LEARNING TO RETRIEVE MULTIMODAL EVIDENCE FROM THE ANSWER ALONE: LATENT EVIDENCE-SET RETRIEVAL FOR MULTIMODAL MULTI-HOP QA

Anonymous authors
Paper under double-blind review

ABSTRACT

In multimodal multi-hop question answering, a retriever selects images, passages, and tables from a question's candidate pool as evidence for a vision–language model (VLM) reader. Training such retrieval-augmented systems typically requires evidence annotations alongside questions and answers, but these annotations are costly and scarce. Learning retrieval from answers alone is difficult: answer quality does not identify which sources were useful or which work together, and pretrained retrievers favor particular modalities regardless of what the reader can use. We introduce LEAF, a two-stage framework for building multimodal RAG systems without supporting-source supervision. First, we fine-tune the reader on reference answers and retrieved contexts of varying sizes and diversity, creating a task-adapted answer generator and teacher. Second, we freeze it and use answer rewards to adapt the retriever encoder and to train a selector that chooses evidence diversity and size per question. Across Qwen and VLM2Vec retrievers and Qwen and Gemma readers, LEAF outperforms the two standard annotation-free policies, Top-K and fixed MMR, on MultiModalQA and WebQA, reaching 76.02 F1 and 53.75 Overall QA respectively. These scores exceed published evidence-supervised SKURG and Solar results on both benchmarks and MuRAG on WebQA, and our best WebQA system ranks fifth all-time by Overall QA on the official leaderboard.

1 INTRODUCTION

In multimodal multi-hop question answering, the evidence required to answer a question can be distributed across sources of different types—images, text passages, and tables—so that no single source suffices (Talmor et al., 2021). Vision–language models (VLMs) are the natural candidate for the task, since answering may require these modalities to be interpreted together. Retrieval-augmented generation provides access to evidence beyond a model's parametric knowledge (Lewis et al., 2020). In the multimodal setting, reading every candidate can be costly: images consume substantial context, sources may repeat the same information, and irrelevant candidates can distract from the evidence needed to answer. The retrieval problem is therefore not simply to collect more context, but to select a useful combination of sources for the reader.

A retrieval-augmented system pairs a *retriever*, which selects evidence, with a *reader*, which generates an answer. Pretrained multimodal encoders and VLMs provide reusable components, but combining them does not automatically align retrieval with answering. A source can be relevant to a question yet difficult for a particular reader to use: retriever and reader hold different parametric knowledge and encode the same source into different representations, so a high similarity score need not indicate evidence the reader can turn into an answer. The mismatch is systematic in the multimodal case, where a text query is scored against images, passages, and tables in a space that separates modalities (Liang et al., 2022) and therefore favors some of them. The goal is to learn which evidence helps that reader, rather than rely solely on pretrained similarity.

Prominent multimodal QA systems supervise retrieval with annotated supporting sources. SKURG trains its retrieval and stopping decisions against annotated sources (Yang et al., 2022), and Solar trains retrieval and ranking with ground-truth evidence and provenance (Yu et al., 2023). This super-

vision is expensive: it requires identifying which images, passages, or tables support each answer, rather than recording the answer alone, and it limits applicability wherever question–answer pairs exist but their supporting sources do not.

Answer feedback offers supervision tied directly to the reader's task, but useful individual sources are only part of the problem: sources can be useful alone yet redundant together, or informative only in combination. Taking the K highest-scoring candidates (Top-K) ignores redundancy altogether, and maximal marginal relevance (MMR) (Goldstein & Carbonell, 1998) balances relevance against redundancy but, in its fixed-policy form, learns neither the diversity weight nor the context size for each question and reader. Two levels of learning follow: candidate rankings from singleton rewards, and evidence-set choices from whole-context rewards.

Latent retrieval learns from downstream answers without evidence annotations (Lee et al., 2019b). We apply that principle to reader-specific multimodal evidence selection through LEAF (Latent Evidence from Answer Feedback), a two-stage procedure for aligning a multimodal retriever with a reader. The reader is first adapted to answer from retrieved contexts, then frozen, so that the quality of its generated answers against references supplies retrieval rewards. At the item level, LEAF-R uses singleton rewards to adapt candidate ordering without prescribing a modality quota. At the set level, LEAF-S compares contexts of different relevance–diversity trade-offs and sizes using the reader's pre-answer states. These are distinct decisions: singleton usefulness need not add up to set usefulness, so the selector learns from within-question whole-context reward differences rather than summed candidate utilities.

We apply LEAF separately on MultiModalQA (Talmor et al., 2021) and WebQA (Chang et al., 2022), two benchmarks with different evidence and answer structures, using readers of different scales and families and two families of universal multimodal retriever. Our primary comparison holds the reader fixed and varies only retrieval, so that any difference is attributable to the evidence selected. Against Top-K retrieval and fixed MMR, the two most widely used annotation-free policies, LEAF answers more accurately on both benchmarks, reaching 76.02 F1 on MultiModalQA and 53.75 Overall QA on WebQA. These scores also exceed the published results of SKURG and Solar on both benchmarks and MuRAG on WebQA, all of which train retrieval against annotated sources, and our best WebQA system ranks fifth all-time by Overall QA on the official leaderboard. Our contributions are:

- We provide a two-stage procedure for learning multimodal retrieval from answer feedback: fine-tune a reader on retrieved contexts, then freeze it to supervise retrieval, with no supporting-source targets and no source-based checkpoint selection.
- We propose LEAF-R, an answer-supervised adaptation of a multimodal retriever, and LEAF-S, a reader-conditioned selector that chooses evidence diversity and cardinality jointly from answer feedback, without gold supporting-source or retrieval-termination targets.
- Matched controls isolate reader fine-tuning, encoder adaptation, and set selection, showing that LEAF outperforms annotation-free retrieval on both benchmarks and surpasses published evidence-supervised systems. We report source retrieval separately from answer quality, since the two need not improve together.

2 RELATED WORK

Latent and answer-supervised retrieval. ORQA learns a retriever and reader jointly from question–answer pairs, treating evidence as a latent variable (Lee et al., 2019b). Later work models retrieved sets as latent variables trained by expectation–maximization (Sachan et al., 2021), optimizes a variational bound on the task marginal likelihood (Liévin et al., 2023), mines evidence iteratively from question–answer pairs (Zhao et al., 2021), or adapts a retriever to the preferences of a frozen language model (Shi et al., 2023). Our encoder objective follows this distillation principle with generated-answer rewards, applied to image, text, and table candidates and coupled with reader-conditioned selection over evidence-set sizes and relevance–diversity trade-offs.

Multimodal multi-hop question answering. SKURG aligns sources of different modalities through shared entities and integrates iterative retrieval with answer generation (Yang et al., 2022);

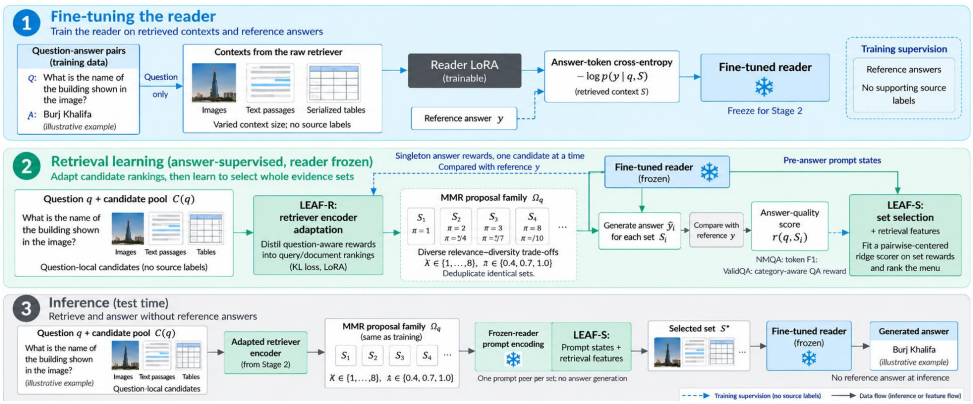

Figure 1: LEAF training and inference. Reference answers supervise reader fine-tuning and separate singleton/set reward passes for LEAF-R/LEAF-S. Inference uses pre-answer prompt states and retrieval features, never reference answers or generated-answer rewards.

