EFFICIENT UNIFIED MULTIMODAL UNDERSTANDING AND GENERATION WITH GATED HYBRID ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal learning requires attention mechanisms that are both efficient and expressive. Softmax attention provides strong modeling capacity but suffers from quadratic complexity, while linear attention achieves near-linear efficiency at the cost of weaker expressivity. We identify two major expressivity challenges in efficient unified multimodal models: (*i*) modality imbalance, where dominant signals suppress weaker modalities during fusion, and (*ii*) loss of global context, as efficient variants tend to over-smooth long sequences. We propose **Gated Hybrid Attention (GHA)**, a multimodal-specialized operator that augments linear attention with (*i*) a selective gating mechanism to balance modality contributions and stabilize training, and (*ii*) agent-token softmax aggregation to restore adaptive global context while preserving near-linear complexity. To demonstrate generality, we validate GHA in two representative paradigms: autoregressive-only(AR-only) and autoregressive+diffusion(AR+Diffusion). In both settings, GHA consistently improves multimodal alignment, long-context retention, and efficiency over comparable Transformer and efficient attention baselines. These cross-paradigm results highlight that GHA functions as a plug-and-play building block, offering a lightweight and extensible approach that is orthogonal to scaling trends and modality complexity.

1 INTRODUCTION

Recent advances in Large Language Models (LLMs) Vaswani et al. (2017); Brown et al. (2020); Chowdhery et al. (2022); OpenAI et al. (2024) have spurred interest in extending their capabilities beyond text, leading to the emergence of Multimodal Large Language Models (MLLMs) Caffagni et al. (2024); Wang et al. (2024a); Zhang et al. (2024). Early efforts such as Flamingo Alayrac et al. (2022), GIT Wang et al. (2022a), and OFA Wang et al. (2022b) demonstrated the feasibility of joint vision–language modeling but were often specialized for narrow tasks, motivating the development of more unified architectures. Recent unified systems including Show-o Xie et al. (2024a), JanusFlow Ma et al. (2025), Janus-Pro Chen et al. (2025) and VILA-U Wu et al. (2025) extend Transformer-based pipelines to support both multimodal understanding and generation. However, these models inherit the quadratic complexity of self-attention, which poses scalability bottlenecks for high-resolution inputs and long multimodal sequences.

An alternative line of work replaces Transformers with state-space architectures such as Mamba, exemplified by OmniMamba Zou et al. (2025). While these approaches achieve linear efficiency in both computation and memory, their strictly sequential modeling and limited cross-layer interactions constrain the ability to capture complex long-range dependencies, resulting in weaker expressivity compared to Transformer counterparts.

In contrast, our work addresses the complementary challenge of balancing efficiency and expressivity through architectural design. We introduce *Gated Hybrid Attention (GHA)*, a lightweight and theoretically grounded drop-in substitute for standard Transformer attention, tailored for unified multimodal models. Unlike approaches that rely on large-scale data scaling or specialized backbones, GHA directly enhances the attention mechanism itself, extending linear attention with two key components: (i) a gating mechanism that stabilizes optimization and alleviates modality

imbalance, and (ii) a softmax-based agent-token bottleneck that reintroduces global context while maintaining near-linear complexity. This architectural advance complements large-scale scaling and complex unified backbones, offering a practical direction toward building multimodal systems that are both efficient and expressive.

Our contributions can be summarized as follows:

- **Gated Hybrid Attention.** We introduce GHA, which extends linear attention with a key–value gating module and a softmax-based agent-token bottleneck. This design significantly enhances expressivity, long-context reasoning, and cross-modal integration, while preserving near-linear computational complexity.
- **Hardware-efficient Triton Kernel.** To ensure the theoretical efficiency of GHA translates into practice, we provide a hardware-optimized implementation based on FlashAttention-style chunking and a custom Triton kernel.
- **Empirical Evaluation.** We evaluate GHA on both AR-only and AR+Diffusion unified multimodal pipelines. Results show consistent gains in multimodal alignment, long-context retention, and generation quality compared to Transformer and efficient attention baselines. These cross-paradigm results highlight that GHA functions as a plug-and-play building block, complementary to large-scale scaling and specialized backbones.

2 Related Work

2.1 Unified Multimodal Models

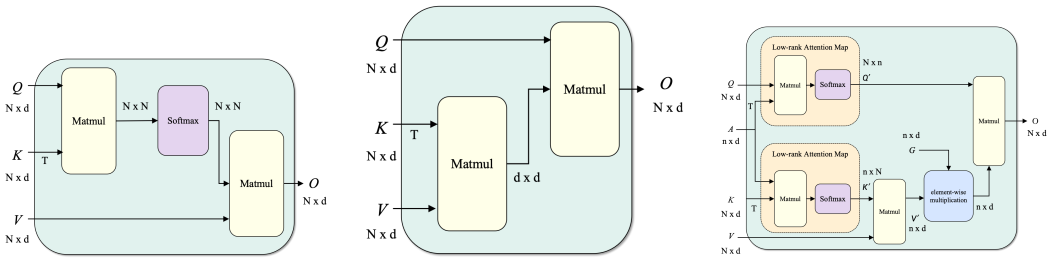
Early frameworks such as Flamingo, GIT, and OFA demonstrated the feasibility of coupling vision and language but were limited to narrow task formats. Subsequent research has sought fully unified architectures that support both multimodal understanding and generation within a single backbone. As summarized in Show-o Xie et al. (2024a), existing systems fall into two paradigms: *AR-only*, where all tasks are cast as next-token prediction (e.g., Emu3 Wang et al. (2024c), VILA-U Wu et al. (2025)); and *AR+Diffusion*, where an AR backbone is augmented with a discrete diffusion module for higher-fidelity generation (e.g., Show-o Xie et al. (2024a), JanusFlow Ma et al. (2025), Janus-Pro Chen et al. (2025)). Concurrent works such as Show-o2 Xie et al. (2025) and BAGEL Deng et al. (2025) push performance further by scaling data and adopting increasingly sophisticated multimodal backbones (e.g., decoupled visual encoders, 3D causal latent spaces, mixture-of-transformer-experts). While these systems substantially improve multimodal generation quality, they also reveal persistent issues of cross-modal imbalance, where dominant modalities (e.g., text) suppress weaker ones during fusion. Moreover, most progress has come from scaling data, model size, and backbone complexity, leaving the core attention operator largely unchanged.

2.2 Efficient Sequence Modeling

To mitigate the quadratic cost of attention, linear variants (e.g., Linear Transformer Katharopoulos et al. (2020), Performer Choromanski et al. (2020), Reformer Kitaev et al. (2020)) approximate softmax kernels or restrict interactions, reducing complexity to linear time and memory. In parallel, state-space models such as Mamba Gu & Dao (2024) and Mamba-2 Dao & Gu (2024) achieve similar efficiency through recurrent parameterizations, while hybrid designs like Jamba Datta et al. (2024) combine state-space recurrence with sparse attention. While these architectures excel in efficiency, they often struggle to maintain global context in long sequences and thus sacrifice expressivity. In particular, approximations in linear attention can lead to over-smoothing, while strictly recurrent models limit cross-token interactions across layers. As a result, efficient sequence modeling provides strong scalability but remains insufficient for tasks that require rich long-range dependencies.

