LMTC: QUANTIZING KV CACHES FOR HOW ATTENTION READS THEM IN OMNI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal Large Language Models (OmniLLMs) unify the understanding of video, audio, and text, yet their continuous streaming inputs generate an overwhelming volume of tokens, making the Key-Value (KV) cache a major memory bottleneck. Existing KV cache quantization methods are suboptimal for OmniLLMs because they either minimize simple reconstruction error or rely on query activations that are unavailable when streaming tokens are cached, ignoring both the attention reading mechanism and cross-modal heterogeneity. To address these challenges, we propose Logit-Metric Transform Coding (LMTC), a training-free and query-free KV cache quantization framework that minimizes the expected attention error using only the query projection weights. Specifically, LMTC quantizes Keys using a Modality-Conditioned Transform Coder guided by a Query-Free Logit Metric derived directly from query projection weights, thereby encoding each modality in its own basis and allocating precision via reverse water-filling to the directions attention reads. Simultaneously, LMTC handles Values through Phase-Offset Rounding, a metadata-free mechanism that offsets the rounding grids of spatio-temporally adjacent visual tokens so that rounding errors cancel out during attention aggregation rather than accumulate. Extensive evaluations across three OmniLLMs on eight benchmarks show that LMTC is virtually lossless at 3 effective bits per scalar, retaining over 99% of BF16 average accuracy. Compared to BF16 FlashDecoding-v2, LMTC achieves up to a $1.94\times$ decoding speedup, a $5.28\times$ KV cache memory reduction, and a $2.08\times$ throughput boost.

1 INTRODUCTION

While recent advances in omni-modal large language models (OmniLLMs) have successfully unified text, audio, and video understanding within a single architecture (Xu et al., 2025a;b), this paradigm shift introduces a severe GPU memory bottleneck. Unlike traditional vision-language models that process static image-text pairs, OmniLLMs continuously ingest multi-modal streaming inputs to enable real-time, interactive assistants (Lin et al., 2024a). High-resolution visual frames and uninterrupted audio signals continuously generate massive sequence lengths, forcing their Key-Value (KV) cache to be retained in memory throughout generation. Consequently, the KV cache footprint rapidly scales to rival or even surpass model weights, severely constraining context capacity and batch throughput (Di et al., 2025; Yang et al., 2025a; Jiang et al., 2026; Sun et al., 2026).

Low-bit KV cache quantization offers a direct remedy to this memory overhead (Duanmu et al., 2024; Sharma et al., 2025). The cache is written once and read later by attention, where a Key error affects generation only through its inner products with future queries. Existing quantizers take two stances toward this reader. The first ignores it and minimizes reconstruction error (Liu et al., 2024; Hooper et al., 2024; Su et al., 2025a;b; Staniszewski & Łańcucki, 2025), treating every error direction as equally harmful. The second aligns with the reader by steering quantization with attention scores (He et al., 2024) or query subspaces (Wang et al., 2025), but applies the right criterion at the wrong time, since the queries arrive after the cache is written and at unpredictable positions. OmniLLMs expose the limits of both paradigms. Streaming input caches audio-visual tokens before any question arrives, leaving future readers unobservable. Moreover, the cache mixes text, audio, and video tokens with heterogeneous distributions, which text-tailored quantizers code alike.

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.

To overcome the limitations of reconstruction objectives and static text-tailored quantizers, we first uncover two overlooked intrinsic properties of OmniLLM KV caches: (i) **Weight-governed read directions.** Challenging raw reconstruction goals, we observe that Key quantization damage concentrates along a few directions: a tiny fraction of error along critical read directions causes nearly half the attention distortion, whereas high-energy Key principal components cause minimal damage. Crucially, because these directions depend on query projection weights, they can be characterized *a priori* without observing runtime queries. (ii) **Modality-driven distribution heterogeneity.** Contradicting uniform token assumptions, we find that text, audio, and video Keys form distinct statistical clusters with disparate means and principal axes. Consequently, a modality-agnostic codebook becomes heavily dominated by video tokens, which form the vast majority of streaming inputs, degrading text and audio precision while shifting attention mass across modalities.

Motivated by these insights, we introduce **LMTC** (Logit-Metric Transform Coding), a training-free, query-free KV cache quantization framework for OmniLLMs. At its core, LMTC formulates a Query-Free Logit Metric that computes the expected logit distortion directly from query projection weights by taking an expectation over arbitrary reader positions. Built upon this metric, a Modality-Conditioned Transform Coder whitens the Keys of each modality in its own basis, allocating bits via reverse water-filling along the directions attention reads. For Values, we introduce Phase-Offset Rounding, which systematically shifts the quantization grid of visual tokens based on their space-time coordinates, enabling quantization errors across spatio-temporally redundant tokens to cancel out during attention aggregation rather than accumulate. Because every token is quantized in a single pass upon entry, LMTC applies directly to streaming input.

Our main contributions are summarized as follows:

- We reveal two critical insights into the OmniLLM KV cache: that attention read directions are determined by static projection weights rather than unobserved runtime queries, and that distinct modalities require dedicated bases against majority-modality dominance.
- We propose LMTC, a training-free and query-free KV cache quantizer that optimizes Key quantization via a weight-derived logit metric in modality-specific subspaces and cancels redundant visual Value errors through position-offset rounding grids.
- Evaluations on three OmniLLMs and eight benchmarks show that LMTC is virtually lossless at 3 effective bits per scalar (over 99% of BF16 average accuracy), with up to $1.94\times$ decoding speedup, $5.28\times$ KV cache reduction, and $2.08\times$ throughput increase over BF16 FlashDecoding-v2.

2 Motivation

Existing KV cache quantizers are designed primarily for text LLMs. They minimize the reconstruction error of Keys and Values and quantize all tokens identically (Liu et al., 2024; Hooper et al., 2024; Su et al., 2025b; Staniszewski & Łańcucki, 2025). In OmniLLMs, however, the cache is dominated by redundant audio and video tokens, and reconstruction error no longer reflects how attention reads the cache. We analyze the attention of Qwen2.5-Omni (Xu et al., 2025a) under INT2 quantization and identify two mismatches, concerning how Keys are read and how the modalities share the cache, respectively.

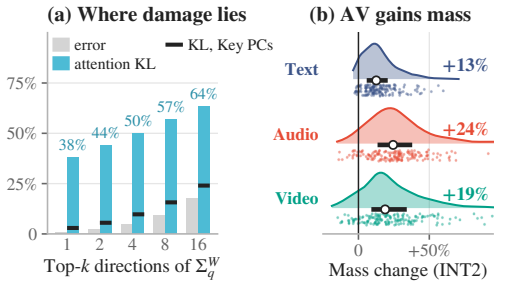


Figure 1: **Few read directions; modalities read differently.** (a) Share of the INT2 Key error inside the top-k directions of the weight-derived logit metric (Sec. 3.2; gray) and share of the attention KL it causes (blue); ticks: KL share for the top-k principal directions of the Keys. (b) Change in attention mass of each modality under INT2.

Observation (i): Attention reads Keys along a few weight-determined directions and responds differently to each modality. A Key error e_i shifts the logit of a reader q by $q^\top e_i/\sqrt{d}$ and therefore affects attention only through its components along the directions that queries read. These directions are few and largely determined by the query projection weights.

As shown in Figure 1(a), the small share of INT2 error that falls into the top weight-derived directions accounts for nearly half of the attention KL, whereas the principal directions of the Keys, which reconstruction-based quantizers favor, contain more error but contribute little KL. Moreover, the same Key noise affects the modalities differently (Figure 1(b)). Text, which receives the largest share of attention, is affected mainly through re-ranking, whereas audio and video mainly gain spurious attention mass and, since their logits are closer to ties, flip their top-ranked Keys more often.

Insight (i): Key quantization should concentrate its precision on the few directions that queries read, which the weights determine without any query, and should allocate this precision per modality, since the same noise re-ranks text but inflates the attention mass of audio and video.

Observation (ii): The three modalities form distinct populations in one cache, and a shared code favors the majority. Under the read metric, the Keys of the three modalities form separate populations, each with its own mean and principal axis (Figure 2(a)). Quantizing one modality with the code fitted to another substantially increases the attention KL, even between audio and video, which prior methods treat identically. A single code fitted to the whole cache is dominated by video, the majority of tokens, and consequently serves video well but text and audio poorly (Figure 2(b), top). Quantization groups that span modality boundaries exhibit the same bias. The mean introduces a further mismatch. With a single global mean, the uncoded directions retain a modality-level offset that shifts attention mass across modalities, a failure absent from single-modality caches (Figure 2(b), bottom).

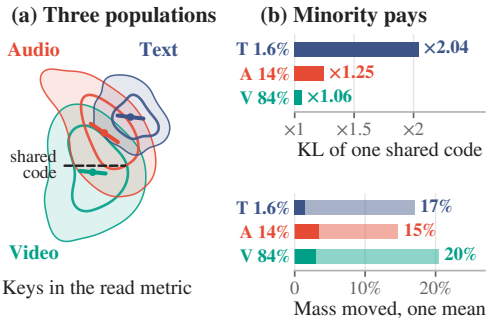


Figure 2: **One cache, three populations.** (a) Keys in the read metric; each modality has its own principal axis (solid), whereas a shared code follows that of the whole cache (dashed). (b) KL of one shared code relative to per-modality codes (top) and attention mass moved by one global mean (bottom); labels: share of cache tokens.