Solar converts images and tables into text so that retrieval, ranking, and generation occur in one language space (Yu et al., 2023); and MuRAG augments generation from a multimodal memory (Chen et al., 2022); the VLMT preprint trains a multimodal reranker the same way (Lim et al., 2025). All train retrieval on annotated supporting sources, where we use answer rewards.

Universal multimodal retrieval and the modality gap. Universal embedding models map queries and candidates of every supported modality into one space, by contrastively training vision–language models as encoders (Jiang et al., 2024) or fine-tuning multimodal LLMs for retrieval (Lin et al., 2025). A shared space does not guarantee calibrated cross-modal rankings: contrastive models such as CLIP show geometric separation between modalities (Liang et al., 2022) and ranking bias in mixed-modality search, motivating post-hoc calibration (Li et al., 2025). Rather than imposing a modality quota, LEAF-R adapts the ordering through answer utility; whether that improves modality-specific source recovery we evaluate separately from answer quality.

Set-level and adaptive evidence selection. We use MMR's relevance–redundancy trade-off (Goldstein & Carbonell, 1998) to propose sets, not as a contribution. Set Transformer models set interactions with permutation-invariant attention (Lee et al., 2019a); our prompt-state scorer is order-sensitive. Adaptive-RAG routes among retrieval strategies by predicted question complexity (Jeong et al., 2024); LEAF-S instead compares evidence compositions and cardinalities for a fixed reader.

3 METHOD

LEAF has two source-label-free stages. Answer quality can measure a context's usefulness only if the reader uses what it is given, so the first stage fine-tunes a reader on question–answer pairs and retrieved contexts. The second freezes that reader and turns its answer quality into a training signal for retrieval: it adapts the retriever encoder so that similarity reflects usefulness, constructs a family of candidate evidence sets, and learns to select one set per question. We describe the three in turn, after fixing notation and the supervision boundary. Figure 1 shows the training and inference paths.

3.1 PROBLEM SETUP

For a question q, let $\mathcal{C}(q)$ be a pool of candidate sources containing images, text passages, and serialized tables. A retrieval policy selects an evidence set $S \subseteq \mathcal{C}(q)$, and the reader R, a VLM, generates the answer $R(q, S)$. Our task is to learn that policy from training questions paired with reference answers, and to do so for a *given* reader, since the evidence a policy should select depends on which reader must use it.

Training consumes questions, reference answers, and the content and modality of candidates. Gold supporting-source labels supply no training target, no proposal admission rule, and no checkpoint-

selection criterion; they are opened only to evaluate evidence retrieval, after answers are sealed. At inference the system sees only q and $\mathcal{C}(q)$.

3.2 Stage 1: fine-tuning the reader on retrieved contexts

A base VLM—the released checkpoint, before any fine-tuning of ours—is fine-tuned with LoRA on question–answer pairs paired with contexts retrieved by the unadapted encoder, minimizing cross-entropy on reference-answer tokens with prompt and padding positions masked. Training contexts span a range of cardinalities under a fixed relevance–diversity mixture, together with a bounded number of image-support examples; no context is admitted because it contains annotated evidence. The fine-tuned reader is then frozen; it is both the reward model for retrieval learning and the final answer generator, and its answer quality is a reader-specific, imperfect signal of evidence usefulness.

3.3 Stage 2a: answer-supervised encoder adaptation

Let E_ϕ be the retriever encoder: a single model that embeds the question and candidates of every modality into one normalized space. For each fitting question and candidate c, the frozen reader answers from the singleton context $\{c\}$, and the benchmark's answer metric scores that answer, giving a candidate utility $u_1(q, c)$. These utilities define a teacher distribution over the pool,

$$t(c \mid q) = \frac{\exp(u_1(q,c)/\tau_t)}{\sum_{c' \in \mathcal{C}(q)} \exp(u_1(q,c')/\tau_t)}, \tag{1}$$

where MMQA uses token F1. For WebQA, we first form this distribution from the locally computed category-aware Overall-QA reward, then refine exact reward ties using token F1 multiplied by normalized QA-FL. For a tie group G, its total probability is preserved and distributed within G in proportion to the composite score's softmax, also at temperature 0.2. This is a dataset-specific teacher construction, not plain Overall-QA distillation. The resulting teacher is distilled into the encoder's own similarity distribution

$$p_\phi(c \mid q) = \frac{\exp(\cos(E_\phi(q), E_\phi(c))/\tau_s)}{\sum_{c' \in \mathcal{C}(q)} \exp(\cos(E_\phi(q), E_\phi(c'))/\tau_s)} \tag{2}$$

by minimizing $\mathbb{E}_q[D_{\mathrm{KL}}(t(\cdot \mid q) \,\|\, p_\phi(\cdot \mid q))]$ with equal weight per question. Question and candidate encodings share one LoRA adaptation and the vision tower is frozen, so the objective reshapes the relative ordering of candidates rather than raising all similarities.

Conservative adapter application. Stage 2a optimizes the utility of candidates taken *individually*, whereas a multi-hop question is answered from a *set*. Driving the encoder too far toward singleton utility can therefore distort the pretrained geometry in ways that do not serve set retrieval, so we apply the learned update as $W_{\mathrm{use}} = W_{\mathrm{base}} + \gamma \Delta W_{\mathrm{LoRA}}$, which preserves more of that geometry while retaining reader-specific adaptation. This is the WiSE-FT interpolation between pretrained and fine-tuned weights (Wortsman et al., 2022), since $W_{\mathrm{base}} + \gamma \Delta W = (1-\gamma) W_{\mathrm{base}} + \gamma W_{\mathrm{tuned}}$, though applied for a different reason than robustness to distribution shift. We use $\gamma = 1$ on MMQA, which applies the adapter unchanged, and $\gamma = 0.5$ on WebQA, scaling the question and candidate branches identically; both values were set by empirical calibration.

3.4 Stage 2b: the proposal family

For each question we construct a *proposal family*: a small collection of complete evidence sets, any one of which could be given to the reader, from which the selector of Stage 2c will choose. Learning that choice requires several scored contexts per question, each costing a reader pass, so enumerating subsets of the pool is infeasible. The family must be cheap to build, use no annotation, and still contain sets the reader can answer from.

Maximal marginal relevance (MMR) (Goldstein & Carbonell, 1998) builds such a family from two terms. Its relevance term ranks candidates that match the question highly, which after Stage 2a means candidates this reader can use. Its redundancy term penalizes a candidate in proportion to its similarity to sources *already selected*, so each proposal accumulates complementary evidence instead of near-duplicates of its own first source. Every set in the family is thus relevant and internally complementary by construction, rather than an arbitrary subset.

MMR leaves two quantities to be fixed: how many sources to keep, K, and how strongly to penalize redundancy, λ. We sweep both. Questions differ in how much evidence they need—some are answered from a single source, others require several—so we retain every prefix length K; they also differ in how much redundancy is tolerable, so at each K we retain the sets produced under several values of λ. The sweep covers the length and composition of a useful set at a cost of a few contexts per question.

Concretely, greedy MMR selects the candidate maximizing

$$\lambda\, r(q, c) - (1 - \lambda) \max_{s \in S} \mathrm{sim}(c, s), \tag{3}$$

where r and sim are adapted-space cosine similarities; the first source maximizes $r(q, c)$. Sweeping $\lambda \in \{0.4, 0.7, 1.0\}$ and retaining every prefix $K \in \{1, \ldots, 8\}$ yields at most 24 proposed contexts per question. Physically identical sets are deduplicated.

3.5 Stage 2c: reader-conditioned set selection

A single scorer ranks the whole family, assigning each candidate evidence set S a score. To assess a set jointly rather than sum singleton utilities, we score it through the frozen reader itself: we take the reader's final attended state of the question-and-evidence prompt, *before* answer generation. Selection is therefore *reader-conditioned*—the same set is represented differently for different readers—rather than a function of retrieval similarity alone. These states are normalized and projected by a PCA basis fitted on training questions only, after centering states within each question so that the projection captures differences between contexts rather than between questions. The projected state is concatenated with retrieval statistics, modality counts, the cardinality K, indicators for the values of λ that produced the set, since deduplication can leave one set reachable under several, and interactions among these terms.

