A2Quant-QJL: Pairwise A_2 Lattice Quantization for Low-Variance Unbiased Attention Recovery

Abstract

KV-cache compression for autoregressive transformers can be factorized into three modular layers: a transform layer that regularizes the representation, a base quantizer layer that stores the compressed keys and values, and a residual correction layer that restores attention fidelity. This manuscript studies that decomposition in the specific setting of unbiased score recovery. The main method, A2Quant-QJL, keeps the transform layer modular, replaces the usual separable scalar base quantizer by a finite pairwise two-dimensional A_2-style lattice quantizer with 30 usable states per coordinate pair, and then applies residual QJL to preserve unbiased inner-product estimation. The key claim is deliberately conditional rather than universal: if post-mixing pairwise marginals are locally near-isotropic and saturation pressure is low, then pairwise lattice cells can reduce deterministic base residual energy relative to a matched-budget separable 32-state product baseline drawn from a restricted searched factorization family. Since the conditional variance of the residual-QJL estimator scales with the residual norm, lower base residual energy yields a direct mechanism for lower estimator variance.

The present paper is a theory-plus-methods preprint. The quantizer, baseline search, calibration rule, and Phase-1/Phase-2 evaluation harnesses are specified at implementation level, but no executed empirical results are claimed. The manuscript therefore contributes a sharpened novelty boundary, explicit notation, four worked derivations, code-grounded pseudocode, analytical sensitivity tables, and a disciplined comparison among three categories: the originating A2Quant design, compatible upgrades that preserve its novelty lane, and replacement-level alternatives that solve neighboring problems by different principles. The strongest conclusion is conservative: A2Quant remains the best-defended main method for this draft, while adaptive pairing is the most compelling compatible upgrade and transform-coding or codebook-vector-quantization families are the strongest replacement-level alternatives if future evidence fails to support the pairwise lattice premise.

1 Introduction

Long-context autoregressive inference is frequently limited not by parameter storage but by the key-value cache. For a model with L layers, H attention heads, head width d, context length T, and per-channel storage b bits, the cache memory scales as

$$M_{\mathrm{KV}} = \frac{2LHTdb}{8} \text{ bytes.} \tag{1}$$

The factor of two accounts for keys and values. As T grows, the KV cache can dominate both device memory and memory bandwidth.

The recent KV-compression literature has diversified quickly, but many methods can still be organized by a three-layer decomposition:

1. **Transform layer:** a mixer or preconditioner that makes channels easier to quantize.
2. **Base quantizer layer:** the primary discretization geometry that stores most of the compressed state.
3. **Residual correction layer:** a mechanism that restores attention fidelity, ideally with unbiasedness or a direct error certificate.

This decomposition is useful because different papers innovate at different layers, and the novelty boundaries otherwise blur. A2Quant-QJL deliberately occupies a narrow lane: it is a base-quantizer paper inside a residual-QJL pipeline.

Item	Statement in this manuscript
Main method	A2Quant base quantizer plus residual QJL
Scientific posture	Theory-plus-methods preprint
Executed evidence	None claimed in this draft
Claim strength	Local and conditional, not universal
Fairness baseline	Best searched separable 32-state product quantizer within the chosen factorization family, at matched 5 bits per pair
Transform stance	Modular; not the novelty lane

Table 1: Scope of the adopted main design.

raw q, k
$\downarrow$
transform layer: $q' = Sq$, $y = Sk$, with orthogonal S
examples: dense random orthogonal, Hadamard-style alignment,
structured pairwise rotations, blockwise $SO(4)$
$\downarrow$
base quantizer layer: A2Quant on RMS-normalized pairs of y
output: 5-bit pair codes plus base reconstruction $\hat{y}$
$\downarrow$
residual correction layer: QJL on $e = y - \hat{y}$
output: unbiased correction to $\langle q', e \rangle$
$\downarrow$
final score estimator: $\hat{s} = \langle q', \hat{y} \rangle + \widehat{\langle q', e \rangle}_{\mathrm{QJL}}$

Figure 1: Three-layer decomposition and where A2Quant intervenes.

Let $q, k \in \mathbb{R}^d$ denote one query-key pair inside a fixed layer-head. In column-vector notation,

$$q' = Sq, \qquad y = Sk. \tag{2}$$

Throughout this manuscript, the admissible mixer class is restricted to orthogonal transforms, so $S^\top S = I$ and $\langle q', y \rangle = \langle Sq, Sk \rangle = \langle q, k \rangle$. A more general invertible transform class would require the compatible query transform $q' = S^{-T} q$, which is outside the present scope. The quantizer constructs a base reconstruction $\hat{y}$, defines the residual

$$e = y - \hat{y}, \tag{3}$$

and estimates the attention score by the two-stage rule

$$\hat{s} = \langle q', \hat{y} \rangle + \widehat{\langle q', e \rangle}_{\mathrm{QJL}}. \tag{4}$$

Equation (4) is the core lens of the paper. A2Quant does not alter the residual estimator; it attempts to reduce $\|e\|_2^2$ before that estimator is invoked.

The originating design already proposed this pairwise A_2 lattice inside a residual-QJL pipeline. The revised main formulation adopted here keeps the same algorithmic core but sharpens four things that matter for paper defensibility: the admissible regime is made explicit through diagnostics, the transform class is restricted to orthogonal mixers, the baseline is fixed to a fair matched-budget separable product search family, and stronger alternatives are separated into compatible upgrades versus replacement-level redesigns.

Category	Role in this draft	What changes
Originating design	Main algorithmic seed	Pairwise A_2 base lattice plus residual QJL
Adopted main design	Main method in this paper	Same algorithm, but with diagnostic-gated theory and fairness-locked baselines
Compatible upgrades	Preserved as extensions	Adaptive pairing; calibration-free alignment before A2Quant
Replacement-level alternatives	Discussed, not adopted	Polar-angle base quantization; codebook VQ; transform coding; low-rank attention compression

Table 2: Originating design, revised paper formulation, and nearby alternatives.

Three contributions follow. First, the paper isolates *base quantizer geometry* as an under-separated design variable in unbiased KV-cache recovery. Second, it turns the existing project materials into a single paper-shaped theory-plus-methods manuscript with explicit formulas, derivations, and algorithms. Third, it searches beyond the original design and concludes that the strongest nearby improvement is still conservative: keep A2Quant as the main method, add adaptive pairing if needed, and treat more radical families as explicit alternatives rather than silently replacing the method.

Contribution	Status
A2Quant finite pairwise lattice base quantizer	Specified and code-grounded
Residual-QJL unbiased estimator in the A2Quant pipeline	Specified and code-grounded
$\Delta^\star$ calibration rule	Specified and code-grounded
Best searched separable 32-state product baseline search	Specified and code-grounded
Real-activation diagnostics and experimental phases	Planned only

Table 3: Contributions and current status.

