HeatKV: Head-tuned KV-cache Compression for Visual Autoregressive Modeling

Jonathan Cederlund[1,2†], Axel Berg[2], Durmus Alp Emre Acar[2], Chuteng Zhou[2], Pontus Giselsson[1]

[1]Dept. of Automatic Control, Lund University,

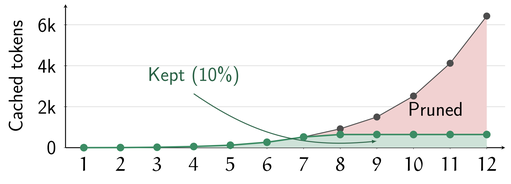
[2]Arm

Visual Autoregressive Modeling

- Generates an image coarse-to-fine: a 1×1 token map, then 2×2, 4×4, …, up to 64×64.
- Each step predicts a whole token map at once, conditioned on the prompt and all earlier maps.
- Earlier maps are stored in a KV-cache, so only 13 steps are needed instead of one per token.
- The catch: the cache grows with every scale, up to ~10 GB for a single 1024×1024 image.

Motivation & Key Observations

- The cache size grows superlinearly with scale with the last 2 scales taking over 60% of the final cache size.
- Attention patterns remain stable between prompts.
- Each head follows its own structure with some heads attending to recent scales while some attend more to early scales.
- Binary classification of heads is too coarse.

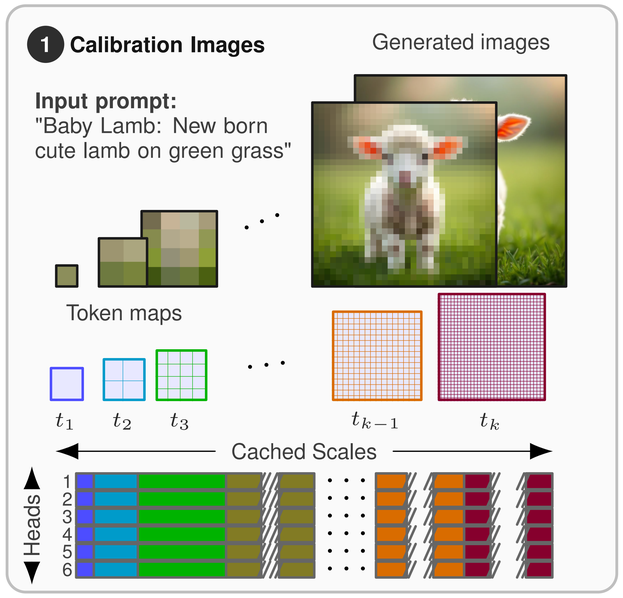


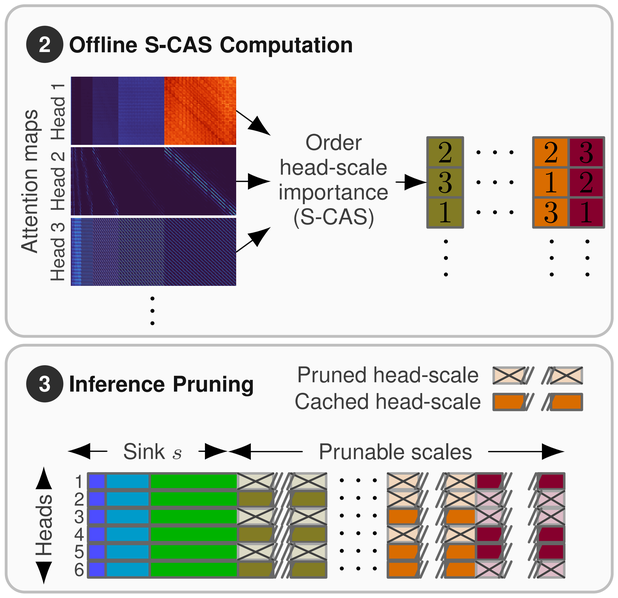
Our Contribution

- Per head-scale cache budgets instead of binary.
- Scale-dependent Cumulative Attention Score (S-CAS) ranking each head scale on attention mass.
- Greedy pruning strategy optimal in the binary case.
- Up to 2x higher compression than previous methods.

