From Intrinsic Quantization Difficulty to Ranking-Aware Distillation for Low-Bit Multimodal Retrieval

Anonymous authors
Paper under double-blind review

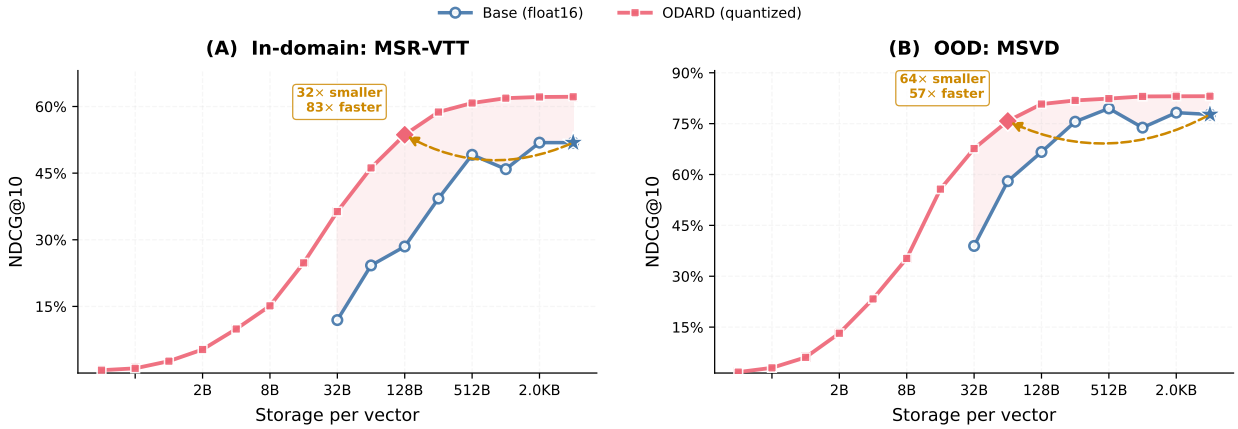

Figure 1: Storage–quality–speed trade-off. ODARD (RAD with online difficulty weighting; Section 4.4) matches the full-precision baseline at 16–64× less storage and up to 83× faster search. **How it is built:** each marker sweeps embedding dimension (128–2048) and bit-width (sign 1-bit, scalar 2/4/8-bit); the x-axis is bits stored per vector and the speed axis is measured wall-clock search time on identical hardware, not an estimate. The full construction and timing protocol are detailed in the appendix.

Abstract

Vector embeddings power retrieval and ranking in modern systems, and low-bit quantization is the standard way to make them cheap to store and search. The dominant recipe minimizes the reconstruction error (MSE) between full-precision and quantized embeddings, assuming that preserving embedding geometry preserves ranking. We show this assumption fails. Under a fixed bit budget, per-sample ranking degradation varies by more than an order of magnitude (up to 21.3×), and reconstruction error is negatively correlated with it: the embeddings that MSE protects most are often not the ones whose rankings break. We name this property Intrinsic Quantization Difficulty (IQD). No static float-space predictor tracks it reliably across regimes. Only a simple online signal computed during training does. IQD shows directly that reconstruction is the wrong objective and that ranking must be preserved, which motivates our core method, Ranking-Aware Distillation (RAD): distill a KL divergence over similarity distributions instead of MSE, which provably preserves pairwise margins in the zero-divergence limit and admits a finite-bit stability bound. RAD wins most settings (+11.7 R@1 on image), transfers zero-shot across domains, and matches or exceeds the full-precision baseline at 16–64× compression. A simple per-sample difficulty weighting derived from IQD adds further, secondary gains where difficulty is most heterogeneous. We validate across three backbones (Qwen3-VL, VLM2Vec, CLIP) and image and video retrieval.

Table 1: Nine predictors of per-sample quantization difficulty. TIER 1 use only float embeddings; TIER 2 require quantized similarities. All evaluated by Spearman $|\rho|$ against `rank_drop`.

#	Predictor	Definition	Quant?
1	`quant_error`	$\|z - q(z)\|_2$	No
2	`density`	Mean cos-sim to k-NN	No
3	`LID`	Local Intrinsic Dimensionality	No
4	`margin_float`	$s(q, t^+) - \max_{t^-} s(q, t^-)$	No
5	`inv_margin`	1/`margin_float`	No
6	`iqd_combined`	`quant_error` × `density`	No
7	`focal_weight`	$(1 - p^{\text{quant}}_{\text{correct}})^\gamma$	Yes
8	`margin_erosion`	`margin_float` − `margin_quant`	Yes
9	`online_diff`	$\mathcal{L}_{\text{quant}} - \mathcal{L}_{\text{float}}$	Yes

1 Introduction

Low-bit quantization is standard for compressing the embeddings that power large-scale retrieval, and the dominant way to train for it is distillation that minimizes Mean Squared Error (MSE) between the float and quantized embeddings (Jégou et al., 2011; Guo et al., 2020). MSE distillation implicitly assumes that all samples degrade roughly equally and that preserving embedding geometry preserves ranking. We find that neither holds. Under a fixed quantization budget, different samples suffer widely different ranking degradation, and, counter-intuitively, per-sample reconstruction error (ℓ_2) is *negatively* correlated with that degradation: the samples MSE tries hardest to protect are not the ones whose rankings break. We call this phenomenon *Intrinsic Quantization Difficulty* (IQD), and we treat fidelity as preserving ranking structure rather than minimizing reconstruction error.

The setting is not hypothetical. Production embedding models now ship quantization-aware training for their output vectors (Li et al., 2026), and the report accompanying our own backbone shows retrieval quality degrading sharply once embeddings drop below a few bits per dimension. IQD explains where that damage concentrates, and why the standard recipe cannot see it coming.

Under the same 2-bit budget, the easiest third of video queries loses 2.7% recall while the hardest third loses 57.4%, a 21.3× gap (Fig. 2B). Across 9 predictors, 2 models, 2 modalities, and 70 settings (Section 3), three findings emerge:

1. **Difficulty variance is large**: up to 21.3× (6.7× on CLIP), consistent across models and modalities (Section 3.2).
2. **Float-space predictors are not regime-robust**: on video all six fail (mean $|\rho| \leq 0.12$); the strongest, `margin_float`, is high only on high-dimensional image and near-zero on video and at low dim; difficulty is emergent wherever quantization is severe (Section 3.3).
3. **Online discovery succeeds**: the loss gap $\Delta\mathcal{L} = \mathcal{L}_{\text{quant}} - \mathcal{L}_{\text{float}}$, computed for free during training, tracks IQD ($\rho = 0.79$–0.90 peak) (Section 3.4).

MSE reconstruction error *anti-correlates* with ranking degradation in some settings ($\rho \approx -0.19$): the embeddings with the highest distortion are not the ones whose rankings collapse.

From phenomenon to method. These findings leave little room for design choices. Our primary method, *Ranking-Aware Distillation* (RAD), replaces MSE with a KL divergence on similarity distributions, which provably preserves all pairwise margins in the zero-divergence limit (Proposition 3) and tightens a finite-bit retrieval bound (Section 4.3). On top of RAD, an *online difficulty weighting* reweights each sample by its loss gap $\Delta\mathcal{L}$; together they add no parameters beyond two scalars and no auxiliary networks. We validate on two training arms (video: MSR-VTT; image: MSCOCO) with a 6-way progressive ablation, bootstrap confidence intervals, post-hoc quantizer baselines, and zero-shot transfer to MSVD (Chen & Dolan, 2011), DiDeMo (Hendricks et al., 2017), and Flickr30k (Plummer et al., 2017). Across the full storage and speed range it matches or exceeds full precision (Fig. 1).

Contributions.

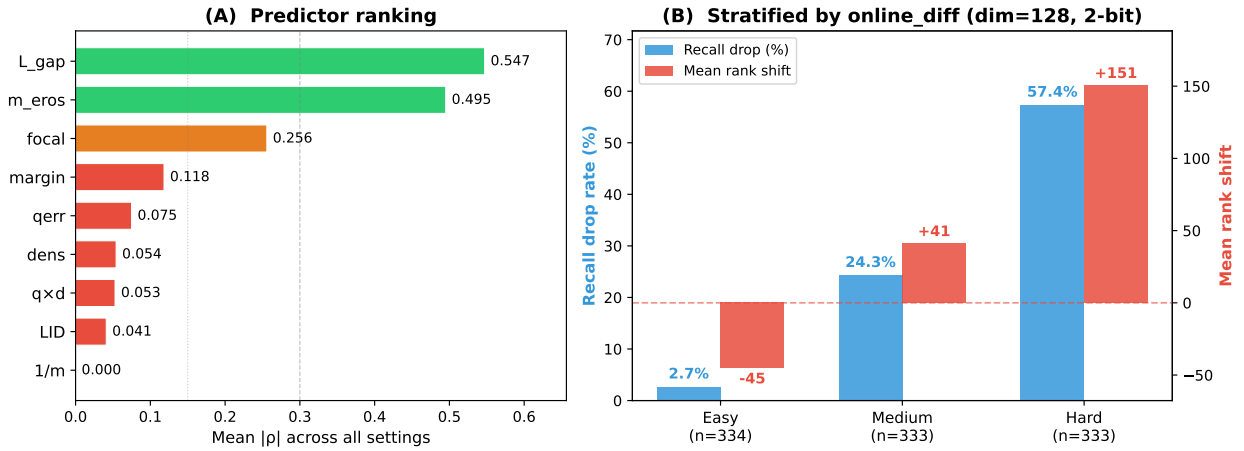

Figure 2: IQD is real but emergent. (A) Nine predictors of per-sample ranking degradation (Table 1); only the online loss gap `online_diff` correlates strongly. (B) Stratified by `online_diff`: easy samples lose 2.7% recall, hard samples lose 57.4% (Qwen3-VL, MSR-VTT, dim=128, 2-bit).

1. **IQD is real and emergent** (Section 3). Per-sample ranking degradation varies by up to 21.3×, yet no float-space predictor anticipates it across regimes; only the online loss gap discovers it everywhere ($\rho = 0.79$–0.90).
2. **Ranking-Aware Distillation (primary)** (Section 4). KL on similarity distributions preserves pairwise margins (Proposition 3) and tightens a per-query Pinsker bound; +11.7 R@1 on image, wins 42 of 45 video-arm evaluation settings, transfers zero-shot across domains.
3. **Cross-backbone validation** (Section 5). We test RAD on three backbones spanning two families (Qwen3-VL, VLM2Vec, CLIP). The large gain is backbone-dependent: it concentrates on Qwen3-VL and shrinks but stays positive (small but visible) on VLM2Vec and CLIP, tracking how much recoverable ranking damage low-bit quantization inflicts, and we characterize the boundary cases with controls (a CLIP hard-vs-soft 2×2).
4. **Conditional difficulty weighting** (Section 7). Online reweighting adds a further gain precisely where IQD heterogeneity is high (in-domain image, +1.7–2.1 R@1, significant at float and 4-bit) and is neutral on MSR-VTT video ($p > 0.18$); far from the training domain it can hurt. We characterize when to enable it and default to RAD alone otherwise.