2.3 Expressivity Enhancements

Several works aim to strengthen attention capacity. Gated designs such as GLA Yang et al. (2024) and GTrXL Parisotto et al. (2019) improve stability and selectivity by modulating key–value or residual pathways. Another family of approaches strengthens expressivity through *agent tokens*,



(a) Softmax Attention (b) Linear Attention (c) Gated Hybrid Attention

Figure 1: Difference between softmax attention ($\mathcal{O}(N^2d)$), linear attention ($\mathcal{O}(Nd^2)$), and GHA ($\mathcal{O}(N\,n\,d)$).

where a small set of latent units mediate communication between input tokens. Representative examples include Set Transformer Lee et al. (2019) with inducing points, Perceiver and Perceiver IO Jaegle et al. (2021; 2022) with latent bottleneck arrays, and Agent Attention Wang et al. (2024b) with explicitly defined agent tokens. These agents compress token interactions into a compact latent space and then redistribute information back to the sequence, typically using content-dependent weighting such as softmax aggregation. While effective, such designs have been studied mainly in unimodal contexts and remain underexplored in unified multimodal systems. Overall, unified multimodal models have advanced primarily through scaling data and backbone complexity, with some architectural innovations targeting specialized multimodal components. In parallel, efficient sequence modeling has been explored mostly in unimodal contexts and remains rarely applied to unified multimodal systems. As a result, the attention operator at the core of unified models is still underexplored, facing persistent challenges of cross-modal imbalance (dominant modalities suppressing weaker ones), loss of global context in long sequences, and quadratic computational complexity that limits scalability. To address these challenges, we propose GHA, which extends linear attention with gating and agent-token–based aggregation. This balances efficiency and expressivity, yielding improvements across both AR-only and AR+Diffusion paradigms. This architectural direction is complementary to scaling- and backbone-oriented advances and points toward future integration into more complex unified frameworks.

3 PRELIMINARY

Softmax Attention. Given queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, standard attention computes

$$\mathbf{O} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{1}$$

which provides strong expressivity but incurs $\mathcal{O}(N^2d)$ complexity for sequence length N and hidden dimension d.

Linear Attention. To alleviate this bottleneck, Linear Attention Katharopoulos et al. (2020) reformulates the exponential similarity $\exp(q^\top k)$ as a kernelized dot-product $\phi(q)^\top\phi(k)$, where $\phi(\cdot)$ is a positive feature mapping (e.g., $\phi(x) = \mathrm{ELU}(x) + 1$ or randomized Fourier features). This enables computing attention as

$$\mathbf{O} \;\approx\; \phi(\mathbf{Q})\big(\phi(\mathbf{K})^\top\mathbf{V}\big), \tag{2}$$

which reduces complexity to $\mathcal{O}(Nd^2)$ and admits an equivalent recurrent form for streaming updates. In practice, however, many implementations simplify $\phi(\cdot)$ to the identity mapping, which abandons strict kernel approximation but often yields more stable and competitive performance in downstream tasks Sun et al. (2023).

Our proposed GHA builds on Linear Attention by (i) incorporating a gating mechanism to dynamically filter memory updates, and (ii) reintroducing content-dependent softmax weighting through agent tokens, which aggregate and redistribute global context. These additions retain linear-time efficiency while substantially improving expressivity and stability in long-sequence modeling.

4 GATED HYBRID ATTENTION

As shown in Figure 1, we propose a three-stage hybrid architecture, termed GHA. This model augments linear attention through two key components: a selective gating mechanism applied to the key-value pathway and a softmax-based aggregation over agent tokens. The proposed design effectively addresses the two primary challenges in unified multimodal modeling: efficiency, attained by maintaining linear computational complexity, and expressivity, enhanced by the introduced agent-token aggregation and gating mechanism.

4.1 PARALLEL FORMULATION

(1) Agent-token based Softmax Aggregation. To avoid the quadratic query–key interactions, GHA introduces a small set of $n \ll N$ agent tokens $\mathbf{A} \in \mathbb{R}^{n\times d}$, which serve as an information bottleneck that aggregates signals from all tokens before redistributing them. All tokens send their query and key representations to these agents via softmax attention:

$$\mathbf{K}' = \mathrm{Softmax}(\mathbf{A}\mathbf{K}^\top) \in \mathbb{R}^{n\times N}, \qquad \mathbf{Q}' = \mathrm{Softmax}(\mathbf{Q}\mathbf{A}^\top) \in \mathbb{R}^{N\times n}. \tag{3}$$

Compared to standard softmax attention, our design preserves content-dependent weighting while reducing the complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(Nn)$. Compared to plain linear attention, it reintroduces softmax aggregation in a compressed latent space, thereby restoring adaptive global context modeling.

(2) Linear Accumulation. Following the practical form of linear attention, we adopt the identity mapping $\phi(x) = x$, which is widely used for its stability. During training, the attention output can be computed in parallel as

$$\mathbf{O} = \mathbf{Q}'\,(\mathbf{K}'\mathbf{V}), \tag{4}$$

here $(\mathbf{K}'\mathbf{V}) \in \mathbb{R}^{n\times d}$ and the computation costs $\mathcal{O}(Nnd)$. Compared to softmax attention, this accumulation is more efficient; in GHA we incorporate gating and agent-token based softmax aggregation to mitigate over-smoothing and stabilize long-sequence modeling.

(3) Selective Gating. To further stabilize training and suppress over-smoothing, GHA applies gating directly on the aggregated key–value pathway. Formally, we compute

$$\mathbf{O} = \mathbf{Q}'\,\big(\mathbf{G} \odot (\mathbf{K}'\mathbf{V})\big), \tag{5}$$

where $\mathbf{G} \in (0,1)^{n\times d}$ is a learnable, data-dependent gating matrix applied to the compressed key–value representation, and $\odot$ denotes element-wise multiplication. To make this data-dependent nature explicit and avoid interpreting $\mathbf{G}$ as a static parameter matrix, we write it as $\mathbf{G} = f(x)$, where $f(x)$ is an input-conditioned projection that produces the dynamic gating weights. This design keeps the parallel complexity at $\mathcal{O}(Nnd)$, while enabling selective control over memory updates. By filtering noisy or redundant updates, gating improves long-sequence stability and prevents modality dominance, a common challenge in unified multimodal training, thus yielding more fine-grained and robust cross-modal alignment.

