MEKV: CONTROLLING CONCENTRATION IN QUERY-AGNOSTIC VISUAL KV COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models require compressing visual KV caches for efficient multi-query serving. When compression happens before user queries are known, a natural approach is to retain tokens according to a relevance signal—yet we show that high-quality signals can degrade accuracy when over-concentrated, starving unseen questions of needed evidence. We develop a concentration theory along the maximum-entropy temperature path, proving that a surrogate accuracy objective is single-peaked with the optimum given by a moment-matching condition. This yields MEKV, a simple, training-free method that separates signal quality from concentration control via a single temperature parameter. The theory further predicts that the right strategy is model-specific: models whose optimum requires strong tempering need concentration control first, while those near the raw signal benefit more from fine-grained token selection. On full validation across three benchmarks, MEKV improves over coverage baselines by up to 17 pp, with the temperature transferring across tasks without re-tuning. The prediction is borne out by a strategy reversal: same-signal per-head top-k falls far behind MEKV on Qwen yet excels on InternVL, and placing per-head selection inside calibrated quotas repairs the failure while preserving the gains.

1 INTRODUCTION

Serving a vision-language model on document-understanding tasks often involves multiple user queries against the same image, motivating shared-context caching (Kwon et al., 2023; Zheng et al., 2024; Li et al., 2025). Each query attends to visual KV entries that may represent thousands of tokens under dynamic-resolution encoding (Wang et al., 2024; Chen et al., 2024b). Compressing this visual KV cache is a practical necessity, but the compression must happen *before* the user's questions are known — *query-agnostic visual KV eviction*.

A natural approach is to score visual regions using synthetic proxy questions (SQA) and retain those with the highest scores. This produces a relevance signal that positively correlates with true downstream evidence needs. Yet we observe a counterintuitive failure: on some models, retaining tokens strictly according to this signal *underperforms* a simple coverage baseline that allocates tokens proportionally to spatial density (Figure 1). The signal is good; the problem is how aggressively it is used.

We trace this failure to *over-concentration*. When the retention distribution tracks the raw signal too closely, it places too much budget on a few high-scoring regions and starves others. The unseen query may need exactly those starved regions. Coverage baselines avoid this by spreading tokens uniformly, but they waste budget on regions the signal correctly identifies as unimportant. The question is: *what is the right concentration?*

We develop a theoretical framework that answers this question. Along the maximum-entropy temperature path connecting the coverage distribution and the raw signal distribution, we prove that a surrogate accuracy objective is single-peaked, with the optimum determined by a moment-matching condition. This yields MEKV, a training-free method that controls concentration with a single temperature, requires only a one-time scoring pass per image, and makes no model modification.

The theory also makes a sharper, architecture-level prediction. The cost of leaving concentration uncontrolled depends on how far the model's optimum λ^* lies from the untempered signal: where

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

$\lambda^* \ll 1$, concentration control should dominate other design choices, while near $\lambda^* \approx 1$ finer-grained token selection has room to help. Same-signal SnapKV-style top-k shows exactly this reversal on DocVQA. It falls 14–16 pp below MEKV on Qwen, whose optimum is strongly tempered, yet leads on InternVL, whose optimum lies near the untempered end (Figure 6). Placing per-head selection inside calibrated quotas repairs the failure on Qwen by up to 14 pp while preserving the gains on InternVL.

We evaluate MEKV across three architectures and three benchmarks. Our contributions:

1. **Over-concentration as a failure mode.** Across three architectures, informative relevance signals lose accuracy when over-concentrated, and which endpoint fails depends on the architecture.
2. **Concentration theory.** We prove single-peakedness of a surrogate accuracy objective along the temperature path and derive the moment-matching optimum. The theory explains how the optimal concentration shifts across models and budgets.
3. **Temperature-controlled allocation.** A simple, training-free method that improves over coverage baselines by up to 17 pp on full validation.
4. **Architecture reversal.** The framework predicts, and experiments confirm, that same-signal token top-k fails where $\lambda^* \ll 1$ and helps where $\lambda^* \approx 1$; placing fine-grained selection inside calibrated quotas repairs Qwen while preserving InternVL's gains.

2 Related Work

KV cache compression. Attention-based retention uses persistent importance and head statistics (Zhang et al., 2023; Liu et al., 2023; Ge et al., 2024). Other approaches adapt head budgets (Fu et al., 2025), quantize KV entries (Liu et al., 2024), retrieve query-relevant entries (Tang et al., 2024), or separate retrieval and streaming heads (Xiao et al., 2025).

Visual token compression in VLMs. FastV (Chen et al., 2024a) and visual pruning or merging methods (Bolya et al., 2022; Zhang et al., 2024; Xing et al., 2025; Yang et al., 2025) reduce tokens using attention or similarity cues. SnapKV (Li et al., 2024) selects KV entries per head, PyramidKV (Cai et al., 2024) varies layer budgets, and MixKV (Liu et al., 2026) adds head-wise diversity to importance scores. Diversity-based pruning (Alvar et al., 2025) and multimodal KV compression (Wan et al., 2024) likewise address redundancy and coverage. We study how a query-agnostic cache should distribute evidence for unseen questions, and why signal reliance shifts across architectures and budgets.

Query-agnostic compression. KVzip (Kim et al., 2026) builds a reusable cache through context reconstruction. Its reconstruction target is the textual context, which visual tokens do not provide. Complementary work studies robust eviction (Guo et al., 2026; An et al., 2026) and future-token importance (Ahn et al., 2026). Chen et al. (2026) show that current attention can misrank evidence needed in future dialog turns. Our focus is concentration control: even informative query-agnostic signals require calibrated retention.

3 Over-Concentration in Query-Agnostic Eviction

Task. Given a VLM and an image, construct a compressed visual KV cache of at most r fraction of the original visual tokens, before any user query is known. All subsequent queries are served from this single cache.

Signal. We use Synthetic Question Attention (SQA): generate a small set of proxy questions about the image, capture the model's attention distribution over visual tokens during these forward passes, and aggregate into a per-region score. The SQA signal correlates positively with true downstream attention needs (measured ρ: 0.66–0.78 across models) but is constructed without access to actual evaluation questions or answers.

Spatial bins. Visual tokens are grouped into a grid over the visual token layout ($16\times16 = 256$ bins for Qwen and InternVL; $4\times4 = 16$ bins for Gemma; Appendix B.4). Each bin's SQA score is computed from the tokens it contains. The allocation decision operates at bin granularity.

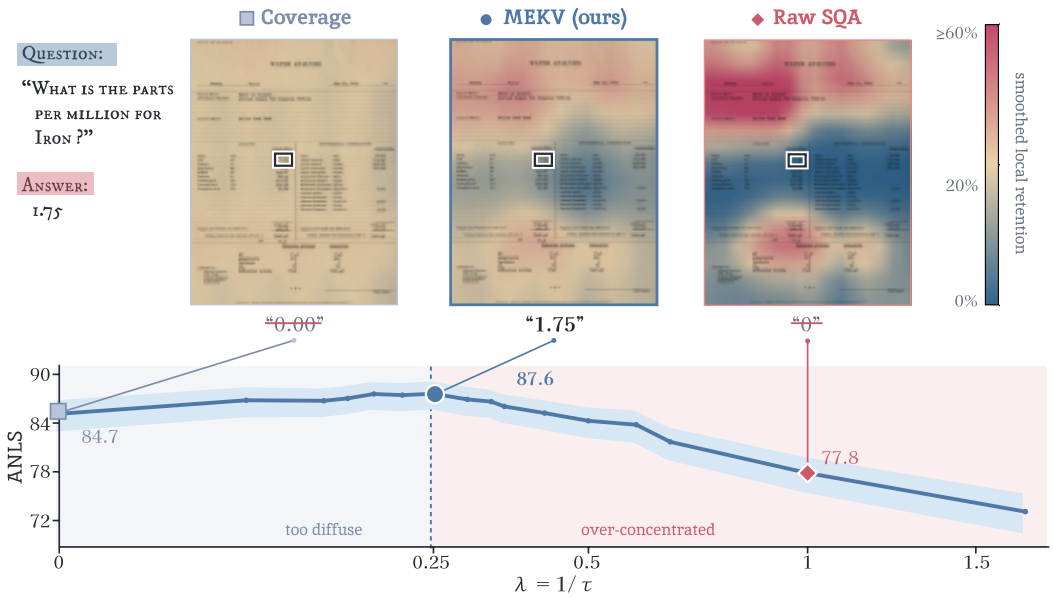


Figure 1: An example document query and the over-concentration effect. MEKV retains the evidence needed to answer the highlighted iron-content question, while the coverage and raw-SQA endpoints miss it. The Qwen DocVQA response below rises to an interior optimum and falls under raw SQA at 20% retention.