Let $z(q, S)$ denote these features and $u(q, S)$ the answer reward obtained by reading S. Both are centered within each question's family, and a ridge scorer is fitted with equal total weight per question and no intercept:

$$\min_{w} \frac{1}{N} \sum_{q} \frac{1}{|\mathcal{M}(q)|} \sum_{S \in \mathcal{M}(q)} \Big(w^{\top} \bar{z}(q, S) - \bar{u}(q, S) \Big)^2 + \rho \|w\|^2, \tag{4}$$

where N is the number of fitting questions, $\mathcal{M}(q)$ is the proposal family, and bars denote within-question centering. Centering matters because absolute reward levels differ across questions for reasons unrelated to evidence choice, so the scorer is asked only to rank sets within a question. At inference the policy selects $\arg\max_S w^{\top} z(q, S)$ over the whole family, across cardinalities and λ values jointly, and the reader then generates the answer from the selected set.

3.6 Inference

At test time the adapted encoder embeds the question and its candidate pool, and greedy MMR builds the proposal family over the same λ and K grid used in training, collapsing identical sets. For each distinct proposal the frozen reader encodes the question-and-evidence prompt without generating, yielding the state from which $z(q, S)$ is formed; the scorer ranks the family, and the reader then generates the answer from the highest-scoring set. Neither reference answers nor source annotations enter this path.

4 Experimental Setup

4.1 Datasets and splits

MultiModalQA (Talmor et al., 2021) pairs each question with a pool of images, text passages, and Wikipedia tables, and was built so that answering generally requires combining sources across modalities rather than within one: of its sixteen question types, thirteen call for cross-modal retrieval and reasoning. The official release provides 23,817 training and 2,441 development questions. Because the test split is not released, we train on the training split and report development results, as

SKURG (Yang et al., 2022) and Solar (Yu et al., 2023) also do. Reader fine-tuning uses all training questions. Retriever adaptation and selector fitting each draw 3,072 training questions.

WebQA (Chang et al., 2022) provides question-local pools of image and text candidates and sentence-form reference answers. The public release provides 36,766 training, 4,966 validation, and 7,540 test questions; test answers and source annotations are withheld, so test scores are obtained only by uploading predictions to the official WebQA evaluation server. Reader fine-tuning uses all training questions. Retriever adaptation uses a 512-question training panel per reader, and selector fitting uses a disjoint 3,072-question training panel. Questions with flat teacher logits are excluded from optimizer updates. We use the validation split for method development, and all reported WebQA answer and retrieval scores are those returned by the evaluation server over the full test split.

4.2 Models and preprocessing

The primary retriever is Qwen3-VL-Embedding-8B. Each reader independently adapts a fresh copy of this raw retriever from its own answer-derived rewards. We pair Qwen3-VL readers at 2B and 8B scales and Gemma 4 12B with the retriever. Reader fine-tuning additionally compares contexts supplied by the raw Qwen3-VL-Embedding-2B and 8B retrievers; retrieval learning uses the 8B retriever throughout. To test whether LEAF transfers to a second retriever family, we repeat the WebQA comparison with VLM2Vec-V2.0 and the prepared Qwen3-2B reader, reported in Appendix G.

Question and candidate embeddings are L_2 normalized, and retrieval-learning proposals use global cosine without modality quotas. Generation is deterministic greedy decoding. Appendix D gives the preprocessing, decoding, and token-cap settings, all held fixed within each dataset's paired comparisons.

4.3 Training protocol

Retriever adaptation uses $\tau_t = 0.2$ and $\tau_s = 0.1$. A rank-8, alpha-16 LoRA is fitted for three epochs with AdamW, learning rate 10^{-4}, weight decay 0.01, and batches of 16 questions. The LoRA updates shared language-attention Q/V projections used by question and candidate encoding; the vision tower is frozen.

On each dataset, selector fitting splits the 3,072 training questions into four outer folds, with fold-local feature transformations and PCA. Inner training-side reward selects regularization (three inner folds for WebQA); final scorers refit on all 3,072 questions. Neither development answers nor source labels enter fitting. All experiments run on NVIDIA A100 80GB GPUs.

4.4 Controls and ablations

With the fine-tuned reader and adapted encoder fixed, we compare global Top-K, MMR ($\lambda = 0.7$), and learned joint λ/K selection. Every within-reader comparison also holds the prompt formatter, candidate pool, image policy, table serialization, decoder, and evaluator fixed. Fixed controls use $K = 5$ on MMQA and $K = 4$ on WebQA; Table 10 reports an earlier WebQA comparison at $K = 5$ separately. A separate raw-versus-adapted fixed-MMR control isolates encoder adaptation. Reader fine-tuning compares base and fine-tuned readers under identical raw MMR contexts—$K = 5$, $\lambda = 0.7$, and at most three images on MMQA—with the same prompts and decoding before and after adaptation; question-only controls omit evidence. Appendix E details matched λ-only, K-only, reader-state removal, and best-constant selector ablations.

4.5 Metrics

We report official MMQA token F1 and exact match. After selection and answer generation are frozen, source annotations measure complete retrieval, source and modality recall, and mean K; source agreement is a diagnostic, not a training target or an exact surrogate for answer utility. Paired question-bootstrap intervals quantify key contrasts. WebQA evaluation reports Overall QA, QA-Acc, and QA-FL (Chang et al., 2022), where Overall QA averages the per-question product of category-aware answer accuracy and normalized fluency and is not the product of the two reported dataset means. Because local rewards use the available answer and category annotations while

Table 1: Main comparison. Answer scores (%) with the reader, candidate pool, decoder, and evaluator held fixed, so columns differ only in retrieval: Top-K, fixed MMR ($\lambda = 0.7$), and LEAF's learned joint λ/K selector. $K = 5$ on MMQA; $K = 4$ and adapter strength 0.5 on WebQA. The lower panel lists published systems that train retrieval on annotated supporting sources; they differ in architecture and training data, and MuRAG's MMQA subset is not comparable (n/c).

MMQA dev: F1 / EM			
Reader	Top5	MMR5	Learned
Qwen3-2B	63.04 / 57.31	64.37 / 58.54	65.13 / 59.36
Qwen3-8B	71.33 / 65.71	72.43 / 66.20	74.94 / 69.23
Gemma-12B	73.77 / 67.96	74.81 / 68.87	76.02 / 70.59

WebQA test: Overall QA / QA-Acc / QA-FL			
Reader	Top4	MMR4	Learned
Qwen3-2B	45.38 / 65.81 / 60.53	44.54 / 64.63 / 60.11	46.37 / 66.96 / 61.37
Qwen3-8B	50.13 / 71.00 / 63.77	49.81 / 69.91 / 63.48	52.24 / 73.62 / 64.76
Gemma-12B	52.11 / 72.76 / 65.17	51.26 / 71.45 / 64.66	53.75 / 74.88 / 66.22

Published systems trained on annotated sources		
System	MMQA F1 / EM	WebQA QA / Acc / FL
MuRAG (Chen et al., 2022)	n/c	36.1 / 54.6 / 55.7
SKURG (Yang et al., 2022)	64.0 / 59.8	37.7 / 57.1 / 55.4
Solar (Yu et al., 2023)	66.1 / 59.8	40.9 / 58.9 / 60.9

the official test evaluator additionally uses answer-keyword annotations, we keep the local reward implementation and the official measurement distinct, even though both are called Overall QA.

5 RESULTS

5.1 LEARNED SELECTION AGAINST TOP-K AND FIXED MMR

Table 1 is the paper's main comparison. Within a row the reader, candidate pool, decoder, and evaluator are fixed, so the columns differ only in the evidence the reader is given: Top-K, fixed MMR at $\lambda = 0.7$, and LEAF's learned joint λ/K selection over the same proposal family. Learned selection is the best of the three for every reader on both benchmarks.