2 Related Work

Compressing retrieval embeddings. MRL (Kusupati et al., 2022) provides dimensional elasticity; product quantization (Jégou et al., 2011), ScaNN (Guo et al., 2020), and deep hashing (Cao et al., 2017) compress embeddings post-hoc. ScaNN is the closest in spirit: its *anisotropic*, score-aware loss already shows that reconstruction error is misaligned with maximum-inner-product search. That insight operates at the codebook level, with a query distribution known a priori; we show the difficulty is instead emergent and per-sample, surfacing reliably only after quantization. Training the encoder itself for low-bit output embeddings is by now established practice, from binary passage retrieval (Yamada et al., 2021) and cross-view video quantization (Wang et al., 2022) to matryoshka-plus-quantization adaptation (Ahemad, 2025) and the quantization-aware training that production embedding models ship, including our backbone (Li et al., 2026).[1] All of these report aggregate metrics only; the per-sample structure of quantization damage, which is our subject, is not examined in any of them.

Ranking-oriented quantization objectives. For dense text retrieval, MoPQ (Xiao et al., 2021), JPQ (Zhan et al., 2021), and RepCONC (Zhan et al., 2022) established that reconstruction objectives

[1] Non-archival industrial recipes, such as Jina's output QAT of a Qwen-VL-based embedder and mixedbread's binary MRL, occupy the same setting with plain contrastive fine-tuning.

are the wrong target for a learned codebook, and Distill-VQ (Xiao et al., 2022) distills the float embeddings' relevance distribution into the codebook, observing that listwise objectives beat MSE there. We take this objective one level deeper: instead of learning a codebook against a frozen encoder, RAD trains the encoder itself through STE so that its embeddings survive scalar quantization at 1–4 bits, in image and video retrieval, and we supply the margin-preservation and finite-KL analysis this line has lacked. We compare against post-hoc PQ and against Distill-VQ and MoPQ transferred to our setting (Section 6.3).

Quantization-aware training and distillation. STE (Bengio et al., 2013) passes gradients through discretization; LSQ (Esser et al., 2020) learns scales; MatQuant (Nair et al., 2025) co-trains bit-widths for LLM *weights*. KL-based quantization-aware distillation is established for classification logits (SQAKD (Zhao & Zhao, 2024), NVFP4-QAD (NVIDIA Nemotron Team, 2026)), and distillation in the image-text similarity space is the standard primitive for compressing CLIP-style models (Wu et al., 2023; Yang et al., 2024), though never with a quantized branch as the student. Our loss is the retrieval-specific composition of the two: KL between the float and quantized branches' similarity distributions over a candidate pool, tied to pairwise-margin preservation. That KL pins logits only up to a per-sample translation (Kim et al., 2021) is exactly the invariance ranking needs (Proposition 3). QReg (AskariHemmat et al., 2022) and QT-DoG (Javed et al., 2025) note that weight quantization regularizes classification; we observe an analogous effect on embeddings (Section 6.2).

Per-sample structure of compression damage. That compression concentrates its damage on a narrow slice of examples is known for classification under weight pruning and quantization: Hooker et al.'s pruning-identified exemplars over-index on atypical, long-tail examples (Hooker et al., 2019; 2020), and quantized LLMs preserve aggregate accuracy while flipping many individual answers (Dutta et al., 2024). Two things separate IQD from this line. The damaged quantity is a *ranking* over a corpus, not a class label. And the damage is unpredictable where it matters: pruning-identified exemplars correlate with atypicality, and per-input bit allocation in vision succeeds precisely because activation-quantization difficulty is readable from cheap content features (Hong et al., 2022), whereas we find no float-space feature that tracks ranking damage where quantization is severe. On the weighting side, loss-based reweighting is a mature template, from focal loss (Lin et al., 2017) to the reducible-loss prioritization of RHO-Loss (Mindermann et al., 2022); our weighting is its QAT instantiation, the gap between a model's own quantized and float branches, which Section 3.4 shows is the one signal that faithfully reads out IQD. FLIPD (Kamkari et al., 2024) predicts image compressibility via local intrinsic dimensionality; LID fails for retrieval ranking ($\rho = 0.041$). **Gap:** no prior work measures per-sample ranking degradation under output-embedding quantization. It varies by 21.3×, and reconstruction error points the wrong way.

3 Intrinsic Quantization Difficulty

Before designing a method, we ask three questions: *does* per-sample quantization difficulty exist? Can it be *predicted* offline? And if not, can it be *discovered* during training?

3.1 Setup

We study two architecturally distinct models, Qwen3-VL-Embedding-2B (Li et al., 2026) (decoder-VLM, dim=2048) and CLIP ViT-B/32 (Radford et al., 2021) (dual-encoder, dim=512), on MSR-VTT (Xu et al., 2016) video retrieval and MSCOCO (Lin et al., 2014) image retrieval. We sweep 70 (model, data, dim, bit) configurations with uniform scalar quantization and measure per-sample ranking degradation $\texttt{rank_drop}_i = \text{rank}_{\text{quant}}(i) - \text{rank}_{\text{float}}(i)$. Video pools are one-to-one paired; MSCOCO is gallery retrieval, where the correct candidate is matched by ground-truth label rather than by position. CLIP video retrieval scores frame-level embeddings without temporal mean-pooling, so its cells serve as internal comparisons across quantization settings rather than as comparisons to published CLIP video numbers (Section C). Nine candidate predictors (Table 1) split into Tier 1 (float-only, 6 predictors) and Tier 2 (require quantized similarities, 3 predictors). Among Tier 2, `online_diff` is the only signal available during training at zero extra cost and with gradients.

Table 2: Per-sample quantization difficulty varies widely. Settings: Qwen/MSR-VTT dim=128, 2-bit; CLIP/MSR-VTT dim=128, 4-bit; Qwen/MSCOCO dim=256, 1-bit.

Model / Data	Group	n	Recall drop	Mean rank shift
Qwen / MSR-VTT	Easy	334	2.7%	−45
	Med	333	24.3%	+41
	Hard	333	57.4%	+151
CLIP / MSR-VTT	Easy	334	4.2%	−69
	Med	333	13.8%	+9
	Hard	333	28.2%	+165
Qwen / MSCOCO	Easy	334	0.0%	−281
	Med	333	0.0%	−20
	Hard	333	2.1%	+253

Table 3: Predictor comparison at the mean level (mean $|\rho|$ across all (dim, bit) cells of the paired video sweeps). All six Tier 1 float-space predictors fail; only `online_diff` correlates. The image sweeps preserve the same ordering, with `online_diff` at 0.62/0.72 (see text and Section C.1); Fig. 3 maps the per-cell structure of all four sweeps.

	Predictor	Qwen/Vid	CLIP/Vid
Tier 2	`online_diff`	**0.547**	**0.668**
	`margin_erosion`	0.495	0.579
	`focal_weight`	0.256	0.389
Tier 1	`margin_float`	0.118	0.080
	`inv_margin`	0.097	0.088
	`quant_error`	0.075	0.043
	`density`	0.054	0.078
	`iqd_combined`	0.053	0.083
	`LID`	0.041	0.022

3.2 Finding 1: Difficulty variance is real and large

We stratify the test pool into easy/medium/hard thirds by `online_diff` (Table 2). Here the signal serves only as a stratification oracle; whether difficulty can be *predicted* is the subject of Findings 2 and 3. Easy samples barely degrade while hard samples lose over half their recall under the same bit budget. The 21.3× gap on Qwen/MSR-VTT is not model-specific: CLIP shows 6.7×, and on MSCOCO, where absolute recall barely moves at this operating point, the rank-shift stratification is equally sharp (−281 for the easy third versus +253 for the hard third).

3.3 Finding 2: Float-space predictors fail

Proposition 1 (MSE ≠ Ranking)**.** *Minimizing embedding reconstruction error does not guarantee ranking preservation.*

Evidence. On both video backbones (Table 3, 1:1-paired and ground-truth-safe), all six Tier 1 float-space predictors fail (mean $|\rho| \leq 0.12$, with `quant_error` at 0.075), while `online_diff` reaches 0.55–0.67. At the mean level the ordering is unchanged on image: under the GT-correct gallery protocol, `online_diff` leads at 0.62 (Qwen) and 0.72 (CLIP), with the strongest static predictor at 0.45 and 0.11 (Section C.1). What changes on image is the *per-cell* structure (Fig. 3): the float margin `margin_float` becomes strong only at high dimensionality ($\rho \approx 0.88$ in the cells shown, up to 0.92 at the native d=2048; Section C), where quantization is mild enough that the margin survives, and collapses on video and at low dimensionality. The pocket is also backbone-specific: on CLIP image no float-space predictor wins a single cell (`margin_float` mean $|\rho| = 0.106$; Fig. 3D). No static float predictor is robust across regimes; only `online_diff` is. Difficulty is thus *emergent* exactly where quantization bites: it arises from how an embedding interacts with its corpus under quantization, not from geometry alone. A constructive counterexample is in Section B. Concretely, `quant_error`, the quantity MSE minimizes, *anti-correlates* with `rank_drop` at $\rho \approx -0.19$ (Qwen, dim≥512): larger reconstruction error does not imply worse ranking.

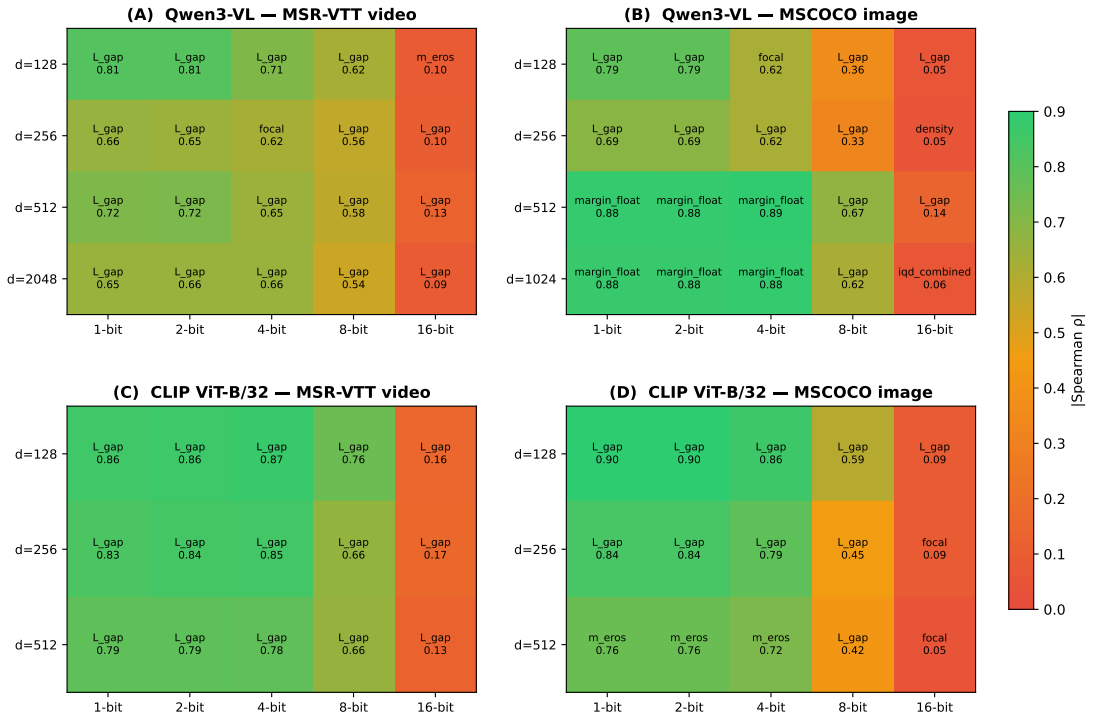

Figure 3: Best predictor per quantization setting, over the full backbone × modality matrix (rows: Qwen3-VL, CLIP; columns: MSR-VTT video, MSCOCO image). Each cell shows the winning predictor and its $|\rho|$ against `rank_drop`. `online_diff` dominates video (18/20 and 15/15 cells) and leads on image (11/20 and 10/15); where it does not win, the winner is almost always another quantization-aware signal. The only static-predictor pocket is six high-dimensional Qwen image cells (`margin_float`, $d \geq 512$, $\leq$4-bit; $\rho \approx 0.88$ here, up to 0.92 at the native d=2048, Section C), where quantization is mild enough that the float margin survives. The pocket does not replicate on CLIP image, where no float-space predictor wins a single cell. `online_diff` is the only regime-robust predictor.

