

AudioKV: KV Cache Eviction in Efficient Large Audio Language Models

Yuxuan Wang^{*1}, He Peize², Xiyan Gui³, Xiaoqian Liu⁴, Junhao He⁵, Xuyang Liu⁶, Xuming Hu^{†7}, Linfeng Zhang^{†5}¹ Xidian University · ² UESTC · ³ HUST · ⁴ Northeastern University · ⁵ Shanghai Jiao Tong University
⁶ Sichuan University · ⁷ HKUST (Guangzhou)

* Work conducted during an internship at SJTU. † Corresponding authors.

Preserve the acoustic heads and smooth noisy token scores - aggressive KV compression without long-form ASR collapse.

SCAN TO READ



PAPER / DOI



CODE

10.1145/3767308.3835763

yxwang1215@gmail.com

MM '26

40%

KV RETAINED

60% cache reduction

0.45%

ABS. ACCURACY DROP

Qwen3-Omni-30B

5

AUDIO-LANGUAGE MODELS

Qwen + Gemma

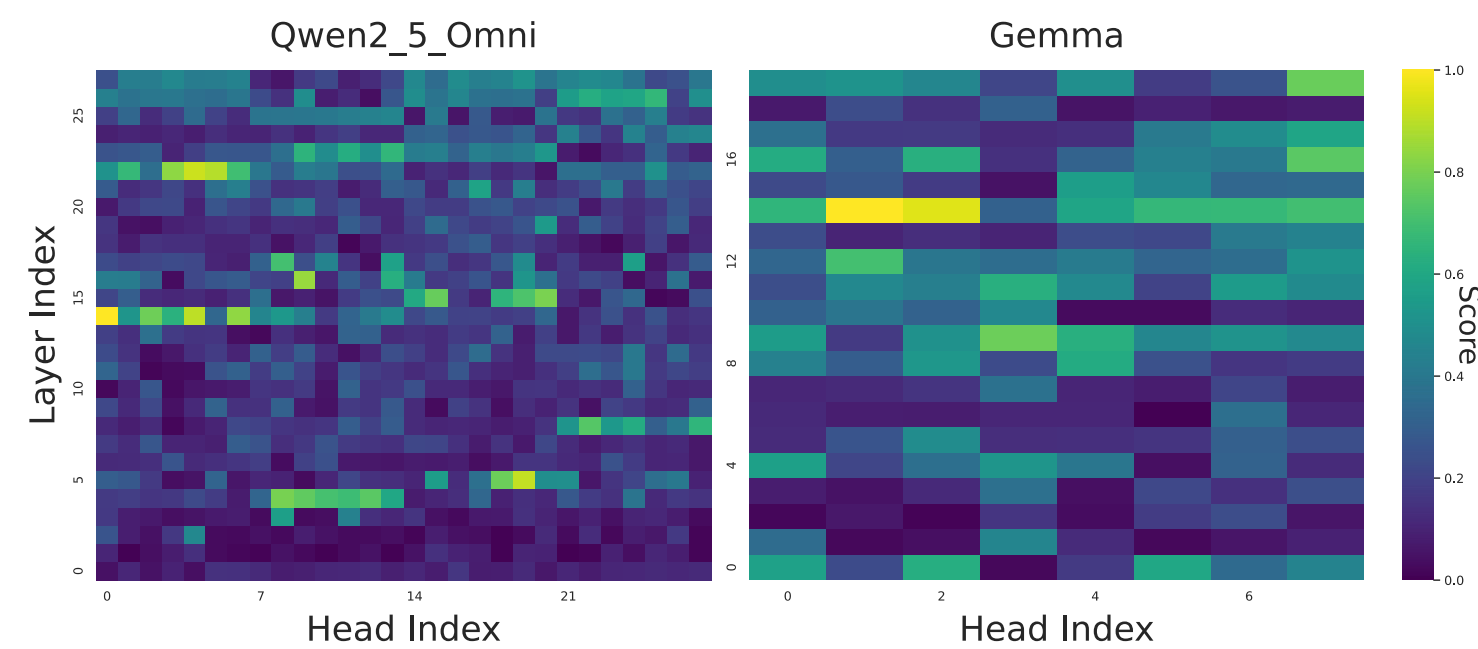
11

ASR / ST / AQA BENCHMARKS

