WHEN LESS RETRIEVES MORE: SELECTIVE TOKEN RETENTION IN VISUAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Can discarding visual patch vectors produce a better retrieval representation than aggregating them? We study this question with frozen encoders, full query tokens and exact MaxSim scoring, developing the study on revisited Oxford. Retaining 39 of 196 database patches by CLS attention improves DINOv2 Medium mAP from 63.91% to 72.30%, with 80.1% fewer stored patch vectors. Holding those attention-selected anchors fixed, averaging all patches into their nearest anchors reduces mAP to 69.13%; normalizing those means reduces it further to 64.57%. Retention also outperforms nine specified hierarchical-pooling adaptations at the same exact count. An additional Paris evaluation improves Medium mAP from 85.92% to 88.23% and Hard mAP by 4.20 points relative to the full representation. The attention-versus-full gain also recurs with DINOv3 on Oxford, but dense CLIP changes by -1.61 points; native DINOv3 CLS remains a strong single-vector alternative. Ordering diagnostics show asymmetric repair and damage, while CLIP demonstrates that improved mean margins against fixed competitive negatives do not suffice to explain improved mAP. The contribution is a controlled selection–aggregation contrast in DINOv2 patch retrieval, not a new compression algorithm or a universal benefit of deletion. The Paris evaluation extends the direction of the attention-versus-full gain, not the Oxford-only same-anchor finding.

1 INTRODUCTION

A database compressor is often evaluated against an uncompressed representation: does it preserve the reference scores or rankings with fewer vectors? For a frozen representation not trained for its retrieval scorer, the reference itself may make avoidable relevance errors. This raises a different empirical question: when compression improves retrieval, does aggregating the discarded vectors preserve the gain, or does retaining selected vectors work better? The answer matters when choosing between a compact subset and synthesized representatives at the same stored-vector count within a fixed patch interface.

We examine patch-based image retrieval. Each query patch matches its most similar database patch, and MaxSim sums the matches (Khattab & Zaharia, 2020). We keep every query patch and intervene only on database representations. On our Oxford development protocol, selecting 39 of 196 DINOv2 patches by CLS attention raises Medium mAP from 63.91% to 72.30%. This is not a benefit of arbitrary downsampling: random, spatial, norm-based and original clustering controls all underperform attention at the same count.

Our central control fixes the attention-selected anchors and changes what is stored at them. Retaining the original vectors reaches 72.30% Medium mAP. Assigning every patch to its nearest anchor and storing raw cluster means reaches 69.13%; unit-normalizing those means reaches 64.57%. Both use the same 39 anchors and full queries. The chosen anchors alone therefore do not explain the retention result. Nor does aggregation remove the entire gain: raw pooling still beats the full representation. The evidence supports an operator-specific ordering, not a claim that discarded patches are universally harmful.

A later Paris evaluation gives an additional test of the attention-versus-full comparison: Medium improves by 2.31 points and Hard by 4.20. It was conducted after Oxford development, not as untouched confirmation. The same-anchor and Ward controls remain Oxford-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.

Two boundaries constrain interpretation. Attention pruning improves the tested DINOv3 patch representation on Oxford, but not the dense CLIP adaptation. Native DINOv3 CLS retrieval remains competitive with attention on Medium and better on Hard. A patch-query/CLS-database bridge confirms that query interface matters, without establishing a practical advantage for multi-vector retrieval. These results delimit the scientific observation rather than being exceptions to a presumed universal rule.

Our contributions are empirical:

- Fixed-query, exact-score attention-pruning evidence on Oxford and an additional Paris evaluation, with two DINO generations and a dense CLIP boundary on Oxford. Query and landmark losses remain reported.
- A retention-versus-aggregation comparison at identical attention anchors, complemented by nine exact-count hierarchical-pooling variants and a separate CLS query-interface control.
- An ordering analysis with explicit populations at risk, fidelity measurements and a counterexample to a simple mean-margin explanation.

Attention selection, pooling, and the distinction between reference fidelity and relevance are established ideas (Noh et al., 2017; Liu et al., 2026; Mahdizadeh et al., 2026; Jha et al., 2026). Our contribution is the bounded empirical contrast and its controls, not priority for those ideas, a new theorem or a predictor of when to prune.

2 Retrieval setting and interventions

For query vectors $Q = (q_1, \ldots, q_m)$ and database vectors $D = (d_1, \ldots, d_n)$, MaxSim is

$$S(Q, D) = \sum_{i=1}^{m} \max_{1 \le j \le n} q_i^\top d_j. \tag{1}$$

Patch vectors are L2-normalized before float16 storage and restored to float32 for scoring. Original full representations have $m = n = 196$. A database operator C acts once at indexing time. Patch comparisons keep all $m = 196$ query vectors and exactly $|C(D)| = 39$ stored vectors. Neither relevance labels nor test-query scores enter the attention selector.

Retention and same-anchor aggregation. Let A contain the k highest final-layer CLS-to-patch attention values, averaged across heads. Stored attention ties use native patch order. Retention stores $(d_a)_{a \in A}$. Aggregation assigns each non-anchor patch by cosine similarity to its nearest anchor. Each anchor belongs to its own cluster, ensuring nonempty clusters. For cluster V_a, the raw representative is $c_a = |V_a|^{-1} \sum_{j \in V_a} d_j$; the unit variant stores $c_a / \|c_a\|_2$. The anchor set, assignment rule and count remain fixed. Only retention versus these specified synthesized representatives changes. Retention reuses its unchanged original scores.

Ordering corrections. Define score loss $L_C(Q, D) = S(Q, D) - S(Q, C(D))$. For a relevant image D^+ and an irrelevant image D^-, the score-margin change satisfies

$$\Delta M_C = L_C(Q, D^-) - L_C(Q, D^+). \tag{2}$$

A correction moves a relevant image above an irrelevant image that outranked it in the full index; damage reverses a previously correct order. This identity is accounting, not theoretical novelty or a semantic explanation. A selected-pair average does not determine AP over the full ranking. Subset selection cannot increase MaxSim in exact arithmetic. Unnormalized convex representatives share this bound; unit normalization need not (Appendix B). This motivates separate raw and unit controls without identifying a cause of relevance improvement.

3 Experimental protocol

Dataset and metrics. For development we use 4,993 database images and 70 ROI queries from revisited Oxford (Radenovic et al., 2018), excluding the query-source images from the original

image list. Queries form 11 semantic landmark groups. Medium treats easy and hard positives as relevant; Hard uses hard positives and ignores easy positives. We compute official junk-filtered trapezoidal AP over complete rankings, with stable candidate-order ties, and retain mP@1/5/10. This is a controlled Oxford protocol, not an external leaderboard submission. Oxford influenced development, including the operator focus and budget choice. The second-benchmark revisited Paris evaluation uses 6,322 database images and 70 queries with DINOv2 and the same full-query scoring interface. It follows Oxford development; the internal admission gate was bypassed under deadline pressure, and no untouched or preregistered confirmation is claimed. Its completed independent audit covers all 33 conditions and a CPU MaxSim oracle for Full and all three attention arms. We do not pool Oxford and Paris into a new significance test (Appendix I).

Backbones and interfaces. DINOv2-B/14 uses 196×196 inputs and 768-dimensional patches (Oquab et al., 2023). DINOv3-B/16 uses 224×224, 768 dimensions and excludes four register tokens (Siméoni et al., 2025). CLIP ViT-B/16 uses 224×224 and 512-dimensional projected patches (Radford et al., 2021). All yield a 14×14 grid. Dense CLIP applies its visual post-layer normalization and projection to patches; it is not trained for MaxSim retrieval. Checkpoint, preprocessing, projection and training differences confound causal encoder comparisons. CLS and mean descriptors use single-vector queries and databases. A separate DINOv3 bridge keeps the full patch query but uses one native CLS database vector. It is an interface control, not a 39-vector budget match.

