

# MoEZip: Routing-Aware KV Cache Compression for Sparse Mixture-of-Experts LLMs

Minsang Kim<sup>1,2</sup>   Seung Jun Baek<sup>1</sup>   <sup>1</sup> Korea University   <sup>2</sup> SK Telecom

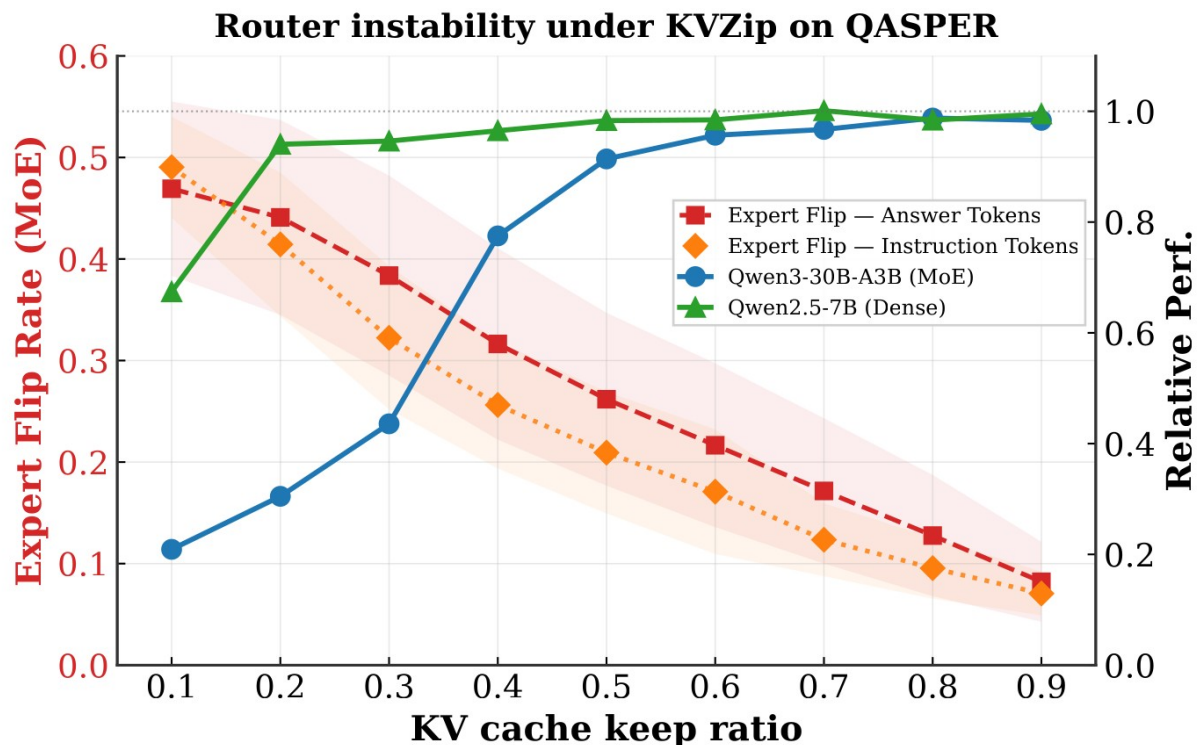
**TL;DR** KV eviction in MoE LLMs **flips which experts fire**. MoEZip keeps the context that keeps routers stable.



[github.com/kmswin1/MoEZip](https://github.com/kmswin1/MoEZip)