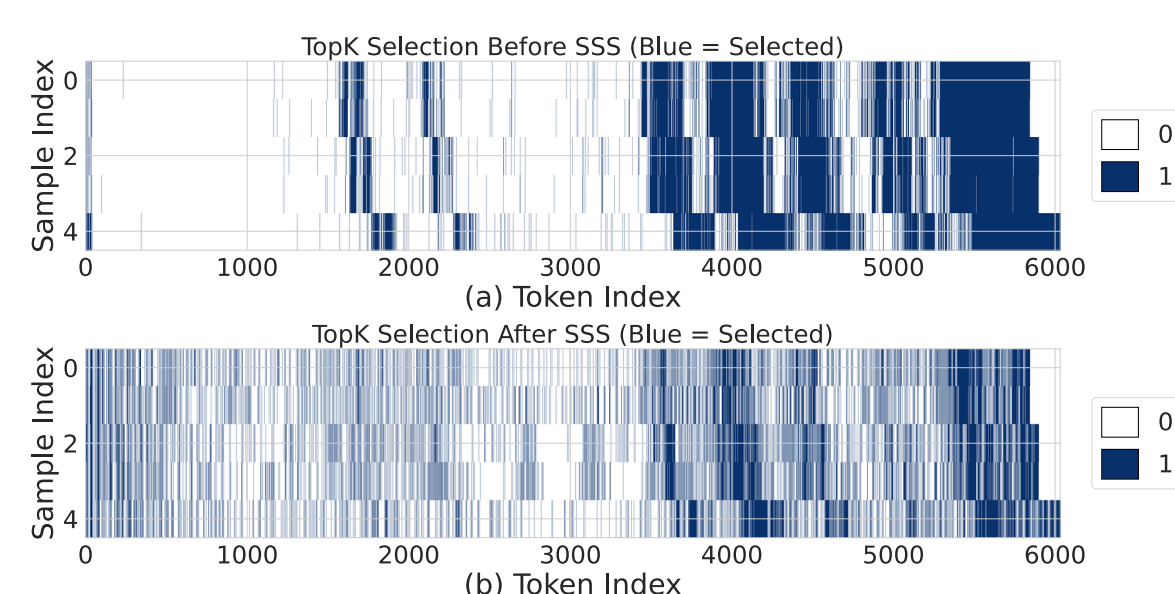
multilingual + long-form

Why audio needs its own KV policy

1 Head specialization: only a small subset consistently grounds decoded words to audio.



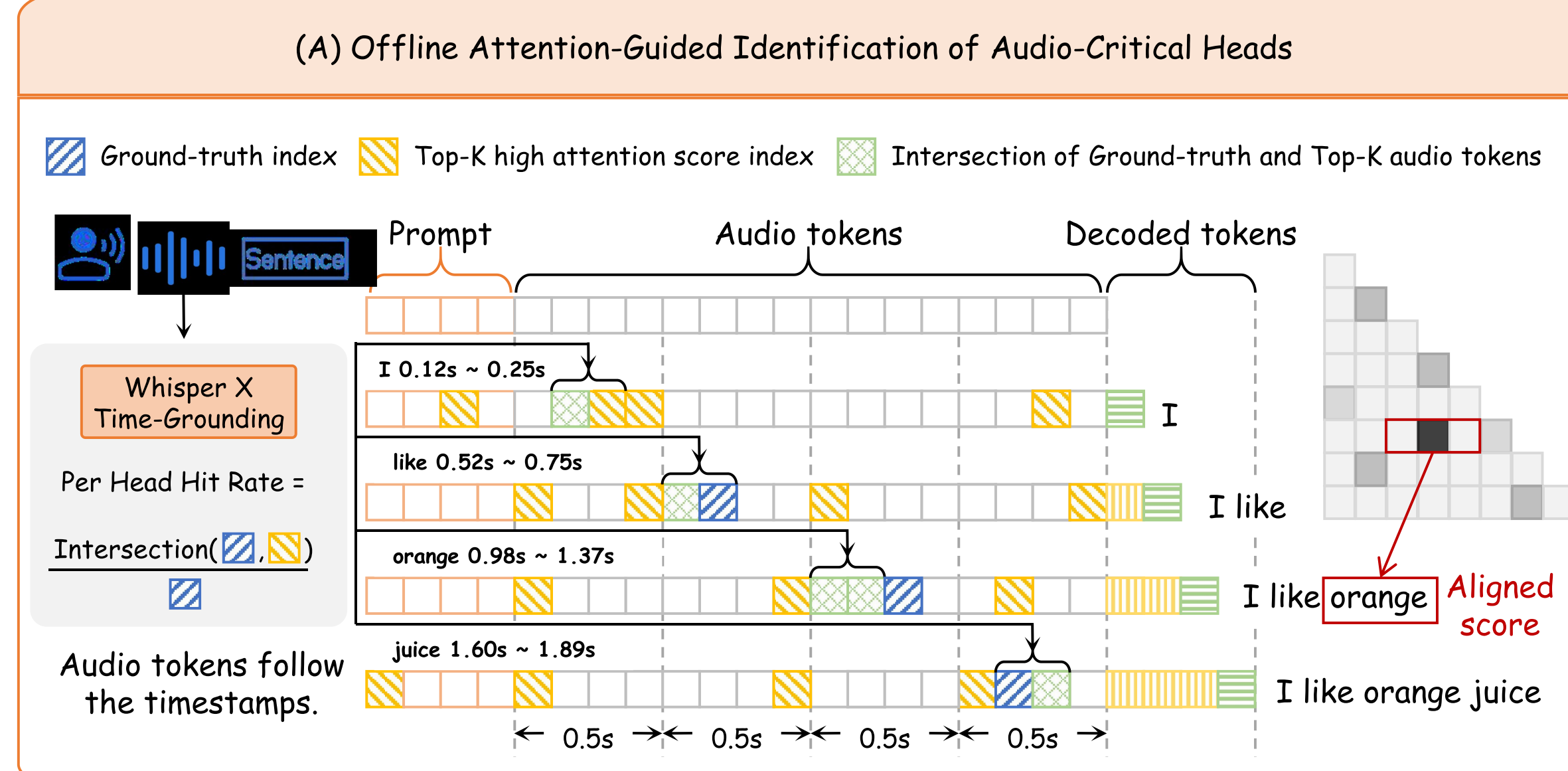
2 Temporal continuity: raw top-K scores cluster in narrow regions and lose global context.



SSS redistributes selected tokens across the full acoustic sequence.