2 Related Work and Novelty Boundary

2.1 Transform-layer precedents

Modern KV-cache compression rarely quantizes raw channels directly. Instead, a transform layer reshapes the distribution before discretization. Dense randomized orthogonal mixing is the canonical example. Structured alternatives now push toward hardware-aligned local mixing, calibration-free alignment, or sparse pairwise rotations. This matters because A2Quant is intentionally not a transform paper. Its claim starts only after a transform has produced pairwise marginals that are regular enough for local geometric arguments to be meaningful.

2.2 Base-quantizer precedents

The base layer is more crowded than it first appears. Separable scalar quantization remains the most common practical default. Polar-coordinate schemes show that non-Cartesian geometry can help after preconditioning. Recent codebook-based methods pursue higher-dimensional vector quantization directly. Coupled or commutative quantizers exploit inter-channel dependence in a more global way than A2Quant does. A2Quant differs from these in two ways. It is intentionally local, operating on independent two-dimensional pairs rather than large learned codebooks, and it is embedded in a residual-QJL estimator where the downstream effect of base residual norm can be analyzed cleanly.

2.3 Residual, exact, and fidelity-oriented alternatives

Residual-QJL gives A2Quant its clean estimator story. But unbiased recovery is not the only way to preserve quality. Some methods target exact reparameterization, some compress the attention matrix more directly, and some move from quantization to transform coding or low-rank approximation. These are legitimate neighboring projects, but they occupy different novelty lanes.

Method family	Transform layer	Base layer	Residual / fidelity layer	Relation to A2Quant
TurboQuant	Dense random orthogonal mixing	Separable scalar quantization	Residual QJL	Defines the two-stage unbiased pipeline that A2Quant keeps
PolarQuant	Random preconditioning	Polar-angle or recursive geometry	Optional residual handling	Closest base-geometry alternative
KVQuant	Practical quantization pipeline with system-aware design	Scalar low-bit quantization	No unbiased residual lane	Strong systems baseline with a different goal
NSNQuant	Double-normalization plus Hadamard alignment	Low-bit VQ	No residual-QJL lane	Compatible transform inspiration
CQ / CommVQ / VQKV	May include learned or commutative transforms	Coupled or vector-codebook quantization	Reconstruction-based	Replacement-level vector-quantization alternatives
SpinQuant / ParoQuant / IsoQuant	Learned or structured rotations	Usually separable scalar base	Reconstruction-based	Transform-side alternatives
KQ-SVD / low-rank fidelity methods	Sometimes head reordering or online adaptation	Low-rank or factorized compression	Attention-fidelity focus	Replacement-level approximation family
GaugeKV / KVTC / FreqKV	Exact reparameterization or transform coding	Not A2Quant-style	Exactness or transform-coded fidelity	Different problem formulation

Table 4: Layer-wise literature map.

KV compression design space		
Transform-heavy	Base-geometry-heavy	Fidelity / exactness
SpinQuant	PolarQuant	KQ-SVD
ParoQuant	CQ / CommVQ / VQKV	GaugeKV
IsoQuant	A2Quant	KVTC / FreqKV
NSNQuant		

Figure 2: Novelty boundary in the three-layer space. A2Quant occupies the middle column and borrows only the residual-QJL lane from TurboQuant/QJL.

The novelty boundary can therefore be stated precisely: A2Quant does not claim a new transform, a new unbiased estimator, or a new global rate-distortion optimum. It claims that, inside an existing unbiased recovery architecture, the base geometry itself is a meaningful and under-isolated design variable.

This manuscript claims	This manuscript does not claim
Base geometry matters inside residual-QJL pipelines	A2Quant universally beats all scalar quantizers
Pairwise A_2 geometry is promising under diagnostics A1–A5	The transform layer is solved
Lower base residual norm lowers the QJL variance bound	Real-model gains have already been demonstrated
Adaptive pairing is the strongest compatible upgrade	Codebook VQ or low-rank methods are invalid

Table 5: Claims and explicit non-claims.

3 Method

3.1 Notation and symbol ledger

Symbol	Meaning	Shape / units
d	Head dimension	scalar
S	Orthogonal transform or mixer	$\mathbb{R}^{d\times d}$
q, k	Query and key vectors	$\mathbb{R}^d$
q', y	Transformed query and key	$\mathbb{R}^d$
s	Per-vector RMS scale	scalar
z_j	Normalized coordinate pair j	$\mathbb{R}^2$
Δ	A2Quant spacing	scalar
$\hat{y}$	Base reconstruction of y	$\mathbb{R}^d$
e	Base residual $y - \hat{y}$	$\mathbb{R}^d$
γ	Residual norm $\|e\|_2$	scalar
u	Normalized residual direction	$\mathbb{R}^d$
G	Gaussian QJL sketch matrix	$\mathbb{R}^{d\times d}$
c_j	5-bit pair code	integer in $\{0, \ldots, 29\}$

Table 6: Symbol ledger.

The method factorizes into a transform, a base quantizer, and a residual estimator.

$$q' = Sq, \qquad y = Sk. \tag{5}$$

Here again S is restricted to be orthogonal, which keeps the transformed score $\langle q', y\rangle$ equal to the original score $\langle q, k\rangle$. The base layer begins with per-vector RMS normalization,

$$s = \sqrt{\frac{1}{d}\sum_{i=1}^{d} y_i^2}, \tag{6}$$

$$\tilde{y} = \frac{y}{s}, \tag{7}$$

and pair formation,

$$z_j = (\tilde{y}_{2j-1}, \tilde{y}_{2j}), \qquad j = 1, \ldots, \frac{d}{2}. \tag{8}$$

3.2 A2Quant base geometry

A2Quant uses two interleaved cosets of a scaled A_2-style lattice, truncated to five horizontal positions and three vertical positions per coset. For coset 0,

$$p_0(a, b) = (a\Delta, b\sqrt{3}\,\Delta), \tag{9}$$

with $a \in \{-2, -1, 0, 1, 2\}$ and $b \in \{-1, 0, 1\}$. For coset 1,

$$p_1(a, b) = \left(\left(a + \frac{1}{2}\right)\Delta, \left(b + \frac{1}{2}\right)\sqrt{3}\,\Delta\right). \tag{10}$$

Each coset contributes $5 \times 3 = 15$ points, so the usable set contains 30 code points per pair.

