# ResidualKV

## Residual-Based KV Cache Compression for Efficient Long-Context Inference

Paper by Jitai Hao, Qiang Huang, Yaowei Wang,
Min Zhang, and Jun Yu

Research group discussion | 8 October 2026

arXiv:2602.08005v2 (30 September 2026)

Shared historical references + compact token-specific residuals

Reading presentation; not authored by the paper's authors. https://arxiv.org/abs/2602.08005v2

# Long-context inference has three coupled costs

**Goal**

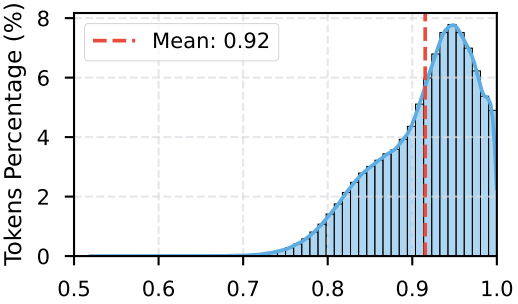
Preserve useful history while reducing both stored KV bytes and active attention reads.

| Family | Main saving | Remaining cost |
|---|---|---|
| Static eviction | Fewer stored tokens | Deleted history cannot return |
| Dynamic selection | Fewer active tokens | Full history remains accessible |
| KV compression | Fewer bytes per token | Full-history processing |

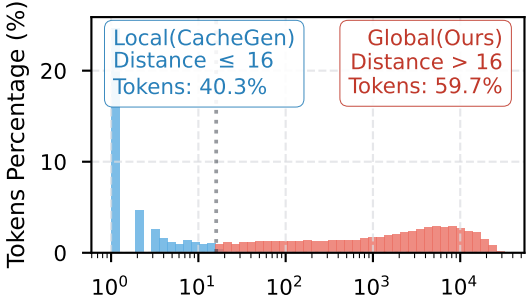
Full prefill attention is quadratic in length; a cached decode step reads history linearly.

Paper, Secs. I–III, PDF pp. 1–3. Complexity distinction stated here for clarity.

# Useful reference tokens are often far away



(a) Max Cosine Similarity with History Token

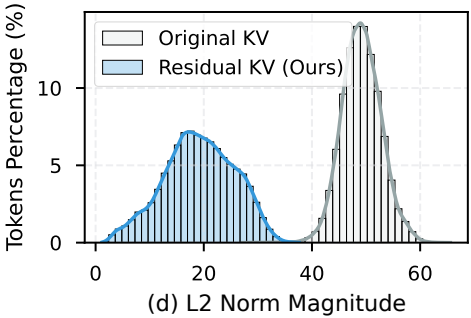


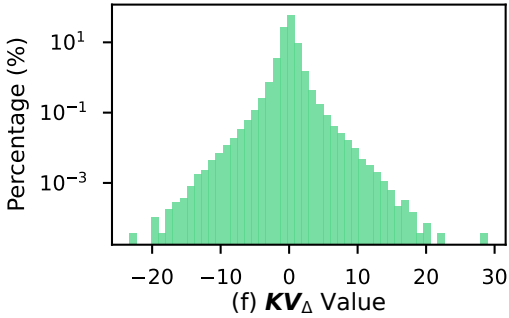
(b) Log Distance to Most Similar Token

- Mean *best historical* cosine similarity: 0.92.
- 59.7% of nearest matches are more than 16 positions away.

Paper, Sec. III, Fig. 3(a–b), PDF pp. 3–4; Qwen2.5-7B-Instruct-1M.

