You Only Cache Once

Decoder-Decoder Architectures for Language Models

Paper: Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang,
Wenhui Wang, Shuming Ma, Quanlu Zhang,
Jianyong Wang, and Furu Wei

arXiv:2405.05254v2 (9 May 2024)
Reading presentation: 20 September 2026

One global memory, many layers of queries

- A causal self-decoder builds a shared, token-addressable KV memory.
- A causal cross-decoder repeatedly queries that same memory.
- History needs no separate full-length KV cache at every layer.
- Main payoff: lower memory and much faster long-prompt prefill.

Paper, Sections 1–2, PDF pp. 2–5.

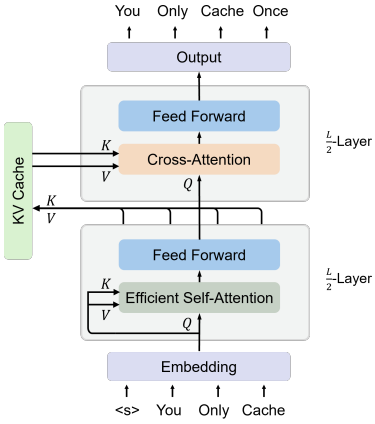
The bottleneck grows with both depth and length

$$\text{Transformer KV memory} \propto L \times N \times d_{\mathrm{KV}}$$

- L: layers; N: cached tokens; d_{KV}: KV width.
- Grouped-query attention reduces KV width, but retains the depth factor.
- Pure recurrent compression is cheap, yet exact historical recall is difficult.
- YOCO keeps global attention while sharing its memory across depth.

Paper, Sections 1–2, PDF pp. 2–4; interpretation of the design tradeoff.

The architecture separates memory building and reading



- First $L/2$ layers: efficient causal self-attention.
- Last $L/2$ layers: global causal cross-attention.
- Different queries at each layer; shared keys and values.
- RMSNorm, SwiGLU, residuals, and GQA remain.

Paper, Figure 2 and Section 2, PDF p. 3; original figure.

Shared KV does not mean identical attention

$$M = X^{L/2}, \qquad \hat{K} = \mathrm{LN}(M)W_K, \quad \hat{V} = \mathrm{LN}(M)W_V$$
$$\hat{Q}^l = \mathrm{LN}(X^l)W_Q^l, \qquad Y^l = \mathrm{Attention}(\hat{Q}^l, \hat{K}, \hat{V}) + X^l$$
$$X^{l+1} = \mathrm{SwiGLU}(\mathrm{LN}(Y^l)) + Y^l$$

- $M \in \mathbb{R}^{N \times d}$; actual KV width depends on GQA.
- Queries and hidden states evolve; memory remains shared.
- Causal masking prevents future-token leakage during training.

Paper, Equations (2)–(3), PDF p. 4. LN denotes RMSNorm.

Gated retention compresses the self-decoder history

$$\gamma_n = \text{sigmoid}(X_n W_\gamma)^{1/\tau}, \quad S_n = \gamma_n S_{n-1} + K_n^\top V_n, \quad O_n = Q_n S_n$$

- Per-head state $S_n \in \mathbb{R}^{d_k \times d_v}$ has fixed size in sequence length.
- Input-dependent, head-wise decay controls forgetting.
- Position-rotated queries/keys supply relative positional information.
- The full module also uses head-wise GroupNorm and a swish output gate.

Paper, Section 3.1, Equations (4), (5), (7), PDF pp. 5–6.

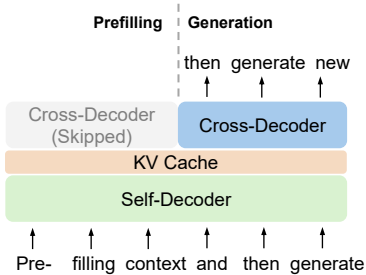
Three equivalent views serve different workloads

$$O = (QK^\top \odot D)V, \qquad D_{nm} = \mathbf{1}_{m \le n} \prod_{i=m+1}^{n} \gamma_i$$

- **Parallel:** expose all pair interactions; convenient for training.
- **Recurrent:** update one state per token; efficient generation.
- **Chunkwise:** parallel within a block, recurrent between blocks.
- A sliding-window self-decoder is an alternative; $C = 1{,}024$ in scaling experiments.

Paper, Sections 3.1–3.2 and 4.2, PDF pp. 5–8. Chunk-index caveat in backup.

Prefill skips historical cross-decoder outputs



- Build all prompt KV with the self-decoder; old cross outputs are not future memory.
- The last prompt position still needs cross-decoder computation for next-token logits.
- Full training loss and generated tokens still use the cross-decoder.

Paper, Figure 3 and Section 2.3, PDF pp. 4–5. Last-position detail follows the dependency graph.