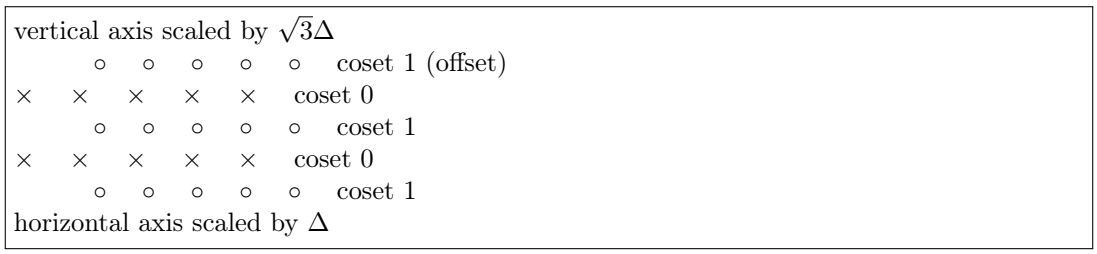

Figure 3: Schematic of the two-coset A2Quant geometry.

For a given pair $z = (z_1, z_2)$, the reference implementation computes one candidate on each coset and chooses the nearer one. The axis-aligned candidate is given by

$$a_0 = \mathrm{clip}\Big(\mathrm{round}\Big(\frac{z_1}{\Delta}\Big), -2, 2\Big), \tag{11}$$

$$b_0 = \mathrm{clip}\left(\mathrm{round}\left(\frac{z_2}{\sqrt{3}\,\Delta}\right), -1, 1\right), \tag{12}$$

$$d_0 = (z_1 - a_0\Delta)^2 + (z_2 - b_0\sqrt{3}\,\Delta)^2. \tag{13}$$

For the offset coset,

$$a_1 = \mathrm{clip}\left(\mathrm{round}\left(\frac{z_1}{\Delta} - \frac{1}{2}\right), -2, 2\right), \tag{14}$$

$$b_1 = \mathrm{clip}\left(\mathrm{round}\left(\frac{z_2}{\sqrt{3}\,\Delta} - \frac{1}{2}\right), -1, 1\right), \tag{15}$$

$$d_1 = \left(z_1 - \left(a_1 + \frac{1}{2}\right)\Delta\right)^2 + \left(z_2 - \left(b_1 + \frac{1}{2}\right)\sqrt{3}\,\Delta\right)^2. \tag{16}$$

The chosen triple is

$$(a, b, c) = \begin{cases} (a_0, b_0, 0), & d_0 \le d_1, \\ (a_1, b_1, 1), & d_1 < d_0. \end{cases} \tag{17}$$

The packed pair code is

$$c_j = c + 2\big((a + 2) + 5(b + 1)\big). \tag{18}$$

Decoding uses a lookup table of unscaled two-dimensional reference points:

$$\hat{z}_j = \Delta\,\mathrm{LUT}[c_j], \tag{19}$$

$$\hat{y} = s\,\mathrm{flatten}(\hat{z}_1, \ldots, \hat{z}_{d/2}). \tag{20}$$

Component	Value
States per pair	30
Bits per pair index	5
Pair layout	Adjacent pairs in the current implementation
Geometry	Two-coset finite A_2-style lattice
Decode path	Lookup-table decode followed by rescaling
Arithmetic cost	$\mathcal{O}(d)$ per vector

Table 7: A2Quant summary.

Worked numeric example 1. Take $\Delta = 0.5$ and $z = (0.74, 0.55)$. Then $\sqrt{3}\,\Delta \approx 0.866$. On coset 0, one gets $a_0 = 1$, $b_0 = 1$, and

$$d_0 = (0.74 - 0.50)^2 + (0.55 - 0.866)^2 \approx 0.1575. \tag{21}$$

On coset 1, one gets $a_1 = 1$, $b_1 = 0$, and

$$d_1 = (0.74 - 0.75)^2 + (0.55 - 0.433)^2 \approx 0.0138. \tag{22}$$

Hence coset 1 is chosen with $(a, b, c) = (1, 0, 1)$, the packed code is

$$c_j = 1 + 2[(1 + 2) + 5(0 + 1)] = 17, \tag{23}$$

and the decoded normalized pair is approximately

$$\hat{z} = (0.75, 0.433). \tag{24}$$

3.3 Residual-QJL layer

After base reconstruction, define

$$e = y - \hat{y}, \tag{25}$$

$$\gamma = \|e\|_2 . \tag{26}$$

If $\gamma = 0$, the base path is exact. Otherwise define the unit residual direction

$$u = \frac{e}{\gamma}. \tag{27}$$

The current reference harness uses a dense Gaussian sketch matrix $G \in \mathbb{R}^{d \times d}$ and a 1-bit sign sketch

$$z = \mathrm{sign}(Gu) \in \{-1, +1\}^d. \tag{28}$$

The sketch matrix G is shared public randomness fixed by the protocol rather than stored per vector, so it does not enter the dynamic bit accounting. The dequantized residual direction estimator is

$$\hat{u} = \frac{\sqrt{\pi/2}}{d} G^\top z. \tag{29}$$

The residual score estimate is then

$$\hat{r} = \gamma \left\langle q', \hat{u} \right\rangle, \tag{30}$$

and the full estimator becomes

$$\hat{s} = \left\langle q', \hat{y} \right\rangle + \gamma \left\langle q', \hat{u} \right\rangle. \tag{31}$$

This residual-QJL construction preserves the existing unbiasedness mechanism. A2Quant changes only the base path that produces $\hat{y}$ and hence the residual magnitude γ.

3.4 Effective bits per channel

Under the flagship configuration used in the project materials, base indices cost $(d/2) \cdot 5 = 5d/2$ bits, residual signs cost d bits, and metadata costs B_{meta} bits per vector. Therefore,

$$B_{\text{total}}(d) = \frac{5d}{2} + d + B_{\text{meta}} = 3.5d + B_{\text{meta}}, \tag{32}$$

$$\text{EBC}(d) = \frac{B_{\text{total}}(d)}{d} = 3.5 + \frac{B_{\text{meta}}}{d}. \tag{33}$$

Here EBC refers to dynamic per-vector payload storage only. It excludes transform storage, shared protocol randomness, and end-to-end systems costs such as kernel overhead or mixer application cost. With fp16 storage for the RMS scale s and residual norm γ,

$$B_{\text{meta}} = 16 + 16 = 32 \text{ bits}, \tag{34}$$

so that

$$\text{EBC}(128) = 3.5 + \frac{32}{128} = 3.75 \text{ bits/channel}. \tag{35}$$

d	$B_{\text{meta}} = 16$	$B_{\text{meta}} = 32$	$B_{\text{meta}} = 48$
64	3.75	4.00	4.25
128	3.625	3.75	3.875
256	3.5625	3.625	3.6875
512	3.53125	3.5625	3.59375

Table 8: Analytical EBC sensitivity over head width and metadata size, for payload storage only.