AudioKV: align, allocate, smooth

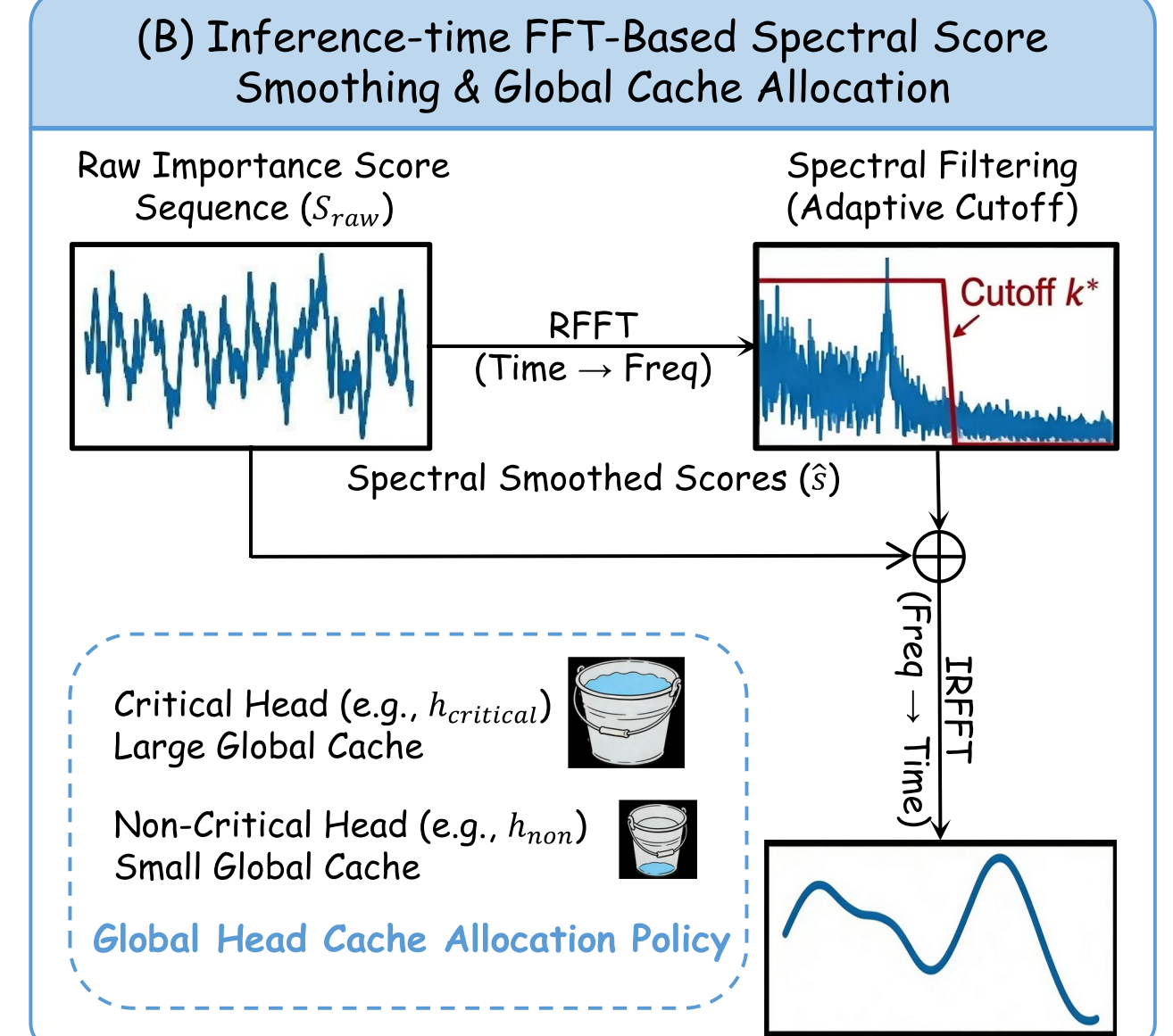
Offline head identification; lightweight inference-time cache allocation and global spectral filtering.