The complexity gain has a precise scope

	Transformer	YOCO
Cache memory	$O(LND)$	$O((N+L)D)$
Prefill attention	$O(LN^2D)$	$O(LND)$

- Paper notation suppresses fixed window/head-state constants.
- "Once" means one global cache; per-layer local states still exist.
- Weights, FFNs, activations, and communication are separate costs.
- Generation still reads global KV; full training attention is not made linear.

Paper, Tables 1–2, PDF p. 4; scope clarified from the architecture.

Four experiments answer four different questions

1. **Quality at scale:** 3B model trained up to 1.6T tokens.
2. **Parameter scaling:** 160M–13B, each trained on 10B tokens.
3. **Long context:** extend 3B progressively to 1M tokens.
4. **Deployment:** memory, prefill, and total inference throughput.

Public-model comparisons and controlled scaling runs offer different evidence.

Paper, Section 4, PDF pp. 6–11.

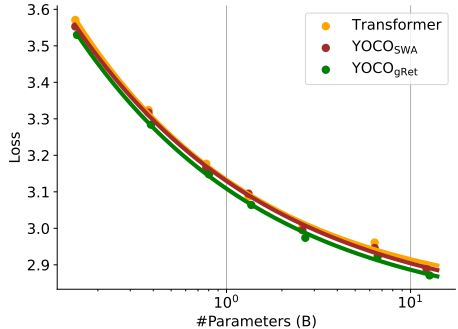
The 3B model reaches competitive downstream quality

Model	Training tokens	8-task average ↑
OpenLLaMA-3B-v2	1T	0.619
StableLM-base-alpha-3B-v2	1T	0.612
YOCO-3B	1T	**0.634**
YOCO-3B	1.6T	0.636
YOCO-3B-1M	extended	0.645

- At 1T: $+1.5$ percentage points over OpenLLaMA on the mean.
- Not a per-task sweep; training data and recipes are not fully controlled.
- 5T is the learning-rate schedule, not the actual completed training.

Paper, Table 3, PDF p. 7; Appendix C, p. 17. Zero-shot LM Eval Harness.

Fixed-budget scaling favors the hybrid design



- Same data/settings; 10B tokens for every size; validation loss is lower-is-better.
- FFN widths align parameter counts. This is not compute-optimal scaling.

Paper, Figure 4, PDF p. 7; Section 4.2, p. 8; Appendix D, pp. 17–18.

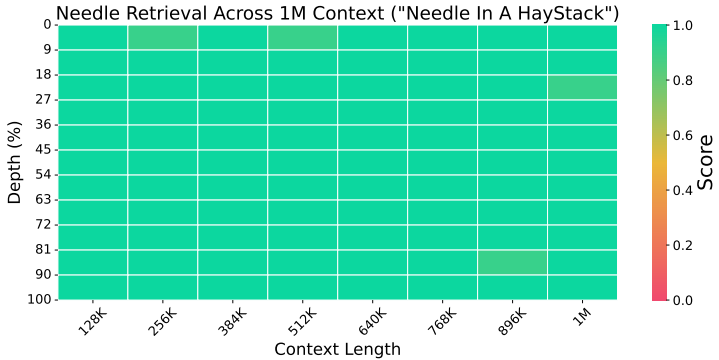
One million tokens requires continued training

Context length	Tokens	Learning rate	RoPE θ
65,536	6B	8×10^{-5}	640K
262,144	4B	4×10^{-5}	5M
1,048,576	1.5B	2×10^{-5}	80M

- An additional 11.5B tokens, with long-document upsampling.
- No long-instruction tuning data in this extension.
- Not zero-training extrapolation from the original 4K model.

Paper, Section 4.3, PDF p. 8; Appendix E, Table 7, p. 18.

Single-needle recall stays strong up to 1M



- Axes: context length and insertion depth; color: accuracy, higher is better.
- 10 trials per depth/length; city + magic-number retrieval. Not general reasoning.

Paper, Figure 5 and Section 4.3, PDF p. 8; original figure.

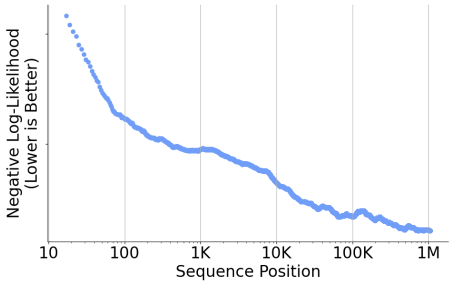
Multi-needle retrieval reveals a harder boundary

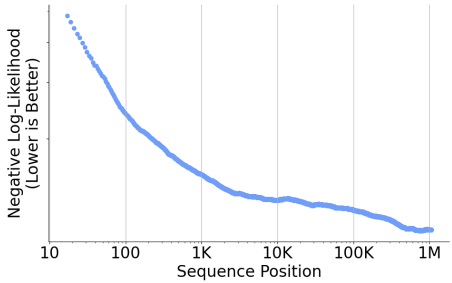
Model	1 needle	2	4	8
LWM-1M-text (7B)	1.00	0.90	0.76	**0.62**
MiniCPM-128K (2.4B)	1.00	1.00	0.54	0.56
ChatGLM3-128K (6B)	0.94	0.72	0.52	0.44
YOCO-3B-1M	0.98	0.98	**0.84**	0.56

- All these retrieval tests use **128K context**, not 1M.
- YOCO leads at four needles, but trails LWM at eight.
- Model sizes and context-training histories differ.