Figure 1: Link to the paper on Arxiv

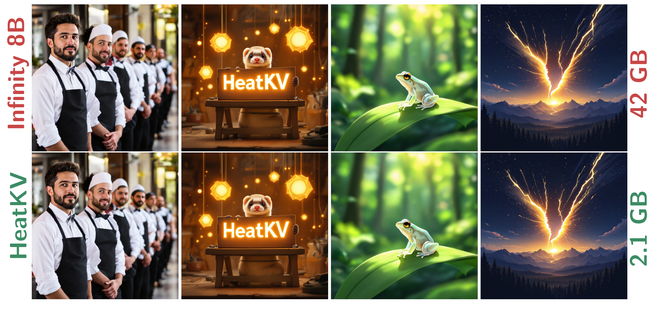
HeatKV





42 GB → 2.1 GB

KV-cache for Infinity-8B, with nearly identical images



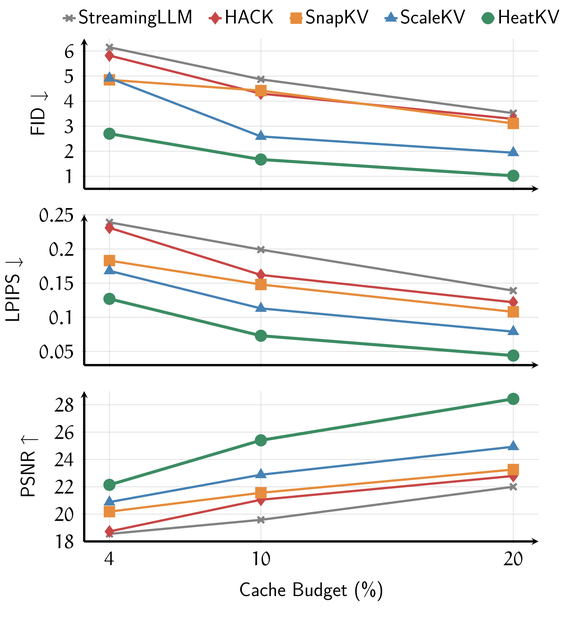
S-CAS

$$\text{S-CAS}_{\ell,h,k} = \underbrace{\frac{1}{K-k}}_{\text{avg. over future scales}} \sum_{\tau=k+1}^{K} \underbrace{\beta_{\ell,h}[\tau,k]}_{\substack{\text{attention on scale } \tau \\ \text{at scale } k \text{ in head } h}}$$

ℓ: layer, h: head, k: cached scale, averaged over calibration prompts.

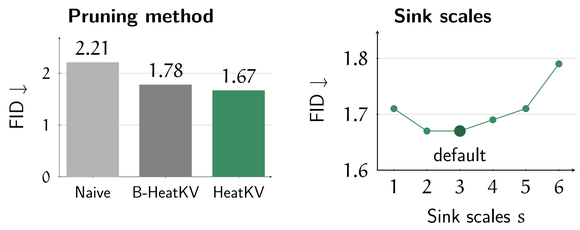
- **Low S-CAS** ⇒ future scales rarely look here ⇒ **prune first**.

Results



Ablations

- **One calibration prompt is enough:** FID 1.67 with $|\mathcal{D}| = 1$ and with $|\mathcal{D}| = 10$.



- **Robust to the calibration domain:** every calibration set beats ScaleKV on every test set.

Calibrated on	FID↓ tested on			
	Car	Portrait	Landscape	COCO
Car	0.94	0.99	2.49	1.70
Portrait	0.95	0.97	2.46	1.68
Landscape	0.94	0.99	2.48	1.68
COCO	0.93	0.91	2.46	1.67
ScaleKV	1.68	2.11	3.66	2.59

Summary

- VAR KV-cache memory grows superlinearly.
- HeatKV prunes per head-scale with an offline schedule.
- Up to 2x higher compression with similar or better image quality.
- HeatKV enables VAR models on consumer hardware.

Acknowledgement

We extend a massive thank you to William Isaksson who designing and implementing the custom Triton kernel that enables HeatKV to achieve on par with the full-cache baseline, as well as for developing the infrastructure supporting the latency measurements. We thank Danny Loh for supporting this work and, Eva Li for her feedback and comments. Finally, we thank Kunjun Li for sharing the code for ScaleKV with the research community. P. Giselsson acknowledges support from the ELLIIT Strategic Research Area, the Wallenberg AI, Autonomous Systems, and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Swedish Research Council.

†Work completed during an internship at Arm.

Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS 2026)