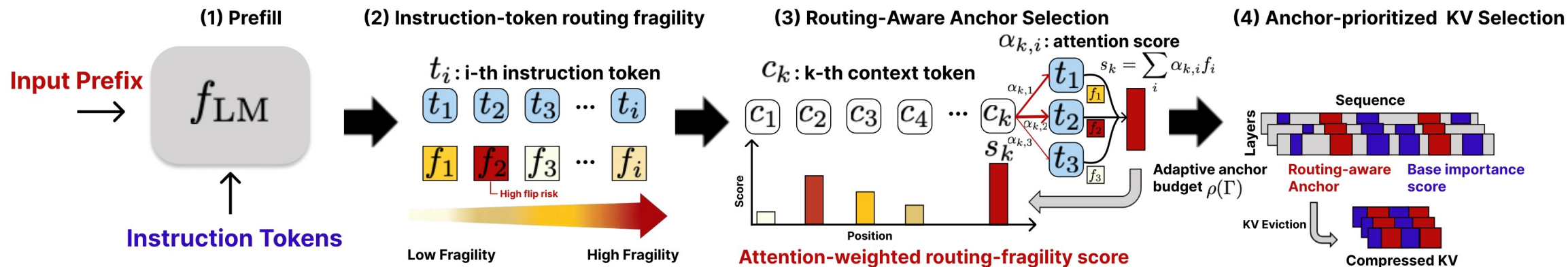
# KV eviction breaks expert routing

Pruning prefix KV perturbs the attention output → router logits → expert choice. We compare experts on the same teacher-forced sequence with full vs. compressed KV (Qwen3-30B-A3B, QASPER, KVZip).



- MoE quality collapses** under aggressive compression; dense Qwen2.5-7B degrades far less.
- Answer-token flips Flip(A)** rise in lockstep with the quality drop.
- Instruction-token flips Flip(I)** track **Flip(A)**, and are observable before eviction, so they serve as our proxy.

# MoEZip: routing-aware anchors



A

## Fisher-trace fragility

Router sensitivity of each instruction token, in  $O(E)$ .

B

## Attention propagation

Context attended by fragile-router tokens scores high.

C

## Adaptive anchor budget

Fragile samples spend more of budget  $B$  on anchors.

D

## Base scorer fills the rest

KVZip / SnapKV / ...; cache size & decoding unchanged.

# Fisher trace: a cheap fragility score

Prop. 1 — under isotropic logit noise  $\delta z$ , the expected KL shift of the routing distribution is proportional to the trace of its Fisher information.

$$\mathbb{E}_{\delta z} \left[ D(p(\cdot|z) \parallel p(\cdot|z + \delta z)) \right] \propto \text{tr } F(z)$$

## Softmax routing

$$f_{t,\ell} = 1 - \sum_{e=1}^E g_{t,\ell,e}^2$$

## Sigmoid routing

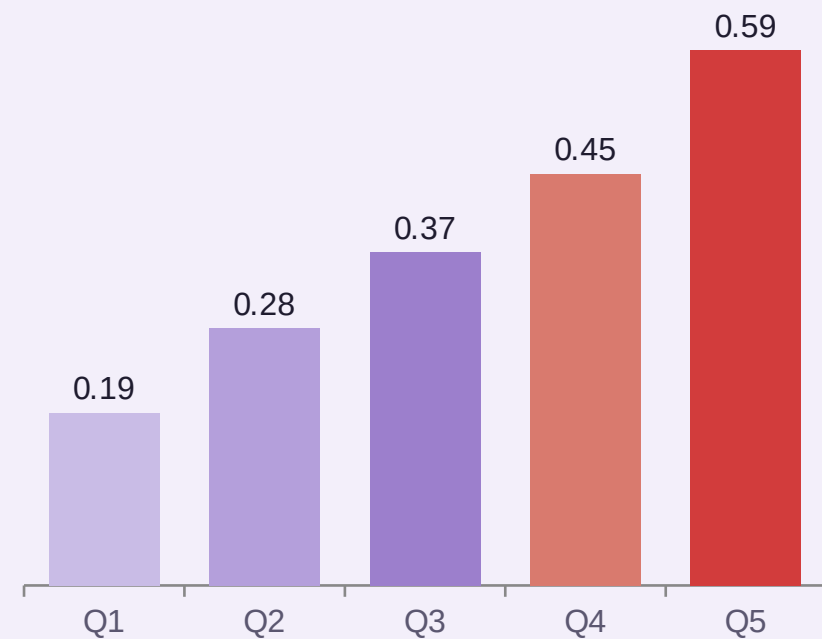
$$f_{t,\ell} = \sum_{e=1}^E g_{t,\ell,e} (1 - g_{t,\ell,e})$$

## Routing-fragility score & anchors

$$r(i) = \sum_{\ell \in \mathcal{M}} \sum_{t \in \mathcal{I}} \bar{\alpha}_{t,\ell,i} f_{t,\ell}, \quad \bar{\alpha}_{t,\ell,i} = \max_h \alpha_{t,\ell,h,i}$$

$$S_{\ell,h} = \mathcal{A}_{\text{anc}} \cup S_{\ell,h}^{\text{base}}(B - B_f; \{1, \dots, L\} \setminus \mathcal{A}_{\text{anc}})$$

