ICML International Conference On Machine Learning

INTERNATIONAL CONFERENCE ON MACHINE LEARNING · 2026

Beyond Single-View Indexing

Structure-Aware Multi-View Retrieval for Knowledge-Based VQA

Hao Wang[1] **Xujia Li**[2*] **Lei Chen**[1,2]

[1]Data Science and Analytics, Information Hub, HKUST(GZ), Guangzhou, China · [2]HKUST, Hong Kong, China

香港科技大学(广州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Code & details
github.com/seraveea/SCAR

BACKGROUND

KB-VQA stands or falls on coarse-grained retrieval

Knowledge-Based VQA (KB-VQA) answers questions that need external, encyclopedic facts beyond the image, via a multi-modal Retrieval-Augmented Generation (mRAG) pipeline:

Coarse Retrieval → Re-ranking → MLLM Answering

Prior work pours effort into stronger **rerankers** and **MLLMs**, yet treats the first **coarse-grained retrieval** stage simplistically.

But downstream stages can only re-rank what retrieval already found — **misses here are unrecoverable**.

The blind spot of single-view indexing

Most systems build the entity index from a **single view** — entity image, title, *or* summary — or fuse views with naive score averaging.

The best view **varies across datasets and queries**; each captures only partial evidence, leaving systematic coverage gaps.

Retrieval recall sets the **ceiling** for the entire KB-VQA system.

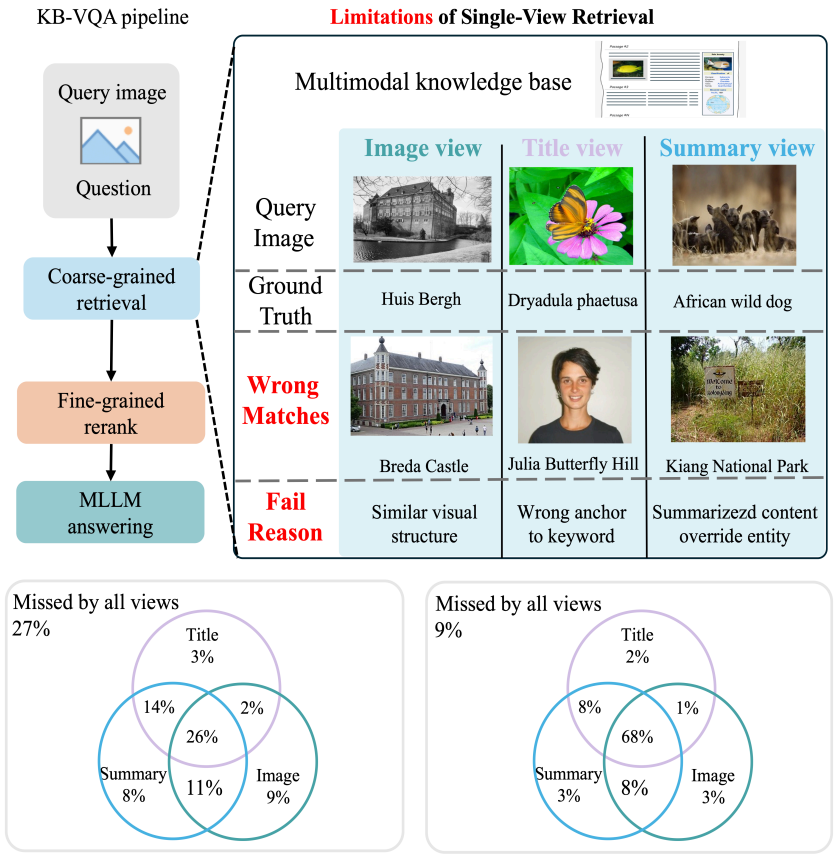
MOTIVATION

Preliminary Test — index has blind spots

We build three view-specific indexes — **Image**, **Title**, **Summary** — and study their Recall@20 on E-VQA and InfoSeek.

- **Strong complementarity:** large non-overlapping regions — many entities are retrievable by only one view.
- **View-specific failures:** e.g. a butterfly image under the *title* view returns "Ms. Butterfly" — noise from limited encoders.
- **No single winner:** even the best overall view loses on many individual queries.

Single-view indexing is **unstable**; the union of views holds far more correct entities than any one alone.



View-specific failure cases and Recall@20 overlap (Venn) of Image / Title / Summary views on E-VQA (left) and InfoSeek (right).