On MMQA dev, learned selection improves F1 over adapted MMR5 by 0.76 points for Qwen3-2B, 2.51 for Qwen3-8B, and 1.21 for Gemma-12B. The paired 95% intervals are $[-0.47, 2.00]$, $[1.39, 3.63]$, and $[0.16, 2.28]$, respectively: the smaller-reader gain is not statistically resolved. The best configuration reaches 76.02 F1 and 70.59 EM. These rows use the frozen evaluation prompt, so they cannot be subtracted from Table 2, which applies the short-answer correction.

On WebQA test, learned selection improves Overall QA over adapted MMR4 by 1.83 points for Qwen3-2B, 2.43 for Qwen3-8B, and 2.49 for Gemma-12B, and over Top4 by 0.99, 2.11, and 1.64 points respectively, reaching 53.75 Overall QA with Gemma-12B. Aggregate server scores do not permit a paired significance claim. Official retrieval does not follow the answer metrics; Section 5.5 reports it separately.

Comparison with evidence-supervised systems. The lower panel of Table 1 lists published systems that train retrieval against annotated supporting sources. Our best configurations exceed them on both benchmarks: 76.02 F1 on MMQA against 66.1 for Solar and 64.0 for SKURG, and 53.75 Overall QA on WebQA against 40.9, 37.7, and 36.1 for Solar, SKURG, and MuRAG. They differ from ours in architecture, reader, and training data, so this compares published results rather than a controlled ablation, and MuRAG's MMQA subset is not comparable. It does establish that removing supporting-source supervision did not cost answer quality relative to published results on these benchmarks.

Table 2: Reader fine-tuning on MMQA dev (F1 / EM, 2,441 questions), in percentages. Base/fine-tuned pairs share raw MMR contexts, prompts, and decoding; $\lambda = 0.7$, $K = 5$. Q-only omits evidence; its retriever labels the training-context provider. The matching WebQA test panel is Table 5.

		Base		Fine-tuned	
Retriever	Reader	Q-only	MMR5	Q-only	MMR5
Qwen-2B	Qwen3-2B	15.94 / 12.99	41.57 / 34.94	20.39 / 18.23	58.41 / 53.01
Qwen-2B	Qwen3-8B	25.91 / 22.20	58.37 / 51.99	28.88 / 25.36	65.63 / 59.73
Qwen-2B	Gemma-12B	7.76 / 6.60	56.51 / 49.86	27.25 / 24.09	67.02 / 61.41
Qwen-8B	Qwen3-2B	15.94 / 12.99	42.35 / 35.85	20.68 / 18.64	60.27 / 54.81
Qwen-8B	Qwen3-8B	25.91 / 22.20	60.14 / 53.63	28.27 / 24.83	68.06 / 62.60
Qwen-8B	Gemma-12B	7.76 / 6.60	57.80 / 51.25	27.34 / 24.01	69.69 / 63.91

Table 3: LEAF-R alone on MMQA dev: fixed MMR5 ($\lambda = 0.7$), the same fine-tuned reader, and no set selector. Values are F1 / EM percentages. Zero-image rates cover all 2,441 questions, not just image-supported questions.

Reader	Raw	Adapted	Zero-image: raw / adapted
Qwen3-2B	60.44 / 54.94	64.37 / 58.54	63.50 / 19.50
Qwen3-8B	67.96 / 62.64	72.43 / 66.20	63.50 / 17.29
Gemma-12B	70.00 / 64.32	74.81 / 68.87	63.50 / 21.02

5.2 STAGE 1: FINE-TUNING THE READER

Stage 1 must produce a reader that answers from the evidence it is given, since everything afterwards is supervised by that reader's answers. Fine-tuning improves fixed-context F1 in all six pairs; Gemma-12B gains 10.51 and 11.89 points with Qwen-2B and Qwen-8B retrieval, respectively (paired 95% intervals $[9.05, 12.00]$ and $[10.40, 13.43]$), and the matched Gemma WebQA pair also improves with unchanged evidence. These are fine-tuning, not retrieval-learning, gains, and they may include answer-format adaptation; the prompt protocol, the shared question-only controls, and training-side reward diagnostics are given in Appendix A.

5.3 LEAF-R: RETRIEVER ADAPTATION ALONE

Holding the fine-tuned reader and the retrieval policy fixed isolates the encoder. Adaptation improves fixed-MMR answers for every reader, and it corrects the modality bias described in Section 1: the raw encoder returns no image at all for 63.50% of questions, which answer feedback alone reduces to 17–21%, while annotated image recovery rises (Figure 2; Appendix A). These controls use the retrieval-learning prompt, not the short-answer correction of Table 2; adding images is not a goal.

With only the first-ranked source ($K = 1$), LEAF-R raises dev F1 from 44.50 to 53.86 for Qwen3-2B, 51.52 to 60.82 for Qwen3-8B, and 53.56 to 61.17 for Gemma-12B. All paired gain intervals exclude zero (Table 9). These matched controls isolate improved single-source usefulness, not sufficiency for multi-hop reasoning. An earlier WebQA comparison at $K = 5$, including its negative adaptation-only effect, is reported separately in Table 10.

5.4 LEAF-S: SET SELECTION ALONE

Matched ablations hold the fine-tuned Qwen3-8B reader and its adapted encoder fixed and vary only what the selector may choose (Table 7). On MMQA dev, joint selection reaches 74.94 F1, ahead of adaptive K alone by 1.26 points, adaptive λ alone by 1.79, the best training-selected constant by 3.11, and the same joint selector without reader states by 3.53; the paired F1 intervals against K-only, λ-only, and reader-state removal are positive. Two conclusions follow: the two decisions are worth making jointly, and the reader-conditioned features carry more of the effect than either

Table 4: Annotated-evidence retrieval on MMQA dev for the current policies in Table 1. Complete retrieval means recovering all annotated sources at the policy's selected K; recall is averaged over questions. Retrieval metrics are percentages.

Reader	Policy	Complete	Recall	Mean K
Qwen3-2B	Top5	81.44	89.51	5.00
	MMR5	78.74	87.39	5.00
	Learned	70.59	81.58	4.05
Qwen3-8B	Top5	79.15	88.53	5.00
	MMR5	77.35	86.98	5.00
	Learned	78.16	86.55	4.82
Gemma-12B	Top5	82.67	90.17	5.00
	MMR5	80.01	87.79	5.00
	Learned	78.86	86.68	4.69

decision alone. On WebQA test the ordering is the same, with joint selection at 52.24 Overall QA (Table 8). This evidence is reader-specific.

5.5 Annotated-evidence retrieval

Answer gains do not establish evidence-retrieval gains (Table 4). Qwen3-2B selects fewer sources and loses 8.15 complete-retrieval points versus MMR5 (95% interval $[-9.87, -6.47]$), despite its unresolved positive answer difference. Qwen3-8B gains 2.51 F1 points, but changes complete retrieval by only 0.82 ($[-0.70, 2.38]$) and recall by -0.44 ($[-1.45, 0.59]$). Gemma improves answer F1 but reduces recall by 1.11 points (interval $[-2.12, -0.07]$ for learned minus MMR5).

The same divergence appears on WebQA. Official retrieval falls from 40.42 (MMR4) and 45.09 (Top4) to 40.32 for Qwen3-2B, from 40.79 and 44.70 to 37.31 for Qwen3-8B, and from 41.25 and 45.81 to 38.28 for Gemma-12B, while Overall QA rises in every case. Answer improvement does not imply better annotated-source recovery, on either benchmark.

6 Limitations

We evaluate question-local pools, not collection-wide search. Finite MMR menus omit arbitrary combinations and offer one singleton per question, and singleton rewards do not measure complementarity. Reward generation, adaptation, and reader passes remain costly, and adapter strength and category-aware WebQA rewards are dataset-specific choices. MMQA dev informed development and is not untouched testing; bootstrap intervals exclude training-seed variation and adaptive selection bias, and retrieval learning uses training subsets. Answer gains do not establish faithful reasoning or exclude parametric-memory and spurious-cue reliance.

