Enabling 8B Bitwise Autoregressive Image Generation on Edge GPUs

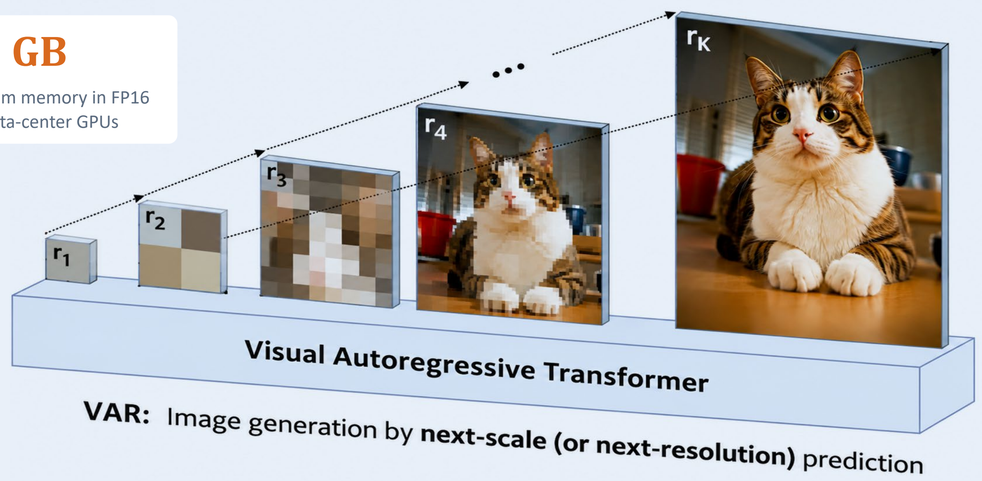
Enrico Vezzali[1,3] Federico Bolelli[1] Costantino Grana[1] Luca Benini[2] Yawei Li[2]

[1]AImageLab, University of Modena and Reggio Emilia, Italy · [2]Integrated Systems Laboratory, ETH Zürich, Switzerland · [3]Datalogic S.p.A., Bologna, Italy

The Memory Wall in Visual Autoregressive Models

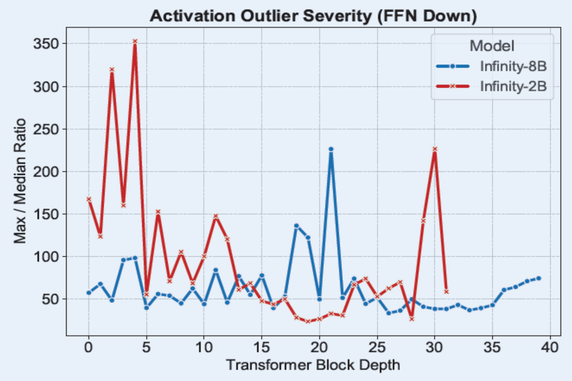
- Visual Autoregressive (VAR) models such as Infinity reach state-of-the-art image fidelity by scaling next-scale prediction up to 8B parameters, rivaling diffusion transformers.
- Unlike diffusion models' constant VRAM footprint, VAR needs a Key-Value cache that grows every generation step — on top of large static weights. Infinity-8B peaks at 37.1 GB, far beyond any edge GPU.
- Real-time robotic visual planning, bandwidth-limited semantic compression, and privacy-sensitive on-device generation are all locked out by these memory demands.

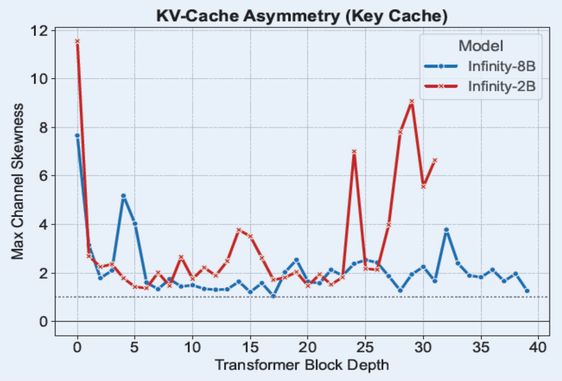
37.1 GB
Infinity-8B peak system memory in FP16 — confined to data-center GPUs



Activation Outliers

- **Our plan is to quantize weights and activations to INT4 and KV-Cache to INT8**. Activations are the toughest challenge.
- VAR models exhibit the **"Massive Activation"** phenomenon known from LLMs: FFN Down-Projection activations reach up to 353× the median value (Infinity-2B) — collapsing naive INT4 quantization.
- KV-Cache variance is strictly channel-driven (5× higher across channels than tokens), with per-channel skewness up to γ = 11.6 — symmetric clipping wastes bit-width and destroys signal.