Paper, selected rows from Table 4, PDF p. 8. Accuracy: higher is better.

Long-document prediction improves with length

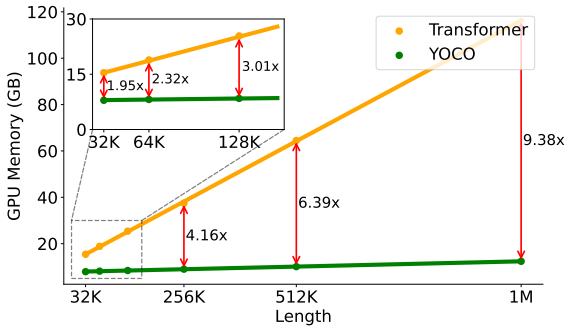




- Books (left), repository-level code (right); examples longer than 1M.
- Metric is cumulative average NLL, lower is better.
- Position/content changes also affect this curve; not a fixed-target context ablation.

Paper, Figure 6, PDF p. 9; original panels. Last point is reading analysis.

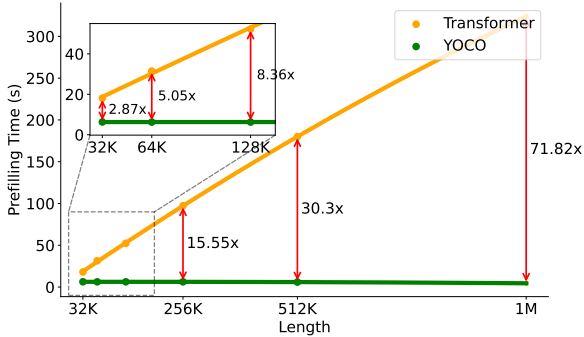
Memory savings grow with context length



- Default 3B configuration: 12.4 GB at 1M; Transformer uses about 9.4 times more.
- Total inference memory includes weights, activations, and KV.

Paper, Figure 7(a) and Section 4.4, PDF p. 10; original figure.

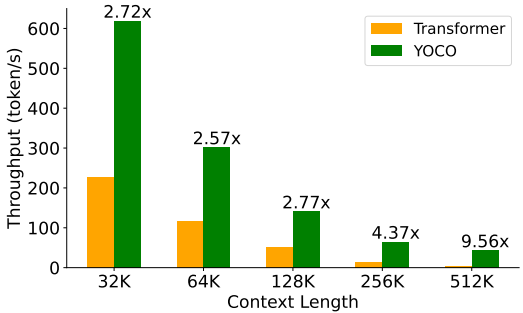
Prefill benefits directly from the dependency change



- 512K: about 180 seconds to less than 6 seconds; 32K: 2.87 times faster.
- 1M: figure reports 71.82 times speedup; hardware/kernel conditions matter.

Paper, Figure 9, PDF p. 11; Section 4.4, pp. 9–10. H100-80GB profiling.

Throughput includes more than token decoding



- 512K: 4.5 to 43.1 tokens/s, about 9.6 times higher.
- Includes prefill and generation; final 1,024 tokens are generated.
- Lower memory permits larger batches. Not fixed-batch decode speedup.

Paper, Figure 10 and Section 4.4, PDF pp. 9–11.

Small-model diagnostics support global memory

160M model	Validation ↓	AR-Hit ↓	First-Occur ↓
Mamba	3.645	1.555	4.126
gRetNet	3.600	1.354	4.116
Transformer	3.564	1.219	4.104
YOCO-gRet	**3.530**	**1.199**	**4.067**

- AR-Hit: target bigram already appeared in the context.
- First-Occur: target cannot be directly recalled from the context.
- Values and the paper's "perplexity" label are retained as reported.

Paper, selected rows from Table 8, Appendix G.1, PDF p. 19.

What is established, and what remains open

- **Established here:** competitive quality and large long-prompt deployment gains.
- **Architectural limit:** global KV still grows with length; deep layers cannot rewrite it.
- **Evidence limit:** synthetic retrieval is not general long-context reasoning.
- **Reproduction limit:** per-length batch/device details and uncertainty are incomplete.
- BitNet/Groq integration, multimodal fusion, and indexed memory remain proposals.

