EMM-QA Workshop @ ICML 2026 • Seoul, South Korea

Proceedings of Efficient Multimodal Question Answering Workshop, PMLR 306, 2026

Salient Knowledge Pathways:

Sparse Cross-Modal Routing for Efficient Knowledge-Intensive Multimodal QA

Noor Islam S. Mohammad* Uluğ Bayazıt

Istanbul Technical University, Department of Computer Science / Computer Engineering

islam23@itu.edu.tr

Code & models: https://pmlrbd.github.io/skip/

Outline

Knowledge-Intensive Multimodal QA (KI-MMQA)

What is KI-MMQA?

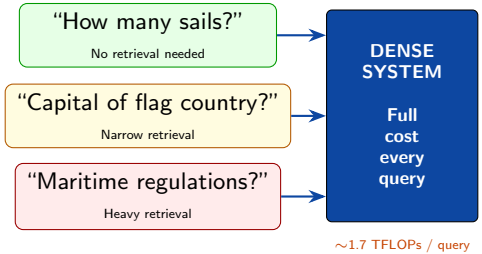
Answer visual questions whose correct answers *cannot* be derived from the image alone — **external knowledge retrieval is mandatory**.

Three tightly coupled sub-problems:

- **Visual localization** — which patches are relevant?
- **Knowledge retrieval** — find the right passage from millions
- **Cross-modal fusion** — bind image evidence with retrieved text

The Cost Problem

Each sub-problem is individually expensive.
Dense systems **pay all three costs uniformly** per query — regardless of difficulty.



The Compute Crisis in KI-MMQA

Inference cost of a single query (A100-80GB):

Compute	$\approx$ **1.7 TFLOPs**
Speed	$>$ **800ms latency**
Memory	> 12 **GB KV-cache memory**

Inference cost formula:

$$\mathcal{O}(V \cdot T \cdot K \cdot d)$$

V=576 tokens, T=20, K=16 chunks, d=4096

Root Cause: Unconditional Dense Pricing

Dense systems treat *every* query identically. But informational requirements are **highly non-uniform**:

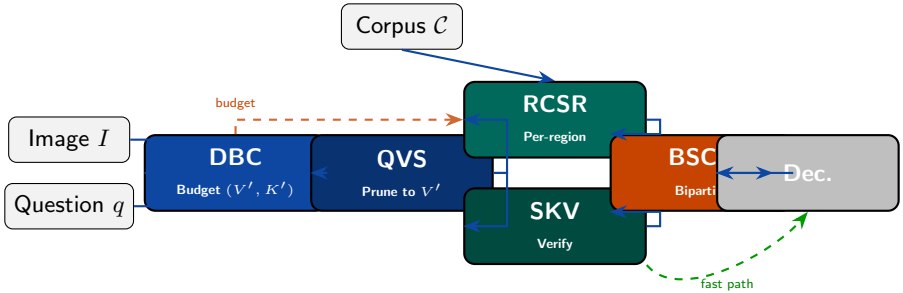
- *"How many sails?"* — no retrieval needed
- *"Capital of flag country?"* — 1 fact needed
- *"Maritime regulations?"* — heavy retrieval

The opportunity: Route compute *conditionally* — jointly across modalities.

Our Answer: SKIP

Sparse, question-conditional pathways spanning visual encoding, retrieval, *and* fusion.

SKIP: Five Learned Components



DBC Difficulty-Adaptive Budget Controller

QVS Question-Guided Visual Saliency

RCSR Region-Conditional Sparse Retrieval

BSCA Bipartite Sparse Cross-Attention

SKV Speculative Knowledge Verif.

Question-Guided Visual Saliency

Goal: Prune irrelevant visual tokens *before* retrieval.

Saliency score per token:

$$s_i = \mathrm{MLP}\big(\boldsymbol{v}_i \,\|\, \bar{\boldsymbol{q}} \,\|\, (\boldsymbol{v}_i \odot \bar{\boldsymbol{q}})\big)$$

The Hadamard term captures multiplicative interactions (ablating it costs -1.4 pts).