# Reference subtraction reduces residual magnitude

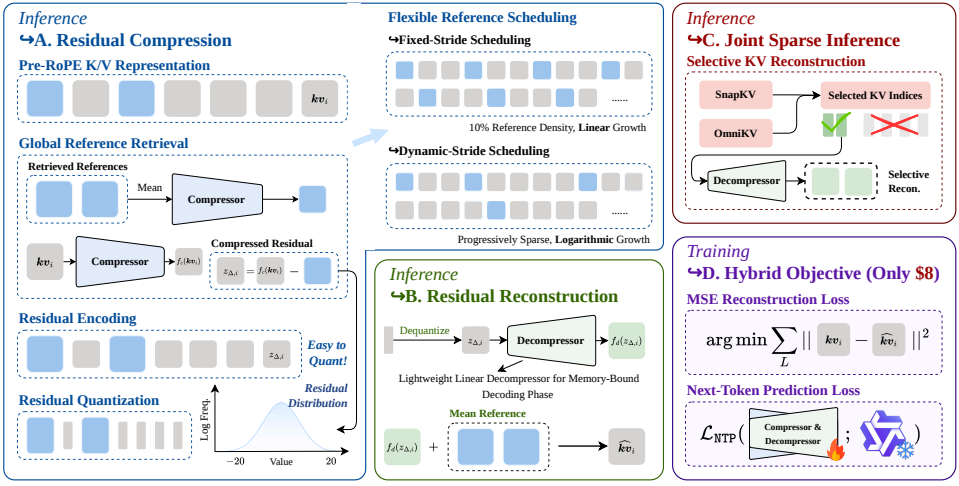



- Smaller norms and fewer extreme values favor low-bit encoding.
- Lower energy does not imply lower rank: the residual spectrum is flatter.

Paper, Fig. 3(c–f), PDF p. 4. Rank/energy distinction is the reading's interpretation.

# The cache separates reusable structure from deviations




Paper, Fig. 4, PDF p. 5. A: encode; B: reconstruct; C: select; D: train.

# The actual code subtracts two nonlinear encodings

1. Concatenate pre-RoPE K and V: $\mathbf{kv}_i \in \mathbb{R}^{2d_k}$.
2. Retrieve $k_i$ nearest references by squared $L_2$ distance; average them.
3. Encode the token and the reference mean separately:

$$f_c(x) = (\mathrm{Swish}(xW_{c1}) \odot xW_{c2})\, W_{c3}, \qquad f_c : \mathbb{R}^{2d_k} \to \mathbb{R}^{d_c}$$

## Residual code

$$\mathbf{z}_{\Delta,i} = f_c(\mathbf{kv}_i) - f_c(\overline{\mathbf{KV}}_{\mathrm{ref},i})$$

This generally differs from $f_c(\mathbf{kv}_i - \overline{\mathbf{KV}}_{\mathrm{ref},i})$.

Paper, Sec. IV-A, PDF pp. 4–5. Llama-3.1-8B: $2d_k = 2048$, $d_c = 512$.

# Token-wise quantization enables selective reconstruction

**Store**

- Groups of at most 32 latent coordinates, within each token.
- Asymmetric min–max quantization.
- INT4 residuals by default; INT2 in multimodal budget experiments.
- References stored in FP8 in deployment.

**Recover**

$$\widetilde{\mathbf{z}_{\Delta,i}} = Q(\mathbf{z}_{\Delta,i})$$

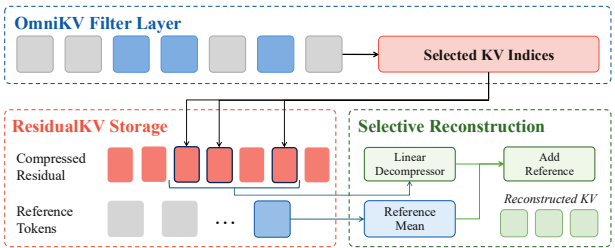
$$\widehat{\mathbf{KV}}_i = \widetilde{\mathbf{z}_{\Delta,i}} W_d + \overline{\mathbf{KV}}_{\mathrm{ref},i}$$

**Cheap repeated read**

One bias-free linear map; split K/V and use the original token's position.

Paper, Secs. IV-A–B and V-A, PDF pp. 5–6, 9. Full-attention layers use 4-bit KV.

# OmniKV selects tokens; ResidualKV reconstructs them




1. Filter layers scan their full history.
2. Selected indices guide sparse layers.
3. Fetch and decode only those residuals and references.

Unselected history remains available in the dynamic combination.

Paper, Sec. IV-C, Fig. 7, PDF p. 7. Static-eviction combinations only retain their selected subsets plus references.

# Dynamic stride limits growth of the reference pool

Fixed stride keeps every $s$-th token; default $s = 10$.

$$p_{m+1} = p_m + s_0 + \lfloor \alpha(p_m - n_{\text{sink}}) \rfloor$$

For fixed $\alpha > 0$, ignoring rounding:

$$N_{\text{ref}}(T) \approx \log_{1+\alpha}\left(1 + \frac{\alpha(T - n_{\text{sink}})}{s_0}\right).$$

- $\alpha = 0$: fixed-stride limit; reference count grows linearly.
- $\alpha > 0$: approximately logarithmic reference count.
- Total cache still grows linearly: token codes and full layers remain.