3.4 Finding 3: Online discovery succeeds

Observation 2 (Online proxy)**.** *The per-sample contrastive loss gap* $\Delta\mathcal{L}_i = \mathcal{L}_{\text{quant}}(i) - \mathcal{L}_{\text{float}}(i)$ *is an effective proxy for ranking degradation under quantization.*

Connection. Under InfoNCE, $\Delta\mathcal{L}_i = \log\big(p_{\text{float}}(i)/p_{\text{quant}}(i)\big)$: the log-likelihood ratio of how much quantization damages sample i's retrieval probability. It is available at zero extra cost (both losses are already computed for STE-QAT) and carries gradients through the quantized branch.

To be precise about what is claimed: $\Delta\mathcal{L}$ is a *measurement* of quantization damage, not a prediction from float space; it requires running the quantized branch. That is the point. Findings 1 and 2 show the damage cannot be anticipated before quantizing; Finding 3 shows the training loop already computes a faithful, batch-local readout of it for free. The non-trivial content is empirical: a mini-batch loss gap tracks corpus-level rank degradation, which is what makes difficulty usable during training (Section 4.4) without any offline predictor.

Evidence. Peak Spearman correlations are $\rho = 0.81$ (Qwen video, 1-bit), 0.87 (CLIP video, 4-bit), 0.79 (Qwen image, 1-bit), and 0.90 (CLIP image, 1–2-bit). It leads every Tier 1 predictor on both video backbones and at low dimensionality; the single regime where a float predictor edges it is high-dimensional image at $\leq$4-bit (`margin_float`, where quantization is mild). Full regime-dependence is in Section C; Fig. 3 maps the winning predictor for every (dim, bit) cell.

Cross-modality validation (MSCOCO). On MSCOCO image-text retrieval, `online_diff` remains the strongest online predictor on both backbones ($\rho = 0.79$ at 1-bit on Qwen, 0.90 on CLIP), matching or exceeding its video performance and confirming the signal generalizes across modality. Unlike the diagonally-

paired video benchmarks, MSCOCO is a gallery-retrieval task (one caption against the full image pool), so correlations are computed against the ground-truth matched item rather than a diagonal assumption.

From analysis to method. Difficulty variance exists (21.3×); no static float predictor is robust across regimes (Fig. 3); it is discoverable online ($\rho = 0.79$–0.90). Combined with the MSE anti-correlation, the design is forced: distill *ranking structure*, and weight by *online difficulty*.

4 Ranking-Aware Distillation

Every component follows from the IQD analysis (Section 3). The method has two loss terms and two scalar hyperparameters.

4.1 Quantization-aware backbone

We fine-tune a pretrained multimodal embedding model with LoRA (Hu et al., 2022). A forward pass produces a float embedding $z \in \mathbb{R}^d$; quantized branches at $b \in \{2, 4, 8\}$ bits use the straight-through estimator (Bengio et al., 2013),

$$q_b(z) \;=\; z + \big(\mathrm{round}_b(z) - z\big)_{\mathrm{detach}}, \tag{1}$$

discrete in the forward pass and transparent in the backward pass. The symmetric InfoNCE contrastive loss is summed over the float and quantized branches ($\mathcal{L}_{\mathrm{cont}}$), making the backbone quantization-aware. Without quantization in the forward pass the difficulty signal never arises: IQD is emergent, not pre-existing.

4.2 Ranking distillation (primary)

Rather than MSE on embeddings (Proposition 1), we distill *ranking structure* via a KL divergence on similarity distributions:

$$\mathcal{L}^{(b)}_{\mathrm{distill}} \;=\; \tfrac{1}{2}\Big(\mathrm{KL}\big(p^{\mathrm{f}}_v \,\|\, p^{(b)}_v\big) + \mathrm{KL}\big(p^{\mathrm{f}}_t \,\|\, p^{(b)}_t\big)\Big), \tag{2}$$

where $p^{\mathrm{f}}_v = \mathrm{softmax}(z_v \cdot Z_t^\top/\tau_d)$ is the float-branch teacher (detached), $p^{(b)}_v = \mathrm{softmax}(q_b(z_v) \cdot Z_t^\top/\tau_d)$ the quantized student, and τ_d a distillation temperature distinct from the contrastive one. Listwise distillation of a float teacher's relevance distribution is established for codebook learning (Xiao et al., 2022); RAD applies it one level deeper, through STE to the encoder itself, so the model learns to place its embeddings where the quantizer can preserve their ranking. The total objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{cont}} + \lambda_d\, \mathcal{L}_{\mathrm{distill}}. \tag{3}$$

Proposition 3 (Margin equivalence)**.** *When* $\mathrm{KL}(P_f \| P_q) \to 0$*, the quantized score vector equals the float score vector up to a global constant,* $\mathbf{s}_q = \mathbf{s}_f + C$*; consequently every pairwise margin is preserved,* $\Delta_q = \Delta_f$*.*

Proof sketch. $P_f = P_q$ implies $P_q(d_i)/P_q(d_j) = P_f(d_i)/P_f(d_j)$ for all pairs. Substituting $P(d) \propto \exp(s/\tau_d)$ and taking logarithms gives $s^i_q - s^i_f = C$ for all i, so $\Delta_q = (s^+_f + C) - (s^-_f + C) = \Delta_f$. (Full derivation in the appendix, Section B.) □

4.3 A finite-bit retrieval bound

Proposition 3 is an idealization: it holds only in the limit of zero divergence, which a b-bit code can never reach, since matching all $O(N^2)$ pairwise constraints exceeds the 2^b levels available per dimension. The question that matters in practice is therefore different: *when the divergence is small but nonzero, how far can retrieval drift?* We answer it with a bound whose only input is the quantity RAD already minimizes.

Setup. Fix a query i and its pool of N candidates. Let

$$p^{(i)}_T = \mathrm{softmax}\big(s^{(i)}_T/\tau_d\big), \qquad p^{(i)}_S = \mathrm{softmax}\big(s^{(i)}_S/\tau_d\big) \tag{4}$$

be the teacher (float) and student (b-bit) retrieval distributions over the pool, where $s_T^{(i)}, s_S^{(i)} \in \mathbb{R}^N$ are the similarity scores. The per-query divergence

$$\varepsilon_i \;=\; \mathrm{KL}\big(p_T^{(i)} \,\|\, p_S^{(i)}\big) \tag{5}$$

is exactly the per-sample term that Eq. (2) drives toward zero.

Total variation: the right notion of "how different." We measure the gap between two retrieval distributions by their *total-variation* (TV) distance: the largest amount by which they can disagree on the probability of *any* event $A \subseteq \{1, \ldots, N\}$ (any subset of candidates),

$$\begin{aligned} \mathrm{TV}\big(p_T^{(i)}, p_S^{(i)}\big) \;&=\; \max_A \big|P_T(A) - P_S(A)\big| \\ &=\; \tfrac{1}{2}\,\big\|p_T^{(i)} - p_S^{(i)}\big\|_1, \end{aligned} \tag{6}$$

where $P_T(A) = \sum_{j \in A} p_T^{(i)}(j)$ is the probability the float model retrieves *some* candidate in A, and likewise $P_S(A)$. Intuitively, TV $= 0$ means the two models retrieve identically, while TV $= 0.1$ means no query's retrieval probability for any candidate set can move by more than 0.1. This is precisely the quantity we want to keep small.

The bound. Pinsker's inequality (Cover & Thomas, 2006) relates TV to KL,

$$\big\|p_T^{(i)} - p_S^{(i)}\big\|_1 \;\le\; \sqrt{2\,\varepsilon_i} \quad\Longleftrightarrow\quad \mathrm{TV}\big(p_T^{(i)}, p_S^{(i)}\big) \;\le\; \sqrt{\varepsilon_i/2}, \tag{7}$$

which immediately yields a per-query retrieval guarantee.

Proposition 4 (Finite-bit retrieval stability)**.** *For every query i and every candidate set A,*

$$\big|\, P_T(A) - P_S(A) \,\big| \;\le\; \sqrt{\varepsilon_i/2}. \tag{8}$$

In particular, the probability mass on the gold candidate, i.e. the top-1 retrieval probability, shifts between the float and b-bit models by at most $\sqrt{\varepsilon_i/2}$.

Proof. By definition $\mathrm{TV}(p_T^{(i)}, p_S^{(i)}) = \max_A |P_T(A) - P_S(A)|$. Pinsker's inequality Eq. (7) bounds this maximum by $\sqrt{\varepsilon_i/2}$. Specializing A to the gold singleton gives the top-1 statement. □

A worked example. Suppose a query's float and 1-bit distributions differ by $\varepsilon_i = 0.002$ (a typical value at 1-bit; see below). Then

$$\mathrm{TV} \;\le\; \sqrt{0.002/2} \;=\; \sqrt{0.001} \;\approx\; 0.032. \tag{9}$$

So the probability the 1-bit model assigns to the gold candidate differs from the float model's by at most 3.2 percentage points: if the float model retrieves the gold item with probability 0.80, the 1-bit model retrieves it with probability at least 0.768, for *every* such query, with no distributional assumption on the data. Averaged over our Qwen/Flickr30k pool the measured $\bar{\varepsilon} = 0.0019$ gives the same ≈3% guarantee; at the looser CLIP/MSCOCO soft-quant point $\bar{\varepsilon} = 0.015$ the bound is $\sqrt{0.015/2} \approx 0.087$, an 8.7% worst-case shift. Because RAD's loss Eq. (2) is exactly ε_i, training *directly* tightens this guarantee.