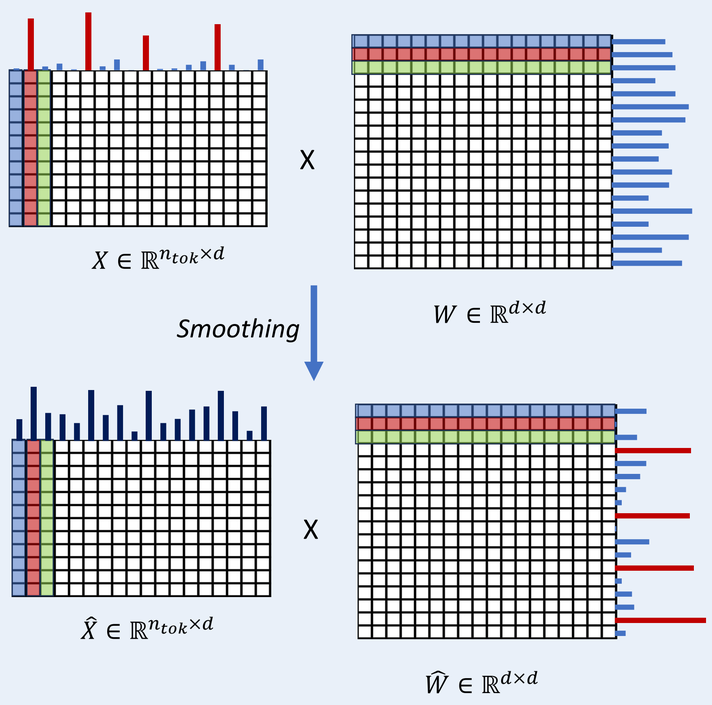


353×
Peak activation outlier vs. median (FFN Down-Proj, Infinity-2B)

11.6
Max per-channel skewness in KV-Cache

Step 1 – Smoothing Activations

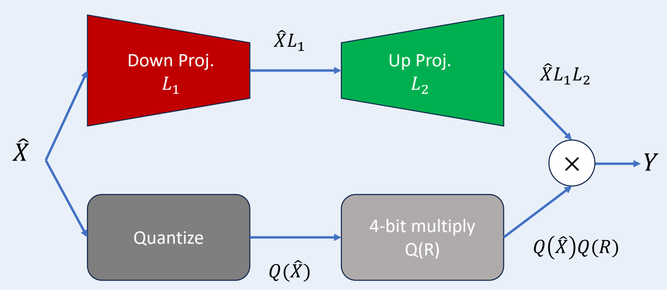
- Block-wise INT4 collapses Infinity-2B (IR 0.981 → 0.616): few outliers dominate each block, zeroing the rest.
- SmoothQuant migrates outlier energy from activations into weights using channel scaling via **diagonal preconditioning**.
- Only partial recovery (IR 0.667).



Step 2 – SVDQuant: A Low-Rank Fix

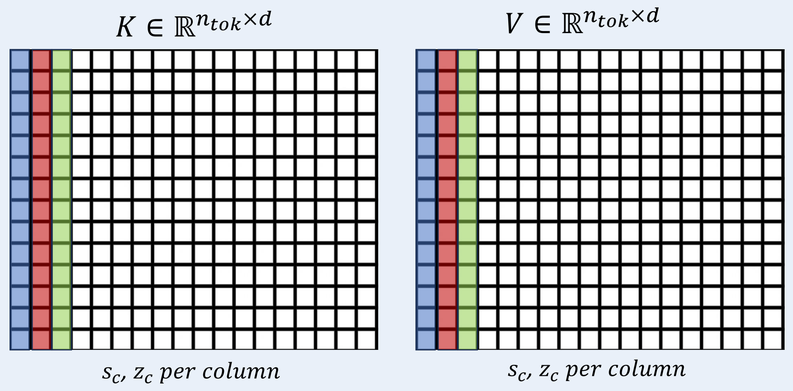
- Singular Value Decomposition isolates these newly formed high-amplitude weight rows into a compact, 16-bit low-rank branch $(L_1 L_2) \rightarrow \widehat{W} = L_1 L_2 + R$
- Optimal by Eckart–Young: the INT4 residual has the minimum possible magnitude for its rank.
- Two branches run in parallel; low-rank branch adds <3% compute and even less memory overhead.