Insight (ii): Each modality should be coded with its own mean, basis, and bit allocation, and no quantization group should span a modality boundary, as mixed groups favor the majority.

3 Methodology

3.1 Overview

Building on these observations, we propose LMTC, a training-free KV cache quantizer with two components (Figure 3): (i) *logit-metric transform coding* for Keys (Sec. 3.2), which codes each modality in its own basis and allocates bits to the directions that queries read; and (ii) *phase-offset rounding* for Values (Sec. 3.3), which makes the rounding errors of redundant video tokens cancel in the attention readout. Neither component requires query activations. The Key metric is derived from the model weights, and all statistics are estimated offline and are position-independent. Each token is thus encoded once upon entering the cache, and the same coder applies to both offline and streaming inputs. Keys and Values share a budget of B bits per scalar, including metadata.

3.2 Logit-Metric Transform Coding for Keys

Following Insights (i) and (ii), we measure a Key error by the logit error it induces, under a metric derived solely from the model weights, and code each modality in its own basis under this metric.

Query-Free Logit Metric. A Key error $e_i = \widehat{k}_i - k_i$ shifts the logit of a reader q by $q^\top e_i/\sqrt{d}$. Since attention is invariant to a common shift of all logits, the resulting distortion depends only on how this shift varies across candidate Keys (Appendix D.1). As the readers are unknown when the cache is written, we minimize the expected variation

$$\mathbb{E}_q \mathrm{Var}_i\big(q^\top e_i\big) = \mathrm{tr}\big(\Sigma_q \,\mathrm{Cov}_i(e)\big), \qquad \Sigma_q = \mathbb{E}[qq^\top]. \tag{1}$$

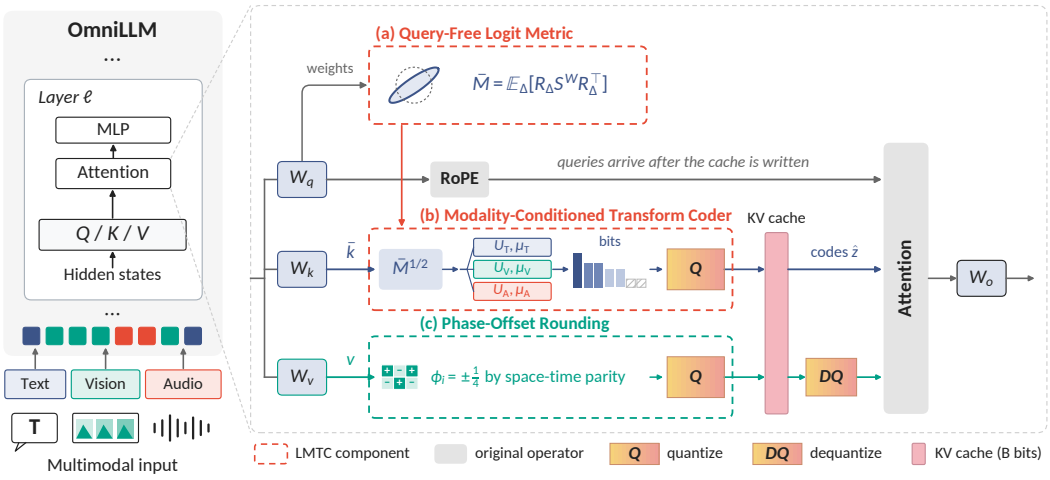


Figure 3: **Overview of LMTC** within the attention block (dashed frames: LMTC components). (a) The Query-Free Logit Metric $\bar{M}$ is computed offline from the query projection weights, since queries arrive only after the cache is written. (b) Each pre-rotary Key is whitened by $\bar{M}$, coded in the basis of its modality with bits allocated by reverse water-filling. (c) Video Values are rounded on grids offset by space-time parity, so that redundant errors cancel.

This objective weights each Key direction by the query energy along it, and it reduces to the reconstruction error minimized by standard Key quantizers when $\Sigma_q \propto I$.

To obtain Σ_q without observing queries, we write the query of head g before the rotary embedding as $\bar{q} = W_{q,g}(\gamma \odot \tilde{x}) + b_{q,g}$, where $\tilde{x} = x/\mathrm{rms}(x)$ is the normalized hidden state and γ is the RMSNorm gain. Because RMSNorm fixes the norm of $\tilde{x}$, only its direction is unknown, and we model it with the maximum-entropy prior, namely the uniform distribution on the sphere. For a KV head h shared by the query heads $g \in h$, the second moment of $\bar{q}$ under this prior is

$$S^W = \frac{1}{|h|} \sum_{g \in h} \left(W_{q,g} \,\mathrm{diag}(\gamma^2) W_{q,g}^\top + b_{q,g} b_{q,g}^\top \right). \tag{2}$$

Although real hidden states are not isotropic, S^W closely approximates the second moment of real queries for two reasons. First, the bias term, which the prior retains exactly, accounts for most of the query energy. Second, the input-dependent component is close to isotropic after projection by $W_q \,\mathrm{diag}(\gamma)$.

The rotary embedding makes this metric position-dependent. A reader at position n and a Key at position p produce the score $q^\top k = \bar{q}^\top R_{p-n} \bar{k}$, where $\bar{k}$ is the pre-rotary Key and R_Δ rotates coordinate pair j by angle $\theta_j \Delta$. We therefore code Keys in the pre-rotary frame, de-rotating each Key by its own position. In this frame, Key statistics are independent of position, and a pre-rotary error $\bar{e}$ changes the logit by $(R_\Delta \bar{q})^\top \bar{e}$ with $\Delta = n - p$. Since the offset is unknown when the Key is written, we take the expectation over a prior w:

$$\bar{M} = \mathbb{E}_{\Delta \sim w} \left[R_\Delta S^W R_\Delta^\top \right], \tag{3}$$

where w is log-uniform between the length W_0 of the BF16 residual window and the maximum context length. For coordinate pairs that rotate substantially over this range, the orientation of the reader is undetermined, and $\bar{M}$ is accordingly isotropic within each such pair; this loss is inherent to encoding a Key before its readers are known. Post-rotary statistics, in contrast, would depend on the positions of the calibration prompts (Appendix D.3).

Modality-Conditioned Transform Coder. For each layer, KV head, and modality m, we estimate the mean μ_m and covariance C_m of pre-rotary Keys offline. Let $A = \bar{M}^{1/2}$ and let $AC_mA = U_m \Lambda_m U_m^\top$ be an eigendecomposition. A Key of modality m is coded as

$$z_i = U_m^\top A \,(\bar{k}_i - \mu_m), \qquad \bar{k}_i = P_m z_i + \mu_m, \quad P_m = A^{-1} U_m. \tag{4}$$

The coefficients are uncorrelated with variances $\lambda_{m,j}$; an error ϵ yields $\bar{e} = P_m\epsilon$ with $\bar{e}^\top \bar{M}\bar{e} = \|\epsilon\|^2$, so Euclidean error in these coordinates is the expected logit error under the prior (Appendix D.2). Coding a modality in the basis of another never reduces this error (Appendix D.5).

Since both the metric and the Key covariance are strongly anisotropic, the variances $\lambda_{m,j}$ differ widely, so we allocate bits by reverse water-filling, assigning each coefficient $b_j \in \{0, 2, 3, 4\}$ bits:

$$\min_b \sum_j \lambda_{m,j}\, \kappa(b_j) \quad \text{s.t.} \quad \sum_j \left(b_j + \mathbb{1}[b_j > 0]\, \tfrac{32}{G}\right) \le Bd, \tag{5}$$

where $\kappa(b)$ denotes the relative distortion of b-bit min–max rounding over G tokens, with $\kappa(0) = 1$, and the second term accounts for the BF16 scale and zero point of each coded coefficient. Coefficients assigned zero bits are reconstructed at the modality mean and incur no storage, and their metadata budget is reassigned to high-variance coefficients. Each coded coefficient is quantized over G consecutive tokens of the same modality, so that no group spans a modality boundary.

Scoring Without Reconstruction. Scoring pre-rotary codes directly would require rotating the query once for every Key. We avoid this cost by partitioning the coordinate pairs according to their rotation frequency. For *fast* pairs, whose orientation is averaged out by the prior, $\bar{M}$ is isotropic within each pair and decoupled across pairs, and Keys at different positions are decorrelated in the post-rotary frame. These pairs therefore need no basis, and we quantize their post-rotary coordinates directly after subtracting the rotated mean $R_{p_i}\mu_m$. *Slow* pairs are coded with Equation 4 restricted to these coordinates, applied to Keys de-rotated by a reference position r_s shared by a segment s of consecutive tokens, within which the slow pairs rotate only negligibly. Both blocks share a single allocation (Equation 5), and a query q scores a Key i of modality m in segment s as

$$q^\top k_i = q^\top R_{p_i}\mu_m + q_{\mathcal{F}}^\top \widehat{z}_i^{\mathcal{F}} + \left(P_{m,\mathcal{S}}^\top\, [R_{r_s}^\top q]_{\mathcal{S}}\right)^\top \widehat{z}_i^{\mathcal{S}}, \tag{6}$$

where $\mathcal{F}$ and $\mathcal{S}$ index the fast and slow coordinates, respectively, and zero-bit coefficients are skipped. The query is thus rotated and projected once per segment rather than once per Key.