Paper, Sec. IV-A, PDF pp. 5–6. $T$ denotes sequence length here.

# Closer references tighten a KV reconstruction bound

Let $\Delta_i = \mathbf{kv}_i - \overline{\mathbf{KV}}_{\mathrm{ref},i}$. The retrieval radius satisfies

$$\rho_i^2 = \frac{1}{k_i} \sum_{j \in \mathcal{R}_i} \|\mathbf{kv}_i - \mathbf{kv}_j\|_2^2 = \|\Delta_i\|_2^2 + V_i,$$

where $V_i$ is reference dispersion around their mean.

**Theorem IV.1**

$$\|\mathbf{kv}_i - \widehat{\mathbf{KV}}_i\|_2 \le (\eta_{\mathcal{S}} + C_b)\sqrt{\rho_i^2 - V_i} \le (\eta_{\mathcal{S}} + C_b)\rho_i.$$

- $\eta_{\mathcal{S}}$: codec relative error on the residual support.
- $C_b$: quantization term, depending on code gain and decoder norm.
- Top-$k$ retrieval minimizes $\rho_i$; task accuracy is evaluated empirically.

Paper, Sec. IV-B, Theorem IV.1, PDF pp. 6–7. Proof and assumptions in backup slides.

# Train the codec for both reconstruction and prediction

$$\mathcal{L} = \underbrace{\sum_{l,t} \|\mathbf{KV}_{l,t} - \widehat{\mathbf{KV}}_{l,t}\|_2^2}_{\mathcal{L}_{\mathrm{rec}}} + \underbrace{\mathcal{L}_{\mathrm{ntp}}(\phi;\theta)}_{\text{prediction through recovered KV}}$$

- Backbone $\theta$ frozen; only codec parameters $\phi$ updated.
- Reconstruction preserves values; NTP protects prediction-sensitive directions.
- Text 7B/8B: 160M FineWeb-Edu tokens, 8,192-token sequences.
- About 8 GPU-hours on one RTX PRO 6000 Blackwell; 32B: 14 hours.

Qwen3-VL uses a separate multimodal setup: 30k steps on streamed LLaVA-OV data.

Paper, Sec. IV-D and Sec. V-A, PDF p. 8. The 8-hour figure is not a universal training cost.

# The evaluation spans different uses of history

| Evaluation | Question | Protocol |
|---|---|---|
| LongBench | Long-input understanding | 16 datasets; $\alpha = 0.001$ |
| SCBench | Repeated context queries | SCDQ subset; $\alpha = 0$ |
| AIME | Long reasoning traces | Up to 32,768 output tokens |
| Multimodal | Visual understanding | 7 benchmarks; INT2 |
| Throughput | Real decode benefit | One H100, batch size 1 |

Baselines cover static eviction, dynamic selection, quantization, low-rank compression, and visual-token reduction.

Paper, Sec. V-A, PDF pp. 8–9. Task scores and throughput: higher is better. KR and CR: lower is better.

# Stored bytes, attention reads, and latency are distinct

Let $\beta$ be the fraction of full-attention layers, $r$ the sparse-layer access fraction.

$$q_b = \frac{b}{16} + \frac{2}{32}, \qquad q_4 = 31.25\%, \quad q_2 = 18.75\%$$

$$\mathrm{KR} = \beta q_4 + (1-\beta)\left(\frac{1}{2}\rho_{\mathrm{ref}} + q_b \frac{d_c}{2d_k}\right), \qquad \mathrm{CR} = \beta + (1-\beta)r$$

- KR: KV storage relative to full FP16; includes quantizer extrema.
- CR: layer-averaged historical attention access; $\mathrm{CR} \geq \beta$.
- Llama, 5/32 full layers, 10% references, $d_c/(2d_k) = 1/4$: KR $\approx 15.7\%$.

Paper, Sec. V-A, PDF p. 9. The formula does not itemize reference indices, workspaces, or codec weights.