THEORETICAL ANALYSIS

An information-theoretic view of retrieval bounds

Single-view upper bound

$$p_m \le \frac{I^{(m)}(x_e^{(m)}; E(x_e^{(m)})) + I(e; E(q)) + \delta_{q,e}^{(m)} + 1}{H_m - \log_2 k}$$

Recall is bounded by the **view-specific mutual information** preserved in the embedding, and hurt by intrinsic entropy H_m.

Multi-view upper bound

$$p_{ora} \le \frac{\frac{M_{eff}}{M} \sum_j I^{(j)} + I(e; E(q)) + \sum_j \delta_{q,e}^{(j)} - \varkappa \Delta H + 1}{H - \log_2 k}$$

Gain depends on $M_{eff} = 1 + (M - 1)(1 - \varrho)$: the **effective number of non-redundant views**.

Challenge 1 — Intra-view noise

Single-view effectiveness is limited by how much view-specific information survives in the embedding space. View-specific noise creates **spurious "hallucinated" matches**.

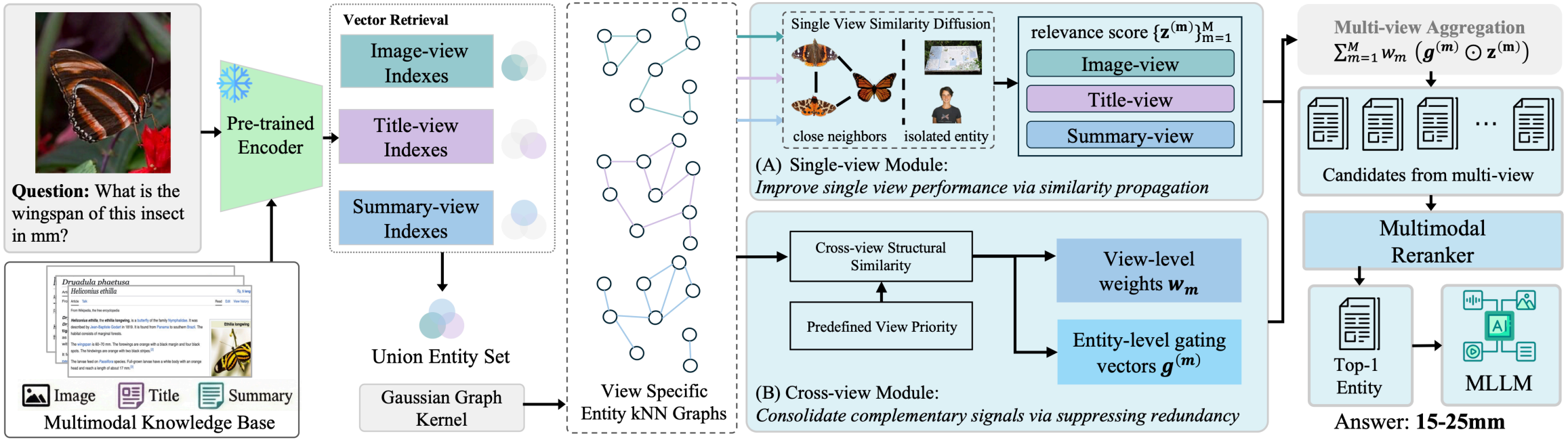
Challenge 2 — Cross-view redundancy

Multi-view benefit hinges on **non-redundant** information. Highly overlapping views (e.g. title & summary) shrink M_{eff}, so naive fusion over-amplifies redundant evidence.

These two challenges directly motivate the two modules of **SCAR**.

METHOD · OVERVIEW

SCAR: Structure-aware Cross-view Alignment for Retrieval



SCAR propagates similarity over per-view entity kNN graphs, then regulates cross-view redundancy from structural overlap — aggregating complementary knowledge views.

Training-free | No extra models / supervision | Inference-time only | < 0.5% latency overhead

METHOD · MODULE 1

Similarity Propagation via Entity kNN Graphs

Goal: tackle Challenge 1 — reduce intra-view noise using the **structure** among retrieved entities.

Over the union E_{union} of retrieved entities, build a per-view kNN graph with a Gaussian affinity kernel:

$$W_{ij}^{(m)} = \exp\left(-\frac{d_{ij}^{(m)2}}{2\sigma_m^2}\right)$$

Entities absent from view m are disconnected ($W_{i,:} = W_{:,i} = 0$); keep top-n neighbors.

Symmetric normalization then closed-form **manifold propagation**:

$$S^{(m)} = D^{-1/2} W^{(m)} D^{-1/2}$$

$$z^{(m)} = (I - \alpha S^{(m)})^{-1} y$$

Effect

Entities in **coherent neighborhoods reinforce each other**; isolated hallucinated matches lack supporting inflow and **decay** — "wisdom of the crowd".

