errors in the correctable region.

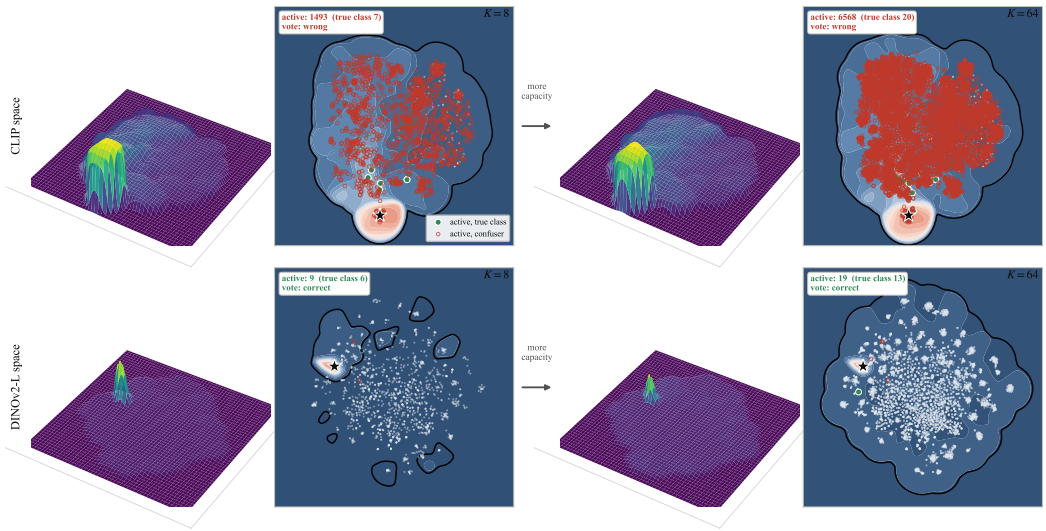


Figure A5: Kernel weights for one ImageNet-A query that CLIP retrieval misses and DINOv2-L retrieval corrects. CLIP activates 1,493 of 1,586 keys and is dominated by confusers; DINOv2-L activates nine keys, six with the true label.

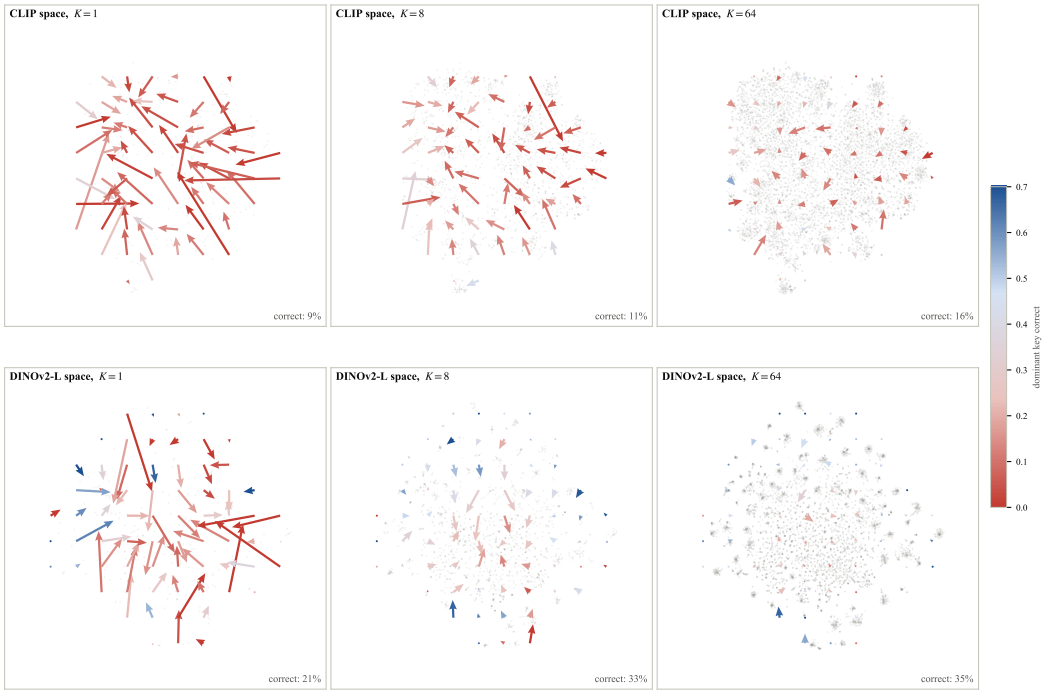


Figure A6: Dominant keys for ImageNet-A base errors. The strongest key has the true label for 11% of queries in CLIP space and 33% in DINOv2-L space; increasing K makes retrieval local in both spaces but does not repair CLIP's label disagreement.

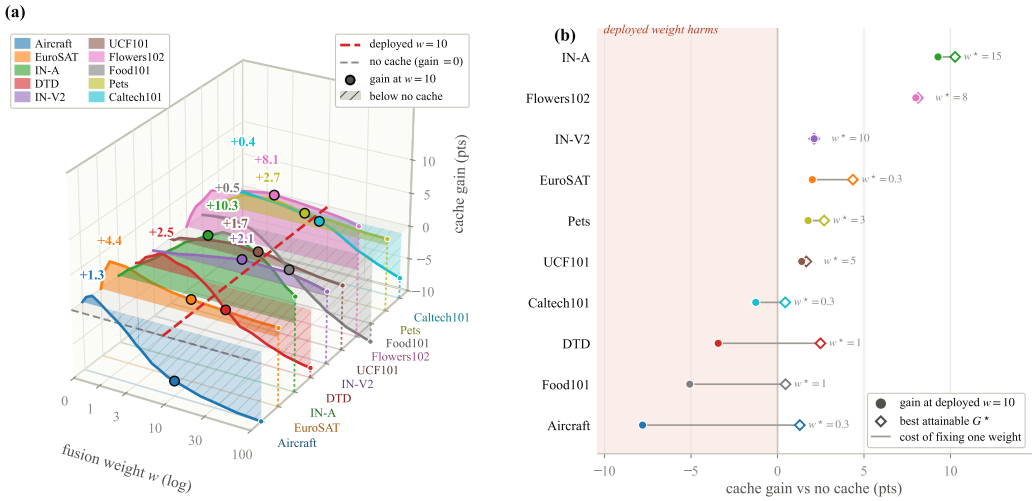


Figure A7: Cache gain across fusion weights on ten streams. The selected w=10 lies near the best region on the ImageNet shifts but is too large for Aircraft and Food101, motivating the held-out transfer analysis below.

Together, these views show the same mechanism at four scales: label-consistent encoder-induced similarities create positive cache margins by concentrating weight on suitable neighbors, while diffuse similarities spread weight across confusers and can amplify base errors.

F LIMITS OF A COMPOSITE GEOMETRY DIAGNOSTIC

This negative result bounds the rule in Sec. 5.2. On ImageNet-A, purity reverses one pair: DINOv2-S has lower purity than CLIP (26.1 vs. 29.7) but a higher ceiling (+1.33 vs. +0.62). Its lower anti-hub fraction (0.7% vs. 6.3%) motivates a hubness-discounted purity score, which perfectly orders the five spaces on ImageNet-A (ρ=1.00). On ImageNet-V2, however, it falls to 0.87, below purity alone at 0.98. We therefore discard the composite as overfit to a single inversion.

The fitted family

$$\mathcal{C}_\lambda(s) = P_s(\kappa) - \lambda A_s(\kappa), \qquad \lambda \geq 0, \tag{16}$$

can change order only when a pair of affine scores crosses. Five spaces therefore permit at most $\binom{5}{2} + 1 = 11$ rankings. Selecting λ to repair the one ImageNet-A inversion and evaluating on the same points gives an in-sample ρ=1.00; without refitting on ImageNet-V2 it falls to 0.87, below purity alone at 0.98. We discard the composite rather than report this overfit improvement.

Definition 4 (Resolution-aware separation)**.** If $s_{(1)}$ has the largest observed purity $\widehat{P}_s$ and e_s are simultaneous measurement-error radii, its lead is resolved when

$$\widehat{P}_{s_{(1)}} - \widehat{P}_s > e_{s_{(1)}} + e_s \quad \text{for every } s \neq s_{(1)}. \tag{17}$$

We use B=300 calibration-subsample bootstrap replicates at each $n \in \{250, 500, 1000, 2000\}$, resampling purity against the fixed measured gains. Table A7 reports central 95% intervals for ρ and the frequency with which the selected encoder attains zero regret. ImageNet-A selection is exact at every tested size; ImageNet-V2 reaches probability 0.99 at n=1000 and 1.00 at n=2000.

Table A7: Calibration-subsample uncertainty over sixteen retrieval spaces using B=300 replicates. Intervals are central 95% intervals; Pr(Reg=0) is the fraction of replicates selecting an oracle-gain leader.

Stream	n	ρ [95% interval]	Pr(Reg=0)
ImageNet-A	250	0.953 [0.925, 0.974]	1.00
	500	0.959 [0.944, 0.971]	1.00
	1000	0.959 [0.950, 0.968]	1.00
	2000	0.957 [0.956, 0.962]	1.00
ImageNet-V2	250	0.749 [0.420, 0.932]	0.60
	500	0.871 [0.753, 0.951]	0.88
	1000	0.917 [0.864, 0.952]	0.99
	2000	0.937 [0.906, 0.958]	1.00

Evaluation-blind joint selection. We select s using pseudo-purity on the unlabeled target batch, select (K, w) on a disjoint labeled calibration stream, and transfer those hyperparameters unchanged to the evaluation stream. Evaluation-stream labels are used only to compute final accuracy. Table A8 compares this protocol with the oracle over the evaluated configuration set for each stream.

Table A8: Evaluation-blind joint selection of (s, K, w). ImageNet-A receives (K, w) calibrated on ImageNet-V2, and ImageNet-V2 receives (K, w) calibrated on ImageNet-A. The oracle column is the envelope over the evaluated configuration set (58 configurations for ImageNet-A and 56 for ImageNet-V2).

