DIAGNOSING LEARNER–CODEC MISMATCH IN LOW-BIT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval comparisons at a fixed code rate can depend on the export codec as well as the learner. On Stanford Online Products, FIT-only range calibration improves frozen twenty-dimensional four-bit triplet descriptors by 14.22 average-precision points at the same ten-byte payload, changing the allocation effect against ten-dimensional eight-bit descriptors from -5.61 to $+8.50$ points. At common four-bit output, a historical four-bit-training advantage falls from $+9.74$ to -0.36 points; learned-scale DINOv2 and SigLIP2 contrasts have conditional intervals including zero. Product quantization fitted separately to each learned representation supplies a further comparison. In one additional paired training realization, native contrastive descriptors trail triplet descriptors by 2.081 points, whereas CAL-selected ten-byte codecs favor contrastive descriptors by 1.854 points; both adjusted intervals exclude zero. Range calibration alone also reverses this pair's learner ordering. The finding concerns codec-dependent rankings on an exposed population, not universal contrastive superiority. Training-step-preserving and range-reset exports remain distinct, and CUB retains adverse allocation comparisons. These controls diagnose learner–codec compatibility without establishing a new quantizer, equivalence, or a minimum-bit rule.

1 INTRODUCTION

Compact retrieval couples learned geometry to a finite code. At eighty bits per image, ten eight-bit coordinates and twenty four-bit coordinates store the same payload but impose different constraints. The central question is therefore *which apparent allocation preferences survive competent, equally available codec repair?* Native triplet descriptors initially favor precision, whereas memory-contrastive descriptors favor dimension. Yet a FIT-only range fit improves the same triplet Q4 checkpoint from 36.49% to 50.71% AP. That recovery requires neither additional item bytes nor retraining. Native rankings consequently describe particular learner–codec pipelines, not intrinsic bit requirements.

Three frozen-checkpoint interventions make these explanations distinguishable. A decoder-only change replaces grid midpoints with training-set centroids while preserving every stored index. A subsequent reassignment changes indices under the same centroid table. Finally, global or coordinate-wise range calibration changes the uniform grid using only training representations. The first intervention tests reconstruction values; the second adds assignment flexibility; the third tests a particularly simple scale mismatch. None requires retraining the image encoder (Figure 1).

The resulting picture goes beyond an allocation reversal. Training-only range calibration removes the opposite-sign pattern on Stanford Online Products (SOP), including two fresh paired heads. At fixed twenty dimensions, an equally calibrated four-bit export also changes the comparison between four-bit and eight-bit training. This second contrast separates a deployment-compatible representation from an apparent advantage caused by the inherited export grid. Eight-bit training remains quantized training, not a full-precision control.

The contribution is a matched-checkpoint diagnosis using standard codec operations. Assignment/decoder crosses, exact zero-epsilon evaluation and sampled margin transitions distinguish reconstruction, reassignment and range mismatch. Learned-scale baselines and a second pretrained family test the common-export finding beyond the original fixed grid. CUB, normalization, memory,

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.

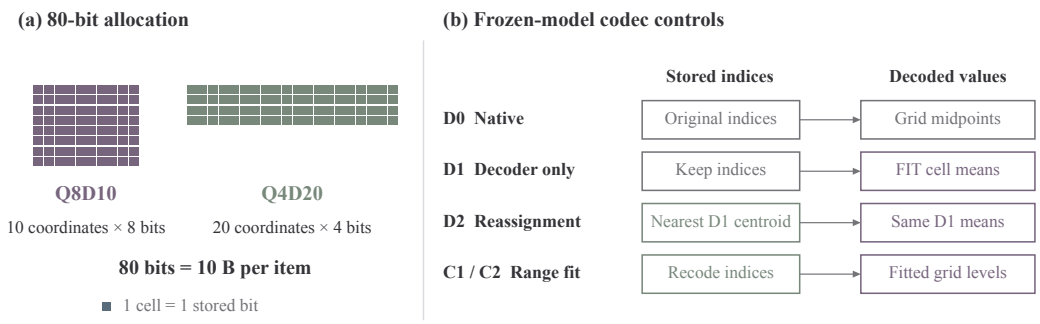

Figure 1: The same payload permits different allocations and codec controls. (a) Each cell is one stored bit: ten eight-bit coordinates and twenty four-bit coordinates both occupy ten bytes. (b) Arrows map stored indices to decoded values under fixed checkpoints. D1 tests reconstruction values with indices fixed; D2 tests reassignment under the same fitted table. C1/C2 independently fit global or coordinate-wise uniform-grid ranges on FIT and recode the same floating vectors. Shared decoder state, recoding and index rebuilding remain separate costs.

and stronger teacher-compression comparisons retain unfavorable outcomes. The practical implication is to align the output interface before assigning a quality difference to insufficient storage or to the training precision alone.

2 Related work and scope

Quantization and retrieval. Centroid fitting and nearest-centroid assignment are classical least-squares quantization operations (Lloyd, 1982). Product quantization separates a vector into subspaces (Jégou et al., 2011); optimized product quantization and Cartesian k-means improve the representation–partition fit (Ge et al., 2013; Norouzi & Fleet, 2013). Iterative quantization similarly emphasizes alignment before binary coding (Gong & Lazebnik, 2011). Supervised product quantization integrates representation learning and quantization (Klein & Wolf, 2019), while anisotropic quantization shows that downstream search quality need not track ordinary reconstruction error (Guo et al., 2020). Scalar controls isolate which stored or decoded object changes; own-latent PQ/OPQ additionally test stronger standard codecs.

Learning under low precision. Straight-through gradient estimators permit optimization through discrete operations (Bengio et al., 2013); learned step sizes explicitly adapt quantizer scale (Esser et al., 2020). Finite scalar quantization provides another use of bounded scalar representations (Mentzer et al., 2024). These methods motivate separating the training grid from the final export grid. The comparison here concerns compact retrieval descriptors, rather than network-weight compression or generative tokenization. A trained low-bit representation may adapt to a coarse grid, but such adaptation does not imply that an uncalibrated grid is the best deployment interface.

Metric learning and evaluation. Triplet, lifted, N-pair, and multi-similarity losses offer different ways to use labeled pairs (Schroff et al., 2015; Song et al., 2016; Sohn, 2016; Wang et al., 2019). Sampling also changes the effective learning problem (Wu et al., 2017). Cross-batch memory expands available embedding pairs (Wang et al., 2020). The tested adaptation fine-tunes the same pretrained backbone under a common optimizer and training-time budget. Controlled evaluation is important because implementation and selection differences can change method rankings (Musgrave et al., 2020). Ties receive explicit treatment following the concern formalized by He et al. (2018). Budgeted training motivates matching a declared resource contract (Li et al., 2020), but equal time is not equal update count. Selection uncertainty and dataset reuse remain distinct from conditional test-set uncertainty (Cawley & Talbot, 2010; Recht et al., 2019).

Scope and limitations. The evidence is narrower than an allocation law. Two datasets, finite menus, small paired cohorts, and a triplet-only SigLIP2 extension do not identify all low-bit regimes. Public test populations are reused. Conditional intervals exclude seed variability, selection and

sequential exposure; a small contrast is not equivalence. Pretraining overlap remains possible (Wah et al., 2011).

Additional diagnostics use zero-epsilon scoring, eight-bit export of each CUB model's own pre-round vectors, and sampled margin transitions. The remaining limits are inferential: sampled flips do not explain total AP causally; global-scale invariance does not cover arbitrary decoders; and reconstruction-selected C1 is not an optimal retrieval codec. The two-head Q8 follow-up at margin .7 removes the margin mismatch in that comparison, but heterogeneous effects do not isolate bit width from dimension, scale initialization or quantizer derivative.

Matched codec opportunities test whether an observed ranking belongs to the learner or to a particular deployment interface; finite menus do not characterize all possible codecs.

3 A MATCHED COMPARISON AT TEN BYTES

Objects and estimands. For an image x, a trained encoder and projection head produce a pre-round vector $z(x) \in \mathbb{R}^d$. An encoder e maps z to a finite index vector $c = e(z)$; a decoder d_c maps indices to reconstruction values. Retrieval ranks gallery items by the declared decoded distance. Let $Q4D20$ denote twenty four-bit coordinates and $Q8D10$ ten eight-bit coordinates. Both occupy exactly ten bytes per image, excluding shared decoder state and index metadata. The paired allocation effect under learner L and codec k is

$$\Delta_{L,k} = \mathrm{AP}(L, Q4D20, k) - \mathrm{AP}(L, Q8D10, k). \tag{1}$$

A positive effect favors more coordinates; a negative effect favors finer coordinate precision. Comparisons of Δ across codecs keep the image model fixed. Comparisons across learners also change the objective and associated training path; such contrasts cannot identify one isolated gradient mechanism.

Learning and pairing. The original learners fine-tune the same DINOv2-base vision-transformer family (Oquab et al., 2024; Dosovitskiy et al., 2021). Triplet learner U retains its norm penalty. The contrastive learner uses normalized embeddings, current-batch pairs, and a detached FIFO of 8,192 embeddings after a 2,000-update warmup. The global batch contains 192 triplets, or 576 endpoint images. For the original SOP U/XBM allocation fits, body/head rates are $10^{-5}f(t)/10^{-3}f(t)$ with multiplier 1 and synchronized linear wall-time decay after setup (Appendix B.1). A paired twenty-row initialization supplies the same first ten rows for the ten-dimensional model; optimizer state starts empty.

The fit contract allocates 4,500 seconds on eight GPUs, including setup and a 180-second completion reserve, with synchronized wall-clock learning-rate decay. Equal time does not imply equal updates. CAL selects the declared recipe menu, never TEST. Original XBM uses margin .9 at both allocations; the selected recipe uses Q4 margin .7 and Q8 margin .9. Same-margin and selected-pipeline effects remain separate.