3.3 Phase-Offset Rounding for Values

Values require a different objective. Given attention weights a_i and errors $e_i = \widehat{v}_i - v_i$, the readout error $\Delta o = \sum_i a_i e_i$ satisfies

$$\|\Delta o\|_2^2 = \textstyle\sum_i a_i^2 \|e_i\|_2^2 + 2\sum_{i<j} a_i a_j \langle e_i, e_j\rangle, \tag{7}$$

where per-token MSE accounts only for the first term. Unlike Keys, Values cannot discard directions, since a Value direction without bits replaces the corresponding component of the readout with its mean, removing the signal itself. We therefore quantize each Value token to INT2 in its original coordinates, with a BF16 scale s_i and minimum β_i per G channels, and target the cross term.

The cross term is significant for video. Static regions produce nearly identical Values across frames, to which round-to-nearest assigns nearly identical errors; attention then accumulates these errors coherently rather than averaging them out. To decorrelate them, we shift the rounding grid of each video token by a quarter step, with the sign given by the parity of its space-time position (t_i, h_i, w_i):

$$\phi_i = \pm\tfrac{1}{4},\ \text{by the parity of } t_i + h_i + w_i, \qquad \widehat{v}_{ic} = \beta_i + s_i\Big(\phi_i + \operatorname{clip}\left(\left\lfloor \tfrac{v_{ic}-\beta_i}{s_i} - \phi_i \right\rceil, 0, 3\right)\Big). \tag{8}$$

Neighboring tokens in space and time are thus rounded on grids offset by half a step. The errors of two equal Values become negatively correlated, and their average behaves as if rounded with half the step (Appendix D.4). The worst-case error remains $s_i/2$, and the phase, being fixed by the token position, requires no metadata. Audio and text tokens exhibit less redundancy and use $\phi_i = 0$.

4 Experiments

4.1 Setups

Models and benchmarks. We evaluate Qwen2.5-Omni-3B/7B (Xu et al., 2025a) and OmniVinci-9B (Ye et al., 2025) on Daily-Omni (Zhou et al., 2025), WorldSense (Hong et al., 2025), OmniVideoBench (Li et al., 2025a), MMAU-Pro (Kumar et al., 2025), AVUT (Yang et al., 2025b),

Table 1: **Task performance of INT2 KV cache methods** on three OmniLLMs. All quantized methods use the same storage budget (three effective bits per scalar, residual 128).

		Audio-Visual						Audio	
	Method	Daily-Omni	WorldSense	OmniVideo	AVUT	VideoMME	OmniBench	MMAU-Pro	Avg
Qwen2.5 Omni-3B	BF16	59.0	46.0	31.9	62.5	62.7	45.9	47.1	50.7
	KIVI	58.7	44.5	31.7	61.2	60.8	44.6	46.3	49.7
	OTT	58.4	44.8	31.6	59.9	61.2	44.3	46.0	49.5
	KVTC	58.5	44.9	31.5	59.7	61.1	44.1	45.9	49.4
	OScaR	56.1	44.4	30.9	61.4	61.3	44.3	46.5	49.3
	LMTC	59.2	45.9	32.7	62.0	61.9	45.4	47.0	50.6
Qwen2.5 Omni-7B	BF16	61.5	46.6	34.0	65.1	64.8	49.1	50.0	53.0
	KIVI	58.2	45.8	33.5	62.8	63.2	48.3	47.8	51.4
	OTT	59.6	45.5	33.2	64.0	63.5	47.8	48.0	51.7
	KVTC	59.4	45.7	33.5	64.1	63.8	47.6	47.9	51.7
	OScaR	61.0	46.4	32.9	64.0	64.0	47.6	49.2	52.2
	LMTC	60.8	46.7	34.4	64.8	64.8	49.0	50.0	52.9
OmniVinci 9B	BF16	61.1	50.0	35.4	65.5	66.7	45.7	54.8	54.2
	KIVI	60.9	48.8	34.5	64.8	65.1	45.1	53.3	53.2
	OTT	60.4	48.1	33.9	64.3	65.0	44.8	52.6	52.7
	KVTC	60.3	48.5	34.3	64.6	65.2	45.0	52.8	53.0
	OScaR	59.7	48.1	33.7	64.4	65.1	44.3	54.2	52.8
	LMTC	61.1	49.0	34.6	65.3	66.4	45.4	54.8	53.8

Video-MME (Fu et al., 2024a), OmniBench (Li et al., 2025c), and StreamingBench (Lin et al., 2024a). We report question accuracy, except for MMAU-Pro's official aggregate score.

Baselines and settings. We compare KIVI (Liu et al., 2024), OTT (Su et al., 2025a), KVTC (Staniszewski & Łańcucki, 2025), and OScaR (Su et al., 2026) against BF16. For a fair comparison, all quantized methods use the same budget of three effective bits per scalar and a BF16 residual of the most recent 128 tokens across text, audio, and visual KV states. KIVI, OTT, and OScaR store INT2 codes with a BF16 scale and zero point per group of 32, that is, $2 + 32/32 = 3$ bits per scalar. For KVTC, we set the budget of its dynamic-programming bit allocation to three bits per scalar including its per-group scales and shifts. For OTT, the number of high-precision outlier tokens is set to 5. Each method retains its own cache commit and protected-token policy. OmniVinci uses native 128-frame inputs.

4.2 Main Results

Performance across model scales. Table 1 compares LMTC with existing INT2 methods on Qwen2.5-Omni-7B and 3B. On the 3B model, LMTC achieves an average score of 50.6, retaining 99.8% of the BF16 performance and outperforming the strongest quantized baseline by 0.9 points. It ranks first among the quantized methods on all seven benchmarks. This advantage persists at the 7B scale: LMTC obtains the highest average score among the quantized methods at 52.9, retaining 99.8% of BF16 performance and exceeding the strongest quantized baseline by 0.7 points, and leads the quantized methods on six of the seven benchmarks.

Generalization across architectures. On OmniVinci-9B, LMTC achieves the best average score among the INT2 methods at 53.8, exceeding the strongest baseline by 0.6 points while retaining 99.3% of BF16 performance. It also matches BF16 on Daily-Omni and MMAU-Pro, indicating that the design transfers beyond the Qwen2.5-Omni architecture.

Performance on streaming understanding. We further evaluate LMTC on streaming understanding with the Omni-Source subset of StreamingBench (Lin et al., 2024a). As shown in Table 2, LMTC retains 99.4% of BF16 accuracy and leads all quantized baselines on every task, exceeding OScaR by 1.2 and KIVI by 1.8 points, which shows its query-free design carries over to streaming input.

Table 2: **Performance of StreamingBench Omni-Source** on Qwen2.5-Omni-7B.

Method	ER	SCU	SD	MA	Avg.
BF16	48.0	22.4	50.0	70.0	47.6
KIVI	45.4	20.8	47.6	68.3	45.5
OScaR	46.3	21.2	48.2	68.7	46.1
LMTC	**47.6**	**22.1**	**49.6**	**69.7**	**47.3**

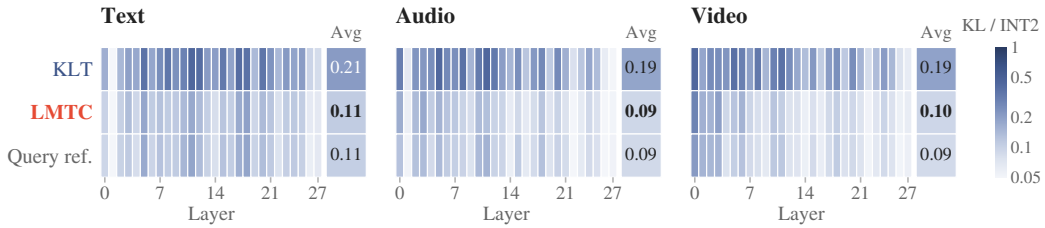


Figure 4: **Key metric across layers** (Qwen2.5-Omni-7B, $B = 3$, only the shown modality quantized). Each cell is the attention KL of one layer relative to per-channel INT2 (lower is better); all rows share the same transform coder, and the query metric is a reference unavailable to LMTC.

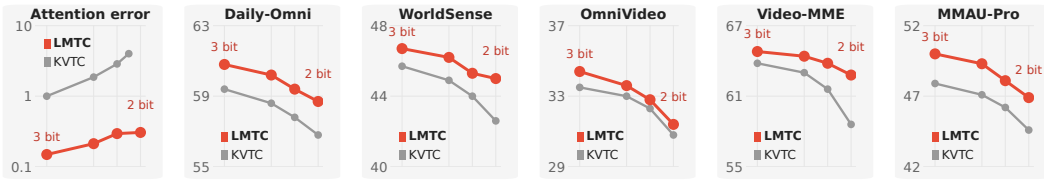


Figure 5: **LMTC vs. KVTC** on Qwen2.5-Omni-7B as the budget falls from three to two bits. Left: attention output error relative to KVTC at $B = 3$ (log scale). Others: task accuracy.

