EVIE: EVIDENCE-ALIGNED VISUAL DOCUMENT RETRIEVAL WITH COMPACT MULTI-VECTOR INDEXES

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual document retrieval depends on local page evidence, yet both learning and indexing can erase the distinctions that late interaction is meant to preserve. A visually similar page mined as a negative may contain the answer, while narrowing or pooling page tokens can change which region wins a query token's MaxSim operation. We introduce **EVIE**, a family of visual document retrievers that coordinates supervision, representation width, and stored token count around this common matching interface. A multimodal judge grades mined pages by answer completeness, expanding full-evidence positives and masking unreliable negatives. A bidirectional teacher transfers candidate rankings to six nested student widths, applying retrieval and listwise objectives to every prefix. Hierarchical agglomerative compression (HAC) then uses semantic–spatial features for cluster assignment but stores only semantic centroids, preserving direct MaxSim retrieval. Across 138 tasks from four evaluation suites, EVIE-8B reaches 66.75 nDCG@10 on public ViDoRe V3 and a four-suite average of 79.51. Compressed EVIE-4.5B exceeds WeMM-9B by 4.64 V3 points at the same 7.63 GiB vector payload per million pages. At 3.81 GiB, it retains 59.58 while using $128\times$ less payload than a configured 1024×256 two-byte image-token representation. Controlled experiments attribute these trade-offs to evidence supervision, prefix-wise ranking transfer, and position-guided compression.

1 INTRODUCTION

A page can mention the right entities without containing the evidence needed to answer a query. Retrieving a margin trend from a financial report, for example, may require binding a legend to a plotted series and comparing values across quarters. A text serialization can retain region names and axis labels while omitting the visual relation that determines the answer. Native visual document retrieval instead represents the page image directly, preserving information carried by tables, charts, and layout (Ma et al., 2024; Faysse et al., 2025).

The granularity of the representation determines both what can be matched and what must be stored. A single vector provides a compact page summary. Late-interaction models retain multiple token vectors and let each query token select its best page match through MaxSim (Khattab & Zaharia, 2020; Faysse et al., 2025), exposing local evidence that a global summary may obscure. This interface also couples three decisions that are often treated separately. The training labels determine which page-level distinctions token vectors must learn. The embedding width determines the geometry of each local comparison. The number of stored page vectors determines which candidate regions remain available to the maximum. Errors at any of these points can remove an answer-bearing match.

The resulting learning and storage problem is concrete. Mined hard negatives are selected for similarity, not verified absence of an answer, so they can contain complete or partial evidence. At deployment, a million pages represented by 1024 image vectors with 256 two-byte components require 488.28 GiB of vector payload before index overhead. Narrowing the vectors changes local similarities; pooling them changes which page region wins each maximum. We therefore ask how to learn the right page distinctions and preserve their useful local matches as both vector width and token count decrease.

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.

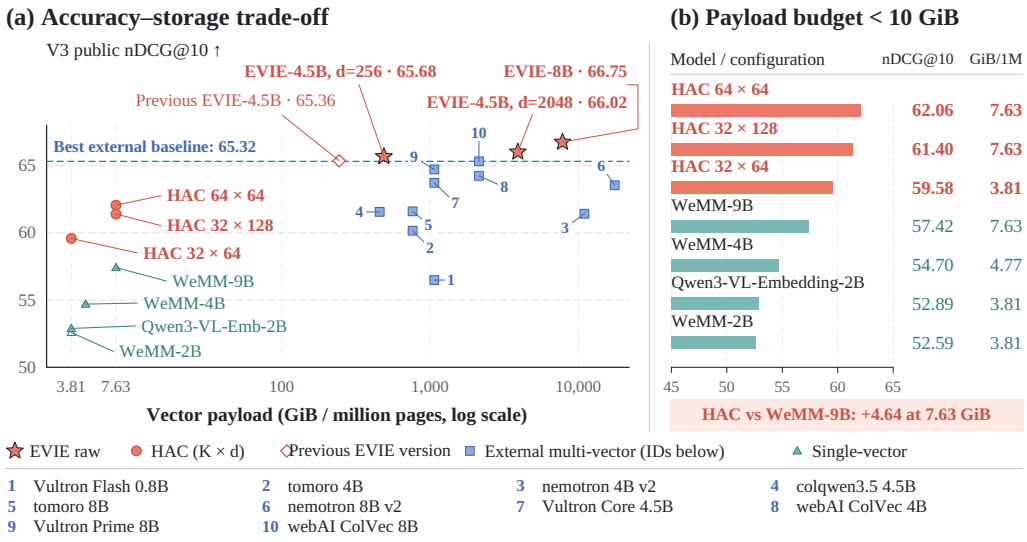

Figure 1: Accuracy versus vector payload on public ViDoRe V3. All points follow Appendix Table 10 and use two-byte components. Panel (b) expands configurations below 10 GiB per million pages; HAC labels give $K \times d$.

EVIE (Evidence-Vector-Informed Embeddings) addresses these coupled decisions while retaining one scoring interface. A multimodal judge first grades mined pages by answer completeness: full evidence becomes positive supervision, partial evidence is excluded from its annotated negative slot, and absent evidence remains negative. Full- and partial-evidence decisions account for 14.4% of pre-expansion candidate slots. A bidirectional EVIE-8B teacher then transfers candidate rankings to EVIE-4.5B. Prefix-based Matryoshka representation learning (Prefix-MRL) applies retrieval supervision and bidirectional listwise distillation at six nested widths, making each prefix a trained MaxSim representation rather than a post-hoc truncation. Finally, hierarchical agglomerative compression (HAC) reduces the stored page-token set. Spatial features guide cluster assignments, while the index stores semantic centroids that support direct MaxSim scoring with unclustered query tokens. Figure 2 summarizes the complete pipeline.

We evaluate the resulting model family on 138 tasks across ViDoRe V1, V2, V3, and JinaVDR. Figure 1 shows the resulting accuracy–payload frontier and the dense set of external operating points. The teacher reaches 66.75 nDCG@10 on public ViDoRe V3, and the full-width student remains within 0.73 points. At the vector payload of WeMM-9B (Zhou et al., 2026), HAC reaches 62.06 versus 57.42; at half that payload, it retains 59.58. Matched controls separate evidence relabeling, prefix supervision, ranking transfer, and spatial assignment from the heterogeneous model comparison. With vector precision and HAC construction fixed, the tested equal-payload allocations favor retaining more narrow centroids over fewer wide ones.

The contributions follow the three decisions at the MaxSim interface. First, we introduce an answer-completeness policy for mined candidates and evaluate it with sequential supervision controls and a stratified human audit. Second, we train six late-interaction widths in one student checkpoint through prefix-wise retrieval and ranking transfer. Third, we introduce spatially guided semantic pooling for direct retrieval and test its assignment rule, transfer across encoders, and allocation under fixed payloads. Together, these components connect page-evidence supervision to compact multi-vector indexing without changing the final retrieval operator.

2 Related Work

Visual Document Retrieval. Native visual retrievers encode page images without an OCR-dependent retrieval representation. DSE and Qwen3-VL-Embedding produce one vector per page (Ma et al., 2024; Li et al., 2026), whereas ColPali applies ColBERT-style late interaction to vi-

sual document tokens (Khattab & Zaharia, 2020; Faysse et al., 2025). Jina v4 supports both output forms; Nemotron-ColEmbed and ModernVBERT develop stronger or smaller multi-vector encoders (Günther et al., 2025; Moreira et al., 2026; Teiletche et al., 2025). Universal multimodal encoders extend the same retrieval objective beyond document pages (Jiang et al., 2024; Zhang et al., 2024; Meng et al., 2025). EVIE retains late interaction and studies a complementary question: how the supervision and stored representation should preserve answer-bearing local matches under a changing index budget.

Supervision and Ranking Transfer. Hard-negative mining improves contrastive training but can select unlabeled relevant documents. ColBERTv2 combines mined negatives with cross-encoder score distillation, Unveil transfers visual–textual knowledge, and ColMate adds region-linked masked-text learning (Santhanam et al., 2022b; Sun et al., 2025; Masry et al., 2025). These methods strengthen the ranking signal or token representation. EVIE instead makes answer completeness an explicit policy for the status of each mined page: complete evidence expands the positive set, partial evidence is withheld from the annotated negative slot, and absent evidence remains negative. The judgments are page-level relevance decisions rather than token-localization targets. The same candidate sets then support teacher–student ranking transfer in both query-to-page and page-to-query directions.

Nested Representations and Compact Indexes. Matryoshka representation learning trains nested embedding prefixes (Kusupati et al., 2022); Qwen3-VL-Embedding and ColNetraEmbed bring dimension flexibility to multimodal retrieval (Li et al., 2026; Kolavi & Jain, 2025). For late interaction, truncating every token changes the local similarities and maxima used for ranking. Prefix-MRL therefore applies retrieval and teacher-ranking objectives to every trained width under MaxSim. Width reduction is complementary to reducing the number of stored vectors. ColBERTv2 compresses individual vectors, ColPali studies hierarchical mean pooling, and PLAID uses centroid interaction and pruning for efficient search (Santhanam et al., 2022b; Faysse et al., 2025; Santhanam et al., 2022a). ColParse encodes layout-informed regions, CausalEmbed generates latent retrieval tokens, and MetaEmbed varies learned token counts (Yan et al., 2026; Huo et al., 2026; Xiao et al., 2026). HAC uses the processor grid to guide hierarchical assignments but averages only semantic token vectors into the index. The resulting centroids participate directly in final MaxSim retrieval, allowing width and token count to be controlled independently.

3 EVIE: Learning and Indexing Page Evidence

We use the same late-interaction interface throughout learning and retrieval. Evidence judgments determine which page matches training should reward; prefix-wise objectives train those matches at different widths; HAC reduces the token set over which they are computed.

The stages act at different granularities. The judge assigns a page-level relevance tier, retrieval and ranking losses learn the token-level matches, and HAC groups page vectors independently of the query. Thus evidence defines the supervision criterion, while local alignment remains an outcome of representation learning.

3.1 Independent Encoding and Nested Token Representations

Within each model, query and page encoders share parameters but process their inputs independently. The teacher and student use Qwen3.5-family backbones (Qwen Team, 2026), with hidden widths 4096 and 2560 and deployed parameter counts of 8.41B and 4.54B. We remove the causal mask in the eight full-attention layers of each 32-layer language backbone while retaining padding masks; the remaining linear/recurrent layers are unchanged. Both models train rank-32 LoRA adapters (Hu et al., 2022) and a shared affine retrieval head, with a frozen visual encoder. Appendix A gives initialization, tokenization, and optimization details.