Two later controls address different axes. Common-export comparisons fix $d = 20$: T4/T8 denote four-/eight-bit training grids and E4/E8 four-/eight-bit output codes. Each T4/T8 model's own pre-round vectors are re-encoded into E4 (ten bytes), using its FIT-selected range. No rounded T8 code is re-rounded. An LSQ-style embedding-step adaptation (Esser et al., 2020) supplies eight paired DINOv2 allocation fits and two additional U T8D20 fits. Four SigLIP2-base/16 U fits provide two further T4/T8 pairs (Zhai et al., 2023; Tschannen et al., 2025). Body, native pooling and preprocessing change together; this transport bundle does not isolate a backbone effect.

The floating latent is $z = \tanh(Wf + b)$, where f is the normalized body feature. The learned scalar s is a *step*, with endpoint $7.5s$ for T4 and $127s$ for T8; the FIT export parameter R is instead a *range*. Training-step-preserving, range-reset and FIT-calibrated exports are distinct (Appendix B.2). D0 denotes the original fixed-grid allocation codec; common-export D0 resets the range to 1 and discards the learned step. A poor reset export cannot establish that the trained quantizer failed.

Data roles and metrics. SOP (Song et al., 2016) has 59,551 training images and 60,502 test images. Product-disjoint training roles contain 35,569 FIT, 12,056 CAL, and 11,926 readout images. FIT supports optimization and codec fitting; CAL supports the specified recipe menu. The final

gallery contains 11,316 products. SOP reports product-macro average precision: query AP is averaged within product and then across products. Self-matches are removed. Native integer distance keys preserve the original scoring convention, and exact ties are averaged rather than broken by array order.

CUB (Wah et al., 2011) uses 5,864 images from the first 100 species for training and 5,924 images from the remaining 100 species for evaluation. CUB reports query-mean AP. All 18 CUB models receive the same FIT-only codec opportunities: twelve original models from two learners, two allocations, and three paired heads, plus six contrastive models transferring the SOP-selected .7 margin to both allocations. No CUB result selects a new margin or range-search menu. CUB has nevertheless been examined during this research program; the extension is an exposed-dataset transfer analysis, not a fresh external validation sample.

Uncertainty. Individual paired-head effects are tabulated in Appendix D. Conditional 95% intervals resample SOP products or CUB species after fixing the trained models. The CUB bootstrap retains query weighting within each species-resampled sample. These intervals do not integrate uncertainty from training seeds, recipe selection, or repeated dataset exposure. The two-head SOP cohorts and three-head CUB cohort are reported separately; test queries are not counted as independent training replications.

4 Separating reconstruction, assignment, and range

D0: original interface. The original four-bit grid has sixteen midrise levels on $[-1, 1]$; the eight-bit grid has 255 signed levels. The native squared-distance convention shifts the query by $\eta = 10^{-6}\mathbf{1}$. For signed Q8 integers, an order-equivalent key is $10^6 \sum(q-g)^2 + 254 \sum(q-g)$; for Q4 indices the coefficient is 15. Additional controls use zero-epsilon squared Euclidean distance and expected exact ties, without near-tie tolerance. FP32 midpoint reconstruction, exact native reproduction and zero-epsilon evaluation are distinct contracts.

D1: decoder-only centroids. For each coordinate j and original index a, FIT representations define

$$\mu_{ja} = \frac{\sum_{i \in \mathrm{FIT}} \mathbf{1}[c^0_{ij} = a] z_{ij}}{\sum_{i \in \mathrm{FIT}} \mathbf{1}[c^0_{ij} = a]}. \tag{2}$$

Empty cells retain the original midpoint. Accumulation uses FP64; the stored table uses FP32. Query and gallery indices remain exactly c^0. The intervention can reduce FIT reconstruction error without improving retrieval because squared reconstruction error is not a ranking objective. A table learned on FIT also need not reduce held-out reconstruction error.

D2: reassignment with a frozen table. Each coordinate is reassigned to its nearest D1 reconstruction value,

$$c^1_j(z) = \arg\min_a (z_j - \mu_{ja})^2. \tag{3}$$

The table is not refitted after reassignment. Consequently, D2 isolates the incremental value of changing the assignment under a fixed decoder. Payload remains ten bytes, but every query and gallery descriptor must be regenerated. D1 and D2 are the two classical descent steps for squared error (Lloyd, 1982); Appendix C gives the elementary bounds and explains why AP has no corresponding monotonicity guarantee.

C1/C2: uniform-grid range calibration. C1 selects a positive range R on FIT reconstruction error and evaluates $RQ(z/R)$. C2 performs the same search independently by coordinate. Both use $R \in \{.25, .5, .75, 1, 1.25, 1.5, 2\}$ for both learners and allocations, including the original range. No TEST retrieval label enters fitting. Boundaries and reconstructions both change, unlike D1. Range parameters, physical lookup tables and recoded galleries remain separate deployment costs; R must not be confused with a learned training step.

What the ladder can establish. A D1 recovery would implicate reconstruction values even when assignments remain fixed. A D2-only recovery would point to assignments under the centroid decoder. A C1 recovery establishes sensitivity to a simple scale-compatible codec, but does not estab-

lish that one scalar fully explains the learned representation. None of these comparisons identifies a rate–distortion frontier across all possible codecs. The ladder tests specific alternative explanations while retaining the original checkpoints and their unfavorable results.

5 RESULTS: THE NATIVE RANKING IS NOT CODEC-INVARIANT

Same-margin evidence precedes recipe selection. The original SOP comparison uses contrastive margin .9 for both allocations. The Q4 − Q8 effect is $+0.776$ AP points, with paired-head effects $+0.242$ and $+1.310$ and conditional product interval $[+0.478, +1.063]$. The corresponding U effect is -5.61 points. Thus the native opposite-sign observation does not originate solely from choosing a different Q4 margin. The later CAL-selected contrast is larger, but combines the allocation change with a selected recipe: Q4 uses .7 while Q8 retains .9. The original comparison establishes the local same-margin observation; the selected and fresh cohorts test the resulting practical recipe. Neither contrast substitutes for the other.

Centroids leave most of the native defect intact. The selected SOP cohort gives $\Delta_{U,D0} = -5.61$ and $\Delta_{XBM,D0} = +5.22$ points (Figure 2; Appendix Table 4). D1 changes the U effect to -5.50 and D2 to -3.47. Both remain negative. XBM remains positive under both interventions. Thus preserving indices while repairing reconstruction values is insufficient, and nearest-table reassignment only partially recovers the U comparison. This result is more specific than attributing the entire deficit to either a poorly learned latent or a poor decoder.

A simple range fit changes the conclusion. C1 changes the SOP U effect to $+8.50$ points, whereas XBM remains near $+5.19$. Coordinate-wise C2 yields $+8.23$ and $+5.19$. The initial opposite-sign statement is therefore false for these calibrated codecs. A claim that contrastive learning intrinsically prefers more dimensions while triplet learning intrinsically prefers more precision would be contradicted by the same checkpoints under a training-only range fit. The relevant evidence is the interaction between a learned geometry and its codec, not an invariant learner label.

Absolute quality separates repair from an allocation interaction. Selected U Q4 improves from 36.4854% to 50.7076% AP, a 14.222-point gain; Q8 improves from 42.0941% to 42.2030%, a $.109$-point gain. Their difference is the 14.113-point allocation change, not Q4's individual gain. Fresh-head native effects are $-4.88/+4.51$ points for U/XBM; C1 gives $+8.98/+4.51$. Both heads agree in sign, but the gallery is previously exposed rather than a new population.

CUB supplies a counterexample to both simple stories. With the native interface, CUB favors Q8D10 for both learners: U gives -11.47 points and original-margin XBM -9.04. Transferring margin .7 improves the XBM allocation effect to -8.16 but does not reverse the sign. Those native results reject an unqualified cross-dataset replication of the initial SOP reversal.

Calibration adds a second qualification. C1 changes CUB U to $+1.18$ points, with a conditional species-bootstrap interval of $[+0.46, +1.90]$. Original-margin XBM remains at -9.04; transferred-margin XBM remains at -8.15, with interval $[-9.58, -6.69]$. The external result is therefore not simply "both learners prefer precision." That description applies to the native interface, not to the calibrated triplet representation. Conversely, the calibration result does not establish a universal preference for more coordinates: XBM supplies a direct negative comparison on the same dataset and under the same codec opportunity.

The calibrated U head effects are $-0.19, +2.26, +1.47$ points; the mean does not establish success at every initialization. All transferred XBM effects remain negative ($-7.52, -8.25, -8.68$). The own-latent Q4-to-E8 panel further separates storage from learning: mean CUB U AP is .2644 at D0, .3945 at C1, .4104 at E8, and .4111 before rounding. E8 uses twenty bytes and the floating latent eighty bytes, so neither is a ten-byte competitor.