Models. We use Qwen3.5-9B (Team, 2026) (8 full-attention KV layers), InternVL3.5-8B (Wang et al., 2025) (36 full-attention layers), and Gemma-4-12B-it (Team et al., 2026) (8 of 48 layers are full-attention, carrying 4.76% of visual KV capacity). These models span distinct visual KV geometries.

Retaining tokens in proportion to raw SQA scores can be actively harmful. Figure 1 illustrates how concentration changes the spatial retention pattern.

Table 1: Over-concentration across architectures (DocVQA development, 20% budget, ANLS × 100). Coverage is the stratified baseline; λ^* is the best tested path point.

Model	ρ(SQA, oracle)	Raw SQA	Coverage	λ^*	Best
Qwen3.5-9B	0.66	77.8	84.7	0.25	87.6
InternVL3.5-8B	0.78	74.0	70.6	0.42	75.3
Gemma-4-12B (full8)	0.71	61.6	62.5	0.026	63.1

The failure is systematic (Table 1). In all three architectures the signal is positively aligned with oracle attention, yet the best allocation lies strictly between the two endpoints. Which endpoint fails worse depends on the architecture: raw SQA on Qwen and Gemma, coverage on InternVL. This difference foreshadows the regime dependence studied in §5.3.

4 A Concentration Theory of Retention

4.1 The Temperature Path and Single-Peakedness

Let u denote the token-uniform distribution over spatial bins, s the SQA score distribution, and $\ell_i = \log s_i - \log u_i$ the log-ratio (signal strength per bin). The temperature family is:

$$p_\lambda \propto u \cdot \exp(\lambda \cdot \ell), \quad \lambda = 1/\tau \in [0, \infty)$$

At $\lambda = 0$ (τ=∞), $p_\lambda = u$ (pure coverage). At $\lambda = 1$ (τ=1), $p_\lambda = s$ (uncontrolled SQA). This family traces the *maximum-entropy path*: among all distributions with a given expected log-ratio moment $\mathbb{E}_p[\ell]$, the member p_λ minimizes KL divergence to coverage (Lemma A.1). This construction is the exponential tilting, or I-projection, of coverage onto a moment constraint (Jaynes, 1957; Csiszár, 1975). The continuous concentration $C(\lambda) = \mathrm{KL}(p_\lambda \| u)$ is a strictly increasing bijection of λ (Lemma A.2), so concentration and temperature are interchangeable coordinates.

Theorem 1. Let a be an oracle distribution on $\mathrm{supp}(u)$ and ℓ a finite, non-constant log-ratio on that support. The surrogate accuracy $-\mathrm{KL}(a\|p_\lambda)$ is strictly concave in $\lambda \geq 0$, so it has at most one maximizer. Any finite interior maximizer satisfies the moment-matching condition

$$\mathbb{E}_{p_{\lambda^*}}[\ell] = \mathbb{E}_a[\ell].$$

Under the sandwich condition $\mathbb{E}_u[\ell] < \mathbb{E}_a[\ell] < \mathbb{E}_s[\ell]$, the maximizer exists and lies in $(0, 1)$.

That is, the optimal allocation matches the signal moment of the oracle. Coverage is too diffuse (moment below oracle); uncontrolled SQA is too concentrated (moment above oracle). Temperature provides continuous control, and an interior optimum exists whenever the oracle lies between coverage and raw signal. We verify in §5 that single-peakedness transfers to the real metrics.

Corollary 1 (four regimes). The optimum is: coverage-optimal ($\lambda^* = 0$) when the oracle moment is below coverage; interior when the sandwich holds; calibrated-SQA ($\lambda^* = 1$) when moments match exactly; and beyond-SQA ($\lambda^* > 1$) when the oracle is sharper than SQA. We empirically observe three of these four regimes (§5).

Corollary 2 (budget shift). If tightening the budget increases the oracle's log-signal moment — intuitively, scarcer budget makes the remaining evidence more concentrated — then the optimal concentration increases with budget tightness: $\mathrm{KL}^*(r_1) > \mathrm{KL}^*(r_2)$ for $r_1 < r_2$.

4.2 Off-Path Structure and an Architecture-Level Prediction

Proposition 1 (Pythagorean concentration decomposition). Let q be an allocation supported on $\mathrm{supp}(u)$, with finite log-ratios ℓ_i on that support. Suppose a finite $\lambda(q) \geq 0$ satisfies $\mathbb{E}_q[\ell] = \mathbb{E}_{p_{\lambda(q)}}[\ell]$. With $u = p_0$, its total concentration decomposes as

$$\mathrm{KL}(q\|u) = \mathrm{KL}(p_{\lambda(q)}\|u) + \mathrm{KL}(q\|p_{\lambda(q)}).$$

The first term is the *on-path concentration*, C_{sig}: the KL divergence of the moment-matched family member from the coverage distribution u. The second is the *off-path structural concentration*, C_{exc}: the KL divergence of q from that family member. Together, they separate total concentration into a component along the signal direction and a component beyond the temperature family. C_{exc} measures how far a policy's spatial allocation departs from the temperature family; it is a geometric coordinate, and whether such departures help is an empirical question addressed in §5.3.

Empirical concentrations in §5 are measured on realized integer allocations (Appendix A.6).

Moment matching also quantifies the surrogate cost of an off-optimum path point. With log-partition function $\psi(\lambda) = \log \sum_i u_i e^{\lambda \ell_i}$, we have $A(\lambda) = \mathrm{const} + \lambda \mathbb{E}_a[\ell] - \psi(\lambda)$ and $\psi'(\lambda^*) = \mathbb{E}_a[\ell]$, so

$$\begin{aligned} A(\lambda^*) - A(\lambda) &= \psi(\lambda) - \psi(\lambda^*) - \psi'(\lambda^*)(\lambda - \lambda^*) \\ &\approx \tfrac{1}{2} \mathrm{Var}_{p_{\lambda^*}}[\ell]\,(\lambda - \lambda^*)^2. \end{aligned}$$

This Bregman divergence gives a locally quadratic on-path cost in $\lambda(q) - \lambda^*$. Policies that do not control concentration have no mechanism to place $\lambda(q)$ near λ^*. We predict temperature control to be the priority when $\lambda^* \ll 1$, where uncontrolled policies are most likely to overshoot, while near $\lambda^* \approx 1$ fine-grained per-head, per-layer selection has greater opportunity to improve on spatial allocation. Section 5.3 tests this prediction by comparing MEKV and same-signal token top-k across two architectures with different calibration regimes.

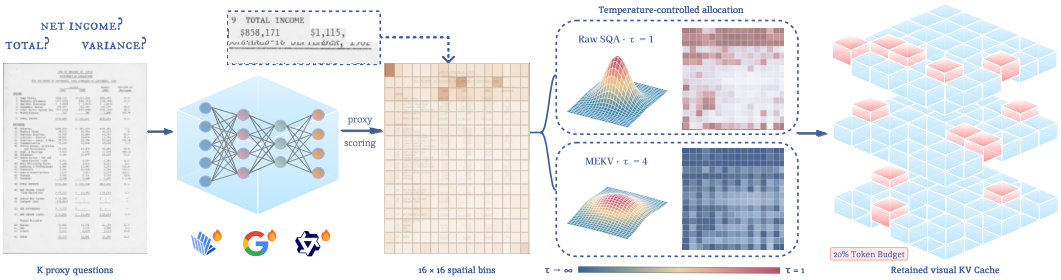


Figure 2: MEKV pipeline. Synthetic proxy questions produce spatial relevance scores before user queries are known. Temperature controls how strongly allocation follows this signal: tempering broadens retention while preserving informative regions. The resulting token allocation constructs a compressed visual KV cache that is reused for subsequent queries.

4.3 MEKV

The theory suggests a direct recipe: move along the temperature path and stop at the calibrated point. MEKV converts SQA scores into a temperature-controlled retention distribution. Figure 2 shows the scoring, allocation, and retention stages.

Step 1: Score. For each image, generate a small set of synthetic questions and capture the model's attention over visual tokens. Aggregate into a per-bin SQA score s_b over the grid.

Step 2: Allocate. Given budget r and temperature τ:

$$p_b(\tau) \propto u_b \cdot \exp\left(\frac{\ell_b}{\tau}\right), \quad \ell_b = \log \max(s_b, \epsilon) - \log \max(u_b, \epsilon)$$

Convert $p_b(\tau)$ to integer token counts under the total budget, capping each bin at its token count. At $\tau = 1$, this recovers the raw SQA distribution; as $\tau \to \infty$, it approaches stratified coverage.

Step 3: Retain. Within each bin, retain the allocated number of tokens by uniform sampling without replacement.

