BitMar: Low-Bit Multimodal Fusion with Episodic Memory for Edge Devices

Euhid Aman
NTUST Taiwan
M11315803@mail.ntust.edu.tw

Esteban Carlin
NTUST Taiwan
M11302809@mail.ntust.edu.tw

Hsing-Kuo Pao
NTUST Taiwan
pao@mail.ntust.edu.tw

Giovanni Beltrame
Polytechnique Montréal
giovanni.beltrame@polymtl.ca

Ghaluh Indah Permata Sari
NTUST Taiwan
d11115804@mail.ntust.edu.tw

Yie-Tarng Chen
NTUST Taiwan
ytchen@mail.ntust.edu.tw

Abstract

Cross-attention transformers and other multimodal vision-language models excel at grounding and generation; however, their extensive, full-precision backbones make it challenging to deploy them on edge devices. Memory-augmented architectures enhance the utilization of past context; however, most works rarely pair them with aggressive edge-oriented quantization. We introduce BitMar, a quantized multimodal transformer that proposes an external human-like episodic memory for effective image-text generation on hardware with limited resources. BitMar utilizes 1.58-bit encoders, one for text (BitNet-style) and one for vision (DiNOv2-based), to create compact embeddings that are combined and used to query a fixed-size key-value episodic memory. During vector retrieval, the BitNet decoder applies per-layer conditioning, which increases the contextual relevance of generated content. The decoder also employs attention sinks with a sliding-window mechanism to process long or streaming inputs under tight memory budgets. The combination of per-layer conditioning and sliding-window attention achieves a strong quality–speed trade–off, delivering competitive captioning and multimodal understanding at low latency with a small model footprint. These characteristics make BitMar well-suited for edge deployment.

Keywords: TinyVLM, Episodic memory, EdgeAI, Quantization.

1 Introduction

Visual Language Models (VLMs) have made rapid progress in recent years, excelling at tasks such as image captioning (Chen et al., 2015), visual question answering (Anderson et al., 2018; Li et al., 2022). Large-scale architectures such as BLIP-2 (Li et al., 2023), Flamingo (Alayrac et al., 2022), and Kosmos-2 (Peng et al., 2023) demonstrate that cross-attention transformers can synchronize modalities for grounded language generation. However, their full-precision, extensive backbones incur significant computational and memory expenses, which restricts their implementation on devices with resource limitations.

A growing body of work targets efficient multimodal processing, such as low-bit quantization (Dettmers et al., 2021; Frantar et al., 2022) and compact language models (Wang et al., 2023), to reduce memory/latency. Quantized ViTs (Jacob et al., 2018; Stock et al., 2019), and self-supervised vision encoders, such as DiNOv2 (Oquab et al., 2024), lower the cost of vision. Multimodal fusion ranges from early concatenation (Lu et al., 2019) to learned query transformers (Li et al., 2023) to bridge frozen vision and language models. Memory-augmented transformers (Graves et al., 2016; Borgeaud et al., 2022) retrieve past context to improve coherence. Yet no existing tiny language model effectively unifies low-bit multimodal encoding with an episodic memory system for edge deployment.

To fill this gap, we propose a compact four-stage pipeline optimized for efficient on-device execution: (1) **1.58-bit text and vision encoders** generate lightweight, quantized embeddings; (2) **a cross-modal fusion module** aligns the modalities within a shared latent space; (3) **an episodic memory** with 512 key–value slots retrieves relevant multimodal context; and (4) **a BitNet-based decoder** conditions each transformer layer on the retrieved memory for context-aware generation. Both encoders output 128-dimensional representations, and DiNOv2's original 768-D vision features are compressed to 128-D before fusion. The fused embedding queries an episodic memory of size $K = 512$, $C = 128$, whose retrieved vectors condition each decoder layer. This architecture maintains all modules in a consistent 768-dimensional space, simplifying integration and minimizing projection overhead while ensuring low-latency, memory-efficient

Proceedings of the First BabyLM Workshop, pages 147–154

operation on edge hardware.