Budgets and operators. We compare actual integer counts without duplicate padding. Equal count and dtype within a backbone match coordinate storage, not deployed disk, latency or index overhead. Different embedding widths prevent cross-backbone byte matching. The original DINOv2 study includes $k \in \{20, 39, 98\}$, five random seeds, spatial selection, highest-raw-norm retention, a k-means representative-member coreset and greedy merging. The coreset selects an original member nearest each centroid; it is not globally optimized k-medoids. Greedy merging repeatedly joins the most similar pair with size weights and normalizes the result.

We add same-anchor raw/unit means and nine frozen Ward-clustering variants from a pinned H-Pool kernel (Qin et al., 2026). Three partition rules are crossed with raw means, unit means and original-member representatives: published distance-profile Ward with a maximum-cluster cut, the same profile with an exact-k cut, and unit-Euclidean Ward with an exact-k cut. The first cut permits at most k clusters; every item in this run actually produced 39. All variants remain reported, including coincident profile results. The strongest exact-count variant is selected by Medium, with Hard taken from the same row. These are explicit kernel adaptations, not trained AGC or complete published models. DINOv3 and CLIP do not have the full sweep.

Verification and inference. Scoring is exhaustive MaxSim, not ANN or candidate-truncated scoring. Separate evaluators authenticate arrays, reconstruct stable rankings and recompute official per-query metrics. The new control audit additionally reconstructs representatives and MaxSim on a fixed three-query/seven-image source subset, not the full corpus or encoder. CLIP's numerical replay passes, but an inconsistent extraction-time receipt leaves its overall protocol status provisional (Appendix H).

Inference here is exploratory development analysis. Exact two-sided tests enumerate 2^{11} landmark sign assignments of paired AP differences, preserving the query-weighted estimand. Marginal 95% intervals use 10,000 whole-landmark bootstrap draws, seed 2509. Holm adjustment covers six backbone-by-protocol attention/full contrasts. A separate family covers all 22 DINOv2 attention/control contrasts, not only the best pooling row. These do not replace earlier frozen C1–C4 tests. Group independence and sign symmetry remain assumptions; eleven groups limit inference. These intervals describe variation across Oxford landmark groups; they do not account for development-driven choices or establish transfer. Dependent query-image pairs are not independent binomial trials.

4 Results: selection, aggregation and representation

The gain recurs in DINO, not in every representation. DINOv2 improves from 63.91% to 72.30% Medium and 39.08% to 49.92% Hard. DINOv3 improves from 66.98% to 75.65% and 42.73% to 55.40%. Each DINO contrast is positive in ten landmark means and tied in one, with

Operator	k	Oxford M	Oxford H	Paris M	Paris H
Full	196	63.91	39.08	85.92	69.42
Attention20	20	70.85	49.61	88.23	73.94
Attention39	39	72.30	49.92	88.23	73.62
Attention98	98	70.58	50.26	86.99	71.32
Frozen greedy merge39	39	62.00	37.71	85.66	69.30
Medoid39	39	60.00	34.79	85.37	68.95
CLS	1	63.80	43.66	88.39	75.20

Table 1: DINOv2 Medium/Hard mAP (%). Patch rows keep all query patches; CLS changes to single-vector queries and databases. Paris is second-benchmark evaluation after Oxford development. Its frozen merge comparator is the original greedy merge, not the later strongest Oxford pooling control. Budgets were not selected by the Paris results.

Dataset	Encoder	Full M	Att39 M	ΔM	ΔH
Oxford	DINOv2	63.91	72.30	+8.39	+10.84
Oxford	DINOv3	66.98	75.65	+8.67	+12.67
Oxford	CLIP*	37.14	35.53	-1.61	-2.39
Paris	DINOv2	85.92	88.23	+2.31	+4.20

Table 2: Full versus Attention39: Medium mAP (%) and paired aggregate changes in Medium/Hard points. *CLIP numerical replay passes; its timing receipt leaves overall protocol compliance provisional. Paris values come from the completed independent audit. No cross-backbone byte match is implied.

individual query losses retained (Appendix G). All four DINO tests survive the six-test Holm adjustment, each with adjusted $p = 0.01171875$.

Dense CLIP changes from 37.14% to 35.53% Medium and 11.06% to 8.68% Hard. Its grouped intervals include zero and adjusted tests do not resolve harm. Negative point estimates nevertheless prevent claiming improvement for every tested representation. They do not establish that pruning generally harms native CLIP.

The direction extends to Paris, not yet the full control story. On Paris, Attention39 improves Medium from 85.92% to 88.23%, a 2.31-point gain. Hard improves from 69.42% to 73.62%, a 4.20-point gain. All three attention budgets improve Medium and Hard over Full (Table 1). Attention20 and Attention39 are close in Medium point estimates; this is not an equivalence result or grounds to retune the selected budget. The frozen greedy-merge comparator does not reproduce the gain. Same-anchor pooling and the stronger Ward variants were not evaluated on Paris; their Oxford ordering is not promoted to a cross-dataset claim.

Same-anchor aggregation weakens retention. Raw pooling reaches 69.13% Medium and 46.08% Hard, below retention by 3.17 and 3.84 points. Unit pooling reaches 64.57% and 40.14%, gaps of 7.73 and 9.78 points. Figure 1 shows both protocols; every paired control contrast is retained in Appendix F. The raw-pooling Medium gap has a marginal grouped 95% interval of [1.51, 5.15] points and a Holm22-adjusted exploratory $p = 0.023438$. The corresponding Hard interval is [1.80, 5.70] points. These are development contrasts, not independent confirmation of a universal operator ranking. Holding anchor choice fixed, these two aggregation rules do not preserve the full retention gain. This does not establish semantic irrelevance of the omitted patches or necessity of deletion: raw pooling remains above Full196.

Beyond the original weak merge control. The strongest Medium result among the nine hierarchical variants is unit-Euclidean exact-k clustering with unit means: 61.13% Medium and 36.62% Hard. All fall below retention, but this is not superiority to learned synthesis or every published pooling system. Same-anchor raw pooling is the strongest tested aggregation control, above both this Ward winner and original greedy merge. The anchor choice and normalization must therefore be specified rather than treating all synthesized representations as one operator.

Operator	k	Medium	Hard
Full	196	63.91	39.08
Attention retention	39	72.30	49.92
Same-anchor raw mean	39	69.13	46.08
Same-anchor unit mean	39	64.57	40.14
Best Ward variant	39	61.13	36.62
Original greedy merge	39	62.00	37.71
Original member coreset	39	60.00	34.79
Random, five-seed mean	39	54.69	28.19

Table 3: DINOv2 full-query comparison. Each compressed row stores exactly 39 vectors/image. Best Ward is selected by Medium among all nine specified variants, with Hard from the same row. Random averages all five seeds. These are coordinate-count comparisons, not measured index-system savings.