The temperature τ is selected once per model on a coarse development grid (DocVQA-300) and frozen for all budgets and benchmarks; the finer sweeps in §5 serve analysis only. For Qwen, which has 8 KV layers requiring separate eviction, the allocator operates per-layer with a shared τ; the multi-layer extension of Theorem 1 applies exactly (Appendix A).

5 Experiments

We evaluate DocVQA (Mathew et al., 2021) with ANLS (Biten et al., 2019), ChartQA (Masry et al., 2022) with relaxed accuracy, and TextVQA (Singh et al., 2019) with accuracy. The primary evaluation uses full validation sets with image-level paired bootstrap (10,000 resamples).

Baselines: (1) *Stratified coverage*: allocate proportional to spatial density, the coverage endpoint of our temperature path. (2) *Random coverage*: uniform random retention. (3) *SnapKV-style*: per-head, per-layer token-level top-k using the same SQA signal as observation window — the strongest same-signal baseline, and simultaneously the sharpest ablation of our allocation policy. (4) *PyramidKV-style*: layered pyramid allocation with SQA. (5) *FastV*: as-published decoder pruning in no-question mode. (6) *Real-query oracle*: per-query attention top-k, an oracle reference that uses the actual questions.

All baselines are budget-matched per image (Appendix B.3).

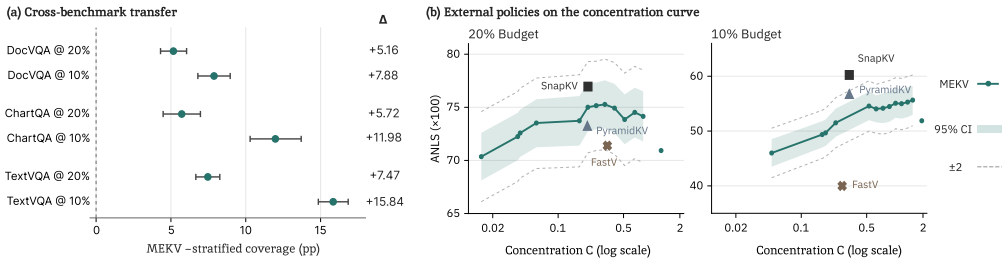


Figure 3: Task transfer and external-policy structure on InternVL. **(a)** MEKV gains over stratified coverage on full validation across DocVQA, ChartQA, and TextVQA using the DocVQA-selected temperature without re-tuning. **(b)** External policies on the DocVQA development concentration–accuracy plane at 20% and 10% budgets; shading denotes the MEKV bootstrap band and dashed curves the ± 2 pp tolerance.

5.1 MEKV Across Tasks and Budgets

Figure 3(a) summarizes InternVL3.5-8B full validation at $\tau = 2$; Appendix G reports absolute scores, paired comparisons, and oracle headroom.

Task transfer. On InternVL, the temperature selected on DocVQA ($\tau = 2$) transfers directly to ChartQA and TextVQA without re-tuning, so a single temperature serves all three benchmarks. Gains increase as budget tightens, consistent with Corollary 2. On Qwen3.5-9B, MEKV improves over stratified coverage by +2.9 on DocVQA at 20% budget and by +10.8 / +17.0 on TextVQA at 20% / 10% budget.

Why the exponential path. Replacing the temperature path with optimized linear mixing $p_\alpha = (1-\alpha)u+\alpha s$ costs 1–2 pp on Qwen and matches it on InternVL (Appendix H). This is the pattern the theory anticipates: for any given signal moment, the exponential path spends the least concentration (Lemma A.1), an advantage that matters most when the optimum is strongly tempered. MEKV stays within $\pm 0.4\%$ of stratified eviction in serve time across 24 timing configurations and reduces peak visual KV cache memory by 77–89% relative to full cache.

5.2 The Theory's Predictions in the Data

Every tested curve is single-peaked, consistent with Theorem 1. Figure 4 shows the DocVQA temperature sweeps; we test for secondary local maxima on all eight measured curves — three architectures at two budgets, plus Gemma's all-layer arm. None shows a significant secondary peak (Appendix D.3).

Optimal concentration rises as budgets tighten, as predicted by Corollary 2 (Figure 5). The trend is monotone on Qwen and InternVL across three budgets (20%, 10%, 5%). We use the concentration relationship fitted on Qwen to predict InternVL's optimal concentration at a held-out 5% budget. The predicted $\ln C^*$ is 0.709 against an interpolated measurement of 0.836, a log-space error of 0.13 (within a factor of 1.14 in concentration). This out-of-sample result indicates that the concentration framework transfers across model families. On Gemma-4-12B, both budgets show interior optima on the primary arm, with the budget shift reaching approximately $54\times$ in concentration ratio. The 10% optimum lies beyond the untempered SQA point ($\lambda^*_{\text{grid}} \approx 1.19$), an instance of Corollary 1's beyond-SQA regime — under a tight budget on its full-attention layers, sharpening past the raw signal is optimal.

The moment-matching estimator recovers the architecture ordering ($\hat{\lambda}_{\text{IV}} = 0.996 > \hat{\lambda}_{\text{Qwen}} = 0.61$) and places both optima inside the path, matching the regime assignment used in §5.3. It solves $\psi'(\hat{\lambda}) = \widehat{\mathbb{E}}_a[\ell]$ from oracle attention moments measured on development data, with no free parameters. Its gap to the grid optimum narrows as budgets tighten, reaching 0.004 in log space on InternVL at 5% (Figure 8, Appendix C). A single scalar solve thus supplies the regime diagnosis that the reversal requires.

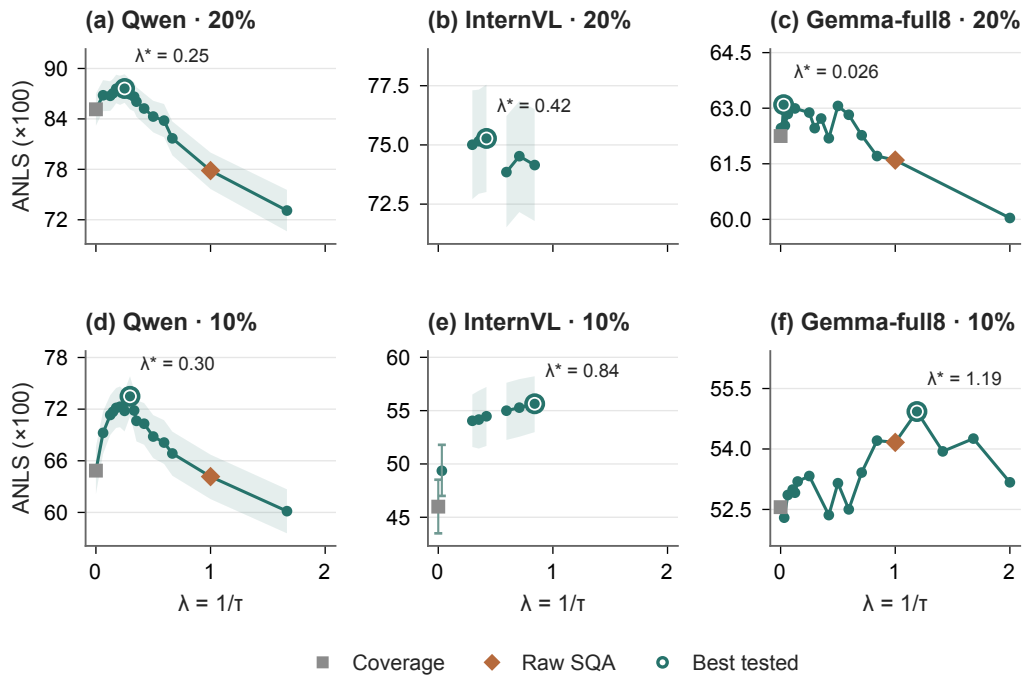


Figure 4: Accuracy along the temperature path (DocVQA development). Every curve peaks inside the path, and tighter budgets favor greater signal reliance (top to bottom). Rings mark the best tested setting; shading shows 95% intervals.