For a valid hidden token $h_{x,i}$ of input $x \in \{q, p\}$, a shared affine projection produces a retrieval vector at width d:

$$e_{x,i}^{(d)} = \mathrm{L2}((W h_{x,i} + b)_{1:d}), \qquad s_d(q,p) = \frac{1}{n_q} \sum_{i \in I_q} \max_{j \in I_p} \langle e_{q,i}^{(d)}, e_{p,j}^{(d)} \rangle. \tag{1}$$

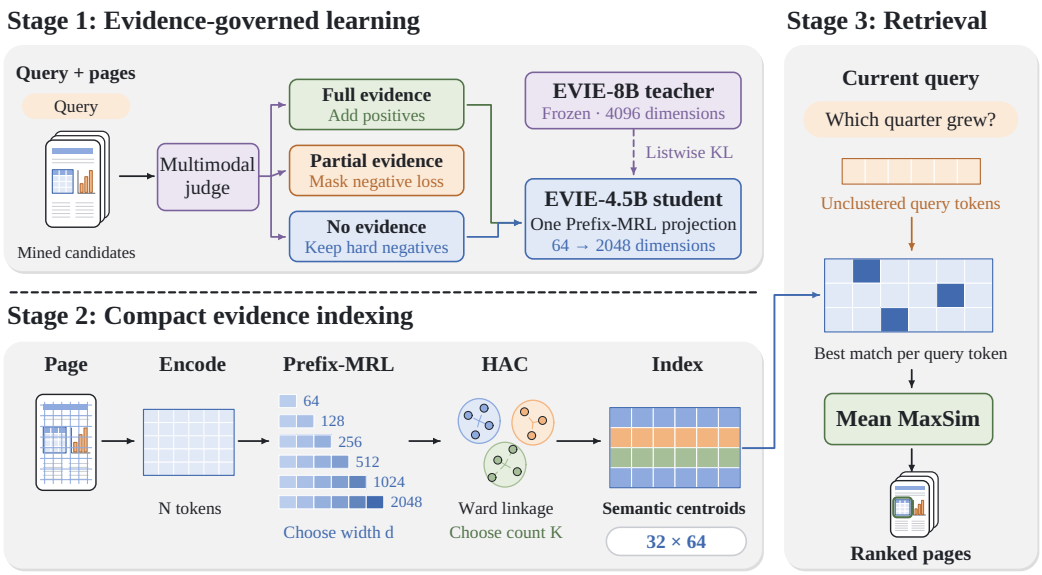

Figure 2: EVIE training, indexing, and retrieval. Evidence judgments determine page supervision, Prefix-MRL transfers teacher rankings to six widths, and HAC stores semantic centroids for direct MaxSim scoring.

Here I_q, I_p index valid tokens, $n_q = |I_q|$, and L2 normalizes nonzero vectors. The teacher uses $d = 4096$; the student uses $\mathcal{D} = \{64, 128, 256, 512, 1024, 2048\}$. Every width shares the leading rows of one projection. A cached full-width projection can therefore supply any trained prefix by slicing and renormalizing. Training uses the query mean in Eq. 1; evaluation uses the sum, which preserves the ranking for a fixed query. Invalid page positions are excluded before the maximum.

Pages are dynamically resized into at most 1024 image tokens. Raw retrieval also retains context tokens, whereas HAC operates on image tokens only. Appendix A specifies the complete input sequences and masks.

3.2 Evidence-Judged Supervision

A mined page is useful as a negative only if it lacks the evidence needed to answer the query. GLM-5V-Turbo (GLM-5V-Turbo Team, 2026) serves as the multimodal judge: it sees the query, registered page, and one mined candidate, and assigns three tiers. *Full evidence* augments the registered positives, *partial evidence* is excluded from its negative slot, and *no evidence* remains negative. Unjudged candidates retain their original status. The corpus contains 775,635 original query groups, including 179,295 without evidence judgments. Before positive expansion, 223,877 of 1,551,270 candidate slots are masked (14.4%), and 175,137 groups have multiple positives.

Let $\mathcal{P}_g$ contain the positive pages for query group g. The all-positive view creates one record i for each $p_i \in \mathcal{P}_{g(i)}$, with weight $\omega_i = 1/|\mathcal{P}_{g(i)}|$ so that each group's nominal weight sums to one. This yields 1,050,270 records. Each record also carries two mined negative slots n_{ik} with validity masks m_{ik}; full- and partial-evidence slots are masked. At an active width, retrieval supervision combines in-batch cross-entropy with masked pairwise ranking:

$$\ell_{\mathrm{ret},i} = \frac{1}{2}\left[-\log \frac{\exp(s(q_i,p_i)/\tau)}{\sum_{j\in\mathcal{C}_i}\exp(\widetilde{s}_{ij}/\tau)} + \frac{\sum_k m_{ik}\,\mathrm{softplus}(-\Delta_{ik}/\tau)}{\max(1,\sum_k m_{ik})}\right], \tag{2}$$

where $\Delta_{ik} = s(q_i,p_i) - s(q_i,n_{ik})$ and $\tau = 0.02$. The permitted in-batch set $\mathcal{C}_i$ and attenuated scores $\widetilde{s}_{ij}$ prevent known or likely positives from dominating the negative loss (Appendix A.4). Full- and partial-evidence policies act on annotated slots; they do not establish global relevance labels for every query–page pair.

Two teacher branches share initialization: one samples a positive from each group, and the other uses all weighted positives. Their trained parameters are combined as $\theta_T = 0.4\theta_{\text{sampled}} + 0.6\theta_{\text{all}}$ and frozen for student training.

3.3 Prefix-wise Ranking Transfer

Nested parameters alone do not ensure that a narrow prefix preserves the useful rankings of a wide representation. We train all six prefixes with Eq. 2 and teacher supervision. Query-to-page distillation normalizes scores over gathered positives and explicit negative slots. Page-to-query distillation normalizes the positive-page score matrix over gathered queries. Teacher and student share candidate masks, excluding non-target known positives and invalid negative slots as detailed in Appendix A.4.

At temperature $\tau_{\text{kd}} = 0.13$, these scores define teacher distributions $P^{\rho}_{T,i}$ and student distributions $P^{(d),\rho}_{S,i}$ for retrieval direction ρ. Ranking transfer averages the two directions:

$$\ell^{(d)}_{\text{rel},i} = \frac{1}{2} \sum_{\rho \in \{q\to p, p\to q\}} c^{\rho}_i \, \text{KL}\Big(P^{\rho}_{T,i} \parallel P^{(d),\rho}_{S,i}\Big). \tag{3}$$

The confidence factor c^{ρ}_i downweights high-entropy or target-inconsistent teacher rankings. Both KL terms run from teacher to student: "bidirectional" refers to the two candidate normalizations, and does not make MaxSim itself symmetric.

The complete student objective also transfers teacher positive–negative margins and anchors the initial 128-dimensional retrieval geometry:

$$\mathcal{L}_S = \overline{\mathcal{L}}_{\text{ret}} + a_t \left(\overline{\mathcal{L}}_{\text{rel}} + 0.25\overline{\mathcal{L}}_{\text{mar}} + 0.25\mathcal{L}^{(128)}_{\text{anc}}\right). \tag{4}$$

An overbar denotes the mean over all six widths after weighted record reduction. Margin transfer uses teacher-confidence-weighted Smooth-L1 on explicit-negative score margins; the anchor uses bidirectional KL to a frozen copy of the initial d128 retriever. Appendix A.4 defines both terms, the calibration of a_t, candidate masks, and the projection warmup.

3.4 Spatially Guided Semantic Compression

After training, HAC selects width d and retains only the page's N image vectors $v_i = e^{(d)}_{p,i}$. Processor-grid coordinates supply normalized positions $(x_i, y_i) \in [0,1]^2$. A unit-normalized 2D sinusoidal encoding $p_i \in \mathbb{R}^d$ is combined with the semantic vector for cluster assignment:

$$z_i = \text{L2}((1 - w_{\text{pos}})v_i + w_{\text{pos}}p_i), \qquad \mu_c = \text{L2}(\frac{1}{|c|}\sum_{i\in c} v_i). \tag{5}$$

Euclidean Ward linkage (Ward, 1963) clusters z_i into $\min(K, N)$ groups, minimizing the increase in within-cluster squared error at each merge. We use $w_{\text{pos}} = 0.1$. The stored centroid μ_c averages the original semantic vectors, so positional features affect assignments without adding channels to the index. Queries remain unclustered and score directly against these centroids by Eq. 1.

This separation reduces the candidate token set without adding positional channels to retrieval. Assignments are recomputed for each reported (d, K) configuration. Appendix B gives the positional encoding, algorithm, and export details.

With two-byte components, P pages storing K centroids of width d require

$$S_{\text{payload}} = \frac{2PKd}{2^{30}} \text{ GiB}. \tag{6}$$

Every page in the reported HAC runs has $N \geq K$. Thus $K = 32, d = 64$ requires 4 KiB per page, or 3.81 GiB per million pages. Width and token count provide multiplicative controls over payload, but may have different effects on retrieval. Our fixed-payload experiments test how to allocate this budget. All payload figures exclude model weights, identifiers, offsets, and index-container overhead.

Table 1: Main nDCG@10 results across four suites. V3 covers eight public domains with six languages each; Avg4 weights the suites equally. External V3 values are reported scores, while the remaining values follow Section 4.1. HAC denotes hierarchical agglomerative compression.

Model / configuration	V1	V2	V3	JinaVDR	Avg4
WeMM-9B (Zhou et al., 2026)	90.95	65.62	57.42	72.05	71.51
nemotron-colembed-vl-8b-v2 (NVIDIA, 2026b)	92.27	66.74	63.54	81.41	75.99
webAI-ColVec1.1-8b (webAI, 2026b)	91.29	69.77	65.32	82.04	77.11
jina-embeddings-v4 (single-vec) (Günther et al., 2025)	84.67	55.74	49.61	68.74	64.69
jina-embeddings-v4 (multi-vec) (Günther et al., 2025)	89.73	60.43	56.99	76.17	70.83
EVIE-8B ($d = 4096$)	92.64	75.35	66.75	83.30	**79.51**
EVIE-4.5B Prefix-MRL ($d = 2048$)	92.53	74.91	66.02	82.48	78.98
EVIE-4.5B Prefix-MRL ($d = 64$)	92.16	73.18	64.51	81.00	77.71
EVIE-4.5B + HAC ($d = 64, K = 64$)	90.63	70.71	62.06	78.90	75.57
EVIE-4.5B + HAC ($d = 64, K = 32$)	89.27	67.83	59.58	77.21	73.47

4 EXPERIMENTS

We evaluate three questions: whether evidence judgments improve relevance learning, whether prefix-wise transfer preserves accuracy at narrow widths, and how spatial token pooling changes retrieval at a fixed storage budget.

4.1 EVALUATION SETUP

Benchmarks and metrics. The evaluation contains 138 tasks: 10 from ViDoRe V1 (Faysse et al., 2025), four from V2 (Macé et al., 2025b), 48 from V3 (Macé et al., 2025a), and 76 from JinaVDR (Günther et al., 2025). They contain 3,943, 1,152, 14,514, and 43,375 queries, respectively. V3 covers eight public domains with six query languages each; private domains are excluded. None of the four suites is included in the training data. We report nDCG@10 on a 0–100 scale, macro-averaged over tasks within each suite. Avg4 is the unweighted mean of the four suite scores. Dataset revisions and complete nDCG, Recall, MAP, and MRR results at cutoffs 1, 5, and 10 appear in Appendices C and E.

Comparisons and accounting. EVIE uses at most 1024 image tokens and single-stage MaxSim without reranking. Raw configurations retain image and context tokens; HAC retains image-token centroids. External V3 values are reported scores from official model cards or benchmark releases on the eight public domains; WeMM, the other suites, and all controlled comparisons use local measurements. The external comparison supplies model-level context, while matched controls test individual EVIE components. Payloads use two-byte components. Raw multi-vector values are configured or published maximum image-token capacities; HAC and single-vector values use fixed vector counts. Appendix D lists every configuration and score source.

4.2 RETRIEVAL QUALITY ACROSS SUITES AND BUDGETS

Figure 1 situates the operating points across payload budgets. EVIE-8B reaches 66.75 on V3 and 79.51 Avg4 (Table 1), 1.43 V3 points above the highest reported external score, webAI-ColVec1.1-8b at 65.32 (webAI, 2026b). The student reaches 66.02 and 78.98, within 0.73 V3 points and 0.53 Avg4 points of the teacher. At $d = 256$, it retains 65.68 V3 and 78.45 Avg4 with a configured maximum visual-token payload of 488.28 GiB per million pages, compared with the published maximum capacity of 2136.23 GiB for webAI.