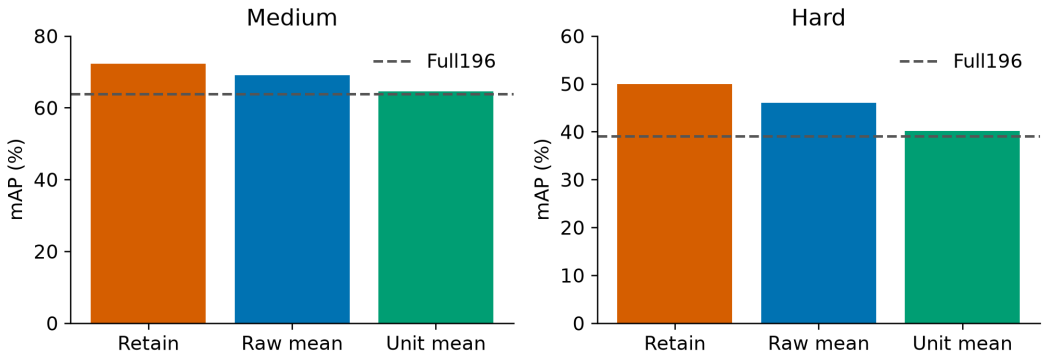

Figure 1: Same attention anchors, different stored representatives. Retention, raw means and unit means share 39 anchors and full query tokens. Full196 is the reference line. Aggregation lowers the retention result; raw means retain a benefit over Full. Bars are descriptive; grouped inference is in the appendix.

CLS remains a practical challenger. DINOv3 native CLS reaches 75.68% Medium and 57.71% Hard, against attention's 75.65% and 55.40%. Keeping database CLS but using full patch queries gives 60.84% and 41.16%. Changing the query interface materially changes the result; replacing only database patches with CLS does not reproduce native CLS retrieval. This does not establish practical superiority for the 39-vector system over native single-vector retrieval. Paris native CLS reaches 88.39% Medium and 75.20% Hard, above Attention39's 88.23% and 73.62% (Table 1). This further limits a practical-superiority claim. For a single database vector c, MaxSim collapses to $S(Q, \{c\}) = (\sum_i q_i)^\top c$. The bridge therefore probes the summed patch query against CLS database vectors, whereas native CLS changes the query direction itself. Its loss cannot be attributed to a query-token count effect.

5 What changes in the ranking?

Repair and damage have different denominators. For each query, fix the first five eligible irrelevant images in the full ranking and pair each with every relevant image. Repair divides corrected pairs by initially wrong pairs; damage divides damaged pairs by initially correct pairs. Stable rank order, including the original tie rule, determines events. The two rates have different populations at risk.

DINOv2 Medium attention repairs 3,611/11,732 initially wrong pairs (30.78%) and damages 449/15,898 initially correct pairs (2.82%). DINOv3 gives 24.00% and 2.80%; CLIP gives 12.42% and 11.05%, with a less favorable Hard pattern (Figure 2). These are pooled rates in a fixed competitive stratum, not independent trials or estimates over all relevant–irrelevant pairs. Appendix tables retain counts, undefined denominators and mean query rates.

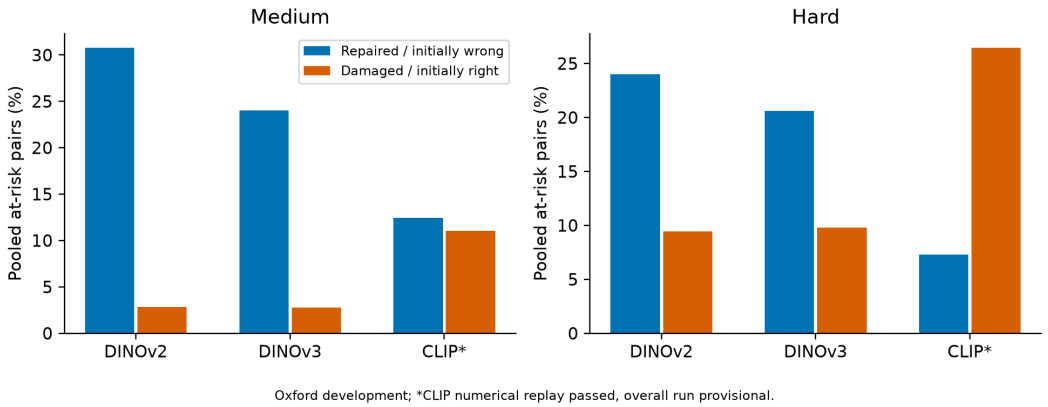

Figure 2: Attention39 repairs/initially wrong pairs and damages/initially right pairs in each full-top-five-negative stratum. Strata differ across backbones and protocols. Pairs are dependent. *CLIP retains its protocol caveat.

Reference agreement reverses the relevance ordering. DINOv2 Medium mean prefix agreement at depths 1–10 is 57.36% for attention, 82.55% for greedy merge and 82.24% for the representative-member coreset. Attention has the highest mAP despite the lowest reference agreement among these operators. The appendix also reports top-10 overlap, average-rank Spearman and score distortion. This is a contrast among tested operators, not a new discovery that fidelity and relevance differ (Liu et al., 2026; Mahdizadeh et al., 2026), or an experiment demonstrating that a trained teacher-distillation objective fails.

Positive mean competitive margins are insufficient. Equation 2 compares mean relevant-image loss with loss on fixed full-top-five negatives and, separately, all eligible negatives. The competitive contrast is positive for attention in all three backbones; the all-negative contrast is negative. This does not support preferential suppression of irrelevant images in general.

CLIP provides a counterexample to a stronger explanation. Its Medium mean competitive margin improves by 1.22 score units while mAP changes by -1.61 points. Repairs also outnumber damages in that stratum. Omitted negatives, rank positions and variation across queries matter to AP. Neither the selected-stratum mean nor a repair surplus suffices to explain the DINO gain. These diagnostics describe where corrections occur; they do not identify a causal semantic mechanism or an index-time predictor.

6 Related Work

Established selection and aggregation. DELF learns attentive local features, DELG combines local and global representations, and HOW studies descriptor learning and aggregation (Noh et al., 2017; Cao et al., 2020; Tolias et al., 2020). MaxSim follows late interaction (Khattab & Zaharia, 2020). Light-ColPali studies patch-embedding reduction, DocPruner adaptive pruning, and Prune-then-Merge filtering and synthesis (Ma et al., 2025; Yan et al., 2025; 2026). Attention pruning and memory–quality tradeoffs are not our invention.

Closest comparisons. Jha et al. compare selected-anchor pruning and pooling (Jha et al., 2026); the fixed-anchor design is therefore not new. Our DINOv2 result supplies a controlled opposite ordering to their pooling-favoring settings, not a general refutation of pooling. Qin et al. distinguish heuristic H-Pool from trained attention-guided clustering, AGC (Qin et al., 2026). We test kernel adaptations, not trained AGC. MarginMerge includes anchor/synthesis and ranking-objective controls (Mahdizadeh et al., 2026); its learned synthesis is outside our comparison.

Fidelity and representation artifacts. Structural Anchor Pruning explicitly distinguishes score retention from relevance (Liu et al., 2026). High-norm artifacts and registers are established repre-

sentation concerns (Darcet et al., 2023); our checkpoint contrasts do not isolate their causes or the training objective. The remaining contribution is the controlled Oxford observation: attention retention exceeds same-anchor raw/unit aggregation in one DINO interface. The full-index gain recurs in another DINO checkpoint and in additional DINOv2 Paris evaluation, while dense CLIP and strong CLS delimit its implications.

7 LIMITATIONS

Oxford influenced development; Paris was evaluated later with an admission-gate deviation, not as untouched confirmation. The same-anchor and strong-pooling comparison remains single-corpus. Two DINO generations do not establish broad family transfer, and no result predicts which new dataset or encoder will benefit. Dense CLIP is not native text–image retrieval; kernel adaptations are not trained pooling systems. Representative directions, norms and assignments are not separately randomized interventions. We do not identify semantic noise or a universal deletion mechanism. Strong DINOv3 CLS limits the practical case.

Saved-score audits can share upstream feature errors. The bounded control reconstruction is not full-corpus independent rescoring. CLIP timing and DINOv2 numerical-gate failures remain in Appendix H. We claim no SOTA, deployed-byte or latency benefit, new compression algorithm, or equivalence from a nonsignificant test.