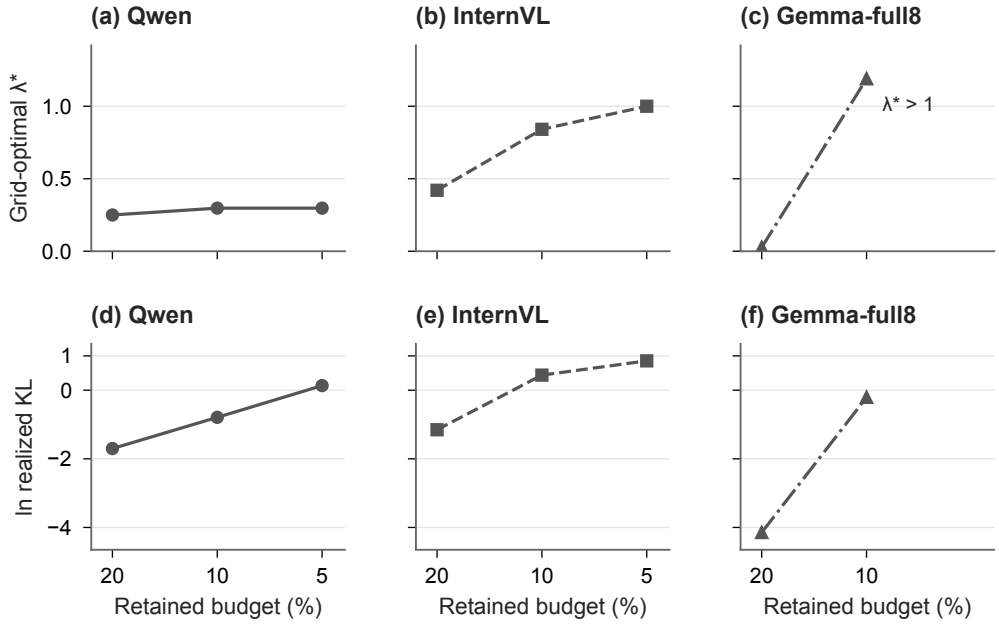


Figure 5: Optimal temperature and realized concentration shift with the retained budget. Top: λ^*; bottom: log realized KL. Gemma-full8 enters the beyond-SQA regime ($\lambda^* > 1$) at 10% budget.

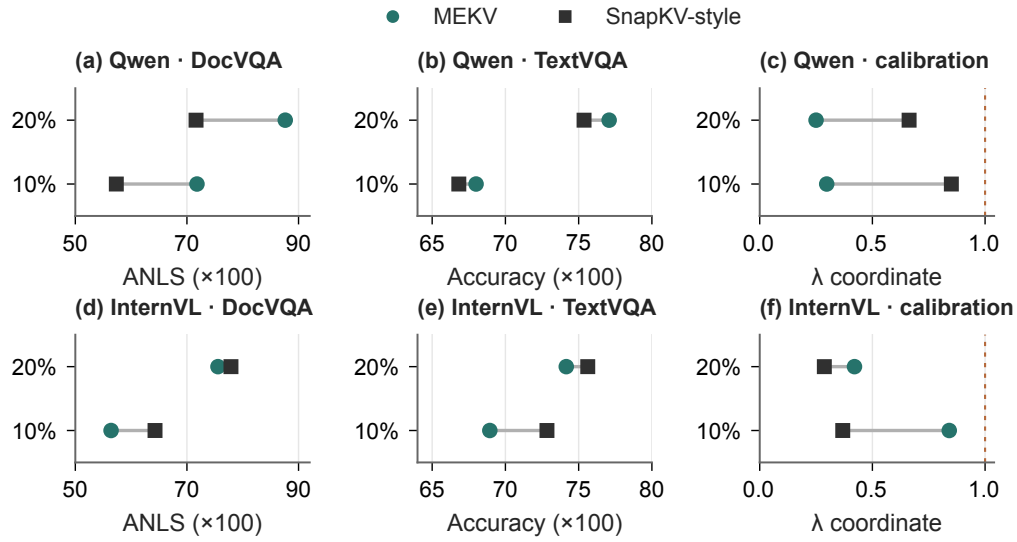


Figure 6: Strategy reversal on DocVQA (a,d) and TextVQA (b,e), with DocVQA calibration diagnostics (c,f). Calibration panels separately show development λ^* (circles), SnapKV $\lambda(q)$ (squares), and raw SQA (dashed). Absolute scores and evaluation splits: Table 13.

5.3 The Architecture Reversal

We test the architecture-dependent prediction in §4.2 using published methods adapted to query-agnostic compression with the same SQA signal and matched budgets. Figure 6 links the performance comparisons to the DocVQA development calibration diagnostics. The concentration framework organizes external methods (Figure 3b). FastV at 10% falls 12.4 pp below the curve, the cost of token pruning under matched budgets, while SnapKV lies above its peak at both budgets. SnapKV's positive off-path concentration under Proposition 1 identifies structure beyond the temperature family; Appendix E gives the decomposition and accuracy gains.

The ordering reverses across architectures. On DocVQA, MEKV scores 87.6/71.8 versus SnapKV's 71.6/57.3 on Qwen — a gap of 14–16 pp — but 75.6/56.4 versus 77.9/64.3 on InternVL (20%/10%; Figure 6). The same signal and token-level policy produce opposite orderings. TextVQA replicates the reversal: MEKV leads by 1–2 pp on Qwen and trails by 1.5–4 pp on InternVL, consistent with the calibration-regime prediction.

The calibration regimes match the prediction. On InternVL, the calibrated optimum sits near the untempered end of the path ($\lambda^*_{\text{grid}} = 0.42/0.84$; moment-matching estimate $\hat{\lambda} = 0.996$). SnapKV-style top-k operates at $\lambda(q) = 0.29/0.37$, below the calibrated optimum, so it does not over-concentrate along the signal direction; its positive off-path concentration then adds structure the temperature family cannot express.

On Qwen, the calibrated optimum sits deep in the tempered range ($\lambda^*_{\text{grid}} = 0.25/0.30$; $\hat{\lambda} = 0.61$). SnapKV-style top-k operates at $\lambda(q) = 0.66/0.85$, far past the optimum — the over-concentration failure of §3 — and its additional off-path structure does not recover the loss.

This ordering holds across all three benchmarks on InternVL (Appendix G). Figure 7 separates the two design axes. On InternVL, finer allocation granularity helps, from head-pooled to per-head selection. On Qwen, removing per-head granularity does not rescue token-level selection: both variants remain 14–19 pp below MEKV, which locates the loss in concentration rather than granularity.

Composing the two axes. The separation suggests placing fine-grained selection inside calibrated quotas rather than trading one against the other. We test this with a controlled pair that shares the SQA signal, per-head selection, and budget: per-head top-k over the whole image, versus per-head

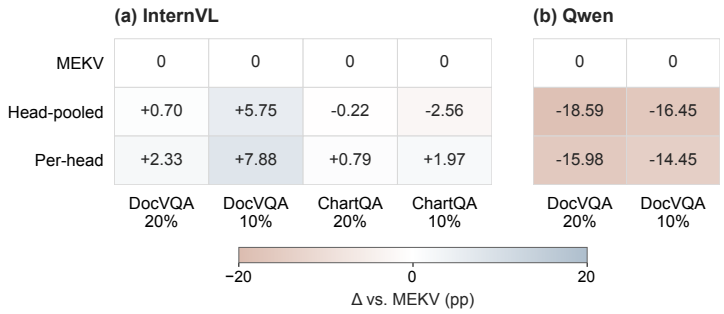


Figure 7: Allocation granularity versus concentration control. Differences relative to MEKV (pp) for same-signal token-level policies with per-head or head-pooled (per-layer) selection. On InternVL (full validation), finer granularity adds accuracy; on Qwen (DocVQA, frozen $\tau = 4$), both variants fall 14–19 pp below MEKV, so the loss stems from over-concentration rather than granularity.

Table 2: Adding calibrated quotas to per-head selection. Differences in pp, calibrated per-head selection (regional quotas with tempered within-region sampling) minus per-head top-k. DocVQA uses validation images disjoint from the development set; ChartQA uses full validation. All Qwen differences and two of four InternVL differences are statistically significant; all point estimates are positive (Appendix I).

Model	**DocVQA**		**ChartQA**	
	20%	10%	20%	10%
Qwen3.5-9B	+11.7	+13.9	+8.9	+13.8
InternVL3.5-8B	+1.7	+1.2	+0.1	+1.9

selection within MEKV's temperature-calibrated regional quotas, with tempered sampling inside each region (Appendix I). Adding concentration control at both levels repairs Qwen by 9–14 pp across both datasets and budgets, while on InternVL it preserves and slightly extends the gains of fine-grained selection (Table 2). Concentration calibration and allocation granularity are therefore complementary: calibrated quotas with fine-grained selection inside them repair Qwen while preserving InternVL's gains.

6 Discussion and Conclusion

Two separable design axes. The reversal identifies concentration calibration and allocation granularity as separable design choices. Token-ranking policies implicitly couple signal reliance with head, layer, and token selection. The framework separates these choices, and composing them, with fine-grained selection inside calibrated quotas, repairs Qwen while preserving InternVL's gains (§5.3).