Worked numeric example 2. For $d = 128$, base indices consume $64 \times 5 = 320$ bits, residual signs consume 128 bits, and metadata consumes 32 bits. Hence

$$B_{\text{total}}(128) = 320 + 128 + 32 = 480 \text{ bits}, \tag{36}$$

which yields $480/128 = 3.75$ bits per channel.

3.5 Calibration and matched baselines

The spacing parameter Δ is calibrated by line search over a calibration batch $Y_{\text{calib}} \in \mathbb{R}^{N \times d}$. Let s_i be the per-vector RMS scales. For a candidate Δ, the reconstructed batch is $\hat{Y}(\Delta)$. The reference rule chooses

$$\Delta^\star = \arg\min_{\Delta} \frac{\left\| Y_{\text{calib}} - \hat{Y}(\Delta) \right\|_F^2}{\|Y_{\text{calib}}\|_F^2}. \tag{37}$$

The fairness baseline is the best searched separable 32-state product quantizer within the chosen factorization family and under the same base-code budget. The implementation searches the factorization set

$$(n_1, n_2) \in \{(1, 32), (2, 16), (4, 8)\}, \tag{38}$$

learns 1D Lloyd–Max centroids for each factor on pooled RMS-normalized coordinates, and minimizes

$$\varepsilon_{\text{base}}(Q) = \frac{1}{N}\sum_{i=1}^{N}\frac{\|y_i - \hat{y}_i(Q)\|_2^2}{\|y_i\|_2^2}. \tag{39}$$

Baseline	States per pair	Bits per pair	Learned from calibration?	Role
A2Quant	30	5	$\Delta^\star$ only	Main method
Best searched separable product	32	5	Yes, via 1D Lloyd–Max	Fair scalar opponent within the searched family
Fixed hardware grid	32	5	No	Kernel-friendly proxy

Table 9: Baseline fairness design.

4 Theory and Analytical Mechanism

4.1 Diagnostics for the admissible regime

The A2Quant claim is local and conditional. For a fixed layer-head, let z_j denote normalized coordinate pairs after mixing. The intended operating regime is described by five diagnostics. Covariance anisotropy is measured by

$$\kappa_j = \frac{\lambda_{\max}(\Sigma_j)}{\lambda_{\min}(\Sigma_j)}. \tag{40}$$

Whitened angular non-uniformity is measured schematically by

$$V_j = \text{Kuiper}(\{\phi_n^{(j)}\}, \text{Uniform}[0, 2\pi)). \tag{41}$$

Radial-tail mismatch relative to a reference isotropic radial law is

$$T_j = \sup_{\tau}\left|\Pr(r_j > \tau) - \Pr_{\text{ref}}(r > \tau)\right|. \tag{42}$$

Edge-hit and clamp-hit rates are summarized by

$$\beta_{\text{edge},j} = \Pr\big((a,b) \text{ lands on the truncated boundary set}\big), \tag{43}$$

$$\beta_{\text{clamp},j} = \Pr\big(\text{pre-clipped candidate index lies outside the interior range}\big). \tag{44}$$

Pair-position heterogeneity can be summarized as the average normalized discrepancy over all $\binom{d/2}{2} = d(d-2)/8$ unordered pair-pairs:

$$U = \frac{8}{d(d-2)}\sum_{j<j'}\frac{\left\|\Sigma_j - \Sigma_{j'}\right\|_F}{\left\|\bar{\Sigma}\right\|_F}. \tag{45}$$

The diagnostics are screening variables rather than hard thresholds. Intuitively, A2Quant should help only when $\kappa_j \approx 1$, $V_j \approx 0$, T_j is small, saturation rates are low, and heterogeneity across pairs is limited.

Diagnostic regime	What it means	Expected failure if violated
$\kappa \approx 1$, $V \approx 0$	Local pair distribution is near-isotropic	Separable scalar cells may be equally good or better
Small T	Radial tails are tame	Truncation and clipping dominate
Small β_{edge}, β_{clamp}	Quantizer operates in the interior	Finite-lattice boundary effects dominate
Small U	One pairing pattern and one $\Delta^\star$ are meaningful	Adaptive pairing or zoned calibration may be needed

Table 10: Assumption-to-failure map.

4.2 Worked derivation 1: local geometry proxy

Let Q be any pairwise quantizer with code points p_c and cells C_c. The exact pairwise distortion is

$$\mathbb{E}\ \|z - Q(z)\|_2^2 = \sum_c \int_{C_c} \|x - p_c\|_2^2 \, f(x)\, dx. \tag{46}$$

If the density is slowly varying over each cell, a first-order high-resolution proxy gives

$$\mathbb{E}\ \|z - Q(z)\|_2^2 \approx \sum_c f(p_c) \int_{C_c} \|x - p_c\|_2^2 \, dx. \tag{47}$$

If, after whitening, the local density is approximately isotropic and nearly constant over neighboring cells, the leading term reduces to the cell second moment,

$$J(C_c) = \frac{1}{\text{area}(C_c)} \int_{C_c} \|x - p_c\|_2^2 \, dx. \tag{48}$$

Thus the base-geometry comparison becomes a comparison among cell shapes under equal area. This does not prove that the finite, truncated A2Quant globally dominates separable rectangles. It only identifies the mechanism by which an isotropic pairwise geometry can help in the local interior regime.

The next statement should be read as a heuristic design claim motivated by the high-resolution proxy above, not as a formal theorem about the finite truncated 30-point codebook.

Proposition 1 (Heuristic local geometry claim)**.** *If post-mixing pairwise marginals are locally near-isotropic and interior-cell operation dominates boundary hits, then A2Quant is expected to reduce base residual energy relative to a matched searched separable 32-state product quantizer operating on the same RMS-normalized pair distribution.*

4.3 Worked derivation 2: QJL dequantization constant

Let $g \sim \mathcal{N}(0, I_d)$ and $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$. By rotational symmetry one may take $u = e_1$. Then

$$\mathbb{E}[\text{sign}(g^\top u) g] = \mathbb{E}[\text{sign}(g_1) g], \tag{49}$$

$$\mathbb{E}[\operatorname{sign}(g_1)g] = (\mathbb{E}|g_1|, 0, \ldots, 0), \tag{50}$$

$$(\mathbb{E}|g_1|, 0, \ldots, 0) = \sqrt{\frac{2}{\pi}}\, u. \tag{51}$$

If G has i.i.d. Gaussian rows, then

$$\mathbb{E}[G^\top \operatorname{sign}(Gu)] = d\sqrt{\frac{2}{\pi}}\, u. \tag{52}$$

Therefore the inverse scaling used in the dequantized residual direction estimator satisfies

$$\mathbb{E}\left[\frac{\sqrt{\pi/2}}{d} G^\top \operatorname{sign}(Gu)\right] = u. \tag{53}$$

This identity is the reason the residual direction estimator in the code is unbiased.