8 CONCLUSION

Retaining 39 original DINOv2 attention vectors improves over the full 196-vector index on Oxford and in additional Paris evaluation. On Oxford it also beats aggregating all patches into those same anchors. The result depends on the operator and interface: raw pooling retains some improvement, dense CLIP does not show the gain, and native DINOv3 CLS is a strong alternative. This motivates controlled relevance evaluation of retention and synthesis, with explicit checks on what ranking diagnostics explain. It does not imply that less information or lower fidelity is generally better.

REPRODUCIBILITY STATEMENT

Sections 2–3 define the interfaces and interventions. The appendix gives checkpoint revisions, preprocessing, all original/pooling conditions, grouped inference and numerical failures. Figures and tables derive from hash-bound per-query evidence. Separate evaluators reconstruct rankings and official metrics; the control audit reconstructs a fixed source subset. Full-corpus re-extraction and a clean anonymous artifact build are not claimed.

AI USE STATEMENT

AI assistants were used to refine hypotheses, design experiments, formulate the mathematical framing, search literature, develop code and tests, check evidence, and revise the manuscript. They were not research subjects. Human authors are responsible for verifying results, citations, originality and the final submitted text.

REFERENCES

Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, 2020. URL https://arxiv.org/abs/2001.05027.

Timothee Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv:2309.16588*, 2023. URL https://arxiv.org/abs/2309.16588.

Rohan Jha, Chunsheng Zuo, Reno Kriz, and Benjamin Van Durme. A brief comparison of training-free multi-vector sequence compression methods. *arXiv:2603.22434v1*, 2026. URL https://arxiv.org/html/2603.22434v1.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *arXiv:2004.12832*, 2020. URL https://arxiv.org/abs/2004.12832.

Zhuchenyang Liu, Ziyu Hu, Yao Zhang, and Yu Xiao. Structural anchor pruning: Training-free multi-vector compression for visual document retrieval. *arXiv:2601.20107v3*, 2026. URL https://arxiv.org/html/2601.20107v3.

Yubo Ma, Jinsong Li, Yuhang Zang, Xiaobao Wu, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Jiaqi Wang, Yixin Cao, and Aixin Sun. Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings. In *Findings of ACL*, pp. 19568–19580, 2025. URL https://aclanthology.org/2025.findings-acl.1003/.

Ailar Mahdizadeh, Aria Salari, Sohail Rajabi, Shahriar Mirabbasi, Panos Nasiopoulos, and Alireza Morsali. Coverage matters: Marginmerge for compressing multi-vector visual document retrievers. *arXiv:2608.02969v1*, 2026. URL https://arxiv.org/html/2608.02969v1.

Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Noh_Large-Scale_Image_Retrieval_ICCV_2017_paper.html.

Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Hanxiang Qin, Alexander Martin, Rohan Jha, Chunsheng Zuo, Reno Kriz, and Benjamin Van Durme. Multi-vector index compression in any modality. *arXiv:2602.21202v1*, 2026. URL https://arxiv.org/html/2602.21202v1.

Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Radenovic_Revisiting_Oxford_and_CVPR_2018_paper.html.

Alec Radford et al. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Oriane Siméoni et al. Dinov3. *arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Giorgos Tolias, Tomas Jenicek, and Ondrej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *ECCV*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1793_ECCV_2020_paper.php.

Yibo Yan, Guangwei Xu, Xin Zou, Shuliang Liu, James Kwok, and Xuming Hu. Docpruner: A storage-efficient framework for multi-vector visual document retrieval via adaptive patch-level embedding pruning. *arXiv:2509.23883*, 2025. URL https://arxiv.org/abs/2509.23883.

Yibo Yan et al. Sculpting the vector space: Towards efficient multi-vector visual document retrieval via prune-then-merge framework. *arXiv:2602.19549v2*, 2026. URL https://arxiv.org/abs/2602.19549v2.

A IMPLEMENTATION CONTRACT

Recorded checkpoints and revisions are:

- `facebook/dinov2-base` `f9e44c814b77203eaa57a6bdbbd535f21ede1415`.
- `facebook/dinov3-vitb16-pretrain-lvd1689m` `5931719e67bbdb9737e363e781fb0c67687896bc`.
- `openai/clip-vit-base-patch16` `57c216476eefef5ab752ec549e440a49ae4ae5f3`.

All inference is frozen evaluation-mode inference. ROI crops precede square PIL bicubic resizing; recorded numeric crop coordinates are truncated to integers. DINO uses means $(.485, .456, .406)$ and standard deviations $(.229, .224, .225)$. CLIP uses its checkpoint normalization. The square-resize recipe is disclosed rather than equated with every checkpoint's standard evaluation protocol.

Attention comes from the final eager-attention layer's CLS-to-patch entries, averaged across heads and min–max scaled per image before storage. Stored attention is float16, so native-index tie handling matters. Original patch vectors are retained without a learned gate. DINOv3 register vectors are excluded. CLS and patches are L2-normalized; raw norms are recorded separately. Mean descriptors average patches and normalize the resulting vector. Patch queries always use their own full ROI features.

The original random seeds are 42, 1051, 2060, 3069 and 4078, not five independent query corpora. The medoid-labelled coreset uses k-means with four initializations. Greedy merging repeatedly renormalizes merged vectors, which distinguishes it from a raw centroid of a fixed original-token partition.

B OPERATOR DEFINITIONS AND LIMITS OF THE CONTROLS

For an unnormalized convex representative of the original database vectors,

$$q^\top \Big(\sum_j \alpha_j d_j \Big) = \sum_j \alpha_j q^\top d_j \le \max_j q^\top d_j, \qquad \alpha_j \ge 0, \quad \sum_j \alpha_j = 1.$$

Maximizing over representatives and summing over query vectors preserves this bound. Subsets are a special case. Unit normalization can move a vector outside the convex hull, and float16 rounding requires numerical checks. This standard inequality is not a retrieval-quality guarantee.

Same-anchor assignment resolves cosine ties by ordered anchors and forces each anchor into its own cluster. The mean includes the anchor. Means use restored original vectors; raw means are not renormalized at scoring. Unit means are normalized before storage. Representatives return to float16, and scores use restored float32 values.

The hierarchical kernel associated with Qin et al. (2026) is pinned to revision `4a559677bbc8a3ea0c10322a721b52bb70d382ec`. The published-profile rule supplies rows of $1 - XX^\top$ to Ward linkage, not a condensed cosine distance vector. We preserve that behavior as a named profile adaptation. The geometric control instead supplies unit-normalized feature rows. The maximum-cluster cut permits fewer than k clusters; exact-k cuts enforce k without padding. Every observed output has 39 clusters. Each partition uses raw means, normalized means or an original member selected nearest its mean by the specified cosine rule. These are frozen patch-interface adaptations, not the complete trained model or AGC, and do not claim to reproduce those systems' published benchmark scores.