Diagnose the regime before choosing the strategy. For a new model, first diagnose its calibration regime. Given SQA and development oracle statistics, the scalar moment-matching solve provides an inexpensive initial guide: $\hat{\lambda}$ near 1 suggests prioritizing fine-grained selection; $\hat{\lambda}$ well below 1 flags over-concentration risk and calls for temperature control first. This directs which design axis to explore; a short grid search guided by $\hat{\lambda}$ then fixes the deployment temperature.

Query-agnostic visual KV eviction is a concentration control problem: the same signal can help or hurt depending on its concentration, the model, and the budget. MEKV turns this insight into gains over coverage baselines, while the framework explains the architecture-dependent ordering of allocation strategies. Extending the diagnostic to SQA-only statistics is a natural next step.

Reproducibility Statement

All models (Qwen3.5-9B, InternVL3.5-8B, Gemma-4-12B-it) and benchmarks (DocVQA, ChartQA, TextVQA) are publicly available. Section 4.3 and Appendix B specify the full MEKV procedure, including proxy-question generation, attention aggregation, binning, the log-ratio clamp, and capacity-constrained integer allocation, together with the selected temperatures and the development split used to choose them. Evaluation uses the validation splits and metrics described in Section 5, byte-matched budgets, and image-level paired bootstrap with 10,000 resamples; stochastic policies average multiple random seeds (Appendix B). Complete proofs of all theoretical results are in Appendix A. We will release code for scoring, allocation, and evaluation upon publication.

AI Use Disclosure

We used generative AI tools (large language models) in supporting roles. They assisted with code implementation, including evaluation scripts and experiment orchestration; with parts of the data analysis, such as aggregating per-image results and computing bootstrap statistics; and with drafting and editing portions of the manuscript. We also used them to check derivations, which led to the corrected statement of Proposition 1 and to the Bregman expansion in Section 4.2; the authors verified both. The research question, the method, the experimental protocol, and the conclusions are the authors'. All AI-assisted code, analyses, and text were reviewed by the authors, who take full responsibility for the content of this paper.

References

Jinwoo Ahn, Ingyu Seong, Akhil Kedia, Junhan Kim, Hyemi Jang, Kangwook Lee, and Yongkweon Jeon. Lookaheadkv: Fast and accurate kv cache eviction by glimpsing into the future without generation. *arXiv preprint arXiv:2603.10899*, 2026.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Yongqi An, Chang Lu, Kuan Zhu, Tao Yu, Chaoyang Zhao, Hong Wu, Ming Tang, and Jinqiao Wang. Rest-kv: Robust kv cache eviction with layer-wise output reconstruction and spatial-temporal smoothing. *arXiv preprint arXiv:2605.08840*, 2026.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling, 2024. *URL https://arxiv. org/abs/2406.02069*, 2024.

Hong Chen, Kang Chen, Yuxuan Fan, Bo Wang, Yubo Gao, Yuanlin Chu, and Xuming Hu. Seen, said, or forgotten? a causal audit of visual kv memory across dialog turns. *arXiv preprint arXiv:2607.25467*, 2026.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024b.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. Not all heads matter: A head-level kv cache compression method with integrated retrieval and reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 99269–99290, 2025.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. In *International Conference on Learning Representations*, volume 2024, pp. 22975–22988, 2024.

Haoyu Guo, Junlin Lv, S Kevin Zhou, Xike Xie, et al. Defensivekv: Taming the fragility of kv cache eviction in llm inference. In *International Conference on Learning Representations*, volume 2026, pp. 43407–43432, 2026.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W Lee, Sangdoo Yun, and Hyun Oh Song. Kvzip: Query-agnostic kv cache compression with context reconstruction. *Advances in Neural Information Processing Systems*, 38:167563–167591, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, et al. Scbench: A kv cache-centric analysis of long-context methods. In *International Conference on Learning Representations*, volume 2025, pp. 66063–66093, 2025.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Xuyang Liu, Xiyan Gui, Yuchao Zhang, and Linfeng Zhang. Mixing importance with diversity: Joint optimization for kv cache compression in large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 125061–125081, 2026.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. *arXiv preprint arXiv:2406.10774*, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4065–4078, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. In *International Conference on Learning Representations*, volume 2025, pp. 37228–37253, 2025.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Conical visual concentration for efficient large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603. IEEE, 2025.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in neural information processing systems*, 36:34661–34710, 2023.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.

A Theoretical Framework

A.1 The Maximum-Entropy Characterization

Lemma A.1 (Gibbs variational characterization). Among all distributions p over n bins with $\text{supp}(p) \subseteq \text{supp}(u)$ and a given expected log-ratio moment $\mathbb{E}_p[\ell] = m$, the distribution minimizing $\text{KL}(p\|u)$ is $p_\lambda \propto u \cdot \exp(\lambda\ell)$ where λ is the Lagrange multiplier enforcing the moment constraint.

Proof. Standard exponential-family variational argument (Amari & Nagaoka, 2000). The KL minimization with a linear moment constraint yields the Gibbs form. The partition function $Z(\lambda) = \sum_i u_i \exp(\lambda\ell_i)$ is strictly log-convex, ensuring uniqueness.

Lemma A.2 (Strict monotonicity of concentration). The continuous concentration $C(\lambda) = \text{KL}(p_\lambda\|u) = \lambda\psi'(\lambda) - \psi(\lambda)$, where $\psi(\lambda) = \log Z(\lambda)$, is strictly increasing in $\lambda \geq 0$ whenever ℓ is non-constant on $\text{supp}(u)$.

Proof. $C'(\lambda) = \lambda\psi''(\lambda)$. Since $\psi''(\lambda) = \text{Var}_{p_\lambda}[\ell] > 0$ for non-constant ℓ, we have $C'(\lambda) > 0$ for $\lambda > 0$, and $C(0) = 0$.

A.2 Proof of Theorem 1

The surrogate accuracy is $A(\lambda) = -\text{KL}(a\|p_\lambda)$, where a is the oracle distribution. Expanding:

$$A(\lambda) = -\sum_i a_i \log \frac{a_i}{p_\lambda(i)} = \text{const} + \lambda\mathbb{E}_a[\ell] - \psi(\lambda)$$

where the constant absorbs $-\sum_i a_i \log a_i + \sum_i a_i \log u_i$.

Taking derivatives: $A'(\lambda) = \mathbb{E}_a[\ell] - \psi'(\lambda)$ and $A''(\lambda) = -\psi''(\lambda) = -\text{Var}_{p_\lambda}[\ell] < 0$.

For non-constant ℓ on $\text{supp}(u)$, strict concavity holds independently of the sandwich condition. Hence A has at most one maximizer on $[0, \infty)$, and any finite interior maximizer satisfies $\psi'(\lambda^*) = \mathbb{E}_a[\ell]$, equivalently $\mathbb{E}_{p_{\lambda^*}}[\ell] = \mathbb{E}_a[\ell]$.

Under the sandwich condition $\mathbb{E}_u[\ell] < \mathbb{E}_a[\ell] < \mathbb{E}_s[\ell]$, we have $A'(0) > 0$ and $A'(1) < 0$. Continuity and strict monotonicity of ψ' give a unique moment-matching maximizer $\lambda^* \in (0, 1)$. The same strict concavity applies beyond $\lambda = 1$, so any finite root there is also the unique maximizer. If $\mathbb{E}_a[\ell] \leq \mathbb{E}_u[\ell]$, the constrained maximum is instead at $\lambda = 0$; moment matching need not hold at that boundary. □

For any finite moment-matched optimum λ^*, the same expansion gives the pathwise calibration penalty

$$\begin{aligned}\text{KL}(a\|p_\lambda) - \text{KL}(a\|p_{\lambda^*}) &= \psi(\lambda) - \psi(\lambda^*) - \psi'(\lambda^*)(\lambda - \lambda^*) \\ &=: D_\psi(\lambda, \lambda^*) \geq 0.\end{aligned}$$

This Bregman divergence is determined by the convex log-partition function ψ, following the exponential-family/Bregman correspondence (Banerjee et al., 2005). Its local curvature is $\psi''(\lambda) = \text{Var}_{p_\lambda}[\ell]$, so distance in λ alone does not determine the size of the penalty across architectures. The identity compares members of the temperature family; Proposition 1 separately describes the concentration of general allocations.

A.3 Multi-Layer Extension

For Qwen, eviction is performed separately on each of L KV layers using per-layer log-ratios $\ell^{(l)}$ at a shared temperature λ. With nonnegative layer weights, the pooled surrogate $A_{\text{pool}}(\lambda) = -\sum_l w_l \text{KL}(a_l\|p_\lambda^{(l)})$ is strictly concave whenever a positively weighted layer has non-constant $\ell^{(l)}$. Its finite interior optimum solves

$$\sum_l w_l \, {\psi^{(l)}}'(\lambda) = \sum_l w_l \, \mathbb{E}_{a_l}[\ell^{(l)}].$$

Strict concavity, uniqueness of the moment-matching root, and monotonicity of pooled concentration therefore carry over to the pooled objective.