The fitted ranges identify an interface mismatch. The reconstruction records provide a concrete explanation for the asymmetry, without requiring a new encoder. Both selected SOP Q4 U heads choose the smallest allowed global range, .25, while both Q4 XBM heads retain range 1. The same choices occur in the fresh SOP heads and the CUB heads. At Q4, C1 changes every U item code but leaves every XBM item code unchanged. Thus C1's apparent simplicity must not be mistaken for a

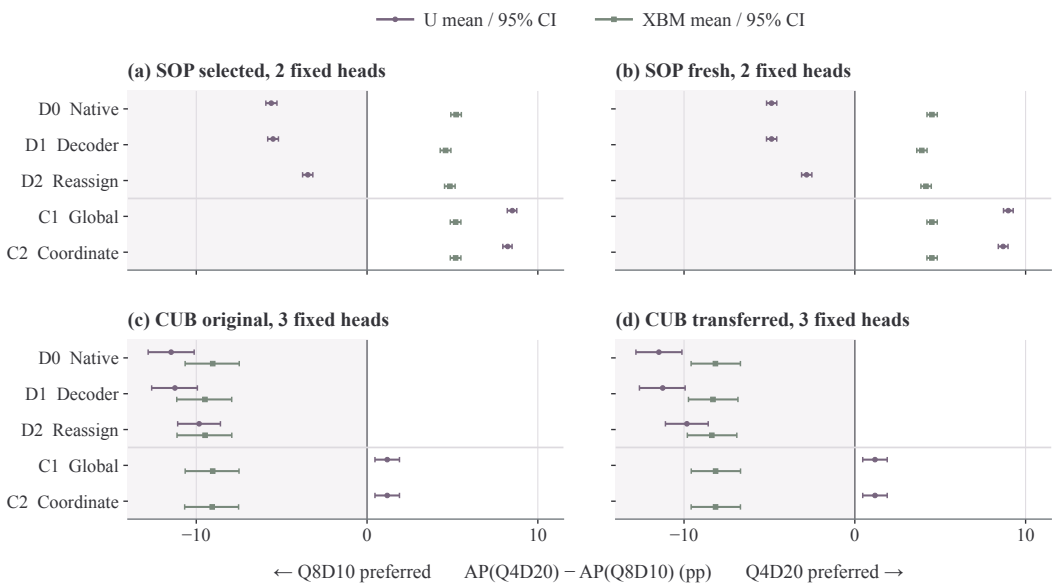

Figure 2: Allocation preference depends on the codec. Four panels show triplet U and contrastive XBM effects, $\mathrm{AP}(Q4D20) - \mathrm{AP}(Q8D10)$ in percentage points, on a common scale. Filled markers are means; capped lines retain the true widths of pointwise conditional 95% product/species-bootstrap intervals. Individual head values are in Appendix Table 10. SOP selected and fresh each use two heads; CUB uses three. The transferred CUB panel changes only the contrastive margin to .7 and reuses U. C1/C2 reverse the U mean on SOP and CUB; XBM remains positive on SOP and negative on CUB. Intervals exclude training-seed and selection uncertainty.

decoder-only operation. The intervention changes the effective resolution in the region occupied by the U representation.

The assignment/decoder cross sharpens this interpretation. With a common positive global scale s and zero epsilon, multiplying both decoded roles by s multiplies every squared distance by s^2, preserving weak orders and ties. Across 80 recorded assignment rows, decoder multipliers 1 and s give identical AP. Thus C1's zero-epsilon ranking change arises from changed assignments, not the final scalar multiplier. This result does not extend to coordinate-wise C2, asymmetric role scales, or the native epsilon convention.

Reconstruction evidence remains useful but limited (Appendix Table 9). Centroids improve FIT MSE by construction, yet CUB transferred-XBM D1 distortion worsens on TEST; D2 repairs distortion without recovering the SOP allocation preference. Range .25 is the finite menu's lower boundary, not a globally optimal scale. Neither post hoc menu expansion nor a comparison of raw latent MSE across learners follows from these findings.

Why reconstruction error cannot identify ranking damage. Write a decoded vector as $\hat{z}_i = z_i + e_i$. For query q and gallery item g, the error in squared distance is

$$\epsilon(q, g) = 2(z_q - z_g)^\top (e_q - e_g) + \|e_q - e_g\|^2. \tag{4}$$

For relevant p and irrelevant n, the margin $m = \|z_q - z_n\|^2 - \|z_q - z_p\|^2$ becomes $m + \epsilon(q, n) - \epsilon(q, p)$. The native version adds η to both query differences. The diagnostics use 4,096 uniformly sampled query/positive/negative triples per model and four absolute-pre-margin strata of the same panel. Across 160 model/scorer/codec rows, the maximum expansion residual is 2.36×10^{-14}. For one selected Q4 U head, destructive zero-epsilon flips fall from 21 to 13 under C1, while restorative flips also fall from 5 to 2. Both directions and ties are retained. These sampled, retrospective transitions are not a full-gallery AP decomposition or an independently validated predictor of retrieval gain.

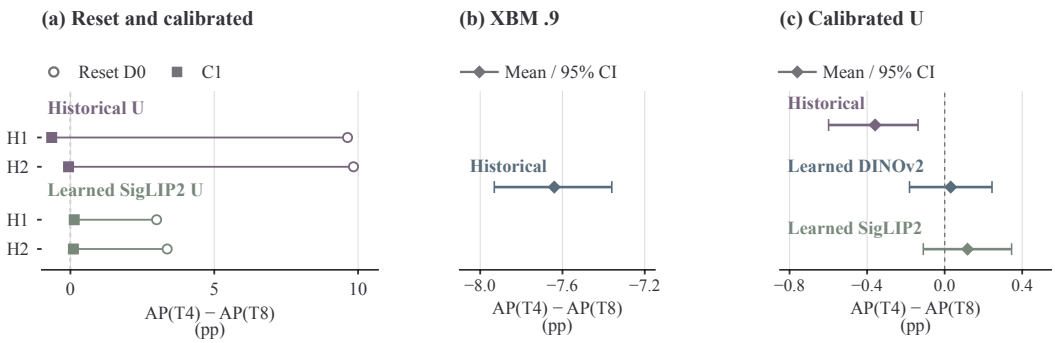

Figure 3: Common-output calibration changes the training-grid comparison. All panels compare T4 − T8 at twenty dimensions and common E4 ten-byte export under zero-epsilon scoring. (a) Each row tracks one fixed paired-head effect from range-reset D0 to FIT-calibrated C1; these are not gains over training-step-preserving export. (b,c) Diamonds and capped lines show two-head means and unadjusted conditional 95% product-bootstrap intervals on separate XBM and U scales. Individual head values are in Appendix Table 11. XBM retains an adverse contrast; small learned-scale U contrasts do not establish equivalence.

Selected-margin gains mostly precede rounding. For Q4D20 XBM, the selected .7 checkpoint is re-exported from its own pre-round latent, not from a different model or an already rounded code. Relative to the original .9 checkpoint, the two-head SOP mean improves by 4.289 points before rounding and 4.439 points under E4. E8 gives 4.292 points. The change in the pre-round-to-E4 gap is only -0.150 points. Most of the selected-recipe improvement is already present in the continuous geometry.

For pre-round AP R and exported AP Q, define the signed export gap $D = R - Q$. The identity

$$Q_B - Q_A = (R_B - R_A) - (D_B - D_A) \tag{5}$$

is descriptive accounting, not causal mediation. Pre-round and E8 exports cost more storage than the E4 deployment and are not ten-byte competitors. The diagnosis excludes a narrow explanation—that the selected margin works mainly by reducing final rounding damage—without identifying which gradient or sampling effect improved the latent.

6 Alternative explanations and deployment cost

Common export changes the training-grid comparison. Figure 3 and Table 1 separate the same-dimensional training-grid contrast from the same-byte allocation contrast. Historical U's T4 advantage of $+9.74$ points under inherited E4 becomes -0.36 under equal FIT calibration. Historical XBM, with margin .9 at both grids, instead favors T8 by 7.64 points after C1. Consequently, range calibration and the training trajectory are not interchangeable remedies. The result refutes a universal coarse-training advantage, without implying that eight-bit training is equivalent to full precision.

Learned-step controls across the two allocations. After equal C1 opportunities, learned-scale U changes neither layout clearly; XBM improves Q4 at margin .7 by 2.305 points and worsens Q8 at margin .9 by $.485$. A same-margin .7 Q8 follow-up remains adverse at -9.205 points, with head effects $-.560/-17.850$. Matching the margin narrows one recipe difference, but the heterogeneous heads do not identify a shared failure mechanism. Dimension, bit width, FIT initialization and the gradient used for the clipped quantizer remain coupled.

A second pretrained family reproduces the calibration qualification. SigLIP2 U favors T4 under range-reset E4 ($+3.00, +3.36$ points); C1 effects shrink to $+.133, +.103$. The mean $+.118$-point interval includes zero, as does DINOv2's $+.031$-point contrast. Neither establishes equivalence. SigLIP2 training-step-preserving T4 instead reaches 62.816% AP, versus 62.036% under C1: a paired $-.781$-point effect $[-.895, -.672]$ despite lower FIT reconstruction error. An endpoint-preserving T8-to-E4 conversion also loses to its own C1 export. Appendix B.2 defines these exports;

Table 1: Does a training-grid preference survive a common calibrated output? T4 − T8 at twenty dimensions and E4 ten-byte export, in AP points. D0 resets the export range to 1; C1 fits the range on FIT. Intervals are unadjusted conditional 95% product-bootstrap intervals for two fixed paired heads, not seed-population intervals. Within-layout learned − fixed effects are separate (Appendix Table 12).

Training family	Contrast	Mean	Conditional interval
Historical U	D0 E4	$+9.737$	$[+9.395, +10.049]$
Historical U	C1 E4	-0.359	$[-0.599, -0.137]$
Historical XBM .9	C1 E4	-7.640	$[-7.931, -7.360]$
Learned-scale DINOv2 U	C1 E4	$+0.031$	$[-0.181, +0.244]$
Learned-scale SigLIP2 U	D0 E4	$+3.180$	$[+2.836, +3.523]$
Learned-scale SigLIP2 U	C1 E4	$+0.118$	$[-0.110, +0.345]$

a CAL-labelled choice between native and C1 has positive, null and adverse outcomes without reselection on a separate SOP readout (Appendix B.5).