Table 2 isolates the compression trade-off with representative systems and EVIE operating points. At 7.63 GiB per million pages, HAC ($d = 64, K = 64$) reaches 62.06 V3, compared with 61.40 for ($d = 128, K = 32$) and 57.42 for WeMM-9B (Zhou et al., 2026). At 3.81 GiB, HAC ($d = 64, K = 32$) reaches 59.58, 6.69 points above Qwen3-VL-Embedding-2B (Li et al., 2026) at the same payload. Relative to raw EVIE-4.5B at $d = 256$, this setting reduces payload from 488.28 to 3.81 GiB per million pages, a $128\times$ reduction, while changing V3 from 65.68 to 59.58. Table 1 shows that the d64 HAC configurations also lead WeMM-9B on V2 and JinaVDR.

Table 2: ViDoRe V3 nDCG@10 on eight public domains and embedding payload. K is fixed for single-vector and HAC configurations and is the maximum image-token count for raw multi-vector configurations. Payload assumes two-byte components and excludes index overhead; Appendix Table 10 reports all configurations.

Model / configuration	Params	K	d	GiB / 1M	V3
Single-vector baselines					
Qwen3-VL-Embedding-2B (Li et al., 2026)	2.1B	1	2048	3.81	52.89
WeMM-9B (Zhou et al., 2026)	9.0B	1	4096	7.63	57.42
Raw multi-vector baselines					
colqwen3.5-4.5B-v3 (athrael-soju, 2026)	4.6B	768	320	457.76	61.56
webAI-ColVec1.1-8b (webAI, 2026b)	8.4B	1792	640	2136.23	65.32
EVIE raw and compressed					
EVIE-8B teacher	8.4B	1024	4096	7812.50	**66.75**
EVIE-4.5B Prefix-MRL ($d = 2048$)	4.5B	1024	2048	3906.25	66.02
EVIE-4.5B Prefix-MRL ($d = 256$)	4.5B	1024	256	488.28	65.68
EVIE-4.5B + HAC ($d = 128$, $K = 32$)	4.5B	32	128	7.63	61.40
EVIE-4.5B + HAC ($d = 64$, $K = 64$)	4.5B	64	64	7.63	62.06
EVIE-4.5B + HAC ($d = 64$, $K = 32$)	4.5B	32	64	3.81	59.58

Table 3: Evidence supervision: (a) sequential policy ablation; (b) stratified three-annotator audit.

(a) Sequential Policy Ablation

Variant	Mask full	Expand full	Mask partial	V3	Avg4
G0: Original	No	No	No	62.36	75.57
G1: Mask full	Yes	No	No	64.23	77.71
G2: Expand	Yes	Yes	No	65.77	78.98
G3: Mask partial	Yes	Yes	Yes	66.33	79.17

(b) Human Evidence Audit

Judge tier	n	Full	Partial	None
Full	300	92.67	6.67	0.67
Partial	300	9.67	80.67	9.67
No	300	0.67	3.67	95.67

Domain and ranking-depth behavior. Across the eight public domains, the full-width teacher–student gap is 0.14–1.50 V3 points, while the d2048-to-HAC loss ranges from 1.89 on Physics to 9.51 on Human Resources (Appendix Table 11). On V3, compression changes Recall@10 from 69.67 to 64.00 and MRR@10 from 75.43 to 69.73, affecting both relevant-page coverage and the first relevant rank (Appendix Tables 15 and 16).

4.3 Controlled Ablation Studies

Evidence completeness improves relevance learning. Table 3(a) adds three governance steps to the sampled-positive teacher branch before parameter averaging. All four runs share initialization, candidate pool, data-source mixture, query-group order, optimizer schedule, update budget, effective batch size, and seed. Each scheduled group contributes one sampled positive; expanding the eligible positive set does not change the group schedule. Removing full-evidence negatives gives the largest first-step improvement: +1.87 V3 and +2.14 Avg4. Promoting them to positives adds 1.54 and 1.27 points, showing that masking alone leaves useful supervision unused. Excluding partial evidence adds a further 0.56 and 0.19 points. The complete policy raises V3 from 62.36 to 66.33 and Avg4 from 75.57 to 79.17. Each increment is conditional on the policies in the preceding row and the fixed seed.

A stratified human audit covers 900 candidates, with 300 per judge tier (Table 3(b); protocol in Appendix F.1). Three volunteer annotators independently apply the same full-, partial-, and no-evidence rubric, and their majority vote defines the human label for each candidate. These labels match the judge tier for 92.67% of full-evidence candidates, 80.67% of partial-evidence candidates, and 95.67% of no-evidence candidates. Direct full/none confusion is 0.67% in either direction. The lower partial-tier agreement supports the conservative masking policy. Because sampling is balanced across judge tiers, these are tier-conditional rates rather than prevalence-weighted accuracy over the mined-candidate pool.

Table 4: Teacher and supervision controls: (a) parameter averaging; (b) governance effects on HAC loss.

(a) Teacher Construction at $d = 4096$

Teacher checkpoint	V3	Avg4
Sampled-positive branch	66.33	79.17
All-positive branch	66.23	79.06
Final mixture $(0.4/0.6)$	**66.75**	**79.51**

(b) Governance and Compression at $d = 64$

Supervision	Raw image only	HAC $K = 32$	Compression loss
Original supervision	61.38	54.80	6.58
Full governance	64.39	59.58	4.81

Table 5: Matched student ablations on V3. B0 truncates a maximum-width model; B1–B3 add prefix supervision and the two KL directions; B4 adds margin transfer and the d128 anchor jointly.

Training variant	$d = 64$	$d = 128$	$d = 2048$
B0: Maximum-width retrieval + truncation	62.04	63.84	65.03
B1: Retrieval at all six prefixes	62.54	63.90	65.06
B2: B1 + query-to-page listwise KL	63.19	64.40	65.56
B3: B2 + page-to-query listwise KL	63.44	64.65	65.81
B4: Full objective	64.51	65.27	66.02

The teacher mixture provides a separate gain. After evidence governance, the sampled-positive and all-positive teacher branches reach 66.33 and 66.23 V3 nDCG@10, respectively. Their $0.4/0.6$ parameter mixture adds 0.42 V3 and 0.34 Avg4 points over the stronger branch (Table 4(a)). The governance ablation measures the label policy before this separate checkpoint combination. Table 4(b) further shows that full governance raises raw and HAC V3 by 3.01 and 4.78 points, reducing the raw-to-HAC loss from 6.58 to 4.81 at this operating point. This single-run interaction is descriptive.

Ranking transfer supports narrow prefixes. Table 5 holds student initialization, training records and order, optimizer schedule, update count, batch size, and seed fixed. B0 trains only at maximum width and truncates at evaluation. B1 applies retrieval supervision at all prefixes; B2 and B3 add the two listwise directions; B4 uses the complete objective. Direct prefix supervision adds 0.50 V3 points at $d = 64$, compared with 0.06 and 0.03 at $d = 128$ and $d = 2048$. Query-to-page KL then adds 0.65, 0.50, and 0.50 points; page-to-query KL adds 0.25 at each width. Relative to B1, the full objective gains 1.97, 1.37, and 0.96 points. The widening gap toward $d = 64$ shows that ranking transfer contributes most under the tightest width budget. B4 evaluates margin transfer and the anchor jointly.

Spatial features improve semantic pooling. Table 6(a) fixes the final student at $d = 64$ and compares pooling methods at $K = 32$. Removing context tokens costs only 0.12 V3 points; subsequent HAC pooling costs 4.81. Most of the raw-to-HAC loss at this setting therefore comes from pooling the image vectors. Semantic-only Ward improves over grid pooling by 8.75 V3 points. Adding the positional feature yields a further 1.46 points and 0.88 Avg4 points without increasing centroid count or width. The same within-encoder comparison gives gains of 1.28 points for Jina v4 and 0.78 for Nemotron-ColEmbed-V2-4B at $K = 32$ (Table 6(b)), showing that the assignment prior transfers beyond EVIE.

4.4 Allocating a Fixed Payload

The final student's six prefixes offer a controlled width axis (Table 7(a)): reducing d from 2048 to 64 costs 1.51 V3 points and 1.27 Avg4 points for a $32\times$ width reduction. Token pooling has a larger effect along the path to the smallest index: $66.02 \to 64.51 \to 59.58$ for full width, raw d64, and HAC d64/$K = 32$, respectively. This motivates testing whether a fixed payload is better spent on width or token coverage. Table 7(b) holds vector precision and HAC construction fixed, so equal payload is equivalent to fixed Kd within each of the three budgets. At 3.81, 7.63, and 15.26 GiB per million pages, the $d = 64$ configurations achieve the highest V3 and Avg4 point estimates among the tested allocations. Their V3 gains over $d = 128$ are 0.88, 0.66, and 0.07 points; gains over $d = 256$

Table 6: HAC ablations at $K = 32$: (a) pooling components; (b) transfer across three encoders.

(a) Pooling Components

Representation / pooling	Vectors	w_{pos}	V3	Avg4
Raw, img and ctx	N_{all}	—	64.51	77.71
Raw, img only	N	—	64.39	77.63
Grid pooling	32	—	49.37	63.87
Semantic-only Ward	32	0	58.12	72.59
Position-guided Ward (HAC)	32	0.1	59.58	73.47

(b) Transfer Across Encoders

Encoder	d	Raw image only	Semantic Ward	HAC
EVIE-4.5B (final)	64	64.39	58.12	59.58
Jina v4 (mt-vec)	128	56.38	49.60	50.88
Nemotron-ColEmbed-V2-4B	2560	61.62	57.98	58.76

Table 7: Width/count trade-offs: (a) Prefix-MRL scaling; (b) equal-payload HAC allocations at two-byte precision.

(a) Prefix Width Scaling

Embedding width d	GiB / 1M pages	V3	Avg4
64	122.07	64.51	77.71
128	244.14	65.27	78.18
256	488.28	65.68	78.45
512	976.56	65.90	78.77
1024	1953.13	65.94	78.82
2048	3906.25	66.02	78.98

(b) Fixed-Payload HAC Allocations

GiB / 1M pages	Embedding width d	Centroids K	V3	Avg4
3.81	64	32	59.58	73.47
3.81	128	16	58.70	72.38
3.81	256	8	55.19	67.63
7.63	64	64	62.06	75.57
7.63	128	32	61.40	74.87
7.63	256	16	59.28	72.63
15.26	64	128	63.40	76.86
15.26	128	64	63.33	76.56
15.26	256	32	61.98	75.33

range from 1.42 to 4.39. Under this common-precision setting, allocating the same payload to more narrow centroids is most beneficial under severe compression, and the advantage approaches parity as the budget grows.

Serving and failure cases. On 1000 stratified V3 pages using one NVIDIA H20, HAC $K = 32$ reduces median scoring time from 0.909 to 0.160 ms and GPU-resident index memory from 0.2444 to 0.0076 GiB relative to raw d64. Batch-16 throughput rises from 105.26 to 115.73 queries/s, while query encoding remains 99.67 ms. WeMM-9B scores in 0.125 ms at the same measured index memory; the EVIE cascade stores both representations (0.2520 GiB) and scores in 0.335 ms. On this subset, 49 queries move from raw rank ≤ 3 to HAC rank > 10; Appendix G gives the protocol and examples.