Our main contributions are summarized as follows:

- **Low-bit multimodal encoding framework.** We propose a unified architecture that integrates a 1.58-bit quantized BitNet text encoder with a quantized ViT-based vision encoder, enabling efficient and compact multimodal feature extraction.
- **Memory-augmented decoding mechanism.** We design a lightweight episodic memory module that retrieves contextual representations and injects them into each transformer layer through per-layer conditioning, enhancing coherence and contextual relevance during generation.
- **Edge-efficient multimodal reasoning.** We demonstrate that BitMar achieves competitive performance in image captioning and multimodal understanding under extreme compression, maintaining low latency and a minimal memory footprint suitable for on-device deployment.

2 Related Work

Different VLMs and Tiny LLM architectures have emerged that enable deployment and applications of multimodal AI on resource-constrained devices. Recent developments in small VLMs, such as H2OVL-Mississippi (0.8B parameters) (Galib et al., 2024), TinyGPT-V (Yuan et al., 2024), and MiniCPM-V (Yao et al., 2024), demonstrate that compact multimodal models can achieve competitive performance while maintaining efficient deployment characteristics. Similarly, Tiny LLMs, such as MobileLLM (Liu et al., 2024) and TinyLLM (Zhang et al., 2024), have shown that sub-billion parameter models can be quantized and optimized for their deployment on edge devices. These highlight the feasibility of on-device multimodal processing, with models providing meaningful performance while addressing security, latency, and connectivity constraints.

Furthermore, memory-augmented neural networks and language models inspired by cognitive thinking, such as humans, have also garnered significant attention for their ability to store and retrieve contextual information related to specific things across certain short periods of time. Memory-augmented neural networks (MANNs) (Graves et al., 2016), use decoupled key-value structures to store and retrieve contextual information. Recent works, such as EGO (Mattar and Daw, 2024) and selective episodic memory strategies (Mattar and Daw, 2022), have extended these ideas for flexible knowledge transfer and context-based memory access. However, these models face limitations in combining memory systems with low-bit quantized multimodal encoders, often sacrificing either memory capacity or model precision.

BitMar overcomes these challenges by integrating 1.58-bit quantization across text and vision encoders, alongside a cross-modal memory retrieval system. The design enables BitMar to store and retrieve both textual and visual context, improving memory interactions and enhancing multimodal generation tasks, all while maintaining computational efficiency for edge deployment.

3 Method

We introduce **BitMar**, a deployable quantized multimodal LM for efficient image–text generation under tight resources. The four-stage pipeline is: (1) **parallel low-bit text/vision encoders**; (2) **cross-modal fusion** in a shared latent space; (3) **context augmentation** via external episodic memory; (4) **autoregressive decoding** conditioned on fused and retrieved signals. Text uses a BitNet transformer at 1.58-bit precision; vision uses DiNOv2 features plus quantization-aware compression. Fusion aligns 768-D modality latents via lightweight attention. A fixed-size episodic memory stores prior multimodal contexts and injects retrieved vectors into the decoder per layer. Unlike classic MANNs (Graves et al., 2016), BitMar integrates cross-modal retrieval under low-bit constraints. The decoder is a BitNet-based autoregressive transformer with streaming attention via attention sinks for low-latency, long-context generation.

3.1 Text Encoders

Architecture. A 4-layer quantized Transformer (d=128, h=4) supports up to 256 tokens, balancing expressiveness and efficiency.

Quantization. *Weights:* all MHSA/FFN projections use ternary $\{-1, 0, +1\}$ with learned per-layer scales (1.58-bit). *Activations:* token-wise 8-bit using per-token max-abs scaling to $[-127, 127]$, preserving local detail and stable training/inference.