Stronger codecs also change learner rankings. PQ/OPQ fitted to each representation improve competent ten-byte native exports, with CAL selecting among native, C1, PQ and OPQ (Appendix F). An exploratory reversal across two existing Q4 pairs motivated one additional paired fit. In that new pair, XBM − U AP is -2.081 points under native export, with 97.5% interval $[-2.548, -1.621]$, but $+1.854$ under CAL-selected codecs, $[1.415, 2.302]$. Both sign conditions hold within a two-contrast family. C1 alone also reverses this pair; the result neither identifies a unique PQ/OPQ mechanism nor establishes universal XBM superiority. The intervals condition on one new trained pair and the previously exposed SOP products.

Memory alone does not explain the allocation pattern. A same-loss control switches cross-batch memory on or off at margin .9, using the same sampler, wall-clock training budget and time-based learning-rate schedule. On − off effects are negative for both ten-byte allocations in both heads. The interaction between these memory effects is -0.744 AP points, with a two-head standard deviation of 0.933 points. Batch-only contrastive learning still favors Q4D20 by $+1.52$ points on average. The positive allocation effect under batch-only training shows that memory is unnecessary for this tested preference. Two heads do not establish equivalence between memory settings or a broad null effect.

The batch-only head effects are $+1.646, +1.395$ points. Memory changes pair availability and achievable updates, so the negative effect cannot isolate pair quality from computational opportunity cost.

Normalization is not a sufficient repair. Two fresh paired heads compare the ordinary U pathway with a normalized-triplet variant under the same inherited margin, triplets, initialization, and clock. Native AP falls by 7.44 points for Q8D10 and 8.85 points for Q4D20. The allocation interaction is -1.40 points. These outcomes reject normalization alone as a successful drop-in explanation under the tested recipe. A separately retuned normalized objective would be a different experiment, not a reinterpretation of these negative controls.

Conditional intervals are $[-7.73, -7.16]$ and $[-9.16, -8.52]$ points. Retaining the inherited norm penalty and margin defines a drop-in control, not an optimized normalized-triplet baseline.

A strong compressed teacher remains competitive. A separate two-head, 5,400-second cohort compares direct ten-byte descriptors with compression of an eighty-dimensional teacher (Table 2). Teacher OPQ exceeds both direct learning and a distilled student at the same stored item payload. Distillation is a relevant route to compact models (Hinton et al., 2015; Roth et al., 2021), but the teacher is an additional training cost, not a free target. These experiments are not pooled with the 4,500-second allocation study, and no direct-learning superiority over competent compression is claimed.

A CAL-only quality constraint allowing at most $.01$ AP loss relative to the teacher selects the uncompressed teacher in both heads. Hence an allocation ranking among ten-byte candidates does not

Table 2: Two-head mean SOP product-macro AP in a separate 5,400-second cohort. These values do not share the main cohort's training budget. The comparison retains the stronger compressed-teacher result rather than identifying direct low-bit learning with the best available ten-byte system.

Representation	AP (%)	Item bytes	Required training / shared state
Direct compact code	42.44	10	One direct fit
Teacher FP32	59.30	320	One teacher fit
Teacher PQ	40.42	10	Teacher fit plus codec fitting
Teacher OPQ	52.58	10	Teacher fit plus rotation/codebooks
Distilled compact code	45.82	10	Teacher and student fits

Table 3: Partial recoding of two fresh Q4 U heads. Product-macro AP and expected-tie Recall@1/10 are percentages under common zero-epsilon scoring. The gallery column records whether stored gallery codes change.

Recoded roles	Gallery changes	AP (%)	R@1 (%)	R@10 (%)	AP gain (pp)
Neither	No	37.069	46.310	74.300	0.000
Query only	No	41.769	51.131	78.201	4.700
Gallery only	Yes	44.760	54.586	78.970	7.691
Both	Yes	51.153	63.286	82.566	14.085

establish that ten bytes meets the application requirement. The measured study supplies no compact solution to that particular constraint. The conclusion is conditional on the specified quality threshold and codec family; weaker thresholds or other codec families were not evaluated in this decision comparison.

Payload is only one cost. Each packed descriptor occupies ten bytes, but a dense FP32 D1 table occupies 1,280 bytes for Q4D20 and 10,200 bytes for Q8D10. A global range requires one scalar; a coordinate-wise range requires twenty or ten scalars. Materialized tables may exceed those minimal parameter sizes. Changes to a decoder or metric can require rebuilding a search index even when stored codes remain unchanged. Large-scale search libraries make such distinctions operationally important (Johnson et al., 2021; Douze et al., 2024). The experiments establish exact full-gallery rankings, not an end-to-end approximate-index latency improvement.

Partial recoding offers an intermediate quality/cost choice (Table 3). Query-only repair leaves the gallery unchanged but recovers only part of joint repair. Joint recoding under zero-epsilon scoring gains 14.085 points in this fresh cohort, not the selected native cohort's 14.222. All roles retain ten bytes; no serving-speed advantage is measured.

Conditional intervals and role-specific scoring conventions are given in Appendix E.

Block decoding reduces the measured scoring allocation from 4.636 GiB to .479 GiB but increases duration in a fixed-order 256-query profile (Appendix E). Ten-byte payload, scoring memory and search speed remain separate claims.

7 Conclusion

Matching the codec recovers 14.22 AP points from frozen triplet descriptors without exceeding ten bytes. Assignment/decoder crosses locate the global-range repair in recoding; common E4 separates export mismatch from training-grid differences. Separately fitted codecs additionally reverse U–XBM ordering in one new paired realization. C1 also reverses that pair, while SigLIP2 native exports and adverse CUB comparisons constrain generality. These results concern learner–codec compatibility under explicit costs and exposure, not universal learner superiority, a bit requirement, or the dispensability of low-bit training.

AI USE STATEMENT

Generative AI tools assisted hypothesis refinement, mathematical exposition, experimental design and implementation, numerical analysis, literature search and reference verification, figure preparation, and manuscript drafting and editing. The authors are responsible for the final content.

ETHICS STATEMENT

The experiments evaluate compact product and bird-image retrieval using existing benchmarks and pretrained encoders. Repeated dataset exposure and possible pretraining overlap limit conclusions about unseen populations. Smaller stored descriptors alone do not establish lower total computation or environmental impact; decoder state, recoding, and serving costs remain separate.

REPRODUCIBILITY STATEMENT

Appendix B specifies data roles, matched training and selection, and tie-aware evaluation. Appendices C and D provide analytical checks and per-head effects with conditional intervals; Appendix E records adverse controls and storage and serving qualifications. Appendix F specifies the separately fitted codec comparison and additional paired fit.

REFERENCES

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. URL https://arxiv.org/abs/1308.3432.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. URL https://arxiv.org/abs/2010.11929.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. *arXiv preprint arXiv:2401.08281*, 2024. URL https://arxiv.org/abs/2401.08281.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *ICLR*, 2020. URL https://openreview.net/forum?id=rkgO66VKDS.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *CVPR*, pp. 2946–2953, 2013. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Ge_Optimized_Product_Quantization_2013_CVPR_paper.html.

Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, pp. 817–824, 2011. URL https://slazebni.cs.illinois.edu/publications/cvpr11_small_code.pdf.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *ICML*, pp. 3887–3896, 2020. URL https://proceedings.mlr.press/v119/guo20h.html.

Kun He, Fatih Cakir, Sarah Adel Bargal, and Stan Sclaroff. Hashing as tie-aware learning to rank. In *CVPR*, pp. 4023–4032, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/He_Hashing_as_Tie-Aware_CVPR_2018_paper.pdf.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. doi: 10.1109/TPAMI.2010.57. URL https://inria.hal.science/inria-00514462.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. URL https://arxiv.org/abs/1702.08734. Published online in 2019.

Benjamin Klein and Lior Wolf. End-to-end supervised product quantization for image search and retrieval. In *CVPR*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Klein_End-To-End_Supervised_Product_Quantization_for_Image_Search_and_Retrieval_CVPR_2019_paper.pdf.

Mengtian Li, Ersin Yumer, and Deva Ramanan. Budgeted training: Rethinking deep neural network training under resource constraints. In *ICLR*, 2020. URL https://www.cs.cmu.edu/~mengtial/proj/budgetnn/.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *ICLR*, 2024. URL https://arxiv.org/abs/2309.15505.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *ECCV*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123700681.pdf.

Mohammad Norouzi and David J. Fleet. Cartesian k-means. In *CVPR*, pp. 3017–3024, 2013. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Norouzi_Cartesian_K-Means_2013_CVPR_paper.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Karsten Roth, Timo Milbich, Bjorn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based self-distillation for deep metric learning. In *ICML*, pp. 9095–9106, 2021. URL https://proceedings.mlr.press/v139/roth21a.html.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. URL https://arxiv.org/abs/1503.03832.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. URL https://arxiv.org/abs/1511.06452.

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

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://www.vision.caltech.edu/datasets/cub_200_2011/.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, 2019. URL https://arxiv.org/abs/1904.06627.

Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R. Scott. Cross-batch memory for embedding learning. In *CVPR*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Cross-Batch_Memory_for_Embedding_Learning_CVPR_2020_paper.html.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, 2017. URL https://arxiv.org/abs/1706.07567.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. URL https://arxiv.org/abs/2303.15343.

Table 4: Mean allocation effects in AP points. SOP selected/fresh are different two-head cohorts. CUB original uses the same margin .9 for both contrastive allocations; CUB transferred uses .7 for both. The U row is unchanged by that transfer. All columns use the same ten-byte payload; decoder state and recoding requirements differ.