5 Discussion and Conclusion

EVIE treats supervision, representation width, and token storage as coupled decisions at the MaxSim interface. Evidence judgments correct mined-candidate supervision, prefix-wise ranking transfer trains six widths, and HAC stores a compact set of semantic page regions. One teacher and one multi-width student span the measured accuracy–payload range from raw d2048 tokens to d64 centroids.

The experiments separate width reduction from token pooling: narrowing from 2048 to 64 dimensions costs less V3 accuracy than pooling the sequence to 32 centroids. With precision and HAC construction fixed, more narrow centroids achieve the highest point estimates among the tested equal-payload allocations, although the advantage diminishes at the largest budget. Suite- and domain-level variation makes the final operating point selectable for each retrieval setting.

AI USE STATEMENT

GLM-5V-Turbo (GLM-5V-Turbo Team, 2026) supplies the full-, partial-, and no-evidence judgments used to construct retrieval supervision (Section 3.2); Appendix F.1 reports the stratified human audit of those judgments. A generative AI assistant was also used to edit the prose, prepare the LaTeX source, and check formatting and numerical consistency. This manuscript preparation reused the supplied experimental results; it did not generate new empirical measurements. Responsibility for the submitted text, claims, references, and artifacts rests with the authors.

ETHICS STATEMENT

This work uses existing public document-retrieval datasets. Preprocessing removes empty queries and images; evidence judgments assess relevance and determine positive expansion and negative masking. These checks do not constitute a comprehensive privacy or harmful-content audit. Public pages may contain personally identifiable information. Use and redistribution remain subject to dataset licenses and underlying document restrictions; public availability does not remove these obligations. Retrieval deployments should respect corpus access permissions, and compressed retrieval should not be assumed to preserve every answer-bearing page.

REPRODUCIBILITY STATEMENT

Section 3 defines the representation, supervision policy, distillation objective, and compressed scoring interface. Appendix A specifies initialization, token masks, optimization, and training algorithms. Appendices C and E record dataset revisions, aggregation, and secondary metric definitions. Appendices D–G give the additional operating points, controlled comparisons, and serving setup. The submission also includes an anonymized supplementary code archive implementing the evidence-governance processing, training, HAC indexing, and evaluation pipelines. These materials document the reported experiments; external released scores and local measurements are distinguished in Section 4.1.

REFERENCES

athrael-soju. colqwen3.5-4.5B-v3 model card. Hugging Face model repository, 2026. URL https://huggingface.co/athrael-soju/colqwen3.5-4.5B-v3.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2407.01449. arXiv:2407.01449; introduces the ViDoRe benchmark.

GLM-5V-Turbo Team. GLM-5V-Turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026. URL https://arxiv.org/abs/2604.26752.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *arXiv preprint arXiv:2506.18902*, 2025. URL https://arxiv.org/abs/2506.18902.

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

Jiahao Huo, Yu Huang, Yibo Yan, Ye Pan, Kening Zheng, Wei-Chieh Huang, Yi Cao, Mingdong Ou, Philip S. Yu, and Xuming Hu. CausalEmbed: Auto-regressive multi-vector generation in latent space for visual document embedding. *arXiv preprint arXiv:2601.21262*, 2026. URL https://arxiv.org/abs/2601.21262.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language

models. *arXiv preprint arXiv:2407.12580*, 2024. URL https://arxiv.org/abs/2407.12580.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Adithya S Kolavi and Vyoman Jain. M3DR: Towards universal multilingual multimodal document retrieval. *arXiv preprint arXiv:2512.03514*, 2025. URL https://arxiv.org/abs/2512.03514. releases NetraEmbed and ColNetraEmbed.

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 *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. URL https://arxiv.org/abs/2205.13147.

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. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6492–6505, 2024. doi: 10.18653/v1/2024.emnlp-main.373. URL https://arxiv.org/abs/2406.11251. arXiv:2406.11251.

Quentin Macé, Antonio Loison, Antoine Edy, Victor Xing, and Gautier Viaud. ViDoRe V3: A comprehensive evaluation of retrieval for enterprise use-cases. Hugging Face Blog, November 2025a. URL https://huggingface.co/blog/QuentinJG/introducing-vidore-v3.

Quentin Macé, António Loison, and Manuel Faysse. ViDoRe benchmark V2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*, 2025b. URL https://arxiv.org/abs/2505.17166.

Ahmed Masry, Megh Thakkar, Patrice Bechard, Sathwik Tejaswi Madhusudhan, Rabiul Awal, Shambhavi Mishra, Akshay Kalkunte Suresh, Srivatsava Daruru, Enamul Hoque, Spandana Gella, Torsten Scholak, and Sai Rajeswar. ColMate: Contrastive late interaction and masked text for multimodal document retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2071–2080, 2025. doi: 10.18653/v1/2025.emnlp-industry.145. URL https://arxiv.org/abs/2511.00903. arXiv:2511.00903.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. URL https://arxiv.org/abs/2507.04590. introduces MMEB-V2.

Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, Bo Liu, and Even Oldridge. Nemotron ColEmbed v2: Top-performing late interaction embedding models for visual document retrieval. *arXiv preprint arXiv:2602.03992*, 2026. URL https://arxiv.org/abs/2602.03992.

NVIDIA. nemotron-colembed-vl-4b-v2 model card. Hugging Face model repository, 2026a. URL https://huggingface.co/nvidia/nemotron-colembed-vl-4b-v2.

NVIDIA. nemotron-colembed-vl-8b-v2 model card. Hugging Face model repository, 2026b. URL https://huggingface.co/nvidia/nemotron-colembed-vl-8b-v2.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official model-family release, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. PLAID: An efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 1747–1756, 2022a. doi: 10.1145/3511808.3557325. URL https://arxiv.org/abs/2205.09707.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022b. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

Hao Sun, Yingyan Hou, Jiayan Guo, Bo Wang, Chunyu Yang, Jinsong Ni, and Yan Zhang. Unveil: Unified visual-textual integration and distillation for multi-modal document retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23935–23945, 2025. doi: 10.18653/v1/2025.acl-long.1166. URL https://aclanthology.org/2025.acl-long.1166/.

Paul Teiletche, Quentin Macé, Max Conti, Antonio Loison, Gautier Viaud, Pierre Colombo, and Manuel Faysse. ModernVBERT: Towards smaller visual document retrievers. *arXiv preprint arXiv:2510.01149*, 2025. URL https://arxiv.org/abs/2510.01149. ColModernVBERT is the late-interaction variant released in this paper.

TomoroAI. tomoro-colqwen3-embed-4b model card. Hugging Face model repository, 2026a. URL https://huggingface.co/TomoroAI/tomoro-colqwen3-embed-4b.

TomoroAI. tomoro-colqwen3-embed-8b model card. Hugging Face model repository, 2026b. URL https://huggingface.co/TomoroAI/tomoro-colqwen3-embed-8b.

Vultr. VultronRetrieverCore-Qwen3.5-4.5B model card. Hugging Face model repository, 2026a. URL https://huggingface.co/vultr/VultronRetrieverCore-Qwen3.5-4.5B.

Vultr. VultronRetrieverFlash-Qwen3.5-0.8B model card. Hugging Face model repository, 2026b. URL https://huggingface.co/vultr/VultronRetrieverFlash-Qwen3.5-0.8B.

Vultr. VultronRetrieverPrime-Qwen3.5-8B model card. Hugging Face model repository, 2026c. URL https://huggingface.co/vultr/VultronRetrieverPrime-Qwen3.5-8B.

Joe H. Ward, Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845. URL https://doi.org/10.1080/01621459.1963.10500845.

webAI. webAI-ColVec1.1-4b model card. Hugging Face model repository, 2026a. URL https://huggingface.co/webAI-Official/webAI-ColVec1.1-4b.

webAI. webAI-ColVec1.1-8b model card. Hugging Face model repository, 2026b. URL https://huggingface.co/webAI-Official/webAI-ColVec1.1-8b.

Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. MetaEmbed: Scaling multimodal retrieval at test-time with flexible late interaction. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.18095. arXiv:2509.18095.

Yibo Yan, Mingdong Ou, Yi Cao, Xin Zou, Shuliang Liu, Jiahao Huo, Yu Huang, James Kwok, and Xuming Hu. Beyond the grid: Layout-informed multi-vector retrieval with parsed visual document representations. *arXiv preprint arXiv:2603.01666*, 2026. URL https://arxiv.org/abs/2603.01666. introduces ColParse.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. GME: Improving universal multimodal retrieval by multimodal LLMs. *arXiv preprint arXiv:2412.16855*, 2024. URL https://arxiv.org/abs/2412.16855.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. WeMM-Embedding: WeChat multi-modal embedding technical report. *arXiv preprint arXiv:2608.24053*, 2026. URL https://arxiv.org/abs/2608.24053.

A Implementation and Training Details

This appendix specifies the token representation and training objective used in the main paper. The additional equations make masking, weighting, and teacher confidence explicit.

A.1 Input Sequences and Contextual Encoding

Pages undergo aspect-preserving dynamic resizing, 16×16 patch embedding, and 2×2 spatial merging, yielding $N \leq 1024$ image tokens. The processor records the resulting two-dimensional grid and its raster ordering; HAC derives page coordinates from this metadata without OCR or layout parsing. The page sequence also contains the fixed image-description prompt and special tokens, which participate in raw retrieval but are excluded from HAC pooling. Queries use the checkpoint's BPE tokenizer with ten appended `<|endoftext|>` augmentation tokens. These are valid tokens despite sharing the padding token ID, so validity is determined by the processor's attention mask.

Within each model, query and page encoders share parameters. For input $x \in \{q, p\}$, let $H_x \in \mathbb{R}^{L_x \times h}$ denote the hidden-state matrix and $a_x \in \{0,1\}^{L_x}$ its valid-token mask; L_x includes batch padding, and L_p includes both image and context tokens. To expose tokens to context on both sides, we remove causal masking from the eight full-attention layers of each 32-layer language backbone (Moreira et al., 2026). The other 24 linear/recurrent layers remain unchanged. In the full-attention layers, the additive mask changes as

$$A_{\text{causal}} = A_{\text{pad}} + A_{\text{tri}} \quad \longrightarrow \quad A_{\text{EVIE}} = A_{\text{pad}}, \tag{7}$$

where A_{tri} blocks future positions and A_{pad} blocks padding. Bidirectional context is confined to each independently encoded input; query–page interaction occurs only during scoring.

A.2 Model Initialization and Prefix Projection

EVIE-8B initializes from `Qwen/Qwen3.5-9B`. EVIE-4.5B initializes from a previous EVIE-4.5B retrieval checkpoint whose backbone is `Qwen/Qwen3.5-4B`. Both backbones belong to the Qwen3.5 family (Qwen Team, 2026). Their hidden and maximum retrieval widths $(h, D_{\max})$ are $(4096, 4096)$ and $(2560, 2048)$, respectively. The deployed models contain 8.41B and 4.54B parameters, including the retrieval head after LoRA merging. Both train rank-32 LoRA adapters (Hu et al., 2022) with $\alpha = 128$ on language attention, recurrent, and feed-forward projections, together with the full retrieval head; the visual encoder remains frozen.