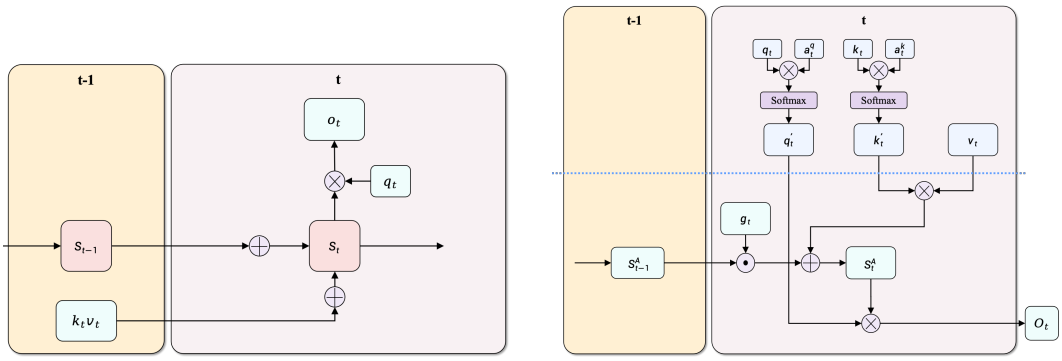
Together, these designs yield an overall complexity of $\mathcal{O}(Nnd)$ in theory, substantially lower than standard softmax attention ($\mathcal{O}(N^2d)$) while significantly enhancing the representational power of plain linear attention.

4.2 RECURRENT FORMULATION

For autoregressive decoding, GHA admits a recurrent formulation that achieves true linear-time complexity during inference as shown in Figure 2. At each time step t, the compressed key–value state $\mathbf{S}_t \in \mathbb{R}^{n\times d}$ is updated as

$$\mathbf{S}_t = \mathbf{g}_t \odot \mathbf{S}_{t-1} + \mathbf{k}'_t\mathbf{v}_t^\top, \qquad \mathbf{o}_t = \mathbf{q}'_t\mathbf{S}_t, \tag{6}$$

where $\mathbf{g}_t \in (0,1)^n$ is a token-wise gating vector (broadcast along the d dimension), $\mathbf{k}'_t, \mathbf{q}'_t \in \mathbb{R}^n$ are the agent-projected key and query (with q'_t and k'_t denoting the t-th row of Q' and the t-th column of K', respectively), and $\mathbf{v}_t \in \mathbb{R}^d$ is the value.



(a) Linear attention (recurrent inference). **(b)** GHA (recurrent inference).

Figure 2: Comparison of (a) linear attention (recurrent inference) and (b) GHA (recurrent inference).

Here $\mathbf{k}'_t \mathbf{v}_t^\top \in \mathbb{R}^{n \times d}$ matches the dimension of $\mathbf{S}_t$. Since each update depends only on $\mathbf{S}_{t-1}$ and the current token $(\mathbf{k}'_t, \mathbf{v}_t)$, this recurrent inference form scales linearly with sequence length as $\mathcal{O}(Nnd)$.

During training, causal masking prohibits direct use of the recurrent update. As with other linear attention mechanisms Katharopoulos et al. (2020); Yang et al. (2024); Ma et al. (2021), the masked parallel form requires handling prefix-restricted interactions, which breaks the associativity trick in linear accumulation and leads to quadratic $\mathcal{O}(N^2 d)$ complexity.

4.3 Hardware-efficient Implementation

Inspired by FlashAttention Dao et al. (2022) and FlashLinearAttention Yang et al. (2023), we design a GPU-friendly chunkwise execution strategy implemented in Triton to bridge the gap between theoretical efficiency and practical training.

Let the sequence of length N be partitioned into $L = N/C$ contiguous chunks of size C. Under causal masking, we treat *inter-chunk* and *intra-chunk* dependencies differently.

Notation. We denote by $t \in \{1, \ldots, N\}$ the *token index* along the input sequence of length N, and by $i \in \{0, \ldots, L-1\}$ the *chunk index* when the sequence is partitioned into $L = N/C$ contiguous chunks of size C. Accordingly, $\mathbf{S}_t$ refers to the recurrent state at token t in the inference (sequential) formulation, while $\mathbf{S}^{(i)}_{\text{start}}$ and $\mathbf{S}^{(i)}_{\text{end}}$ denote the carried states across chunks in the chunkwise (parallel training) formulation.

Inter-chunk dependencies (no mask). Across chunks, all tokens in chunk $i{+}1$ can fully attend to preceding chunks $\{0, \ldots, i\}$ without causal masking. We maintain a carried key–value state $\mathbf{S}^{(i)} \in \mathbb{R}^{n \times d}$ updated as

$$\mathbf{S}^{(i+1)} = \Gamma^{(i+1)} \odot \mathbf{S}^{(i)} + \Delta^{(i+1)}, \qquad \mathbf{O}^{(i+1)}_{\text{inter}} = \mathbf{Q}'[i{+}1]\, \mathbf{S}^{(i+1)}, \tag{7}$$

where $\Gamma^{(i+1)} \in (0,1)^{n \times d}$ encodes the cumulative decay across chunk $i{+}1$, and

$$\Delta^{(i+1)} = \sum_{t \in \text{chunk}(i+1)} \mathbf{k}'_t \mathbf{v}_t^\top \in \mathbb{R}^{n \times d}.$$

This component is fully parallelizable across chunks, with total cost $\mathcal{O}(Nnd)$.

Intra-chunk dependencies (causal mask). Within each chunk of size C, local causality is enforced by applying a lower-triangular mask $\mathbf{M} \in \{0,1\}^{C \times C}$ to restrict interactions to valid prefixes:

$$\mathbf{K}' = \text{Softmax}(\mathbf{A}\mathbf{K}^\top), \qquad \mathbf{Q}' = \text{Softmax}(\mathbf{Q}\mathbf{A}^\top). \tag{8}$$

$$\mathbf{O}_{\text{intra}} = \big(\mathbf{Q}'\mathbf{K}'^\top \odot \mathbf{M}\big)\mathbf{V}. \tag{9}$$

Naively, applying the causal mask incurs $\mathcal{O}(C^2 d)$ cost per chunk, as each token depends on all its predecessors.

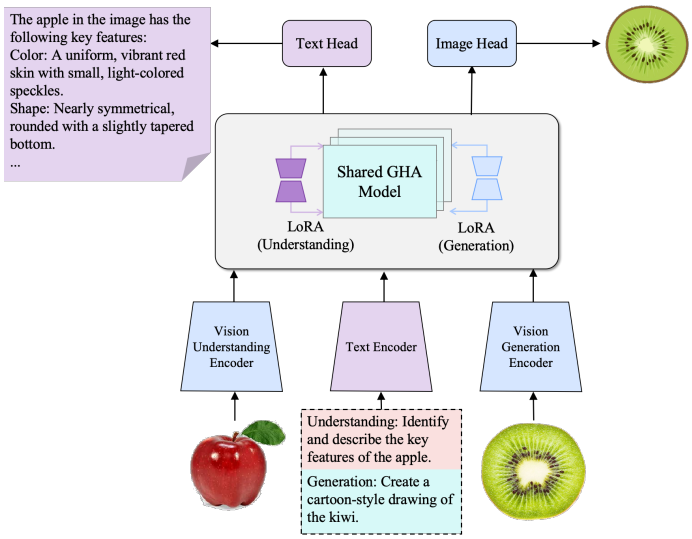


Figure 3: **Architecture of OmniGHA (AR-only).** The system processes three types of inputs: (i) visual features from the *vision understanding encoder*, (ii) visual tokens from the *vision generation encoder*, and (iii) text embeddings from the *text encoder*. All modalities are fed into a shared *Gated Hybrid Attention (GHA)* Transformer decoder, which serves as the core of the architecture. Two modality-specific output heads are employed: a *text head* for understanding tasks and an *image head* for generation tasks, respectively. Task-specific LoRA adapters are applied within the decoder to enable parameter-efficient, modality-specific adaptation while keeping the decoder weights largely shared across tasks.