Evaluation stream	Pseudo-purity choice s	Transferred (K, w)	Accuracy	Gain	Oracle	Regret
ImageNet-A	DINOv2-L	$(4, 20)$ from V2	68.39	+17.55	70.45	2.07
ImageNet-V2	DeiT-III	$(8, 15)$ from A	68.62	+4.96	68.72	0.10

F.1 EUROSAT FAILURE CASE

EuroSAT is the only stream with an unreliable space ranking. Candidate spaces have similar purity, so small run-to-run changes alter their order; on the other five streams, the leading spaces remain separated across seeds. This subsection concerns the five spaces of Tab. A1; over all eight, the best space on EuroSAT is DINO v1 (Sec. N), which does not change the point being made here.

In Fig. A8, DINOv2-S and DINOv2-L have disjoint three-seed ranges on every stream except EuroSAT. There, the ranges overlap ($+4.91$–$+8.36$ versus $+2.89$–$+11.40$): DINOv2-S leads for seeds 0 and 2, and DINOv2-L for seed 1. DINOv2-B, previously omitted for lacking a third seed, is now measured at all three; it tracks DINOv2-S closely on most streams but spans -3.43 to $+1.54$ on DTD, where its rank against DINOv2-S flips with the seed while DINOv2-L stays above it in every seed. The ranking claim this section defends concerns the DINOv2-S/DINOv2-L pair and is unchanged.

Converting these ranges to intervals sharpens the point. EuroSAT carries the two widest seed intervals in the paper: $[47.63, 70.06]$ for DINOv2-L at $M{=}4$, a span of 22.4 points, and a 38.6-point span for DINOv2-B at $M{=}64$. No ranking claim can be supported on a column that wide, which is why EuroSAT is presented as a failure case rather than as a measurement. The cache *gain* on EuroSAT is a separate quantity and is well resolved ($+7.32$ $[+6.21, +8.43]$ at $M{=}4$); it is the ordering among candidate spaces, not the benefit of caching, that EuroSAT cannot settle.

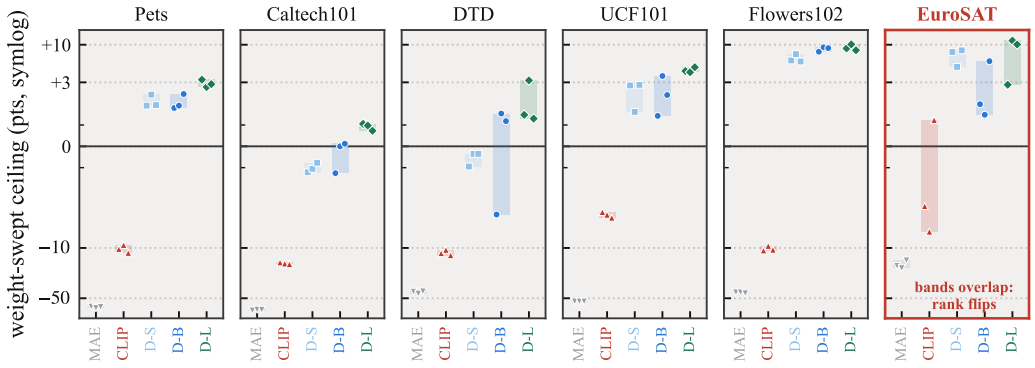


Figure A8: Seed spread of cross-domain space rankings at $M{=}4$, for the five spaces of Tab. A1; markers are seeds and bands show their range. All five spaces now have three seeds, the missing DINOv2-B runs having since completed. The DINOv2-S/DINOv2-L ordering is seed-independent on five streams and flips on EuroSAT, where their bands overlap; the new DINOv2-B measurements add one caveat, a wide band on DTD (-3.43 to $+1.54$) that overlaps both neighbours' bands.

The separation condition distinguishes stable from unstable rankings. EuroSAT's purity range is 14.4 points, versus 47.3–85.6 elsewhere; four spaces lie within 4.4 points, and even MAE reaches 74.6% purity.

We did not estimate the error radii e_s, so we cannot formally certify separation on any stream. EuroSAT's narrow purity range and overlapping gain ranges show that its leading candidates are difficult to distinguish, but do not tell us whether the cause is sampling variation or a failure of the diagnostic. EuroSAT is also a counterexample to a simple scale account: DINOv2-L has the highest purity but the third-best ceiling, 5.0 points below DINOv2-S, despite ranking first on every other stream.

F.2 Limits of Cross-Stream Prediction

Cache gain varies by an order of magnitude across streams, but DINOv2-B purity at $K{=}5$ does not predict it ($\rho{=}+0.03$). Purity saturates at 0.87–1.00 on five streams, versus 0.52–0.56 on ImageNet-A and ImageNet-V2. Anti-hubness and skewness correlate more strongly ($\rho{=}+0.71$ and $+0.66$), but $n{=}6$ supports no conclusion. Thus, Sec. 5.2 ranks candidate *spaces* within a fixed stream; it does not predict which streams benefit. Predictor headroom appears relevant, but these experiments do not establish its causal role.

G HELD-OUT FUSION-WEIGHT TRANSFER

Eq. 10 selects the fusion weight on the evaluation stream and is therefore an oracle. We quantify its advantage by selecting the weight on one stream and applying it unchanged to another, re-scoring stored logits over the same $\mathcal{W}$ without new inference.

Because stored labels and priorities are independent of w, one online pass produces base logits b_t and cache logits L_t that can be re-scored at every tested weight.

Proposition 5 (Offline re-scoring)**.** *For fixed stream, views, retrieval space, capacity, and random draw, the memory trajectory is identical for all w. Hence $b_t + wL_t$ can be evaluated for every $w \in \mathcal{W}$ without rerunning either encoder or changing the cache.*

Proof. The update stores the base pseudo-label and CLIP-derived priority, neither of which depends on the fused prediction. Induction over the stream therefore gives the same retained indices and keys for every w. □

For a common no-cache accuracy A_0, distinguish fixed-weight gain $G_w = \mathrm{Acc}(w) - A_0$, the evaluation-stream oracle $G^\star = \max_{w\in\mathcal{W}} \mathrm{Acc}(w) - A_0$, and held-out gain $G^{\mathrm{tr}} = \mathrm{Acc}(\widehat{w}_{\mathrm{sel}}) - A_0$. Oracle gain is non-negative only because $0 \in \mathcal{W}$; $+0.00$ means no tested weight helps. Held-out selection uses no evaluation labels. Its cost $G^\star - G^{\mathrm{tr}}$ is non-negative when both entries use the same stream, capacity, grid, and seeds; unmatched oracle comparisons do not have that interpretation.

For Table A9, we select w on ImageNet-V2 at a common $K{=}8$ and apply it to ImageNet-A, averaging three stored-logit seeds. This is more conservative than using each space's $K^\star$. MAE and CLIP are harmful on V2 at $K{=}8$, so transfer selects $w{=}0$.

Table A9: Oracle and held-out fusion weights on ImageNet-A at $M{=}1$, averaged over two seeds. Transfer selects w on ImageNet-V2 and preserves both the five-space ordering and its 0.90 rank correlation with Purity @10. Columns report $G^\star$ and G^{tr}, respectively.

Cache space	$G^\star$ oracle ↑	G^{tr} held out ↑	cost
MAE	+0.09	+0.00	0.09
CLIP	+0.56	+0.00	0.56
DINOv2-S	+1.33	+0.45	0.88
DINOv2-B	+9.88	+9.73	0.16
DINOv2-L	+20.06	+19.16	0.90
Span	19.97	19.16	0.81
Spearman with Purity @10	0.90	0.90	0.00

For cross-domain transfer, each space selects w on the mean curve of five streams and applies it to the sixth. Across 30 cells, transfer costs 0.37 points on average. Mean purity–gain correlation falls from 0.87 to 0.73 but remains positive on every stream. The largest loss is -2.44 for CLIP on EuroSAT, the only cell where a CLIP cache contributes over one point and its weight does not transfer.

Tab. A1 therefore reports upper bounds, but ImageNet-A's ordering does not depend on tuning w on the evaluation stream: held-out selection loses under one point of a roughly 20-point retrieval-space span.

H COMPONENT ABLATIONS ACROSS OOD SPLITS

Table A10 gives the cumulative ablation from Sec. 4 by OOD split.

Table A10: Incremental accuracy change from each component at $M{=}1$. Refinement consistently hurts, the negative memory is neutral, and adaptive fusion helps only on ImageNet-Sketch.

Component	A	V2	R	Sketch	Mean
Cross-modal refinement	−3.27	−0.58	−0.10	−3.81	−1.94
Entropy-band neg. memory	+0.01	+0.00	+0.00	+0.00	+0.00
Uncertainty-adaptive fusion	−0.32	−0.22	−1.22	+1.41	−0.09

Entropy and admission-priority controls. Table A11 reports paired three-seed comparisons on ImageNet-A. Replacing the inherited entropy score with entropy of the view posterior preserves accuracy to within 0.04 points. Retrieval-aware priority, computed from fused pre-update evidence, gives closely matched accuracy for both auxiliary encoders. These controls support the shared admission trajectory used throughout the encoder intervention; the admission rows retain CLIP-derived stored pseudo-labels.

Table A11: Paired protocol controls on ImageNet-A at M=1 and w=10, averaged over three seeds. The entropy control uses K=8; the admission-priority controls use K=16. Subscripts are standard deviations across seeds, and Δ is paired variant minus reference.

Control	Retrieval encoder	Reference	Variant	paired Δ
Entropy: inherited $\rightarrow$ standard	DINOv2-B	$60.14_{\pm 0.38}$	$60.10_{\pm 0.41}$	-0.04
Priority: CLIP $\rightarrow$ retrieval-aware	DINOv2-B	$60.14_{\pm 0.38}$	$60.24_{\pm 0.14}$	$+0.11_{\pm 0.26}$
Priority: CLIP $\rightarrow$ retrieval-aware	DINOv2-L	$69.93_{\pm 0.23}$	$70.01_{\pm 0.28}$	$+0.08_{\pm 0.22}$

Table A12 broadens the population intervention to confidence-only CLIP admission, random per-pseudo-class reservoir sampling, and chronological FIFO caching. Across all five policies, the DINOv2-L–DINOv2-B accuracy gap stays near 9.6 points. Random and FIFO retain at least 8.73 points of gain with DINOv2-B and 18.33 with DINOv2-L, establishing the encoder ordering across retained populations.