4.3 Ablation Studies

Effect of each component. Table 3 adds the components cumulatively to the KIVI baseline. On Daily-Omni, AVUT, and VideoMME, coding Keys in the weight-derived metric, still with one shared basis and mean and uniform bits, adds 1.1, 0.7, and 0.6 points; per-modality means and bases add 0.6, 0.6, and 0.3; reverse water-filling, which drops low-variance coefficients and pays their metadata to the rest, adds 0.3, 0.2, and 0.4; and phase-offset Values add 0.6, 0.5, and 0.3, whereas random dither yields 60.1, 64.0, and 64.4.

Key metric. Figure 4 fixes the transform coder and changes only the metric that defines the basis and the allocation. The reconstruction metric of KLT removes about 80% of the INT2 divergence. The weight-derived logit metric removes 88–93%, improves on KLT in 83 of the 84 layer–modality cells, and stays within a factor of 1.4 of the query reference per layer, although it never observes a query. The Keys' own second moment and the diagonal of the weight metric without rotation remain 1.4–2.3 times worse (Table 5), so both the query-side metric and the decorrelating rotation are needed.

Table 3: **Component ablation on Qwen2.5-Omni-7B** at $B = 3$. Components are added cumulatively to the KIVI baseline; the last row only replaces phase-offset rounding with random dither.

Method	Daily-Omni	AVUT	VideoMME
BF16	61.5	65.1	64.8
KIVI	58.2	62.8	63.2
+ Logit-metric transform	59.3	63.5	63.8
+ Per-modality mean, basis	59.9	64.1	64.1
+ Water-filling, zero bits	60.2	64.3	64.5
+ Phase-offset Values	**60.8**	**64.8**	**64.8**
Random dither for Values	60.1	64.0	64.4

Weight-derived vs. query-derived metrics. We replace Σ_q in Equation 1 by the second moment of real queries (Table 4). For the in-domain setting, the calibration queries are drawn from a held-out subset of the evaluated benchmark itself; for the cross-data setting, an equal number of prompts is sampled at random from other benchmarks. In-domain calibration comes close to S^W, but degrades quickly once the calibration data leave the domain. S^W covers directions the query projection can produce, independently of the question.

Table 4: **Weight- vs. query-derived metrics** at $B = 3$. KL relative to INT2 for readers after the prefix (Near) or 256K away (Far). BF16 accuracy: 61.5 / 64.8.

Metric source	Requires	Near	Far	Daily-Omni	VideoMME
Weights, S^W	**nothing**	**0.07**	**0.08**	**60.8**	**64.8**
Calib., in-domain	calibration	0.09	0.10	60.6	64.4
Calib., cross-data	calibration	0.13	0.17	59.6	63.7
Generic prompt	extra prefill	0.16	0.18	59.3	63.5

Budget. Figure 5 lowers the effective budget from three to two bits per scalar. We compare with KVTC, whose bit allocation, like that of LMTC, adapts to any budget; KIVI can lower its budget

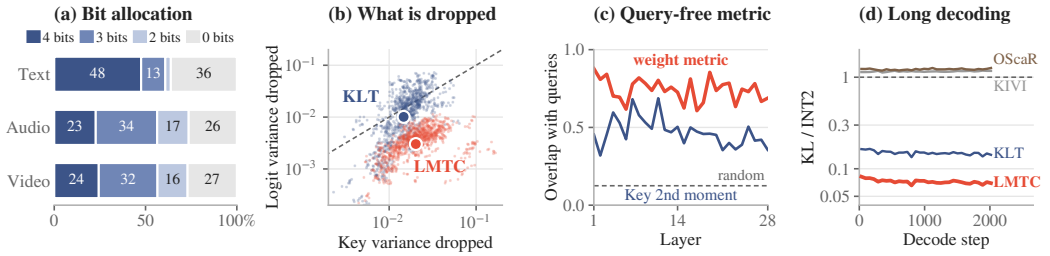


Figure 6: **Analysis of LMTC** (Qwen2.5-Omni-7B, $B = 3$). (a) Share of Key coefficients per bit width. (b) Key variance and real-query logit variance held by the zero-bit coefficients (Audio and Video; dots: cells, circles: geometric means; dashed: $y = x$). (c) Top-16 subspace overlap of each metric with the second moment of real queries. (d) Key KL relative to INT2 along reasoning generations of up to 2K tokens.

only through larger groups and collapses below three bits, falling to 34% on Daily-Omni. At the attention level, LMTC stays well below KVTC at every budget, and at two bits it keeps about a third of the output error of KVTC at three bits (Figure 5, left). Task accuracy follows the same trend: LMTC stays above KVTC at every budget on all five benchmarks. At $B = 2$ it loses 1.6–3.1 points relative to BF16, against 3.2–6.2 points for KVTC, and its margin over KVTC widens on four of the five benchmarks, most on Video-MME (from 1.0 to 4.2 points).

Analysis. Figure 6 looks inside the Key coder. About a quarter of the Audio and Video coefficients and a third of the Text coefficients receive zero bits (a). Dropping coefficients is standard in transform coding; what matters is which ones are dropped. The zero-bit coefficients of LMTC hold slightly more Key variance than those of KLT, but only 0.41% and 0.35% of the logit variance read by real question queries for Audio and Video, against 1.8% and 1.6% for KLT (b). The weight metric shares about three quarters of its top-16 subspace with the second moment of real queries (at least 0.6 in every layer), against about half for the Keys' own second moment and an eighth for a random subspace (c). Finally, the advantage persists when the readers are the tokens of long reasoning chains (200 chain-of-thought answers and 4 multi-turn conversations of about 2K tokens): over all readers, LMTC keeps the Key KL at 0.08 of INT2, against 0.16 for KLT and 0.99–1.23 for OTT, KIVI, and OScaR (d; Appendix D.6).

4.4 Efficiency

LMTC needs no training, and its offline statistics (the per-modality Key moments and the weight term of Equation 2) occupy about 15 MB for the 7B model, so no decomposition runs per request. Per decode step, each query head adds one dot product per Key for the rotated mean and one projection per modality and segment (Equation 6), and scores the coded Keys without reconstruction. We evaluate packed-cache decoding on Qwen2.5-Omni-7B using A100 40GB GPUs, against OScaR and FD2, an uncompressed BF16 cache decoded with FlashDecoding-v2 (Dao et al., 2023). The evaluation measures batch-one latency over 1K–256K contexts and, at a fixed 4K context, throughput over batch sizes 1–48. Timing covers the Thinker decoder, LM head, and cache updates, excluding prefill, media and prefix encoding. As shown in Figure 7, LMTC scales favorably with both context length and batch size. At 256K context it reduces latency from FD2's 75.39 to 38.86 ms/token, a $1.94\times$ speedup. At batch 48, throughput increases from 683.65 to 1,424.88 tokens/s ($2.08\times$). We report KV-cache storage only, excluding model memory: at 4K/batch 48, FD2, OScaR, and LMTC use 11.854/2.609/2.279 GiB, a $5.20\times$ reduction for LMTC; at 256K/batch one, 14.028/2.686/2.658 GiB, a $5.28\times$ reduction.

5 Related Work

Omni-modal Large Language Models. OmniLLMs process interleaved text, audio, images, and video jointly (Xu et al., 2025a;b). Architectural advancements—such as multimodal instruction alignment (Fu et al., 2024b; Li et al., 2025b) and time-aligned token interleaving (Xu et al., 2025a;b;

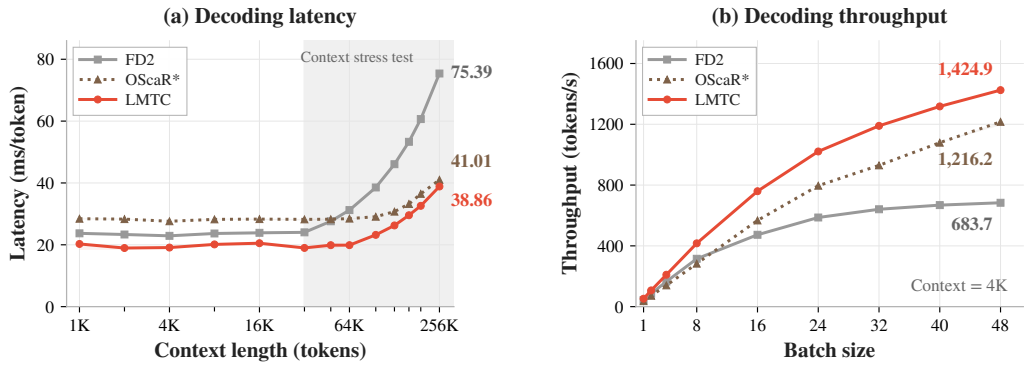


Figure 7: **Decoder efficiency on Qwen2.5-Omni-7B.** Batch-one latency across context lengths (left) and throughput across batch sizes at a 4K context (right). Bands show the observed range; contexts of 32K and above are stress tests. FD2: BF16 cache with FlashDecoding-v2.

Ye et al., 2025)—enable joint reasoning over long audio-visual inputs. However, OmniLLMs face severe KV cache memory bottlenecks due to two core properties. First, dense video and audio tokenization makes the cache grow linearly with input duration, often rivaling or exceeding model weight memory. Existing token-reduction methods prune or merge tokens (Di et al., 2025; Yang et al., 2025a; Jiang et al., 2026; Sun et al., 2026), but leave the per-token memory footprint untouched. Second, audio-visual contexts are typically cached before the question arrives and reused across queries. These challenges call for an OmniLLM-specific KV cache quantizer that operates without unobserved queries while accommodating cross-modal statistical heterogeneity.