# LongBench retains competitive quality at lower storage

| Backbone | KR (%) | CR (%) | ResidualKV | OmniKV |
|---|---|---|---|---|
| Llama-3.1-8B | 13 | 30 | 50.3 | 50.2 |
| Qwen2.5-7B | 16 | 30 | 47.3 | 47.2 |
| Qwen2.5-32B | 13 | 30 | 48.4 | 48.5 |
| Qwen3-4B | 13 | 30 | 47.9 | 47.9 |

- OmniKV retains 100% KV storage at the table's 30% CR.
- Average is over 16 datasets, not an unweighted mean of six categories.
- Other strong rows: Llama ResidualKV$^{\dagger}$ 50.4; Qwen2.5-7B KVzip 47.4.

Paper, Table III, PDF p. 10. †: no neural compressor, with a different storage budget.

# Multi-query gains coexist with retrieval failures

| SCBench backbone | ResidualKV | Best listed alternative | Gain |
|---|---|---|---|
| Qwen2.5-7B | 40.7 | 36.7 (KVzip) | +4.0 |
| Qwen3-4B | 32.9 | 22.8 (SnapKV) | +10.1 |

- Qwen2.5 key-value retrieval: 52.6, vs. KVzip 36.2 and SnapKV 1.4.
- Its English QA remains below KVzip: 29.9 vs. 44.0.
- Qwen3 prefix-suffix retrieval is still 0.0.

Paper, Table IV, PDF p. 10. Reported KR/CR: 19/20% and 16/20%. Selected SCDQ tasks; no full-cache row.

# Multimodal quality is strong; budget accounting conflicts

| Backbone | Full cache | ResidualKV | Reported KR/CR |
|---|---|---|---|
| LLaVA-OV-7B | 73.8 | 73.6 | 8% / 10% |
| Qwen3-VL-8B | 74.5 | 74.4 | 10% / 10% |

Qwen3-VL exceeds DivPrune by 7.1 points (74.4 vs. 67.3).

## A configuration discrepancy in v2

Table I lists 7 full-attention layers for Qwen3-VL-8B. Its official decoder has 36 layers: $\mathrm{CR} \geq 7/36 = 19.4\%$, above the reported 10%.

Paper, Tables I–II and Sec. V-A–B, PDF pp. 8–9. Full-cache scores are in the prose. Official Qwen model config linked in references.

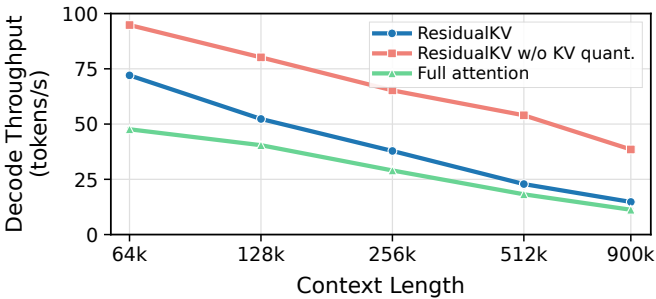
# AIME preserves the measured full-cache score

| Qwen3-4B-Think | Full | SnapKV | OmniKV | ResidualKV |
|---|---|---|---|---|
| AIME score | 80.0 | 60.0 | 76.7 | 80.0 |

- Up to 32,768 generated tokens for reasoning.
- $+20.0$ points over SnapKV; $+3.3$ over OmniKV.
- Table V does not specify KR/CR: do not attach the LongBench budget to this result.

Paper, Table V, PDF p. 10; output-length setting in Sec. V-A, PDF p. 9.

# Quantization changes the practical speed trade-off




**One H100; batch size 1**

- Up to 1.5× with KV quantization.
- Up to 3.4× without KV quantization.
- Full-attention layers still decode their whole low-bit cache.

Paper, Fig. 8 and Sec. V-B, PDF p. 10. Speed maxima and minimum storage do not describe a single configuration.

# A much smaller reference pool can preserve quality

| $\alpha$ | References at 262K | Reference fraction | LongBench |
|---|---|---|---|
| 0 | 26,189 | 9.99% | 47.4 |
| 0.001 | 3,354 | 1.28% | 47.4 |
| 0.020 | 319 | 0.12% | 47.3 |
| 0.050 | 149 | 0.06% | 47.0 |
| 0.100 | 83 | 0.03% | 47.1 |

- $\alpha = 0.001$: about 7.8$\times$ fewer references, same reported score.
- Counts are measured at 262K; the quality column is a LongBench aggregate.