Attention sinks (streaming). With S=4 sink

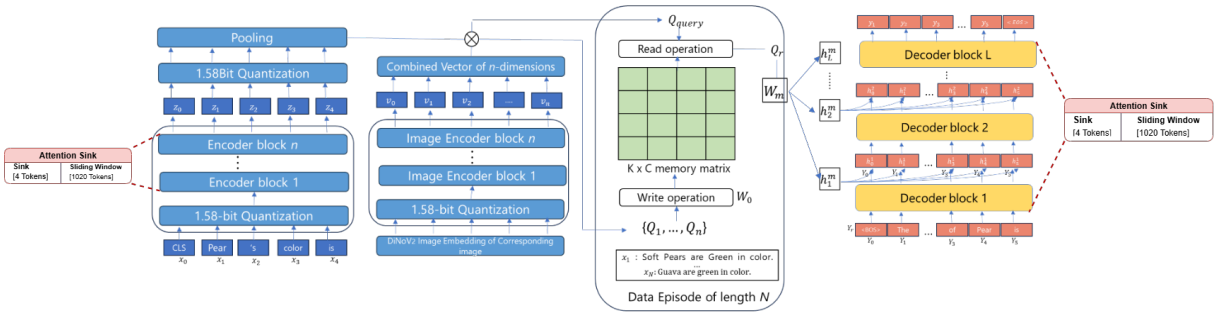


Figure 1: **BitMar Architecture.** The model processes multimodal inputs: text tokens and DiNOv2-compressed image features. Quantized encoders (1.58-bit) generate compact text and vision embeddings (z, v), which are fused via cross-modal attention into shared query representations (Q, Q_{query}). A sliding-window attention mechanism enables long-context processing. A fixed episodic memory matrix ($K \times C$) stores and retrieves multimodal context vectors through quantized read/write weights (W, W_0), supporting optional SD-card offloading for edge deployment.

tokens (never evicted) and window W=1020, the KV cache maintains persistent anchors + recent tokens. On each new token, the oldest in-window token is evicted; sink and window sets are merged; positions are clamped to $[0, S+W-1]$. This yields fixed-memory, long-context attention under low-bit compute.

3.2 Vision Encoders

We use frozen DiNOv2 (Oquab et al., 2024) to extract 768-D patch features offline, avoiding heavy vision backbones at inference. 2×2 average pooling reduces the number of patches 4× while keeping 768-D per patch. 2-layer MLP bottleneck then compresses 768→128 with ReLU and dropout between layers (parameters $\mathbf{W_1} \in \mathbb{R}^{384\times768}$, $\mathbf{W_2} \in \mathbb{R}^{128\times384}$), all subsequent fusion/memory/decoder paths operate in 128-D.

3.3 Cross-Modal Fusion

Given pooled text tokens $\mathbf{Z} \in \mathbb{R}^{n_t\times128}$ and vision tokens $\mathbf{V}_{\text{img}} \in \mathbb{R}^{n_v\times128}$, we apply standard cross-attention (Vaswani et al., 2017) (text queries, vision keys/values; cf. Transformer attention) to obtain the fused sequence $\mathbf{F} \in \mathbb{R}^{n_t\times128}$. All $Q/K/V$ and fusion projections use 1.58-bit ternary weights with learned scales; softmax and residual/LN are in FP32. We then pool $\mathbf{F}$ (mean or learned) to a single vector $\mathbf{q}_{\text{mem}} \in \mathbb{R}^{128}$ to query episodic memory.

3.4 Episodic Memory

We maintain a learnable matrix $\mathbf{M} \in \mathbb{R}^{K\times C}$ (default K=512, C=128) that stores multimodal episode vectors.

Writing. At step t, we compute a pooled query $\mathbf{q}_t \in \mathbb{R}^C$ and learned write weights $\mathbf{W}_w \in \mathbb{R}^K$. We perform soft multi-slot writes with rate α=0.2 via an outer product:

$$\mathbf{M} \leftarrow \mathbf{M} + \alpha\, \mathbf{W}_w\, \mathbf{q}_t^\top. \tag{1}$$

Reading. We use content-based addressing (Graves et al., 2016):

$$\begin{aligned} \mathbf{W}_r &= \text{softmax}(\mathbf{M}\,\mathbf{q}_t) \in \mathbb{R}^K, \\ \mathbf{M}_r &= \mathbf{W}_r^\top \mathbf{M} \in \mathbb{R}^{1\times C}. \end{aligned} \tag{2}$$

Regularization. To avoid thrashing, we penalize abrupt updates to the store with a Frobenius penalty, $\mathcal{L}_{\text{reg}} = \lambda \left\|\Delta\mathbf{M}\right\|_F^2$, $\Delta\mathbf{M} := \mathbf{M}^{(t)} - \mathbf{M}^{(t-1)}$. We additionally apply usage-based forgetting to down-weight stale slots.