C COMPLETE ORIGINAL CONDITIONS

Condition	k	M mAP	M mP@10	H mAP	H mP@10
full_tokens	196	63.91	80.53	39.08	52.40
attn_keep@0.1	20	70.85	89.57	49.61	66.71
attn_keep@0.2	39	72.30	86.71	49.92	64.80
attn_keep@0.5	98	70.58	86.14	50.26	63.64
cls_descriptor	1	63.80	80.86	43.66	55.50
mean_pool	1	39.57	52.57	16.75	24.00
medoid@0.1	20	56.66	73.38	30.80	43.29
medoid@0.2	39	60.00	77.38	34.79	49.16
medoid@0.5	98	63.01	79.60	37.96	51.27
merge@0.1	20	56.44	73.29	28.06	42.24
merge@0.2	39	62.00	79.10	37.71	50.93
merge@0.5	98	63.67	80.50	39.11	51.65
random_keep@0.1_s0	20	48.26	68.86	24.47	37.37
random_keep@0.1_s1	20	49.38	66.46	22.84	35.10
random_keep@0.1_s2	20	50.61	68.43	20.66	35.03
random_keep@0.1_s3	20	49.06	67.29	25.19	35.19
random_keep@0.1_s4	20	47.96	65.71	23.33	35.35
random_keep@0.2_s0	39	52.01	69.20	24.05	36.96
random_keep@0.2_s1	39	54.34	71.75	28.01	41.51
random_keep@0.2_s2	39	55.92	72.10	28.70	40.36
random_keep@0.2_s3	39	55.29	71.38	28.42	41.26
random_keep@0.2_s4	39	55.88	74.07	31.78	43.52
random_keep@0.5_s0	98	60.85	76.81	33.97	46.63
random_keep@0.5_s1	98	60.81	77.96	34.25	48.13
random_keep@0.5_s2	98	61.64	78.03	37.18	49.34
random_keep@0.5_s3	98	61.13	77.32	36.68	49.18
random_keep@0.5_s4	98	61.32	77.81	35.70	48.71
raw_norm@0.1	20	32.41	47.29	16.84	27.19
raw_norm@0.2	39	42.14	61.76	26.83	37.40
raw_norm@0.5	98	59.63	78.95	35.59	45.83
spatial@0.1	20	49.80	67.17	24.76	35.63
spatial@0.2	39	54.95	72.10	29.83	40.43
spatial@0.5	98	61.04	77.82	37.26	49.95

Table 4: All 33 original DINOv2 conditions from the verified intake. Metrics are percentages. CLS/mean use single-vector queries. s0–s4 identify the recorded random seeds. No unfavorable budget or seed is dropped.

D STATISTICAL ESTIMANDS AND PAIRED RESULTS

For query AP differences d_q, the effect is $\bar{d} = \sum_q d_q/N$. Let $s_g = \sum_{q \in g} d_q$. Exact randomization enumerates $N^{-1} \sum_g \epsilon_g s_g$, $\epsilon_g \in \{-1, +1\}$, over all 2^{11} sign patterns. Tail comparisons include ties within 10^{-15}. Duplicate statistic values, including those induced by a zero-effect group, retain their probability mass. Holm adjustment covers only the six exploratory tests below.

Bootstrap draws resample 11 whole groups. With group size c_g, each draw uses $\sum_{g \in \text{draw}} s_g / \sum_{g \in \text{draw}} c_g$, not an unweighted mean of group means. Intervals are marginal percentile intervals, not simultaneous six-contrast intervals. Eleven groups and development exposure limit inference; non-significance is not equivalence.

Encoder	Protocol	Δ pp	Marginal 95% CI	p exact	p Holm	Queries +/–/=
DINOv2	M	+8.39	[5.22, 12.12]	0.001953	0.011719	62/3/5
DINOv2	H	+10.84	[6.42, 15.87]	0.001953	0.011719	63/2/5
DINOv3	M	+8.67	[4.59, 14.45]	0.001953	0.011719	59/6/5
DINOv3	H	+12.67	[5.59, 24.63]	0.001953	0.011719	60/5/5
Dense CLIP*	M	-1.61	[-4.88, 2.49]	0.437500	0.437500	18/52/0
Dense CLIP*	H	-2.39	[-4.52, 0.04]	0.091797	0.183594	22/47/1

Table 5: Attention39 minus Full196. Effects/intervals are mAP points. M/H are Medium/Hard. Query counts are positive/negative/tied AP differences. CLIP remains provisional overall.

Original C1–C4 records retain their Monte Carlo seeds and bootstrap estimands. They are not silently renamed as the six-test family. Manuscript checking found rounding-sensitive equality handling in an archived Monte Carlo tail: group concatenation can change reduction order relative to the observed mean. Reproducing the archived number does not validate its tie handling. Those p-values are not reused as validated inference here. The current exact tests include numerical ties. The random plot is descriptive; the artifact retains prior hierarchical seed/group uncertainty records.

E ORDERING AND FIDELITY DIAGNOSTICS

DB operator	Medium	Hard	Prefix(%)	Spearman
attn_keep@0.2	72.30	49.92	57.36	0.9741
merge@0.2	62.00	37.71	82.55	0.9966
medoid@0.2	60.00	34.79	82.24	0.9954
spatial@0.2	54.95	29.83	68.28	0.9897
raw_norm@0.2	42.14	26.83	46.07	0.9405
Random, five-seed mean	54.69	28.19	68.57	0.9864

Table 6: DINOv2 at $k = 39$. mAP is in percent. Prefix agreement and average-rank Spearman use Medium eligible IDs. Random averages all five seeds, not the best.

Top-heavy agreement is $10^{-1} \sum_{t=1}^{10} |R_{1:t} \cap \widetilde{R}_{1:t}|/t$ after identical junk removal. Spearman uses average ranks for equal scores; constant vectors yield undefined correlation. Normalized absolute distortion divides mean absolute score change by mean absolute full score. Raw retention divides score sums and is undefined near a zero denominator. These are different fidelity diagnostics, not relevance guarantees.

Operator	Top10(%)	Prefix(%)	Spearman	Distortion	Retention
Attention retention	63.57	57.36	0.9741	0.1629	0.8371
Greedy merge	85.86	82.55	0.9966	0.0780	0.9220
Member coreset	85.86	82.24	0.9954	0.1094	0.8906

Table 7: Distinct Medium fidelity diagnostics for original DINOv2 operators at 39 vectors. Distortion and retention are ratios, not percentages.

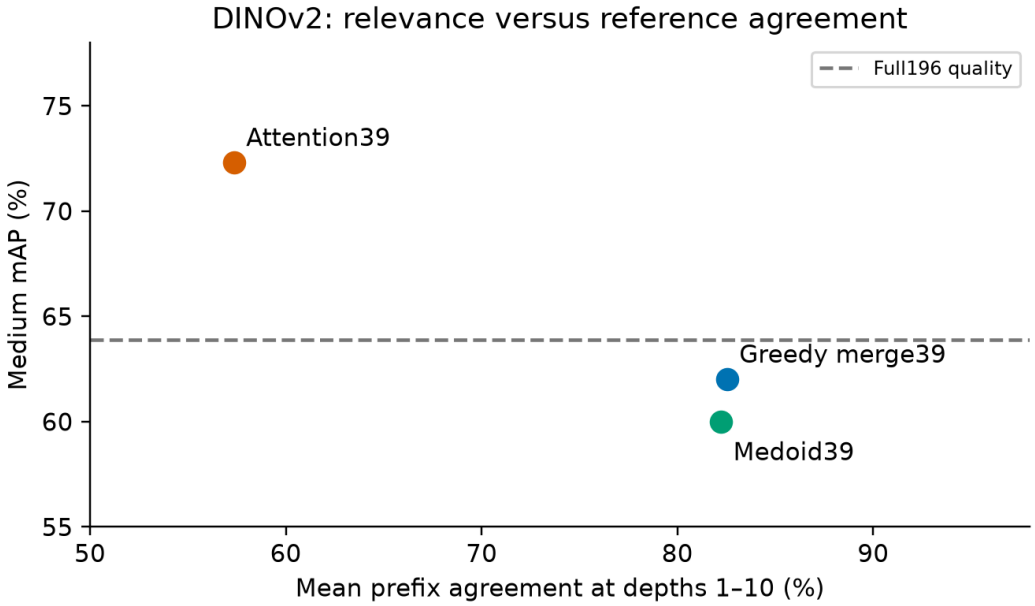

Figure 3: Observed relevance/reference-agreement reversal in DINOv2. These three operator points do not identify a general correlation or a learned objective's behavior.