METHOD · MODULE 2

Cross-View Redundancy Regulation

Goal: tackle Challenge 2 — stop redundant views from over-amplifying the same evidence.

On shared entities $E_{j_1 j_2}$, measure structural similarity of induced subgraphs via the Frobenius inner product:

$$\mathrm{sim}(j_1, j_2) = \frac{\langle S_E^{(j_1)}, S_E^{(j_2)} \rangle_F}{\| S_E^{(j_1)} \|_F \, \| S_E^{(j_2)} \|_F}$$

When $\mathrm{sim}(j_1, j_2) > \tau$, the lower-priority view is suppressed at two granularities.

Dual-granularity suppression

(1) View-level: global down-weight $\omega_{j_{\mathrm{low}}} \leftarrow \gamma \, \omega_{j_{\mathrm{low}}}$.

(2) Entity-level: overlapping entities gated by factor β via $g^{(m)}$.

$$z_{\mathrm{final}} = \sum_{m=1}^{M} w_m \, (g^{(m)} \odot z^{(m)})$$

Preserves **complementary signals** while suppressing redundant overlap. Triggered on **~80%** of queries.

EXPERIMENTS · MAIN RESULTS

SOTA recall, approaching the coverage upper bound

Retrieval Recall@K on KB-VQA benchmarks. SCAR uses EVA-CLIP-8B, no extra training.

Method	E-VQA				InfoSeek			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Multi-view coverage bound	*33.9*	*56.9*	*65.8*	*73.0*	*68.9*	*84.2*	*87.9*	*90.7*
Wiki-LLaVA	3.3	9.9	13.2	17.5	36.9	66.1	71.9	78.4
EchoSight	13.3	31.3	41.0	48.8	45.6	67.1	73.0	77.9
ReflectiVA	15.6	36.1	—	49.8	56.1	77.6	—	86.4
OMGM	18.7	41.2	49.7	58.7	52.2	73.7	79.8	84.7
SCAR (Ours)	**28.8**	**49.9**	**57.0**	**65.1**	**60.8**	**80.0**	**84.6**	**87.7**
% of upper bound	*86.0*	*87.7*	*86.6*	*89.2*	*88.2*	*95.0*	*96.3*	*96.7*

+6.4
R@20 gain over best baseline (E-VQA)

96.7%
of multi-view upper bound at R@20 (InfoSeek)

89.2%
of upper bound at R@20 (E-VQA)

EXPERIMENTS · ANALYSIS

Ablation Study — both modules matter, at near-zero cost

Module ablation (E-VQA, Recall@K)

Module 1	Module 2	R@1	R@5	R@20
—	—	24.6	42.3	56.4
✓	—	28.4	48.3	63.8
✓	✓	**28.8**	**49.9**	**65.1**

View combinations (E-VQA, R@20)

Img	Title	Sum	R@20
✓			48.8
	✓		44.7
		✓	58.7
✓		✓	62.9
✓	✓	✓	**65.1**

Average per-query runtime, ms (E-VQA)

Modality	Retrieval	Total	R@20
Title	44	8447	44.7
Image	59	8535	48.8
Summary	46	8545	58.7
SCAR	**85**	**8569**	**65.1**

Highly favorable trade-off

~40 ms extra retrieval cost = **< 0.5%** of ~8000 ms total latency, for a **+6.4 R@20** gain.

Gains hold with re-ranking (R@20 58.7→65.1) and lift end-to-end VQA across LLaVA, Qwen3-VL, InternVL3.5.

CONCLUSION

Coordinate the views, win the retrieval

- Coarse-grained retrieval is the **critical bottleneck** in KB-VQA — and views are **complementary, not redundant**.
- **SCAR** turns this insight into a **training-free** method: intra-view manifold propagation + cross-view redundancy regulation.
- Achieves **SOTA recall**, approaches the coverage upper bound, and integrates with rerankers & MLLMs at **negligible cost**.

Future: memory-efficient multi-view indexing (cross-view sharing / compression) and lightweight MLLM internal signals for adaptive retrieval.

One-line summary

SCAR is a training-free, inference-time multi-view retriever that runs manifold propagation over per-view entity kNN graphs to denoise each view, then regulates cross-view redundancy by Frobenius structural similarity — unifying image, title & summary indexes to reach **+6.4 R@20** and **~96.7%** of the coverage upper bound at **<0.5%** latency.

Code: github.com/seraveea/SCAR

Thank you! Questions welcome.