Paper, Table VI, PDF p. 11; Qwen2.5-7B-Ins. Fig. 5 additionally tests RULER variable tracking up to 96K.

# Both learned coding and references matter

| Llama-3.1-8B ablation | KR / CR | Average | Change |
|---|---|---|---|
| ResidualKV | 45% / 30% | 50.3 | — |
| Without $f_c, f_d$ | 45% / 30% | 46.7 | −3.6 |
| Without reference tokens | 47% / 30% | 45.9 | −4.4 |

- Additional KV quantization disabled; $d_c = 768$ instead of 512.
- Table VIII reports 45–47% KR; adjacent prose says about 35%.
- Interpret the matched-budget claim with this discrepancy in view.

Paper, Sec. V-D and Table VIII, PDF p. 11. Values here follow the displayed table.

# The representation also works without dynamic selection

| Llama-3.1-8B / LongBench | KR (%) | CR (%) | Average |
|---|---|---|---|
| ResidualKV standalone | 16 | 100 | 50.5 |
| ResidualKV + OmniKV | 13 | 30 | 50.3 |
| ResidualKV + SnapKV | 17.5 | 40 | 50.1 |

- Standalone matches the full-cache average reported in the prose.
- Dynamic selection adds attention savings and selective decompression.
- Static combinations inherit the external policy's information loss.

Paper, Sec. V-C and Table VII, PDF p. 11; composition mechanics in Sec. IV-C, PDF p. 7.

# The main contribution is a reusable cache representation

## Mechanism

Store shared structure once, retain token-specific deviations, and reconstruct only active states.

- **Supported by results:** competitive LongBench quality, multi-query gains, and composability with selection policies.
- **Engineering trade-off:** learned codec, reference retrieval, quantization, and full-layer overhead.
- **What needs clarification:** multimodal CR accounting, ablation KR discrepancy, and exact serving configurations.

Synthesis of Secs. IV–VI, PDF pp. 4–13. Detailed equations, diagnostics, and model settings follow.

# References and supplementary discussion

**Primary paper**

Jitai Hao, Qiang Huang, Yaowei Wang, Min Zhang, and Jun Yu.
*ResidualKV: Residual-Based KV Cache Compression for Efficient Long-Context Inference.*
arXiv:2602.08005v2, 30 September 2026.

`https://arxiv.org/abs/2602.08005v2`

**Official model metadata used for the layer-count check**
Qwen/Qwen3-VL-8B-Instruct: config.json
`text_config.num_hidden_layers = 36` (accessed 8 October 2026).

**Author implementation** `https://github.com/CURRENTF/ResidualKV`

The following slides are supplementary material for questions and discussion.

# Backup: quantization connects geometry to reconstruction

$$\eta_{\mathcal{S}} = \sup_i \frac{\|\Delta_i - f_d(\mathbf{z}_{\Delta,i})\|_2}{\|\Delta_i\|_2}, \qquad \gamma_{\mathcal{S}} = \sup_i \frac{\|\mathbf{z}_{\Delta,i}\|_2}{\|\Delta_i\|_2}$$

Supremum over retrieval-induced support, excluding zero residuals.

$$\|e_{q,i}\|_2 \le \frac{\sqrt{\sum_{\mathcal{G}} |\mathcal{G}|\ \mathrm{range}(z_{\mathcal{G}})^2}}{2(2^b - 1)} \le \frac{\sqrt{g/2}}{2^b - 1}\|\mathbf{z}_{\Delta,i}\|_2$$

$$C_b = \frac{\|W_d\|_{\mathrm{op}}\gamma_{\mathcal{S}}\sqrt{g/2}}{2^b - 1}$$

- Split total error into unquantized codec error and decoded rounding error.
- Apply the decoder operator norm and $\|\Delta_i\|_2^2 = \rho_i^2 - V_i$.
- References denote the same effective stored values in encoder and decoder.

Paper, Theorem IV.1 and proof, PDF pp. 6–7. $\eta_{\mathcal{S}}$ and $\gamma_{\mathcal{S}}$ are support-dependent, not universal constants.

# Backup: averaging references has diminishing returns

Assume compatible states $\mathbf{kv}_i = s_c + \epsilon_i$, with independent, zero-mean perturbations of covariance $\Sigma_c$.