Chunkwise combination. The final outputs combine both contributions:

$$\mathbf{O} = \sum_{i=0}^{L-1} \left(\mathbf{O}_{\text{inter}}^{(i)} + \mathbf{O}_{\text{intra}}^{(i)} \right). \tag{10}$$

Most FLOPs lie in the inter-chunk part (unmasked matmuls over the small agent dimension $n \ll N$), while the intra-chunk part is confined to local $C \times C$ blocks. This decomposition keeps the dominant work parallelizable and yields training that closely approaches the theoretical $\mathcal{O}(Nnd)$ scaling despite causal masking. The persistent state buffers scale as $\mathcal{O}(nd)$ per layer, with $\mathcal{O}(Cnd)$ temporary tiles during chunked training.

4.4 OmniGHA: Unified Multimodal Architecture

While GHA is a general attention mechanism, we demonstrate its effectiveness by instantiating it within a simplified unified multimodal pipeline, which we call *OmniGHA*. As shown in Figure 3, we illustrate OmniGHA in its basic AR-only instantiation for clarity. Frozen text and vision encoders provide embeddings, a shared GHA decoder performs multimodal fusion, and lightweight LoRA adapters and task-specific heads produce outputs. This design concentrates most parameters and compute in the shared backbone, highlighting the plug-and-play nature of GHA. A detailed breakdown of modules and the progressive training schedule is provided in Appendix A.1 and Appendix A.2, and we also describe an AR+Diffusion variant in Appendix A.3, where the shared decoder is coupled with a latent predictor and a diffusion-style generation head (e.g., flow matching). This confirms that GHA extends naturally to pipelines beyond AR-only.

To isolate the architectural contribution of GHA, we first evaluate it in the most basic AR-only setting, where frozen encoders feed embeddings into a single shared decoder and lightweight heads handle downstream tasks. This controlled setup deliberately excludes diffusion modules or sophisticated backbones, ensuring that observed gains can be attributed directly to GHA. We then extend OmniGHA to the AR+Diffusion paradigm by replacing the image head with a latent predictor and diffusion head, validating that GHA integrates seamlessly into more complex generation pipelines. The diffusion pathway is kept entirely unchanged to ensure a fair comparison. Across both paradigms, GHA delivers consistent improvements in multimodal alignment, long-context retention, and generation quality over strong Transformer baselines. These results demonstrate that GHA is

Type	Model	LLM Params	Res.	POPE↑	MME-P↑	$VQAv2_{test}$↑	GQA↑	MMMU↑
Und. Only	LLaVA-Phi Zhu et al. (2024)	Phi-2-2.7B	336	85.0	1335.1	71.4	-	-
	LLaVA Liu et al. (2024c)	Vicuna-7B	224	76.3	809.6	-	-	-
	Emu3-Chat Wang et al. (2024c)	8B from scratch	512	85.2	-	75.1	60.3	31.6
	LLaVA-v1.5 Liu et al. (2024b)	Vicuna-13B	448	86.3	1500.1	81.8	64.7	-
	InstructBLIP Dai et al. (2023)	Vicuna-13B	224	78.9	1212.8	-	49.5	-
	MobileVLM Chu et al. (2023)	MobileLLaMA-1.4B	336	84.5	1196.2	-	56.1	-
	MobileVLM-V2 Chu et al. (2024)	MobileLLaMA-1.4B	336	84.3	1302.8	-	59.3	-
	LLaVA-v1.5-Phi-1.5 Xie et al. (2024b)	Phi-1.5-1.3B	336	84.1	1128.0	75.3	56.5	30.7
Unified	LWM Liu et al. (2024a)	LLaMA2-7B	256	75.2	-	55.8	44.8	-
	Chameleon Team (2024)	7B from scratch	512	-	-	-	-	22.4
	LaVIT Jin et al.	7B from scratch	256	-	-	66.0	46.8	-
	Emu3 Wang et al. (2024c)	8B from scratch	512	85.2	1243.8	75.1	60.3	31.6
	Janus Wu et al. (2024)	DeepSeek-LLM-1.3B	384	87.0	1338.0	77.3	59.1	30.5
	JanusFlow Ma et al. (2024)	DeepSeek-LLM-1.3B	384	88.0	1333.1	79.8	60.3	29.3
	Show-o Xie et al. (2024b)	Phi-1.5-1.3B	512	80.0	1097.2	69.4	58.0	26.7
	OmniMamba Zou et al. (2025)	Mamba-2-1.3B	384	86.3	1290.6	77.7	60.8	30.6
	OmniGHA (AR)	1.3B from scratch	384	**88.7**	**1342.5**	**80.6**	**62.4**	**32.8**
	OmniGHA (AR+Diffusion)	1.3B from scratch	384	**89.6**	**1354.9**	**81.3**	**62.1**	**33.6**

Table 1: **Performance comparison on multimodal understanding benchmarks.** "Und. only" models are trained for understanding only; "Unified" models support both understanding and generation.

effective not only in controlled AR-only pipelines but also transferable to AR+Diffusion settings, providing a lightweight building block orthogonal to scaling trends and backbone complexity.

5 Experiments

5.1 Data and Training Setup

We train a 1.3B-parameter model from a scratch attention mechanism. Images are 384×384 for understanding and 256×256 for generation via a VQ-VAE tokenizer. We use BF16 AdamW and lightweight rank-8 LoRA on input projectors. Unified training follows a three-phase curriculum: (1) representation warm-up on understanding, (2) long-context adaptation on generation, and (3) joint alignment. We deliberately avoid sophisticated visual stacks to isolate the effect of replacing attention with GHA. Full datasets and hyperparameters are in Appendix A.4.

Recent concurrent works such as Omni-Qwen Xu et al. (2025), BAGEL, and Show-o2 advance unified multimodal learning primarily through larger-scale datasets and increasingly sophisticated backbones (e.g., causal 3D latent spaces, mixture-of-experts, or refined training strategies). Our work instead focuses on an orthogonal dimension: enhancing the attention operator itself. To clearly isolate this effect, we deliberately maintain a controlled evaluation setting rather than comparing against scaling- or backbone-heavy systems.