Expert-flip rate by Fisher-trace quintile

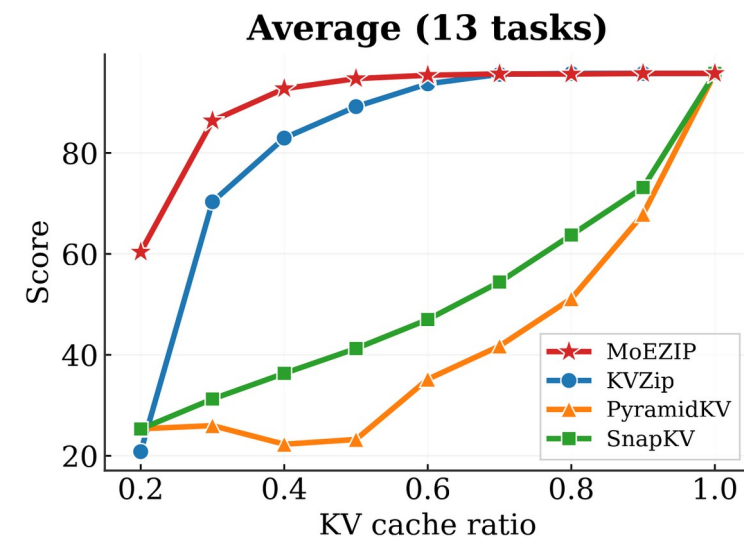
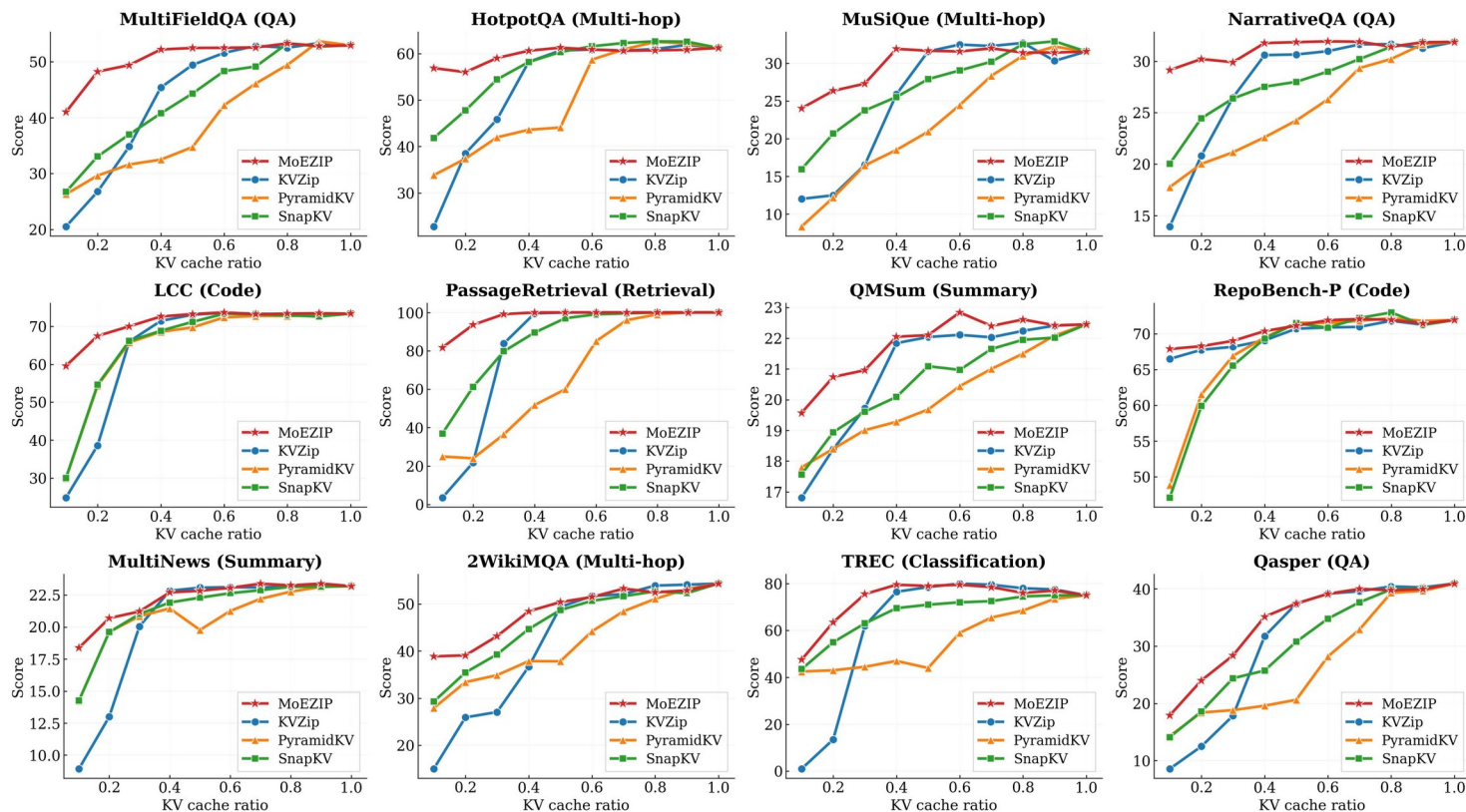