4.4 Worked derivation 3: unbiasedness of the full score estimator

Condition on fixed $q, k, S, \hat{y}, e$, and write $q' = Sq$, $e = \gamma u$ with $\gamma = \|e\|_2$. If $\gamma = 0$, then $e = 0$ and $\hat{s} = \langle q', \hat{y}\rangle = \langle q', y\rangle$. For $\gamma > 0$, substitute the dequantized residual direction estimator into the residual score expression:

$$\mathbb{E}_G[\hat{s}] = \left\langle q', \hat{y}\right\rangle + \gamma\, \mathbb{E}_G\left\langle q', \hat{u}\right\rangle, \tag{54}$$

$$\mathbb{E}_G[\hat{s}] = \left\langle q', \hat{y}\right\rangle + \gamma\, \left\langle q', \mathbb{E}_G[\hat{u}]\right\rangle, \tag{55}$$

$$\mathbb{E}_G[\hat{s}] = \left\langle q', \hat{y}\right\rangle + \gamma\, \left\langle q', u\right\rangle, \tag{56}$$

$$\mathbb{E}_G[\hat{s}] = \left\langle q', \hat{y} + e\right\rangle, \tag{57}$$

$$\mathbb{E}_G[\hat{s}] = \left\langle q', y\right\rangle = \langle q, k\rangle. \tag{58}$$

Hence the estimator is conditionally unbiased.

4.5 Worked derivation 4: conditional variance and normalized variance

Define the residual-only estimator

$$Z = \gamma\, \left\langle q', \hat{u}\right\rangle. \tag{59}$$

Using the standard QJL variance scaling for the dequantized sign sketch,

$$\operatorname{Var}_G\big(\left\langle q', \hat{u}\right\rangle\big) \le \frac{\pi}{2d} \left\|q'\right\|_2^2. \tag{60}$$

Multiplying by γ^2 gives

$$\operatorname{Var}_G(Z) \le \frac{\pi}{2d} \left\|q'\right\|_2^2 \gamma^2, \tag{61}$$

$$\operatorname{Var}_G(Z) \le \frac{\pi}{2d} \left\|q'\right\|_2^2 \|e\|_2^2. \tag{62}$$

Because the full estimator is conditionally unbiased,

$$\mathbb{E}_G\big[(\hat{s} - \langle q, k\rangle)^2\big] = \operatorname{Var}_G(\hat{s}) = \operatorname{Var}_G(Z). \tag{63}$$

This is the paper's central mechanism: smaller $\|e\|_2^2$ implies a smaller conditional MSE upper bound. The implementation also reports a normalized variance diagnostic,

$$\mathrm{NV} = \frac{2d}{\pi} \frac{\mathrm{Var}_G(\hat{s})}{\|q'\|_2^2}. \tag{64}$$

Substituting the conditional variance bound into the normalized-variance definition gives the analytical comparator

$$\mathrm{NV} \leq \|e\|_2^2 . \tag{65}$$

This is exactly what the Phase-2 harness tests numerically.

Quantity	Definition	Intended interpretation
$\|e\|_2^2$	Base residual energy	Control knob for the residual estimator
$\mathrm{Var}_G(\hat{s})$	Conditional score-estimation variance	Should fall when base residual shrinks
NV	Normalized variance diagnostic	Should not materially exceed $\|e\|_2^2$ in expectation

Table 11: Theory-to-metric consistency check.

Worked numeric example 3. Let $d = 128$, $\|q'\|_2 = 10$, and $\|e\|_2 = 2$. Then the conditional variance bound above yields

$$\mathrm{Var}_G(\hat{s}) \leq \frac{\pi}{256} \cdot 100 \cdot 4 = \frac{400\pi}{256} \approx 4.91. \tag{66}$$

The corresponding normalized-variance bound is simply

$$\mathrm{NV} \leq \|e\|_2^2 = 4. \tag{67}$$

d	$\|e\|_2 = 1$	$\|e\|_2 = 2$	$\|e\|_2 = 3$
64	2.45	9.82	22.09
128	1.23	4.91	11.05
256	0.61	2.45	5.52

Table 12: Analytical variance-bound sensitivity for $\|q'\|_2 = 10$. Values are derived from the conditional variance bound above, not from executed experiments.

4.6 Dimensional checks

Expression	Units check
$q' = Sq$, $y = Sk$	Same units as the original channels if S is dimensionless
s, Δ, γ	Channel-amplitude scale
c_j	Dimensionless index
$\langle q', \hat{y} \rangle$ and $\gamma \langle q', \hat{u} \rangle$	Dot-product units and therefore directly summable
EBC	Bits per stored channel

Table 13: Unit and consistency checks.

5 Implementation and Algorithms

The paper is backed by a compact reference implementation. Table 14 maps the core quantizer, calibration, baseline-search, evaluation, and data-contract objects to their roles in the manuscript. The implementation still uses the historical class name `PairLattice30Quantizer` for the concrete 30-state pairwise lattice object. Throughout the manuscript, that code-level object is referred to at the method level as A2Quant, so Table 14 preserves the implementation name only where a direct source-to-equation mapping is useful.

Implementation object	Role in the manuscript
`PairLattice30Quantizer.encode`	Equations (11)–(18)
`PairLattice30Quantizer.decode`	Equations (19)–(20)
`generate_reference_lut`	Fixed unscaled 2D codebook
`calibrate_lattice_spacing`	Equation (37)
`find_best_separable_baseline`	Equation (39) baseline search
`evaluate_phase1_base_geometry`	Normalized base-error comparison
`evaluate_phase2_conditional_variance`	Equations (64)–(65) harness
`RealActivationLoader`	Future extraction of real activations

Table 14: Code-grounded object map.