Each model uses a shared affine projection $W \in \mathbb{R}^{D_{\max} \times h}$ with bias $b \in \mathbb{R}^{D_{\max}}$ for query and page tokens:

$$\begin{aligned} U_x &= H_x W^\top + \mathbf{1} b^\top \in \mathbb{R}^{L_x \times D_{\max}}, \\ E_x^{(d)} &= \operatorname{diag}(a_x)\, \mathrm{RowL2}(U_x[:, :d]) \in \mathbb{R}^{L_x \times d}. \end{aligned} \tag{8}$$

Here RowL2 normalizes each token vector, with zero vectors remaining zero; padding rows are then zeroed by a_x. The teacher uses $d = 4096$. The student supports $d \in \mathcal{D} = \{64, 128, 256, 512, 1024, 2048\}$ through nested prefixes, retaining the previous checkpoint's first 128 projection rows and bias entries at initialization. Available full-width projections can supply every prefix through slicing and renormalization without rerunning the backbone (Figure 3).

A.3 Late Interaction Scoring (MaxSim)

Writing $e_{x,i}^{(d)}$ for row i of $E_x^{(d)}$, the query-normalized score at width d is

$$s_d(q,p) = \frac{1}{n_q} \sum_{i : a_{q,i}=1} \max_{j : a_{p,j}=1} \left\langle e_{q,i}^{(d)}, e_{p,j}^{(d)} \right\rangle, \qquad n_q = \sum_{i=1}^{L_q} a_{q,i}. \tag{9}$$

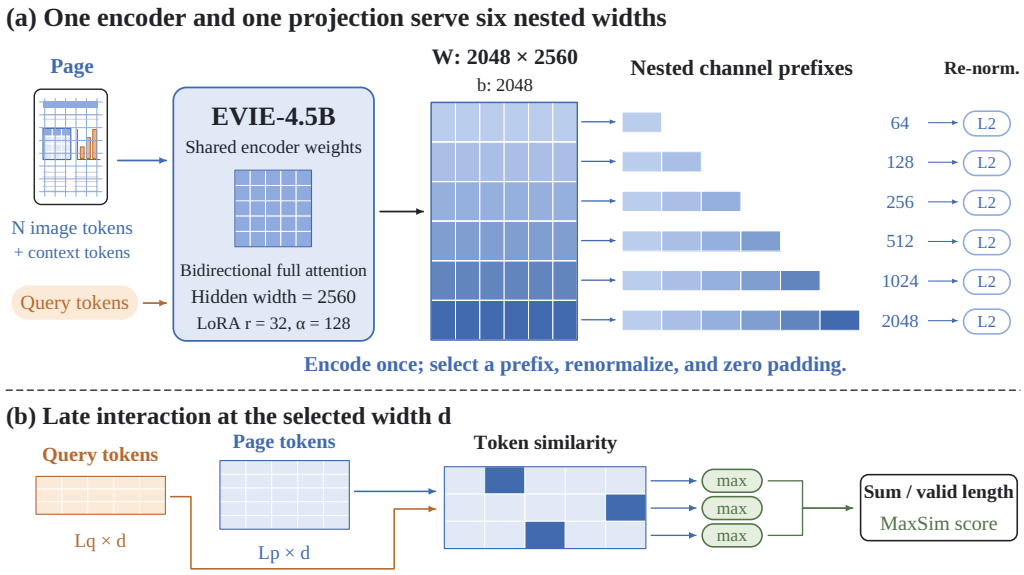

Figure 3: EVIE-4.5B encoding and prefix-wise matching. Independently encoded inputs share an affine projection; every selected prefix is normalized before MaxSim.

Before taking the maximum, invalid page positions are masked with a dtype-safe negative value. Query padding positions are excluded from the aggregation. Student training uses the query mean above; evaluation uses the sum, which gives the same ranking for a fixed query.[1] Nested projection parameters alone do not ensure useful rankings at every width. Section 3.3 specifies the relevance labels and teacher signals that train this matching function at each student prefix; Section 3.4 then replaces page tokens with centroids for compact retrieval.

A.4 Complete Supervision and Distillation Objective

Training records. To avoid penalizing answer-bearing pages, GLM-5V-Turbo grades mined candidates by answer completeness: full-evidence pages augment the registered positives, partial-evidence pages are excluded from their negative slots, and no-evidence pages remain negatives. The corpus contains 775,635 original query groups, including 179,295 records without evidence judgments; unjudged candidates retain their original negative status. Before positive expansion, 223,877 of 1,551,270 candidate slots (14.4%) are masked, and 175,137 query groups have multiple positives.

Let $\mathcal{P}_g$ be the positive-page set of group g. The all-positive training view expands each group into one record i per positive $p_i \in \mathcal{P}_{g(i)}$, yielding 1,050,270 records with weights $\omega_i = 1/|\mathcal{P}_{g(i)}|$. Each group's nominal weights therefore sum to one. Each record also carries two negative slots n_{ik} and binary validity masks m_{ik}; full- and partial-evidence slots have $m_{ik} = 0$. The expanded view contains 462,897 masked slots out of 2,100,540. Group and page identities accompany these records to prevent known positives from becoming competing negatives. These page-level judgments supervise relevance without providing patch-level localization labels. The counts describe the judge's decisions; Section 4.3 evaluates their agreement with human labels and the retrieval gains from each governance step.

Retrieval losses and teacher branches. At an active width d, abbreviate s_d as s and write $s_{ij} = s(q_i, p_j)$, $s_i^+ = s(q_i, p_i)$, and $\Delta_{ik} = s_i^+ - s(q_i, n_{ik})$. Retrieval supervision combines masked

[1] Teacher losses estimate query length from nonzero token outputs; student training obtains it directly from the attention mask.

negative ranking with in-batch cross-entropy:

$$\ell_{\mathrm{hard},i} = \frac{\sum_k m_{ik}\,\mathrm{softplus}(-\Delta_{ik}/\tau)}{\max(1, \sum_k m_{ik})},$$
$$\ell_{\mathrm{CE},i} = -\log \frac{\exp(s_i^+/\tau)}{\sum_{j\in\mathcal{C}_i}\exp(\widetilde{s}_{ij}/\tau)}, \qquad \ell_{\mathrm{ret},i} = \tfrac{1}{2}(\ell_{\mathrm{hard},i} + \ell_{\mathrm{CE},i}). \tag{10}$$

Here $\tau = 0.02$, and $\mathcal{C}_i$ contains permitted entries from the gathered positive-page batch, including p_i. To attenuate possible false negatives, $\widetilde{s}_{ij} = 0.5 s_{ij}$ for non-target entries satisfying $s_{ij} > 0.95 s_i^+$; other scores are unchanged. Empty valid-negative sets contribute zero ranking loss. For weighted records, each rank reduces losses as $\mathcal{R}_\omega(\ell) = \sum_{i\in\mathcal{B}_r}\omega_i\ell_i / \sum_{i\in\mathcal{B}_r}\omega_i$, followed by DDP gradient averaging.

Two 4096-dimensional teacher branches share initialization and retrieval supervision. The sampled-positive branch draws one page uniformly from $\mathcal{P}_g$ per record, averaging ranking loss over all valid slots on each rank and CE over records. The all-positive branch uses $\mathcal{R}_\omega(\ell_{\mathrm{ret}})$ and excludes other records of the same query group from $\mathcal{C}_i$. After training, their full model parameters are combined as $\theta_T = 0.4\theta_{\mathrm{sampled}} + 0.6\theta_{\mathrm{all}}$ and frozen for student supervision.

Student objective. The student uses the all-positive view. At each width $d \in \mathcal{D}$, retrieval labels establish relevance, teacher distributions transfer candidate rankings, and teacher margins constrain difficult positive–negative pairs. The complete objective is

$$\mathcal{L}_S = \overline{\mathcal{L}}_{\mathrm{ret}} + a_t\left(\overline{\mathcal{L}}_{\mathrm{rel}} + 0.25\overline{\mathcal{L}}_{\mathrm{mar}} + 0.25\mathcal{L}_{\mathrm{anc}}^{(128)}\right), \tag{11}$$

where $\overline{\mathcal{L}}_x = |\mathcal{D}|^{-1}\sum_{d\in\mathcal{D}}\mathcal{R}_\omega(\ell_x^{(d)})$ gives equal weight to all six prefixes, and a_t calibrates the combined distillation strength.

Candidate distributions and masks. Query-to-page distillation uses gathered positive pages and explicit negative slots from all ranks. Page-to-query distillation normalizes the same positive-page score matrix over gathered queries. In that matrix, non-target entries sharing a query group or matching a positive in $\mathcal{P}_{g(i)}$ are masked for student CE and both KL directions. The expanded query-to-page pool additionally masks known positives and invalid negative slots. Teacher and student use identical masks; partial-evidence exclusions are limited to annotated slots. Softmax of remaining MaxSim scores at $\tau_{\mathrm{kd}} = 0.13$ defines P_T^ρ and $P_S^{(d),\rho}$ for $\rho \in \{q\to p, p\to q\}$, conditioned on q_i or p_i, respectively.

Ranking transfer and anchor preservation. The two directions receive equal weight:

$$\ell_{\mathrm{rel},i}^{(d)} = \tfrac{1}{2}\sum_\rho c_i^\rho\,\mathrm{KL}\Big(P_{T,i}^\rho \parallel P_{S,i}^{(d),\rho}\Big). \tag{12}$$

Symmetry refers to these retrieval directions: both KL terms run from teacher to student and use the same asymmetric MaxSim score function. The scale a_t controls distillation strength independently of temperature. For V valid candidates, confidence is $c = 0.1 + 0.9[1 - H(P_T)/\log\max(2, V)]$, where H denotes entropy; c is multiplied by 0.25 when the teacher's top-ranked entry is not the designated target. This attenuates uncertain or target-inconsistent supervision. Explicit negatives additionally transfer score margins through

$$\ell_{\mathrm{mar},i}^{(d)} = \frac{\sum_k m_{ik}\,\sigma(\Delta_{T,ik}/\tau_{\mathrm{kd}})\,\mathrm{smoothL1}(\Delta_{S,ik}^{(d)}, \Delta_{T,ik})}{\max(1, \sum_k m_{ik})}, \tag{13}$$

where σ is the sigmoid function and Smooth-L1 uses unit transition scale. To retain the initial d128 geometry, $\mathcal{L}_{\mathrm{anc}}^{(128)}$ applies the weighted bidirectional KL of Eq. (12) with $c = 1$, using the frozen previous EVIE-4.5B checkpoint as target and only in-batch positive pages. This anchor is evaluated in a separate no-gradient pass with the student's adapters disabled.