A.4 Pythagorean Decomposition (Proof of Proposition 1)

Let $I = \mathrm{supp}(u)$, assume ℓ_i is finite on I, and let q be supported on I. Suppose a finite $\lambda(q) \geq 0$ matches q's signal moment. Write $\lambda = \lambda(q)$ and $p = p_\lambda$. The exponential-family form gives

$$p_i = u_i \exp(\lambda \ell_i - \psi(\lambda)), \qquad \psi(\lambda) = \log \sum_{i \in I} u_i \exp(\lambda \ell_i).$$

Since $p_i > 0$ on I, direct expansion yields

$$\begin{aligned} \mathrm{KL}(q\|u) &= \sum_{i \in I} q_i \log \frac{q_i}{p_i} + \sum_{i \in I} q_i \log \frac{p_i}{u_i} \\ &= \mathrm{KL}(q\|p) + \lambda \mathbb{E}_q[\ell] - \psi(\lambda) \\ &= \mathrm{KL}(q\|p) + \lambda \mathbb{E}_p[\ell] - \psi(\lambda) \\ &= \mathrm{KL}(q\|p) + \mathrm{KL}(p\|u). \end{aligned}$$

The third line uses moment matching, and the last uses the same exponential-family expression for $\log(p_i/u_i)$. Terms with $q_i = 0$ use the convention $0 \log 0 = 0$. Substituting $p = p_{\lambda(q)}$ gives

$$\mathrm{KL}(q\|u) = \mathrm{KL}(p_{\lambda(q)}\|u) + \mathrm{KL}(q\|p_{\lambda(q)}).$$

Both components are nonnegative. The first is the minimum concentration among distributions with the prescribed signal moment (Lemma A.1); the second is the additional concentration carried by q beyond this moment-matched family member. It vanishes exactly when $q = p_{\lambda(q)}$. The identity quantifies allocation geometry, while accuracy is evaluated separately in §5. □

A.5 Corollary 2: Budget Shift

Assumption B. As budget r decreases, the oracle log-signal moment $\mathbb{E}_a[\ell]$ strictly increases. Intuitively, under tighter budgets, the remaining evidence concentrates on the most informative regions, increasing the oracle moment.

Under this assumption, the moment-matching condition $\psi'(\lambda^*(r)) = \mathbb{E}_{a(r)}[\ell]$ implies that $\lambda^*(r)$ increases as r decreases (since ψ' is strictly increasing). Since $C(\lambda)$ is also strictly increasing, we have $C^*(r_1) > C^*(r_2)$ for $r_1 < r_2$.

A.6 Continuous and Realized Concentration

The theory uses the continuous coordinate $C(\lambda) = \mathrm{KL}(p_\lambda\|u)$. The empirical coordinate uses retained-token distributions $q_{\lambda,r}$ after integer allocation, in the reported spatial binning coordinate. The logarithm of the image-mean realized concentration is:

$$\ln C_{\mathrm{real}}(\lambda, r) = \ln \left(\frac{1}{N} \sum_{j=1}^{N} \mathrm{KL}\left(q_{\lambda,r}^{(j)} \| u^{(j)} \right) \right).$$

For models with layer-specific allocations, each image's KL is first aggregated using the evaluation's layer convention. These measurements combine changes in the maximizing temperature with budget-dependent rounding; a shift in realized concentration alone is not a direct measurement of the oracle-moment change assumed in Corollary 2.

B Experimental Details

B.1 SQA Signal and Allocation Details

Each evaluated VLM generates up to six proxy questions about the image with greedy decoding, prompted to cover different regions and information types (e.g., titles, dates, names, numbers, tables, and labels); a fixed list of generic questions is used as a fallback when too few questions are produced. SQA scores aggregate the attention from all proxy-question tokens to each visual token over attention heads, scoring layers, and questions, and are summed within bins and normalized. Log-ratios use $\ell_b = \log \max(s_b, \epsilon) - \log \max(u_b, \epsilon)$ with $\epsilon = 10^{-12}$ over non-empty bins. Integer counts are capped at each bin's token count, with any remainder redistributed over bins with remaining capacity; tokens within a bin are sampled uniformly without replacement.

B.2 Temperature Selection

Temperature τ is selected once per model on a coarse development grid and frozen across budgets and benchmarks: Qwen $\tau = 4$, InternVL $\tau = 2$. For analysis, the finite-temperature grid is refined geometrically with ratio $2^{1/4}$ (uniform refinement spacing of $\ln(2)/4 \approx 0.1733$ in log-τ), together with the coarse-grid points and the coverage endpoint $\tau = \infty$.

B.3 Baseline Implementation Details

Budget matching. All comparisons match retained visual KV bytes to within $\pm 5\%$ of the target budget per image.

Evaluation splits. Temperature is selected per model on a development set. Full validation uses the complete DocVQA, ChartQA, and TextVQA validation sets. All baselines share the same evaluation harness, manifest, metric code, and bootstrap protocol. Paired comparisons resample images jointly across arms with replacement (10,000 resamples), following the bootstrap principle (Efron & Tibshirani, 1994).

Stochastic arms are evaluated with multiple random seeds; we report the mean.

SnapKV-style. Uses the SQA bank as observation window. Per KV head, per layer, retains the top-k tokens by attention score. Byte-matched to the method's retention budget per image.

PyramidKV-style. Layered pyramid allocation with higher retention in lower layers, using SQA attention. Byte-matched.

FastV. No-question mode (image tokens only, no instruction). Prunes during decoding based on accumulated attention. Byte-matched by adjusting the pruning threshold.

B.4 Gemma-Specific Details

Operative grid. The 16×16 grid used for Qwen/InternVL is infeasible for Gemma (250–276 visual tokens, many bins empty). A 4×4 grid (16 bins) is used, determined by a feasibility rule requiring all development images at both budgets to have non-degenerate allocations.

Primary arm (full8). Eviction is restricted to the 8 full-attention layers, consistent with the rule applied to Qwen (8 KV layers) and InternVL (all 36 layers are full-attention). The 40 sliding-window layers retain their full visual KV under this arm, meaning only 4.76% of total visual KV capacity is ever at stake.

Secondary arm (all48). Eviction across all 48 layers; coarse grid only.

C Moment Matching: Full Data

C.1 Pooled Estimator

The oracle distribution a is computed from each image's real evaluation questions, aggregated in the same way as the SQA scores.

Table 3: Pooled estimator

Model	$\hat{\lambda}_{\text{pool}}$
Qwen3.5	0.6128
InternVL3.5	0.9958

Aggregated-score sensitivity for Qwen: 0.6226 (same verdicts).

C.2 Per-Cell Comparison

Figure 8 extends the comparison to all three budgets.

Table 4: Per-cell comparison. The InternVL@5% grid optimum is $\lambda^* \approx 1$, hence $\ln \lambda^* \approx 0$.

Cell	$\ln \hat{\lambda}$	$\ln \lambda^*_{\text{grid}}$	**Gap**
Qwen@20%	-0.490	-1.386	0.897
Qwen@10%	-0.490	-1.213	0.723
InternVL@20%	-0.004	-0.866	0.862
InternVL@10%	-0.004	-0.173	0.169
InternVL@5%	-0.004	≈ 0.000	0.004

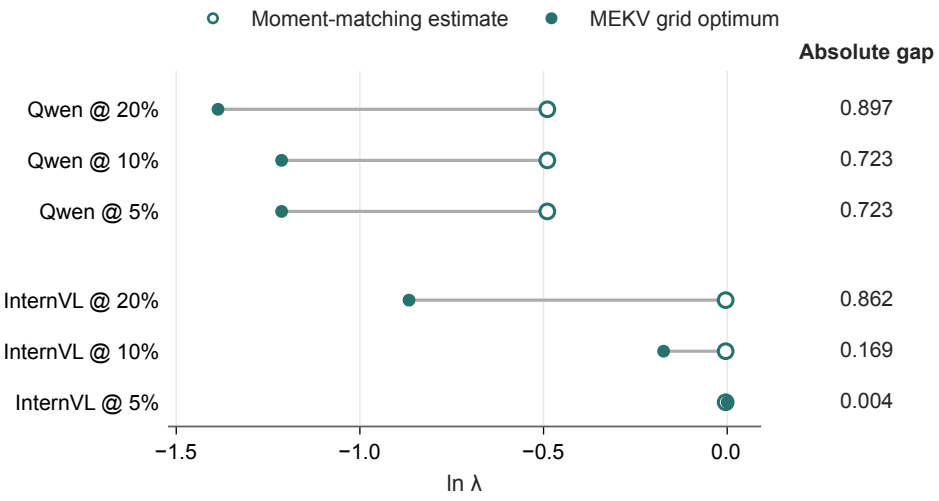


Figure 8: Moment-matching diagnostics on DocVQA development data. Open markers show $\hat{\lambda}$ from SQA and oracle moments; filled markers show the best tested MEKV setting. Connecting segments and the right column report absolute gaps in $\ln \lambda$; the InternVL gap narrows to 0.004 at 5% budget.