7 Conclusion

LEAF learns candidate rankings and evidence-set choices from a fine-tuned reader's answer feedback, without source targets. Encoder adaptation improves MMQA single-source and fixed-MMR answers, and learned selection improves answer scores over adapted fixed MMR on both benchmarks, though answer gains do not uniformly improve annotated-source recovery.

AI use statement

Generative AI tools assisted hypothesis development, experiment design, literature discovery, software implementation, data preparation, experiment execution, result interpretation, and manuscript drafting and revision. The method overview illustration was generated with an image-generation tool; numerical plots were produced from experiment reports. The experimental outputs reported here come from the identified model runs and benchmark evaluators, not from generated estimates.

The authors reviewed the AI-assisted artifacts and take responsibility for the manuscript and reported results.

Ethics statement

Learning from answer rewards does not ensure that an answer is supported by the retrieved evidence. A reader may exploit parametric knowledge or spurious correlations, and its reward can transmit these preferences to retrieval. We therefore report annotated-evidence retrieval separately from answer quality and do not present either benchmark as validating safety-critical use. Model and dataset licenses continue to apply to downstream use and release.

Reproducibility statement

Section 3 specifies the supervision boundary, objectives, and proposal family; Section 4 describes populations and matched controls. The accompanying paper package records result provenance and unresolved inputs. A public, anonymous code-release location has not yet been supplied and is not claimed here.

References

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. WebQA: Multihop and multimodal QA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.375. URL https://aclanthology.org/2022.emnlp-main.375/.

Jade Goldstein and Jaime Carbonell. Summarization: (1) using MMR for diversity-based reranking and (2) evaluating summaries. In *TIPSTER Text Program Phase III: Proceedings of a Workshop Held at Baltimore, Maryland*, pp. 181–195. Association for Computational Linguistics, October 1998. doi: 10.3115/1119089.1119120. URL https://aclanthology.org/X98-1025/.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

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

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019a. URL https://proceedings.mlr.press/v97/lee19d.html.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096. Association for Computational Linguistics, 2019b. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Binxu Li, Yuhui Zhang, Xiaohan Wang, Weixin Liang, Ludwig Schmidt, and Serena Yeung-Levy. Closing the modality gap for mixed modality search, 2025. URL https://arxiv.org/abs/2507.19054.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. Variational open-domain question answering. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20950–20977. PMLR, 2023. URL https://proceedings.mlr.press/v202/lievin23a.html.

Qi Zhi Lim, Chin Poo Lee, Kian Ming Lim, and Kalaiarasi Sonai Muthu Anbananthen. VLMT: Vision-language multimodal transformer for multimodal multi-hop question answering, 2025. URL https://arxiv.org/abs/2504.08269.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. In *International Conference on Learning Representations*, 2025.

Devendra Singh Sachan, Siva Reddy, William Hamilton, Chris Dyer, and Dani Yogatama. End-to-end training of multi-document reader and retriever for open-domain question answering, 2021. URL https://arxiv.org/abs/2106.05346.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models, 2023. URL https://arxiv.org/abs/2301.12652.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: Complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang. Enhancing multi-modal and multi-hop question answering via structured knowledge and unified retrieval-generation, 2022. URL https://arxiv.org/abs/2212.08632.

Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. Unified language representation for question answering over text, tables, and images. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4756–4765. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.292. URL https://aclanthology.org/2023.findings-acl.292/.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. Distantly-supervised evidence retrieval enables question answering without evidence annotation, 2021. URL https://arxiv.org/abs/2110.04889.

A Reader Diagnostics and Component Visualization

Stage-1 measurement protocol. Table 2's MMQA panel applies the same post-hoc short-answer instruction to both readers to control verbosity; original outputs are retained, so gains may include answer-format adaptation. Base question-only outputs are shared across retriever rows and retain the evidence prompt; Gemma often abstains, so this control does not isolate parametric knowledge. Cap hits are reported without exclusions. All three readers show training-side singleton reward discrimination and menu-oracle headroom, neither of which is a deployable selection gain.

Table 5: Reader fine-tuning on WebQA test (7,540 questions): official Overall QA / QA-Acc / QA-FL, in percentages. Base/fine-tuned pairs share raw MMR contexts, prompts, and decoding; $\lambda = 0.7$, $K = 4$. Q-only omits evidence. Only returned results are shown.

			WebQA test			
Retriever	Reader	State	Context	Overall QA	QA-Acc	QA-FL
None	Qwen3-2B	Base	Q-only	16.64	39.10	29.10
Qwen-2B	Qwen3-2B	Base	MMR4	24.49	58.38	31.22
Qwen-2B	Qwen3-2B	Fine-tuned	Q-only	21.99	34.38	42.22
Qwen-2B	Qwen3-2B	Fine-tuned	MMR4	41.38	60.44	58.06
None	Qwen3-8B	Base	Q-only	13.83	44.57	18.05
Qwen-2B	Qwen3-8B	Base	MMR4	21.88	63.51	25.28
Qwen-2B	Qwen3-8B	Fine-tuned	MMR4	46.71	66.32	61.58
Qwen-2B	Qwen3-8B	Fine-tuned	Q-only	29.38	43.20	47.41
None	Gemma-12B	Base	Q-only	12.19	27.45	19.76
Qwen-2B	Gemma-12B	Base	MMR4	25.97	55.66	31.79
Qwen-2B	Gemma-12B	Fine-tuned	Q-only	29.59	43.71	49.25
Qwen-2B	Gemma-12B	Fine-tuned	MMR4	48.74	68.66	63.19
Qwen-8B	Qwen3-2B	Fine-tuned	Q-only	22.62	34.83	42.36
Qwen-8B	Qwen3-2B	Base	MMR4	24.24	60.62	30.00
Qwen-8B	Qwen3-2B	Fine-tuned	MMR4	44.57	63.98	59.71
Qwen-8B	Qwen3-8B	Base	MMR4	22.34	68.18	25.09
Qwen-8B	Qwen3-8B	Fine-tuned	MMR4	50.07	70.08	63.18
Qwen-8B	Qwen3-8B	Fine-tuned	Q-only	28.96	43.09	47.47
Qwen-8B	Gemma-12B	Base	MMR4	29.57	60.93	35.30
Qwen-8B	Gemma-12B	Fine-tuned	MMR4	51.59	71.87	64.35
Qwen-8B	Gemma-12B	Fine-tuned	Q-only	29.83	43.96	48.57

Matched encoder-adaptation controls. Encoder adaptation improves fixed-MMR F1 by 3.93 points for Qwen3-2B and 4.47 for Qwen3-8B (paired 95% intervals $[2.46, 5.38]$ and $[3.11, 5.84]$). Both controls use unrestricted global MMR and the retrieval-learning prompt, not Table 2's short-answer correction. An inherited raw three-image cap was removed by regenerating seven affected questions per reader and verifying unchanged-context replay.

Among 940 questions with annotated image support, image omission falls from 28.40% to 1.49% and 1.17%; image-source micro recall rises from 42.82% to 70.92% and 73.49%, respectively. These evaluation-only annotations confirm greater image-evidence recovery, not a causal answer benefit from every added image. Overall image inclusion alone is not a success criterion because some questions need no images.

WebQA reader fine-tuning. On WebQA hidden test, the completed Gemma-12B pair with raw Qwen-2B MMR4 contexts improves Overall QA from 25.97 to 48.74 (+22.77 points), QA-Acc from 55.66 to 68.66, and QA-FL from 31.79 to 63.19. The official retrieval score is unchanged at 37.14, consistent with using the same evidence selections before and after reader fine-tuning. These are reader fine-tuning results, not gains from retrieval learning. With raw Qwen-8B MMR4 contexts, Gemma-12B improves from 29.57 to 51.59 Overall QA (+22.02 points), while official retrieval remains 44.22 before and after fine-tuning.

MMQA question-only truncation. The 64-token cap is reached on 212 base Qwen3-2B and three base Gemma Q-only outputs, and on 22 and 17 outputs from the Qwen-2B-context and Qwen-8B-context fine-tuned Gemma readers. The fine-tuned Qwen3-2B readers reach the cap on 12 and 19 Q-only outputs for the two context providers. Qwen3-8B reaches it on 85 base outputs and on 18 and 16 fine-tuned outputs, respectively. All cap-reaching questions remain in the evaluation; reaching the cap indicates possible truncation, not verified semantic incompleteness.