5.2 Evaluation Results

Multimodal Understanding. We follow Show-o, JanusFlow, and OmniMamba, and evaluate on standard multimodal understanding benchmarks, including POPE Li et al. (2023), MME Fu et al. (2024), VQAv2 Goyal et al. (2017), GQA Hudson & Manning (2019), and MMMU Yue et al. (2024). As shown in Table 1, OmniGHA consistently improves over models of similar scale ($\sim$1.3B), achieving 88.7 on POPE, 1342.5 on MME, 80.6 on VQAv2, 62.4 on GQA, and 32.8 on MMMU. It surpasses understanding-only systems such as LLaVA-v1.5-Phi-1.5 (84.1/1128.0/75.3/56.5/30.7) and MobileVLM-V2 (84.3/1302.8/–/59.3/–), while also outperforming unified baselines like Show-o (80.0/1097.2/69.4/58.0/26.7) and JanusFlow (88.0/1333.1/79.8/60.3/29.3), despite JanusFlow being trained on over 65M image–text pairs. Interestingly, the AR+Diffusion variant achieves slightly higher understanding scores, which we attribute to the additional conditioning projector acting as a form of representation regularization during joint training.

Visual Generation. We evaluate on MS-COCO Lin et al. (2014) and compare with both generation-only diffusion systems and unified multimodal models (Table 2). Large-scale diffusion models (e.g., DALL·E 2 Ramesh et al. (2022), Imagen Saharia et al. (2022)) progressively reduce

FID to the 7–10 range, with recent refinements such as U-ViT Bao et al. (2023) reaching 5.95. Among unified models, early approaches (CoDI, SEED-X, LWM) lag far behind, while Show-o narrows the gap (9.24). OmniMamba achieves 5.36 when trained on 35M pairs. Our OmniGHA further improves to 5.12, surpassing OmniMamba at the same scale. While the AR+Diffusion variant incurs additional cost, it still surpasses Show-o and JanusFlow in FID, demonstrating that GHA can transfer effectively to diffusion pipelines while retaining high generation quality.

Please note that our paper does not claim SOTA over all understanding- or generation-only models. The contribution lies in the GHA module—an attention operator that provides a more stable and efficient architecture under the unified multimodal setting. For fairness, we use our own Transformer baseline trained from scratch under the same data, compute, and training pipeline as OmniGHA. Additional details are provided in Appendix A.5.

Model	Inference Time (s)					
	4K	8K	16K	32K	64K	128K
Show-o Xie et al. (2024b)	153	218	289	OOM	OOM	OOM
JanusFlow Ma et al. (2024)	91	125	153	182	223	268
OmniMamba Zou et al. (2025)	16	43	85	127	153	187
OmniGHA (AR)	**12**	**38**	**75**	**110**	**132**	**156**
OmniGHA (AR+Diffusion)	**13**	**36**	**79**	**113**	**128**	**151**

Table 3: **Inference efficiency on multimodal understanding.** OmniGHA (AR-only) achieves the best scalability across long contexts, consistently outperforming both Transformer- and Mamba-based baselines. The AR+Diffusion variant likewise surpasses all baseline models across comparable sequence lengths, underscoring its versatility across paradigms. OOM: Out-of-Memory

Model	Speed (Image/s)	Time (s)
Show-o Xie et al. (2024b)	0.81	19.66
JanusFlow Ma et al. (2024)	1.02	15.64
OmniMamba Zou et al. (2025)	5.68	2.81
OmniGHA (AR)	**5.26**	**2.97**
OmniGHA (AR+Diffusion)	**1.95**	**17.2**

Table 4: **Visual generation efficiency.** OmniGHA (AR-only) approaches OmniMamba while being $5\times$ faster than Show-o and JanusFlow. The AR+Diffusion variant is slower, reflecting the inherent overhead of diffusion, but demonstrates that GHA remains compatible and effective when integrated into heavier generative pipelines.

5.3 Efficiency Results

We compare the inference efficiency of OmniGHA with prior Transformer- and Mamba-based unified models on both multimodal understanding and visual generation, using a single NVIDIA A100 GPU in FP16.

For understanding (Table 3), OmniGHA (AR-only) achieves the lowest latency across sequence lengths up to 128k. At 32k and 128k tokens, it is $1.15\times$–$1.2\times$ faster than OmniMamba, and significantly outperforms Show-o ($12.8\times$) and JanusFlow ($2.3\times$). The AR+Diffusion variant also outperforms all baseline systems across comparable lengths, demonstrating that GHA preserves scalability and efficiency even when integrated with diffusion-style modules. Overall, these results confirm that both OmniGHA variants are competitive solutions for long-context multimodal understanding.

Type	Model	Params	Images	FID-30K↓
Gen. Only	DALL·E Ramesh et al. (2021)	12B	250M	27.5
	GLIDE Nichol et al. (2021)	5B	250M	12.24
	DALL·E 2 Ramesh et al. (2022)	6.5B	650M	10.39
	SDv1.5 Rombach et al. (2022)	0.9B	2000M	9.62
	PixArt Chen et al. (2023)	0.6B	25M	7.32
	Imagen Saharia et al. (2022)	7B	960M	7.27
	Parti Yu et al. (2022)	20B	4.8B	7.23
	Re-Imagen Chen et al. (2022)	2.5B	50M	6.88
	U-ViT Bao et al. (2023)	45M	83K(coco)	5.95
Unified	CoDI Tang et al. (2024)	-	400M	22.26
	SEED-X Ge et al. (2024)	17B	-	14.99
	LWM Liu et al. (2024a)	7B	-	12.68
	DreamLLM Dong et al. (2023)	7B	-	8.76
	Show-o Xie et al. (2024b)	1.3B	35M	9.24
	OmniMamba Zou et al. (2025)	1.3B	83K(coco)	5.50
	OmniMamba	1.3B	35M	5.36
	OmniGHA (AR)	**1.3B**	**35M**	**5.12**
	OmniGHA (AR+Diffusion)	**1.3B**	**35M**	**4.66**

Table 2: **Performance on MS-COCO.**

For generation (Table 4), OmniGHA (AR-only) delivers 5.26 images/s, $5.2\times$ faster than Show-o and $4.9\times$ faster than JanusFlow, while remaining close to OmniMamba (5.68 images/s). The AR+Diffusion variant reaches 1.95 images/s—slower due to the diffusion head but still outperforming Show-o and JanusFlow—confirming that the additional cost stems from the generation module rather than the GHA backbone, and that GHA transfers effectively to diffusion pipelines without sacrificing its efficiency advantage in the shared core.

Overall, these results show that GHA not only improves efficiency in AR-only pipelines but also transfers robustly to AR+Diffusion settings, offering a favorable trade-off between speed, accuracy, and extensibility across unified modeling paradigms.

Ablation	POPE↑	MME↑	GQA↑	FID-30K↓
Linear	71.8	893	45.2	23.7
Linear + Gate	80.7	1029	52.6	14.7
Linear + Agent-Softmax	80.9	1048	53.2	10.5
GHA	**82.6**	**1135**	**55.9**	**9.2**

Table 5: **Attention ablation in AR-only OmniGHA.** We ablate the proposed GHA by starting from a linear attention baseline and incrementally adding its two components: a gating mechanism and an agent-token softmax bottleneck. Both additions individually improve multimodal alignment (POPE, MME, GQA) and image fidelity (FID-30K), while combining them in GHA yields the largest gains. All experiments are conducted in the AR-only setting to isolate the effect of the attention operator.