3.4.1 Decoder with Attention Sinks

A 4-layer causal Transformer (d=128, h=4, max length 256) conditions on fused inputs and retrieved memory.

Long-context generation. Each layer, similarly as the text encoder, maintains KV caches of S sink tokens and a window of W recent tokens.

Memory integration. $\mathbf{M_r} \in \mathbb{R}^{1\times128}$ is projected and combined with token embeddings via either concatenation $[x_t; \mathbf{M_r}]$ (then projected) or residual addition x_t+$\mathbf{M_r}$.

Output projection. BitNet-quantized linear layer (128→50,257) maps to GPT-2 vocab logits; logits computed in FP32.

3.4.2 Training Objectives

We complement standard Language Modeling cross-entropy (Vaswani et al., 2017) and an InfoNCE cross-modal (Oord et al., 2018) term with a memory-consistency regularizer Equation 3 that

penalizes changes between successive writes to the episodic store, which discourages oscillatory updates and helps retain slot semantics. The total loss integrates these factors as Equation 4. We set $\mathcal{L}_{\text{cm}} = 1.5$ to prioritize cross-modal alignment, and $\mathcal{L}_{\text{mem}} = 0.1$ as a light stabilizer.

Memory consistency.

$$\mathcal{L}_{\text{mem}} = |\mathbf{M}_{\text{write}}^{(t)} - \mathbf{M}_{\text{write}}^{(t-1)}|_2^2 \tag{3}$$

Total objective.

$$\mathcal{L} = \mathcal{L}_{\text{lm}} + 1.5\mathcal{L}_{\text{cm}} + 0.1\mathcal{L}_{\text{mem}} \tag{4}$$

Adaptive Training Controller. When a 200-step EMA of cross-modal cosine similarity drops by > 0.12 from its recent max (with an $\geq$800-step cooldown), we randomly freeze one encoder or upweight $\mathcal{L}_{\text{cm}}$ for 1,500 steps to prevent modality collapse.

4 Experimental Setup

Our experimental framework systematically evaluates the proposed 14M-parameter BitMar model across several critical dimensions. We first benchmark its performance against established compact and low-bit baselines to assess overall viability (Table 1). We then conduct an analysis of its capabilities across a suite of language understanding and multimodal tasks to identify specific strengths and limitations (Table 2). Beyond task performance, we also investigate the internal dynamics of the model, examining how the episodic memory evolves from diffuse to structured activation patterns during training (Figure 2). Finally, we track the progression of quantization efficacy throughout the training process to validate our low-precision approach (Figure 3).

4.1 Dataset

The corpus comprises 100M tokens, split evenly between multimodal captions and text-only data.

Multimodal (50M). From CC3M (Sharma et al., 2018) and Localized Narratives (Pont-Tuset et al., 2020), aligned with precomputed DiNOv2 features (frozen backbone, reused across training).

Text-only (50M). From BabyLM (Charpentier et al., 2025), spanning six domains (BNC, CHILDES, Gutenberg, OpenSubtitles, Simple English Wikipedia, Switchboard).

Mixture. Uniform 50:50 sampling; a 1M-token hold-out tracks cross-modal alignment (cosine similarity) and perplexity.

Preprocessing. GPT-2 BPE tokenizer, max 256 tokens (truncate/pad). Visual features stored as memory-mapped ".npy" with on-the-fly compression for efficient batching.

4.2 Training Configuration

We trained on an NVIDIA A6000 GPU using FP16 and gradient checkpointing. Each step processed 64 sequences, with two-step gradient accumulation yielding an effective batch size of 128. Optimization used AdamW8bit (2×10^{-4}) with cosine restarts (T_0=1000, T_{mult}=2, η_{min}=0.1lr) for 10 epochs. We logged to Weights & Biases every 500 steps, including losses ($\mathcal{L}_{\text{lm}}, \mathcal{L}_{\text{cm}}, \mathcal{L}_{\text{mem}}$), cross-modal alignment metrics, episodic-memory utilization, attention maps, and FLOPs per step.