KV Cache Quantization. Existing KV cache quantizers fall into two paradigms, neither of which fully suits OmniLLMs. The first paradigm minimizes element-wise reconstruction errors using techniques like channel-wise quantization (Liu et al., 2024), pre-rotary embedding quantization (Hooper et al., 2024; Su et al., 2025b), outlier smoothing/rotation (Lin et al., 2024b; Ashkboos et al., 2024; Su et al., 2025a), low-rank compensation (Kang et al., 2024), or transform coding (Staniszewski & Łańcucki, 2025). Though recently extended to visual inputs (Han et al., 2025; Tao et al., 2025; Su et al., 2026), these reconstruction-based objectives treat all error directions equally, letting unified codebooks be dominated by majority video tokens and degrading text and audio precision. The second paradigm applies attention-aware weighting via runtime queries (He et al., 2024; Wang et al., 2025; Xu et al., 2025c) or offline calibration statistics (Laus et al., 2026; Zhou et al., 2026). However, runtime queries are unavailable during pre-caching, while calibration statistics rely on specific offline datasets and typically aggregate across modalities. LMTC instead introduces a query-free logit metric derived from query projection weights, enabling single-pass token encoding upon cache entry. Additionally, it models modality heterogeneity through dedicated bases and bit allocations for Keys, and uses Phase-Offset Rounding with dithered quantization principles (Roberts, 1962; Schuchman, 1964) to mitigate rounding error accumulation across redundant visual Values.

6 Conclusion

In this work, we present LMTC, a training-free KV cache quantizer for OmniLLMs that uses no query values. We show that attention damage follows the logit error that queries read rather than Key reconstruction error, that modalities differ in the geometry of their Keys rather than in the shape of their distributions, and that redundant video Values accumulate their rounding errors in the readout. Guided by these findings, LMTC codes each modality's Keys in a basis that is optimal for a logit metric derived from the query projection weights, allocates bits by reverse water-filling, and scores attention on the codes, while it keeps Values in their own coordinates and decorrelates the rounding of redundant video tokens with a position-determined phase. Our analysis suggests that low-bit KV caches should be coded for how attention reads them, which for Keys is a question of metric and basis rather than of reconstruction.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, assist with the implementation of methods, support the interpretation of experimental results, and assist with the writing, editing, and translation of the manuscript. We have not used generative AI tools to propose and refine research hypotheses, generate synthetic datasets, formulate mathematical claims, provide ingredients for or assist with mathematical proofs, clean or reformat datasets, or conduct qualitative or thematic data analysis. The remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to improve the organization, clarity, and presentation of the manuscript. All AI-assisted written content were independently reviewed by the authors. AI-assisted code and experimental configurations were verified through manual inspection and testing, and AI-assisted interpretations and research suggestions were checked against experimental evidence and relevant literature.

REPRODUCIBILITY STATEMENT

Appendix A defines the quantization conventions, while Appendices B and C describe the baselines, benchmarks, and evaluation metrics. Appendix D provides the mathematical assumptions, derivations, and evidence for the Key coder.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-Free 4-Bit Inference in Rotated LLMs. *arXiv preprint arXiv:2404.00456*, 2024. URL https://arxiv.org/abs/2404.00456.

Tri Dao, Daniel Haziza, Francisco Massa, and Grigory Sizov. Flash-Decoding for long-context inference. PyTorch Blog, https://pytorch.org/blog/flash-decoding/, 2023.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming Video Question-Answering with In-context Video KV-Cache Retrieval. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8g9fs6mdEG.

Haojie Duanmu, Zhihang Yuan, Xiuhong Li, Jiangfei Duan, Xingcheng Zhang, and Dahua Lin. SKVQ: Sliding-window Key and Value Cache Quantization for Large Language Models. *arXiv preprint arXiv:2405.06219*, 2024. URL https://arxiv.org/abs/2405.06219.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*, 2024a. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. VITA: Towards Open-Source Interactive Omni Multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2024b. URL https://arxiv.org/abs/2408.05211.

Insu Han, Zeliang Zhang, Zhiyuan Wang, Yifan Zhu, Susan Liang, Jiani Liu, Haiting Lin, Mingjie Zhao, Chenliang Xu, Kun Wan, and Wentian Zhao. CalibQuant: 1-Bit KV Cache Quantization for Multimodal LLMs. *arXiv preprint arXiv:2502.14882*, 2025. URL https://arxiv.org/abs/2502.14882.

Yefei He, Luoming Zhang, Weijia Wu, Jing Liu, Hong Zhou, and Bohan Zhuang. ZipCache: Accurate and Efficient KV Cache Quantization with Salient Token Identification. *arXiv preprint arXiv:2405.14256*, 2024. URL https://arxiv.org/abs/2405.14256.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs. *arXiv preprint arXiv:2502.04326*, 2025. URL https://arxiv.org/abs/2502.04326.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization. *arXiv preprint arXiv:2401.18079*, 2024. URL https://arxiv.org/abs/2401.18079.

Zhonghua Jiang, Kui Chen, Kunxi Li, Keting Yin, Yiyun Zhou, Zhaode Wang, Chengfei Lv, and Shengyu Zhang. AccKV: Towards Efficient Audio-Video LLMs Inference via Adaptive-Focusing and Cross-Calibration KV Cache Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(7):5494–5502, 2026. doi: 10.1609/aaai.v40i7.37467. URL https://ojs.aaai.org/index.php/AAAI/article/view/37467.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. GEAR: An Efficient KV Cache Compression Recipe for Near-Lossless Generative Inference of LLM. *arXiv preprint arXiv:2403.05527*, 2024. URL https://arxiv.org/abs/2403.05527.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence. *arXiv preprint arXiv:2508.13992*, 2025. URL https://arxiv.org/abs/2508.13992.

Hannah Laus, Claudio Mayrink Verdun, Hao Wang, Flavio du Pin Calmon, and Felix Krahmer. KV Cache Compression Through the Lens of Transform Coding. *arXiv preprint arXiv:2608.14191*, 2026. URL https://arxiv.org/abs/2608.14191.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, et al. OmniVideoBench: Towards Audio-Visual Understanding Evaluation for Omni MLLMs. *arXiv preprint arXiv:2510.10689*, 2025a. URL https://arxiv.org/abs/2510.10689.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-Omni-1.5 Technical Report. *arXiv preprint arXiv:2501.15368*, 2025b. URL https://arxiv.org/abs/2501.15368.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, et al. OmniBench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 2025c. URL https://arxiv.org/abs/2409.15272.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding. *arXiv preprint arXiv:2411.03628*, 2024a. URL https://arxiv.org/abs/2411.03628.

Yujun Lin, Haotian Tang, Shang Yang, Zhekai Zhang, Guangxuan Xiao, Chuang Gan, and Song Han. QServe: W4A8KV4 Quantization and System Co-design for Efficient LLM Serving. *arXiv preprint arXiv:2405.04532*, 2024b. URL https://arxiv.org/abs/2405.04532.

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. URL https://arxiv.org/abs/2402.02750.

Lawrence G. Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, 1962.

Leonard Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, 12(4):162–165, 1964.

Akshat Sharma, Hangliang Ding, Jianping Li, Neel Dani, and Minjia Zhang. MiniKV: Pushing the Limits of 2-Bit KV Cache via Compression and System Co-Design for Efficient Long Context Inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18506–18523. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.952. URL https://aclanthology.org/2025.findings-acl.952/.

Konrad Staniszewski and Adrian Łańcucki. KV cache transform coding for compact storage in LLM inference. *arXiv preprint arXiv:2511.01815*, 2025.

Yi Su, Yuechi Zhou, Quantong Qiu, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Accurate KV Cache Quantization with Outlier Tokens Tracing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12895–12915. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.631. URL https://aclanthology.org/2025.acl-long.631/.

Zunhai Su, Hanyu Wei, Zhe Chen, Wang Shen, Linge Li, Huangqi Yu, and Kehong Yuan. RotateKV: Accurate and Robust 2-Bit KV Cache Quantization for LLMs via Outlier-Aware Adaptive Rotations. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6200–6208, 2025b. doi: 10.24963/ijcai.2025/690. URL https://www.ijcai.org/proceedings/2025/690.

Zunhai Su, Rui Yang, Chao Zhang, Yaxiu Liu, Yifan Zhang, Wei Wu, Jing Xiong, Dayou Du, Xialie Zhuang, Yulei Qian, Yuchen Xie, Yik-Chung Wu, Hongxia Yang, and Ngai Wong. OScaR: The Occam's Razor for Extreme KV Cache Quantization in LLMs and Beyond. *arXiv preprint arXiv:2605.19660*, 2026. URL https://arxiv.org/abs/2605.19660.

Guangzhi Sun, Yixuan Li, Yudong Yang, and Chao Zhang. OmniMem: Perturbation-aware Memory Compression for Streaming Audio-Visual LLMs. *arXiv preprint arXiv:2606.07577*, 2026. URL https://arxiv.org/abs/2606.07577.

Keda Tao, Haoxuan You, Yang Sui, Can Qin, and Huan Wang. Plug-and-Play 1.x-Bit KV Cache Quantization for Video Large Language Models. *arXiv preprint arXiv:2503.16257*, 2025.