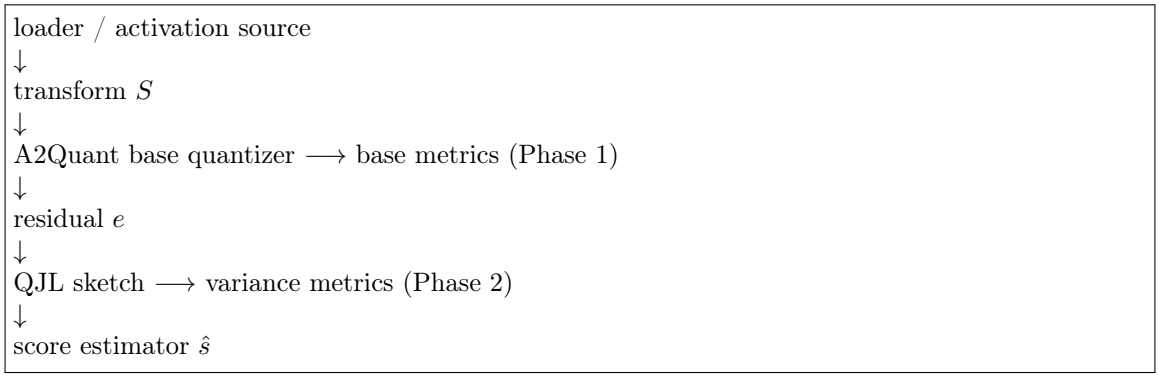

Figure 4: Dependency structure of the implementation.

5.1 Algorithm 1: A2Quant encode/decode

1. Input: $y \in \mathbb{R}^d$, spacing Δ, reference lookup table.
2. Compute the RMS scale $s \leftarrow \sqrt{\text{mean}(y \odot y)}$.
3. Form $\tilde{y} \leftarrow y/s$ and reshape it into pairs $z_j = (\tilde{y}_{2j-1}, \tilde{y}_{2j})$.
4. For each pair $z = (z_1, z_2)$, compute (a_0, b_0, d_0) and (a_1, b_1, d_1) as in Equations (11)–(16).
5. Choose the lower-distance candidate, set (a, b, c) accordingly, and pack the index via Equation (18).
6. Decode each pair by lookup and scaling: $\hat{z}_j = \Delta \cdot \text{LUT}[c_j]$.
7. Reconstruct the vector $\hat{y} = s\,\text{flatten}(\hat{z}_1, \ldots, \hat{z}_{d/2})$.

8. Output the pair codes, the scale s, and the reconstruction $\hat{y}$.

5.2 Algorithm 2: $\Delta^\star$ calibration

1. Input: calibration batch $Y_{\text{calib}} \in \mathbb{R}^{N \times d}$, lookup table, candidate spacing grid $\{\Delta_t\}$.
2. For each vector i, compute the RMS scale s_i.
3. For each candidate Δ_t, reconstruct $\hat{Y}(\Delta_t)$ using Algorithm 1 over the full batch.
4. Compute the normalized Frobenius error $\left\| Y_{\text{calib}} - \hat{Y}(\Delta_t) \right\|_F^2 / \left\| Y_{\text{calib}} \right\|_F^2$.
5. Return the minimizer $\Delta^\star$.

Worked numeric example 4. If $\Delta = 0.40$ gives normalized error 0.022 and $\Delta = 0.55$ gives normalized error 0.019, then the calibration objective selects $\Delta^\star = 0.55$. This is an illustration of the rule, not an executed experiment.

5.3 Algorithm 3: best searched separable baseline search

1. Input: calibration batch $Y \in \mathbb{R}^{N \times d}$.
2. Compute the per-vector scales $s_i = \sqrt{\text{mean}(Y_i \odot Y_i)}$.
3. Form the pooled one-dimensional sample vector by flattening Y/s.
4. Search the candidate layouts $(1, 32)$, $(2, 16)$, and $(4, 8)$.
5. For each layout, learn 1D Lloyd–Max centroids on the pooled data.
6. Quantize each coordinate pair independently to the nearest product centroid.
7. Compute the mean normalized base error over the searched family.
8. Return the lowest-error candidate within that searched family.

Layout	1D centroid counts	Total pair states	What it emphasizes
1×32	one trivial axis, one fine axis	32	Strong anisotropy along one coordinate
2×16	modest asymmetry	32	Moderate anisotropy
4×8	more balanced product grid	32	Near-symmetric searched product baseline

Table 15: Sensitivity of the matched searched separable baseline family.

5.4 Algorithm 4: Phase-2 conditional variance harness

1. Input: $q, k \in \mathbb{R}^d$, orthogonal mixer S, A2Quant quantizer, lookup table, and trials T.
2. Compute $q' \leftarrow qS^\top$ and $y \leftarrow kS^\top$ in the row-vector implementation convention; this is the row-vector equivalent of the column-vector convention $q' = Sq$, $y = Sk$ under orthogonal S.
3. Run A2Quant base quantization on y to obtain $\hat{y}$ and the residual $e = y - \hat{y}$.
4. Set $\gamma \leftarrow \|e\|_2$. If $\gamma = 0$, return zeros.
5. Form $u \leftarrow e/\gamma$ and base $\leftarrow \langle q', \hat{y} \rangle$.
6. For each trial $t = 1, \ldots, T$, draw $G_t \sim \mathcal{N}(0, I)$ elementwise, compute $z_t = \operatorname{sign}(G_t u)$, decode $\hat{u}_t = (\sqrt{\pi/2}/d) G_t^\top z_t$, and set $s_t = \text{base} + \gamma \langle q', \hat{u}_t \rangle$.
7. Compute the empirical variance of $\{s_t\}_{t=1}^T$.
8. Return $\text{NV} = (2d/\pi) \operatorname{Var}(\{s_t\}) / \|q'\|_2^2$, the bound $\|e\|_2^2$, and the residual norm γ.

Operation	Complexity in the current implementation	Comment
A2Quant encode/decode	$\mathcal{O}(d)$	Pairwise vectorized arithmetic only
$\Delta^\star$ search	$\mathcal{O}(\text{steps} \cdot Nd)$	Calibration time only
Lloyd–Max baseline search	Small $\mathcal{O}(\text{iters} \cdot Nd \cdot n)$	Candidate set is tiny
Phase-2 variance harness	$\mathcal{O}(Td^2)$	Dominated by dense Gaussian sketches
Loader contracts	Streaming or I/O bound	Model execution not yet implemented

Table 16: Complexity notes.

6 Planned Evaluation

No executed empirical claims appear in this manuscript. The evaluation section therefore defines what evidence would be required to move from theory-plus-methods to an evidence-backed systems paper.

6.1 Phased evaluation design

Phase 0 measures whether the diagnostic premise holds on real activations. Phase 1 tests the base-geometry claim directly. Phase 2 tests whether residual variance tracks the predicted bound. Phase 4 asks whether the main observation survives across transform families.