Why this matters, and its limits. Proposition 3 says "zero divergence ⇒ ranking preserved"; Proposition 4 is its robust counterpart: retrieval degrades *gracefully*, as $\sqrt{\varepsilon_i}$, rather than abruptly as the code shortens. Minimizing Eq. (2) is therefore not a heuristic surrogate for ranking but a direct, deterministic, per-query guarantee. Pointwise MSE admits no analogous bound, because ℓ_2 embedding error is not monotone in any retrieval probability, the formal echo of the IQD anti-correlation (Proposition 1). The bound controls retrieval *probability*, the quantity recall integrates; it does not by itself pin the integer rank of the gold item (which also depends on the score-gap distribution), and it loosens for top-K at large K through a union bound. These are the honest limits of a distribution-level guarantee.

Table 4: Cross-backbone RAD and difficulty-weighting deltas (MSCOCO image, R@1). RAD = **(d)→(e)**; difficulty weighting = **(e)→(f)**. The large gain concentrates on Qwen3-VL; it is smaller but still positive ("small but visible") on VLM2Vec and CLIP (the latter on its soft-quantization arm); weighting helps only on Qwen3-VL.

Backbone	Family	RAD	weighting
Qwen3-VL-2B	decoder-VLM	+11.7	+2.1
VLM2Vec-LLaVA	decoder-VLM	+0.2–3.6	≈ 0
CLIP ViT-B/32	dual-encoder	+0.2–1.2	≈ 0

4.4 Online difficulty weighting (conditional)

When most samples are already aligned ($P_q \approx P_f$), their per-sample KL gradients are near zero yet still dilute the batch average, washing out the signal from the few collapsing samples. This motivates an *optional* reweighting that concentrates gradient where margins are fragile. Each sample is weighted by its online difficulty,

$$w_i \;=\; \mathrm{softmax}\Big(\frac{\mathcal{L}_{\mathrm{quant}}(i) - \mathcal{L}_{\mathrm{float}}(i)}{\tau_w}\Big) \cdot N, \tag{10}$$

and the distillation loss becomes

$$\mathcal{L}_{\mathrm{distill}} \;=\; \frac{1}{|B|}\sum_{b\in B}\frac{1}{N}\sum_{i=1}^{N} w_i\, \mathcal{L}^{(b)}_{\mathrm{distill}}(i). \tag{11}$$

The weights sum to N, preserving the loss scale, and reduce to uniform when $\Delta\mathcal{L}$ is homogeneous. As Sections 6.4 and 7 show, the benefit of this term is conditional, concentrated in-domain where difficulty is most heterogeneous, so we treat it as a switch that is enabled selectively rather than always-on. For compact figure legends, we denote RAD with this online difficulty weighting enabled as Online Difficulty-Aware Ranking Distillation (ODARD).

5 Cross-Backbone Validation

A natural worry is that ranking distillation helps only on the backbone we tuned. We test RAD on two further backbones spanning two families (Table 4): a second decoder-VLM, VLM2Vec (Jiang et al., 2024), and a dual-encoder, CLIP. The large gain concentrates on Qwen3-VL, where low-bit quantization (and MSE distillation) severely damages ranking; it shrinks but stays positive where that damage is smaller, and is present on all three backbones.

Three architectures. **Qwen3-VL** shows the large RAD gain (+11.7 R@1) and a difficulty-weighting bonus (+2.1): here MSE distillation actively hurts and KL repairs it. **VLM2Vec**, also a decoder-VLM, is largely quantization-robust after fine-tuning, so there is less ranking damage to repair; RAD still gives a modest gain (+0.2–3.6 R@1) while difficulty weighting is a no-op. **CLIP** gives a *small but visible* RAD gain on its soft-quantization arm (+0.2–1.2 R@1), smaller than the decoder-VLMs but nonzero: its dual-tower, contrastively-pretrained embeddings are already largely ranking-shaped, leaving the least damage to repair. The gain thus tracks how much recoverable ranking damage low-bit quantization inflicts on each backbone.

CLIP as a controlled boundary. A small gain on CLIP could mean two different things: little recoverable damage, or a training recipe that fails to transfer. A 2×2 control (hard STE vs. soft pseudo-quantization $\times$ disciplined vs. undisciplined schedule) separates the two: what governs stability on this dual-encoder is the quantizer's smoothness, not the schedule, so we report CLIP on its stable soft arm, where RAD retains its small positive gain (+0.2–1.2). On the same arm, difficulty weighting matches its inverted control ($f \approx fi$): the weighting no-op on CLIP is a measured absence of headroom, consistent with the recoverable-damage reading above, not an untested assertion. Full 2×2 tables are in Section H.

Table 5: Progressive ablation (R@1, dim=2048). Each row adds one variable. Shaded : a quantized column exceeds Float (quantization-as-regularization). †1-bit is eval-only post-hoc sign hashing, not a trained branch; it is reported for video and omitted on image, where sign-hash effects unrelated to the training objective dominate the comparison.

			Video ID				Image ID		
	Method	Var	Fl	1b†	2b	4b	Fl	2b	4b
(a)	Base+PTQ	–	36.5	40.5	35.2	36.3	27.8	28.7	27.9
(b)	Finetune+PTQ	FT	45.4	42.6	42.7	45.6	49.1	48.8	49.2
(c)	STE-QAT	QAT	46.0	42.4	41.9	46.6	51.1	53.0	51.3
(d)	+ MSE Distill	Dist	45.8	43.7	42.8	45.1	47.7	49.3	48.0
(e)	+ KL (RAD)	MSE→KL	**47.2**	43.3	**43.8**	**47.3**	59.4	57.8	60.0
(f)	+ Weighting	`online_diff`	46.8	**44.0**	43.4	46.4	**61.2**	**60.7**	**61.5**

6 Experiments

6.1 Setup

We fine-tune Qwen3-VL-Embedding-2B with LoRA (rank=32, $\alpha = 64$). Two independent arms share hyperparameters: *video* (MSR-VTT, 7,010 pairs) and *image* (MSCOCO, 113K pairs); quantized branches $\{2, 4, 8\}$-bit via STE, $\lambda_d = 0.5$, $\tau_d = 0.1$. Each arm is evaluated in-domain (MSR-VTT / MSCOCO text→image test, 1,000 queries) and zero-shot on MSVD, DiDeMo, and Flickr30k. We report Recall@k and NDCG@10 at float precision, sign-hash (1-bit), and scalar quantization (2/4/8-bit).

6.2 Ablation: each variable in isolation

MSE distillation hurts (**(c)**→**(d)**): −3.4 float on MSCOCO. **RAD dominates** (**(d)**→**(e)**): +11.7 R@1 on MSCOCO at float precision ($p < 10^{-4}$) and positive gains at 2/4-bit (Table 5); on MSR-VTT it gives +1.4 to +2.7 ($p \leq 0.024$), winning 42 of 45 video-arm evaluation settings (3 datasets × 5 dims × 3 quantizers, NDCG@10; Section D). **Weighting helps on image, neutral on MSR-VTT** (**(e)**→**(f)**): MSCOCO +1.7 to +2.1 (significant at float and 4-bit, $p \leq 0.004$), MSR-VTT −0.5 to −0.4 ($p \geq 0.19$). Bootstrap confidence checks are reported in Section D. In 10 of 12 rows a quantized column exceeds float, an implicit-regularization effect (AskariHemmat et al., 2022) strongest where the pretrained model is weaker.

6.3 Comparison to post-hoc and trained quantizers

Vector-quantization methods run on any frozen embedding, so we compare them to RAD at matched per-vector bit-budgets (Table 6). The *post-hoc* cohort applies an off-the-shelf compressor to the fine-tuned (**(b)**) embedding: sign hashing, per-dimension scalar quantization (SQ), the industrial standard PQ (Jégou et al., 2011), and binary ITQ (Gong & Lazebnik, 2011). The *trained-codebook* cohort learns the codebook with a retrieval objective rather than reconstruction: the Distill-VQ-style control (Xiao et al., 2022) distils the float embedding's similarity distribution into the codebook with a listwise KL, and the MoPQ-style control (Xiao et al., 2021) trains it with a contrastive matching loss; both build on the differentiable-PQ line of JPQ (Zhan et al., 2021) and RepCONC (Zhan et al., 2022).

Baseline adaptations. Distill-VQ and MoPQ were proposed for *text* retrieval; we transfer only the codebook objective to our multimodal embedding, and never reuse their text numbers. To keep every method on one fixed embedding we run them with the backbone frozen, in Distill-VQ's native fixed-document regime and MoPQ's optional frozen-encoder mode, learning the codebook on corpus neighbours (eval queries held out) with neighbour surrogates for query supervision; we call them MoPQ-style and Distill-VQ-style controls. Their full backbone-training variants live in the same regime as RAD, where the axis that matters is the objective, already isolated by the **(d)**→**(e)** ablation (+11.7, MSE→KL).

Result. RAD with difficulty weighting (**(f)**) attains the highest R@1 at both budgets (+10.4 over the strongest post-hoc method at 2-bit), exceeding every post-hoc and trained-codebook quantizer while retrieving via Hamming/dot products rather than PQ's asymmetric-distance tables. Sign hashing and ITQ are 1-bit-only, eval-only operating points, so the matched-budget comparison is made at 2 and 4 bits. One entry

Table 6: Post-hoc & trained-codebook quantizers vs. RAD (MSCOCO image R@1, 1,000-query test, matched bits). Post-hoc and trained-codebook rows quantize one fixed export of the fine-tuned (**(b)**) checkpoint [float 48.5]; Table 5 uses separate backend-specific encoding runs. RAD (**(e)**,**(f)**) are the trained QAT arms [float 59.4/61.2]. The MoPQ-style contrastive codebook re-embeds the candidates at 4-bit, lifting it above float (§6.3) but below RAD: a representation-relearning effect rather than quantization.

Method (per-vector bits)	2-bit	4-bit
Post-hoc		
Scalar SQ	50.3	48.8
PQ (Jégou et al., 2011)	47.6	48.8
Trained codebook		
Distill-VQ-style (Xiao et al., 2022)	46.7	48.5
MoPQ-style (Xiao et al., 2021)	39.3	60.9
Ours		
RAD **(e)**	57.8	60.0
+ weighting **(f)**	**60.7**	**61.5**

Table 7: Generalization: larger benchmark and new training arms (Qwen3-VL, R@1 averaged over 2/4/8-bit). RAD =**(d)**→**(e)**; weighting =**(e)**→**(f)**. Per-bit tables in Section E.

Dataset	Regime	RAD	weighting
MSCOCO-5K (Karpathy)	larger eval pool (image)	+4.7	+0.9
Flickr30k	extended training arm (image)	+2.8	+0.5
ActivityNet	extended training arm (video)	+0.2	+1.2

deserves a note: the MoPQ-style contrastive codebook lifts R@1 above float at 4-bit, indicating it *re-embeds* the candidates rather than purely quantizing them, yet it stays below RAD. Weight-quantization methods (MatQuant (Nair et al., 2025) and classification PTQ) operate on a different object, the network weights rather than the deployed embedding, and are discussed in Section 2 rather than run.