4.3 Hyperparameters

The model architecture employs a four-layer text encoder with 128-dimensional hidden states. The episodic memory module comprises 512 slots, each with 128 dimensions, balancing memory footprint with recall capacity. For long-context streaming, we maintain four sink tokens with a sliding window of 1020 tokens. Training utilizes weighted losses with cross-modal and memory consistency coefficients of 1.5 and 0.1, respectively. An adaptive controller triggers memory freezing when alignment metrics drop by 0.12 from their recent maximum, applying 1,500-step freezes with a minimum interval of 800 steps between interventions.

4.4 Benchmarks and Baselines

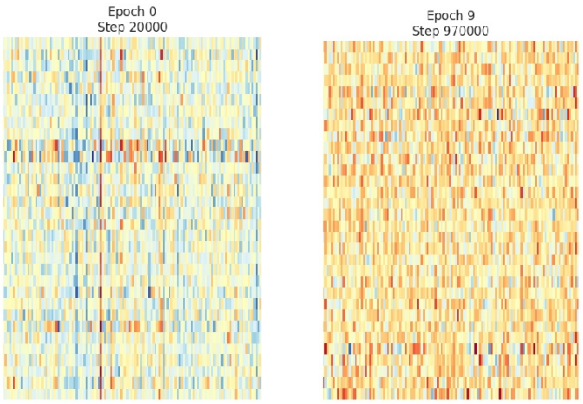
We evaluate on six language benchmarks: *ARC-Easy*, *BoolQ*, *HellaSwag*, *WinoGrande*, *CommonsenseQA*, and *MMLU*, plus multimodal tasks aligned with DiNOv2 features. Outputs are evaluated by accuracy and compared against baselines (*Bonsai 0.5B*, *OLMo-BitNet 1B*, *Falcon3-1.58bit 7B*, *LLaMA3-8B-1.58*, and *BitNet b1.58 2B*). Beyond benchmarks, we track the effectiveness of quantization and episodic activations to assess representational efficiency and memory use.

5 Results and Discussion

5.1 BitMar's performance

Figure 2 shows episodic memory slot activations over training. Early on Figure 2(a), activations are weak and scattered, with minor specialization or proper storage. By late training Figure 2(b), activations strengthen and differentiate, indicating se-

lective storage of contextual features. This progression demonstrates that extended joint optimization enables the memory to evolve into a more structured, capacity-efficient component for long-term context integration.



(a) (b)

Figure 2: **Episodic Memory Activation Patterns.** (a) Early training shows scattered and weak activations with minimal specialization. (b) Late training exhibits stronger and more differentiated activations, reflecting the emergence of structured memory representations.

We measure the quantization effectiveness E_q, inspired by (Zhu et al., 2016), as the zero-weight fraction in ternary weights across BitNet-quantized layers, where a higher value means more compression.

As training progresses (Figure 3), E_q gradually increases and stabilizes at 42.8%, demonstrating effective compression without degrading downstream performance.

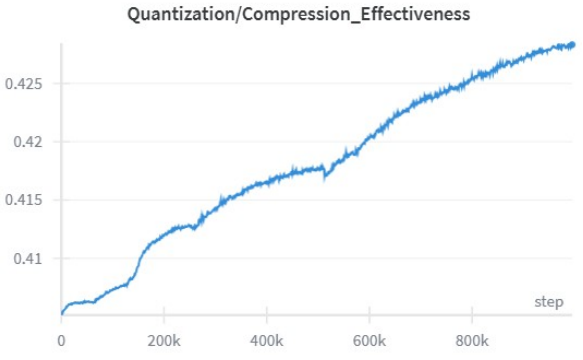


Figure 3: **Quantization effectiveness over training epochs.**

Table 1 compares BitMar-14M with low-bit baselines. Despite its small size (14M parameters), it achieves competitive performance on *BoolQ* (42.8) and *WinoGrande* (54.6), demonstrating strength in binary reasoning and coreference. On *ARC-Easy* (28.3) and *HellaSwag* (30.0), it lags larger models, reflecting limits in multi-step reasoning. *CommonsenseQA* (24.6) and *MMLU* (27.9) remain challenging due to restricted factual coverage. Still, BitMar achieves non-trivial accuracy across all tasks, confirming that extreme compression can yield usable models for targeted workloads, though with expected trade-offs in knowledge-heavy benchmarks.