On a shared 128-question MMQA training panel with raw Qwen-8B contexts, singleton answer rewards rank annotated sources above non-sources with question-macro pairwise AUCs of 0.738, 0.720, and 0.751 for the fine-tuned Qwen3-2B, Qwen3-8B, and Gemma-12B readers, respectively (ties receive half credit). Within the frozen $K = 5$ proposal menu, choosing the set with the highest reference-answer F1 improves over fixed MMR by 7.17, 7.87, and 6.70 F1 points. These are

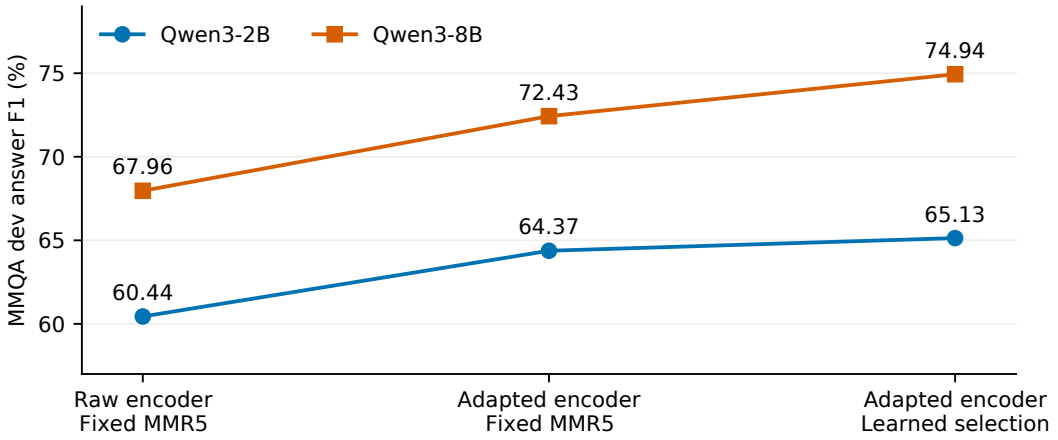

Figure 2: Matched retrieval-learning components on all 2,441 MMQA dev questions. Points are answer-F1 estimates with the same fine-tuned reader within each curve; lines connect experimental conditions, not training steps. Both fixed policies use $K = 5$ and $\lambda = 0.7$. The selector's 2B increment is not statistically resolved; paired intervals are reported in the main text.

training-side diagnostics with expected reader-training overlap, not held-out generalization results. Source labels are used only to report AUC after rewards are frozen; menu-oracle selection uses reference answers and is not deployable. Positive oracle headroom establishes an opportunity, not that the learned selector can realize it.

B MMQA Reader Fine-Tuning Context Recipe

The Qwen-8B-context Qwen3-2B, Qwen3-8B, and Gemma-12B readers share a frozen corpus of 26,121 examples from 23,689 training questions. Each question has one primary MMR context. The deterministic assignment uses cardinality weights $(0.10, 0.15, 0.15, 0.125, 0.125, 0.35)$ for $K = 3, \ldots, 8$ and diversity weights $(0.20, 0.60, 0.20)$ for $\lambda = (0.55, 0.70, 0.85)$. Assignment depends on hashed question IDs, not source annotations. Reader fine-tuning contexts contain at most three images for payload compatibility; this is not a minimum-image quota.

An additional 2,369 image-support examples are drawn from primary contexts without images. They replace the lowest-query-similarity member with the highest-scoring available image, subject to a maximum relevance loss of 0.2 and a minimum image cosine score of 0.27550315856933594. These are extra training contexts, not replacements for the primary examples or a deployment constraint. Exact duplicate sets are removed within a question, and at most two examples are retained per question.

A separate answer-based recoverability check admits only 33 additional $K = 1$ contexts and 30 $K = 2$ contexts. Thus, this recipe exposes the reader to multiple cardinalities but is not balanced over $K = 1, \ldots, 8$. The final counts by increasing K are $(33, 30, 2633, 3961, 3961, 3259, 3242, 9002)$. These are MMQA Qwen-8B-context corpus counts, not WebQA corpus statistics.

The small-K admission check uses the unadapted Qwen2.5-VL-3B reader: generated token F1 must be 1.0 for $K = 1$ or at least 0.8 for $K = 2$, and mean reference-answer log likelihood must be no more than 0.1 below the same question's $K = 8$ context. Nonfinite, truncated, or image-error cases are excluded. This historical corpus-construction check used its frozen 768-token cap, repetition penalty 1.05, and no-repeat 4-gram rule; it is distinct from the neutral 64-token final evaluation. These 63 admitted examples are shared across the three readers, not selected using development answers or source labels.

The readers use one epoch of answer-token supervision, rank-16 LoRA with alpha 32 and dropout 0.05 on language attention Q/K/V/O projections, and a frozen vision tower. AdamW uses learning rate 2×10^{-5}, zero weight decay, cosine scheduling with 3% warmup, batch size one, and eight

gradient accumulation steps. Model-native chat rendering and answer termination are preserved; prompt and padding tokens do not contribute to the loss.

C MMQA Selector Features

The common MMQA scorer has 972 input coordinates. Its 20-dimensional proposal descriptor consists of an eight-way cardinality indicator, $K/8$, three indicators for the retained λ aliases, and eight set statistics: mean/minimum/standard deviation of query cosine, mean/maximum pairwise candidate cosine, and image/text/table counts. Pairwise statistics are zero for singleton sets.

A 46-dimensional question-pool descriptor contains the top eight global query scores; the top five scores in each modality; seven adjacent global score gaps; candidate count divided by 64; image and table fractions; global mean, standard deviation, median, and maximum; and per-modality mean, standard deviation, and maximum. Missing top-score entries use the pool's minimum score. An absent modality uses that minimum for its mean and maximum and zero for its standard deviation. These descriptors summarize retrieval scores, not reference answers or source annotations.

After training-fold standardization, all 46×20 question–proposal products are concatenated with the 20 proposal coordinates, giving 940 scalar features. The remaining 32 coordinates come from normalized pre-answer reader states and a training-only PCA basis fitted to within-question state differences. Proposal descriptors and projected states are expressed relative to adapted MMR with $K = 5, \lambda = 0.7$ before feature construction. This reference defines coordinates only: selection still maximizes the score over the entire menu, with no fallback threshold or privileged deployment choice. The complete design matrix and reward target are then centered within each question for the ridge loss. Exact score ties are resolved by the canonical context hash.

D Training and Inference Procedure

Inputs. Training questions and reference answers, question-local candidate pools, a pretrained universal encoder, and a base multimodal reader. Candidate modality and content are available; supporting-source annotations are not used to construct training targets or choose checkpoints.

Reader fine-tuning. Construct source-label-free contexts using the raw encoder and the frozen context recipe. Fit reader LoRA by reference-answer cross-entropy with prompt tokens masked. Freeze this fine-tuned reader for retrieval learning and final answer generation.

Encoder adaptation. For each fitting question, generate one answer per singleton candidate and compute its answer reward. Form the teacher distribution and fit shared query/document LoRA using KL distillation. Apply the dataset's frozen adapter strength to both branches and encode the candidate pools again.

Set-selector fitting. Construct the adapted MMR proposal family. Encode each question–set prompt with the frozen reader to obtain pre-answer features; generate answers for training proposals to obtain reward targets. Fit preprocessing and the ridge scorer without outer-fold questions, select regularization from training-side reward, and refit the final scorer on its full fitting population.

Inference. Encode the new question and candidate pool with the adapted encoder. Construct the same proposal family, extract the frozen reader's pre-answer features, score every distinct proposal, and generate from the highest-scoring set. Neither a reference answer nor a supporting-source label is required by these steps.

Evaluation boundary. Freeze selections and generated answers before opening annotated source labels. Answer metrics and evidence-retrieval metrics are evaluated separately. Oracle proposal reward is a diagnostic computed with reference answers, never the deployed policy.