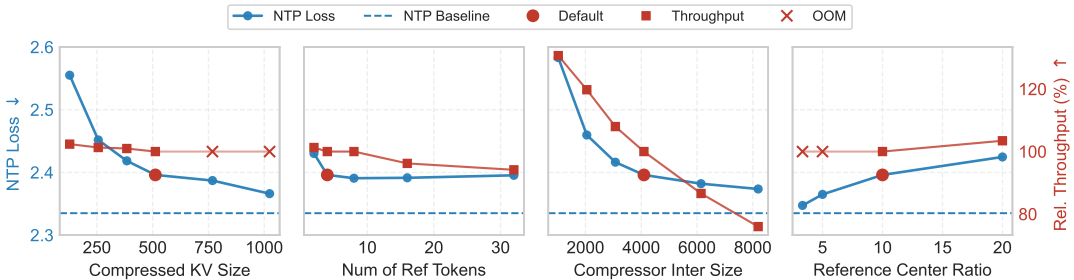
$$\Delta_i = \epsilon_i - \frac{1}{k_i} \sum_{j=1}^{k_i} \epsilon_{r_j}$$

$$\mathrm{Cov}(\Delta_i \mid c) = \left(1 + \frac{1}{k_i}\right) \Sigma_c$$

- Shared component $s_c$ cancels exactly in this idealized model.
- Reference noise falls as $1/k_i$; the current token's variation remains.
- Actual nearest-neighbor selection need not satisfy independence or compatibility.

Paper, Corollary IV.2, PDF p. 7. This is an explanatory model for benefit and saturation.

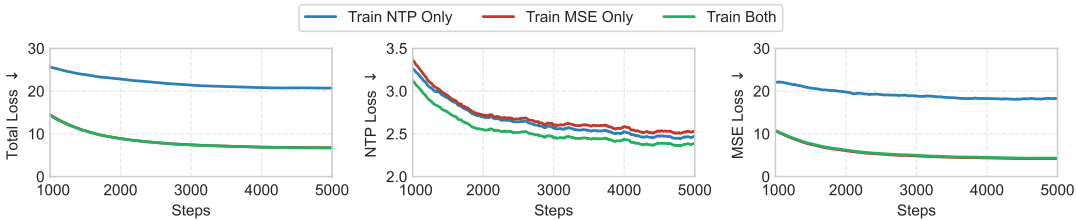
# Backup: hyperparameters exchange quality for throughput




- Larger code/MLP: lower NTP loss, greater memory or compute cost.
- More references saturate; denser reference spacing may cause OOM.

Paper, Fig. 11, PDF p. 12. Throughput: context 128K, batch size 16. Blue: loss; red: throughput; crosses: OOM.

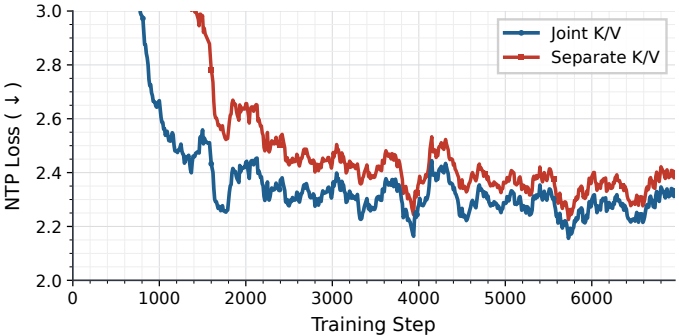
# Backup: joint training addresses two different errors




- MSE-only training also improves next-token prediction.
- NTP-only leaves substantial reconstruction error.
- Joint optimization reaches low reconstruction and prediction losses.

Paper, Fig. 12 and Sec. V-D, PDF p. 12. These are training loss curves, not downstream accuracy measurements.