5.4 Ablation Studies

We conduct ablations to isolate the contribution of the proposed GHA attention module, and provide additional results for LoRA configurations, staged pretraining, and agent-token number in Appendix A.6. All ablations are performed in the AR-only setting, as it provides a controlled environment to attribute gains directly to GHA without confounding factors from diffusion modules or complex backbones. For these ablations, we use a lightweight 0.4B model with a reduced agent-token count ($n = 9$) and a lower-cost training setup (shorter schedule and lower image resolution), allowing efficient experimentation while preserving the relative behavior of different attention variants.

Table 5 shows that while the linear attention baseline provides efficiency, it underperforms on multimodal reasoning and generation fidelity. Adding gating alleviates modality imbalance and stabilizes training, while agent-token softmax aggregation restores adaptive global context in long sequences. The full GHA module combines both components, consistently outperforming all partial variants, which confirms that the gains come from the synergy of gating and agent-token aggregation.

To further examine the role of the gating mechanism, we evaluate GHA under intentionally modality-skewed training regimes, where batches contain predominantly text-heavy or image-heavy samples while keeping the agent-token bottleneck fixed. Table 6 summarizes the results. Enabling the gate provides three consistent benefits: (i) it improves performance in both text-skewed and image-skewed settings; (ii) it reduces the performance gap between the two regimes; and (iii) it substantially lowers sensitivity to which modality dominates the data. These findings provide direct evidence that gating mitigates modality imbalance rather than merely improving overall accuracy.

Interestingly, the image-heavy regime even slightly outperforms the baseline one in Table 5 on certain metrics. This phenomenon is also observed in recent studies on modality imbalance, where increasing visual supervision might strengthen cross-modal alignment without degrading overall balance Park et al. (2025). The key point there is that greater exposure to visual signals can reduce the modality gap and enhance multimodal reasoning, rather than harming the model's ability to integrate modalities.

Our focus here is on efficient attention mechanisms under the unified model, but we view the systematic study of modality data distribution shifts as an important and emerging direction for future work.

Ablation	POPE↑	MME↑	GQA↑	FID-30K↓
GHA w/ gate (more text)	80.5	1059	53.5	10.8
GHA w/ gate (more image)	82.9	1156	56.8	10.3
GHA w/o gate (more text)	78.7	1003	51.7	11.4
GHA w/o gate (more image)	81.3	1094	54.4	10.9

Table 6: **Ablation under modality-imbalanced training.** Gating consistently improves robustness and reduces sensitivity to modality skew.

6 CONCLUSION

We introduced GHA, a hybrid operator that augments linear attention with gated recurrence and agent-token–based softmax aggregation, achieving consistent gains over strong unified baselines such as Show-o and JanusFlow while matching the efficiency of OmniMamba. Our study deliberately focused on a controlled setting with moderate scale and simplified AR-only / AR+Diffusion pipelines, ensuring improvements can be attributed directly to the attention mechanism itself. A key limitation is that we have not yet tested GHA within larger-scale unified systems such as BAGEL or Omni-Qwen, which couple massive datasets with more sophisticated paradigms (e.g., mixture-of-experts backbones). We view these as orthogonal and complementary directions, and future work will explore integrating GHA into such advanced frameworks to further assess its scalability and generality. The plug-and-play nature of GHA makes it especially suitable for such integration, offering a lightweight yet expressive operator for the next generation of unified multimodal models.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. All models were trained using publicly available datasets (MS-COCO, LAION, and GQA) under clearly specified preprocessing steps described in Section 5. We provide detailed hyperparameters, training schedules, and ablation settings in the Appendix. Hardware details, including GPU type, memory footprint, and batch sizes, are explicitly reported in Section 6.2. Random seeds were fixed across all experiments to reduce variance.

For the review phase, we provide an anonymous repository containing the core code for replication purpose: `https://anonymous.4open.science/r/OmniGHA-ICLR2026-63B2`

Upon acceptance, we will release the full codebase, pretrained models, and scripts to reproduce all evaluation results, including ablation studies.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, 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. Flamingo: a visual language model for few-shot learning, 2022. URL `https://arxiv.org/abs/2204.14198`.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL `https://arxiv.org/abs/2005.14165`.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey, 2024. URL `https://arxiv.org/abs/2402.12451`.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. *CoRR*, abs/2009.14794, 2020. URL https://arxiv.org/abs/2009.14794.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL https://arxiv.org/abs/2405.21060.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL https://arxiv.org/abs/2205.14135.

Karan Goel Datta, Amir Zadeh, Antoine Rondepierre, Kathryn Tunyasuvunakool, George Papamakarios, and Ahmed Osman. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2404.03063*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *arxiv*, 2024. URL https://arxiv.org/abs/2306.13394.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019. URL https://arxiv.org/abs/1902.09506.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021. URL https://arxiv.org/abs/2103.03206.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs outputs, 2022. URL https://arxiv.org/abs/2107.14795.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. arxiv 2024. *arXiv preprint arXiv:2309.04669*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020. URL https://arxiv.org/abs/2006.16236.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer, 2020. URL https://arxiv.org/abs/2001.04451.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks, 2019. URL https://arxiv.org/abs/1810.00825.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *arxiv*, 2023. URL https://arxiv.org/abs/2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv e-prints*, pp. arXiv–2402, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024c.

Xuezhe Ma, Xiang Kong, Sinong Wang, Chunting Zhou, Jonathan May, Hao Ma, and Luke Zettlemoyer. Luna: Linear unified nested attention, 2021. URL https://arxiv.org/abs/2106.01540.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2025. URL https://arxiv.org/abs/2411.07975.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Emilio Parisotto, H. Francis Song, Jack W. Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant M. Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, Matthew M.

Botvinick, Nicolas Heess, and Raia Hadsell. Stabilizing transformers for reinforcement learning, 2019. URL https://arxiv.org/abs/1910.06764.

Simon Park, Abhishek Panigrahi, Yun Cheng, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Generalizing from simple to hard visual reasoning: Can we mitigate modality imbalance in vlms?, 2025. URL https://arxiv.org/abs/2501.02669.