1 Identify audio-critical heads

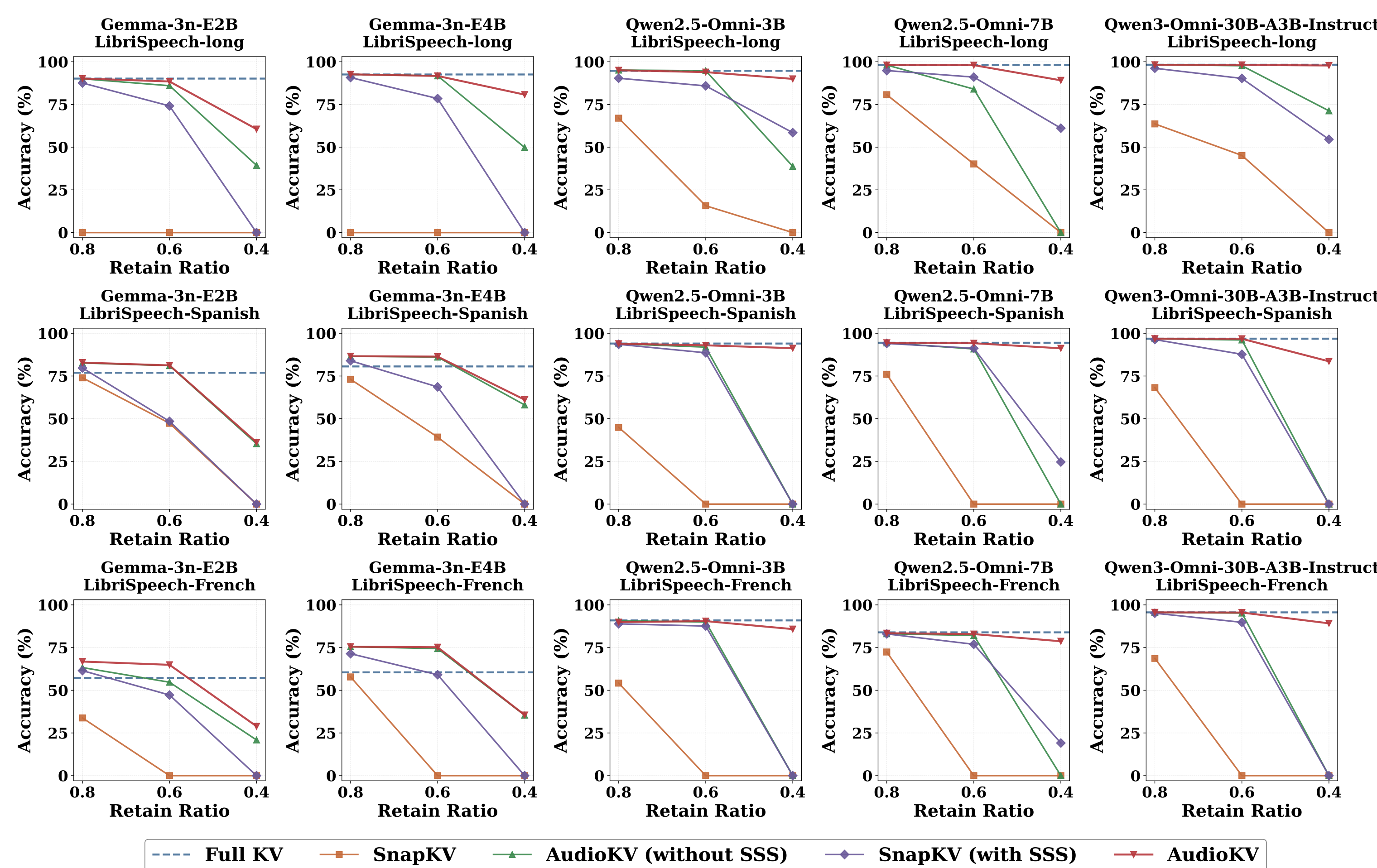
WhisperX-aligned word spans + per-head top-K attention hit rates.

2 Allocate cache + apply SSS

 $\hat{s} = (1-\alpha)s + \alpha \cdot \text{IRFFT}(\text{LP}_k^*(\text{RFFT}(s)))$ with adaptive spectral cutoff.