Higher Fisher trace  $\rightarrow$  more real expert flips (avg. of 12 LongBench tasks,  $Q5/Q1 = 2.3\text{--}4.2\times$ ).

# Results: best accuracy at aggressive compression

LongBench · 12 tasks · Qwen3-30B-A3B

RULER · avg.



**+24**

LongBench @10% KV  
(2.3× KVZip)

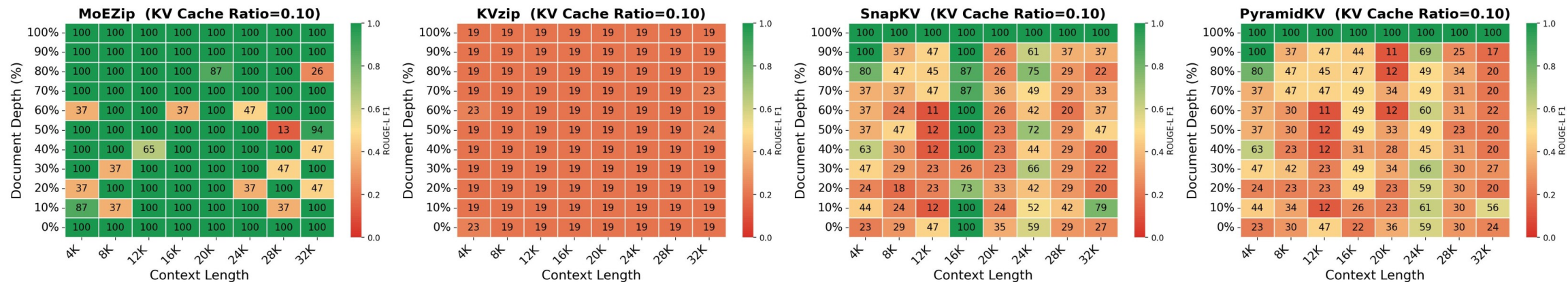
**+20**

RULER @10% KV  
(4.9× KVZip)

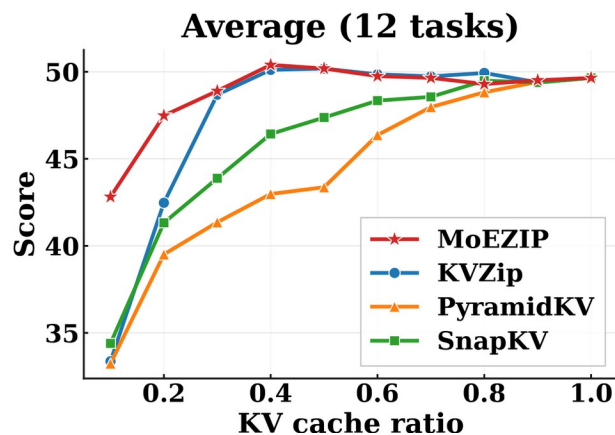
*MoEZip (red) leads at most budgets; KVZip (blue) collapses as the keep ratio shrinks.*

# Robust across positions, routers, and scale

Needle-in-a-Haystack @ 10% KV (context length  $\times$  needle depth)



## Sigmoid router · GLM-4.7-Flash



**27.6 → 44.7**

Qwen3-235B-A22B · 10-task  
LongBench @10% KV

**19.5 → 42.0**

query-agnostic MoEZip-Recon  
(reusable cache) vs. KVZip

**+0.9 s**

scoring overhead;  
same cache & decoding

**Take-away** For MoE LLMs, keep the KV that keeps routers stable, not just the KV that draws attention.