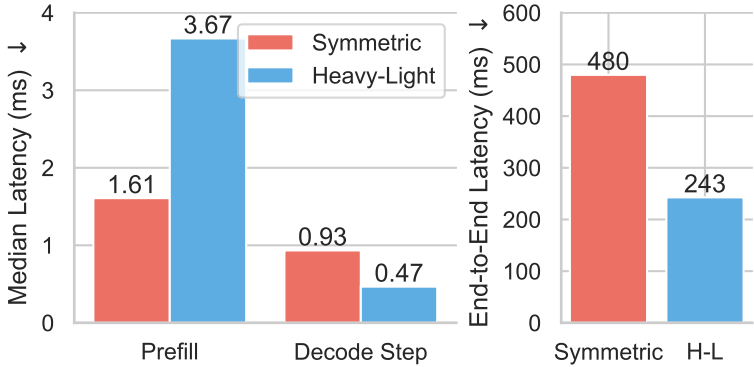
# Backup: K and V benefit from a shared compressor




Paper, Fig. 10, PDF p. 11. Curves are smoothed over 200 steps.

- Matched parameter count, compression ratio, and training steps.
- Joint K/V usually has lower NTP loss.
- Both projections originate from the same hidden state.

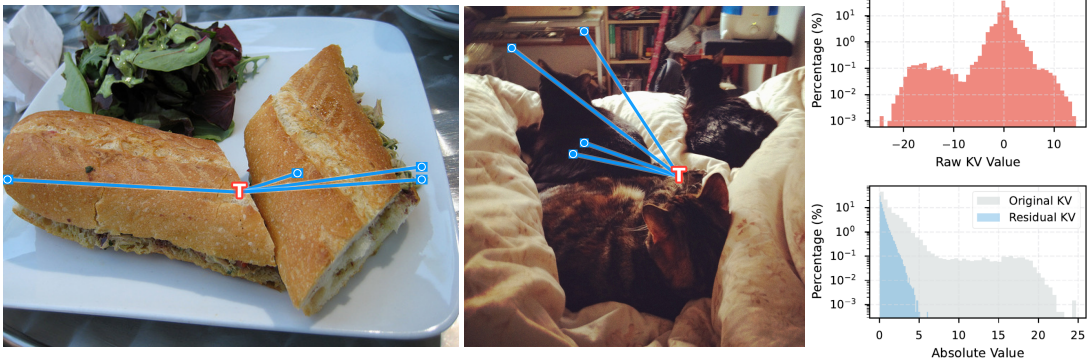
# Backup: a heavier write path can make repeated reads cheaper




Prefill cost rises, but 512 repeated reconstruction steps dominate this test.

Paper, Fig. 6, PDF p. 6. H100; 4,096-token prefill + 512 decode steps; codec-path latency, not whole-model request latency.

# Backup: visual references capture distant structural cues




Targets on object boundaries retrieve distant edges; residuals concentrate near zero.

Paper, Fig. 13, PDF p. 12; case study discussion, PDF pp. 12–13.

# Backup: model dimensions and full-attention layers

| Model | $2d_k$ | $d_c$ | $d_{\text{ff}}$ | Full-attention layer indices |
|---|---|---|---|---|
| LLaVA-OV-7B | 1024 | 384 | 3072 | 0,1,4,7,14 |
| Qwen3-VL-8B | 2048 | 384 | 3072 | 0,1,2,3,10,21,29 |
| Llama-3.1-8B | 2048 | 512 | 3072 | 0,1,2,8,18 |
| Qwen2.5-7B | 1024 | 384 | 3072 | 0,1,2,4,7,14 |
| Qwen2.5-32B | 2048 | 512 | 4096 | 0,1,2,3,4,5,17,29,40 |
| Qwen3-4B Ins/Think | 2048 | 512 | 3072 | 0,3,9,13,16,21,28 |

Default retrieval: $k = 4$, fixed stride $s = 10$. LongBench uses dynamic stride $\alpha = 0.001$.

Paper, Table I and Sec. V-A, PDF pp. 8–9. Separate variants may require different layer configurations.

# Backup: 10% CR cannot follow from the listed VL configuration

$$\beta = \frac{7}{36}, \qquad \mathrm{CR} = \frac{7}{36} + \frac{29}{36} r, \quad r \geq 0$$

## Arithmetic consequence

Even with no sparse-layer history reads, CR is at least 19.4%. If sparse layers read 10%, the same formula gives 27.5% overall.

- Table II reports 10%; Table I lists seven full-attention layers.
- INT2 changes byte storage, not this attention-token ratio.
- A different configuration or denominator could explain the report, but v2 does not specify that mapping.

Reading's calculation using paper Sec. V-A and Tables I–II, PDF pp. 8–9, plus the official 36-layer Qwen3-VL-8B config.