Table A12: Admission-policy comparison on ImageNet-A at M=1, K=16, and w=10. All policies use CLIP pseudo-labels; values are mean $\pm$ standard deviation over three seeds. The cache-free reference is 50.84%.

Admission policy	DINOv2-B	DINOv2-L
Entropy priority	60.14 ± 0.38	69.93 ± 0.23
Confidence-only CLIP p_1	60.28 ± 0.35	70.14 ± 0.17
Random reservoir	59.60 ± 0.28	69.27 ± 0.16
Chronological FIFO	59.57 ± 0.15	69.17 ± 0.37
Retrieval-aware	60.24 ± 0.14	70.01 ± 0.28

Refinement is most harmful where CLIP is weakest: -3.27 on ImageNet-A and -3.81 on ImageNet-Sketch, versus -0.10 on ImageNet-R. Refinement and negative memory both use the weak CLIP neighborhoods identified in Tab. A1; adaptive fusion does not, so retrieval space alone does not determine its effect.

The optional mechanisms have three distinct pathways. With CLIP-memory evidence r_t^C, refinement replaces only the stored retrieval label by $\hat{y}_t^\star = \arg\max_c [b_t + r_t^C]_c$. The entropy-band negative memory instead emits

$$\widetilde{b}_t = b_t + wL^{\mathcal{M}_{t-1}}(q_t) - \lambda_- R_t,$$

and adaptive fusion replaces w by $w_t = w\bar{h}_t^\gamma$, where $\bar{h}_t$ is normalized view entropy. The latter two change the emitted logit but not the positive-cache trajectory. The reduced configuration disables all three output paths while retaining the K_C=16 CLIP memory for admission priority; the strict one-memory control sets K_C=0.

I Full Rank-Correlation Results

Table A13 gives the rank correlations quoted in Sec. 5.2.

Spearman ρ is the Pearson correlation of the two midrank vectors; tied gains receive midranks. We use it descriptively because the encoders form a designed candidate set rather than a random sample, and report no significance tests. With untied rankings, $\rho = 1 - 6\sum_i d_i^2/[n(n^2-1)]$: attainable values are spaced by 0.1 at n=5 and $1/42 \approx 0.024$ at n=8. We therefore interpret the ordering, held-out replication, and selection regret rather than small numerical differences. Midranks matter wherever gains tie at zero: on ImageNet-V2 five of the sixteen spaces do, and breaking those ties by index order instead of averaging them would raise $\rho(-\mathrm{AntiHub}, G^\star)$ from 0.756 to 0.771. Every correlation reported in this paper uses the midrank definition stated above.

Table A13: Spearman ρ between geometry and outcome across five spaces. ImageNet-V2 is held out with no refitting. Green marks the stronger association per outcome. Over eight spaces, Purity @10 correlates with gain at 0.93 on ImageNet-A and 0.98 on ImageNet-V2 (Sec. N). Over sixteen spaces, the correlations are 0.959 and 0.943 (Table 1); retaining positive-gain spaces gives 0.971 and 0.973, respectively.

Dataset	Predictor	vs. ceiling ↑	vs. $K^\star$ ↑
ImageNet-A (predictors selected here)			
	Purity @10	0.90 [0.900, 0.900]	0.36 [0.359, 0.359]
	− AntiHub	0.50 [0.500, 0.718]	0.87 [0.718, 0.872]
ImageNet-V2 (held out, nothing refitted)			
	Purity @10	0.98 [0.975, 0.975]	0.95 [0.949, 0.949]
	− AntiHub	0.95 [0.564, 0.872]	0.97 [0.632, 0.949]

Bracketed values are central 95% calibration-subsample intervals at the largest tested size, n=2000, from B=300 replicates that resample calibration images and recompute the geometry statistic while holding the measured ceilings and $K^\star$ fixed. The encoder set is held fixed throughout: these five spaces are a designed candidate set rather than a draw from a population of encoders, so an interval taken over encoders would have no sampling model behind it, which is why the correlations themselves remain descriptive. What the intervals do establish is that both arms of the dissociation are resolved on ImageNet-A: purity's lead on the ceiling and anti-hubness's lead on $K^\star$ are each separated, and the weak entries are weak precisely rather than uncertainly. Note that Purity is close to size-invariant while AntiHub counts zero-in-degree points in an n-point graph and is therefore not, so the anti-hubness intervals describe the subsample regime and need not bracket the full-split point estimate printed beside them; the two ImageNet-V2 anti-hubness rows are the clearest instance.

Table A14: Three-seed replication of the sixteen-space ImageNet-A oracle-envelope ranking. Each seed uses a coupled stream across all encoders. DINOv2-L leads every seed, and offline purity selects that leader with zero regret. Streams use an explicitly seeded permutation; Table A1 reports an independent three-seed sweep of the same protocol drawn from the global RNG.

Seed	ρ(Purity @10, $G^\star$)	DINOv2-L $G^\star$	selection regret
0	0.959	+19.36	0.00
1	0.961	+19.71	0.00
2	0.967	+20.16	0.00
Mean	0.962	$+19.74 \pm 0.40$	0.00

Table A15: Correlation checks for the sixteen-space intervention. Tie-excluded rows retain spaces with positive oracle gain. The ResNet-50 row repeats the complete intervention with a second frozen predictor and pseudo-label source; the near-matched row uses the six ImageNet-A spaces whose gains lie between 4.4 and 10.0 points.

Predictor	Stream/candidate set	n	ρ(Purity @10, $G^\star$)	Selected encoder	Regret
CLIP ViT-B/16	ImageNet-A, all	16	0.959	DINOv2-L	0.00
CLIP ViT-B/16	ImageNet-A, positive gain	15	0.971	DINOv2-L	0.00
CLIP ViT-B/16	ImageNet-V2, all	16	0.943	DeiT-III	0.00
CLIP ViT-B/16	ImageNet-V2, positive gain	11	0.973	DeiT-III	0.00
CLIP ViT-B/16	ImageNet-A, near-matched	6	0.943	DINOv2-B	0.00
CLIP ResNet-50	ImageNet-A, all	16	0.968	DINOv2-L	0.00

With the ResNet-50 predictor, the cache-free accuracy is 24.1% and the encoder-envelope gain spans 0.00–19.04 points. The same training-regime ordering appears under both predictors, complementing the coupled-seed replication in Table A14.

J Additional Benchmark Baselines

Table A18 collects the parameter-updating OOD references in compact form and reports the larger auxiliary scale.

Table A16: Top-1 accuracy (%) on the cross-domain benchmark with CLIP ViT-B/16 under the protocol of (Karmanov et al., 2024). Green marks the best comparable training-free result per dataset; shaded rows are MARC. Every MARC row is training-free and uses a separate frozen DINOv2-B retrieval encoder. The lower block removes the cache from our pipeline, so Δ isolates its contribution. Subscripts on MARC rows are standard deviations over three seeds; averages and deltas use unrounded values. Aircraft and Food101 are evaluated separately in Table A17. The average here uses the six streams shown; Table A17 also reports the average over all eight.

Method	Caltech101 ↑	DTD ↑	EuroSAT ↑	Flowers102 ↑	Pets ↑	UCF101 ↑	Average ↑
CLIP-ViT-B/16	93.55	45.04	50.42	66.99	86.92	65.16	68.01
Methods with parameter adaptation							
CoOp (Zhou et al., 2022b)	93.70	41.92	46.39	68.71	89.14	66.55	67.74
CoCoOp (Zhou et al., 2022a)	93.79	45.45	39.23	70.85	90.46	68.44	68.04
TPT (Shu et al., 2022)	94.16	47.75	42.44	68.98	87.79	68.04	68.19
DiffTPT (Feng et al., 2023)	92.49	47.00	43.13	70.10	88.22	62.67	67.27
Training-free methods: CLIP visual encoder							
TDA (Karmanov et al., 2024)	94.24	47.40	58.00	71.42	88.63	70.66	71.73
BoostAdapter (Zhang et al., 2024)	94.77	45.69	61.22	71.66	89.51	71.93	72.46
TaTa (Zhang et al., 2026)	93.82	49.57	59.67	71.28	89.91	73.54	72.97
ETTA (Dastmalchi et al., 2025)	94.74	49.64	59.86	73.86	90.08	71.95	73.36
Training-free methods: CLIP + frozen DINOv2-B							
MARC(Ours) (DINOv2-B), M=4	$94.73_{\pm 0.18}$	$53.76_{\pm 0.54}$	$55.60_{\pm 2.93}$	$81.27_{\pm 0.12}$	$92.03_{\pm 0.32}$	$72.36_{\pm 1.01}$	$74.96_{\pm 0.53}$
MARC(Ours) (DINOv2-B), M=64	$94.69_{\pm 0.08}$	$54.37_{\pm 0.87}$	$58.67_{\pm 1.52}$	$80.67_{\pm 0.33}$	$92.39_{\pm 0.19}$	$71.94_{\pm 0.33}$	$75.45_{\pm 0.14}$
Ensemble-only controls: same pipeline without the cache							
Ensemble only, M=4	94.65	52.36	50.37	72.80	89.89	69.73	71.63
Ensemble only, M=64	95.17	53.66	47.47	72.15	89.70	70.39	71.42
Δ *contributed by the cache,* M=64	−0.48	+0.71	+11.20	+8.52	+2.69	+1.55	+4.03

The cross-domain table likewise includes CoOp, CoCoOp, TPT, and DiffTPT as parameter-adapting literature references and separates them from the training-free comparison. On this benchmark, MARC exceeds the best published results by 4.7, 6.8, and 2.3 points on DTD, Flowers102, and Pets, and trails on Caltech101, EuroSAT, and UCF101. Reducing M from 64 to 4 costs 0.50 average points, driven primarily by EuroSAT, which alone loses 3.07 points while three of the other five datasets improve; the ensemble-only baseline instead changes from 71.42 to 71.63. TPT, DiffTPT, and ZERO+Ens average 68.19, 67.27, and 67.77, respectively.

The transferred weight hurts on two additional streams. We extend the six-dataset benchmark with Aircraft and Food101, using three seeds for each configuration. Cars is no longer distributed, and we did not evaluate SUN397. At w=10, the cache reduces accuracy on both added datasets. Including them lowers the mean cache contribution at M=4 from +3.32 to +0.84 points. The cache itself retains a small positive gain when the weight is selected per stream: +1.39 on Aircraft and +0.50 on Food101.