Training:

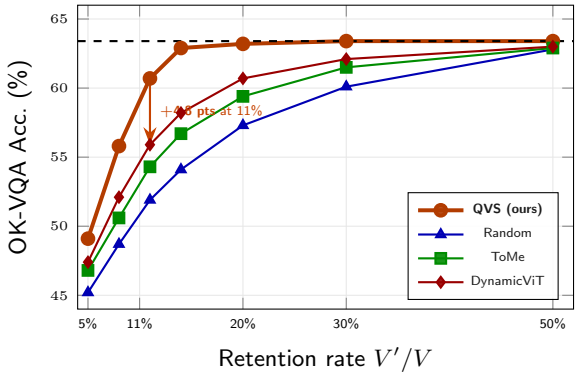
$$\mathcal{L}_{\mathrm{QVS}} = \mathcal{L}_{\mathrm{task}} + \lambda_1 \left| \frac{1}{V} \sum_i \sigma(s_i) - \rho \right|$$

Gumbel-Sigmoid relaxation; temperature annealed $1.0 \to 0.1$.

Why pre-retrieval?

78% of visual tokens encode background with MI < 0.01 bits.
Post-retrieval pruning *cannot* recover a diluted query.

QVS vs. unconditional pruning:



Region-Conditional Sparse Retrieval

Key idea: Issue *separate* retrieval queries per salient image region — not one global query.

Region proposal:

$$\mathcal{R} = \text{SAM-lite}(I, q;\, \theta_r)$$

Per-region query:

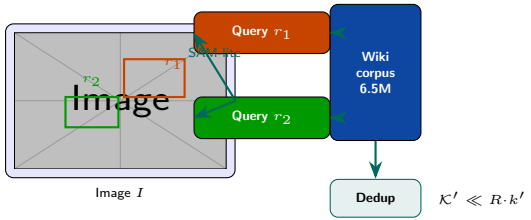
$$\boldsymbol{q}_{r_i} = \text{MLP}_\theta\big([\bar{\boldsymbol{q}};\, \boldsymbol{f}_{r_i};\, \boldsymbol{p}_{r_i}]\big)$$

Deduplicate: $\mathcal{K} = \text{Dedup}\big(\bigcup_i \mathcal{K}_{r_i}\big)$
Deduplication rate 40–60% on real queries.

InfoSeek Hard Split

Method	Recall@4
Global retrieval	43%
RCSR (per-region)	**71%**

Target entities occupy $< 10\%$ image area



- Surfaces knowledge tied to small, discriminative regions
- Logos, text, secondary objects, fine-grained categories
- $R = \min(V'/4, 8)$ regions in practice

BSCA Bipartite Sparse Cross-Attention

Bipartite compatibility matrix:

$$A_{ij} = \sigma\left(\frac{\langle W_v \boldsymbol{v}_i,\, W_k \boldsymbol{k}_j \rangle}{\sqrt{d}}\right)$$

Zero edges below threshold τ (top-β quantile).
Attention computed *only* on surviving edges $\mathcal{E}$.

Sparse attention:

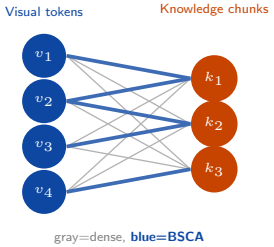
$$\mathrm{BSCA}(v_i) = \frac{\sum_{k_j \in \mathcal{N}(v_i)} \exp(e_{ij})\, W_V k_j}{\sum_{k_j \in \mathcal{N}(v_i)} \exp(e_{ij})}$$

Complexity Reduction

$\mathcal{O}(P \cdot \bar{\Delta})$ vs. $\mathcal{O}(P \cdot M)$ (dense)
9.8$\times$ **reduction** at $\bar{\Delta}$=4, M=16

Why bipartite (not general sparse)?