E SELECTOR ABLATION PROTOCOL AND TRAINING-SIDE DIAGNOSTICS

Measured proposal counts. On all 2,441 MMQA dev questions, the deduplicated common family contains 48,825 contexts for the 2B reader's adapted encoder and 48,932 for the 8B reader's encoder: respectively 20.00 and 20.05 prompt passes per question before final answer generation. The per-question ranges are 13–22 and 12–22. Thus a small linear selector does not make the complete retrieval procedure cheap. These counts exclude encoder computation and final generation; they are not end-to-end latency measurements.

Component scope. The base-versus-fine-tuned reader comparison measures fine-tuning's direct answer-quality effect under matched retrieval contexts. The raw-versus-adapted fixed-MMR comparison measures encoder adaptation with the fine-tuned reader and retrieval policy held fixed. The selector comparisons below isolate decisions after adaptation. These are distinct controls: better fine-tuned-reader answers do not establish that fine-tuning is necessary for learning retrieval. That claim would require separately adapting retrievers with base and fine-tuned teachers under matched training conditions, which we have not evaluated.

For MMQA, the ablation holds the fine-tuned Qwen3-VL-8B reader and its adapted Qwen3-VL-Embedding-8B encoder fixed. All arms use the same training questions, answer-derived rewards, outer and inner folds, regularization grid, and decoding protocol. The proposal family is $\lambda \in \{0.4, 0.7, 1.0\}$ and $K \in \{1, \ldots, 8\}$. The matched arms are:

- **Fixed MMR**: $\lambda = 0.7$, $K = 5$, without a learned selector.
- **Adaptive λ only**: refit selection over the three diversity settings at $K = 5$.
- **Adaptive K only**: refit selection over eight cardinalities at $\lambda = 0.7$.
- **Joint selection**: the full learned policy reported in Table 1.
- **Without reader states**: refit the full-menu selector with retrieval, action, and question-statistic features, excluding reader-state PCA.
- **Best training constant**: choose one (λ, K) by mean training reward and apply it to every evaluation question.

Restricted arms are refitted rather than obtained by masking the joint selector at inference. Duplicate contexts are collapsed, alias features are recomputed, and preprocessing is fitted within each training partition. Questions with one eligible context remain in evaluation and contribute zero centered ranking loss. Such questions supply no contrast for fitting the contrast-based preprocessing transforms. The best constant is selected without development rewards; its out-of-fold version uses only the outer training partition. Comparisons therefore measure the benefit of available decision dimensions, including their proposal support, rather than identifying a unique causal interaction between λ and K.

Completed training-side diagnostics. Table 6 reports four-fold held-out selector reward on 3,072 MMQA training questions. Questions are held out from selector fitting, not necessarily from reader fine-tuning; these are not development-set results or reader-unseen generalization estimates. The full selector exceeds the λ-only arm by 2.06 points (95% paired question-bootstrap interval $[1.14, 2.99]$), and the reader-state removal arm by 0.86 ($[0.06, 1.66]$). Its 0.41-point advantage over K-only selection is unresolved ($[-0.34, 1.18]$). Thus these diagnostics support adaptive cardinality and a contribution from reader states, but do not establish that joint selection outperforms adaptive cardinality alone on this training-side panel. The separate development evaluation follows below.

Development-set ablations. Table 7 evaluates all frozen policies on the same 2,441 MMQA development questions. Joint selection exceeds K-only selection by 1.26 F1 points (95% paired question-bootstrap interval $[0.40, 2.13]$), λ-only selection by 1.79 ($[0.84, 2.78]$), and selection without reader states by 3.53 ($[2.50, 4.60]$). These comparisons support both decision dimensions and reader-conditioned features within the tested proposal family. They do not establish an optimal proposal family or isolate representation quality from its interaction with the fitted readout. The training-selected constant scores below the fixed MMR control on dev, illustrating that training reward alone does not identify the best fixed development policy.

Table 6: MMQA training-side selector ablations, not final dev performance. Mean out-of-fold token-F1 reward is shown on a 0–100 scale. The best constant is chosen separately within each outer training fold.

Policy	OOF reward
Fixed MMR ($\lambda = 0.7$, $K = 5$)	74.06
Best training-selected constant	74.31
Adaptive λ only	75.03
Adaptive K only	76.68
Joint, without reader states	76.22
Joint, with reader states	77.09

Table 7: Matched selector ablations on MMQA dev with a fine-tuned Qwen3-8B reader and adapted Qwen-8B retriever. All rows use the same frozen reader and neutral greedy decoding; only proposal support and selector features vary as specified above. F1 and EM are percentages.

Policy	F1	EM
Fixed MMR ($\lambda = 0.7$, $K = 5$)	72.43	66.20
Best training-selected constant	71.83	65.55
Adaptive λ only	73.15	67.10
Adaptive K only	73.68	67.92
Joint, without reader states	71.41	65.67
Joint, with reader states	74.94	69.23

WebQA test ablations. Table 8 records official EvalAI results for the matched fine-tuned Qwen3-8B reader and adapted Qwen-8B retriever at strength 0.5. Joint selection scores 52.24 Overall QA, exceeding the training-selected constant by 0.45 points, removal of reader states by 0.64, K-only selection by 0.83, and λ-only selection by 1.22. These are aggregate official-test differences, not paired significance tests. Official retrieval is lower than for either single-dimension adaptive policy; the answer-quality gain does not establish better annotated-source recovery.

F Supplementary Single-Source Evaluation

Table 9 tests whether singleton-supervised adaptation improves the usefulness of the source actually ranked first on development questions. All three readers improve, by 7.61–9.37 F1 points. This validates the item-level effect separately from multi-source composition; it does not imply that one source suffices for multi-hop questions. On WebQA, adaptation raises Top1 Overall QA from 35.81 to 39.05 ($+3.24$ points) and official retrieval from 40.15 to 45.46. These aggregate EvalAI results do not provide a paired significance test.

G Secondary-Retriever Evaluation

On all 2,441 MMQA development questions, adapted VLM2Vec-V2.0 with learned joint λ/K selection achieves 63.34 F1 and 57.44 EM, compared with 61.40 F1 and 55.80 EM for adapted fixed MMR5 ($\lambda = 0.7$). The paired gains are 1.94 F1 points (95% question-bootstrap interval $[0.54, 3.31]$) and 1.64 EM points ($[0.25, 3.07]$). Both policies use the same adapted encoder, prepared Qwen3-2B reader, and neutral greedy decoder capped at 64 new tokens. Three answers per policy reach the cap and remain included. The reader was prepared on Qwen-8B contexts and is held fixed, so this tests retrieval-learning transfer to another encoder, not VLM2Vec-specific reader preparation.

On WebQA hidden test, VLM2Vec-V2.0 with a reader-specific adapter applied at strength 0.5 and learned joint λ/K selection achieves 44.91 Overall QA, 64.84 QA-Acc, 60.43 QA-FL, and 32.26 official retrieval. This evaluation holds the prepared Qwen3-2B reader fixed across the secondary-retriever controls; its preparation used Qwen-8B contexts, not VLM2Vec contexts. The matched adapted MMR4 control ($\lambda = 0.7$) scores 40.61 Overall QA, 58.87 QA-Acc, 56.94 QA-FL, and

Table 8: WebQA test selector ablations. All metrics are official EvalAI percentages. Adaptive λ holds $K = 4$; adaptive K holds $\lambda = 0.7$.

Policy	Overall QA	QA-Acc	QA-FL	Retrieval
Adaptive λ only	51.02	71.73	64.20	45.61
Adaptive K only	51.41	72.17	64.32	39.34
Joint, without reader states	51.60	72.77	64.59	34.34
Best training-selected constant	51.79	72.88	64.82	30.58
Joint, with reader states	52.24	73.62	64.76	37.31

Table 9: Single-source answer quality on MMQA dev (upper) and WebQA test (lower). The reader receives the actual top-ranked source from each encoder, not an oracle source or an average over singleton rewards. MMQA values are F1 / EM percentages; WebQA values are official Overall QA / QA-Acc / QA-FL percentages.