Table A17: Results on Aircraft and Food101 at M=4 and w=10, averaged over three seeds. Subscripts are standard deviations. Comparable published baselines are unavailable under this protocol, so we compare with the same pipeline without the cache. The lower rows show how the two datasets change the cross-domain average; $\Delta = G_{10}$. Aircraft has too many incorrect pseudo-labels for reliable retrieval, while Food101 leaves few base errors to correct (Corollary 2). The two arms are separate runs, so Δ here is an unpaired difference of means and differs slightly from the paired per-image losses in App. R. Table 1 uses $M = 1$, so its values should not be compared directly with the Aircraft and Food101 results from this run.

Stream	MARC (DINOv2-B) ↑	Ensemble only, no cache ↑	G_{10}
Streams added to the suite			
Aircraft	$19.25_{\pm 0.93}$	$27.71_{\pm 0.62}$	−8.46
Food101	$78.98_{\pm 0.14}$	$83.77_{\pm 0.05}$	−4.79
Effect on the cross-domain average, M=4			
Six streams (Table A16)	74.96	71.63	+3.32
Eight streams	68.50	67.66	+0.84

Table A18: Compact OOD reference for parameter-updating methods and the larger auxiliary scale. The method groups keep training protocol and auxiliary scale explicit; † marks our runs.

Method	IN	IN-A	IN-V2	IN-R	IN-Sk	OOD Avg
ResNet-50: methods that update parameters						
CoOp (Zhou et al., 2022b)	63.33	23.06	55.40	56.60	34.67	42.43
CoCoOp (Zhou et al., 2022a)	62.81	23.32	55.72	57.74	34.48	42.82
TPT (Shu et al., 2022)	60.74	26.67	54.70	59.11	35.09	43.89
DiffTPT (Feng et al., 2023)	60.80	31.06	55.80	58.80	37.10	45.69
ViT-B/16: methods that update parameters						
TPT (Shu et al., 2022)	68.98	54.77	63.45	77.06	47.94	60.81
DiffTPT (Feng et al., 2023)	70.30	55.68	65.10	75.00	46.80	60.65
ViT-B/16: larger auxiliary scale						
COSMIC (Huang et al., 2025) (DINOv2-L-reg)	78.19	73.32	69.62	85.60	62.79	72.83
MARC (DINOv2-L), M=64†	77.03	72.04	66.95	84.62	59.83	70.86

At DINOv2-L scale, MARC reaches a 70.86 OOD average versus COSMIC's 72.83. The configurations differ in unisolated factors, so we do not attribute the gap. Within MARC, changing only DINOv2-B to DINOv2-L adds 2.95 points (67.91 → 70.86).

K Cross-Domain Retrieval-Space Rankings

Table A20 gives the per-stream ceilings underlying Sec. 5.2.

Table A19: Selecting among sixteen retrieval spaces on three cross-domain streams at M=4. *Range* is the purity spread, and *ties* counts spaces with zero gain. Purity selects the best space on Flowers102 and DTD but not on EuroSAT, where its values are tightly grouped. All gains used for ranking and regret are oracle envelopes $G^\star$.

Stream	range	ties	ρ(purity, $G^\star$) ↑	best space	purity selects	regret ↓
Purity is separated: the rule applies						
Flowers102	85.6	2	+0.922	DINOv2-L	DINOv2-L	0.00
DTD	50.5	4	+0.736	DINOv3-B	DINOv3-B	0.00
Precondition violated: purity has no spread to rank with						
EuroSAT	16.7	1	+0.400	DINOv3-B	EVA-02 CLIP	1.98

Table A20: Oracle-envelope cache gains $G^\star$ for five spaces on six cross-domain streams at M=4. ρ correlates Purity @10 with $G^\star$; *range* is the purity spread. Green marks the best of these five spaces. The three spaces added in Sec. N are not listed here; they enter the eight-space selection analysis of Tab. A25, where DINO v1 takes the lead on EuroSAT.

Stream	MAE	CLIP	D-S	D-B	D-L	ρ ↑	range
Purity is discriminative: the rule applies							
Pets	+0.03	+0.11	+1.94	+2.45	+3.60	1.00	82.2
Caltech101	+0.00	+0.00	+0.28	+0.89	+1.46	1.00	63.1
DTD	+0.00	+0.00	+2.19	+3.25	+4.14	0.90	47.3
UCF101	+0.05	+1.24	+2.80	+2.62	+4.55	0.90	50.0
Flowers102	+0.08	+0.00	+6.33	+8.61	+9.26	0.80	85.6
Mean						0.92	65.6
Precondition violated: purity has no spread to rank with							
EuroSAT	+0.00	+0.07	+8.63	+7.54	+3.60	0.10	**14.4**

EuroSAT is the only stream with a narrow purity range; the next-narrowest is over three times wider. Separation therefore tests ranking stability, not correctness. EuroSAT also remains a counterexample: DINOv2-L has the highest purity but only the third-best gain among these five, and the sixth-best of all eight.

The two reported EuroSAT correlations use different candidate sets: ρ=0.400 in Table A19 ranks sixteen spaces, whereas ρ=0.10 in Table A20 ranks the original five.

Table A21 extends the sweep to the three spaces of Sec. N. Unlike on the ImageNet streams, adding candidates lowers the cross-domain rank correlation: on the matched single-seed basis of this table, the mean ρ over the five well-separated streams falls from 0.91 to 0.82, though it remains at least 0.78 on each. The added spaces introduce many ties in cache gain: BEiT (Bao et al., 2022) contributes exactly $+0.00$ on all five well-separated streams, and DINO v1 does so on two of them. No predictor can order tied outcomes; as a diagnostic, restricting each stream to the spaces that deliver a non-negligible gain returns the mean to 0.89. That restriction conditions on the outcome and is therefore not a correction we adopt, only evidence that the loss is concentrated in unrankable ties.

Despite the lower rank correlation, the rule's selection performance does not degrade. Across the same eight streams, selection from eight spaces has lower regret than selection from the original five (Table A25). Ties among spaces that contribute nothing reduce rank correlation without increasing practical regret.

Table A21: Oracle-envelope cache gains $G^\star$ for the three added spaces on the six cross-domain streams at $M{=}4$, with ρ recomputed over all eight. To keep the eight comparable, every entry here and every ρ in this table is a single seed-0 run; Tab. A20 averages two seeds for DINOv2-B, which is why its ρ column differs by up to 0.03. Green marks the best space over all eight.

Stream	Sup. ViT	DINO v1	BEiT	ρ over 8 ↑
Purity is discriminative				
Pets	$+2.59$	$+1.94$	$+0.00$	0.84
Caltech101	$+0.00$	$+0.00$	$+0.00$	0.79
DTD	$+0.06$	$+0.00$	$+0.00$	0.91
UCF101	$+2.83$	$+2.67$	$+0.00$	0.79
Flowers102	$+5.64$	$+2.15$	$+0.00$	0.78
Mean				0.82
Precondition violated				
EuroSAT	$+5.31$	$+9.72$	$+6.35$	0.26

EuroSAT is again the exception, and the added spaces sharpen rather than soften it. DINO v1 gives $+9.72$ there, the largest gain achieved by any space on that stream and more than it achieves elsewhere; on the other five streams, its gain ranges from $+0.00$ to $+2.67$. BEiT shows an even sharper contrast: $+6.35$ on EuroSAT and $+0.00$ elsewhere. A stream on which otherwise weak spaces become competitive may reward structure other than class-level neighborhood purity, consistent with EuroSAT's purity range of 14.4 points and with its being the one stream where our rule does not apply.

L Qualitative Examples

Four ImageNet-A queries that the base predictor mislabels and the fused rule repairs, chosen by the data rather than by hand: base prediction wrong, cache margin negative in CLIP space, positive in DINOv2-L space, fused prediction correct. Each row of Fig. A9 shows the query and its ten nearest retained keys in the two spaces; the memory is the same $K{=}8$ population in both (Proposition 1), and neighbours are ranked by cosine in the full embedding.

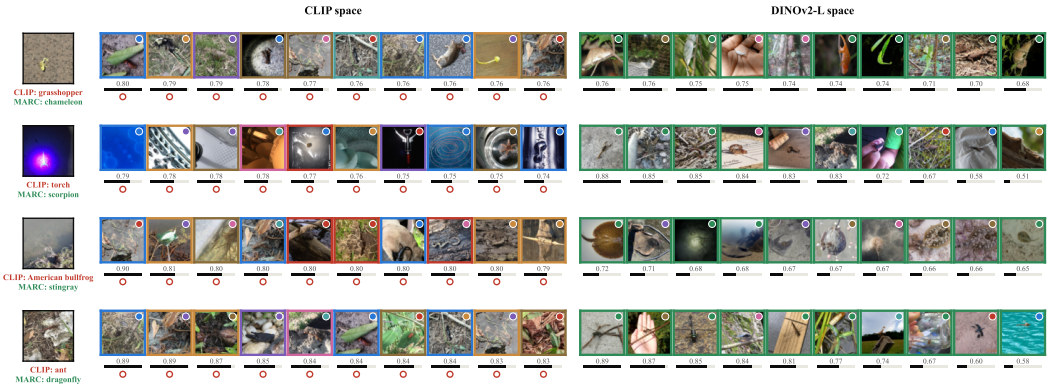


Figure A9: Queries repaired after the retrieval-encoder swap. Left block: the ten nearest retained keys in CLIP space; right block: the same memory re-embedded by DINOv2-L. Border colour is the neighbour's true label (green matches the query), the corner dot its stored pseudo-label, the number the cosine similarity, the bar the kernel weight on a scale shared across both spaces, and a red ring marks a neighbour of a different true class. In CLIP space the nearest keys mix classes at nearly uniform similarity, so the vote goes to a confuser; re-embedded by DINOv2-L, the same memory returns neighbours that are almost all of the query's class.

M Cache Capacity at a Large View Budget

The capacity results in Sec. 5.2 are measured at $M{=}1$, leaving open whether the same optimum holds once many views are averaged. We therefore repeat the six cross-domain streams at $M{=}64$ under both $K{=}8$ and $K{=}16$, with three seeds in each arm, so that capacity is the only factor that varies.