Most informative edges in KI-MMQA are *cross-modal*.
Intra-modal attention contributes little after encoder layers.



Implemented via block-sparse CUDA kernel (FlashAttention-based, two-pass approach).
7.2$\times$ **wall-clock speedup** for fusion layer.

DBC SKV Budget Controller & Speculative Verifier

DBC Difficulty-Adaptive Budget Controller

Lightweight 3-layer MLP predicting (V', K') from $(\bar{\boldsymbol{v}}, \bar{\boldsymbol{q}})$:

$$\mathcal{L}_{\text{DBC}} = \mathbb{E}[\mathcal{L}_{\text{task}} + \lambda_2 \cdot \text{cost}(V', K')]$$

What DBC Learns

Counting/spatial	$K' = 0$ (no retrieval)
Entity ID	high V', moderate K'
Encyclopedic	low V', high K'

Median: V'=64, K'=6

SKV Speculative Knowledge Verification

Small 700M-parameter VLM drafter produces candidate answer a_D + confidence c_D.

- If $c_D > \tau_{\text{SKV}}$: return a_D directly
- Otherwise: full SKIP pipeline

Result

Routes **34–41**% of queries through fast path with $\leq 1\%$ accuracy cost.

40% FLOP reduction at zero accuracy cost.

Fast-path rate: higher for perceptual benchmarks, lower for entity-centric ones — genuine difficulty signal, not spurious confidence.

Sparsity–Accuracy Tradeoff Theorem

Assumption 1 (Piecewise-Lipschitz MI)

$I(\boldsymbol{V}_S;\, a \mid q)$ is L-Lipschitz in $|S|/V$ on intervals of width $\geq 1/\sqrt{V}$.

Assumption 2 (Bounded Saliency Error)

QVS scores satisfy $|s_i - I(\boldsymbol{v}_i;\, a \mid q)| \leq \eta$ uniformly.
Empirically: $\eta \in [0.011, 0.018]$ across benchmarks.

Theorem (Sparsity-Accuracy Tradeoff)

Under both assumptions, retaining top-V' tokens with

$$V' \geq C \cdot \sqrt{V} \cdot \log(1/\varepsilon)$$

ensures $I(\boldsymbol{V}_{S'};\, a \mid q) \geq I(\boldsymbol{V};\, a \mid q) - \varepsilon$.

Proof sketch (5 steps):

1. Bound symmetric difference $|S^* \triangle S'|$ using score gap Δ and η bound
2. MI loss from misranking via Lipschitz condition: $\leq 2L\eta/\Delta$
3. MI loss from sparsification decomposes over non-selected tokens
4. Concentration argument: top $\sqrt{V}$ tokens carry $\geq 1 - O(\varepsilon)$ total MI
5. Combine: set each term $\leq \varepsilon/2$

Corollary

For V=576, ε=0.05:
Bound predicts $V' \approx 51$ tokens (8.8%)
Empirical optimum: $V' = 64$ (11.1%)

First non-vacuous sparsity bound for KI-MMQA.

Experimental Setup

Benchmarks (5 datasets):

- **OK-VQA** — 14k questions
- **A-OKVQA** — 25k + rationales
- **InfoSeek** — 1.4M entity-centric
- **Enc-VQA** — 1M fine-grained
- **ViQuAE** — 3.7k named entities

Baselines:
Retrieval-free: BLIP-2, LLaVA-1.5
RAG: KAT, RAVQA-2, ReVeaL, RA-CM3
All use Vicuna-7B v1.5 backbone.