Reader	Raw Top1	Adapted Top1	Δ F1 [95% CI]
Qwen3-2B	44.50 / 39.78	53.86 / 48.79	9.37 $[7.74, 10.99]$
Qwen3-8B	51.52 / 46.42	60.82 / 55.51	9.30 $[7.73, 10.89]$
Gemma-12B	53.56 / 48.50	61.17 / 55.84	7.61 $[6.07, 9.15]$

WebQA test: Overall QA / QA-Acc / QA-FL

Reader	Raw Top1	Adapted Top1
Qwen3-2B	35.81 / 52.73 / 53.82	39.05 / 57.30 / 56.04

27.75 retrieval. Learned selection therefore improves Overall QA by 4.30 points and retrieval by 4.51 points over this fixed policy. These are aggregate EvalAI differences, not paired significance estimates. Raw MMR4 scores 43.37 Overall QA, 62.14 QA-Acc, 58.71 QA-FL, and 37.71 retrieval. Thus adaptation alone lowers fixed-policy Overall QA by 2.76 points; learned selection recovers this loss and exceeds raw MMR4 by 1.54 points. Its retrieval score remains 5.45 points below raw MMR4. This supports an answer-quality gain for the full pipeline with a second retriever, not a claim that each component improves every metric.

H Historical Cross-Dataset Component Comparison

The following banked comparison predates the main-table evaluation above. Its fixed WebQA controls use $K = 5$, not the $K = 4$ used in the current main table. It is retained as a separate component ablation, not as a substitute for the reader-specific main-table comparisons.

Table 10 isolates retrieval learning: the reader is fine-tuned once and then frozen, so every row differs only in which evidence it is given.

The full method improves answers on both benchmarks. On MultimodalQA the learned policy reaches 73.31 F1 against 66.95 for raw fixed retrieval, a paired gain of 6.36 points (95% interval $[4.90, 7.78]$), and 1.61 points over adapted fixed retrieval ($[0.52, 2.76]$). On WebQA it reaches 46.37 Overall QA against 45.51 and 45.09 for the two fixed policies. The WebQA figures come from the benchmark's hidden test server, so no paired interval is available for them.

The components do not each improve every metric. On MultimodalQA both steps help, and encoder adaptation is the larger of the two. On WebQA adaptation alone *reduces* the fixed-policy test score by 0.42 Overall-QA points, and reduces annotated retrieval by 3.66 points (Table 11); the learned selector then recovers both and finishes above the raw policy. We therefore claim the pathway as a whole rather than claiming that each component is independently beneficial in every setting. This asymmetry is informative: an encoder reshaped toward singleton usefulness need not produce better *sets* under a fixed policy, and choosing the set is what recovers the loss.

Table 10: Answer quality under the common method. All rows within a benchmark share one fine-tuned reader, candidate pool, decoder, and evaluator, so differences are attributable to retrieval. MultimodalQA uses the 2,441-question development set, which is the split reported by prior work because test labels are not released, and which we also used during development. WebQA uses the hidden test set scored by the benchmark server. Overall QA is the mean of per-question QA-Acc×QA-FL products and is not the product of the two aggregate columns.

	MultimodalQA (dev)		WebQA (test)		
Retrieval policy	F1	EM	Overall QA	QA-Acc	QA-FL
Raw fixed MMR (λ=0.7, K=5)	66.95	61.33	45.51	65.26	60.41
LEAF-R + fixed MMR (λ=0.7, K=5)	71.70	65.75	45.09	64.92	60.61
LEAF (LEAF-R + LEAF-S)	73.31	67.39	46.37	66.96	61.37

Table 11: Evidence retrieval for the same three policies and the same sealed selections as Table 10. WebQA reports the benchmark's official retrieval score on the hidden test set. MultimodalQA evidence metrics are pending recomputation on the current selector's selections and are not carried over from a superseded policy. Retrieving annotated evidence is a separate measurement from answering correctly; neither implies the other.

Retrieval policy	MultimodalQA evidence	WebQA retrieval
Raw fixed MMR (λ=0.7, K=5)	–	40.05
LEAF-R + fixed MMR (λ=0.7, K=5)	–	36.39
LEAF (LEAF-R + LEAF-S)	–	40.32

Adaptivity and context budget remain partly confounded. The learned policy selects a mean cardinality of 5.32 on MultimodalQA, above the K=5 of the fixed controls, and its most frequent choice is K=8. A comparison against K=5 alone therefore cannot attribute the gain to adaptivity rather than to context size; this comparison remains a limitation of the historical result. The policy does vary its choice: it selects every cardinality from one to eight, with 296 questions at K=1 and 784 at K=8.

Table 10 shows that answer-only supervision improves answers. This section asks what it changes about the evidence itself. We open the benchmark's annotated sources only after every policy and every generated answer is frozen, and we never use them to fit or select a model.

Modality composition is measured separately. The current reader-specific comparisons in Table 3 show a reduction in zero-image selection under the same fixed MMR policy after adaptation, without a modality quota or modality-specific training target. This observation does not establish that each newly included image causes an answer gain. Those current comparisons must not be substituted for the historical policy's missing MMQA evidence measurements in Table 11.

Answer quality and evidence retrieval move independently. WebQA makes this concrete. Encoder adaptation alone lowers the official retrieval score from 40.05 to 36.39 and Overall QA from 45.51 to 45.09, and the learned selector then returns retrieval to 40.32 and raises Overall QA above both fixed policies. Retrieval score and answer score are therefore not interchangeable, and a gain in one should not be reported as evidence for the other. We do not read the divergence as proof that the annotations are incomplete: parametric knowledge in the reader, precision and recall trade-offs within a set, and spurious cues are alternative explanations that these measurements do not separate.

What the policy selects. On MultimodalQA the policy uses its whole range, selecting K=1 for 296 questions and K=8 for 784, with a mean of 5.32. The choice is genuinely question-dependent rather than a near-constant budget with rare deviations, but it is also larger on average than the K=5 controls, so part of the gain plausibly reflects reading more. Separating adaptivity from budget requires the matched constant-policy control. The newer Qwen3-8B ablations include such a control, but do not retroactively resolve this historical comparison.

Cost. The selector itself is one linear readout, but every proposed set must be encoded by the frozen reader before one is chosen, so cost scales with the number of distinct proposals per question rather than with the size of the fitted model. The family contains at most 24 proposed contexts before deduplication; each distinct context requires a prompt pass, followed by answer generation for the selected context. A matched wall-clock inference comparison is not yet available, so we make no empirical efficiency claim.

I Comparison with Published Source-Supervised Systems

Table 12 records published reference results rather than matched reimplementations. SKURG reports full-MMQA validation results; Solar reports MMQA results with development data and no test split listed in its dataset statistics. MuRAG evaluates only the TextQ/ImageQ subset, so we do not compare its subset scores with our full-MMQA results. All three WebQA rows use the official test-set distractor setting, not full-corpus retrieval.

Table 12: Published source-supervised reference systems. MMQA reports EM/F1; WebQA reports Overall QA, QA-Acc, and QA-FL. All values are on a 0–100 scale. A dash denotes an incompatible evaluation population, not a missing score to be estimated. These are cross-paper comparisons with different architectures and pretraining, not controlled supervision ablations.

	MMQA		WebQA test		
Method	EM	F1	QA	QA-Acc	QA-FL
MuRAG (Chen et al., 2022)	–	–	36.10	54.60	55.70
SKURG (Yang et al., 2022)	59.80	64.00	37.70	57.10	55.40
Solar (Yu et al., 2023)	59.80	66.10	40.90	58.90	60.90

Our Qwen3-8B and Gemma-12B configurations exceed these full-MMQA answer scores, and our completed Qwen3-2B, Qwen3-8B, and Gemma-12B WebQA systems exceed all three published Overall-QA scores. This does not hold for every reader and metric: our Qwen3-2B MMQA F1 exceeds SKURG but not Solar, and its EM is below both. Our WebQA annotated-source retrieval scores are also below the published source-supervised systems. The comparison therefore supports strong answer quality without source-label supervision, not a claim that removing source labels causes an improvement or that our systems dominate all evidence-retrieval metrics. Our matched controls and ablations provide the separate evidence for the contribution of the proposed procedure.