Qwen.ai. Qwen3-next: Towards ultimate training inference efficiency. https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd&from=research.latest-advancements-list, 2025. Accessed: 2025-12-02.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models, 2023. URL https://arxiv.org/abs/2307.08621.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, Wentao Li, Enzhe Lu, Weizhou Liu, Yanru Chen, Weixin Xu, Longhui Yu, Yejie Wang, Yu Fan, Longguang Zhong, Enming Yuan, Dehao Zhang, Yizhi Zhang, T. Y. Liu, Haiming Wang, Shengjun Fang, Weiran He, Shaowei Liu, Yiwei Li, Jianlin Su, Jiezhong Qiu, Bo Pang, Junjie Yan, Zhejun Jiang, Weixiao Huang, Bohong Yin, Jiacheng You, Chu Wei, Zhengtao Wang, Chao Hong, Yutian Chen, Guanduo Chen, Yucheng Wang, Huabin Zheng, Feng Wang, Yibo Liu, Mengnan Dong, Zheng Zhang, Siyuan Pan, Wenhao Wu, Yuhao Wu, Longyu Guan, Jiawen Tao, Guohong Fu, Xinran Xu, Yuzhi Wang, Guokun Lai, Yuxin Wu, Xinyu Zhou, Zhilin Yang, and Yulun Du. Kimi linear: An expressive, efficient attention architecture, 2025. URL https://arxiv.org/abs/2510.26692.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022a. URL https://arxiv.org/abs/2205.14100.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. A comprehensive review of multimodal large language models: Performance and challenges across different tasks, 2024a. URL https://arxiv.org/abs/2408.01319.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022b. URL https://arxiv.org/abs/2202.03052.

Xin Wang, Yang Li, Xiaoli Li, Yinan He, Wenxin Liu, Fan Yang, Weilin Huang, and Xiaodan Wang. Agent attention: On the integration of softmax and linear attention. *arXiv preprint arXiv:2403.09702*, 2024b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2025. URL https://arxiv.org/abs/2409.04429.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation, 2024a. URL https://arxiv.org/abs/2408.12528.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models, 2025. URL https://arxiv.org/abs/2506.15564.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training, 2024. URL https://arxiv.org/abs/2312.06635.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models, 2024. URL https://arxiv.org/abs/2401.13601.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. Llava-phi: Efficient multi-modal assistant with small language model. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*, pp. 18–22, 2024.

Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models, 2025. URL https://arxiv.org/abs/2503.08686.

A APPENDIX

A.1 OMNIGHA MODULES

The modules are shown in Table 7.

Table 7: OmniGHA architecture components and parameter sharing strategy.

Component	Role	Sharing	Trainable
Text Encoder (SigLIP)	Encode prompts	Shared	Frozen
Vision Encoder (DINOv2)	Visual understanding	Task-specific	Frozen
Image Tokenizer (LlamaGen)	Visual generation	Task-specific	Frozen
GHA Decoder (stacked)	Multimodal fusion	Shared	Trainable
Output Heads (linear)	Token/logit mapping	Task-specific	Trainable
LoRA Adapters	Task adaptation	Task-specific	Trainable

A.2 TRAINING STRATEGY

Phase 1: Representation Warm-up (Understanding). We begin by training the shared GHA decoder together with the text head and a task-specific LoRA branch on multimodal understanding tasks. These objectives provide dense token-level supervision, which encourages the gating mechanism to learn stable memory retention and prevents early degeneration.

Phase 2: Long-context Adaptation (Generation). Next, we activate the generation-specific LoRA branch and the image head while continuing to update the shared GHA decoder. This phase adapts the gated recurrence to handle long-sequence visual token generation, which exhibits higher variance and requires stable accumulation dynamics.

Phase 3: Joint Alignment (Unified Fine-tuning). Finally, we jointly optimize both LoRA branches and modality-specific heads with the shared GHA backbone. Encoders remain frozen. This stage aligns understanding and generation in the shared decoder space, consolidating cross-modal reasoning and ensuring that the gating mechanism generalizes across tasks.

Our design is tailored to the gated–agent structure of GHA. Empirically, we find that direct joint training leads to unstable loss and saturated gates, whereas the proposed progressive schedule provides denser early gradients, stabilizes gating dynamics in long-context generation, and consolidates cross-modal reasoning during fine-tuning. This curriculum not only improves downstream accuracy but also allows us to freeze heavy encoders and concentrate compute on the shared GHA decoder, yielding a more efficient training pipeline.

A.3 OMNIGHA VARIANTS: AR+DIFFUSION

To demonstrate extensibility beyond the AR-only setting, we also instantiate OmniGHA in an AR+Diffusion variant. In this variant, the shared GHA decoder produces unified multimodal fusion features that are projected via a lightweight latent predictor to form conditioning signals. These conditioning features guide a diffusion-style generation head based on flow matching, which performs efficient velocity field prediction to progressively sample high-fidelity images. The final output image is reconstructed through a pretrained VAE decoder that decodes latent representations into pixels. This design inherits the core principles of the Show-o approach, integrating autoregressive multimodal fusion with a diffusion-based generative mechanism, while demonstrating the flexibility of GHA in supporting diverse unified multimodal pipelines.

A.4 TRAINING DETAILS AND DATASETS

We constructed our corpus from publicly available multimodal understanding and visual generation datasets and trained the model from scratch without relying on a pretrained backbone. The architecture consists of 48 transformer layers, using a lightweight agent-token configuration of $n = 49$.

For multimodal understanding, DINOv2 and SigLIP encoders are employed (images resized to 384×384), and a VQVAE tokenizer (LlamaGen) is used for visual generation (images downsampled to 256×256). LoRA modules (rank 8, adding only 0.65% parameters) were applied to the input projectors of each block. Training was conducted on 64 NVIDIA A100 GPUs in BF16 precision using AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$), cosine annealing with warm-up, weight decay $= 0$, and gradient clipping $= 1.0$.

Our training followed a progressive three-phase curriculum. In *Phase 1 (Representation Warm-up)*, we trained only on multimodal understanding data, including 118K COCO images and 558K samples from the LLaVA-1.5 pretraining set, with batch size 32, learning rate 1×10^{-3}, 5K steps, and 100 warm-up steps. In *Phase 2 (Long-context Adaptation)*, we switched to text-to-image generation using 35M image–text pairs from CC12M, SA1B, and LAION-aesthetics-12M (excluding MS-COCO 2014), with batch size 90, learning rate 8×10^{-4}, 100K steps, and 1K warm-up steps. Finally, in *Phase 3 (Joint Alignment)*, we fine-tuned jointly on multimodal understanding (665K LLaVA-1.5 conversations, 220K LVIS-Instruct-4V with GPT-4V instructions, and 400K LRV-Instruct for hallucination mitigation) and text-to-image data, using batch size 48 for generation and 3 for understanding, learning rate 1×10^{-4}, 150K steps, and no warm-up.

This curriculum provides dense supervision early on, stabilizes gate dynamics during long-context generation, and consolidates cross-modal reasoning during unified fine-tuning while keeping heavy encoders frozen.

A.5 Additional Evaluation Details

To isolate the architectural contribution of GHA, we train a pure Transformer baseline that mirrors OmniGHA in all aspects of the pipeline. The baseline matches OmniGHA in model size (1.3B), component configuration (Table 7), training corpus, preprocessing, and the full three-phase curriculum. The only difference is that the decoder uses standard multi-head softmax attention instead of GHA.

Table 8 summarizes the understanding and generation results. OmniGHA improves over the Transformer baseline across all metrics (+5.1 POPE, +208 MME-P, +7.3 VQAv2, +1.3 GQA) and reduces FID-30K from 6.15 to 5.12.