Table A22: Cross-domain accuracy at $M{=}64$ under two cache capacities, three seeds per arm. *Ratio* is the change divided by the pooled dispersion of the two arms. Means and deltas are computed from unrounded values.

Stream	$K{=}8$	$K{=}16$	Δ	ratio
Caltech101	$94.69_{\pm 0.08}$	$93.51_{\pm 0.14}$	-1.18	$10.3\times$
DTD	$54.37_{\pm 0.87}$	$50.18_{\pm 1.24}$	-4.20	$3.9\times$
Flowers102	$80.67_{\pm 0.33}$	$79.71_{\pm 0.39}$	-0.96	$2.6\times$
Pets	$92.39_{\pm 0.19}$	$91.66_{\pm 0.05}$	-0.73	$5.2\times$
UCF101	$71.94_{\pm 0.33}$	$70.65_{\pm 0.52}$	-1.29	$3.0\times$
EuroSAT	$58.67_{\pm 1.52}$	$58.67_{\pm 0.42}$	$+0.00$	$0.0\times$
Mean	75.45	74.06	-1.39	

Increasing capacity from $K{=}8$ to $K{=}16$ remains harmful at a large view budget, costing 1.39 points on average. Five streams lose accuracy and every loss exceeds twice the pooled dispersion of its two arms, reaching ten times on Caltech101. EuroSAT is the exception for the third time: the two capacities agree to two decimals, and its $K{=}8$ dispersion of 1.52 points is the widest in the table, so the stream that defeats the geometry ranking is also the one whose capacity cannot be resolved. On the other five streams, the $M{=}1$ preference for $K{=}8$ over $K{=}16$ transfers to $M{=}64$. This was not obvious in advance: averaging many views could have compensated for a diluted cache, but it does not.

N Additional ViT-B/16 Pretraining Regimes

The five spaces compared in Sec. 5.2 jointly vary the pretraining objective, data source, dataset scale, and encoder scale. DINOv2 is self-supervised on curated LVD-142M, CLIP is language-supervised on WIT-400M, and MAE uses masked image modeling on ImageNet-1k (Deng et al., 2009); three of the five are scale variants of DINOv2. The observed association between encoder-induced retrieval

similarities and gain may therefore reflect the pretraining data and recipe as well as the objective category.

We add three retrieval spaces that hold the architecture fixed at ViT-B/16 with 768-dimensional features while broadening the pretraining regimes represented. They are measured under the same protocol as the other spaces: frozen encoder, CLS token, ℓ_2-normalized keys, and the same cache and fusion configuration. Each model uses its own input normalization rather than a shared one.

Table A23: Additional ViT-B/16 retrieval spaces spanning different pretraining regimes.

Space	Pretraining	Contrast provided
Supervised ViT-B/16	labels, IN-21k $\rightarrow$ IN-1k	supervised objective with fixed architecture
DINO v1 ViT-B/16	self-distillation, IN-1k	related objective family, $\sim$100$\times$ less data than DINOv2
BEiT-B/16	masked image modeling, IN-22k	alternative masked-image pretraining recipe

DINO v1 provides the closest comparison to DINOv2 among the added spaces. It uses a related self-distillation approach but is trained on ImageNet-1k rather than LVD-142M. A result near DINOv2-B would support an explanation based on the broad objective family, whereas a result near MAE would point to differences in data scale or training recipe. This comparison is informative but cannot isolate those factors completely.

Table A24: Comparison of the three added spaces with the five spaces already reported, on ImageNet-A at $M{=}1$. Purity and anti-hubness are measured at $K{=}10$ before adaptation, and the ceiling is the oracle envelope $G^\star$ at each space's own $K^\star$. Rows are ordered by ceiling. $K^\star$ and the ceiling are computed from seed 0; the three-seed figures in the text use the common $K{=}8$. Because those figures average different stream orders, DINO v1's three-seed mean at $K{=}8$ can exceed its seed-0 ceiling.

Cache space	Purity @10$\uparrow$	AntiHub$\downarrow$	$K^\star$	$G^\star$ ceiling$\uparrow$
BEiT-B/16	19.1	11.1	1	+0.01
MAE ViT-B/16	3.2	3.4	1	+0.09
DINO v1	16.8	1.9	1	+0.17
CLIP ViT-B/16	29.7	6.3	2	+0.62
DINOv2-S	26.1	0.7	16	+1.33
Supervised ViT	35.0	1.4	8	+4.75
DINOv2-B	48.9	1.2	8	+9.88
DINOv2-L	67.3	1.7	8	+20.06

DINO v1 reaches $+0.17$ points at its best seed-0 capacity and $+0.25 \pm 0.20$ over three seeds at the matched $K{=}8$, placing it with MAE and BEiT rather than with DINOv2. Thus, membership in the same broad self-distillation family is not sufficient to produce a useful retrieval space. The supervised ViT reaches $+4.75$ at its seed-0 optimum and $+4.82 \pm 0.20$ over three seeds at $K{=}8$, roughly three times the DINOv2-S gain despite using no self-supervision. The same ordering appears on ImageNet-V2, where the supervised ViT reaches $+4.76$ and DINO v1 reaches $+0.51$. Together, these results implicate pretraining data scale and recipe, rather than the objective category alone, but do not separate their individual effects.

The geometry rule remains effective after the three structurally different spaces are added. Across eight spaces, purity ranks the gain at $\rho{=}0.93$ on ImageNet-A and $\rho{=}0.98$ on ImageNet-V2, against 0.90 and 0.98 over the original five. Selection regret falls from 0.98 to 0.35 points, and its worst case from 5.02 to 1.65 (Table A25). Rather than degrading as the candidate set broadens, the association strengthens on ImageNet-A, supporting the interpretation that purity tracks retrieval quality rather than merely DINOv2 model identity. With sixteen spaces, the correlation remains $\rho{=}0.959$, and purity selects the best space on both ImageNet streams (Sec. O).

Table A25: Selecting a retrieval space by a single measurement, over eight streams and eight candidate spaces. *Hits* counts streams where the rule picks the best space; *regret* is $G^\star$ lost relative to oracle selection. Among the tabulated rules, only purity requires labels.

Rule	Labels	hits	mean regret ↓	worst ↓
Purity @5	yes	7/8	0.14	1.14
Purity @10	yes	6/8	0.35	1.65
lowest skewness	no	4/8	4.14	19.89
lowest AntiHub	no	0/8	6.45	18.71

Among the original label-free geometry summaries, none selects reliably. Anti-hubness prefers MAE on almost every stream because a space whose neighborhoods carry no class information can still leave few unretrieved points. Purity at $K{=}5$ is slightly better than at $K{=}10$, although the paper reports $K{=}10$ consistently. For a measurement ϕ, let $\sigma_\phi(\mathcal{D})$ be its selected space. Selection regret is

$$\mathrm{Reg}_\phi(\mathcal{D}) = \max_s G^\star(s, \mathcal{D}) - G^\star(\sigma_\phi(\mathcal{D}), \mathcal{D}), \tag{18}$$

and a hit has zero regret. Replacing ground-truth agreement with agreement against CLIP predictions gives label-free *pseudo-purity*. It reproduces the labeled ordering closely ($\rho{=}0.98$ on ImageNet-A, 1.00 on ImageNet-V2, and 0.93 averaged over eight streams), selects the best space on $5/8$ streams, and incurs 1.48 points of mean regret. This is worse than labeled Purity @10 (0.35) but substantially better than anti-hubness (6.45); its worst case is EuroSAT, where even the labeled rule is unresolved.

O The Full Sixteen-Space Candidate Set

Table 1 compares sixteen retrieval spaces. Eight were introduced earlier: CLIP ViT-B/16, MAE, three DINOv2 variants, a supervised ViT, DINO v1, and BEiT. The remaining eight test whether the result depends on a particular training objective, architecture, or scale.

OpenCLIP (Cherti et al., 2023), SigLIP (Zhai et al., 2023), and EVA-02 CLIP (Sun et al., 2023) test three additional forms of cross-modal training, while CLIP ViT-L/14 tests the effect of scale. ConvNeXt-B (Liu et al., 2022) and ResNet-50 (He et al., 2016) test convolutional architectures. DINOv3-B (Siméoni et al., 2025) adds a newer self-supervised model, and DeiT-III (Touvron et al., 2022) adds a second label-supervised transformer.

Every space is used under the protocol of Section 4: frozen encoder, a single pooled or CLS embedding, ℓ_2-normalised keys, each model's own input normalisation, and the same cache, kernel and fusion configuration. Only the retrieval encoder changes, so the controlled estimand is the resulting replacement of the query–key similarity matrix.

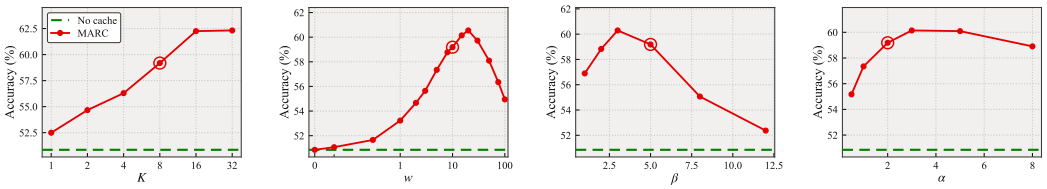
Table A26: The eight retrieval spaces added to reach the sixteen of Table 1, with the purpose of each addition in the controlled comparison.

Retrieval space	Purpose in the comparison
OpenCLIP ViT-B/16	same objective and architecture as CLIP, trained on LAION-2B (Schuhmann et al., 2022) rather than WIT-400M
SigLIP ViT-B/16	tests a cross-modal model trained with a different loss
EVA-02 CLIP	adds an independent cross-modal training recipe
CLIP ViT-L/14	tests whether greater model scale improves cross-modal retrieval
DINOv3-B	adds a later generation of self-supervised training
DeiT-III ViT-B/16	adds a second label-supervised vision transformer
ConvNeXt-B	tests whether the result extends beyond transformer architectures
ResNet-50 (sup.)	repeats that test with a smaller convolutional model

CLIP ViT-L/14 is the only cross-modal space with a substantial improvement, reaching $+5.52$ on ImageNet-A; scale therefore explains part of the gap. The supervised ResNet-50 also changes rank across streams, with gains of $+0.04$ on ImageNet-A and $+1.54$ on ImageNet-V2. This is another example of the stream dependence studied in Table A19.