Encoder	Mode	Repaired/risk	%	Damaged/risk	%	Mean repair (U)	Mean damage (U)
DINOv2	M	3611/11732	30.78	449/15898	2.82	39.24 (5)	3.16 (1)
DINOv2	H	2173/9059	23.99	324/3426	9.46	34.16 (5)	4.57 (8)
DINOv3	M	2669/11121	24.00	463/16509	2.80	33.24 (5)	2.56 (1)
DINOv3	H	1794/8716	20.58	370/3769	9.82	29.43 (5)	4.93 (9)
Dense CLIP*	M	2448/19708	12.42	875/7922	11.05	14.17 (0)	21.61 (8)
Dense CLIP*	H	876/12039	7.28	118/446	26.46	8.02 (0)	27.69 (24)

Table 8: At-risk counts and pooled percentages. Final columns give mean defined per-query percentages, with undefined-query counts U in parentheses. Dependent pairs do not support independent-trial binomial intervals.

Encoder	Mode	L_+	L_{-5}	L_{-all}	ΔM_5	ΔM_{all}
DINOv2	M	16.4248	22.5462	9.6973	6.1214	-6.7275
DINOv2	H	13.3050	22.5462	9.6973	9.2412	-3.6077
DINOv3	M	11.8634	16.6411	9.2676	4.7777	-2.5958
DINOv3	H	9.8348	16.6411	9.2676	6.8063	-0.5672
Dense CLIP*	M	10.3467	11.5695	6.6467	1.2227	-3.7000
Dense CLIP*	H	9.9279	11.5695	6.6467	1.6416	-3.2811

Table 9: Mean attention score losses and margin changes. Each query first averages its relevant or negative stratum, then queries are averaged. Subscripts 5/all refer to fixed full-top-five/all eligible negatives. Raw units are not controlled causal scales across encoders.

F COMPLETE NEW CONTROLS AND PAIRED UNCERTAINTY

All nine Ward variants were fixed before execution. The maximum-Medium rule selects W8; Hard is reported without changing that choice. W1–W3 and W4–W6 coincide on this corpus; these are not nine independent replications. Post-hoc inference includes all eleven new DINOv2 controls

ID	Partition	Representative	k	Medium	Hard
W1	Profile maxclust	Raw mean	39	58.88	32.29
W2	Profile maxclust	Unit mean	39	60.22	35.43
W3	Profile maxclust	Member	39	59.41	33.58
W4	Profile exact-k	Raw mean	39	58.88	32.29
W5	Profile exact-k	Unit mean	39	60.22	35.43
W6	Profile exact-k	Member	39	59.41	33.58
W7	Unit-Euclidean exact-k	Raw mean	39	60.23	34.67
W8	Unit-Euclidean exact-k	Unit mean	39	61.13	36.62
W9	Unit-Euclidean exact-k	Member	39	60.14	35.78

Table 10: All hierarchical controls on DINOv2: 39 vectors/image, 194,727 total. Member denotes an original-vector representative, not a centroid.

Interface	Query tokens	DB tokens	Medium	Hard
Native CLS	1	1	75.68	57.71
Full patches	196	196	66.98	42.73
Attention retention	196	39	75.65	55.40
Patch-query / CLS-DB	196	1	60.84	41.16

Table 11: DINOv3 interfaces. Native CLS and its bridge share the one-vector database, not the query representation. Neither is a 39-token comparison.

and both protocols. This 22-test family is separate from the six backbone/full contrasts and original C1–C4. All two-sided exact tests receive Holm adjustment. Marginal whole-group intervals use the same query-weighted bootstrap defined above, not an unweighted mean of landmark means. The best pooling row receives no selectively chosen test. Neither grouping nor multiplicity correction removes development bias.

Control	Mode	Δ pp	Marginal 95% CI	p exact	p Holm22
Anchor raw	M	+3.17	[1.51, 5.15]	0.009766	0.023438
Anchor raw	H	+3.84	[1.80, 5.70]	0.011719	0.023438
Anchor unit	M	+7.73	[4.71, 11.43]	0.001953	0.023438
Anchor unit	H	+9.78	[6.34, 13.80]	0.001953	0.023438
W1	M	+13.42	[9.77, 17.06]	0.000977	0.021484
W1	H	+17.63	[13.33, 21.98]	0.000977	0.021484
W2	M	+12.08	[7.94, 16.70]	0.001953	0.023438
W2	H	+14.49	[10.05, 19.06]	0.001953	0.023438
W3	M	+12.89	[9.10, 16.86]	0.000977	0.021484
W3	H	+16.34	[12.59, 20.13]	0.000977	0.021484
W4	M	+13.42	[9.77, 17.06]	0.000977	0.021484
W4	H	+17.63	[13.33, 21.98]	0.000977	0.021484
W5	M	+12.08	[7.94, 16.70]	0.001953	0.023438
W5	H	+14.49	[10.05, 19.06]	0.001953	0.023438
W6	M	+12.89	[9.10, 16.86]	0.000977	0.021484
W6	H	+16.34	[12.59, 20.13]	0.000977	0.021484
W7	M	+12.06	[8.52, 15.65]	0.000977	0.021484
W7	H	+15.25	[11.56, 18.91]	0.000977	0.021484
W8	M	+11.17	[7.12, 15.71]	0.001953	0.023438
W8	H	+13.30	[8.99, 17.82]	0.001953	0.023438
W9	M	+12.16	[7.98, 16.77]	0.001953	0.023438
W9	H	+14.14	[9.84, 18.33]	0.001953	0.023438

Table 12: Attention retention minus each new DINOv2 control in mAP points. W1–W9 refer to the table above. Intervals are marginal, not simultaneous; tests are exploratory, not untouched confirmation.

G EVERY LANDMARK AND QUERY

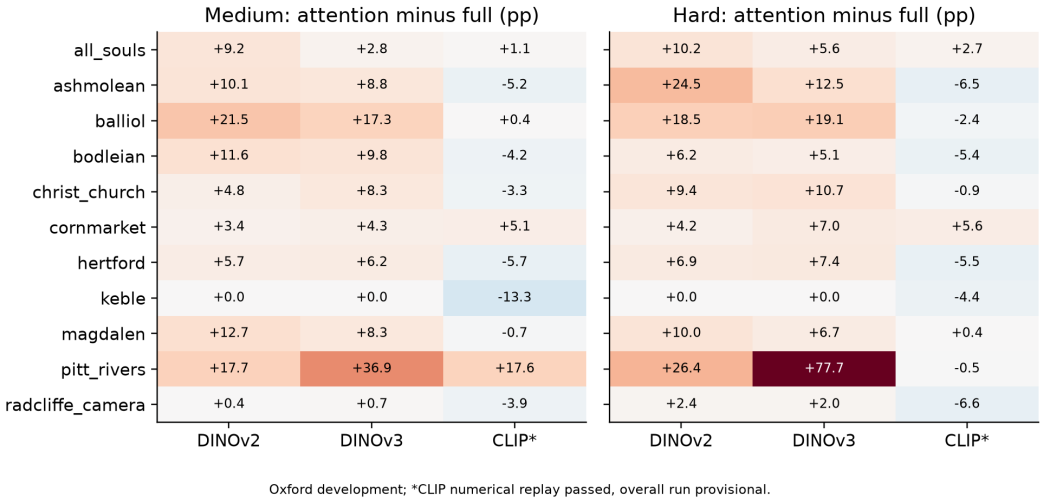

Figure 4: All landmark effects on a common diverging scale. Values are mAP point changes. Groups are not independent datasets. CLIP remains visible.