$$XW = \hat{X}\widehat{W} = \hat{X}L_1L_2 + \hat{X}R \approx \underbrace{\hat{X}L_1L_2}_{\text{16-bit low-rank branch}} + \underbrace{Q(\hat{X})Q(R)}_{\text{4-bit residual}}$$



0.667 → 0.848
Image Reward recovered by SVDQuant vs. SmoothQuant (Infinity-2B, MJHQ-30K)

Step 3 – Asymmetric KV-Cache Quantization

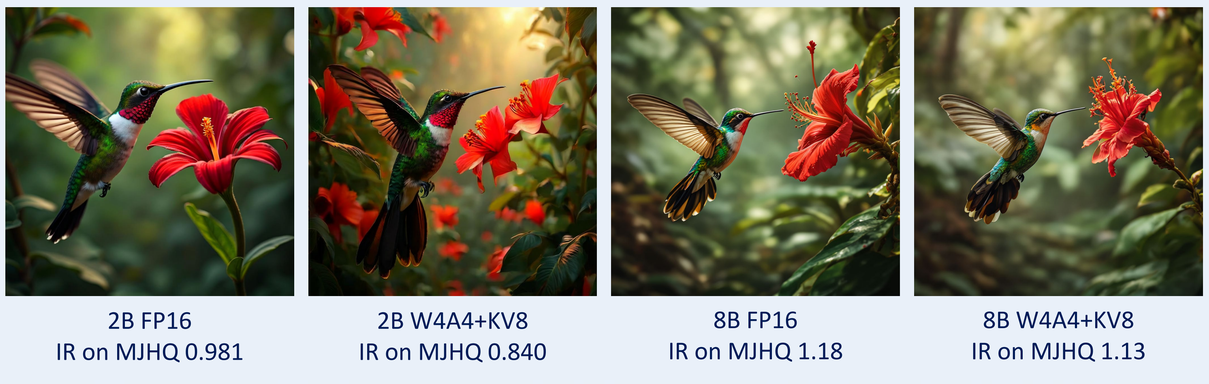


We choose per-channel quantization because the coefficient of variation (CV = $\frac{\sigma}{\mu}$) is much larger across channels than across tokens. Asymmetric, because some channels have high skewness (γ = 11.6).

Cache Type	Model	Coefficient of Variation CV Channel	CV Token
Key Cache	2B	**0.28**	0.07
	8B	**0.31**	0.06
Value Cache	2B	**0.25**	0.17
	8B	**0.20**	0.17

Cache quantization has a negligible effect on generative quality (IR 1.129 → 1.128 for Infinity 8B).

Compression Results



2B FP16 IR on MJHQ 0.981 | 2B W4A4+KV8 IR on MJHQ 0.840 | 8B FP16 IR on MJHQ 1.18 | 8B W4A4+KV8 IR on MJHQ 1.13

Model	Method	Peak Memory	Reduction vs FP16	Feasible Hardware
Infinity 2B (2 Billion Params)	FP16	16.0 GB	–	AGX Orin (32GB)
	W4A4	12.3 GB	↓ 23%	Orin NX (16GB)
	W4A4+KV8	7.7 GB	↓ 52%	Orin Nano (8GB)
Infinity 8B (8 Billion Params)	FP16	37.1 GB	–	AGX Orin (64GB)
	W4A4	23.6 GB	↓ 36%	AGX Orin (32GB)
	W4A4+KV8	13.3 GB	↓ 64%	Orin NX (16GB)

64 %
Footprint reduction for the 8B model

70 %
Hardware cost reduction (Orin NX (16GB) vs AGX Orin 64GB)

Only 0.05 IR Loss
Quality preserved for Infinity 8B

Comparison with SOTA diffusion model: FLUX.1-dev

Metric	Flux FP16	Flux W4A4	Infinity 8B W4A4+KV8
Image Reward	0.953	0.935	1.13
CLIP Score	26.0	25.8	27.0
Latency	246 s	112 s	27 s
Memory	33.8 GB	11.8 GB	13.3 GB

Prompt 1: *Portrait, photograph, Canon 5D, magazine, editorial, full profile shot, photorealism, Annie Leibovitz, middle-aged man, realistic, accurate*

Prompt 2: *detailed close-up photograph of a small shrine against a solid black background. A weathered stone statue in the center, surrounded by fresh colorful flowers, several burning candles casting warm light, and fruit wrapped in clear cellophane plastic. Macro lens, highly textured, cinematic lighting, 8K*

4.1x
Faster than FLUX.1-dev (W4A4)

IR +0.195
Higher generation quality than FLUX.1-dev (W4A4)

Paper Link

GitHub Link