Implementation:

Vision: EVA-CLIP-G (336×336, V=576)
Retrieval: SPLADE++ over 6.5M Wikipedia passages
Backbone: Vicuna-7B v1.5 + LoRA rank-64
Training: 6-stage training (∼4,800 A100-hours)
Index: FAISS IVF-PQ, 768-dim SPLADE vectors

Metrics: VQA accuracy • TFLOPs/query • Wall-clock latency (A100-80GB)

Main Results

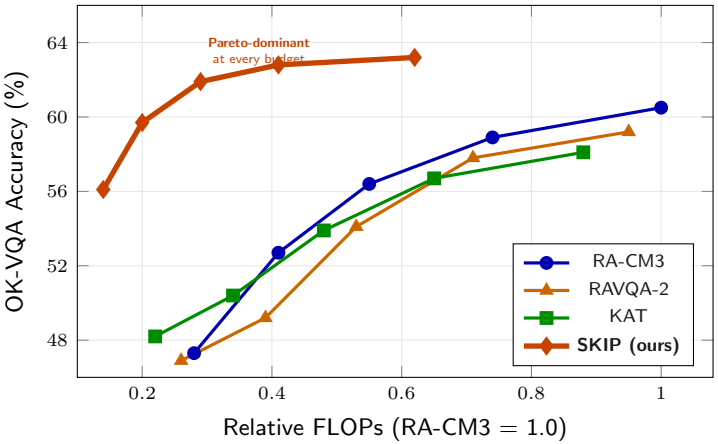
Model	OK-VQA	A-OKVQA	InfoSeek	Enc-VQA	ViQuAE	TFLOPs / Lat. (ms)
BLIP-2 (no retrieval)	45.9	39.7	11.2	8.4	19.1	0.42 / 210
LLaVA-1.5 (no retrieval)	50.3	44.1	13.7	10.2	22.3	0.51 / 240
KAT	54.4	48.2	18.6	14.7	28.9	0.93 / 410
RAVQA-2	56.8	50.6	21.3	16.4	31.7	1.18 / 520
ReVeaL	58.0	52.4	23.8	18.1	33.2	1.34 / 610
RA-CM3	60.5	55.3	26.4	20.9	35.6	1.71 / 820
SKIP (ours)	**63.2**	**58.4**	**30.7**	**24.3**	**37.4**	**0.42 / 305**
vs. RA-CM3	+2.7	+3.1	+4.3	+3.4	+1.8	**4.07× less / 2.69× faster**

Best accuracy on ALL 5 benchmarks

4.07× fewer FLOPs vs RA-CM3

2.69× faster end-to-end

Accuracy–Compute Frontier



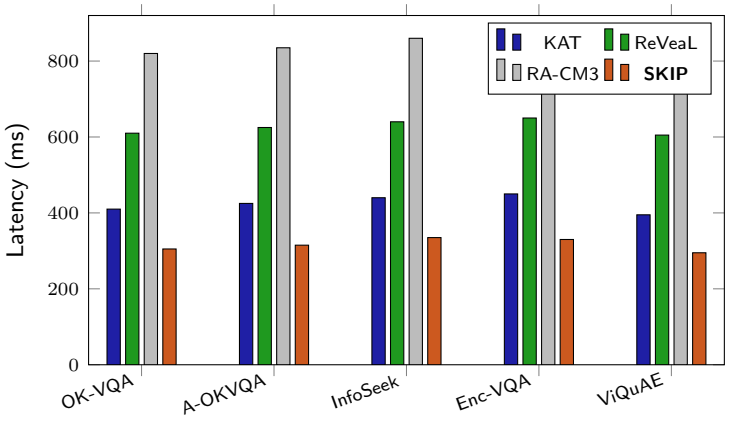
Key findings:

- SKIP **Pareto-dominates** all baselines at every measured FLOP budget
- Gap *widens* at lower budgets
- At $0.20\times$ RA-CM3 cost:
  SKIP retains **98.4%** of its peak accuracy
  RA-CM3 drops to **87.1%** of its peak
- At matched FLOPs (0.42 TFLOPs): **+17.3 pts** over BLIP-2 on OK-VQA

Core insight

Sparse routing **converts saved compute into accuracy**: the system can afford a much larger effective corpus and richer visual context.

End-to-End Latency



Consistent 2.5–2.9× speedup across all 5 benchmarks.