Algorithm 1 EVIE-4.5B Prefix-wise Training Step

Require: Records $(q_i, p_i, \{n_{ik}, m_{ik}\}_k, \omega_i, g(i), \mathcal{P}_{g(i)})$; frozen teacher T; student S; widths $\mathcal{D}$

1: Gather group/page identities; construct CE and row/column KL masks
2: Encode queries, positives, and negative slots once through the student backbone: H_q, H_p, H_n
3: Without gradients, score the same inputs with T and query/positive inputs with the adapter-disabled d128 anchor
4: **for all** $d \in \mathcal{D}$ **do**
5: Project each token matrix H: $E^{(d)} \leftarrow \mathrm{diag}(a)\,\mathrm{RowL2}(HW_{:d}^{\top} + \mathbf{1}b_{:d}^{\top})$; compute gathered MaxSim scores
6: Apply masks and compute $\mathcal{R}_\omega(\ell_{\mathrm{ret}}^{(d)}), \mathcal{R}_\omega(\ell_{\mathrm{rel}}^{(d)}), \mathcal{R}_\omega(\ell_{\mathrm{mar}}^{(d)})$
7: **end for**
8: Average each loss across widths; compute $\mathcal{L}_{\mathrm{anc}}^{(128)}$ with the same weighted reduction
9: Calibrate a_t during warmup; combine losses by Eq. (11)
10: Backward and update student parameters, respecting the projection warmup schedule

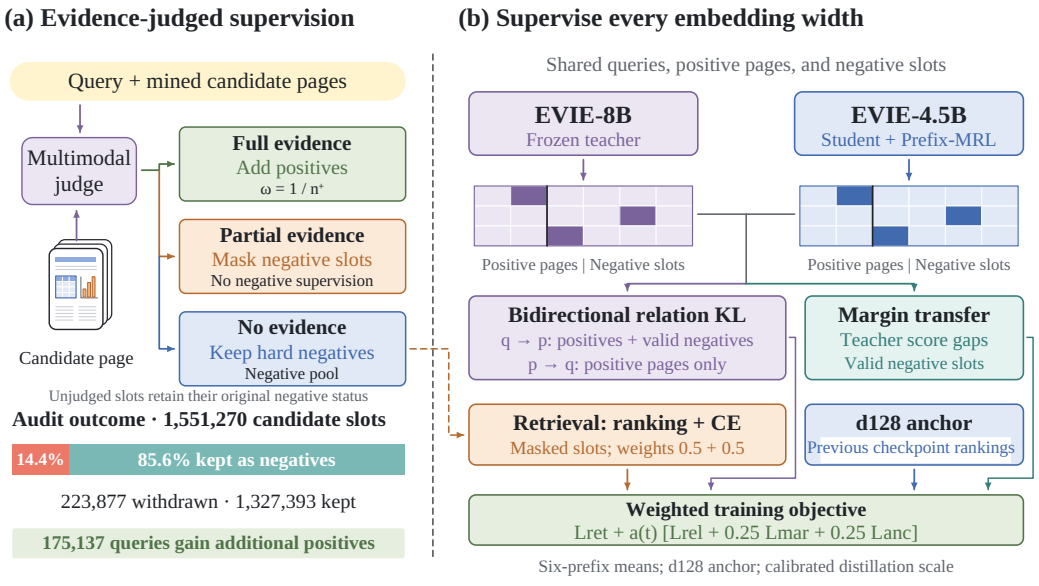

Figure 4: Evidence supervision and prefix-wise ranking transfer. The student learns retrieval labels, teacher ranking distributions, and positive–negative margins at every width; the frozen previous checkpoint anchors d128.

A.5 Optimization Details

The student initializes from the previous EVIE-4.5B checkpoint, retaining its first 128 projection weight rows and bias entries and initializing the additional channels. Only added channels update for the first 100 optimizer steps; language-backbone LoRA and the full projection then train jointly. During these 100 steps, a_t is calibrated every ten steps using query/positive d128 representation gradients: the target is $0.5\|\nabla\mathcal{L}_{\text{ret}}\|/\|\nabla\mathcal{L}_{\text{aux}}\|$, clipped to $[0.05, 20]$, where $\mathcal{L}_{\text{aux}}$ is the bracketed term in Eq. (11). Gradient norms aggregate squared magnitudes across ranks; the first target initializes a_t, subsequent targets use EMA decay 0.9, and the resulting scale is held fixed thereafter.

Each teacher branch and the student train for one epoch with an effective optimization batch of 512; the gathered contrastive batch is 64. Each student micro-batch is drawn from one data source. AdamW uses weight decay 0.02, cosine decay with 8% learning-rate warmup, and peak rates 4.57×10^{-5} for the teacher and 1.5×10^{-5} for the student. The visual token budget is 1024; LoRA uses rank 32, $\alpha = 128$, and dropout 0.197. Figure 4 summarizes training on uncompressed tokens at each prefix. HAC subsequently reduces the stored page vectors while retaining the query representation and MaxSim aggregation.

B HAC Specification and Storage Accounting

The positional feature is used only to assign clusters. Neither the stored centroids nor the query embeddings include the added positional feature.

B.1 Position-Guided Clustering and Semantic Centroids

HAC extends hierarchical mean token pooling in ColPali (Faysse et al., 2025) with page-grid positional features for cluster assignment at the selected prefix width. Whereas ColParse (Yan et al., 2026) encodes parsed sub-images and combines their vectors with a global-page representation, HAC pools existing page tokens using the processor grid, without region parsing or re-encoding.

HAC takes the image-token rows of $E_p^{(d)}$ as $v_1, \ldots, v_N \in \mathbb{R}^d$ in raster order, for $d \in \mathcal{D}$; raw evaluation additionally retains non-image context tokens. For a single-image grid $(1, g_h, g_w)$, the visual encoder's 2×2 merge yields $H = g_h/2$, $W = g_w/2$, and $N = HW$ image tokens. In raster order, token i at row r and column c has coordinates $x_i = c/\max(W-1, 1)$ and $y_i = r/\max(H-1, 1)$, with zero-based indices.

To condition assignments on page position, we use $m = d/4$ frequencies per coordinate, $\gamma_k = 10000^{k/(m-1)}$ for $k = 0, \ldots, m-1$. The four sinusoidal blocks are concatenated and normalized:

$$\begin{aligned} \widetilde{p}_i &= \left[(\sin(\gamma_k x_i))_k;\ (\cos(\gamma_k x_i))_k;\ (\sin(\gamma_k y_i))_k;\ (\cos(\gamma_k y_i))_k\right]^\top, \\ p_i &= \mathrm{L2}(\widetilde{p}_i) \in \mathbb{R}^d. \end{aligned} \tag{14}$$

The clustering feature combines unit-norm semantic and positional vectors:

$$z_i = \mathrm{L2}\left((1-w_{\text{pos}})\mathrm{L2}(v_i) + w_{\text{pos}} p_i\right), \tag{15}$$

We fix $w_{\text{pos}} = 0.1$. Section 4.3 compares this setting with semantic-only clustering on the final student. Euclidean Ward linkage (Ward, 1963) repeatedly merges the cluster pair with the smallest increase in within-cluster squared error over z_i, stopping at $K' = \min(K, N)$ clusters. This objective does not optimize MaxSim directly. The positional encoding biases assignments without imposing spatial connectivity or a monotonic page-distance penalty. Cluster membership determines which original retrieval vectors enter each centroid:

$$\mu_c = \mathrm{L2}\left(\frac{1}{|c|}\sum_{i\in c} v_i\right). \tag{16}$$

No additional positional feature is averaged into μ_c. Queries use the same prefix width, remain unclustered, and score directly against these centroids by MaxSim. Thus the query representation and scoring rule are unchanged, although pooling can alter the best-matching vector and its score. The raw-to-HAC comparison also includes removal of the context tokens.

Assignments are recomputed for each reported (d, K) configuration. For the configurations in Table 13, the $d = 64$ and $d = 128$ runs encode separately, while the two $d = 64$ budgets share one token export. Prefix-MRL can derive narrower token representations from cached full-width projections without another backbone pass, but changing width still requires reclustering because both token geometry and the positional frequency grid depend on d. Algorithm 2 and Figure 5 specify this indexing procedure.

Algorithm 2 HAC Indexing at a Selected Prefix Width

Require: Projected page sequence U, image-token indices $\mathcal{I}$, merged image grid (H, W)
Require: Prefix width $d \in \mathcal{D}$, centroid budget K, positional weight $w_{\text{pos}} = 0.1$
Ensure: $K' = \min(K, |\mathcal{I}|)$ semantic centroids of width d
1: Select image tokens and normalize prefixes: $V \leftarrow [\text{L2}(U_{i,:d})]_{i \in \mathcal{I}}$
2: Set $N \leftarrow |\mathcal{I}|$; verify $N = HW$; $K' \leftarrow \min(K, N)$
3: **if** $N \leq K$ **then**
4: $\quad \mathcal{C} \leftarrow \{\{1\}, \ldots, \{N\}\}$
5: **else**
6: $\quad$ Construct raster coordinates and normalized p_i by Eq. (14)
7: $\quad$ Form $Z = [z_i]_i$ by Eq. (15); $\mathcal{C} \leftarrow \text{Ward}(Z, K')$
8: **end if**
9: Compute each μ_c from V by Eq. (16) in FP32
10: **return** $\text{FP16}([\mu_c]_{c \in \mathcal{C}})$

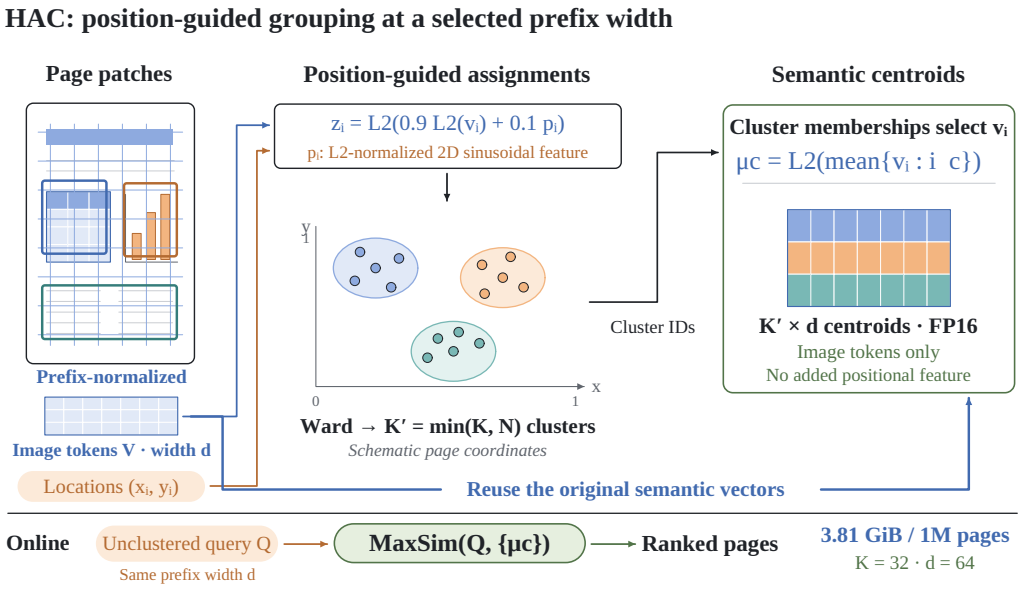

Figure 5: HAC position-guided assignment and semantic storage. Spatial features affect Ward assignments, while stored centroids average only the original retrieval vectors for direct MaxSim.

B.2 Index Storage Accounting

Storage estimates count the embedding payload, excluding document IDs, offsets, serialization overhead, and model weights. The exported vectors use FP16, or two bytes per component; all storage estimates use 2^{30}-byte units (GiB). For P pages with K vectors of width d,

$$S_{\text{payload}}\ (\text{GiB}) = \frac{2PKd}{2^{30}}. \tag{17}$$

Every page in the reported HAC runs has $N \geq K$ and therefore contributes exactly K centroids. The primary configuration ($K = 32, d = 64$) uses 4 KiB per page, or 3.81 GiB per million pages: a 128× reduction relative to the configured 1024×256 maximum visual-token payload of EVIE in Table 8. Width d and centroid count K are independently selectable indexing budgets whose costs

multiply. Consequently, $(d = 128, K = 32)$ and $(d = 64, K = 64)$ both use 7.63 GiB per million pages; Section 4 compares their retrieval accuracy.

Table 8: Embedding payload per million pages at 16-bit precision. Uncompressed EVIE rows use the configured maximum of 1024 image vectors per page (`max_num_visual_tokens`). Payload compression is relative to the first row's maximum visual-token payload. The primary HAC configuration is bold.