Hao Wang, Ligong Han, Kai Xu, and Akash Srivastava. SQuat: Subspace-orthogonal KV Cache Quantization. *arXiv preprint arXiv:2503.24358*, 2025. URL https://arxiv.org/abs/2503.24358.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025b. URL https://arxiv.org/abs/2509.17765.

Yuhui Xu, Zhanming Jie, Hanze Dong, Lei Wang, Xudong Lu, Aojun Zhou, Amrita Saha, Caiming Xiong, and Doyen Sahoo. ThinK: Thinner key cache by query-driven pruning. In *International Conference on Learning Representations*, 2025c.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, Zongyuan Ge, Muhammad Haris Khan, and Imran Razzak. StreamAgent: Towards Anticipatory Agents for Streaming Video Understanding. *arXiv preprint arXiv:2508.01875*, 2025a. URL https://arxiv.org/abs/2508.01875.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6569–6587, 2025b. doi: 10.18653/v1/2025.emnlp-main.333. URL https://aclanthology.org/2025.emnlp-main.333/.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, et al. OmniVinci: Enhancing architecture and data for omni-modal understanding LLM. *arXiv preprint arXiv:2510.15870*, 2025. URL https://arxiv.org/abs/2510.15870.

Zhongzhu Zhou, Donglin Zhuang, Jisen Li, Ziyan Chen, Shuaiwen Leon Song, Ben Athiwaratkun, and Xiaoxia Wu. OSCAR: Offline Spectral Covariance-Aware Rotation for 2-bit KV Cache Quantization. *arXiv preprint arXiv:2605.17757*, 2026. URL https://arxiv.org/abs/2605.17757.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities. *arXiv preprint arXiv:2505.17862*, 2025. URL https://arxiv.org/abs/2505.17862.

A PRELIMINARIES ON LOW-BIT QUANTIZATION

Low-bit quantization represents floating-point tensors using a finite set of discrete values. For KV cache compression, it reduces the storage required for previously computed keys and values during autoregressive inference. We introduce the basic formulation below; the bit-width and grouping policy are parameters of the quantizer rather than fixed experimental settings.

Uniform affine quantization. Let $\mathcal{G}$ denote a group of tensor elements sharing quantization parameters, and let b be the number of bits per element. A uniform affine quantizer maps each $x \in \mathcal{G}$ to an unsigned integer code:

$$q(x) = \mathrm{clip}\left(\left\lfloor \frac{x-\ell}{s} \right\rceil, 0, 2^b - 1\right), \qquad \widehat{x} = s\, q(x) + \ell, \tag{9}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, $\mathrm{clip}(\cdot, a, c)$ restricts its argument to $[a, c]$, $s > 0$ is the quantization step, and ℓ is the reconstruction offset. A simple min–max parameterization uses

$$\ell = \min_{x \in \mathcal{G}} x, \qquad u = \max_{x \in \mathcal{G}} x, \qquad s = \max\left(\frac{u-\ell}{2^b - 1}, \epsilon\right), \tag{10}$$

with a small positive ϵ to avoid division by zero for constant groups. The codes are stored at low precision, while the scale and offset are stored as auxiliary metadata. For values within the representable interval, rounding introduces an absolute reconstruction error of at most $s/2$. Consequently, a wider range or a lower bit-width increases the error bound. Clipping the range can reduce the step size but introduces additional error for values outside the selected interval.

Symmetric quantization. An alternative centers the reconstruction grid at zero. For $b \geq 2$, one common convention sets $q_{\max} = 2^{b-1} - 1$ and uses

$$\begin{aligned} s_{\mathrm{sym}} = \max\left(\frac{\max_{x \in \mathcal{G}} |x|}{q_{\max}}, \epsilon\right), \qquad \widehat{x} = s_{\mathrm{sym}} q_{\mathrm{sym}}(x), \\ q_{\mathrm{sym}}(x) = \mathrm{clip}\left(\left\lfloor \frac{x}{s_{\mathrm{sym}}} \right\rceil, -q_{\max}, q_{\max}\right). \end{aligned} \tag{11}$$

This convention uses $2^b - 1$ reconstruction levels and leaves one signed integer code unused. It requires no offset, whereas affine quantization can adapt the grid to a range that is not centered at zero. Non-uniform quantizers instead use unequally spaced reconstruction levels, as explored for KV caches by KVQuant (Hooper et al., 2024).

Quantization granularity. For a single attention head, let $K, V \in \mathbb{R}^{T \times d}$ denote the cached keys and values, where T is the number of cached tokens and d is the head dimension. Per-token quantization shares parameters across feature coordinates within a token, whereas per-channel quantization shares them across tokens within a feature channel. Group-wise variants partition the corresponding axis into smaller blocks with separate parameters. Smaller groups can better accommodate local ranges at the cost of additional metadata. KIVI (Liu et al., 2024), for example, uses different grouping axes for keys and values to account for their different distributions.

Quantized KV caches in attention. Given a query $\mathbf{q} \in \mathbb{R}^{1 \times d}$, attention with reconstructed caches is written as

$$\widehat{\mathbf{o}} = \mathrm{softmax}\left(\frac{\mathbf{q}\widehat{K}^\top}{\sqrt{d}}\right)\widehat{V}, \qquad \widehat{K} = K + E_K, \quad \widehat{V} = V + E_V, \tag{12}$$

where E_K and E_V are reconstruction errors. Key errors perturb the attention logits by $\mathbf{q}E_K^\top/\sqrt{d}$, while value errors affect the weighted aggregation. Thus, reconstruction error and downstream task quality measure different aspects of quantization. Integer cache storage also does not require all attention arithmetic to use integer operations; dequantization can be incorporated into the attention kernel.

Storage cost. For N cached scalar elements, a b-bit representation requires Nb bits for the codes under ideal packing. If the original precision is p bits, the ideal compression ratio is p/b. Accounting for scales, offsets, padding, and any high-precision exceptions gives the effective ratio

$$\mathrm{CR} = \frac{Np}{Nb + M_{\mathrm{aux}}}, \tag{13}$$

where M_{aux} is the additional storage in bits beyond the nominal low-bit codes. Practical comparisons should therefore account for auxiliary storage as well as the specified bit-width.

B Baseline Methods

We evaluate four KV cache quantization baselines at the task level, KIVI, OTT, KVTC, and OScaR. At the attention level we add KLT and, for long reasoning readers (Appendix D.6), the Key quantizer of VidKV (Tao et al., 2025). We briefly summarize their core designs below.

OScaR. OScaR (Su et al., 2026) addresses token norm imbalance, which can impair per-channel key quantization when tokens with different magnitudes share quantization parameters. It combines Canalized Rotation with Omni-Token Scaling to improve low-bit quantization, with evaluations covering language, multimodal, and omni-modal models.

KIVI. KIVI (Liu et al., 2024) is a tuning-free asymmetric KV cache quantization method. Motivated by the different distributions of keys and values, it quantizes keys per channel and values per token, enabling 2-bit cache compression without model fine-tuning.

OTT. Outlier Tokens Tracing (OTT) (Su et al., 2025a) identifies a small subset of tokens whose key distributions deviate from the assumptions underlying channel-wise quantization. It traces these tokens during decoding and excludes them from quantization, improving low-bit accuracy by preserving their representations at full precision.

KVTC. KVTC (Staniszewski & Łańcucki, 2025) compresses the KV cache by transform coding. It decorrelates the cache with PCA, allocates bits to the principal components by dynamic programming under a storage budget, and minimizes the reconstruction error of the cache, treating tokens of all modalities alike.

KLT. KLT isolates the objective of transform coders such as KVTC (Staniszewski & Łańcucki, 2025). For every layer, KV head, and modality it subtracts the modality mean of the pre-rotary Keys, rotates them into the eigenbasis of their covariance, allocates $\{0, 2, 3, 4\}$ bits per coefficient by reverse water-filling on the eigenvalues under the same budget and metadata accounting as LMTC, and quantizes each coefficient by min–max rounding along G tokens. It differs from LMTC only in the metric, $\Sigma_q \propto I$. Unlike the full KVTC baseline, it does not pool layers and heads into one basis, use dynamic-programming allocation, or add entropy coding.

C Benchmark Descriptions

We evaluate six audio-visual benchmarks and one audio benchmark in the main comparison, and one streaming benchmark.

Daily-Omni. Daily-Omni (Zhou et al., 2025) evaluates audio-visual reasoning with temporal alignment across modalities. We evaluate all 1,197 multiple-choice questions and report answer accuracy.

WorldSense. WorldSense (Hong et al., 2025) evaluates real-world omni-modal understanding using synchronized audio-visual videos and multiple-choice questions. The reported runs cover all 3,172 questions and use accuracy.

OmniVideoBench. OmniVideoBench (Li et al., 2025a) tests reasoning over complementary audio and visual evidence, including temporal reasoning, spatial localization, counting, and causal inference. We report accuracy on all 1,000 questions.

MMAU-Pro. MMAU-Pro (Kumar et al., 2025) evaluates audio understanding across speech, environmental sounds, music, and their combinations. We report the official aggregate score, which includes closed-form, open-ended, and instruction-following tasks. Each run contains 5,305 predictions; the official scorer excludes one single-choice item.

AVUT. AVUT (Yang et al., 2025b) emphasizes auditory information and audio-visual interactions. We evaluate all 1,734 released questions in the human subset with a fixed input, prompt, and answer-extraction protocol, with available audio and no subtitle input.