		Original / centroid interface			Range calibration	
Cohort	Learner	D0	D1	D2	C1	C2
SOP selected	U	-5.61	-5.50	-3.47	+8.50	+8.23
	XBM	+5.22	+4.60	+4.85	+5.19	+5.19
SOP fresh	U	-4.88	-4.87	-2.82	+8.98	+8.68
	XBM	+4.51	+3.92	+4.16	+4.51	+4.51
CUB original	U	-11.47	-11.25	-9.83	+1.18	+1.18
	XBM	-9.04	-9.49	-9.48	-9.04	-9.07
CUB transferred	XBM	-8.16	-8.31	-8.37	-8.15	-8.15

A SUMMARY OF CONTROLS

The main-text ladder separates decoder replacement, reassignment, and FIT range calibration while preserving each model. The appendix specifies the actual training/export correspondence, exact-tie metrics, reconstruction bounds, paired effects, and physical costs. Common-dimensional grid effects and same-byte allocation effects remain separate throughout; neither a failed native interface nor a repaired one defines an intrinsic bit requirement.

B TRAINING, SELECTION, AND EVALUATION CONTRACTS

B.1 WHAT REMAINS MATCHED

The central comparison changes the storage allocation while retaining the pretrained body, paired initialization, product-triplet sampler, wall-clock training budget and time-based learning-rate schedule. Pairing does not make gradients identical across losses or dimensions. The twenty-dimensional head provides the initial ten-dimensional prefix, but every fit subsequently optimizes its own model. All decoder controls reuse the exact trained model for the corresponding learner, allocation, and head. No codec intervention changes the optimizer trajectory.

The contrastive recipe is an adaptation to the common DINOv2 and Adam pipeline. The detached memory contains embeddings from previous updates, has capacity 8,192, and starts after 2,000 updates. A positive/negative pair objective with a selected margin replaces the ordinary triplet pathway. Comparison with U therefore combines loss, normalization, and pair-availability differences. The separate memory and normalization controls narrow this bundle but do not factor every component of the training algorithm.

Original SOP learning-rate schedule. For seeds 7202/7203, both Q4D20/Q8D10 and U/XBM at margin .9 use multiplier 1 for the body/head Adam rates $10^{-5}/10^{-3}$. Let t be elapsed wall time from the synchronized rank-zero training-function entry clock and s the synchronized post-cache setup time. The factor is

$$f(t) = \max\{0, 1 - (t - s)/(4500 - 180 - s)\}, \qquad s \le t < 4320.$$

Setup is charged to the 4,500-second allocation but excluded from the decay span. New updates begin only before 4,320 seconds, also subject to the 32,768-update cap; 180 seconds are reserved for completion and export. No learning-rate warmup is used. The 2,000-update XBM memory warmup is separate. The last update can end after the start-time cutoff, so the last applied learning rate need not be zero. The eight fits have setup times 100.283–104.937 seconds and 14,419–14,711 updates; matched wall time does not imply equal updates or equal rates at the same update index. This setting confirmation concerns the original eight SOP fits, not the later selected-margin, CUB, learned-scale, SigLIP2 or teacher cohorts.

Table 5: A resource and role contract, not a claim that all objectives receive equal image presentations. The separate teacher-compression cohort uses a 5,400-second contract and is not pooled into the main comparison.

Quantity	Main contract
Backbone	DINOv2-base vision transformer; body and projection head trained
Paired head	Twenty-row initial head; ten-dimensional model uses first ten rows
Global image workload	192 product triplets / 576 endpoint images per update
Base learning rates	Body 10^{-5}; projection head 10^{-3}
Clock	4,500 seconds on eight GPUs, including setup and 180-second completion reserve
Schedule	Synchronized wall-clock learning-rate decay; no equal-update assertion
Contrastive memory	8,192 detached embeddings; 2,000-update warmup
Original / selected XBM	Margin .9 / Q4 margin .7; Q8 remains .9 on SOP
CUB transfer	Margin .7 for both Q4 and Q8, without CUB selection
Codec fitting	FIT pre-round vectors only; no TEST reconstruction or label optimization

B.2 Training quantizers and export correspondence

For normalized body feature f, the head forms $z = \tanh(Wf+b)$ before quantization. SigLIP2 uses its native vision pooling and preprocessing; DINOv2 uses its declared pooled feature. No rounded training code supplies a new export's input. With learned positive step s, the forward reconstructions are

$$q_4(z;s) = s\left[\mathrm{clip}_{[0,15]}\big(\mathrm{round}(z/s+7.5)\big) - 7.5\right], \tag{6}$$

$$q_8(z;s) = s\,\mathrm{clip}_{[-127,127]}\big(\mathrm{round}(z/s)\big). \tag{7}$$

Rounding uses ties to even. T4 has sixteen midrise levels, including no zero; T8 has 255 signed levels. Endpoint magnitude is $Q_p s$, with $Q_p = 7.5$ or 127. The fixed-grid endpoints are 1. A learned step is initialized as $2\,\mathrm{mean}_{\mathrm{FIT}}\,|z|/\sqrt{Q_p}$ on 1,024 specified FIT representations. The head's Adam rate, initially 10^{-3}, also applies to the step; body rate is 10^{-5}. Adam uses $(.9, .999)$, $\epsilon = 10^{-8}$ and zero weight decay. The step is projected to at least 10^{-8} after each update. All four SigLIP2 fits record zero positivity projections.

Writing $u = z/s$ and $k = q/s$, the clipped straight-through derivative in z is one strictly inside $|u| < Q_p$ and zero outside. The step-size gradient uses $k - u$ inside and k outside, with scale $1/\sqrt{576dQ_p}$. This is the tested embedding adaptation, not a full reproduction of LSQ network quantization. The learned and fixed recipes therefore differ in initialization and gradient rule as well as in the presence of a trainable scalar.

Table 6: SigLIP2 step and export correspondence from recorded training values. The final endpoint is $Q_p s$, not the step itself. Every C1 range is .25. Stored training codes are not the common-E4 D0 control.

Head	Training	Initial step	Final step	Final endpoint	C1 range	Train/export bytes
1	T4	.023949	.016376	.122818	.25	10/10
1	T8	.005820	.003340	.424132	.25	20/10
2	T4	.017490	.016821	.126155	.25	10/10
2	T8	.004250	.002046	.259782	.25	20/10

Four exports must remain distinct. A training-step-preserving export retains s and the training alphabet; T8D20 requires twenty bytes. Common-E4 D0 resets the floating representation's endpoint to 1, while C1 uses the FIT-selected endpoint. A T8 endpoint-preserving E4 instead sets $s'_4 = (127 s_8)/7.5$ and directly quantizes the floating representation to sixteen levels. Rounded T8 codes are not re-rounded. The new FP32 step preserves the endpoint up to rounding; all E4 variants

use ten bytes. Each deployed head additionally stores one shared FP32 scalar (4 B), excluding optional decoder tables and serialization. Endpoint-preserving E4 is a new export, not the native T8 code.

Table 7: SigLIP2 full-gallery product-macro percentages, averaged over two fixed heads with zero-epsilon expected ties. Reset E4 discards the learned step; native T8 is a separate twenty-byte reference. Shared scalar state is excluded from item bytes.

Grid	Export	Item B	AP (%)	R@1 (%)	R@10 (%)
T4	Learned-step native	10	62.816	72.803	89.143
T4	Reset E4, range 1	10	34.926	42.434	72.627
T4	FIT-calibrated E4	10	62.036	71.650	88.999
T8	Learned-step native	20	65.264	75.387	89.877
T8	Reset E4, range 1	10	31.747	38.900	69.444
T8	Endpoint-preserving E4	10	58.562	67.675	87.720
T8	FIT-calibrated E4	10	61.918	71.486	88.966

C1 $-$ native T4 has head effects $-.783/-.778$ AP points and mean $-.781$ with conditional product interval $[-.895, -.672]$. Recall@1 changes by -1.153 points $[-1.346, -.958]$ and Recall@10 by $-.144$ $[-.263, -.022]$. Products are resampled after averaging fixed heads; these intervals do not estimate seed-population uncertainty. C1 reduces FIT coordinate MSE from $(2.494, 2.209) \times 10^{-4}$ to $(.926, .928) \times 10^{-4}$, reductions of $62.86\%/57.99\%$. The native endpoints lie outside the seven-range menu but have higher MSE than selected C1. Adding only those endpoints would not change the MSE-minimizing choice. Lower reconstruction error therefore coexists with worse retrieval in both heads.

Native T8 exceeds C1 by 3.346 AP points but uses twenty bytes. Native T4 $-$ ten-byte endpoint-preserving T8 is $+4.255$ points $[+4.019, +4.499]$, with head effects $+7.354/+1.155$. The conversion itself loses 3.356 points to its own C1 export $[-3.506, -3.213]$, with effects $-6.438/-.274$. The former is the specified primary contrast; the own-C1 comparison is secondary with pointwise conditional intervals. Neither replaces the common-C1 T4 $-$ T8 effect or isolates training bit width.

The DINOv2 Q4 U native-code comparison retains ten bytes in both arms: C1 changes AP from 50.858% to 51.099%, Recall@1 from 62.811% to 62.761%, and Recall@10 from 82.383% to 82.735%. A small AP improvement need not improve every retrieval depth.