Latency breakdown (SKIP):

- ∼60% LLM decoder generation
- ∼25% FAISS retrieval
- <5% Routing overhead (QVS + DBC + SKV)

Memory savings

RA-CM3	24.0 GB
SKIP	**15.2 GB**

KV cache: 12.7 GB → 4.8 GB

Deployable on 16 GB consumer GPUs!

Further speedups should target the *decoder*, not routing or retrieval.

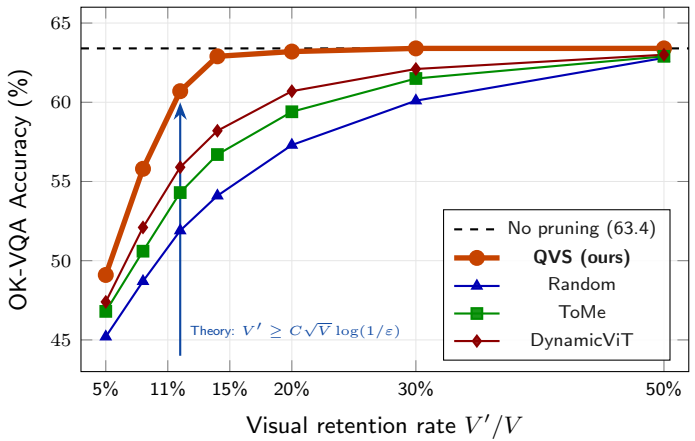
Ablation Study

Configuration	Acc. (%)	TFLOPs
Full SKIP	**63.2**	**0.42**
– QVS (random pruning)	54.1	0.42
– QVS (no pruning)	62.4	1.38
– RCSR (global retr.)	59.7	0.48
– BSCA ($\rightarrow$ dense)	62.9	1.21
– DBC (fixed $V'{=}64$)	61.8	0.46
– SKV (no fast path)	63.1	0.71
QVS only (others dense)	58.3	0.95
RCSR only (others dense)	60.1	1.05
BSCA only (others dense)	60.8	0.88

Key ablation findings:

1. **QVS is critical:** Random pruning -9.1 pts at same FLOPs. *What* we prune matters more than *how much*.
2. **RCSR contributes 3.5 pts** over global retrieval (gap widens to -4.3 on InfoSeek).
3. **BSCA is the efficiency lever:** Dense replacement triples FLOPs for only $+0.3$ pts.
4. **DBC matters for hard queries:** -2.7 pts on InfoSeek hard split (vs. -1.4 here).
5. **SKV is free savings:** -40% FLOPs, zero accuracy cost.
6. **Gains compound** — no single component recovers the full effect.

Sparsity Analysis vs. Theory



Sparsity analysis findings:

- QVS reaches **99% of no-pruning accuracy** at just **11% retention**
- Random/DynamicViT need 30–50% retention for comparable accuracy
- **5–10 pt gap** at 11% retention vs. unconditional methods

Theory–Practice Agreement

Theoretical prediction:

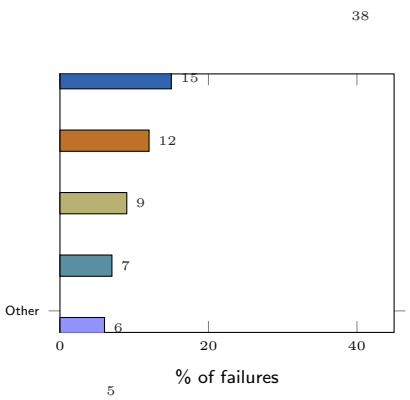
$$\sqrt{576}/576 \approx 0.042$$

With constant $C \in [1.5, 3]$: sweet spot in $[0.06, 0.13]$ range.

Empirical optimum: **11%** ✓

Failure Modes & Error Analysis

36.8% of OK-VQA queries missed.