C.3 Sandwich Diagnostics

Qwen: sandwich condition holds for 100% of images (all images have $\mathbb{E}_u[\ell] < \mathbb{E}_a[\ell] < \mathbb{E}_s[\ell]$). Let $\hat{m} = \widehat{\mathbb{E}}_a[\ell]$ denote the sample oracle moment. On InternVL, 46.7% of images are over-sharp ($\hat{m} = \widehat{\mathbb{E}}_a[\ell] > \mathbb{E}_s[\ell]$, Corollary 1 case-4 population). InternVL's interiority is a knife-edge pass ($\hat{\lambda} = 0.9958$, margin 0.004 from the boundary).

C.4 Budget Refinement at 5%

For InternVL at 5%, the tested-grid optimum sits at $\tau \approx 1.0$ ($\lambda^* \approx 1.0$), the calibrated-SQA point. The moment-matching estimate $\hat{\lambda} = 0.9958$ predicts this to within 0.004 in log-space — the first instance of the instrument achieving near-exact quantitative agreement, supporting the interpretation of $\hat{\lambda}$ as a tight-budget limit.

For Qwen at 5%, the tested-grid argmax is the same as at 10% ($\lambda^*_{\text{grid}} \approx 0.30$), so the gap does not shrink further at grid resolution.

D Budget Shift: Three-Budget Additive Analysis

Figure 5 shows the measured budget trajectories, including Gemma's primary arm.

Table 5: Budget shift and cross-model prediction.

Model	$\ln C^*_{\text{real}}(20\%)$	$\ln C^*_{\text{real}}(10\%)$	$\ln C^*_{\text{real}}(5\%)$
Qwen3.5	-1.758	-0.830	+0.149
InternVL3.5	-1.198	+0.386	+0.836

D.1 Three-Budget Fit

Fitted budget effects (ln C*): -1.075 (20%) / +0.180 (10%) / +0.895 (5%). The three-point monotone pattern is consistent with the conditional budget trend in Corollary 2; the realized coordinate also includes discretization effects (Appendix A.6).

D.2 Interaction Analysis

Table 6: Interaction analysis

Free dof	**Point estimate**	**95% CI**	**vs** ± 0.1733
$\gamma_{Q,20}$	+0.131	$[-0.310, +0.254]$	not contained
$\gamma_{Q,10}$	-0.198	$[-0.438, +0.207]$	not contained

Additivity survives unrejected (both CIs contain zero, $\approx 4\times$ tighter than the two-budget fit). Relative to a fixed reference tolerance of ± 0.1733 in log-concentration, neither interval is contained within that range. This tolerance is a numerical reference, not the spacing of the realized log-KL grid. The log-additive structure is a useful descriptive model, with precision still limited at the current sample size. The all48 Gemma ensemble has interaction SS of only 3.6%, suggesting near-additive structure across the widest model span.

D.3 Secondary-Peak Check

For each tested curve, secondary local maxima are compared with the intervening valley toward the global maximum. A secondary peak is significant when the 95% image-bootstrap interval for its prominence lies strictly above zero. No tested curve has a significant secondary peak under this criterion.

Table 7: Largest secondary-peak prominence per curve (DocVQA development). All intervals include zero.

Curve	**Prominence (pp)**	**95% CI (pp)**
Qwen 20%	0.1	$[-0.7, 1.0]$
Qwen 10%	0.5	$[-1.4, 2.5]$
InternVL 20%	0.7	$[-0.7, 2.1]$
InternVL 10%	0.5	$[-1.5, 2.6]$
Gemma full8 20%	0.9	$[-0.0, 1.8]$
Gemma full8 10%	1.0	$[-0.3, 2.3]$
Gemma all48 20%	no secondary peak	—
Gemma all48 10%	no secondary peak	—

Qwen query-mean sensitivity curves give the same verdict.

E SnapKV Reversal: Detailed Composition Analysis

E.1 Composition Columns

For each baseline arm, the retained-index counts define an allocation q on the spatial bins. We compute its moment-matched member $p_{\lambda(q)}$, on-path concentration $C_{\text{sig}} = \text{KL}(p_{\lambda(q)} \| u)$, and off-

path structural concentration $C_{\text{exc}} = \text{KL}(q\|p_{\lambda(q)})$. InternVL uses the layer- and head-summed allocation; Qwen is decomposed separately on each KV layer before averaging. The identity in Proposition 1 is checked within each decomposition, with maximum absolute residual below 5.6×10^{-11} in the development panels.

If an allocation's signal moment is below the coverage moment, the implementation clips $\lambda(q)$ to zero. Then $p_0 = u$, $C_{\text{sig}} = 0$, and $C_{\text{exc}} = \text{KL}(q\|u)$; the identity is immediate even though this boundary case does not match moments.

We also record the dimensionless structural diagnostic

$$\chi^2(q\|p_{\lambda(q)}) = \sum_{i:p_{\lambda(q)}(i)>0} \frac{(q_i - p_{\lambda(q)}(i))^2}{p_{\lambda(q)}(i)}.$$

Table 8 reports ANLS gains in percentage points relative to the mean MEKV temperature curve at each arm's total concentration, separately from these geometric quantities.

E.2 InternVL Panel

Table 8: InternVL panel (DocVQA development set, ANLS $\times$ 100).

Arm	**Budget**	**Score**	λ**(q)**	C_{exc}	**Gain vs. curve (pp)**
Method (τ=2)	20%	74.93	—	0 (on path)	—
SnapKV	20%	76.95	0.287	0.140	+1.9
SnapKV	10%	60.23	0.368	0.210	+7.0
PyramidKV	10%	56.74	0.361	0.215	+3.5

The corresponding $\chi^2(q\|p_{\lambda(q)})$ values are 0.283 for SnapKV at 20%, 0.448 for SnapKV at 10%, and 0.456 for PyramidKV at 10%. SnapKV's moment-matched temperature is below the tested-grid optimum at both budgets, while its ANLS exceeds the peak of the mean MEKV temperature curve. Positive C_{exc} identifies structure beyond the family; the ablation in §E.4 tests how head and token granularity contribute to the gain.

E.3 Qwen Panel

All Qwen MEKV scores in this comparison use the frozen $\tau = 4$. Table 9 reports the DocVQA development scores and SnapKV-style calibration coordinates; Table 13 reports absolute scores multiplied by 100.

Table 9: Qwen panel

Arm	**Budget**	**Score**	**MEKV** ($\tau = 4$)	$\lambda(q)$
SnapKV	20%	0.7163	0.8762	0.663
SnapKV	10%	0.5735	0.7180	0.850

SnapKV-style scores 14–16 pp below MEKV at the frozen temperature. On Qwen, SnapKV's $\lambda(q) = 0.66/0.85$ is far above the calibrated optimum ($\lambda^*_{\text{grid}} = 0.25/0.30$), placing its signal-aligned component in the over-concentrated regime despite positive off-path structure.

E.4 Decoupling: Per-Head vs Token-Level

A head-pooled variant (same SQA, per-layer token top-k, no per-head granularity) separates the per-head contribution:

- @10%: head-pooled $\approx$ SnapKV (0.6024 vs 0.6023) — per-head granularity contributes nothing; the gain is from token-level per-layer freedom.
- @20%: head-pooled falls back to the curve (0.7485) — the gain is entirely per-head.

At full validation, the attribution resolves along the dataset axis: head-pooled clears the method on DocVQA@10% (+5.75 pp) but loses on ChartQA@10% (-2.56 pp). Per-head granularity is decisive specifically on ChartQA.

Figure 7 in §5.3 summarizes the full-validation InternVL and frozen-$\tau = 4$ development Qwen granularity comparisons.

F Efficiency Details

F.1 Serve-Time Comparison (24 timing configurations)

Method and stratified eviction share identical eviction mechanics at serve time. Across 24 configurations (2 models $\times$ 2 budgets $\times$ 2 batch sizes $\times$ 3 repetitions), the relative median difference is -0.31% to +0.34%. Only InternVL@10% batch-1 shows a small significant overhead (+0.34%). All measurements taken on a single GPU in controlled conditions (3 repetitions, median reported, warm-up excluded).

F.2 Peak Visual KV Cache Memory

Table 10: Peak visual KV cache memory relative to the full cache.