The U objective acts on the dequantized q, without an additional normalization. For 192 anchor/negative/positive triplets and Q_r the corresponding $192 \times d$ matrices,

$$L_U = \frac{1}{192} \sum_i [\|q_{a_i} - q_{p_i} + \eta\|_2 - \|q_{a_i} - q_{n_i} + \eta\|_2 + .2]_+ + \frac{.005}{3\sqrt{192}} \sum_{r \in \{a,n,p\}} \|Q_r\|_F. \quad (8)$$

Here $\eta = 10^{-6}\mathbf{1}$. XBM instead normalizes dequantized codes, $f_q = q/\max(\|q\|_2, 10^{-12})$, and computes FP64 similarities $a_{ij} = f_{q_i}^\top f_{q_j}$. The pair objective includes $1-a_{ij}$ terms for retained same-label pairs and a_{ij} terms for different-label pairs with $a_{ij} > m$. The positive-pair filter excludes near-unit similarities. The margin gates negative pairs; the summand is a_{ij}, not $[a_{ij} - m]_+$. Current-batch references receive gradients on both sides. After warmup, detached FIFO references contribute with weight one. These normalization, pair-selection and memory differences prevent a loss-only causal interpretation of U–XBM contrasts.

B.3 Split and exposure accounting

SOP training roles are disjoint by product. FIT contains 35,569 images, CAL 12,056, and a training-side readout 11,926, summing to 59,551. The official 60,502-image TEST gallery is separate by product from these roles. Recipe choice uses CAL; reconstruction tables and ranges use FIT only. A new head can be independent of the recipe-selection head while sharing the same exposed test population. Fresh optimization and fresh population sampling are therefore different properties.

CUB training species 1–100 contain 5,864 images; test species 101–200 contain 5,924. The three original paired initializations train U and margin-.9 contrastive models at both allocations. The .7 transfer adds only the six contrastive fits. The U models are reused in the transfer comparison,

not counted again. Codec fitting uses every eligible training representation under a fixed, common procedure, including the same empty-cell fallback. A test outcome never changes a range grid, a codebook fit, or the transferred margin.

These separations prevent direct label leakage but do not undo prior research exposure. The later decoder, range, normalized-loss, and transferred-margin analyses are explicitly diagnostic extensions on previously inspected benchmarks. A broader statement about unseen image populations would require new datasets or prospectively withheld cohorts. Benchmark overlap with public backbone pretraining remains an additional limitation rather than an assumption of complete data independence.

B.4 Metric definitions and exact ties

For a query with r_q relevant gallery items after self-exclusion, average precision is the mean precision at the relevant ranks. Under a distance tie, the evaluator averages over uniformly ordered permutations of the tied group. Let a group contain n items and m relevant items, with N items and A relevant items preceding the group. For $n > 1$, the group's expected numerator contribution is

$$\sum_{j=1}^{n} \frac{m}{n} \frac{A + 1 + (j-1)(m-1)/(n-1)}{N + j}. \tag{9}$$

The singleton case uses the ordinary precision contribution. Division by r_q gives the group's contribution to query AP. The conditional expectation in Equation 9 follows by first conditioning on relevance at within-group position j; the remaining $m - 1$ relevant items are uniformly distributed over the remaining $n - 1$ positions. This eliminates arbitrary file-order tie breaking, which can be consequential for low-precision codes (He et al., 2018).

Recall@k is the probability of at least one relevant item among the first k ranks under the same uniform tie ordering. When the cutoff includes t positions from a group of n items containing m positives, with no earlier positive, the contribution is $1 - \binom{n-m}{t} / \binom{n}{t}$ (equal to one when $t > n-m$). Product averaging then follows the AP convention. This metric is not the fraction of all relevant items retrieved.

SOP first averages query AP inside each product and then averages products. CUB uses the ordinary mean over queries. For CUB conditional bootstrap intervals, species are resampled with replacement, and each resampled species contributes its original query set; normalization by the resulting query count preserves the stated query-weighted estimand. These protocols need not yield the same numerical effect as an unweighted mean over CUB species or a query-micro SOP score.

B.5 CAL-labelled codec selection

A post-hoc follow-up chooses between each learned-native export and its fixed FIT-selected C1 export by product-macro AP on 12,056 CAL images/2,264 products, disjoint from FIT. Exact AP ties choose native; Recall does not break ties. Both query and gallery use the selected ten-byte codec under zero-epsilon expected-tie scoring. All eight learned-scale DINOv2 heads and both SigLIP2 T4 U heads are included. Earlier fixed-grid caches lack CAL floating latents and are not re-quantized from rounded bytes.

All ten decisions precede the new evaluation of existing TEST arrays. Both heads choose C1 for DINOv2 Q4 U, Q4 XBM and Q8 XBM, and native for DINOv2 Q8 U and SigLIP2 Q4 U. The same choices are evaluated without reselection on the separate 11,926-image SOP readout gallery. Selection adds CAL relevance labels beyond label-free FIT reconstruction fitting. Earlier TEST exposure motivated the comparison; the selection rule is not prospectively designed on untouched data. Two fixed heads are averaged within product before 4,000 paired product-bootstrap draws. TEST has 11,316 products and readout 2,264; roles use separate draws and are not pooled. Each role has five primary AP intervals; no historical-experiment adjustment or model/selection uncertainty is included. Only Q4 XBM retains a positive primary interval on readout. The SigLIP2 advantage over C1 is secondary, not a gain over native. DINO Q4 U TEST Recall@1 falls $.050$ points $[-.257, +.165]$ despite positive AP. Readout Q8 U loses $.064$ Recall@10 points relative to C1 with an interval spanning zero; SigLIP2's analogous effect is $-.025$ $[-.190, +.148]$. AP selection does not establish improvement at every retrieval depth.

Table 8: CAL-selected native/C1 exports. AP is percent; differences are AP points. Primary selected − native intervals are nominal 99%; secondary selected − C1 intervals are pointwise 95%, conditional on fixed heads and choices. Native-choice zeros are policy identities.

Role	Model	Choice	AP	Δ native	Δ C1
TEST	DINO Q4 U	C1	51.099	$+.242[+.085, +.402]$	$0[0, 0]$
	DINO Q4 XBM	C1	50.452	$+2.629[+2.408, +2.842]$	$0[0, 0]$
	DINO Q8 U	Native	42.145	$0[0, 0]$	$+.019[-.010, +.050]$
	DINO Q8 XBM	C1	43.151	$+.039[-.001, +.079]$	$0[0, 0]$
	SigLIP2 Q4 U	Native	62.816	$0[0, 0]$	$+.781[+.666, +.893]$
Readout	DINO Q4 U	C1	66.609	$+.265[-.038, +.565]$	$0[0, 0]$
	DINO Q4 XBM	C1	64.674	$+2.619[+2.158, +3.081]$	$0[0, 0]$
	DINO Q8 U	Native	59.742	$0[0, 0]$	$-.006[-.073, +.066]$
	DINO Q8 XBM	C1	60.310	$+.011[-.069, +.086]$	$0[0, 0]$
	SigLIP2 Q4 U	Native	77.828	$0[0, 0]$	$+.446[+.222, +.655]$

Readout products and images are disjoint from FIT, CAL and official TEST and were not used for gradients, checkpoint selection or range fitting. Historical readout exposure may still have informed recipe design. This is another SOP gallery, not a new domain or untouched research population. No gallery-size-only explanation or comparison with an OPQ/teacher budget follows.

C Classical bounds and limits of reconstruction diagnosis

Table 9: Held-out coordinate reconstruction MSE, multiplied by 10^3, under frozen Q4D20 checkpoints. Means use the same two SOP or three CUB heads. Range selection uses FIT only. Reconstruction MSE is a diagnostic, not the retrieval objective.

Dataset	Q4 learner	Range	D0 FP32	D1	D2	C1
SOP	U	0.25	1.482	1.359	1.249	0.218
SOP	XBM	1.00	1.485	1.316	1.259	1.485
CUB	U	0.25	1.510	1.322	1.250	0.093
CUB	XBM .7	1.00	1.347	1.470	1.241	1.347

C.1 Fixed-assignment centroid bound

For one nonempty FIT cell, let $S = \{i : c^0_{ij} = a\}$ and $\bar{z} = |S|^{-1} \sum_{i \in S} z_{ij}$. For any reconstruction value v,

$$\sum_{i \in S} (z_{ij} - v)^2 = \sum_{i \in S} (z_{ij} - \bar{z})^2 + |S|(v - \bar{z})^2. \tag{10}$$

The cross term vanishes because $\sum_{i \in S} (z_{ij} - \bar{z}) = 0$. Hence the centroid minimizes squared reconstruction error with fixed assignments. Summing over coordinates and nonempty cells proves the D1 FIT bound. Empty cells contribute zero to the FIT objective, so retaining the old midpoint is a deterministic fallback rather than an estimated centroid. FP32 table storage can introduce a small numerical discrepancy relative to the exact FP64 minimizer; the physical table, not an unstored ideal value, defines evaluation.

C.2 Fixed-table reassignment bound

For each FIT vector and coordinate, the previous index c^0_{ij} belongs to the same candidate set used by Equation 3. Therefore

$$(z_{ij} - \mu_{j,c^1_{ij}})^2 \le (z_{ij} - \mu_{j,c^0_{ij}})^2. \tag{11}$$

Summation gives the D2 reconstruction bound under the D1 table. No second centroid fit is required. This is the ordinary nearest-centroid step. A full Lloyd iteration would refit after reassignment and would define a different intervention.

C.3 PRE-ROUND AND EXPORTED AP

The identity in Equation 5 compares two separately evaluated rankings. The signed export gap D may be negative: discretization can occasionally improve AP. Calling D a nonnegative quantization loss would add an assumption absent from the metric. Likewise, the decomposition is not additive at the individual-distance level; AP changes discontinuously when rankings or ties change. The .7 margin analysis uses the exact checkpoint-specific pre-round vectors and exports both gallery and queries under each grid, avoiding the invalid shortcut of applying a finer quantizer to an already rounded coarse code.