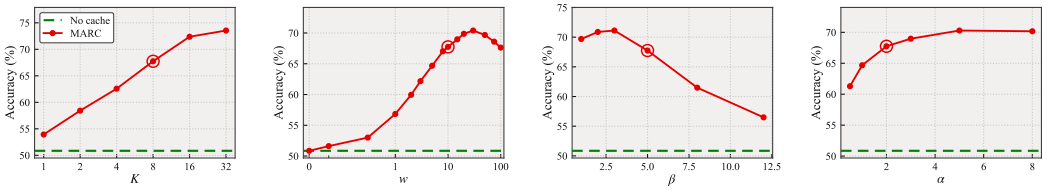
P HYPERPARAMETER SENSITIVITY

The following one-at-a-time sweeps vary (K, w, β, α) around the deployed setting $(8, 10, 5, 2)$. They are placed after the broader space comparisons because they test local robustness of the fixed scoring rule rather than the paper's central retrieval-space claim.



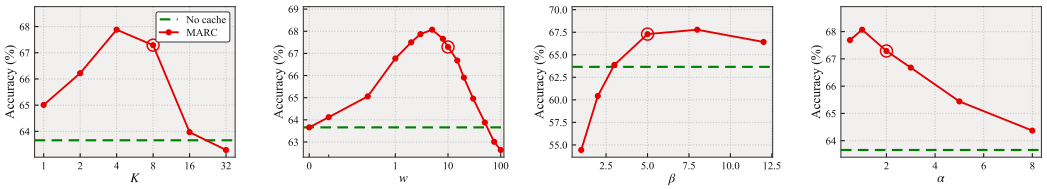
(a) Cache capacity. (b) Fusion weight. (c) Kernel sharpness. (d) Kernel scale.

Figure A10: ImageNet-A sensitivity with DINOv2-B retrieval. The deployed setting lies in the broad high-accuracy region; capacity and fusion weight dominate the response.



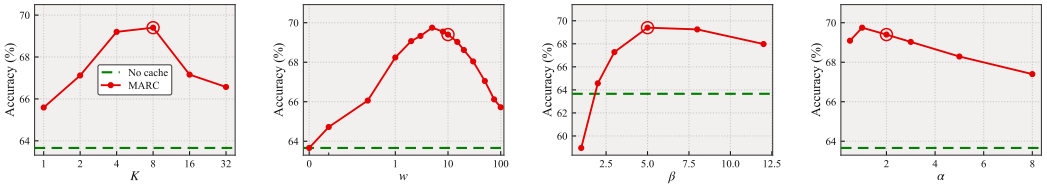
(a) Cache capacity. (b) Fusion weight. (c) Kernel sharpness. (d) Kernel scale.

Figure A11: ImageNet-A sensitivity with DINOv2-L retrieval. Its similarities support a wider range of cache weights while retaining a large gain.



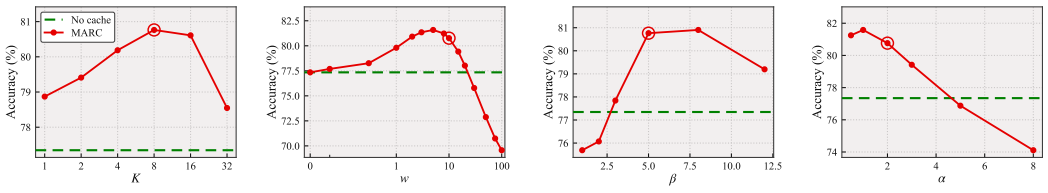
(a) Cache capacity. (b) Fusion weight. (c) Kernel sharpness. (d) Kernel scale.

Figure A12: ImageNet-V2 sensitivity with DINOv2-B retrieval. Smaller predictor headroom yields a flatter and lower-gain response than ImageNet-A.



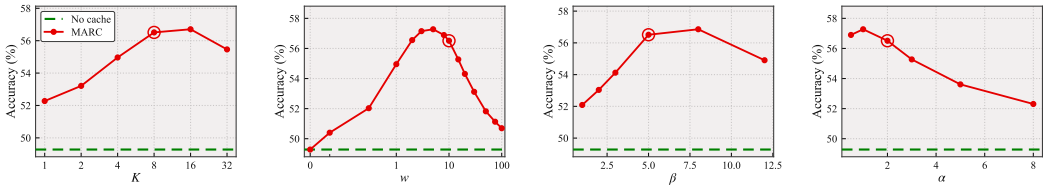
(a) Cache capacity. (b) Fusion weight. (c) Kernel sharpness. (d) Kernel scale.

Figure A13: ImageNet-V2 sensitivity with DINOv2-L retrieval. The useful region remains broad, although the attainable gain is smaller than on ImageNet-A.



(a) Cache capacity. (b) Fusion weight. (c) Kernel sharpness. (d) Kernel scale.

Figure A14: ImageNet-R sensitivity with DINOv2-B retrieval under the same one-factor sweeps.



(a) Cache capacity. (b) Fusion weight. (c) Kernel sharpness. (d) Kernel scale.

Figure A15: ImageNet-Sketch sensitivity with DINOv2-B retrieval under the same one-factor sweeps.

Q Full Experimental Protocol and Isolation Proof

Data and models. We evaluate ImageNet-A (7,500 images, 200 classes), ImageNet-V2 (10,000) (Recht et al., 2019), ImageNet-R (30,000, 200 classes), ImageNet-Sketch (50,889), ImageNet-val (50,000), and the cross-domain suite of Karmanov et al. (2024). The predictor is frozen CLIP ViT-B/16 or ResNet-50 with fixed prompts. Retrieval encoders are frozen, use one pooled or CLS embedding with native normalization, and are listed in Sections N and O.

Geometry calibration. The reported purity, pseudo-purity, and hubness measurements use the complete target test split as the calibration batch: 7,500 images for ImageNet-A, 10,000 for ImageNet-V2, 30,000 for ImageNet-R, and 50,889 for ImageNet-Sketch, with the analogous complete split for each cross-domain dataset. These batches are the evaluation splits rather than held-out sets. Ground-truth labels enter only the offline purity analysis; pseudo-purity uses frozen CLIP top-1 predictions on the same unlabeled images.

Views and online update. From M views we retain $r_M = \max(1, \lfloor 0.1M \rfloor)$ lowest-entropy views and average their raw CLIP logits. The unaugmented view supplies each query and candidate key. Both memories are queried before insertion and store the ensemble top-1 pseudo-label. The CLIP memory has per-class capacity K_C=16, zero output weight, and supplies retrieval-admission priority. Within each pseudo-class, the retrieval cache retains the K highest-priority arrivals; a new item replaces the current minimum only when its priority is larger. All keys are ℓ_2-normalized. Setting K_C=0 gives the strict one-memory control.

Query-before-insert prevents a current-sample self-vote. Strict priority replacement makes each pseudo-class contain the top-K arrivals in every stream prefix; under distinct priorities, capacity-K entries are nested in those at K+1. Because stored pseudo-labels and priorities do not use the fusion weight, the same trajectory supports the complete offline sweep of Proposition 5.

Proof of Proposition 1. Couple two retrieval-space runs with the same stream and augmentation draws. Their CLIP quantities $(b_t, \hat{y}_t, e_t, k_t^C)$ coincide. By induction, their CLIP-memory states, pre-update evidence, and retrieval priorities also coincide. Since admission compares only pseudo-class and priority, both runs retain the same stream indices at every time; only the vectors attached to those indices differ.

For a retained index i, write $u_{ti}^{s} = g_s(x_t)^\top g_s(x_i)$. Substitution into Eq. (3) gives the logit identity in Eq. (6). Moreover, $\phi'_\beta(u) = \beta\phi_\beta(u) \le \beta$ on $[-1, 1]$, so the mean-value theorem gives

$$|L_{t,c}^{s} - L_{t,c}^{s'}| \le \alpha\beta \sum_{i \in I_{t-1}^{K}(c)} |u_{ti}^{s} - u_{ti}^{s'}| = D_{t,c}^{s,s'}.$$

Because b_t is shared, subtracting the largest possible perturbations from the winning-class gap yields the sufficient stability condition in Eq. (7). □

Search and reporting. Capacity uses $K \in \{1, 2, 3, 4, 8, 16, 32, 64\}$ and fusion uses $\mathcal{W} = \{0, 0.1, 0.3, 1, 2, 3, 5, 8, 10, 15, 20, 30, 50, 75, 100\}$. Including $w{=}0$ makes the oracle envelope a non-negative upper bound. Retrieval-space comparisons, benchmark rows, component ablations, and capacity sweeps use three seeds unless a caption states otherwise. Seeds change stream order and, for $M > 1$, augmentation draws; matched configurations share seeds. Subscripts report sample standard deviations with the $n - 1$ divisor, and averages use unrounded values. A standard deviation is not a confidence interval: at three seeds the 95% t-interval is $2.484\times$ the printed subscript. Interval estimates, and the exact tests behind every comparative claim, are reported separately in App. R.

Timing. We time a 3,000-image prefix on one thermally settled RTX 3060 after 100 warm-up samples. TDA, COSMIC, and MARC are measured in the same session, repeated three times in Section S. We interpret within-session ratios rather than absolute consumer-GPU latency.

R Confidence Intervals for the Reported Gains

We quantify uncertainty using four complementary procedures, each associated with a distinct source of variation. Absolute accuracies are accompanied by *sampling intervals*, computed as Wilson intervals over the evaluation split. Differences between methods are summarized by *paired intervals* over the same evaluation images, together with exact two-sided McNemar tests on the discordant predictions; these intervals support the comparative claims in this section. For the oracle gain $G^\star$, we report an *envelope interval* obtained by bootstrapping evaluation examples and re-optimizing over $\mathcal{W}$ within each replicate, thereby incorporating uncertainty due to weight selection. Finally, variation across stream orders is summarized by a *seed interval*: a t-interval with half-width $t_{.975,n-1}s/\sqrt{n}$, where s is the sample standard deviation across orders. For three orders, this half-width is $2.484s$. Sampling and seed intervals address different estimands. The former characterizes uncertainty over evaluation examples, whereas the latter characterizes sensitivity to stream order. Similarly, overlap between two marginal accuracy intervals does not determine whether their difference is statistically distinguishable: because the methods are evaluated on identical images, inference about their difference should use the paired results. Calibration-subsample intervals for Spearman's ρ and selection regret are reported separately in Table A7.