6.4 Generalization: larger pools, new domains, extended training

We test whether the trends survive three changes of regime: a larger evaluation pool (the image arm evaluated on the MSCOCO Karpathy 5K split), zero-shot domain transfer (the fixed arms evaluated on datasets never seen in training), and extended training coverage (two further arms trained on Flickr30k image and ActivityNet video). Flickr30k therefore appears in two distinct roles: as a zero-shot target of the MSCOCO arm (Table 8) and as a training source in its own right (Table 7).

RAD improves in all three regimes. On the larger 5K pool it gains +4.7 R@1 (Table 7). Zero-shot, its margin guarantee is scale-invariant, so the gain transfers without OOD tuning: over the base **(a)**, **(e)** gains +10.7 R@1 on MSR-VTT, +7.3 on MSVD, and +35.2 on Flickr30k (Table 8); training on Flickr30k directly lifts RAD further still (77.0 → 82.0 float R@1). The difficulty-weighting gain is *dataset-dependent*: positive on image and on ActivityNet video (+1.2), neutral on MSR-VTT, so the video-neutral observation is specific to MSR-VTT, not to video as a whole. On video **(e)**≈**(f)**; weighting adds a further +1.3 on near-OOD image at 2-bit. Full per-bit tables for all three additions are in Section E.

6.5 Directional controls

If the difficulty signal is real, corrupting it should degrade performance. Table 15 confirms a monotonic ordering **(f)**>**(e)**≈**(ft)**>**(fs)**>**(fi)**: inverting the weights (concentrating gradient on the wrong samples) is worst, and shuffling (which preserves the weight distribution but destroys the signal) lands between. On MSR-VTT all variants fall within ±0.8 R@1: the signal is directional on image but inert on MSR-VTT, consistent with the IQD-heterogeneity asymmetry of Section 5.

6.6 The storage–quality–speed frontier

Returning to the trade-off that opened the paper (Fig. 1): across the full storage and speed sweep (Fig. 4), RAD with weighting sits above the PTQ baseline at every operating point and matches the float baseline at 64× compression (dim=256, 4-bit).

Table 8: Zero-shot cross-dataset (R@1, dim=2048). Video arm trained on MSR-VTT; image arm on MSCOCO; OOD columns evaluated zero-shot.

	MSVD (vid, OOD)		Flickr (img, OOD)	
Method	Float	2-bit	Float	2-bit
(a)	62.7	61.3	41.8	43.5
(e)	70.0	**69.9**	77.0	75.1
(f)	**70.3**	**69.9**	**77.1**	**76.4**

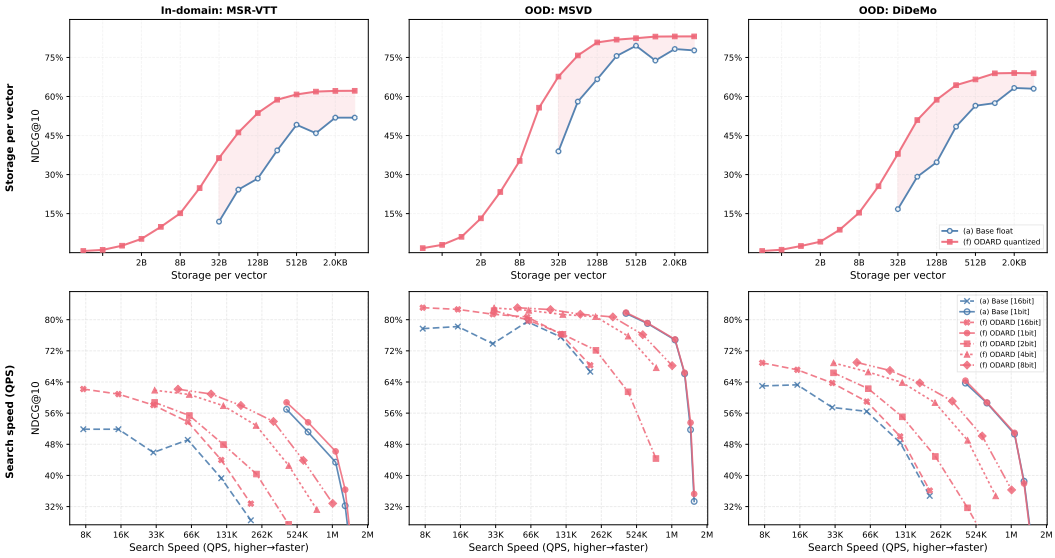

Figure 4: Pareto frontiers (storage and speed vs. NDCG@10). Sweeps sign 1-bit, SQ 2/4/8-bit × dims 128–2048. Blue: Base+PTQ **(a)**; red: RAD+weighting **(f)**.

7 Discussion and Conclusion

When does difficulty weighting help? When the evaluation domain is close to training and the measured IQD heterogeneity is high. On in-domain MSCOCO, weighting adds +1.7 to +2.1 R@1 (significant at float and 4-bit, $p \leq 0.004$; Section D), and it is likewise positive on in-domain ActivityNet video (+1.2, Section 6.4), so the switch tracks domain proximity rather than modality; on near-OOD Flickr30k the gain shrinks to +1.3; on far-OOD VisualNews it inverts: normal weighting hurts (−0.8 to −2.0) while inverted weighting recovers RAD or better (Section E). One reading is that `online_diff` mixes domain-invariant hardness with domain-specific embedding weakness. In-domain the two point the same way; far from the training domain the second component is noise. The practical recipe is therefore simple: deploy RAD universally, and enable weighting in-domain where the measured heterogeneity is high; default to RAD alone otherwise.

Outlook. The largest opportunity IQD opens is non-uniform precision. Our stratification shows that two-thirds of samples lose at most 4% recall at 2-bit while the remaining third loses over 24%, so a uniform-precision index spends most of its bits on samples that do not need them. Allocating precision by per-sample difficulty, with an offline router that approximates `online_diff` at index time, is a promising path to order-of-magnitude index savings.

Conclusion. Per-sample quantization difficulty varies by up to 21.3× in retrieval, no static float-space predictor detects it across regimes, and the online loss gap reads it out for free. MSE distillation allocates capacity blind; KL on similarity distributions allocates it by ranking relevance, wins 42 of 45 video-arm evaluation settings, and transfers across domains. Deploy it by default. The further gain from difficulty weighting is conditional, concentrated where difficulty is most heterogeneous and the deployment domain is close to training, so we expose it as a switch rather than a default. And when per-sample difficulty spans an order of magnitude, uniform-precision indices are wasting bits.

References

Faizan Ahemad. Quantization aware matryoshka adaptation: Leveraging matryoshka learning, quantization, and bitwise operations for reduced storage and improved retrieval speed. In *CIKM*, 2025.

MohammadHossein AskariHemmat, Reyhane Askari Hemmat, Alex Hoffman, Ivan Lazarevich, Ehsan Saboori, Olivier Mastropietro, Sudhakar Sah, Yvon Savaria, and Jean-Pierre David. QReg: On regularization effects of quantization. 2022. URL https://arxiv.org/abs/2206.12372.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation, 2013. URL https://arxiv.org/abs/1308.3432v1.

Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. HashNet: Deep Learning to Hash by Continuation. In *ICCV*, 2017. URL https://arxiv.org/abs/1702.00758v4.

David Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pp. 190–200, 2011. URL https://aclanthology.org/P11-1020/.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *NeurIPS*, 2024.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned Step Size Quantization, 2020. URL https://arxiv.org/abs/1902.08153. ICLR 2020.

Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating Large-Scale Inference with Anisotropic Vector Quantization, 2020. URL http://arxiv.org/abs/1908.10396.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pp. 5803–5812, 2017. URL https://arxiv.org/abs/1708.01641.

Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Cadyq: Content-aware dynamic quantization for image super-resolution. In *ECCV*, 2022.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2022. URL https://arxiv.org/abs/2106.09685. ICLR 2022.

Saqib Javed, Hieu Le, and Mathieu Salzmann. QT-DoG: Quantization-aware training for domain generalization. In *ICML*, 2025. doi: 10.48550/arXiv.2410.06020. URL https://arxiv.org/abs/2410.06020.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product Quantization for Nearest Neighbor Search. *IEEE TPAMI*, 33(1):117–128, 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.57. URL https://ieeexplore.ieee.org/document/5432202.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Hamidreza Kamkari, Brendan Leigh Ross, Rasa Hosseinzadeh, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A Geometric View of Data Complexity: Efficient Local Intrinsic Dimension Estimation with Diffusion Models, 2024. URL http://arxiv.org/abs/2406.03537. NeurIPS 2024 (spotlight).

Taehyeon Kim, Jaehoon Oh, Nakyil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. In *IJCAI*, 2021.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka Representation Learning. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2205.13147v4.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking, 2026. URL http://arxiv.org/abs/2601.04720.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://arxiv.org/abs/1405.0312.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, 2017. URL https://arxiv.org/abs/1708.02002v2.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *ICML*, 2022.

Pranav Ajit Nair, Puranjay Datta, Jeff Dean, Prateek Jain, and Aditya Kusupati. Matryoshka quantization. In *ICML*, pp. 45484–45506, 2025. URL https://proceedings.mlr.press/v267/nair25a.html.

NVIDIA Nemotron Team. Quantization-aware distillation for NVFP4 inference accuracy recovery. *arXiv preprint arXiv:2601.20088*, 2026.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017. doi: 10.1007/s11263-016-0965-7.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://arxiv.org/abs/2103.00020.

Jinpeng Wang, Bin Chen, Dongliang Liao, Ziyun Zeng, Gongfu Li, Shu-Tao Xia, and Jin Xu. Hybrid contrastive quantization for efficient cross-view video retrieval. In *The Web Conference (WWW)*, 2022.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Chen, Xinggang Wang, Hongyang Chao, and Han Hu. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *ICCV*, 2023.

Shitao Xiao, Zheng Liu, Yingxia Shao, Defu Lian, and Xing Xie. Matching-oriented embedding quantization for ad-hoc retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8119–8129, 2021. doi: 10.18653/v1/2021.emnlp-main.640. URL https://aclanthology.org/2021.emnlp-main.640/.

Shitao Xiao, Zheng Liu, Weihao Han, Jianjin Zhang, Defu Lian, Yeyun Gong, Qi Chen, Fan Yang, Hao Sun, Yingxia Shao, and Xing Xie. Distill-VQ: Learning retrieval oriented vector quantization by distilling knowledge from dense embeddings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1513–1523, 2022. doi: 10.1145/3477495.3531799. URL https://doi.org/10.1145/3477495.3531799.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pp. 5288–5296, 2016. URL `https://openaccess.thecvf.com/content_cvpr_2016/papers/Xu_MSR-VTT_A_Large_CVPR_2016_paper.pdf`.

Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. Efficient passage retrieval with hashing for open-domain question answering. In *ACL*, 2021.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *CVPR*, 2024.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Jointly optimizing query encoder and product quantization to improve retrieval performance. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, 2021.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Learning discrete representations via constrained clustering for effective and efficient dense retrieval. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM)*, 2022.

Kaiqi Zhao and Ming Zhao. Self-supervised quantization-aware knowledge distillation. In *AISTATS*, 2024.

Appendix

The appendix is organized as follows.