D PAIRED RESULTS AND CONDITIONAL INTERVALS

Table 10 reports individual paired-head effects alongside conditional intervals. Within each cohort, head numbers identify paired initializations; each head receives the same codec comparisons. The initial and fresh SOP cohorts are not pooled. CUB uses the same three paired initializations across original and transferred margins; the triplet row is not duplicated for the transferred family. All values are percentage points of the stated AP metric.

Table 10: Per-head allocation effects and conditional 95% intervals for all five codecs in Figure 2, rounded to two decimals. The interval resamples products/species conditional on the displayed models, not heads.

Cohort	Learner	Codec	Individual heads	Conditional interval
SOP selected	U	D0	-5.99, -5.23	[-5.93, -5.28]
SOP selected	U	D1	-5.78, -5.22	[-5.82, -5.19]
SOP selected	U	D2	-3.70, -3.23	[-3.77, -3.17]
SOP selected	U	C1	+8.01, +9.00	[+8.21, +8.77]
SOP selected	U	C2	+7.53, +8.94	[+7.95, +8.49]
SOP selected	XBM	D0	+4.59, +5.84	[+4.90, +5.52]
SOP selected	XBM	D1	+3.98, +5.22	[+4.28, +4.91]
SOP selected	XBM	D2	+4.21, +5.49	[+4.54, +5.16]
SOP selected	XBM	C1	+4.57, +5.80	[+4.87, +5.50]
SOP selected	XBM	C2	+4.57, +5.80	[+4.87, +5.50]
SOP fresh	U	D0	-4.82, -4.93	[-5.17, -4.57]
SOP fresh	U	D1	-4.87, -4.87	[-5.17, -4.57]
SOP fresh	U	D2	-2.95, -2.70	[-3.12, -2.52]
SOP fresh	U	C1	+8.74, +9.23	[+8.71, +9.27]
SOP fresh	U	C2	+8.40, +8.96	[+8.40, +8.97]
SOP fresh	XBM	D0	+3.96, +5.07	[+4.23, +4.83]
SOP fresh	XBM	D1	+3.38, +4.46	[+3.63, +4.23]
SOP fresh	XBM	D2	+3.59, +4.73	[+3.87, +4.48]
SOP fresh	XBM	C1	+3.94, +5.08	[+4.22, +4.83]
SOP fresh	XBM	C2	+3.94, +5.08	[+4.22, +4.83]
CUB original	U	D0	-12.84, -10.43, -11.15	[-12.82, -10.12]
CUB original	U	D1	-12.54, -10.29, -10.92	[-12.61, -9.94]
CUB original	U	D2	-10.92, -9.18, -9.39	[-11.09, -8.58]
CUB original	U	C1	-0.19, +2.26, +1.47	[+0.46, +1.90]
CUB original	U	C2	-0.19, +2.26, +1.47	[+0.46, +1.90]
CUB original	XBM	D0	-10.02, -8.86, -8.23	[-10.65, -7.48]
CUB original	XBM	D1	-10.28, -9.64, -8.56	[-11.14, -7.93]
CUB original	XBM	D2	-10.30, -9.49, -8.67	[-11.13, -7.92]
CUB original	XBM	C1	-10.02, -8.85, -8.24	[-10.65, -7.50]
CUB original	XBM	C2	-10.05, -8.89, -8.25	[-10.68, -7.52]
CUB transferred	XBM	D0	-7.54, -8.25, -8.69	[-9.59, -6.70]
CUB transferred	XBM	D1	-7.61, -8.40, -8.91	[-9.74, -6.84]
CUB transferred	XBM	D2	-7.71, -8.47, -8.94	[-9.81, -6.91]
CUB transferred	XBM	C1	-7.52, -8.25, -8.68	[-9.58, -6.69]

Table 10 continued

Cohort	Learner	Codec	Individual heads	Conditional interval
CUB transferred	XBM	C2	-7.54, -8.24, -8.69	[-9.58, -6.70]

For the selected Q4 checkpoint, the per-head pre-round improvements are $+4.175$ and $+4.403$ points; E4 improvements are $+4.348$ and $+4.530$. The corresponding means are $+4.289$ and $+4.439$. A change of roughly 0.15 points in the export gap is much smaller than the change in pre-round AP. Both the continuous improvement and its retained low-bit benefit are observations; the responsible training mechanism remains unresolved.

Table 11: Per-head effects of the stronger controls, in AP points. Each pair uses the same 11,316 SOP products. Reusing products, export variants or historical comparators does not increase the number of independent trained heads. The main-text intervals condition on these models.

Contrast	Head 1	Head 2
Historical U, T4 − T8, common C1 E4	$-.652$	$-.066$
Historical XBM .9, T4 − T8, common C1 E4	-7.411	-7.869
Learned DINOv2 U, T4 − T8, common C1 E4	$+.333$	$-.270$
Learned SigLIP2 U, T4 − T8, common C1 E4	$+.133$	$+.103$
XBM Q4 .7, learned − fixed, C1	$+2.373$	$+2.236$
XBM Q8 .9, learned − fixed, C1	$-.717$	$-.253$
U Q4, learned − fixed, C1	$+.223$	$-.331$
U Q8, learned − fixed, C1	$-.330$	$+.221$

Table 12: Within-layout learned-scale − fixed-grid effects after equal C1 opportunities. These AP-point effects differ from the same-dimensional T4 − T8 estimand. Intervals condition on two paired heads and the exposed SOP products.

Learner and layout	Mean	Conditional 95% interval
U, Q4D20	$-.054$	$[-.273, +.166]$
U, Q8D10	$-.054$	$[-.276, +.159]$
XBM .7, Q4D20	$+2.305$	$[+2.025, +2.582]$
XBM .9, Q8D10	$-.485$	$[-.696, -.255]$

The same-margin .7 Q8D10 follow-up gives a learned − fixed C1 effect of -9.205 AP points, with head effects $-.560/-17.850$. Relative to the .9 scale effect, the mean difference is -8.721 points, but head differences are $+.156/-17.597$. The Q4 − Q8 difference between scale effects at margin .7 is $+11.510$ points. These paired descriptive results remove the margin mismatch without isolating bit width from dimension, scale initialization or quantizer derivative. No stable seed-population effect is inferred from the heterogeneous pair.

The SigLIP2 family comprises four encoder-fine-tuning fits, not frozen-feature probes. Both training grids use twenty dimensions; the final common E4 payload is ten bytes. Independent FIT range fits use the same seven-value menu. The DINOv2 learned T8 extension reuses the trained T4 pair, rather than counting reused models as new fits. Neither extension creates a new test population.

D.1 A FIXED POSITIVE-RANK AND WEAK-ORDER PANEL

A deterministic, outcome-independent ordering fixes 256 SOP queries before comparison. Both fresh Q4 U heads use every positive among all 60,501 non-self gallery items, under D0, C1 and each checkpoint's pre-round floating representation. The floating reference costs eighty bytes, so the diagnostic is not a rate-matched competitor. FP32 values are promoted before direct FP64 squared-distance calculation; D0/C1 use exact integer distances and zero epsilon.

The 1,494 unique query–positive pairs yield 2,988 head-specific pairs and 8,964 three-arm records. These repetitions are not new independent positives. Exact-tie AP and Recall use the same metric definitions as the full-gallery analysis. For a first-positive tie block with n items, m positives and N

preceding items, expected first-positive rank is $N + (n+1)/(m+1)$. Means below first average heads within query and then queries, not SOP products; no new population interval is inferred. C1

Table 13: Fixed 256-query panel. AP, Recall and AUC are percentages; first-positive rank is an expected position (lower is better). Floating vectors cost eighty bytes rather than ten.

Head	Export	AP	R@1	R@10	First rank	AUC
1	Floating	50.471287	67.734375	85.156250	108.054688	99.258985
	D0	36.304388	50.921689	78.388520	99.085612	99.067728
	C1	48.522354	66.367188	83.007812	106.782552	99.241979
2	Floating	51.267729	66.171875	86.328125	59.859375	99.255088
	D0	36.036422	47.926432	78.780899	65.050326	99.013058
	C1	49.251666	64.674479	85.546875	51.419271	99.221024
Mean	Floating	50.869508	66.953125	85.742188	83.957031	99.257036
	D0	36.170405	49.424061	78.584709	82.067969	99.040393
	C1	48.887010	65.520833	84.277344	79.100911	99.231501

gains 12.716605 AP points over D0 but remains 1.982498 below the floating reference. C1 $-$ D0 first-positive rank changes by $+7.696940$ positions in head 1 and -13.631055 in head 2; the median paired rank change is $-.5$. C1's first positive is earlier than the floating reference in both heads, so the reference is not an upper bound for every metric.

Every positive is compared with every irrelevant item. The 180,754,392 head-specific positive–negative pairs fall into 27 states of floating, D0 and C1 weak orders, each in $\{<, =, >\}$. Credits $1, .5, 0$ define pairwise AUC. Table 14 exhaustively groups the states; its contributions are an AUC identity, not an AP decomposition or causal explanation. Only $1,270,126$ pairs $(.703\%)$

Table 14: Complete weak-order partition. F, D, C denote floating, D0 and C1 orders. Counts weight all pairs; query percentages and AUC contributions average heads within query first. AUC points are not AP points.

Category	Pair count	Pair %	Query %	Δ AUC (pp)
Stable: $F = D = C$	179,346,103	99.220883	99.243437	.000000
Restored: $F = C \neq D$	1,058,358	.585523	.562598	+.212094
Departed: $F = D \neq C$	179,245	.099165	.099428	-.020472
Retained deviation: $D = C \neq F$	138,163	.076437	.076559	.000000
All different	32,523	.017993	.017978	-.000513

change between D0 and C1. Among those pairs, 83.327% restore the floating order, 14.112% depart and 2.561% take a third order. These are pair fractions, not query fractions or shares of AP gain. Equal-query conditional summaries instead give restoration/departure/third fractions $86.077/10.995/2.928\%$ and $87.810/9.488/2.702\%$ for the two heads. Their denominators are 244/243 changed queries; the remaining 12/13 have undefined conditional fractions.