Stable under aggressive compression

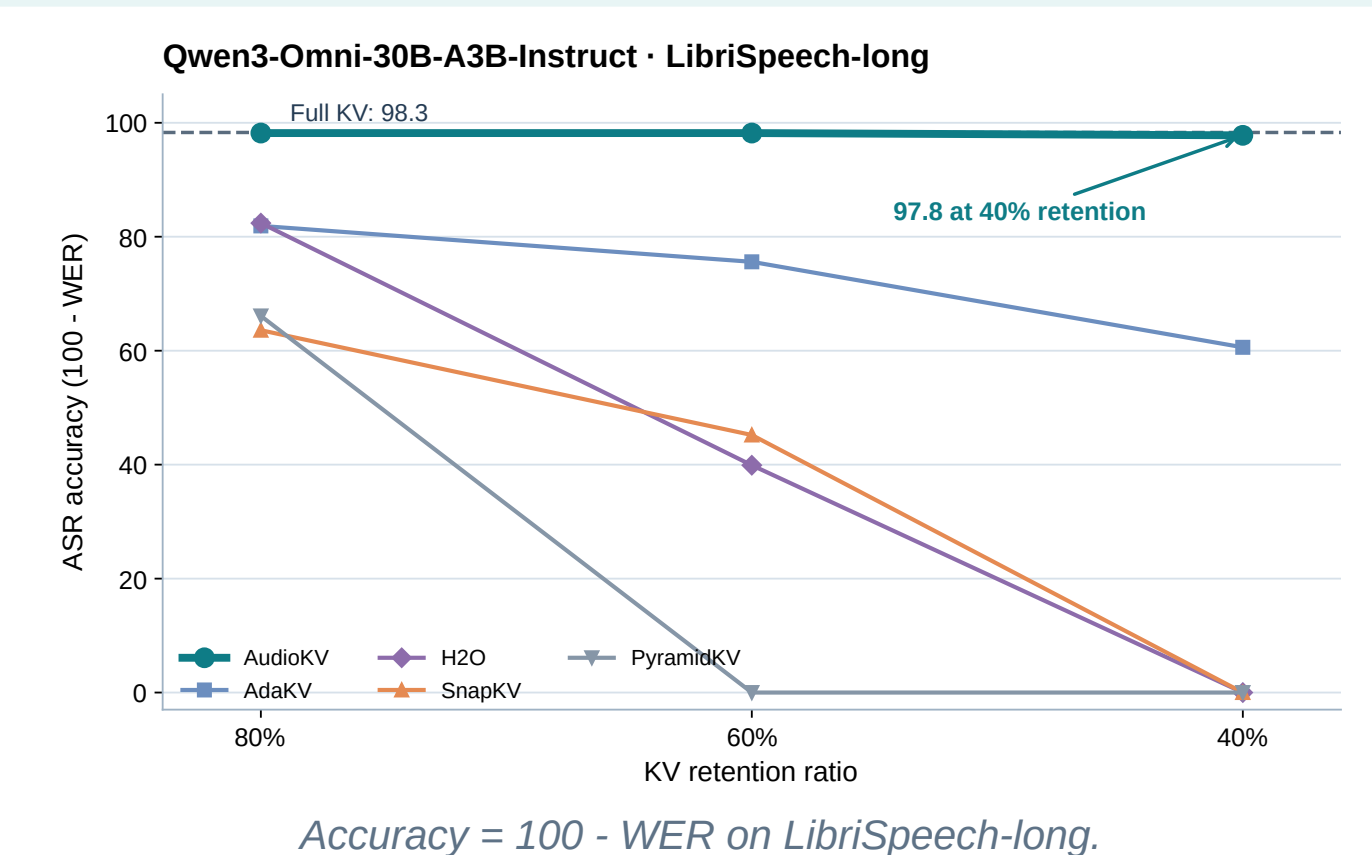
Five LALMs · long-form and multilingual ASR · speech translation · retain ratios 0.8 / 0.6 / 0.4



Competing methods frequently collapse at 40% retention; AudioKV remains robust across Qwen and Gemma.

Qwen3-Omni long-form ASR

97.8% accuracy at 40% KV retention | Full KV: 98.3%



Generalizes to Audio QA

AudioKV score and gain over the best baseline at a 40% KV budget.