Configuration	Vectors / Page	Dim (d)	GiB / 1M Pages	Compression
EVIE-4.5B Raw Max-Visual ($d = 256$)	1024	256	488.28	1.0×
ColPali 1024-Vector Setting (Faysse et al., 2025)	1024	128	244.14	2.0×
EVIE-4.5B Raw Max-Visual ($d = 128$)	1024	128	244.14	2.0×
EVIE-4.5B Raw Max-Visual ($d = 64$)	1024	64	122.07	4.0×
HAC ($d = 128, K = 32$)	32	128	7.63	64.0×
HAC ($d = 64, K = 64$)	64	64	7.63	64.0×
HAC ($d = 64, K = 32$) [Ours]	**32**	**64**	**3.81**	**128.0×**

Positional-weight provenance. The weight $w_{\text{pos}} = 0.1$ was fixed from a sweep over $\{0, 0.1, \ldots, 0.9\}$ using an earlier student checkpoint at $d = 128, K = 32$, then applied unchanged to the final student. Table 6 evaluates this fixed setting without retuning it for the final checkpoint.

C Dataset Provenance and Metric Conventions

Dataset snapshots. Table 9 identifies the 22 ViDoRe repositories used by the local evaluator: ten V1 datasets, four V2 datasets, and eight public V3 domain datasets. The V3 exports provide six query languages per domain; the evaluator shares the English corpus across those language tasks. All listed configurations use the `test` split. Revisions were recovered from the per-file Hugging Face download metadata accompanying the local snapshots, rather than from the repositories' current default branches. These metadata establish snapshot provenance; they do not replace an audit of the historical evaluator and run artifacts.

Table 9: ViDoRe dataset sources and archived revisions. Every repository is under the `vidore/` namespace. Abbreviated commit identifiers link to the full immutable revision. V1 uses the default data configuration; V2 uses `corpus`, `queries`, and `qrels`; V3 uses `english-corpus` and the six language-specific query/qrels configurations.

Suite	Dataset repository	Revision
V1	`arxivqa_test_subsampled`	`b8a106812c86`
V1	`docvqa_test_subsampled`	`49bf8f13e13c`
V1	`infovqa_test_subsampled`	`f793e830aaea`
V1	`shiftproject_test`	`6e6223b3839a`
V1	`syntheticDocQA_artificial_intelligence_test`	`5694fec64c57`
V1	`syntheticDocQA_energy_test`	`a2f2f358463b`
V1	`syntheticDocQA_government_reports_test`	`91cf66572d89`
V1	`syntheticDocQA_healthcare_industry_test`	`d973f3da3e60`
V1	`tabfquad_test_subsampled`	`16c8e633612f`
V1	`tatdqa_test`	`b46fea43695e`
V2	`biomedical_lectures_v2`	`c4754665734e`
V2	`economics_reports_v2`	`76fe40166ba0`
V2	`esg_reports_v2`	`87538b12b20b`
V2	`esg_reports_human_labeled_v2`	`5a338c329bf1`
V3	`vidore_v3_computer_science_mteb_format`	`fb7fb69f81f7`
V3	`vidore_v3_energy_mteb_format`	`84fca99e5978`
V3	`vidore_v3_finance_en_mteb_format`	`fa78cb14152b`
V3	`vidore_v3_finance_fr_mteb_format`	`8a2adfda85a7`
V3	`vidore_v3_hr_mteb_format`	`bc7d43d64815`
V3	`vidore_v3_industrial_mteb_format`	`f732b725cf4a`
V3	`vidore_v3_pharmaceuticals_mteb_format`	`237ed4f43c7f`
V3	`vidore_v3_physics_mteb_format`	`2c18ef90ab3e`

Aggregation and released scores. V3 comprises 48 public domain–language tasks over eight domains, six languages each, with 14,514 queries and 115,512 language-expanded document evalu-

ations over more than 19,000 unique page images. We average over languages within domains and then over domains. These repeated corpora and translated queries are not independent replications. External V3 values retain their official release protocols; their scores were not recomputed for the main comparison. Matching public-domain coverage does not by itself establish identical evaluator implementations. In particular, local nDCG uses the exponential-gain convention detailed in Appendix E. On the same WeMM-9B V3 rankings, the two nDCG calculation paths produce 57.41 and 57.42, respectively, a difference of 0.01 points. Thus the metric convention has negligible effect in this checked case, while the external model comparison remains a comparison of reported results because the complete release evaluators were not rerun locally. Matched local controls provide the evidence about individual EVIE components.

Secondary metric interpretation. V1 and JinaVDR have one relevant page per query; V2 and V3 can have multiple relevant pages. Recall measures coverage, whereas MRR measures the first relevant hit. With binary relevant-page sets, a query with r_q relevant pages has a Recall@1 ceiling of $1/r_q$. A suite ceiling requires averaging these values with the same evaluation weights; the teacher's observed V2 Recall@1 of 37.87 is not itself a ceiling.

D Extended Retrieval Results and Operating Points

D.1 All Compared V3 Configurations

Table 10 preserves the full comparison on the eight public V3 domains. Single-vector and HAC configurations use fixed vector counts; raw multi-vector rows use configured or published maxima. Adaptive tokenization can produce smaller page-dependent counts. For example, published examples contain 1,251 vectors for tomoro-colqwen3 and 755 for colqwen3.5. The table does not measure serialized index size or resident memory.

Table 10: ViDoRe V3 nDCG@10 on eight public domains and embedding payload. K is fixed for single-vector and HAC configurations and denotes the maximum visual-token count for raw multi-vector configurations. Payloads assume two-byte components and exclude index overhead.

Model	Params	Reported K	d	GiB / 1M	V3
Single-vector baselines (fixed K)					
WeMM-2B (Zhou et al., 2026)	2.1B	1	2048	3.81	52.59
Qwen3-VL-Embedding-2B (Li et al., 2026)	2.1B	1	2048	3.81	52.89
WeMM-4B (Zhou et al., 2026)	4.1B	1	2560	4.77	54.70
WeMM-9B (Zhou et al., 2026)	9.0B	1	4096	7.63	57.42
Multi-vector baselines (published $K_{\max}$)					
VultronRetrieverFlash-0.8B (Vultr, 2026b)	0.85B	1792	320	1068.12	56.49
tomoro-colqwen3-embed-4b (TomoroAI, 2026a)	4.0B	1280	320	762.94	60.16
nemotron-colembed-vl-4b-v2 (NVIDIA, 2026a)	4.8B	2304	2560	10986.33	61.42
colqwen3.5-4.5B-v3 (athrael-soju, 2026)	4.6B	768	320	457.76	61.56
tomoro-colqwen3-embed-8b (TomoroAI, 2026b)	8.0B	1280	320	762.94	61.60
nemotron-colembed-vl-8b-v2 (NVIDIA, 2026b)	8.8B	2304	4096	17578.13	63.54
VultronRetrieverCore-4.5B (Vultr, 2026a)	4.5B	1792	320	1068.12	63.72
webAI-ColVec1.1-4b (webAI, 2026a)	4.5B	1792	640	2136.23	64.24
VultronRetrieverPrime-8B (Vultr, 2026c)	8.4B	1792	320	1068.12	64.72
webAI-ColVec1.1-8b (webAI, 2026b)	8.4B	1792	640	2136.23	65.32
EVIE HAC measurements (fixed K)					
HAC ($d = 64, K = 32$)	4.5B	32	64	3.81	59.58
HAC ($d = 64, K = 64$)	4.5B	64	64	7.63	62.06
HAC ($d = 128, K = 32$)	4.5B	32	128	7.63	61.40
EVIE raw scores (configured image-token $K_{\max} = 1024$)					
Previous EVIE-4.5B checkpoint	4.5B	1024	128	244.14	65.36
EVIE-4.5B ($d = 256$)	4.5B	1024	256	488.28	65.68
EVIE-4.5B ($d = 2048$)	4.5B	1024	2048	3906.25	66.02
EVIE-8B (teacher)	8.4B	1024	4096	7812.50	**66.75**

D.2 Domain-level Results

The teacher–student V3 gap ranges from 0.14 to 1.50 points across domains. Relative to the full-width student, HAC at $d = 64, K = 32$ loses 1.89 points on Physics and 9.51 on Human Resources.

Table 11: ViDoRe V3 nDCG@10 on the eight public domains, macro-averaged over six languages per domain. HAC uses the final EVIE-4.5B checkpoint at $d = 64$ and $K = 32$.

Domain	Teacher	Student ($d = 2048$)	HAC ($d = 64, K = 32$)
Computer Science	81.86	81.72	76.26
Energy	72.51	72.32	67.11
Finance (EN corpus)	71.23	70.00	61.47
Finance (FR corpus)	56.40	54.90	46.21
Human Resources	69.29	67.82	58.31
Industrial	59.77	59.40	50.66
Pharmaceuticals	70.81	70.27	66.82
Physics	52.11	51.69	49.80
Overall	66.75	66.02	59.58

These differences show that compression sensitivity varies across domains; the aggregate score alone does not identify which local evidence is lost.

D.3 Six Prefix Widths and Three HAC Configurations

Table 12: Prefix-MRL results, all nDCG@10, from one student checkpoint. Maximum visual-token payload uses the configured EVIE value `max_num_visual_tokens`=1024 at two bytes per component. Nested prefix extraction can reuse cached full-width projections.

d	GiB / 1M	V1	V2	V3	JinaVDR	Avg4
64	122.07	92.16	73.18	64.51	81.00	77.71
128	244.14	92.28	73.37	65.27	81.83	78.18
256	488.28	92.20	73.83	65.68	82.09	78.45
512	976.56	92.38	74.53	65.90	82.27	78.77
1024	1953.13	92.39	74.53	65.94	82.43	78.82
2048	3906.25	92.53	74.91	66.02	82.48	78.98

Table 13: Raw and HAC operating points, all nDCG@10. Raw payloads use the configured EVIE maximum of 1024 image tokens; HAC payloads use the fixed centroid count in the reported runs. Rows at the same d distinguish width changes from the subsequent compression procedure.

Configuration	GiB / 1M	V1	V2	V3	JinaVDR	Avg4
Raw ($d = 2048$)	3906.25	92.53	74.91	66.02	82.48	78.98
Raw ($d = 128$)	244.14	92.28	73.37	65.27	81.83	78.18
Raw ($d = 64$)	122.07	92.16	73.18	64.51	81.00	77.71
HAC ($d = 128, K = 32$)	7.63	90.00	69.45	61.40	78.64	74.87
HAC ($d = 64, K = 64$)	7.63	90.63	70.71	62.06	78.90	75.57
HAC ($d = 64, K = 32$)	3.81	89.27	67.83	59.58	77.21	73.47

V3 and Avg4 increase monotonically across the six evaluated widths, although individual suites need not do so. At $d = 128$, the student's V3 score of 65.27 is 0.09 below the previous EVIE-4.5B checkpoint. That comparison measures retention at the initialization width and does not isolate a training component. The full-width-to-HAC losses at $d = 64, K = 32$ are 3.26, 7.08, 6.44, and 5.27 points on V1, V2, V3, and JinaVDR, respectively.

At 7.63 GiB per million pages, $d = 64, K = 64$ exceeds $d = 128, K = 32$ by 0.63, 1.26, 0.66, and 0.26 points across the four suites, and by 0.70 Avg4 points. Figure 6 complements the expanded fixed-budget comparison in Table 7.