Section	Contents
Section A	Overview and evidence hierarchy
Section B	Proofs and the MSE $\neq$ ranking counterexample
Section C	IQD protocol and full predictor tables
Section D	Full ablation and bootstrap confidence details
Section E	Additional datasets and OOD protocols
Section F	Cross-backbone results
Section G	Evaluation methodology, speed, and calibration
Section H	Training and implementation details

A Overview and Evidence Hierarchy

This appendix follows an explicit evidence hierarchy. Qwen3-VL is the primary in-depth backbone: it carries the full IQD analysis, complete ablations, and the main quantitative claim for RAD. CLIP and VLM2Vec are included as minimal cross-backbone validation to test transferability boundaries, not to define new primary claims. This keeps the contribution centered on the pipeline IQD $\rightarrow$ RAD, with difficulty weighting as a conditional extension, while still reporting where the signal weakens under other encoder families.

The Tier 1 predictors of the main paper are static proxies computed from float embeddings alone, without using quantized similarities; proxies that require the quantized branch are online signals, in the same category as `online_diff`. Across the six Tier 1 families tested, none is regime-robust, which motivates online difficulty discovery rather than offline proxy selection.

Table 9: Float-only (Tier 1) predictors tested.

Predictor	Family	What it measures
`margin_float`	Geometric	Score gap between the matched positive and hardest negative. Larger margin implies safer ranking.
`inv_margin`	Geometric	Reciprocal of margin. Larger value implies ranking is easier to flip under quantization noise.
`density`	Density	Mean cosine similarity to k-nearest neighbors. Crowded neighborhoods are more collision-prone after quantization.
`LID`	Density	Local intrinsic dimensionality: growth rate of neighborhood volume with radius.
`quant_error`	Reconstruction	Per-sample ℓ_2 distortion after quantization.
`iqd_combined`	Hybrid	Product `quant_error` $\times$ `density`.

Relation to retrieval-oriented quantization. Prior retrieval-oriented quantizers replace reconstruction with relevance-aware objectives for codebook learning. Our contribution is complementary: we show why ranking structure should be distilled at the similarity-distribution level and why sample difficulty must be discovered online rather than predicted from float-space geometry alone. These ideas are orthogonal to the specific deployed quantizer (sign, SQ, or PQ).

B Theory Details

B.1 Zero-KL Margin Equivalence

Recall Proposition 3: when the KL divergence between the float retrieval distribution and the quantized retrieval distribution vanishes, all pairwise margins are preserved.

Proof. Let P_f and P_q be the teacher (float) and student (quantized) similarity distributions over N candidates, defined via softmax with temperature τ_d:

$$P_f(d_j) = \frac{\exp(s_f^j/\tau_d)}{\sum_{k=1}^{N}\exp(s_f^k/\tau_d)}, \qquad P_q(d_j) = \frac{\exp(s_q^j/\tau_d)}{\sum_{k=1}^{N}\exp(s_q^k/\tau_d)}. \tag{12}$$

If $\mathrm{KL}(P_f\|P_q) = 0$, then $P_f(d_j) = P_q(d_j)$ for all candidates j. Taking the ratio of any two entries (i, j) gives

$$\frac{P_q(d_i)}{P_q(d_j)} = \frac{P_f(d_i)}{P_f(d_j)}. \tag{13}$$

Substituting the softmax definition cancels the partition functions:

$$\frac{\exp(s_q^i/\tau_d)}{\exp(s_q^j/\tau_d)} = \frac{\exp(s_f^i/\tau_d)}{\exp(s_f^j/\tau_d)}. \tag{14}$$

Taking logarithms yields

$$s_q^i - s_q^j = s_f^i - s_f^j \quad \Longrightarrow \quad s_q^i - s_f^i = s_q^j - s_f^j \quad \forall i, j. \tag{15}$$

Thus there exists a constant C such that

$$\mathbf{s}_q = \mathbf{s}_f + C\mathbf{1}. \tag{16}$$

For any positive-negative pair (d^+, d^-),

$$\Delta_q = s_q^+ - s_q^- = (s_f^+ + C) - (s_f^- + C) = \Delta_f. \tag{17}$$

Since ranks depend only on pairwise margin signs, the ranking is preserved. □

Remark 1 (Why score KL is less over-constrained than pointwise MSE)**.** *Softmax KL depends on score differences, not absolute score values. It therefore permits the global translation $\mathbf{s}_q = \mathbf{s}_f + C\mathbf{1}$, the minimum sufficient condition for preserving pairwise margins. MSE on embeddings or scores asks the low-bit student to match pointwise geometry, which is stricter than ranking preservation and can allocate capacity to samples whose retrieval ordering is already safe.*

The finite-KL counterpart of this statement, a Pinsker bound on retrieval-probability drift, is developed in full in the main paper (Section 4.3).

B.2 MSE $\neq$ Ranking: Constructive Counterexample

We use a synthetic construction: candidates are placed in $\mathbb{R}^6$ so that the ranking margin concentrates in dimensions 5–6. Uniformly minimizing reconstruction error can allocate quantization capacity to high-energy but rank-irrelevant coordinates, corrupting the low-energy margin-bearing coordinates. The result is a lower embedding MSE but more rank inversions.

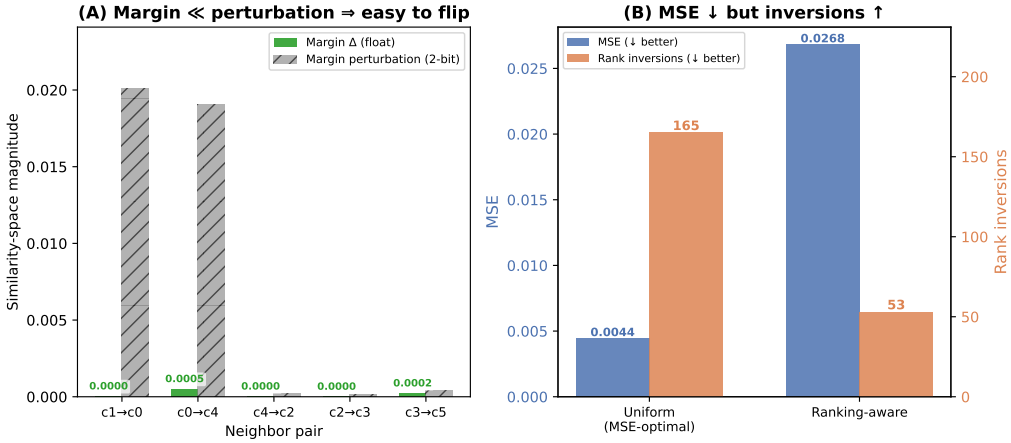

Figure 5: Counterexample for MSE $\neq$ ranking. MSE-optimal quantization can have lower reconstruction error but more rank inversions when the rank-relevant dimensions are not the high-variance dimensions.

C IQD Measurement Protocol and Full Predictor Tables

Protocol. For each query i, we compute the float rank of the matched candidate and the rank after quantization. The per-sample degradation is

$$\texttt{rank_drop}_i = \text{rank}_{\text{quant}}(i) - \text{rank}_{\text{float}}(i). \tag{18}$$

We evaluate six float-only predictors and three quantization-aware predictors by Spearman correlation with `rank_drop`. For video datasets, query and candidate orders are one-to-one paired. For MSCOCO-style gallery retrieval, we construct the correct candidate from the stored `label_name`; this avoids the diagonal-pairing artifact.

Cell composition. The main paper's 70 configurations are the heatmap cells of Fig. 3, the full backbone × modality matrix: 20 Qwen/MSR-VTT, 15 CLIP/MSR-VTT, 20 Qwen/MSCOCO ($d \in \{128, 256, 512, 1024\}$), and 15 CLIP/MSCOCO ($d \in \{128, 256, 512\}$) (dim, bit) cells. The Qwen/MSCOCO full-heatmap figure below comes from a complementary GT-correct export whose sweep includes the native d=2048 in place of d=1024; the two exports agree cell-for-cell on their shared dimensions, and the d=2048 row is where `margin_float` peaks (ρ up to 0.92), as quoted in Section 3.3. As disclosed in Section 3.1, the CLIP video cells score frame-level embeddings without temporal mean-pooling and serve as internal comparisons across quantization settings. The additional predictor tables in this appendix focus on the regimes with the strictest ground-truth construction: the paired video protocol (Qwen/MSR-VTT) and the GT-correct gallery protocol (Qwen/MSCOCO, CLIP/MSCOCO).

C.1 Predictor Comparison Across GT-Correct Regimes

Table 10: Predictor comparison (mean $|\rho|$), all four regimes. The video columns report the paired-protocol values of the main paper's Table 3. Image columns are GT-correct gallery-retrieval values (`label_name` matching, non-diagonal). Note that the image panels of Fig. 3 (B, D) report the *winning* predictor *per cell*, while this table reports each predictor's *mean* across all tested image cells, so values are not expected to match the cell labels. Video means are over all cells including the 16-bit control; image sweeps are Qwen ($d \in \{128, 256, 512, 1024, 2048\}, b \in \{1, 2, 4, 8\}$) and CLIP ($d \in \{128, 256, 512\}, b \in \{1, 2, 4, 8\}$).

	Predictor	Qwen/Vid	CLIP/Vid	Qwen/Img	CLIP/Img
Tier 2	`online_diff`	**0.547**	**0.668**	**0.620**	**0.719**
	`margin_erosion`	0.495	0.579	0.575	0.678
	`focal_weight`	0.256	0.389	0.294	0.568
Tier 1	`margin_float`	0.118	0.080	0.454	0.106
	`inv_margin`	0.097	0.088	0.346	0.117
	`quant_error`	0.075	0.043	0.126	0.052
	`density`	0.054	0.078	0.222	0.112
	`iqd_combined`	0.053	0.083	0.200	0.095
	`LID`	0.041	0.022	0.040	0.105

C.2 Full ρ Heatmaps

The main paper reports the best predictor per setting; the full heatmaps below show every predictor's surface for the protocol-strict regimes.

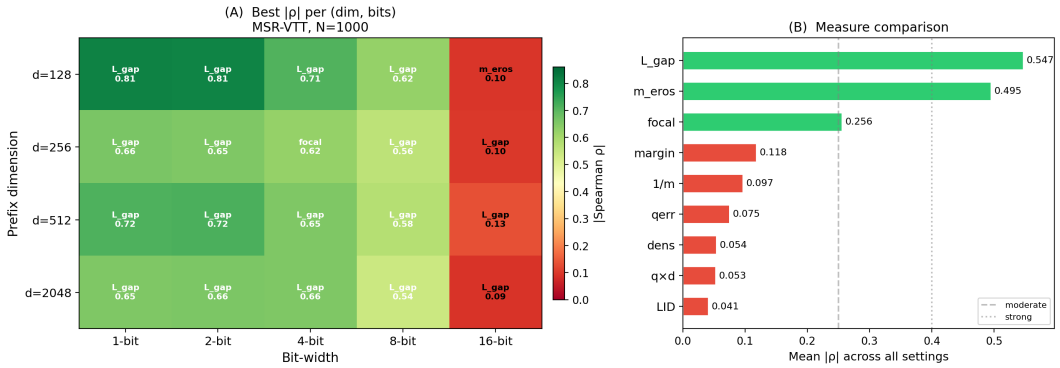

Figure 6: Qwen3-VL on MSR-VTT (paired protocol).