As shown in Table 2, BitMar achieves an average 60.5% across finetuned NLP benchmarks, with strong results on paraphrase (*QQP*: 70.2%, *MRPC*: 69.1%) and reading comprehension (*BoolQ*: 66.5%), but weaker performance on inference (*MNLI*: 42.3%, *RTE*: 54.0%). Multimodal tasks yield modest scores (21–25%), with the best results on *EWoK* (24.9%), likely benefiting from episodic memory. Linguistic analysis shows reasonable syntax (*BLIMP*: 48.7%) and compositional reasoning (51.5%), but poor morphological productivity (*WUG*: -0.16/-0.22). Overall, BitMar balances extreme efficiency with usable performance, excelling in lightweight reasoning while struggling on complex multimodal and morphological tasks.

5.2 Ablation Study: Episodic Memory

Evaluated under BabyLM 2025 evaluation pipeline (same as Table 2).

Efficiency. As Table 3 reports, a fixed retrieved vector supplies context each step, reducing long-range attention while keeping 1.58-bit compute.

Zero-shot accuracy in Δ (pp). Table 4 reports the performance differences on zero-shot tasks. Overall, the results suggest that incorporating additional contextual information generally enhances task accuracy.

Regressions. We observe two regressions. First, regarding *WUG* morphology, correlations are negative, -0.36 for adjectives and -0.16 for past tense, indicating reduced morphological productivity under extreme quantization. Second, *reading* alignment scores are lower with memory (0.44/0.11) than without (1.11/0.66), suggesting that episodic conditioning can dampen psycholinguistic alignment. Tuning memory capacity or injection strategy may mitigate this.

Fine-tuning. No significant changes on *BoolQ/MNLI/MRPC/MultiRC/QQP/RTE/WSC*, suggesting memory mainly affects generation, not supervised heads.

Model	Native 1-bit	ARC-Easy	BoolQ	HellaSwag	WG	CQA	MMLU
Bonsai 0.5B	✓	58.25	58.44	48.01	54.46	18.43	25.74
OLMo-BitNet 1B	✓	25.38	52.48	25.88	51.54	19.49	25.47
Falcon3-1.58bit 7B	×	65.03	72.14	59.46	60.14	67.08	42.79
LLaMA3-8B-1.58 8B	×	70.71	68.38	68.56	60.93	28.50	35.04
BitNet b1.58 2B	✓	74.79	80.18	68.44	71.90	71.58	53.17
BitMar-14M (Ours)	✓	28.32	42.83	30.04	54.57	24.57	27.90

Table 1: **Benchmark performance on language understanding tasks.** A ✓indicates models trained natively with 1-bit precision. All reported values correspond to task accuracy (%), illustrating BitMar's competitive performance under extreme compression. [WG-WinoGrande; CQA-CommonsenseQA]

Category	Task	Primary Metric	Score
Finetune NLP	BoolQ	Accuracy	66.5%
	MNLI	Accuracy	42.3%
	MRPC	Accuracy	69.1%
	MultiRC	Accuracy	57.6%
	QQP	Accuracy	70.2%
	RTE	Accuracy	54.0%
	WSC	Accuracy	63.5%
Multimodal	DevBench	Visual Vocab Acc.	21.2%
	VQA	Accuracy	21.4%
	Winoground	Accuracy	23.8%
World Knowledge	EWOK	Accuracy	24.9%
Linguistic	BLIMP	Accuracy	48.7%
Reasoning	Compositional	Accuracy	51.5%
	Entity Tracking	Accuracy	31.2%
Psycholing.	Reading Comp.	Score	0.44
Morphology	Wug Adj.	Corr.	-0.16
	Wug Past	Corr.	-0.22

Table 2: BitMar results on BabyLM evaluation tasks.