Restoring the reference can hurt: 209,022 pairs change from positive-ahead under D0 to negative-ahead under both floating and C1, and 56,155 change from tied to negative-ahead. The net contribution is $+.191109$ AUC points, not an explanation of the $+12.716605$ AP-point gain. Literal floating equality gives twelve positive–negative ties; the smallest nonzero gap is 3.32×10^{-8}, with none within eight ULPs. These observed gaps do not establish platform-independent real-arithmetic ordering. No new fit, query selection or confidence interval is added.

E Adverse controls and physical cost

E.1 Separate teacher-compression state

The teacher PQ/OPQ implementation stores 107,520 bytes of shared codec state. A full FP32 decoded gallery would add 320 bytes per item. Block decoding of 2,048 gallery rows requires $2{,}048 \times 80 \times 4 = 655{,}360$ bytes for that reconstruction block, excluding other buffers. Packed stor-

age for all 60,502 SOP items remains 605,020 bytes before identifiers and index structures. Equal item bytes alone would hide these shared and temporary costs.

Role-specific recoding and block decoding. For partial recoding, the query-only gain over original D0 is 4.700 points, with conditional interval $[4.556, 4.853]$; gallery-only gains 7.691 points, $[7.532, 7.853]$. Their deficits relative to joint recoding are -9.385, $[-9.590, -9.202]$, and -6.394, $[-6.557, -6.240]$. Common integer units preserve each role's physical scale; normalizing the roles separately would define a different intervention. Both role payloads remain ten bytes, and no new trained models or independent test population are introduced.

The OPQ comparison also distinguishes packed storage from a decoded scoring cache. A block-decoding evaluator preserves all ten recorded native metrics while avoiding a full 320-byte-per-item cache. In a fixed-order profile of 256 queries on four GPUs, packed decoding takes $.245/.295$ seconds across the two heads versus $.041/.071$ seconds for a materialized reference. Peak reported allocation falls from 4.636 GiB to .479 GiB. These measurements demonstrate a storage–latency trade-off, not a production throughput advantage: the single order does not control cache warming. Consequently, a ten-byte payload, lower scoring memory, and faster search are three separate claims. Only the first is fixed by the central experiment.

E.2 Storage and scoring interfaces

Table 15: Physical payload and minimal parameter state are different quantities. Table sizes are shared per encoder. Index metadata, query representations, buffers, and any rebuilt ANN index are additional.

Interface	Q4D20	Q8D10	Additional work
Packed item payload	10 B	10 B	Same item storage
D1 FP32 decoder table	1,280 B	10,200 B	Shared; fixed codes, changed reconstruction
C1 minimal range state	4 B	4 B	Recoding; physical scorer may materialize tables
C2 minimal range state	80 B	40 B	Per-coordinate recoding and range state
Pre-round item vector	80 B	40 B	Diagnostic, not the ten-byte contract

E.3 Unresolved extensions

The study evaluates CUB E8 export, zero-epsilon scoring, sampled margin transitions, partial recoding, scale-aware allocation controls, SigLIP2 native and common exports, the same-margin Q8 learned-scale pair and a fixed positive-rank panel. Prospectively withheld populations and XBM transfer to another pretrained family remain unmeasured. The 256-query panel does not extend the 4,096-triple analysis to every query or establish causal mediation of AP. Scoring diagnostics use fixed trained models.

F CODECS FITTED TO EACH LEARNED REPRESENTATION

Fit and selection. Ten existing learned-scale heads comprise two paired DINOv2 realizations at both allocations under U/XBM and two SigLIP2 Q4 U heads. Each head supplies its own frozen FIT latent; the separate eighty-dimensional teacher is not reused. FIT has 35,569 images. PQ uses ten subspaces with 256 centroids each, 75 iterations and three restarts; OPQ uses 50 outer iterations followed by the same final PQ fit. The seed is 260919 and fitting uses four CPU threads. Query and gallery are both compressed and scored symmetrically, rather than using uncompressed-query asymmetric distance.

For each head, CAL's 12,056 queries from 2,264 products select literal maximum product-macro expected-tie AP among native, C1, PQ and OPQ, with ties resolved in that order. TEST's 60,502 queries from 11,316 products do not reselect the codec. Each descriptor remains ten bytes; shared D20 PQ/OPQ state costs 20,480/22,080 bytes, and D10 state costs 10,240/10,640 bytes. Native/C1 scalar state is four bytes. Encoder, index and optional decoded-gallery costs are additional.

Table 16: CAL-selected codec − competent native export for existing learned-scale heads. AP and recall changes are percentage points. The five AP contrasts use individual nominal 99% paired-product intervals for a 95% Bonferroni family target.

Panel	Selected codecs	AP [99% interval]	R@1	R@10
DINO Q4 U	OPQ / PQ	.920 [.770, 1.062]	.871	.637
DINO Q4 XBM	PQ / OPQ	4.587 [4.385, 4.780]	4.735	2.990
DINO Q8 U	native / native	0 [0, 0]	0	0
DINO Q8 XBM	PQ / C1	.059 [.019, .098]	.111	.016
SigLIP2 Q4 U	PQ / PQ	.549 [.411, .689]	.376	.383

The two fixed head effects are averaged within product before 4,000 paired-product bootstrap resamples (seed 2609261545). The Q8 U result is an exact no-op because both heads select native. All five pointwise 95% OPQ − PQ AP intervals include zero. These codecs do not reverse the Q4/Q8 allocation winner. A different, exploratory learner contrast does change sign: mean Q4 XBM − U AP is -3.034 points for native, $-.647$ for C1, $+.583$ for PQ, $+.651$ for OPQ and $+.633$ for CAL selection. Both existing pairs reverse under PQ/OPQ. This post-result observation is separate from the five codec-recovery contrasts.

Table 17: Exploratory Q4D20 learner contrasts in the two existing paired realizations. Entries are XBM − U AP in percentage points, not Q4 − Q8 allocation effects.

Codec	Pair 1	Pair 2	Mean
Native	-3.3633	-2.7038	-3.0336
C1	-1.2795	$-.0143$	$-.6469$
PQ	.2284	.9378	.5831
OPQ	.2480	1.0534	.6507
CAL-selected	.3161	.9505	.6333

An additional paired training realization. A new Q4D20 U/XBM pair shares initialization and the SOURCE training stream, under the same eight-GPU, 4,500-second allocation including setup and a 180-second completion reserve. The full body is trained; global batch is 576 images and XBM retains the 2,000-update warmup and 8,192-entry memory. U/XBM complete 14,257/14,320 updates and 8,212,032/8,248,320 image presentations. The first 14,257 update draws agree. Equal wall-clock budgets do not imply equal updates. FIT-only codec fitting and CAL selection precede this pair's TEST export; CAL chooses PQ for U and OPQ for XBM.

Four thousand product-paired resamples use seed 2609269263. The native upper bound is below zero and the selected lower bound above zero; both signs, rather than their interaction alone, establish the reversal for this pair. The selected-minus-native interaction is 3.935 points, with descriptive 95% interval $[3.681, 4.199]$. These intervals condition on fixed models, codecs and previously exposed products, not new training seeds or a new population.

Table 18: Additional Q4D20 pair on the exposed SOP population. AP and recall are percentages; the last column is XBM − U AP in points. Native and selected are the two primary contrasts with individual nominal 97.5% intervals; other codec intervals are descriptive 95%.

Codec	U AP / R@1 / R@10	XBM AP / R@1 / R@10	AP difference [interval]
Native	50.348 / 62.410 / 82.220	48.267 / 60.868 / 80.565	-2.081 $[-2.548, -1.621]$
C1	50.338 / 62.210 / 82.289	51.235 / 63.684 / 82.476	$.897$ $[.496, 1.283]$
PQ	51.036 / 63.042 / 82.414	52.912 / 65.598 / 83.390	1.876 $[1.488, 2.262]$
OPQ	50.885 / 62.907 / 82.569	52.890 / 65.504 / 83.386	2.005 $[1.620, 2.393]$
Selected	51.036 / 63.042 / 82.414	52.890 / 65.504 / 83.386	1.854 $[1.415, 2.302]$

C1 also reverses the new pair, so PQ/OPQ are not uniquely responsible. Applying the same PQ or OPQ codec to both learners also reverses the ordering, excluding unequal selector flexibility as the sole explanation. CAL-selected XBM still loses on 4,645 products. XBM's TEST-best PQ exceeds the selected OPQ by $.0215$ AP points but does not replace that selection. The contrast changes objective and training path, so memory alone is not identified as a cause. No all-seed or all-dataset superiority follows.

Cost boundary. The existing-head comparison adds twenty quantizer fits and no neural training. Its quantizer CPU phase takes 491.060 seconds; successful GPU profiling, CAL and TEST phases total $.063005$ allocated GPU-hours, excluding initial imports and failed attempts. The additional pair's elapsed fitting times are 4,369.943/4,369.151 seconds for U/XBM, with setup 105.325/100.526 seconds and elapsed times at the end of training 4,320.517/4,320.566 seconds. Its codec fitting phase takes 97.318 CPU seconds. Recorded process-allocation time across profiling, fitting, calibration, export and ranking totals 20.194221 GPU-hours, including all measured fitting and inference attempts but excluding CPU checks, waiting between computations and pretrained-encoder training. These are allocation-time accounts, not active-kernel time or lifecycle cost.