Landmark	n	D2 M	D2 H	D3 M	D3 H	CLIP M	CLIP H
all_souls	8	+9.20	+10.23	+2.77	+5.64	+1.12	+2.74
ashmolean	8	+10.14	+24.54	+8.80	+12.48	-5.23	-6.52
balliol	5	+21.46	+18.54	+17.31	+19.10	+0.35	-2.40
bodleian	6	+11.58	+6.25	+9.83	+5.14	-4.17	-5.35
christ_church	8	+4.82	+9.37	+8.34	+10.73	-3.34	-0.85
cornmarket	5	+3.35	+4.23	+4.30	+6.96	+5.11	+5.58
hertford	8	+5.75	+6.92	+6.23	+7.37	-5.69	-5.54
keble	5	+0.00	+0.00	+0.00	+0.00	-13.34	-4.44
magdalen	5	+12.68	+9.97	+8.25	+6.67	-0.70	+0.38
pitt_rivers	5	+17.73	+26.41	+36.87	+77.74	+17.56	-0.55
radcliffe_camera	7	+0.37	+2.40	+0.73	+1.95	-3.91	-6.63

Table 13: Group sizes and changes. Averaging these means without their sizes differs from the reported query-weighted effect.

Query	D2 M	D2 H	D3 M	D3 H	CLIP M	CLIP H
all_souls_000013	+5.00	+7.14	+4.43	+5.86	-0.22	+1.45
all_souls_000026	+13.52	+10.31	+3.07	+9.30	-3.30	+2.69
oxford_002985	+11.86	+8.38	+2.77	+10.31	-0.15	+1.29
all_souls_000051	+1.33	+6.47	+0.71	+5.06	-2.64	-4.34
oxford_003410	+11.86	+4.10	+2.34	+7.23	+2.93	+0.56
ashmolean_000058	+16.52	+15.24	+22.42	+16.58	-2.66	-0.04
ashmolean_000000	+5.24	+25.36	+3.64	+18.15	-13.41	-15.38
ashmolean_000269	+5.11	+32.28	+1.79	+7.33	-7.58	-17.56
ashmolean_000007	+8.15	+39.69	+1.88	+10.89	+0.94	+6.56
ashmolean_000305	+10.06	+35.34	+6.88	+18.98	-10.73	-23.85
balliol_000051	+11.10	+21.18	+14.30	+18.29	+0.24	+0.37
balliol_000187	+49.59	+49.59	+13.81	+13.81	+0.41	+0.41
balliol_000167	+22.45	+10.50	+20.39	+24.61	+4.65	+0.79
balliol_000194	+16.55	+9.61	+21.89	+21.46	-1.47	-7.04
oxford_001753	+7.62	+1.82	+16.18	+17.32	-2.07	-6.51
bodleian_000107	+15.79	+3.43	+11.21	+16.62	-1.43	-1.46
oxford_002416	+12.04	+9.28	+8.83	+1.48	-6.66	-2.22
bodleian_000108	+13.84	+1.38	+9.94	+2.32	-6.07	-19.75
bodleian_000407	+13.03	+9.05	+12.25	+3.08	-2.96	-4.61
bodleian_000163	+13.59	+7.14	+14.83	+1.15	-4.16	-1.49
christ_church_000179	+5.19	+14.55	+4.98	+6.97	-2.47	-1.62
oxford_002734	+4.32	+10.24	+5.58	+10.64	-5.71	-0.47
christ_church_000999	+5.94	+15.50	+4.20	+5.19	-3.60	-0.56
christ_church_001020	+4.02	+8.30	+5.93	+8.27	-2.98	-1.23
oxford_002562	+2.15	+4.44	+2.71	+2.82	-2.45	+0.28
cornmarket_000047	+6.71	+8.92	+8.28	+13.31	+2.34	-2.92
cornmarket_000105	+1.67	+1.19	+4.19	+7.43	+10.72	+10.85
cornmarket_000019	+2.51	+2.91	+4.91	+8.75	+3.09	+2.93
oxford_000545	+2.50	+2.95	+3.22	+5.05	+4.18	+7.37
cornmarket_000131	+3.38	+5.21	+0.89	+0.26	+5.22	+9.66
hertford_000015	+1.01	+1.58	-0.44	+1.10	-14.68	-1.72
oxford_001752	+0.89	+1.51	-1.14	-0.06	-8.23	-9.75
oxford_000317	+1.51	+8.94	-0.17	+2.34	-4.04	-6.33
hertford_000027	+2.01	+4.00	-0.70	-0.21	-9.91	-13.32
hertford_000063	+1.94	+5.90	-0.99	-0.38	-8.01	-12.68

Table 14: Every query, 1–35 in the fixed order: attention-minus-full AP changes in percentage points. D2/D3 denote the DINO generations.

Query	D2 M	D2 H	D3 M	D3 H	CLIP M	CLIP H
keble_000245	+0.00	+0.00	+0.00	+0.00	-20.49	-18.75
keble_000214	+0.00	+0.00	+0.00	+0.00	-3.89	-2.98
keble_000227	+0.00	+0.00	+0.00	+0.00	-1.55	-1.15
keble_000028	+0.00	+0.00	+0.00	+0.00	+0.90	+0.69
keble_000055	+0.00	+0.00	+0.00	+0.00	-41.67	+0.00
magdalen_000078	+13.42	+9.80	+9.22	+5.81	-1.36	-0.01
oxford_003335	+15.52	+12.29	+7.07	+6.61	+0.12	+1.46
magdalen_000058	+11.45	+10.18	+7.23	+6.04	-0.40	+0.35
oxford_001115	+12.55	+11.02	+7.76	+8.64	-1.77	-0.10
magdalen_000560	+10.44	+6.57	+9.98	+6.23	-0.12	+0.22
pitt_rivers_000033	+6.43	+5.36	+20.83	+75.00	-1.08	-0.73
pitt_rivers_000119	+17.52	+14.39	+47.55	+98.33	+6.74	-0.51
pitt_rivers_000153	+8.47	+6.94	+48.97	+99.31	-0.43	-0.29
pitt_rivers_000087	+14.93	+11.61	+28.29	+24.41	+46.09	-0.53
pitt_rivers_000058	+41.32	+93.75	+38.69	+91.67	+36.47	-0.69
radcliffe_camera_000519	+0.19	+0.72	+0.94	+2.65	-3.01	-6.67
oxford_002904	-0.15	-1.44	+0.36	-0.91	-2.16	-12.61
radcliffe_camera_000523	+0.58	+2.75	+1.28	+3.69	-2.11	-5.33
radcliffe_camera_000095	+0.21	+1.47	+0.71	+0.33	-2.23	-10.86
bodleian_000132	+1.19	+7.21	+1.90	+6.22	-3.73	-2.59
all_souls_000021	+5.71	+8.17	+5.84	+5.85	-3.96	+1.02
all_souls_000022	-1.82	+0.40	+2.98	+2.04	-1.36	-0.32
all_souls_000152	+26.12	+36.88	+0.06	-0.58	+17.64	+19.55
ashmolean_000078	+7.40	+8.66	+5.02	+5.77	-3.32	-1.20
ashmolean_000005	+10.56	+8.45	+16.73	+12.74	+5.72	+2.14
ashmolean_000301	+18.12	+31.33	+12.06	+9.42	-10.79	-2.82
christ_church_000262	+10.80	+13.79	+19.98	+21.19	-2.64	+0.06
oxford_000736	+3.77	+4.61	+11.89	+16.88	-1.89	-0.46
oxford_002364	+2.40	+3.50	+11.41	+13.87	-4.98	-2.82
hertford_000078	+16.04	+13.28	+15.57	+16.80	-0.09	-0.07
hertford_000044	+13.82	+11.43	+18.75	+20.42	-0.32	-0.26
hertford_000082	+8.74	+8.74	+18.98	+18.98	-0.22	-0.22
radcliffe_camera_000290	+1.93	+10.17	+1.87	+7.36	+3.44	+3.39
radcliffe_camera_000286	+0.03	+3.34	-0.38	+0.26	-8.87	-1.87
radcliffe_camera_000031	-0.18	-0.20	+0.32	+0.29	-12.39	-12.49

Table 15: Queries 36–70. Negative changes and ties remain. Rows are dependent within landmark groups.