Video-MME. Video-MME (Fu et al., 2024a) covers video understanding across multiple durations. We evaluate all 2,700 questions with the same fixed protocol, sampled video and available audio, without subtitles.

OmniBench. OmniBench (Li et al., 2025c) evaluates joint image, audio, and text understanding. We use all 1,142 questions, original images and audio, and the author's multimodal user prompt. Canonical option answers and a corrected final-answer parser determine accuracy.

StreamingBench. StreamingBench (Lin et al., 2024a) evaluates streaming video understanding with questions posed at given timestamps. We use all 1,000 questions of its audio-visual Omni-Source subset, covering Emotion Recognition (ER), Scene Understanding (SCU), Source Discrimination (SD), and Multimodal Alignment (MA), and report per-task and average accuracy.

D METHOD: DERIVATIONS AND EVIDENCE

Unless stated otherwise, measurements use Qwen2.5-Omni-7B, all 28 layers and four KV heads. Readers are the question rows of all GQA heads; they serve evaluation only, and no method component sees them. Ratios are geometric means over layers and heads relative to the stated baseline.

D.1 SECOND-ORDER ATTENTION DIVERGENCE

Let ℓ be the clean logits on the visible support, $p = \mathrm{softmax}(\ell)$ and $\widehat{p} = \mathrm{softmax}(\ell+\delta)$. Expanding $\log \sum_i p_i e^{\delta_i}$ to second order gives

$$D_{\mathrm{KL}}(p\|\widehat{p}) = \log \textstyle\sum_i p_i e^{\delta_i} - \sum_i p_i \delta_i = \tfrac{1}{2}\mathrm{Var}_p(\delta) + O(\|\delta\|_\infty^3). \tag{14}$$

Across all layers and modalities, the logarithms of the exact KL and of $\frac{1}{2}\mathrm{Var}_p(\delta)$ correlate at 0.99 or more.

D.2 WEIGHT-INDUCED METRIC AND METRIC TRANSFORM CODING

Objective. Equation 14 weights the logit shifts by the attention p, which depends on the reader. Equation 1 replaces p by uniform weights over the candidates, which assumes that the attention weights do not depend on the quantization error, and drops the constant $1/d$. The coder minimizes the second moment $\mathbb{E}_q(q^\top e)^2 = e^\top \Sigma_q e$ per token, which upper-bounds the variance in Equation 1 and is tight when the errors have zero mean over the tokens of a modality.

Metric. Here D is the hidden size and d is the head dimension. Under the prior, $\tilde{x} = x/\mathrm{rms}(x)$ is uniform on the sphere of radius $\sqrt{D}$, so $\mathbb{E}\tilde{x} = 0$ and $\mathbb{E}\tilde{x}\tilde{x}^\top = I$; the gain enters afterwards as $\gamma \odot \tilde{x}$. The pre-rotary query second moment of head g under the prior is then $W_{q,g}\,\mathrm{diag}(\gamma^2)W_{q,g}^\top + b_{q,g}b_{q,g}^\top$, since the cross term $2\,b_{q,g}^\top W_{q,g}\,\mathrm{diag}(\gamma)\mathbb{E}\tilde{x}$ vanishes under the prior. Text readers carry the same position n in all three TMRoPE components (trivially so for 1-D RoPE), so the relative rotation R_Δ turns pair $(j, j+d/2)$ by $\theta_j\Delta$; for a Key, each pair uses the position component of its TMRoPE section. Write the 2×2 block of S^W between pairs j and j' as a rotation-commuting part $\alpha I + \beta J$, with $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, plus a reflection part. Under Equation 3, the first is multiplied by $\mathbb{E}_w e^{\mathrm{i}(\theta_j - \theta_{j'})\Delta}$ and the second by $\mathbb{E}_w e^{\mathrm{i}(\theta_j + \theta_{j'})\Delta}$, so $\bar{M}$ is closed-form; we evaluate the two expectations on a log-spaced grid of 4,096 offsets. Pairs whose coherence $|\mathbb{E}_w e^{2\mathrm{i}\theta_j\Delta}|$ is small become isotropic within their block and decouple from the other pairs; these are the fast pairs of Equation 6. We add $10^{-6}\lambda_{\max}$ to the eigenvalues of $\bar{M}$ before taking A and A^{-1}.

Equivalence. With $z = U_m^\top A(k - \mu_m)$ and $k = A^{-1}U_m z + \mu_m$, a coefficient error ϵ gives $e = A^{-1}U_m\epsilon$ and

$$e^\top \bar{M} e = \epsilon^\top U_m^\top A^{-1} A A A^{-1} U_m \epsilon = \|\epsilon\|^2. \tag{15}$$

The covariance of z is $U_m^\top A C_m A U_m = \Lambda_m$, so the coefficients are uncorrelated with variances $\lambda_{m,j}$ and the expected logit error under the prior decomposes additively over coefficients.

Allocation. For a coefficient of variance λ coded with b bits by min–max rounding over G tokens, the mean squared error is $\lambda\kappa(b)$ with $\kappa(b) = \bar{r}_G^2/\big(12(2^b-1)^2\big)$, where $\bar{r}_G$ is the expected range of G standard normal samples ($\bar{r}_{32} \approx 4.1$, $\bar{r}_{64} \approx 4.8$), and $\kappa(0) = 1$. Since $\kappa(1) > 1$, one bit is never used. Equation 5 is a multiple-choice knapsack; the increments $0 \to 2$, $2 \to 3$ and

$3 \to 4$ have decreasing gain per bit for each coefficient, and we add them greedily in order of gain per bit. At high rate and without the metadata term, the optimum is reverse water-filling, $b_j = \max\{0, \frac{1}{2}\log_2(\lambda_{m,j}/\theta)\}$ with the water level θ set by the budget.

D.3 Why Pre-Rotary Statistics

Post-rotary Key statistics depend on where the Keys sit. We estimate μ_m and C_m on post-rotary Keys of the calibration prompts, whose prefixes occupy the first 793 (Qwen2.5-Omni-7B) or about 4.9K (OmniVinci-9B) positions, and move each test request by Δ positions, rotating its Keys and readers by R_Δ. The clean logits are unchanged, and the metric follows the true reader position, so only the drift of the statistics is measured. Relative to the unshifted request, the Key attention KL (Text, Audio, Video) grows to 4.1, 2.9, 4.1 at $\Delta = 1$K, 5.5, 4.3, 5.8 at 16K, and 10.0, 9.8, 13.4 at 256K on Qwen2.5-Omni-7B, and to 3.9, 2.6, 1.2, then 4.8, 4.6, 3.3, then 7.5, 12.0, 9.3 on OmniVinci-9B. At 256K this is 1.3–2.2 times the KL of per-channel INT2. Replacing only the mean by the request's own mean recovers about half (2.0–3.1$\times$ at 16K–256K). Pre-rotary statistics remove the shift exactly, since each Key is de-rotated by its own position.

D.4 Phase-Offset Rounding

Measure a Value in units of the step s and let u be its position within the step, uniform on $[0, 1)$. Nearest rounding has error $\epsilon(u) = u - \lfloor u \rceil$, uniform on $[-\frac{1}{2}, \frac{1}{2})$ with variance $\frac{1}{12}$. Two tokens with the same Value and phases $\pm\frac{1}{4}$ have errors $\epsilon(u - \frac{1}{4})$ and $\epsilon(u + \frac{1}{4})$, which differ by exactly $\pm\frac{1}{2}$; their covariance is $2\int_0^{1/2} x(x - \frac{1}{2})\,dx = -\frac{1}{24}$, a correlation of $-\frac{1}{2}$. The mean of the two errors therefore has variance $\frac{1}{4}\big(\frac{1}{12} + \frac{1}{12} - \frac{2}{24}\big) = \frac{1}{48}$, against $\frac{1}{12}$ for aligned rounding and $\frac{1}{24}$ for independent errors: the pair behaves like one token rounded on a grid of half the step. With two phases, N identical tokens split evenly between the grids reach the same $\frac{1}{48}$ and no less, whereas independent errors would keep shrinking as $\frac{1}{12N}$.

D.5 Modality Mismatch

Proposition 1. *Let $M = AC_mA$ have eigenvalues Λ_m and let U' be any orthogonal basis. At high rate, the distortion of optimal allocation in basis U' is proportional to $\mathrm{GM}\big(\mathrm{diag}(U'^\top M U')\big)$, and*

$$\mathrm{GM}\big(\mathrm{diag}(U'^\top M U')\big) \;\geq\; \mathrm{GM}(\Lambda_m) = \det(M)^{1/d}, \tag{16}$$

with equality if and only if $U'^\top M U'$ is diagonal.

Proof. Reverse water-filling at high rate gives distortion $d\,2^{-2\bar{b}}\mathrm{GM}(\mathrm{diag}(U'^\top M U'))$ up to a constant. Hadamard's inequality gives $\prod_j (U'^\top M U')_{jj} \geq \det(U'^\top M U') = \det M$ for positive semidefinite M, with equality exactly for diagonal matrices. □

A basis fitted to a different modality or to a pooled mixture therefore loses coding gain whenever the subspaces differ.

D.6 Evidence for the Key Coder

Metric. Table 5 gives the per-modality values behind Figure 4 and adds two further metrics, at $B = 3$ with only the listed modality quantized.