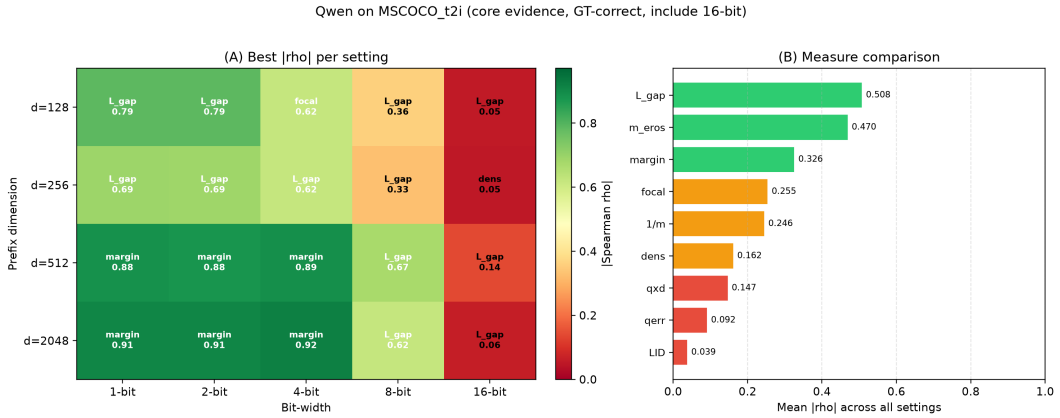

Figure 7: Qwen3-VL on MSCOCO (GT-correct `label_name` protocol). This export sweeps $d \in \{128, 256, 512, 2048\}$; it agrees cell-for-cell with the main paper's Fig. 3C on the shared dimensions, and adds the native d=2048 row where `margin_float` peaks ($\rho = 0.91$–0.92 at $\leq$4-bit).

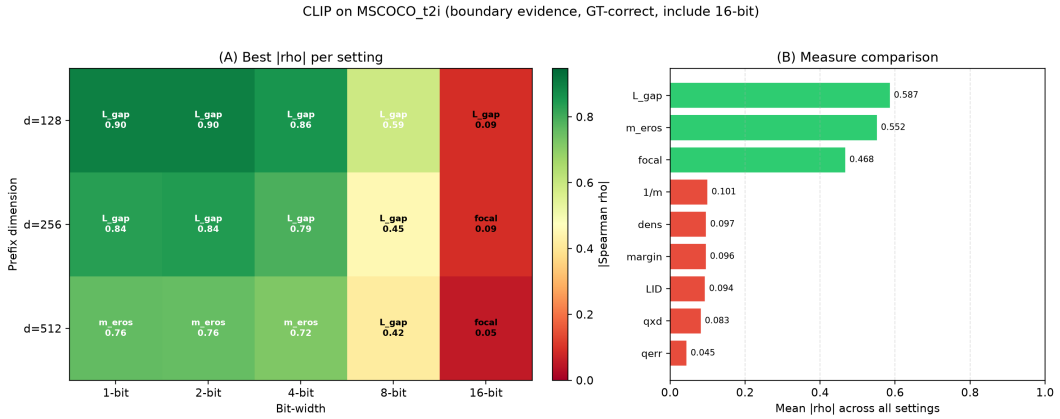

Figure 8: CLIP ViT-B/32 on MSCOCO (GT-correct `label_name` protocol).

C.3 MSCOCO IQD Numeric Table

To keep the table comparable across many operating points, we report condensed mean/min/max $|\rho|$ over all tested (dim, bit) cells under the GT-correct protocol.

Table 11: MSCOCO IQD predictor correlations (GT-correct). Values are computed against the label-name ground truth rather than the diagonal ordering, over the image sweeps stated in Section C.1 (Qwen: 20 settings; CLIP: 12 settings). The `margin_float` maximum (0.922) is the native-dimension d=2048, 4-bit cell.

Backbone	Predictor	mean $\|\rho\|$	min $\|\rho\|$	max $\|\rho\|$
Qwen	`online_diff`	0.620	0.329	0.788
	`margin_erosion`	0.575	0.237	0.724
	`margin_float`	0.454	0.010	0.922
	`quant_error`	0.126	0.004	0.279
	`density`	0.222	0.001	0.569
	`LID`	0.040	0.004	0.115
CLIP	`online_diff`	0.719	0.417	0.897
	`margin_erosion`	0.678	0.349	0.814
	`margin_float`	0.106	0.035	0.176
	`quant_error`	0.052	0.003	0.095
	`density`	0.112	0.011	0.187
	`LID`	0.105	0.007	0.200

D Full Ablation and Confidence Details

Claim scope for this section. The core contribution is the path IQD → RAD, with difficulty weighting treated as a conditional enhancement rather than a universal add-on. Accordingly, the quantitative counts in this section are anchored to the Qwen main ablation grid used by the main IQD analysis.

D.1 42-of-45 Breakdown

The "45 settings" refer to the Qwen video-arm evaluation grid (3 datasets × 5 dimensions × 3 quantization points). KL ranking distillation **(e)** beats MSE distillation **(d)** in 42 of 45 evaluation settings by NDCG@10. The three exceptions are all within noise:

Table 12: Settings where RAD **(e)** does not exceed MSE **(d)**.

Dataset	Dim	Quant	**(d)**	**(e)**
DiDeMo	1024	8-bit	67.0	66.8
DiDeMo	1024	float	67.0	66.9
MSVD	2048	2-bit	81.9	81.8

D.2 Bootstrap Confidence Intervals

All confidence intervals use 10,000 bootstrap resamples with seed 42. Per-sample hit/miss vectors are resampled with replacement; p-values are two-tailed. Quantized rows are listed first because they are the target operating points; float rows are a quantization-free control.

Table 13: Bootstrap confidence intervals for replacing MSE with ranking KL, **(d)**→**(e)**.

Dataset	Quant	**(d)** MSE	**(e)** RAD	Diff	95% CI / p
MSCOCO t2i	2-bit	48.4	58.8	+10.4	[8.0, 12.8], $p < 10^{-4}$
MSCOCO t2i	4-bit	48.0	59.5	+11.5	[9.2, 13.8], $p < 10^{-4}$
MSCOCO t2i	float	47.7	59.4	+11.7	[9.5, 13.9], $p < 10^{-4}$
MSR-VTT	2-bit	42.1	44.8	+2.7	[1.0, 4.5], $p = 0.0009$
MSR-VTT	float	45.8	47.2	+1.4	[0.1, 2.8], $p = 0.024$
MSVD	float	69.3	70.0	+0.7	[-0.4, 1.9], $p = 0.14$
DiDeMo	float	53.0	54.7	+1.7	[0.4, 3.0], $p = 0.006$

Table 14: Bootstrap confidence intervals for the optional difficulty-weighting extension, **(e)**→**(f)**.

Dataset	Quant	**(e)** RAD	**(f)** weighting	Diff	95% CI / p
MSCOCO t2i	2-bit	58.8	60.5	+1.7	[-0.1, 3.4], $p = 0.075$
MSCOCO t2i	4-bit	59.5	61.6	+2.1	[0.7, 3.5], $p = 0.002$
MSCOCO t2i	float	59.4	61.2	+1.8	[0.5, 3.1], $p = 0.004$
MSR-VTT	2-bit	44.8	44.3	-0.5	[-1.8, 0.8], $p = 0.24$
MSR-VTT	float	47.2	46.8	-0.4	[-1.2, 0.4], $p = 0.19$

The weighting term is therefore a conditional extension: it is significant on MSCOCO, neutral on MSR-VTT, and not a universal add-on.

Table 15: Directional controls (MSCOCO, dim=2048). Corrupting the weighting signal degrades performance monotonically. **(ft)**: softened τ_w; **(fs)**: shuffled; **(fi)**: inverted.

	Weighting signal	Float	2-bit	4-bit
(f)	`online_diff`	**61.2**	**60.7**	**61.5**
(e)	none (KL only)	59.4	57.8	60.0
(ft)	softened (τ_w=3)	59.4	57.7	59.6
(fs)	shuffled	58.1	57.9	58.7
(fi)	inverted	57.6	57.3	58.2

E Additional Datasets and OOD Protocols

E.1 MSCOCO Karpathy 5K

The MSCOCO-5K table supports the larger-benchmark claim in the main paper. Directional-control arms (**(fs)** shuffled, **(fi)** inverted) are included for completeness.

Table 16: MSCOCO-5K text-to-image retrieval (Qwen3-VL image arm, R@1).

Arm	Float	2-bit	4-bit	8-bit
(a)	5.4	8.6	5.5	5.4
(b)	31.7	34.2	31.8	31.7
(c)	36.2	36.5	36.3	36.2
(d)	33.8	35.7	33.8	33.8
(e)	39.6	38.6	39.3	39.5
(f)	**40.4**	**39.1**	**40.5**	**40.6**
(fs)	39.4	38.2	38.9	39.4
(fi)	38.0	37.8	37.8	38.1

E.2 Flickr30k, VisualNews, ActivityNet

The image arm is trained on MSCOCO and evaluated zero-shot on Flickr30k text-to-image retrieval. The far-OOD VisualNews task uses news photographs and formal captions, unlike the everyday scenes in MSCOCO and Flickr30k.

Table 17: Flickr30k zero-shot image retrieval (R@1 / NDCG@10, dim=2048).

	Float		4-bit		2-bit	
Method	R@1	N@10	R@1	N@10	R@1	N@10
(a) Base+PTQ	41.8	56.7	41.5	56.4	43.5	58.7
(d) +MSE	74.4	85.1	74.8	85.0	72.9	84.1
(e) RAD	77.0	86.7	77.1	86.7	75.1	85.4
(f) weighting	**77.1**	**87.0**	**77.2**	**87.1**	**76.4**	**86.0**
Δ **(e)**−**(d)**	+2.6	+1.6	+2.3	+1.7	+2.2	+1.3
Δ **(f)**−**(e)**	+0.1	+0.3	+0.1	+0.4	+1.3	+0.6

Table 18: VisualNews far-OOD image retrieval (R@1, dim=2048). Difficulty weighting hurts relative to RAD on this far-OOD domain, while inverted weighting recovers or exceeds RAD in some low-bit settings.

Method	Float	4-bit	2-bit
(a) Base+PTQ	32.8	33.1	32.8
(e) RAD	37.2	**38.0**	36.2
(f) weighting	36.0	36.0	35.4
(fi) Inverted	**37.9**	**38.0**	**38.0**
Δ **(f)**−**(e)**	-1.2	-2.0	-0.8
Δ **(fi)**−**(e)**	+0.7	+0.0	+1.8

E.3 Extended Training Arms: Flickr30k and ActivityNet

The two extended training arms behind Table 7. Both tables report R@1 at dim=2048.

Table 19: Flickr30k *training* arm, in-domain evaluation (R@1). Distinct from the zero-shot Flickr30k evaluation of the MSCOCO arm above: here the arm is trained on Flickr30k.