Phase	Goal	Output metrics
Phase 0	Validate admissible regime	$\kappa, V, T, \beta_{\text{edge}}, \beta_{\text{clamp}}, U$
Phase 1	Compare base geometries	Normalized reconstruction error and advantage over the searched separable 32-state product family
Phase 2	Test residual-QJL mechanism	NV, $\|e\|_2$, bound slack
Phase 4	Test modularity across mixers	Phase 0–2 metrics plus encode/decode latency

Table 17: Planned phases and outputs.

Phase 0: diagnostics on real activations
↓
Phase 1: A2Quant versus the searched separable 32-state product family
↓
Phase 2: conditional variance harness
↓
Phase 4: factorial test across transform families

Figure 5: Evaluation flow.

6.2 Metrics

The primary Phase-1 metric is the expected normalized base error,

$$\varepsilon_{\text{base}} = \mathbb{E}\left[\frac{\|y - \hat{y}\|_2^2}{\|y\|_2^2}\right]. \tag{68}$$

The primary Phase-2 metric is the normalized variance diagnostic NV defined in the theory section. For modularity, the paper proposes a factorial study with three transform families and two base quantizers.

Transform family	A2Quant base	Searched separable 32-state product base
Dense random orthogonal	✓	✓
Blockwise hardware-aligned local mixing	✓	✓
Pairwise rotation-and-scale transform	✓	✓

Table 18: Planned factorial design.

6.3 Minimum evidence threshold

Stronger claim one might want	Minimum evidence required
A2Quant reduces real residual energy	Phase 1 on real activations
A2Quant lowers conditional residual variance	Phase 2 on real activations
A2Quant is modular across mixers	Phase 4 factorial study
A2Quant improves long-context model quality	End-to-end downstream evaluation

Table 19: Evidence thresholds for stronger claims.

The paper therefore stops at theory, implementation, and experimental design. Anything stronger would require new evidence not yet present here.

7 Limitations, Compatible Upgrades, and Replacement Alternatives

7.1 Limitations of the retained main design

A2Quant is finite and truncated. If the transform leaves strong outliers or pairwise anisotropy, then boundary effects can dominate any interior-shape advantage. Likewise, if adjacent coordinate pairing is poorly aligned with the actual dependence structure, the base geometry is solving the wrong local problem.

Failure mode	Symptom	Likely cause	Mitigation lane
High clamp-hit rate	Large base-error spikes	Heavy tails after mixing	Stronger transform or hybrid outlier path
Strong anisotropy	Little or no advantage over the searched separable 32-state product family	Poor mixing	Transform redesign
Large pair heterogeneity	Unstable $\Delta^\star$ or inconsistent benefits	One pairing does not fit all positions	Adaptive pairing or zoned calibration
Latency loss	Better geometry but slower kernels	Lookup or gather overhead	Hardware-aware implementation study
Downstream quality loss despite good Phase 1	Benchmark fragility	Reasoning-specific compression pitfalls	End-to-end evaluation

Table 20: Failure modes of the retained main design.

7.2 Compatible upgrade 1: adaptive pairing

The strongest compatible upgrade keeps the A2Quant novelty lane intact while making its premise more realistic. Instead of fixed adjacent pairs $(1,2),(3,4),\ldots$, choose pairs by calibration-time matching. Let R be a channel-similarity matrix, for example using absolute correlation,

$$w_{ab} = \left|\mathrm{corr}(\tilde{y}_a, \tilde{y}_b)\right|. \tag{69}$$

Choose a disjoint pairing $\mathcal{M}$ by maximum-weight matching,

$$\mathcal{M}^\star = \arg\max_{\mathcal{M}} \sum_{(a,b)\in\mathcal{M}} w_{ab}. \tag{70}$$

Then apply A2Quant to the permuted pairs induced by $\mathcal{M}^\star$. Why this is attractive is simple: if A2Quant benefits from locally near-isotropic pairwise marginals, then pair selection should be part of the method whenever pair heterogeneity is nontrivial.

Aspect	Benefit	Cost
Geometry	Makes pairwise isotropy more plausible	Requires offline statistics
Novelty lane	Preserved at the base layer	Introduces permutation metadata
Systems impact	Still pairwise and local	Gather/scatter overhead must be measured
Risk	Low conceptual risk	Depends on stability of the learned pairing

Table 21: Adaptive pairing as a compatible upgrade.

7.3 Compatible upgrade 2: calibration-free alignment before A2Quant

Another upgrade is to make the transform layer less arbitrary. A calibration-free alignment stack can be inserted before A2Quant, using per-vector normalization and a structured orthogonal transform such as a Hadamard-like operator. In abstract form,

$$\tilde{y}_{\text{align}} = HD_2 \,\text{normalize}(HD_1 k), \tag{71}$$

where H is a fast orthogonal transform and D_1, D_2 are sign or scale operators. This retains A2Quant at the base layer while reducing reliance on a single calibration-only $\Delta^\star$.

7.4 Replacement-level alternative 1: polar-angle base quantization

A direct replacement is a PolarQuant-style base layer. For each pair,

$$r = \sqrt{z_1^2 + z_2^2}, \qquad \theta = \text{atan2}(z_2, z_1). \tag{72}$$

One then quantizes θ (and optionally handles r separately), reconstructing

$$\hat{z} = \hat{r}\,(\cos\hat{\theta}, \sin\hat{\theta}). \tag{73}$$

This may be stronger when angle concentration after preconditioning is far more stable than Cartesian or pairwise-lattice cell occupancy. But it changes the novelty lane from finite pairwise lattice geometry to polar-coordinate base design.

7.5 Replacement-level alternative 2: codebook vector quantization

Recent codebook-based KV compressors operate beyond pairwise local geometry. In abstract form,

$$\hat{y} = C\Big[\arg\min_j \|y - C_j\|_2^2\Big], \tag{74}$$

where C is a learned or statelessly initialized codebook. This family includes commutative, coupled, and vector-quantized approaches. It may be stronger at aggressive compression ratios because it models higher-dimensional structure directly, but it is no longer a lightweight pairwise base-layer change inside the current residual-QJL pipeline.

7.6 Replacement-level alternative 3: transform-side pairwise rotations plus scalar quantization

If the real problem is not base geometry but poor channel conditioning, then pairwise rotations may be the better home for pair structure. For a chosen pair (i, j), apply a Givens rotation,

$$\begin{bmatrix} \tilde{y}_i \\ \tilde{y}_j \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} y_i \\ y_j \end{bmatrix}. \tag{75}$$

After rotation, one can revert to separable scalar quantization and keep the residual-QJL stage unchanged. This is a genuine replacement-level competitor because it removes A2Quant altogether and moves novelty into the transform layer.