Task	3B	Δ	7B	Δ
S-Alpaca	57.6	+1.5	66.7	+1.5
Llama-Q	75.0	+0.3	77.3	+0.6
S-Trivia	31.9	+1.3	42.2	+1.4
S-Web	46.1	+0.8	55.5	+0.4

Four UltraEval-Audio tasks; Gemini 2.5 Flash judge.

Failure mode: repetition collapse

A 232.62-second LibriSpeech sample with a 512-token KV budget.

SnapKV WER = 93.62%

AudioKV WER = 1.09%

Case Study · LibriSpeech Ch. 13754 · Duration: 232.62 s · KV budget: 512 tokens

SnapKV WER = 93.62 %

They knew no north, no south... When Brigham Young passed from earth, he was mourned as Moses of Israel. He had led the people into the land of holy promise, and divided unto them their inheritances.

► Failure onset

"...future states shall recognize"

every future state shall recognize

every future state shall recognize

every future state shall recognize

every future state shall recognize

[repeats xhundreds ...]

AudioKV WER = 1.09 %

They knew no north, no south... When Brigham Young passed from earth, he was mourned as Moses of Israel. He had led the people into the land of holy promise, and divided unto them their inheritances.

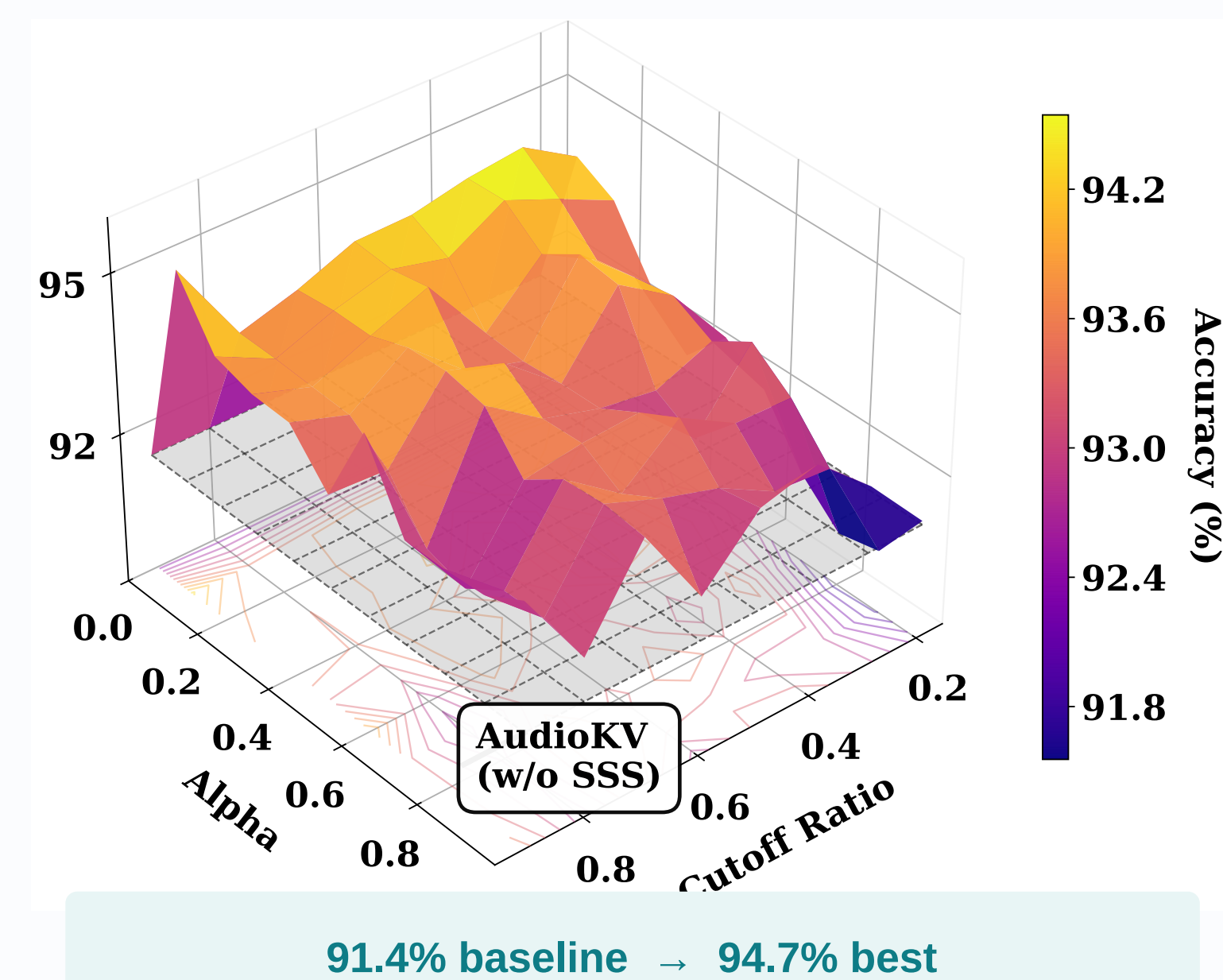
► Coherent continuation

"...be recognized every sanctified and authorized relationship existing here, of parent and child, brother and sister, husband and wife. Contracts of marriage must be sanctioned by a power greater than that of earth."

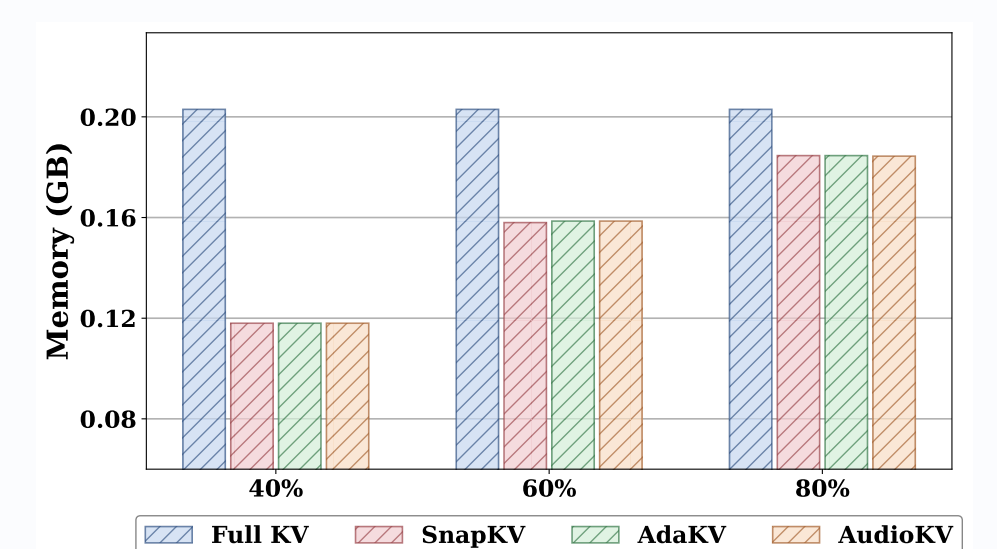
AudioKV preserves discourse-level constraints instead of entering a self-reinforcing repetition loop.

SSS is robust - and costs no extra cache

Spectral smoothing remains effective across a wide parameter region



Same memory budget



BETTER CACHE UTILITY

SnapKV, AdaKV, and AudioKV occupy nearly identical cache sizes at each retain ratio; the gain comes from selecting better context.

Takeaways

01 Allocate by acoustic relevance

Word-aligned attention hits reveal which heads actually ground decoded speech.

02 Smooth globally, not locally

FFT-based SSS suppresses high-frequency noise while preserving global semantic trends.

03 Avoid catastrophic collapse

Across Qwen and Gemma, AudioKV converts tight KV budgets into reliable long-form inference.