Top failure modes:

1. **Long-tail entities (38%):** Sparse Wikipedia coverage. *Not SKIP-specific* — same gap in RA-CM3. **SKIP-ADR** mitigation: +140M Wikidata triples, on-demand Google KG API (+12ms).
2. **Compositional/multi-hop (15%):** BSCA's $\bar{\Delta} \approx 4$ misses second-hop binding. **SKIP-Iter** (H=2): +8.4 pts on multi-hop at 3× fewer FLOPs than RA-CM3.
3. **Counting under occlusion (12%):** SKV over-confident. Mitigation: domain-specific τ_{SKV} calibration.

Inter-annotator agreement $\kappa = 0.79$

Generalization & Qualitative Examples

Beyond KI-MMQA:

	VQAv2	PathVQA
Baseline	81.9%	85.2%
SKIP-NR	**82.1%**	**61.4%**
FLOPs reduction	**72%**	—

SKIP-NR: retrieval suppressed when $d < 0.15$. PathVQA: PubMed corpus, no architecture change.

Per-category OK-VQA highlights:

Category	Δvs. RA-CM3	SKV%
Geography	+4.3	31
Brand & companies	+3.5	22
People & everyday	+3.3	44

Qualitative examples:

Example 1 (entity lookup)

Q: "What is the capital of the flag country?"
QVS: 49/576 tokens (8.5% — flag region)
BSCA: 14/392 edges survive (3.6%)
Output: **"Oslo"** ✓

Example 2 (SKV fast path)

Q: "How many pizzas are visible?"
SKV confidence: $0.93 > \tau_{\text{SKV}}$
Output: **"two"** ✓ (full pipeline skipped)

Failure example

Q: "In what year was this monument erected?"
QVS correct, but entity has 2 Wikipedia mentions.
Retrieval returns nearby (documented) landmark.

Summary

SKIP: Salient Knowledge-Injected Pathways

First system to jointly optimize **visual sparsity** and **retrieval sparsity** under a unified, question-conditional routing policy.

Five Components

QVS	Question-guided visual token pruning
RCSR	Per-region sparse retrieval
BSCA	Bipartite sparse cross-attention
DBC	Difficulty-adaptive budget controller
SKV	Speculative knowledge verification

Theoretical contribution

First non-vacuous sparsity bound for KI-MMQA:
$V' \geq C\sqrt{V}\log(1/\varepsilon)$

Results at a glance:

$4.07\times$
fewer FLOPs

$2.69\times$
faster

$+4.3$ **pts**
InfoSeek vs. RA-CM3

15.2 GB
vs. 24 GB (RA-CM3)

Broader Implications & Future Work

Broader Implication

Knowledge-intensive multimodal inference is **far more sparse** than current systems assume.

Treating sparsity as a *first-class design principle*, jointly across modalities and retrieval, opens headroom that will only grow as models and corpora scale.

SKIP is portable:

- Architecture-agnostic (4× speedup across LLaVA-1.5/1.6, InstructBLIP, mPLUG-Owl2)
- Corpus-agnostic (Wikipedia → PubMed requires only index swap)
- DBC and SKV graftable onto existing RAG-VLMs as lightweight wrappers

Future directions:

1. **Multi-hop retrieval** via k-partite BSCA
2. **Early-exit vision encoder** gated by QVS
3. **Multimodal RCSR** — image+text chunks
4. **Video QA** — temporal sparsity as fourth dimension
5. **Active learning** to expand long-tail entity coverage

Code & Resources

`https://pmlrbd.github.io/skip/`

Stages I–V checkpoints: CC-BY

FAISS index: 240 GB downloadable snapshot

Evaluation harness for all 5 benchmarks

Thank You!

Questions & Discussion

SKIP: Salient Knowledge-Injected Pathways

Noor Islam S. Mohammad Uluğ Bayazıt
Istanbul Technical University, İstanbul, Türkiye

https://pmlrbd.github.io/skip/ · islam23@itu.edu.tr

EMM-QA Workshop @ ICML 2026 • Seoul, South Korea • PMLR 306, 2026