Arm	Float	2-bit	4-bit	8-bit
(a)	41.8	43.5	41.5	41.8
(b)	77.4	76.0	77.2	77.3
(c)	78.8	78.6	79.5	78.7
(d)	78.7	78.2	78.3	78.6
(e)	82.0	80.1	81.4	82.0
(f)	**82.2**	**80.5**	**82.2**	**82.2**

Table 20: ActivityNet *training* arm, in-domain evaluation (R@1).

Arm	Float	2-bit	4-bit	8-bit
(a)	18.9	26.7	18.5	18.8
(b)	47.3	45.3	47.1	47.2
(c)	48.9	46.1	49.0	48.8
(d)	48.7	46.1	48.4	48.7
(e)	49.9	45.8	49.2	48.7
(f)	**50.1**	**47.2**	**49.8**	**50.2**

The extended training arms are boundary analyses; they are not included in the core Qwen-centered claim counts.

F Cross-Backbone Results

This section is intentionally scoped as minimal validation. Qwen3-VL remains the full-reference backbone for method conclusions, while VLM2Vec and CLIP are used to test whether the IQD → RAD logic transfers beyond one model family and to document where difficulty weighting becomes weak or neutral. We do not claim that the CLIP or VLM2Vec settings here are architecture-optimal; they are reported as boundary probes under fixed, lightly tuned recipes.

Table 21: Cross-backbone summary (MSCOCO image, R@1 deltas).

Backbone	Family	RAD	weighting
Qwen3-VL-2B	decoder-VLM	+11.7	+2.1
VLM2Vec-LLaVA	decoder-VLM	+0.2–+3.6	≈ 0
CLIP ViT-B/32	dual-encoder	+0.2–+1.2	≈ 0

Table 22: Cross-backbone RAD arms (MSCOCO t2i, R@1). Per-bit RAD (**(d)**→**(e)**): VLM2Vec Float +0.2, 2b +3.6, 4b +0.4, 8b +1.1; CLIP (soft-quant.) Float +0.2, 2b +1.1, 4b +1.2, 8b +0.2. Difficulty weighting (**(e)**→**(f)**) ≈ 0 on both backbones.

Backbone	Arm	Float	2-bit	4-bit	8-bit
VLM2Vec-LLaVA	**(d)**	57.7	53.3	57.4	56.8
	(e)	**57.9**	**56.9**	**57.8**	**57.9**
CLIP ViT-B/32	**(d)**	62.1	52.8	60.9	62.2
	(e)	**62.3**	**53.9**	**62.1**	**62.4**

Per-bit absolute values for the RAD transition (**(d)**, **(e)**) are above; Section H gives the CLIP hard/soft 2×2 diagnostic.

G Evaluation Methodology

G.1 Dataset Scope

We separate core evaluation datasets from later training-data extensions. Dataset roles, evaluation splits, and ground-truth construction are specified below.

Table 23: Dataset scope (core evaluation and extensions).

Dataset	Modality	Role	Use
MSR-VTT	Video	In-domain	Core Qwen evidence set (IQD analysis and ablations), 1K-query paired retrieval.
MSCOCO t2i	Image	In-domain	Core Qwen evidence set (IQD and image-arm ablations), 1K-query gallery retrieval.
MSVD	Video	OOD	Evaluation-only video OOD check.
DiDeMo	Video	OOD	Evaluation-only video OOD check.
Flickr30k t2i	Image	OOD	Evaluation-only image OOD check (Karpathy 1K test protocol).
VisualNews t2i	Image	far-OOD	Evaluation-only far-OOD image check.
COCO Karpathy 5K	Image	Extra eval split	Additional image evaluation split (Section E).
Flickr30k train split	Image	Train-time extension	Training-data extension; not part of the core 45-setting claim set.
ActivityNet Captions	Video	Train-time extension	Training-time extension used for boundary analysis rather than a primary benchmark table.

G.2 Task Regimes and Ground Truth Construction

We evaluate two retrieval regimes with different ground-truth semantics:

- **Paired video retrieval** (MSR-VTT, MSVD, DiDeMo): query and candidate orders are one-to-one aligned; query i matches candidate i.
- **Gallery image retrieval** (MSCOCO_t2i and image OOD sets): candidates form a global deduplicated pool; query-correct indices are built from stored `label_name` metadata rather than diagonal position.

All MSCOCO IQD analyses use the GT-correct gallery protocol, not the diagonal approximation.

G.3 Metrics and Reporting Rules

The primary ranking metric is R@1; R@5, R@10, and NDCG@10 are reported as secondary metrics for stability checks. For each method and dataset we report float retrieval quality, quantized quality under sign (1-bit) and SQ (2/4/8-bit), and deltas for the key transitions (**(d)**→**(e)**, **(e)**→**(f)**). For predictor tables we aggregate mean $|\rho|$ across all tested settings within the stated regime (video paired or gallery GT-correct).

G.4 Quantization Operating Points

Unless otherwise stated, quantized evaluation sweeps backends sign (1-bit) and SQ (2/4/8-bit); dimensions 128, 256, 512, 1024, 2048 (clamped by backbone max); and a storage axis of bits per vector = dimension × bits per dimension. The same backend definitions are used in both ablation tables and Pareto plots to avoid protocol drift between accuracy and efficiency sections.

G.5 Storage-Speed Pareto Protocol

The storage and speed Pareto frontiers sweep quantization backend (sign 1-bit, SQ 2/4/8-bit) and embedding dimension (128–2048). Storage is computed as dimension times bits per dimension. Speed is measured wall-clock retrieval time on identical hardware, not estimated analytically.

Table 24: Search backends used in speed benchmarking.

Representation	Search backend
Float	FAISS `IndexFlatIP`
Sign 1-bit	FAISS `IndexBinaryFlat`
SQ 2/4/8-bit	FAISS `IndexScalarQuantizer`

Speed benchmarks use the same FAISS backend family across float/sign/SQ with a fixed benchmark protocol (`corpus-size`=1000, `num-queries`=1000, `repeats`=5). The headline points in the main paper correspond to MSR-VTT sign@1024 versus float@2048 (about 83×) and MSVD sq4@128 versus float@2048 (about 57×), measured with identical search code paths.

G.6 Statistical Confidence

Confidence intervals use bootstrap resampling (10,000 draws, seed 42) on per-query outcomes; two-tailed p-values are reported alongside point estimates. Full CI tables are in Section D; this section defines the shared evaluation protocol under which those intervals are interpreted.

G.7 SQ Calibration Sensitivity

We use joint calibration in the main SQ evaluation: per-channel `min`,`max` are computed over query and candidate embeddings ($Q \cup C$) before uniform scalar quantization. Since deployment would normally calibrate the index on candidates only, we rerun the same float embeddings with a corpus-only calibration (C only). The sensitivity table below reports Δ = corpus-only minus joint.

Table 25: SQ calibration sensitivity; Δ = corpus-only − joint.

Arm	Dataset	Bits	ΔR@1	ΔR@5	ΔNDCG@10
(e)	MSR-VTT	2	+2.0	+0.9	+1.0
(e)	MSR-VTT	4	-0.8	+0.0	-0.2
(e)	MSCOCO t2i	2	-1.4	-0.5	-0.9
(e)	MSCOCO t2i	4	-0.1	-0.3	-0.0
(f)	MSR-VTT	2	+1.3	+1.6	+0.8
(f)	MSR-VTT	4	-0.2	+0.9	-0.0
(f)	MSCOCO t2i	2	-0.3	-0.8	-0.5
(f)	MSCOCO t2i	4	+0.2	+0.5	+0.1

Across all 32 metric comparisons, the maximum absolute delta was 2.0 points and the mean was 0.56 points. The largest shift favored corpus-only rather than joint calibration. Deltas are mixed in sign, so we find no systematic inflation from including query embeddings in calibration. All methods share the same quantization pipeline, so residual calibration effects cancel in relative comparisons.

H Training and Implementation Details

Table 26: Training and quantization settings for the Qwen primary evidence chain. Arms **(b)**, **(c)**, **(d)**, **(e)**, **(f)** share model, schedule, and LoRA configuration; only the loss terms differ.

Model	
Base model	Qwen3-VL-Embedding-2B
LoRA	$r = 32,\ \alpha = 64$, attention (+ optional MLP variants)
*Shared training, arms **(b)**–**(f)***	
LR schedule	Cosine decay, linear warmup
Peak LR	10^{-5} backbone, 10^{-4} LoRA/head
Effective batch	128
Training	3,001 steps
Precision	bf16
Contrastive τ	0.07
*Distillation, arms **(d)**, **(e)**, **(f)***	
Distill τ_d	0.1
λ_d	0.5
Distill warmup	Linear ramp
*Difficulty weighting, arm **(f)** only*	
w_i computation	Softmax over $\Delta\mathcal{L}_i$ (Eq. (10))
τ_w	1.0
Quantization	
STE-QAT	Uniform symmetric, {2,4,8}-bit
1-bit	Post-hoc sign hash at evaluation
SQ statistics	Per-channel, computed on test set

CLIP note (boundary-only). For CLIP-family experiments, we treat hard/soft quantization and schedule variants as boundary diagnostics rather than as a second primary evidence chain. The 2×2 control in the main paper should therefore be read conservatively: it tests whether the CLIP observations are dominated by an unstable transferred recipe, not as a universal claim that soft pseudo-quantization is always preferable to hard STE. In our runs, hard STE is brittle for the encoder-style CLIP setting, while soft pseudo-quantization gives the most stable reported CLIP boundary arm; this is also consistent with broader practice in CLIP-style budget-aware retrieval training.[2] Nevertheless, the stabilized CLIP arms remain narrow-band and the difficulty-weighted arm does not reliably separate from shuffled and inverted controls. CLIP thus serves as a low-headroom, recipe-sensitive boundary case, consistent with its low measured IQD heterogeneity, rather than as a second optimization target comparable to the Qwen3-VL sweep. Method conclusions are anchored on the Qwen full sweep, with CLIP used only to test cross-backbone scope.

Table 27: Compact CLIP hard/soft $\times$ schedule diagnostic.

Quantizer	Schedule	Summary
Hard STE	transferred	QAT arms brittle / branch ordering diverges
Hard STE	disciplined	schedule helps little; spread remains narrow
Soft pseudo-quant.	aggressive	smoother but scale-sensitive
Soft pseudo-quant.	disciplined	stabilized narrow band; $f \approx f_s, f_i$

VLM2Vec note (sensitivity-only). VLM2Vec is retrieval-pretrained and differs from Qwen3-VL in initialization and stability profile. Our sensitivity runs indicate that directly transferring a large-batch, low-temperature pure-contrastive recipe into multi-bit QAT can cause branch divergence. A lightly tuned recipe recovers stable behavior; attention-only LoRA is treated as diagnostic, not as a universal best practice.

[2] `https://www.marqo.ai/blog/matryoshka-representation-learning-with-clip-for-multimodal-retrieval-and-ranking`

Accordingly, VLM2Vec is included as a recipe-sensitivity boundary check rather than a second primary optimization target.