Paper, Sections 4–5, PDF pp. 7–12; limitations include reading analysis.

Reference and discussion

You Only Cache Once: Decoder-Decoder Architectures for Language Models

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, Furu Wei.

`https://arxiv.org/abs/2405.05254v2`
Author code link: `https://aka.ms/YOCO`

Discussion: Which tasks need layer-specific historical KV, and which only need deeper queries over a stable memory?

Based on the versioned author TeX and 20-page PDF. Experiments were not reproduced.

Backup: chunkwise recurrence with consistent indices

Let $a = (i-1)B$ and $\beta_j = \prod_{t=1}^{j} \gamma_{a+t}$. With query row j and key column k:

$$D_{jk}^{(i)} = \mathbf{1}_{k \le j} \frac{\beta_j}{\beta_k}$$

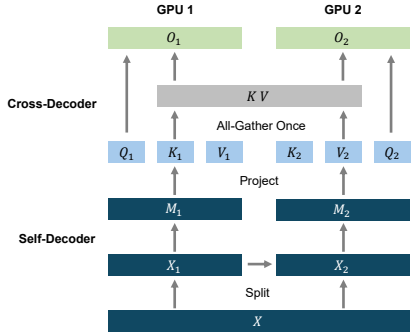
$$O_{[i]} = (Q_{[i]} K_{[i]}^{\top} \odot D^{(i)}) V_{[i]} + \mathrm{diag}(\beta) Q_{[i]} R_{i-1}$$

$$R_i = \beta_B R_{i-1} + \sum_{k=1}^{B} \frac{\beta_B}{\beta_k} K_{a+k}^{\top} V_{a+k}$$

Re-derived from Eqs. (4)–(5). Printed Eq. (6) reverses the mask/ratio under this row convention; Appendix F also has a gt/g typo.

Paper, Equations (4)–(6), pp. 5–6; Appendices B/F, pp. 16–19. Re-derived teaching notation.

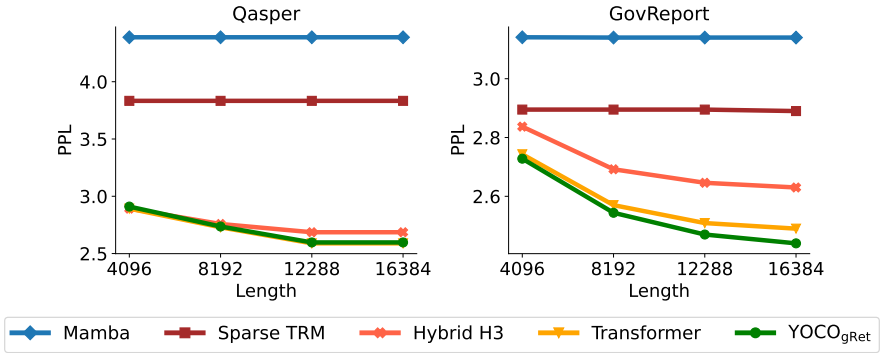
Backup: communicate shared KV once



- Self-decoder exchanges boundary states or local-window information.
- Cross-decoder gathers shared KV once, rather than at every layer.
- Dependency advantage is described; no separate communication-speedup ablation.

Paper, Appendix A, Figure 11, PDF p. 16; original figure.

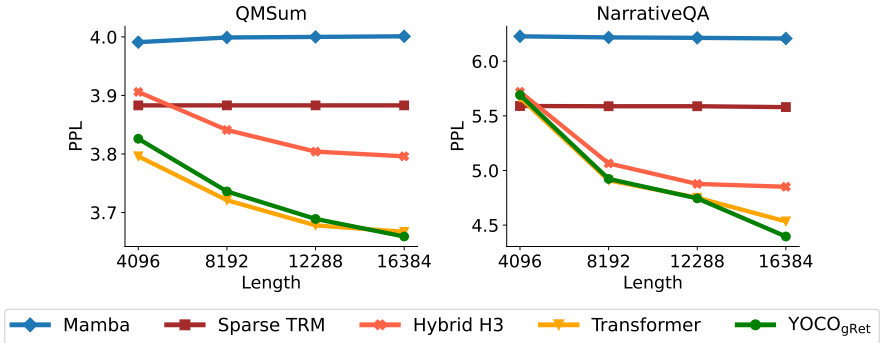
Backup: Qasper and GovReport likelihood



- 160M models; +2B tokens at length 16,384.
- Answer perplexity, lower is better; not task-level answer accuracy.

Paper, Appendix G.2, Figure 12 (top panels), PDF p. 20. Original figure split for readability.

Backup: QMSum and NarrativeQA likelihood



- YOCO and Transformer benefit from longer input context.
- Sparse baseline: 2,048 window and unchanged RoPE base.

Paper, Appendix G.2, Figure 12 (bottom panels), PDF p. 20. Original figure split for readability.