Metric	Mem. On	Mem. Off
Throughput (tok/s)	57.3	7.7
Latency/token (ms)	17.3	129.8
Energy (J)	1.90	9.17
RAM (MB)	956	1,076

Table 3: **Inference ablation metrics.** Comparison of throughput, latency, energy consumption, and memory usage.

Task	Δ (pp)
Entity Tracking (Split 1)	+2.9
Entity Tracking (Split 2)	+4.1
COMPS	+3.4
BLiMP	+0.6
VQA	+3.4
EWoK (Split 1)	-1.6
EWoK (Split 2)	+1.0
Winoground	-1.6
DevBench	No effect

Table 4: **Ablation results on episodic memory.** Performance differences (Δ, in percentage points), positive values indicate improvements when memory is enabled.

Ablation Summary. Episodic Memory is $\sim$7.5$\times$ faster, using 79% less energy and 11% less VRAM in our tests. It delivers 3 – 4 percentage point gains on entity/property reasoning and multimodal QA, though morphology and some psycholinguistic alignment metrics can degrade. Overall, combining attention sinks with episodic memory enables efficient long-context use under tight resource budgets.

6 Conclusion

BitMar-14M is a compact 1.58-bit multimodal language model using BitNet quantization, DiNOv2 vision compression, cross-modal fusion, an attention-sink decoder for efficient long-context reasoning, and an external episodic latent memory for deployment on resource-constrained edge devices. With adaptive training, it maintains stable alignment and memory use despite its tiny size. Though less accurate than larger low-bit models on knowledge-heavy tasks, it performs competitively on binary reasoning and coreference, showing that 1.58-bit compression and efficient design can enable multimodal reasoning with drastically reduced compute and storage.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, Los Alamitos, CA, USA. IEEE Computer Society.

Antoine Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Sebastian Rutherford, Matthew Botvinick, Jean-Baptiste Sifre, and Stan Clark. 2022. Improving language models by retrieving from trillions of tokens. *International Conference on Machine Learning (ICML)*, 162:2209–2226.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. BabyLM turns 3: Call for papers for the 2025 BabyLM workshop. In *BabyLM Turns 3: Call for papers for the 2025 BabyLM workshop*, pages 2–3, Suzhou, China.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *CoRR*, abs/2110.02861.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323.

Shaikat Galib, Shanshan Wang, Guanshuo Xu, Pascal Pfeiffer, Ryan Chesler, Mark Landry, and Sri Satish Ambati. 2024. H2ovl-mississippi vision language models technical report.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. 2024. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *arXiv preprint arXiv:2402.14905*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. *ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*. Curran Associates Inc., Red Hook, NY, USA.

Marcelo G. Mattar and Nathaniel D. Daw. 2022. A neural network model of when to retrieve and encode episodic memories. *eLife*, 11:e74445.

Marcelo G. Mattar and Nathaniel D. Daw. 2024. Toward the emergence of intelligent control: Episodic generalization and optimization. *Open Mind*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*. ArXiv: 1807.03748.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, page 647–664, Berlin, Heidelberg. Springer-Verlag.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, and Hervé Jégou. 2019. And the bit goes down: Revisiting the quantization of neural networks. *CoRR*, abs/1907.05686.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. 2023. Bitnet: Scaling 1-bit transformers for large language models.

Yuanhan Yao, Qinghao Yu, Ao Zhang, Xiaoyi Wang, Zhiyang Xu, Chendong Yuan, Ying Wang, Yaoyao Liu, Kunchang Wang, Yunhai Yu, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhengqing Yuan, Zhaoxu Ren, Lichao Feng, Zhi Zhao, Kai Cui, and Shiliang Jiang. 2024. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862*.

Wei Zhang, Xiaoming Liu, Hao Chen, and Yifan Wang. 2024. Tinyllm: A framework for training and deploying language models at the edge computers. *arXiv preprint arXiv:2412.15304*.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J. Dally. 2016. Trained ternary quantization. *CoRR*, abs/1612.01064.