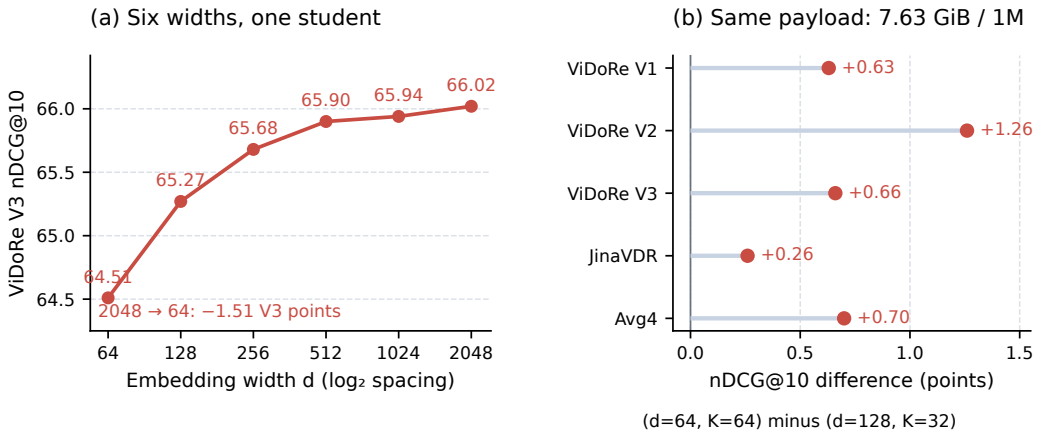

Figure 6: Width scaling and the equal-payload comparison. Left: final-student V3 scores at six raw widths. Right: suite-level differences between $d = 64, K = 64$ and $d = 128, K = 32$ at 7.63 GiB per million pages. The panels use the measured values in Tables 12 and 13.

E Detailed Multi-Metric Results

Tables 14–16 report nDCG, Recall, MAP, and MRR at cutoffs 1, 5, and 10 for three representative EVIE configurations. The V3 columns use the same 48 public domain–language tasks as Section 4. For ViDoRe V1 and JinaVDR, each query has one binary relevant page. For ViDoRe V2 and V3, qrel entries with gain $g > 0$ are relevant for Recall, MAP, and MRR, while nDCG retains the graded gains using $2^g - 1$ with logarithmic rank discount. At cutoff k, MAP is normalized by $\min(k, |R_q|)$. Metrics are averaged over queries within each task and macro-averaged over tasks within each suite; Avg4 is the unweighted mean of the four suite scores. All scores are reported on a 0–100 scale. MAP and MRR coincide at $k = 1$ for all suites and at every cutoff for single-relevant-page tasks.

Table 14: EVIE-8B retrieval results across 138 tasks. Metrics are reported at cutoffs 1, 5, and 10 on each suite; Avg4 weights the four suites equally.

Metric	ViDoRe V1	ViDoRe V2	ViDoRe V3	JinaVDR	Avg4
nDCG@1	88.51	75.58	62.14	75.50	75.43
nDCG@5	92.18	74.23	64.16	82.45	78.26
nDCG@10	92.64	75.35	66.75	83.30	79.51
Recall@1	88.51	37.87	30.79	75.50	58.17
Recall@5	95.07	66.65	58.69	88.07	77.12
Recall@10	96.43	76.34	70.10	90.65	83.38
MAP@1	88.51	75.99	65.94	75.50	76.49
MAP@5	91.21	67.30	55.82	80.56	73.72
MAP@10	91.40	66.81	56.32	80.92	73.86
MRR@1	88.51	75.99	65.94	75.50	76.49
MRR@5	91.21	82.85	75.39	80.56	82.50
MRR@10	91.40	83.23	76.09	80.92	82.91

Table 15: EVIE-4.5B retrieval results at $d = 2048$ across 138 tasks. Metrics are reported at cutoffs 1, 5, and 10; Avg4 weights the four suites equally.

Metric	ViDoRe V1	ViDoRe V2	ViDoRe V3	JinaVDR	Avg4
nDCG@1	88.35	72.76	61.16	74.26	74.13
nDCG@5	92.07	73.38	63.39	81.57	77.60
nDCG@10	92.53	74.91	66.02	82.48	78.98
Recall@1	88.35	36.73	30.14	74.26	57.37
Recall@5	94.95	65.59	58.06	87.39	76.50
Recall@10	96.36	76.56	69.67	90.17	83.19
MAP@1	88.35	73.18	65.03	74.26	75.20
MAP@5	91.09	66.52	55.02	79.61	73.06
MAP@10	91.29	66.25	55.50	79.99	73.25
MRR@1	88.35	73.18	65.03	74.26	75.20
MRR@5	91.09	81.66	74.76	79.61	81.78
MRR@10	91.29	82.11	75.43	79.99	82.20

Table 16: HAC retrieval results at $d = 64, K = 32$ across 138 tasks. Metrics are reported at cutoffs 1, 5, and 10; Avg4 weights the four suites equally.

Metric	ViDoRe V1	ViDoRe V2	ViDoRe V3	JinaVDR	Avg4
nDCG@1	82.95	65.82	54.29	67.50	67.64
nDCG@5	88.74	66.19	56.98	75.89	71.95
nDCG@10	89.27	67.83	59.58	77.21	73.47
Recall@1	82.95	31.16	26.35	67.50	51.99
Recall@5	93.30	60.17	52.82	82.84	72.28
Recall@10	94.94	72.01	64.00	86.90	79.46
MAP@1	82.95	65.82	58.26	67.50	68.63
MAP@5	87.20	58.45	48.34	73.56	66.89
MAP@10	87.42	57.65	48.60	74.11	66.95
MRR@1	82.95	65.82	58.26	67.50	68.63
MRR@5	87.20	75.25	68.89	73.56	76.22
MRR@10	87.42	76.12	69.73	74.11	76.84

F Additional Controlled Comparisons

F.1 Human Audit of Evidence Judgments

The audit contains 900 candidates sampled equally from the three judge tiers. Three volunteer annotators independently assign every candidate to one of the same three answer-completeness categories used by the judge. A page receives a full-evidence label when it independently supports a complete answer, a partial-evidence label when it contains useful but insufficient answer support, and a no-evidence label when it does not support the answer. The majority vote of the three annotations defines the human label used in Table 3(b). Each table cell is the corresponding majority-vote count divided by the 300 candidates sampled from that judge tier. The stratified design supports tier-conditional agreement estimates; prevalence-weighted accuracy over the mined-candidate pool would additionally require its original tier proportions. The lower agreement for partial evidence motivates excluding those candidates from their annotated negative slots.

F.2 Teacher Construction, Encoder Transfer, and Compression Sensitivity

Teacher construction. The final teacher combines the sampled-positive and all-positive branches through parameter averaging, $\theta_T = 0.4\theta_{\text{sampled}} + 0.6\theta_{\text{all}}$. The merged checkpoint is frozen for student supervision; Table 4(a) reports the branch and mixture results. At the evaluated $0.4/0.6$ weighting, the mixture exceeds the stronger branch by 0.42 V3 points and 0.34 Avg4 points. This single mixture comparison does not establish the optimum weights or an improvement averaged over training seeds.

Compression transfer across encoders. Table 6(b) compares semantic-only Ward and position-guided HAC on EVIE-4.5B and two external multi-vector encoders. Spatial coordinates follow each encoder's native processor grid.

HAC improves over Ward by 1.46, 1.28, and 0.78 points on EVIE, Jina, and Nemotron, respectively, with raw-to-HAC losses of 4.81, 5.50, and 2.86 points. The positional gain extends to both external encoders at this budget.

Governance and compression sensitivity. Table 4(b) compares original and governed student supervision at $d = 64$. Compression sensitivity is measured by the within-model score loss from raw image-only retrieval to HAC.

Full governance raises raw and HAC scores by 3.01 and 4.78 points, respectively. Compression loss falls from 6.58 to 4.81 points, giving an interaction estimate of $6.58 - 4.81 = 1.77$ points. The larger gain after compression is consistent with reduced sensitivity at this budget; without uncertainty intervals, the interaction remains descriptive.

G Serving Measurements and Evidence Cases

G.1 Measurement Setup and Scope

The serving experiment uses 1000 stratified V3 pages (seed 42, 125 pages per public domain), all 14,514 V3 queries, one NVIDIA H20, and a maximum visual-token budget of 1024. Latency uses batch size 1; throughput uses batch size 16. Search scores already-encoded query vectors. The four EVIE configurations share one d64 encoding. The cascade first retrieves the HAC $K = 32$ top 100 and reranks them with raw d64 MaxSim; its resident index includes both representations.

Table 17: 1000 stratified ViDoRe V3 pages, 14,514 queries, one NVIDIA H20, MVT=1024. Enc. and Search are medians; Search scores already-encoded query vectors. E2E is batch 1; QPS is end-to-end at batch 16. Cascade: HAC K=32 top-100 then raw d=64 MaxSim. Mem.: GPU-resident index (GiB).

System	Enc. (ms)	Search (ms)	E2E p50	E2E p95	QPS	Mem.
WeMM-9B	106.78	**0.125**	106.90	110.29	96.45	**0.0076**
EVIE raw (d=64)	**99.67**	0.909	100.60	104.35	105.26	0.2444
EVIE HAC (d=64, K=32)	**99.67**	0.160	99.83	103.62	**115.73**	**0.0076**
EVIE HAC (d=64, K=64)	**99.67**	0.150	**99.82**	**103.60**	115.65	0.0153
EVIE cascade (K=32 top-100)	**99.67**	0.335	100.01	103.79	113.27	0.2520

The query encoder dominates end-to-end latency at this corpus size. GPU-resident memory in Table 17 is a measured serving quantity and is distinct from the two-byte exported vector payload used elsewhere. These measurements cannot be extrapolated to million-page search latency, index-construction time, or clustering throughput; those quantities were not measured here.

G.2 Retained and Lost Evidence

On the 1000-page index, 3,588 queries have a relevant page. Of these, 2,521 retain a relevant page in the top three under both raw d64 and HAC $K = 32$, while 49 have raw rank ≤ 3 and HAC rank > 10. Figure 7 shows one case of each type: case A ranks first under both representations, while case B changes from rank 3 to rank 37. Yellow–red indicates the upper MaxSim mass; the box identifies the HAC cluster containing the highest-scoring token. The maps visualize matching behavior and do not constitute independently annotated localization ground truth.

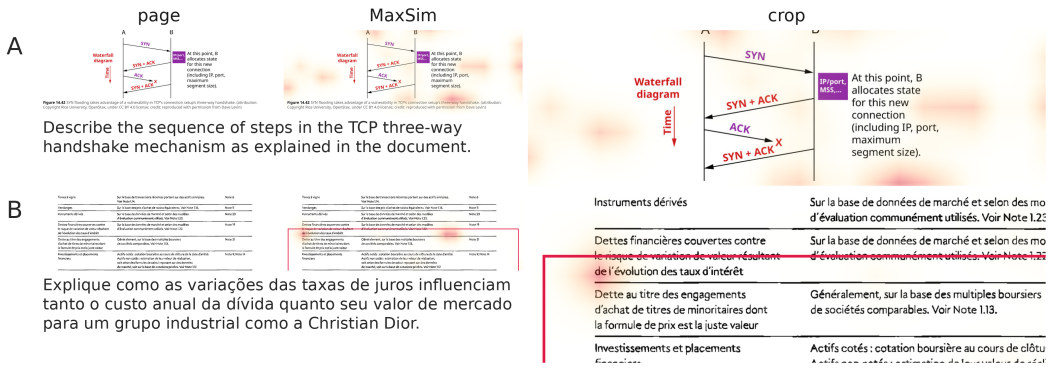

Figure 7: Retained and lost retrieval evidence under HAC. Each case shows the page, its MaxSim map, and a crop around the selected token's HAC cluster. Case A retains rank 1; case B falls from raw rank 3 to HAC rank 37. The examples illustrate that centroid pooling can retain a useful match or lose a distinction required for ranking.