7.7 Replacement-level alternative 4: attention-fidelity factorization

Another family abandons quantization as the first principle and compresses the object that attention cares about more directly. In a schematic low-rank formulation,

$$A \approx UV^\top, \qquad \operatorname{rank}(UV^\top) = r \ll d, \tag{76}$$

or one directly approximates key-query interactions. Such methods can be stronger when the right objective is attention fidelity rather than channel-wise storage geometry.

7.8 Replacement-level alternative 5: wave-inspired ideas translated into defensible transform coding

The uploaded WaveQuant note is conceptually ambitious but too speculative to adopt as the main method. Its defensible mathematical translation is not holographic physics; it is transform coding. Let F be an orthogonal transform acting on a structured axis of the cache, and define

$$\xi = Fy, \tag{77}$$

$$\hat{\xi}_i = Q_i(\xi_i), \tag{78}$$

$$\hat{y} = F^{-1}\hat{\xi}. \tag{79}$$

This replacement-level alternative resembles modern frequency-domain or transform-coded KV compression. It is more plausible than a literal wave-scattering interpretation because it is already expressible as an explicit compression operator with known storage, complexity, and reconstruction rules. It is nevertheless not adopted here because it changes both the geometry and the storage semantics of the original project.

Direction	Category	Why it may be stronger	What it changes	Adopted here?
Adaptive pairing plus A2Quant	Compatible upgrade	Makes pairwise isotropy more plausible	Adds pairing metadata and calibration step	No, but prioritized
Calibration-free alignment plus A2Quant	Compatible upgrade	Mitigates distribution shift and reduces calibration sensitivity	Specifies the transform layer more strongly	No
Polar-angle base	Replacement	May exploit angular concentration better	Replaces A2Quant base geometry	No
Codebook VQ or coupled quantization	Replacement	Models higher-dimensional structure	Replaces local pairwise base with learned codebooks	No
Pairwise rotations plus scalar plus QJL	Replacement	Handles outliers upstream	Shifts novelty from base to transform	No
Low-rank attention-fidelity compression	Replacement	Targets the task objective directly	Leaves the quantization lane entirely	No
Transform coding or frequency-domain compression	Replacement	May compress long-cache redundancy more globally	Changes storage axis and semantics	No

Table 22: Alternatives and adoption decision.

The conservative judgment is therefore as follows: retain A2Quant as the main method, elevate adaptive pairing to the leading follow-on experiment, and keep all other redesigns explicit and separate.

8 Conclusion

This paper sharpens a narrow but defensible hypothesis: in a residual-QJL pipeline for KV-cache compression, the base quantizer geometry is not a detail but a meaningful control knob for conditional estimator variance. A2Quant operationalizes that hypothesis with a finite pairwise A_2-style lattice, the searched separable 32-state product family at matched 5 bits per pair, and a code-grounded implementation path. The theory is intentionally modest. It does not claim universal superiority, only a mechanism: if mixing yields locally near-isotropic pairwise marginals and saturation remains low, then a more isotropic two-dimensional base geometry can reduce base residual

energy, which in turn reduces the residual-QJL variance bound.

Question	Answer in this manuscript
Is the main method retained?	Yes; A2Quant remains the adopted main design
Is the method empirically validated here?	No
What is the strongest compatible upgrade?	Adaptive pairing
What are the strongest replacement-level competitors?	Codebook VQ, transform coding, and attention-fidelity factorization
What is the honest current claim?	A theory-plus-methods mechanism linking base geometry to conditional QJL variance

Table 23: Final takeaways.

The manuscript therefore lands in the right place for the current project state: stronger than a research memo, narrower than a benchmark paper, and explicit about what evidence is still needed.

References

1. *TurboQuant: Online Vector Quantization with Near-optimal Distortion Rate.* ICLR 2026 poster.
2. *PolarQuant: Vector Quantization with Polar Transformation.* AISTATS 2026 poster.
3. *QJL: 1-Bit Quantized JL Transform for KV Cache Quantization with Zero Overhead.* SLLM at ICLR 2025 workshop paper (OpenReview public version).
4. *KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization.* NeurIPS 2024 main conference track.
5. *ParoQuant: Pairwise Rotation Quantization for Efficient Reasoning LLM Inference.* ICLR 2026 poster.
6. *IsoQuant: Hardware-Aligned SO(4) Isoclinic Rotations for LLM KV Cache Compression.* arXiv preprint, 2026.
7. *NSNQuant: A Double Normalization Approach for Calibration-Free Low-Bit Vector Quantization of KV Cache.* NeurIPS 2025 poster.
8. *KV Cache is 1 Bit Per Channel: Efficient Large Language Model Inference with Coupled Quantization* (CQ). NeurIPS 2024 main conference track.
9. *CommVQ: Commutative Vector Quantization for KV Cache Compression.* Proceedings of the 42nd International Conference on Machine Learning (ICML 2025), PMLR 267.
10. *VQKV: High-Fidelity and High-Ratio Cache Compression via Vector-Quantization.* ICLR 2026 withdrawn submission.
11. *KQ-SVD: Compressing the KV Cache with Provable Guarantees on Attention Fidelity.* AISTATS 2026 poster.

12. *KV Cache Transform Coding for Compact Storage in LLM Inference.* ICLR 2026 poster.

13. *FreqKV: Key-Value Compression in Frequency Domain for Context Window Extension.* ICLR 2026 poster.

14. *The Pitfalls of KV Cache Compression.* ICLR 2026 withdrawn submission.

15. *GaugeKV: Composable Exact KV Cache Compression.* ICLR 2026 desk-rejected submission.

16. *SpinQuant: LLM Quantization with Learned Rotations.* ICLR 2025 poster.

A Additional Analytical Notes

A.1 Why the manuscript keeps A2Quant as the main design

The decision is not based on novelty for novelty's sake. A2Quant remains the main design because it has the clearest novelty separation from TurboQuant, QJL, PolarQuant, and transform-centric rotation papers while still fitting the current implementation assets. It is also the only design among the nearby alternatives whose central claim can be phrased entirely through the single scalar quantity $\|e\|_2^2$ without changing the residual estimator.

A.2 Why the manuscript does not silently replace the method with codebook VQ or transform coding

Codebook VQ and transform coding may ultimately be stronger in practice, especially at aggressive compression ratios or for cache reuse. But adopting them as the main method would produce a different paper with a different baseline story, different systems assumptions, and different proofs. The current manuscript therefore treats them as neighboring projects rather than improvements in disguise.

A.3 Honest non-claims

This draft does not establish that A2Quant improves long-context accuracy, reasoning benchmarks, or production latency. It does not establish that dense random mixing is the right transform. It does not establish that residual QJL is the best fidelity-restoration strategy. It establishes only that the project has a coherent theory-plus-methods architecture with a clean base-geometry hypothesis and a concrete implementation path.