Sixteen-space intervention. Table A27 reports paired intervals and exact tests for all retrieval spaces in Table 1 on ImageNet-A. The intervention holds both the stream and retained sample population fixed (Proposition 1); consequently, every row is compared with the same no-cache predictor on the same 7,500 images.

Table A27: Paired uncertainty estimates for the sixteen-space intervention on ImageNet-A at $K{=}8$ and $M{=}1$, relative to the 50.84% no-cache baseline. *Gain at* $w{=}10$ denotes the paired accuracy difference, reported with its 95% confidence interval and the exact two-sided McNemar p-value over discordant predictions. $G^\star$ denotes the oracle envelope; its bootstrap interval re-optimizes the maximum over $\mathcal{W}$ within each replicate. These columns estimate different quantities and therefore need not coincide.

Cache space	gain at $w{=}10$	disc.	McNemar p	$G^\star$
DINOv2-L	$+18.19$ $[+17.21, +19.17]$	1658	$2e{-}285$	$+19.36$ $[+18.36, +20.51]$
DINOv2-B	$+9.31$ $[+8.43, +10.18]$	1186	$4e{-}97$	$+10.27$ $[+9.32, +11.31]$
DeiT-III	$+7.73$ $[+6.80, +8.67]$	1322	$5e{-}59$	$+8.01$ $[+7.00, +9.11]$
DINOv3-B	$+5.72$ $[+4.96, +6.48]$	879	$3e{-}49$	$+6.49$ $[+5.67, +7.49]$
Supervised ViT	$+4.13$ $[+3.38, +4.89]$	844	$5e{-}27$	$+4.43$ $[+3.61, +5.24]$
CLIP ViT-L/14	$+3.69$ $[+2.63, +4.75]$	1657	$1e{-}11$	$+5.52$ $[+4.65, +6.39]$
ConvNeXt-B	$+2.88$ $[+2.20, +3.56]$	684	$1e{-}16$	$+2.85$ $[+2.27, +3.72]$
EVA-02 CLIP	$+1.88$ $[+0.84, +2.92]$	1577	$4e{-}04$	$+4.15$ $[+3.44, +4.93]$
DINOv2-S	$+0.53$ $[-0.24, +1.31]$	882	0.189	$+1.24$ $[+0.85, +1.75]$
DINO v1	-2.27 $[-3.01, -1.52]$	810	$3e{-}09$	$+0.05$ $[+0.00, +0.48]$
OpenCLIP	-6.16 $[-7.17, -5.15]$	1536	$2e{-}32$	$+0.76$ $[+0.39, +1.21]$
ResNet-50 (sup)	-7.05 $[-7.98, -6.13]$	1289	$2e{-}50$	$+0.04$ $[+0.00, +0.32]$
BEiT	-8.60 $[-9.37, -7.83]$	933	$2e{-}108$	$+0.00$ $[+0.00, +0.03]$
SigLIP	-9.40 $[-10.55, -8.25]$	2003	$6e{-}57$	$+0.99$ $[+0.47, +1.56]$
CLIP ViT-B/16	-10.28 $[-11.38, -9.18]$	1861	$2e{-}73$	$+0.44$ $[+0.12, +0.87]$
MAE	-43.68 $[-44.88, -42.48]$	3556	$< 10^{-300}$	$+0.03$ $[+0.00, +0.33]$

The fixed-weight and oracle-envelope results expose different aspects of performance. At $w{=}10$, low-purity retrieval spaces can be strongly detrimental: CLIP ViT-B/16 decreases accuracy by 10.28 points, while MAE decreases it by 43.68 points. By construction, the oracle envelope cannot be negative because $w{=}0$ is included in $\mathcal{W}$. Thus, an envelope gain of $+0.00$ indicates that none of the evaluated weights improves over the no-cache predictor; it does not imply that every nonzero weight is safe. Reporting both quantities is therefore necessary to characterize the deployment risk illustrated in Fig. A7.

The ceiling is an attainable upper envelope, not the fixed deployment outcome. At the reported fixed w = 10, seven of sixteen ImageNet A retrieval spaces reduce accuracy, including CLIP and MAE, as shown in Table Fig. A7.

Results with no detectable effect. Across the 378 tabulated configurations, 50 paired 95% confidence intervals include zero. We consistently interpret these cases as showing no detectable effect, rather than as evidence of a gain or loss. They occur primarily for weaker retrieval spaces and smaller cross-domain evaluation sets. Within the sixteen-space ImageNet-A comparison, DINOv2-S is the only such case: $+0.53$ points $[-0.24, +1.31]$, with $p{=}0.19$, although its oracle envelope is positive at $+1.24$ $[+0.85, +1.75]$. In Table A16, the effects for Caltech101 (-0.04 $[-0.95, +0.87]$, $p{=}1.00$) and DTD ($+0.83$ $[-1.18, +2.83]$, $p{=}0.45$) are not distinguishable from zero at either $M{=}4$ or $M{=}64$, whereas those for EuroSAT, Flowers102, Pets, and UCF101 are. The additional streams in Table A17 instead show statistically resolved losses: -7.92 points $[-9.18, -6.66]$ on Aircraft and -4.74 points $[-5.08, -4.40]$ on Food101, paired per image and averaged over three seeds, consistent with the transfer risk identified in that analysis.

Multiplicity adjustment. We assess multiplicity across the 378 reported tests. Benjamini–Hochberg control at $q{=}0.05$ rejects 328 null hypotheses the same set identified at the uncorrected $p < 0.05$ threshold corresponding to a nominal false-discovery allowance of 16.4 among these rejections. At $q{=}0.01$, 314 hypotheses remain rejected. The more conservative Bonferroni correction at family-wise level 0.05, corresponding to a per-test threshold of 1.32×10^{-4}, retains 296 rejections. The primary comparisons are significant under each correction, often by several orders of magnitude. Moreover, these comparisons are specified by the intervention design rather than selected post hoc from the full set of configurations.

Matched-order comparison with COSMIC. The original $M{=}8$ COSMIC evaluation did not retain the per-example predictions required for paired inference. We therefore evaluated both systems

on three recorded stream orders. Within each order, the methods receive the same permutation and label sequence over all 7,500 images. The resulting matched comparison is reported in Table A28.

Table A28: Matched comparison of MARC and COSMIC on ImageNet-A at M=8 and DINOv2-B scale. Each row uses an identical recorded stream order for both methods. For the pooled estimate, we first average each image's paired difference across the three orders and then construct the interval, so that each image contributes once. Treating all 22,500 predictions as independent would underestimate uncertainty because the same 7,500 images appear in every order.

Order	MARC	COSMIC	paired advantage	disc.	McNemar p
0	63.56	61.80	+1.76 [+0.99, +2.53]	878	9.5e−06
1	63.89	63.52	+0.37 [−0.41, +1.16]	906	0.37
2	64.04	62.81	+1.23 [+0.43, +2.03]	942	3.0e−03
Pooled, clustered by image	63.83	62.71	**+1.12** [+0.66, +1.58]		

Three aspects of this comparison merit clarification. First, the recorded orders used here differ from those in Tables 2 and A30. Because test-time adaptation is order dependent, the corresponding accuracies differ slightly from the values 63.95/64.53/64.03 and 62.89/62.60/62.76 reported in those tables. Second, the seed interval for the mean difference across these three matched orders is +1.12 [−0.62, +2.86] and includes zero. With n=3, the critical value is $t_{.975,2} = 4.303$, so uncertainty over stream order remains substantial even though the image-paired interval excludes zero. These intervals address distinct sources of variation and should be interpreted jointly. Third, COSMIC's reported accuracy is the maximum over a 14×14 grid of fusion weights selected on the evaluation stream, whereas MARC uses the prespecified value w=10. The comparison therefore favors COSMIC with respect to hyperparameter selection and is not fully matched in that dimension.

Across these matched orders, the standard deviation is 0.86 percentage points for COSMIC and 0.25 for MARC. The smaller ±0.15 dispersion reported for COSMIC in Table 2 characterizes its three evaluation orders in that experiment; it should not be interpreted as a general estimate of sensitivity to stream order.

Rank correlations. We quantify uncertainty in the rank correlations from Table 1 using the bootstrap procedure of Table A7. Each of B=300 replicates samples calibration images with replacement and recomputes the geometry statistic, while holding the encoder set and measured gains fixed. Table A29 reports the results at the largest calibration size considered. On both streams, the intervals for purity and anti-hubness are well separated, supporting the conclusion that purity more reliably ranks attainable gain.

Table A29: Calibration-subsample intervals for the rank correlations in Table 1, computed over sixteen retrieval spaces using B=300 bootstrap replicates at n=2000. Full-split point estimates use midranks throughout.

Stream	Measurement	full split	n=2000 [95% interval]
ImageNet-A	$\rho(\text{Purity @10}, G^\star)$	+0.959	+0.957 [+0.956, +0.962]
	$\rho(-\text{AntiHub}, G^\star)$	+0.541	+0.575 [+0.472, +0.682]
ImageNet-V2	$\rho(\text{Purity @10}, G^\star)$	+0.943	+0.937 [+0.906, +0.958]
	$\rho(-\text{AntiHub}, G^\star)$	+0.756	+0.809 [+0.743, +0.883]

Two considerations govern the interpretation of these intervals. First, anti-hubness measures the fraction of zero-indegree points in an n-point probe graph and is therefore intrinsically sensitive to calibration size. Its subsample interval characterizes the n=2000 regime; it is not intended to contain the full-split estimate. Second, the bootstrap resamples images, not encoders. The sixteen retrieval spaces were deliberately selected to span distinct training regimes and do not constitute a random sample from a defined encoder population. Accordingly, the correlations are descriptive summaries of the observed ranking. The more directly decision-relevant quantity is the selection regret in Table A7, which is zero on both streams at every calibration size considered.