Model	Budget	Peak visual KV cache reduction
Qwen3.5	20%	76.7%
Qwen3.5	10%	86.3%
InternVL3.5	20%	79.0%
InternVL3.5	10%	88.9%

These reductions apply to peak visual KV cache memory, not total GPU memory including model weights. Batch-4 inference feasible on 48GB A6000 for all configurations.

F.3 SQA Preprocessing Cost

Table 11: SQA preprocessing cost

Model	Time (seconds)
Qwen3.5	25.0
InternVL3.5	17.8

Times include proxy-question generation.

This is a one-time per-image cost for constructing the SQA signal. It does not scale with the number of downstream queries but represents a significant upfront investment. The method's value proposition is quality improvement under a fixed compression budget, not end-to-end latency reduction.

G Full Baseline Comparison Table

The absolute coverage and strategy-reversal scores are listed in Tables 12 and 13; paired differences and intervals follow in Table 14.

Table 12: Full validation, InternVL3.5-8B, across three benchmarks. All six differences are statistically significant (paired bootstrap; intervals in Appendix G).

Dataset	Budget	Method	Stratified	Δ (pp)
DocVQA	20%	75.56	70.40	+5.16
DocVQA	10%	56.40	48.52	+7.88
ChartQA	20%	79.13	73.41	+5.72
ChartQA	10%	69.17	57.19	+11.98
TextVQA	20%	74.15	66.69	+7.47
TextVQA	10%	68.94	53.10	+15.84

Table 13: Architecture-dependent strategy reversal on DocVQA and TextVQA under the same SQA signal: the method ranks higher on Qwen, SnapKV-style on InternVL. Scores are ANLS (DocVQA) or accuracy (TextVQA), $\times$ 100. Qwen DocVQA uses development evaluation at the frozen $\tau = 4$; all other scores use fixed-temperature full validation. InternVL and both TextVQA model comparisons are statistically significant (Appendix G).

Setting	Qwen: Method	Qwen: SnapKV	InternVL: Method	InternVL: SnapKV
DocVQA @20%	**87.62**	71.63	75.56	**77.89**
DocVQA @10%	**71.80**	57.35	56.40	**64.28**
TextVQA @20%	**77.08**	75.37	74.15	**75.62**
TextVQA @10%	**68.00**	66.82	68.94	**72.84**

Full-validation comparisons (image-level paired bootstrap, 10k resamples; differences and 95% intervals in pp). The upper panel reports InternVL3.5 on DocVQA and ChartQA; the lower panel reports TextVQA on both models.

Table 14: Full baseline comparison table

Comparison	DocVQA@20%	DocVQA@10%	ChartQA@20%	ChartQA@10%
Method − Stratified	+5.16 [4.31, 6.05]	+7.88 [6.80, 8.96]	+5.72 [4.48, 6.97]	+11.98 [10.29, 13.70]
Method − Random	+4.30 [3.42, 5.17]	+7.69 [6.62, 8.88]	+5.65 [4.44, 6.84]	+11.49 [9.81, 13.16]
SnapKV − Method	+2.33 [1.25, 3.40]	+7.88 [6.58, 9.15]	+0.79 [−0.42, 1.97]	+1.97 [0.18, 3.82]
Oracle − Method	+16.1	+34.6	+5.6	+15.1

TextVQA full validation Comparison	Qwen@20%	Qwen@10%	InternVL@20%	InternVL@10%
Method − Stratified	+10.83 [9.85, 11.80]	+17.02 [15.92, 18.14]	+7.47 [6.66, 8.26]	+15.84 [14.85, 16.84]
SnapKV − Method	−1.71 [−2.53, −0.88]	−1.18 [−2.23, −0.11]	+1.47 [0.84, 2.11]	+3.90 [3.01, 4.77]

Oracle reference scores: DocVQA 0.917/0.910; ChartQA 0.848/0.843 at 20%/10%.

H LINEAR SIGNAL-COVERAGE MIXING

We compare the temperature family with the linear alternative $p_\alpha = (1-\alpha)u+\alpha s$, where $\alpha \in [0, 1]$. Table 15 reports the best linear-mixture weight and the best temperature for each model–budget condition. Scores are multiplied by 100; Δ is the temperature method minus the best linear mixture, in percentage points.

Table 15: Best linear-mixture and temperature allocations from the sweeps. Qwen differences are statistically significant; InternVL differences are not.

Condition	**Best α**	**Linear**	**Best τ**	**Temp.**	**Δ (pp)**
Qwen@20%	0.2	86.49	4	87.62	+1.13
Qwen@10%	0.2	71.46	3.364	73.49	+2.03
InternVL@20%	0.7	74.72	2.378	75.27	+0.55
InternVL@10%	0.8	55.64	1.189	55.90	+0.26

95% CIs (pp), in row order: [+0.19, +2.06]; [+0.12, +4.03]; [−0.83, +1.98]; [−1.58, +1.04].

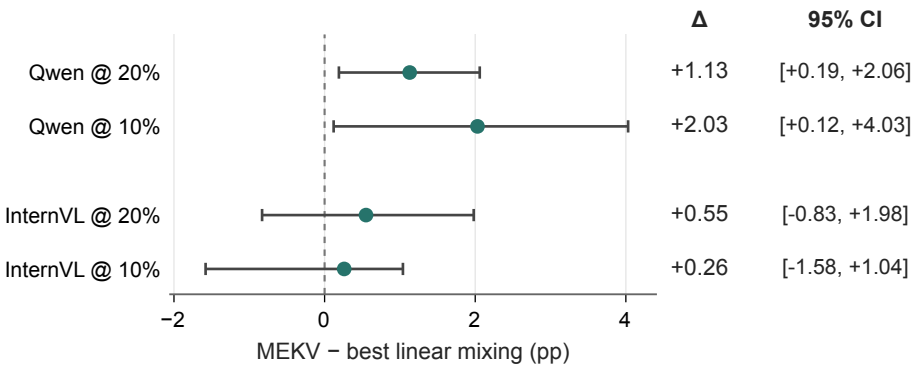


Figure 9: MEKV versus optimized linear signal-coverage mixing, $p_\alpha = (1 - \alpha)u + \alpha s$. Each condition compares the best τ and α from the sweeps; bars are paired 95% intervals in pp. Qwen differences exclude zero; InternVL differences do not (Table 15).

Figure 9 visualizes these paired differences. Both Qwen confidence intervals exclude zero, supporting a significant advantage for the temperature parameterization. Both InternVL intervals include zero, so the difference between the optimized parameterizations is not statistically resolved on this architecture. The temperature family retains its variational characterization: it minimizes KL to coverage at each attainable signal moment (Lemma A.1), whereas linear mixing generally does not attain this minimum.

I CALIBRATED PER-HEAD SELECTION

Policies. Both policies use the same SQA scores (per layer and KV head, kernel-7 pooled) and retain the same number of tokens per head. *Per-head top-k* takes each head's global top-k tokens; ties at the cutoff are broken uniformly at random. *Calibrated per-head selection* first sets regional quotas with MEKV's temperature allocation at τ_R. Within each region, each head then samples its share of tokens without replacement, with probabilities proportional to $\exp(\beta \cdot \text{score})$ (Gumbel-top-k).

Parameters. τ_R and β are selected on the DocVQA development set and fixed across budgets and datasets: Qwen uses $\tau_R = 4$, $\beta = 1/4$; InternVL uses $\tau_R = 2$, $\beta = 1$.

Evaluation. DocVQA uses 786 validation images (2,871 questions) disjoint from the development set. ChartQA uses full validation: 1,056 images (1,920 questions). We use three seeds and image-mean scores, with paired image-level bootstrap over 10,000 resamples. These evaluation images differ from the development panel in §5.3; the per-head top-k scores here are therefore reported separately from Figure 6.

Table 16: Calibrated per-head selection versus per-head top-k. Scores are image-mean ANLS (DocVQA) or relaxed accuracy (ChartQA), $\times$ 100; differences and paired 95% confidence intervals are in pp.

Dataset	Model	Budget	Calibrated	Per-head top-k	Δ [95% CI]
DocVQA	Qwen	20%	89.38	77.66	+11.72 [10.33, 13.16]
DocVQA	Qwen	10%	78.58	64.71	+13.87 [12.37, 15.38]
DocVQA	InternVL	20%	79.93	78.24	+1.68 [0.61, 2.77]
DocVQA	InternVL	10%	66.46	65.26	+1.20 [−0.18, 2.63]
ChartQA	Qwen	20%	70.28	61.37	+8.91 [6.84, 10.97]
ChartQA	Qwen	10%	57.85	44.02	+13.82 [11.44, 16.15]
ChartQA	InternVL	20%	79.84	79.75	+0.10 [−0.93, 1.10]
ChartQA	InternVL	10%	73.52	71.64	+1.88 [0.40, 3.34]