Model	Params	Res.	POPE↑	MME-P↑	VQAv2↑	GQA↑	FID-30K↓
Standard Transformer	1.3B	384	83.6	1134.7	73.3	61.1	6.15
OmniGHA	1.3B	384	88.7	1342.5	80.6	62.4	5.12

Table 8: **Controlled comparison between standard Transformer and OmniGHA.** Both models are trained from scratch under identical data, compute, curriculum, and pipeline setups.

Beyond accuracy, GHA delivers substantial efficiency gains. Table 9 reports inference time across long-context settings. OmniGHA achieves 1.5–7× lower latency for multimodal understanding and 4× higher throughput for image generation.

Model	4K	8K	16K	32K	64K	128K	Avg.	Img/s	Latency (s)
Standard Transformer	87	104	128	167	204	249	156	1.29	15.12
OmniGHA	12	38	75	110	132	156	87	5.26	2.97

Table 9: **Inference efficiency comparison under identical hardware and decoding settings.**

These results confirm that the performance and efficiency gains of OmniGHA arise directly from the attention operator design rather than differences in model scale, data exposure, or training pipeline. Our Transformer baseline is implemented and tuned under exactly the same pipeline as OmniGHA, and both models are trained from scratch under matched compute and corpus. The empirical improvements therefore reflect architectural behavior rather than confounding factors.

Importantly, our observations are consistent with findings in other hybrid-attention studies. Recent works such as Qwen-Next Qwen.ai (2025) and Kimi Linear Team et al. (2025) (though in different domains) have shown that hybrid attention mechanisms can outperform pure Transformers in certain tasks or regimes, demonstrating that hybrid designs can be competitive and compute-efficient archi-

tectural choices. This contextualizes why, under our unified multimodal setting, GHA can deliver stronger results than the pure-Transformer baseline when budgets are matched.

While full softmax attention is undeniably expressive, unified multimodal decoders face an inherent challenge: modality imbalance. Visual tokens are dense and high-dimensional, whereas text tokens are sparse and low-entropy. Prior multimodal systems such as Flamingo and Show-o introduce additional projections, gating modules, or fusion strategies to stabilize cross-modal interactions under pure Transformers, underscoring this difficulty.

As discussed in Subsection 5.4, GHA targets this issue with two dedicated mechanisms: (1) key–value gating, which balances modality contributions by dynamically normalizing visual vs. textual keys/values; and (2) agent-token softmax aggregation, which stabilizes long-context interactions while preserving the global expressiveness and compensating for strengths associated with pure Transformers. Together, these components form a more stable and compute-efficient architecture for the fixed-budget unified multimodal training regime used in this work, explaining the observed gains over the controlled Transformer baseline and aligning with broader empirical trends in hybrid-attention architectures.

A.6 Additional Ablation Studies

Impact of Different Unified Model Architecture. Table 10 reports the results of different LoRA configurations. Although our backbone is designed to be shared across all tasks, we observe that adding lightweight task-specific components can still bring consistent benefits. This is because the backbone must generalize simultaneously to both dense understanding objectives and sparse generative supervision, which inevitably introduces optimization conflicts. Task-specific LoRA modules act as adapters that specialize the shared decoder to each modality without compromising parameter efficiency.

Removing LoRA modules significantly hurts both multimodal understanding and generation performance (POPE: 80.2, FID: 19.5). Introducing LoRA only for the understanding branch improves understanding metrics (82.1/1069/54.8) but only moderately reduces FID (14.3). Conversely, using LoRA only for the generation branch improves generation (FID: 13.9) while leaving understanding weaker (81.7/1045/54.6). Applying both understanding and generation LoRA leads to consistent gains across all benchmarks (82.6/1135/55.9/9.2), and further adding an output-specific LoRA yields the best performance (82.9/1149/56.3/8.8). These results confirm that task-specific LoRA modules help the model efficiently specialize for each modality while maintaining cross-task synergy.

Impact of Pretraining Phases. Table 11 presents an ablation study of the three-phase pretraining pipeline (Phase 1: representation warm-up, Phase 2: long-context adaptation, Phase 3: unified alignment). Using all three phases yields strong performance (82.6/1135/55.9/9.2), while skipping Phases 1 and 2 and training only with a unified objective leads to clear drops in both understanding and generation (80.9/1046/53.2/12.2).

These observations are consistent with prior unified multimodal systems, where multi-stage training is a standard practice. Pipelines such as Show-o Xie et al. (2024a), JanusFlow Ma et al. (2025), and OmniMamba Zou et al. (2025) all adopt staged curricula, as jointly optimizing understanding and generation with a single decoder is challenging under a single-stage recipe. The curriculum provides benefits to GHA: the warm-up phase supplies dense representation supervision, the long-context phase stabilizes the gate dynamics, and the final alignment phase consolidates cross-modal reasoning.

Impact of Agent-Token Number. We further examine the sensitivity of GHA to the number of agent tokens (n). Table 12 summarizes results for $n \in \{4, 9, 16, 25\}$. Performance improves substantially when increasing n from 4 to 9, after which the model enters a stable regime with only marginal gains for $n \in [9, 25]$. This shows that GHA does not rely on a large number of agent tokens—small n already captures most of the benefit—and that the improvements arise from the mechanism itself rather than from a large hidden constant.

To complement accuracy-based analysis, we also measure inference-time scaling at a 64K sequence length. Table 13 shows that runtime grows sublinearly with n, confirming that the cost is softened in practice due to our Triton kernel fusion and persistent KV buffer design.

Ablation	Understanding			Generation
	POPE↑	MME↑	GQA↑	FID-30K↓
No LoRA	80.2	1031	53.3	19.5
Only Understanding LoRA	82.1	1069	54.8	14.3
Only Generation LoRA	81.7	1045	54.6	13.9
Understanding & Generation LoRA	**82.6**	**1135**	**55.9**	**9.2**
Understanding & Generation LoRA + Output LoRA	**82.9**	**1149**	**56.3**	**8.8**

Table 10: Ablation studies on unified model architecture design.

Ablation	Understanding			Generation
	POPE↑	MME↑	GQA↑	FID-30K↓
Stage 1+2+3	82.6	1135	55.9	9.2
Only Stage 3	80.9	1046	53.2	12.2

Table 11: Ablation studies on Pretraining Phases. Phase 1: Understanding Pretraining, Phase 2: Generation Pretraining, Phase 3: Unified Alignment.

n (agent tokens)	POPE↑	MME↑	GQA↑	FID-30K↓
25	83.6	1160	58.7	8.2
16	82.7	1143	56.8	8.8
9	82.6	1135	55.9	9.2
4	78.4	1011	50.6	12.5

Table 12: Ablation on the number of agent tokens (n).

n (agent tokens)	Inference Time @ 64K (s)
25	93
16	78
9	43
4	32

Table 13: Inference-time impact of agent-token number.

A.7 LLM USAGE

This work is entirely original; ChatGPT was used solely for language polishing.