The weight-derived metric matches a metric estimated from real queries, while using the Keys' own second moment as a proxy for queries recovers only part of the gain. The improvement holds in every depth band (LMTC Video: 0.15, 0.09, 0.07 for L0–7, L8–20, L21–27). Its Key MSE (0.29, 0.27, 0.26 of INT2 for Text, Audio, Video) is higher than that of the reconstruction metric (0.23, 0.21, 0.21), although its attention KL is about half, confirming that reconstruction error is the wrong target. At $B = 3$, on average 46, 34, and 35 of the 128 coefficients per head receive zero bits for Text, Audio, and Video (range 25–49). The coded coefficients carry 3.1–3.7 payload bits on average, that is, 2.3–2.4 payload bits per scalar, the rest of the budget paying for their metadata.

Design choices. Table 6 changes one design choice of the final coder at a time. A single basis shared by all modalities raises the KL to 0.22, 0.11, and 0.11 of INT2 for Text, Audio, and Video, and a

Table 5: **Key metric ablation at** $B = 3$ on Qwen2.5-Omni-7B, only the listed modality quantized. Key attention KL relative to per-channel INT2 ($G = 32$), geometric mean over layers and KV heads, and in parentheses the ratio of KL summed over layers. The question-query metric is a reference only.

Key metric for transform and allocation	Queries	Text	Audio	Video
Reconstruction, $\Sigma_q \propto I$ (KLT)	no	0.21 (0.26)	0.19 (0.15)	0.19 (0.14)
Key second moment, $\Sigma_q = \mathbb{E}[kk^\top]$	no	0.16 (0.18)	0.14 (0.11)	0.14 (0.09)
Weights, diagonal only (no rotation)	no	0.25 (0.26)	0.17 (0.13)	0.18 (0.11)
Weights, full (LMTC)	**no**	**0.11 (0.12)**	**0.09 (0.07)**	**0.10 (0.07)**
Question queries (reference)	yes	0.11 (0.12)	0.09 (0.07)	0.09 (0.06)

Table 6: **Key design ablation at** $B = 3$ on Qwen2.5-Omni-7B, with the final pre-rotary coder. Each row changes one design choice of LMTC. Text, Audio, Video: Key attention KL relative to per-channel INT2 with only that modality quantized; All: all three quantized together; Mass: attention mass moved between modalities.

	Variant	Text	Audio	Video	All	Mass (%)
	LMTC Keys	0.11	0.09	0.10	0.07	2.5
Modality	One shared basis	0.22	0.11	0.11	0.13	3.0
	One global mean	0.25	0.20	0.14	0.27	17.3
	Shared basis and mean	0.24	0.12	0.11	0.13	3.5
	Groups across modalities	0.12	0.10	0.11	0.08	3.1
Bits	Uniform 2 bits	0.20	0.16	0.17	0.13	3.4
	$b_j \in \{2, 3, 4\}$, no zero bits	0.14	0.12	0.13	0.09	2.7
	$b_j \in \{0, 2, 4\}$	0.12	0.10	0.11	0.08	2.6

single global mean moves 17.3% of the attention mass between modalities, so Insight (ii) holds for the pre-rotary coder as well as for the simplified coder of Section 2. Uniform bits in the same basis give 0.13 with all modalities quantized, against 0.07 for reverse water-filling, and forbidding zero-bit coefficients gives 0.09.

Rotary design. Table 7 varies the split between fast and slow coordinate pairs, the segment length, the offset prior, and the residual window W_0 that bounds it. Moving pairs from the slow to the fast block changes the KL by at most 0.01 around the default threshold, while longer segments reduce the number of query projections at a cost of 0.01 in KL. The log-uniform prior keeps the far-reader KL at 0.08, against 0.13 for a prior fixed at $\Delta = W_0$, and the result varies by at most 0.01 over $W_0 \in \{32, \ldots, 512\}$.

Long reasoning readers. The question tokens used above directly follow the prefix, whereas tokens generated during decoding read the cache from increasingly distant positions and with different content. We therefore also read the cache with the tokens the model generates. Only the committed prefix is quantized at $B \approx 3$; the prompt and the generated tokens stay BF16, and the attention KL of every generated token is averaged over 64-step windows. Besides INT2, KLT, and LMTC, we apply the Key quantizers of KIVI (per-channel INT2 in token order), OTT (32 lowest-magnitude tokens kept in BF16), OScaR (Hadamard rotation with a per-token norm), and VidKV (INT2 on the widest channels, 1-bit FFT features elsewhere). The symmetric per-token INT2 of QuaRot (Ashkboos et al., 2024) is omitted: on post-RoPE Keys its relative reconstruction error exceeds one. Figure 8 shows that LMTC stays below all other coders at every step. Its ratio stays between 0.07 and 0.08 in the multi-turn runs, and the order of the coders does not change.

Reader offset and streaming input. The pre-rotary coder needs no anchor position. When the readers are moved 0 to 256K positions away from the committed prefix, its Key attention KL stays at 0.09–0.11, 0.08–0.09, and 0.09–0.10 of INT2 for Text, Audio, and Video on Qwen2.5-Omni-7B, and at 0.08–0.10 and 0.08–0.09 for Audio and Video on OmniVinci-9B. Moving a whole request, Keys and readers alike, by 64K positions changes the KL by less than 0.2%.

Budget sweep. Table 8 gives the attention-level values behind Figure 5 (left) at each budget.

Table 7: **Sensitivity to the rotary design** on Qwen2.5-Omni-7B at $B = 3$, all modalities quantized. Key attention KL relative to per-channel INT2 for readers directly after the prefix (Near) and 256K positions away (Far); Cost: query projections per decode step relative to the default (marked *).

	Setting	Near	Far	Cost
Slow pairs	All pairs fast	0.16	0.16	0.00
	Threshold $\times\frac{1}{2}$	0.08	0.09	0.60
	Threshold $\times 1^*$	0.07	0.08	1.00
	Threshold $\times 2$	0.07	0.08	1.50
	All pairs slow	0.12	0.14	3.20
Segment	Length $\times\frac{1}{4}$	0.07	0.08	4.00
	Length $\times\frac{1}{2}$	0.07	0.08	2.00
	Length $\times 1^*$	0.07	0.08	1.00
	Length $\times 2$	0.08	0.09	0.50
Offset prior	Log-uniform on $[W_0, L]^*$	0.07	0.08	1.00
	Uniform on $[W_0, L]$	0.08	0.08	1.00
	Fixed $\Delta = W_0$	0.07	0.13	1.00
	Fixed $\Delta = L$	0.09	0.08	1.00
	No average ($\Delta = 0$)	0.07	0.19	1.00
W_0	32	0.07	0.08	1.00
	64	0.07	0.08	1.00
	128*	0.07	0.08	1.00
	512	0.08	0.08	1.00

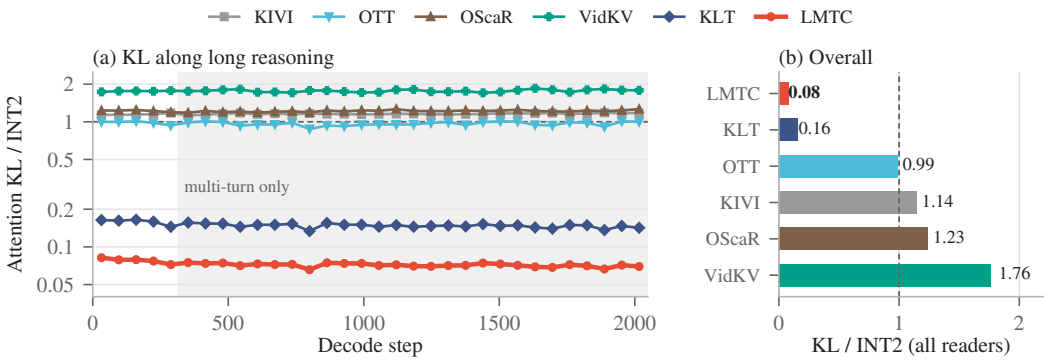


Figure 8: **Key attention KL along long reasoning** (Qwen2.5-Omni-7B, $B \approx 3$), relative to INT2; geometric mean over layers, KV heads, and runs. The shaded range is covered by the multi-turn runs only.

At every budget, LMTC reduces the output error of KIVI to 0.08–0.15, against 0.10–0.18 for KLT. At two bits it stays at 0.31 of the error of KIVI at three bits.

Table 8: **Budget sweep on Qwen2.5-Omni-7B** with all modalities quantized. B counts payload and BF16 metadata per scalar. Entries are the attention output error $\|\widehat{o}-o\|^2/\|o\|^2$ with Keys and Values quantized, relative to KIVI (per-channel INT2 Keys over $G = 32/64/128/256$ tokens, round-to-nearest Values) at the same budget, and in the last column relative to KIVI at $B = 3$. All methods use per-token INT2 Values over 32/64/128/128 channels (2.25 bits in the $B = 2$ row); KLT and LMTC code Keys with $G = 32$.

B	vs. KIVI		LMTC vs. KIVI
	KLT	LMTC	at B=3
3.0	0.18	**0.15**	0.15
2.5	0.14	**0.11**	0.21
2.25	0.13	**0.10**	0.29
2.0	0.10	**0.08**	0.31