Latency. Timing ratios are consistent across the three sessions in Table A31. Computing each ratio within a session and then constructing an interval across sessions gives a latency ratio of 2.58×

$[2.52, 2.64]$ for COSMIC relative to MARC at $M{=}8$, and $5.06\times$ $[4.95, 5.18]$ for the 64-view ensemble relative to single-view MARC-L.

Results without uncertainty estimates. We omit confidence intervals when the available evidence does not support a meaningful estimate. This applies to three groups of results. First, the 216 literature values in Tables 2, A16 and A18 are reproduced from prior work under the respective evaluation protocols; without access to the underlying per-example predictions, we treat them as benchmark context rather than subjecting them to retrospective uncertainty analysis. Second, Tables A23 and A26 report descriptive attributes rather than performance estimates. Third, several configurations are represented by a single run or two stream orders, including the ImageNet-V2 column of Table 1 and the transfer column of Table A1. For these configurations, we report paired sampling intervals over evaluation images when paired predictions are available, but do not report seed intervals. At $n{=}2$, the critical value $t_{.975,1} = 12.706$ yields a half-width of approximately nine sample standard deviations, providing little information about variability across stream orders. We therefore do not use two-order dispersion as inferential evidence.

S TIMING REPRODUCIBILITY

Table A30: Tier (ii), same-budget system comparison: ImageNet-A accuracy versus cost on one RTX 3060, with all rows timed in one session. Green marks the compute–accuracy Pareto frontier. The CLIP reference uses standard prompts; controlled gains use the 50.84% prompt-ensemble baseline. The MARC and COSMIC $M{=}8$ accuracies each average three seeds; the latency ratio is repeated across sessions in Table A31. The remaining MARC rows are the seed-0 runs of the timing session and therefore differ slightly from the three-seed means quoted in Table 2 and in the text.

Configuration	views	s/img ↓	Top-1 ↑
Baselines, same GPU and same session			
CLIP zero-shot (standard prompts)	1	0.021	47.87
Ensemble only, no cache	64	0.310	58.77
TDA (Karmanov et al., 2024)	1	0.190	60.11
COSMIC (Huang et al., 2025) (DINOv2-B)	8	0.181	62.75
MARC: view budget swept at fixed retrieval space			
MARC (DINOv2-B)	1	0.045	59.73
MARC (DINOv2-B)	4	0.053	62.93
MARC (DINOv2-B)	8	0.070	$64.17_{\pm 0.31}$
MARC (DINOv2-B)	64	0.338	65.23
MARC: retrieval space swept at fixed view budget			
MARC (DINOv2-L)	1	0.061	68.99
MARC (DINOv2-L)	4	0.069	70.69
MARC-L $M{=}1$ vs. 64-view ensemble		5.1×	+10.22

Table A30 measures all rows within one session for direct comparison. Two further sessions on an otherwise idle GPU replicate the timing ratios.

Table A31: Steady-state s/img across three independent timing sessions. Session 1 is the one reported in Tab. A30.

Configuration	S1	S2	S3	spread
CLIP zero-shot	0.0207	0.0279	0.0207	0.0072
Ensemble only, $M{=}64$	0.3101	0.3272	0.3003	0.0269
TDA (Karmanov et al., 2024)	0.1903	0.1918	0.1893	0.0025
COSMIC (Huang et al., 2025), $M{=}8$	0.1812	0.1733	0.1787	0.0079
MARC (DINOv2-B), $M{=}1$	0.0450	0.0447	0.0445	0.0005
MARC (DINOv2-B), $M{=}8$	0.0701	0.0679	0.0688	0.0022
MARC (DINOv2-B), $M{=}64$	0.3378	0.3323	0.3355	0.0055
MARC (DINOv2-L), $M{=}1$	0.0614	0.0640	0.0598	0.0042

Every row varies by at most 0.008 s/img across sessions except the ensemble-only row, whose spread is 0.027 s/img. The CLIP reference meets the absolute bound but has the largest relative spread, at 31% of its mean. At roughly 21 ms per image, it is particularly sensitive to fixed per-batch and session-level overhead; no claim depends on this row. The two ratios used in the paper are stable: the matched comparison against COSMIC is $2.6\times$, $2.6\times$, and $2.6\times$ across the three sessions, and single-view MARC-L against the 64-view ensemble is $5.1\times$, $5.1\times$, and $5.0\times$. Constructing intervals from the session-level ratios gives $2.58\times$ $[2.52, 2.64]$ and $5.06\times$ $[4.95, 5.18]$, respectively.

Matched-budget accuracy over seeds. For the comparison in Sec. 5.2, COSMIC reaches 62.89, 62.60, and 62.76 on ImageNet-A at $M{=}8$ over three seeds, yielding a mean of 62.75 ± 0.15. MARC reaches 63.95, 64.53, and 64.03, yielding 64.17 ± 0.31. The paired system advantage is 1.42 ± 0.45 points at the same view budget and auxiliary scale. These runs stored no per-sample predictions for COSMIC, so the advantage they support is a seed-level one; App. R reports a matched re-run of both systems over recorded stream orders, which supplies the per-image paired interval and the exact test.

T Open Questions and Future Research Directions

Our results show retrieval space as an independent design axis, but also raise broader questions about how memory-based adaptation should interact with representation geometry, target shift, and inference compute. Here, we discuss several open questions arising from our findings that may guide future research in this area:

✓ **Can retrieval quality be predicted without labels?** Purity strongly tracks attainable cache gain when candidate spaces are sufficiently separated, while pseudo-purity shows that much of this signal can be recovered without ground-truth labels. The remaining challenge is to identify a fully online criterion that predicts not only neighborhood consistency but whether retrieved evidence will *correct* rather than reinforce the frozen predictor. Such a criterion could combine neighborhood agreement, similarity margins, hubness, prediction uncertainty, and temporal stability. A particularly interesting direction is to estimate the expected utility of a retrieval space directly from the unlabeled stream, allowing systems to decide when a cache is likely to help before committing memory or compute.

✓ **Should retrieval space itself adapt over time?** Our experiments select one frozen retrieval encoder for an entire stream, yet the best space varies across datasets and shifts. Real deployments may be more heterogeneous: a stream can move between environments, acquisition conditions, or semantic regimes over time. This motivates dynamic retrieval systems that select among multiple frozen representations, combine complementary spaces, or locally reshape similarities as the stream evolves. Rather than adapting the predictor, future TTA systems could therefore adapt the *geometry of memory access*, potentially preserving model stability while remaining responsive to non-stationary environments.

✓ **What makes a representation a good cache space?** DINOv2 performs strongly in our experiments, but encoder family or scale alone does not fully explain cache effectiveness. The contrast among CLIP, DINO, DINOv2, MAE, and other representations suggests that useful cache spaces must preserve class-relevant local neighborhoods while avoiding similarity structures that concentrate retrieval on a small set of samples. Purity and anti-hubness capture parts of this behavior, but neither fully characterizes it. Understanding which representation-learning objectives produce favorable memory geometry could connect cache-based adaptation to a broader theory of representation quality beyond linear probing and standard k-NN evaluation.

✓ **How should memory capacity depend on geometry?** Our results suggest that useful cache capacity is representation-dependent rather than a purely system-level hyperparameter. A space with clean, distributed neighborhoods may benefit from retaining more examples, whereas a hub-dominated or contaminated space may saturate earlier. This raises the possibility of geometry-aware memory management in which capacity, eviction, and admission are chosen from observed neighborhood statistics instead of fixed globally. Such mechanisms could allocate memory selectively across classes, regions of feature space, or phases of a changing stream.

✓ **When should retrieval replace additional inference compute?** The diminishing cache gain with increasing view count indicates that retrieval and augmentation provide partly overlapping

evidence. This suggests a broader resource-allocation problem: for each input, should additional compute be spent on more views, deeper retrieval, a larger memory, another retrieval encoder, or no adaptation at all? An adaptive controller could allocate inference compute based on prediction uncertainty and retrieval confidence, using cheap single-view retrieval for easy corrections while reserving expensive augmentation for ambiguous cases. Such conditional computation may be especially useful in latency-, energy-, or throughput-constrained deployments.

✓ **When should the system refuse to use its memory?** Cross-domain results show that cache utility varies substantially across streams, including settings where its contribution is negligible. This is not only a limitation but an opportunity to make cache-based adaptation selective. Future systems could estimate whether the current neighborhood contains sufficiently reliable evidence and suppress cache fusion when it does not. Moving from 'always retrieve" to 'retrieve when useful" could improve robustness under weak neighborhood structure, abrupt shifts, or early stages of a stream when memory is sparse.

✓ **Can multiple retrieval geometries be complementary?** Our controlled study intentionally varies one retrieval encoder at a time to isolate its effect, but different representations may encode complementary notions of similarity. A semantic representation may retrieve class-consistent examples, while another space may better preserve texture, shape, acquisition conditions, or domain-specific structure. Future work could study mixtures of retrieval spaces without collapsing them into a single global similarity, including query-dependent routing or agreement-aware fusion. The central challenge is to obtain complementary evidence without reintroducing the complexity that our reduced design shows is often unnecessary.

✓ **How does retrieval geometry interact with real non-stationarity?** Our online protocol provides a controlled setting for studying memory trajectories, but real streams may contain recurring domains, gradual drift, abrupt transitions, class imbalance, or newly appearing categories. In such settings, the appropriate retrieval geometry and useful memory horizon may change together. Studying this interaction could lead to memories that detect distribution transitions, retain reusable evidence across recurring environments, and forget neighborhoods that no longer reflect the current stream.

✓ **Can the principle extend beyond natural-image classification?** The separation between prediction space and retrieval space is not inherently specific to CLIP or ImageNet. Remote sensing, scientific and medical imaging, embodied perception, and other specialized visual domains often combine pretrained models with streams whose local structure differs from the original training distribution. These settings provide natural tests of whether domain-specific or self-supervised retrieval representations can complement general-purpose predictors. More broadly, the same question arises wherever a frozen model consults an external memory, including multimodal and retrieval-augmented systems: the representation that is best for producing an output need not be the representation that is best for finding the evidence used to produce it.

Taken together, these directions suggest a shift from viewing the cache as a fixed auxiliary component toward treating memory access as an adaptive inference problem. The next generation of cache-based TTA may therefore ask not only *what should be stored?*, but also *in which geometry should it be retrieved, when should that geometry change, and when should retrieved evidence be trusted?*