H VERIFICATION, NUMERICAL FAILURES AND NEGATIVE EVIDENCE

The intake authenticates unchanged seals, source/configuration identities, ordered roles and official GT before parsing arrays. Rankings must equal stable score sorts; metrics are recomputed independently. Partial score-only exports list omitted upstream data and are not called full-cache audits. The reporting build binds tables and figures to verified analysis hashes and its own source. Hashes authenticate identity, not scientific correctness.

CLIP's recorded UTC extraction span is 20,839.323124 seconds, versus a 3,600-second cap and a 308.015913-second timer. Numerical replay passes but the inconsistency remains unresolved. We retain the adverse observation with a provisional label, not claim restored protocol compliance.

A fixed ten-image DINOv2 diagnostic compares CPU-float32 and MPS-float32 with a CPU-double diagnostic. Maximum raw-feature differences are approximately 0.007434 and 0.003468 respectively; CPU-float32 is not uniformly closer. The double diagnostic still uses float32 positional interpolation and is not exact ground truth. Full/Attention winner identities and stable rankings agree on that subset. This supports rounding sensitivity, not a demonstrated semantic cache defect or corpus-wide harmlessness. The original raw-norm gate remains failed. Separately reviewed cache-only CPU controls do not waive it.

The targeted local suite passes, while full Windows collection stops at Unix-only `resource` imports. Test passes are engineering evidence, not discoveries. The proposed public artifact must remove identifying paths and private URLs, provide code/data-use terms and pass a clean authorized build.

Incompletely authenticated register/resolution exports are not promoted to verified robustness. Original query-pruning, wrong crop-attention, informal hubness, winner-coverage and truncated-AP headlines are retired. Highest-norm retention does not match attention, but this does not eliminate every artifact mechanism. No complete-trained-pooling superiority, cross-dataset same-anchor ordering, predictive phase map or measured systems benefit is established. The separate completed-control audit authenticates all twelve arms, exact counts, source/cache identities, selection metadata, full score-to-rank conversions and official per-query AP/mP. Independent representative/MaxSim replay is limited to the fixed three-query/seven-image subset. The largest metric discrepancy is 5.33×10^{-15}, and the largest subset maxima discrepancy is 2.09×10^{-7}. Subset winners and rankings agree. This is not an encoder rerun or a second full-corpus feature-to-score audit.

The first control audit failed because it incorrectly required int32 for the authentic int64 DINOv3 reference ranks. The revised audit uses strict source-specific int32/int64 contracts, preserving the failed record and all scientific output bytes. Eleven independent metric/dtype/order fixture tests accompany it. Completed same-anchor, Ward and CLS-bridge CPU invocations took approximately 76.4, 464.5 and 14.0 seconds; these operational timings are not a deployment benchmark or a waiver of the earlier MPS extraction gate.

I ADDITIONAL PARIS EVALUATION AND REPORTING SCOPE

The later DINOv2 Paris evaluation uses 6,322 database images and 70 queries. It follows Oxford development; internal admission gates were bypassed under deadline pressure. It is not treated as untouched or preregistered confirmation. The completed independent audit reports agreement for official AP, grouping, stable rankings and MaxSim, including CPU-oracle checks for Full and each of the three attention budgets. The same DINOv2 revision listed in Appendix A is used.

The authenticated audit report supplies exact Medium and Hard values for all 33 conditions, retained in Table 16. Its independent CPU oracle reconstructs full score matrices for Full and all three attention budgets from the original float16 token shard. Maximum absolute score differences are at most 1.23×10^{-4}, within the recorded tolerances; the largest AP-mean difference is below 1.90×10^{-7} in fractional mAP. Near-tie rank positions are not identical: approximately99.84–99.86% agree, depending on the condition. This is numerical agreement within tolerance, not bit-exact ranking parity or independent encoder extraction. The reporting build imports the completed audit; it does not rerun model inference or that CPU oracle.

Paris groups were independently reconstructed as eleven landmark groups, including the specified positive-overlap assignment for generic filenames. No Oxford group identities or grouped test results are transferred to Paris. The complete per-query AP export is retained with the audit, but no new significance, equivalence or multiple-dataset randomization test is claimed here. Team-wide prior exposure is unknown. The run's post-Oxford chronology does not establish an untouched benchmark, and the original admission deviation remains visible.

Condition	k	Medium	Hard
full_tokens	196	85.92	69.42
attn_keep@0.5	98	86.99	71.32
random_keep@0.5_s0	98	84.76	67.85
random_keep@0.5_s1	98	85.00	68.30
random_keep@0.5_s2	98	85.05	68.28
random_keep@0.5_s3	98	85.05	68.31
random_keep@0.5_s4	98	84.89	68.08
attn_keep@0.2	39	88.23	73.62
random_keep@0.2_s0	39	82.88	65.93
random_keep@0.2_s1	39	83.00	66.21
random_keep@0.2_s2	39	82.98	65.90
random_keep@0.2_s3	39	83.00	66.12
random_keep@0.2_s4	39	83.21	66.17
attn_keep@0.1	20	88.23	73.94
random_keep@0.1_s0	20	80.96	64.03
random_keep@0.1_s1	20	80.81	63.82
random_keep@0.1_s2	20	80.57	63.52
random_keep@0.1_s3	20	81.14	64.86
random_keep@0.1_s4	20	80.58	63.20
medoid@0.5	98	85.66	69.04
merge@0.5	98	85.76	69.15
spatial@0.5	98	85.11	68.52
raw_norm@0.5	98	82.33	65.63
medoid@0.2	39	85.37	68.95
merge@0.2	39	85.66	69.30
spatial@0.2	39	82.44	65.76
raw_norm@0.2	39	74.30	57.04
medoid@0.1	20	84.15	67.46
merge@0.1	20	85.30	69.46
spatial@0.1	20	80.88	64.57
raw_norm@0.1	20	64.93	48.38
mean_pool	1	77.45	59.93
cls_descriptor	1	88.39	75.20

Table 16: All33 frozen Paris conditions, mAP (%). No seed or unfavorable condition is dropped. CLS/mean use one-vector query interfaces.

Encoder	Representation	k	Medium	Hard
DINOv2	Full	196	63.91	39.08
DINOv2	Attention39	39	72.30	49.92
DINOv2	Cls	1	63.80	43.66
DINOv2	Mean	1	39.57	16.75
DINOv3	Full	196	66.98	42.73
DINOv3	Attention39	39	75.65	55.40
DINOv3	Cls	1	75.68	57.71
DINOv3	Mean	1	44.12	20.89
Dense CLIP*	Full	196	37.14	11.06
Dense CLIP*	Attention39	39	35.53	8.68
Dense CLIP*	Cls	1	27.57	6.24
Dense CLIP*	Mean	1	31.71	9.44

Table 17: Complete Oxford backbone and descriptor table, mAP (%). CLS and Mean use single-vector queries and databases. *CLIP remains provisional overall.