Firebolt-VL: Efficient Vision-Language Understanding with Cross-Modality Modulation

Anonymous CVPR submission

Paper ID *****

Abstract

Recent advances in multimodal large language models (MLLMs) have enabled impressive progress in vision-language understanding, yet their high computational cost limits deployment in resource-constrained scenarios such as personal assistants, document understanding, and smart cameras. Most existing methods rely on Transformer-based cross-attention, whose quadratic complexity hinders efficiency. Moreover, small vision-language models often struggle to precisely capture fine-grained, task-relevant visual regions, leading to degraded performance on fine-grained reasoning tasks that limit their effectiveness in the real world. To address these issues, we introduce Firebolt-VL, an efficient Vision-Language Model that replaces the Transformer-based Decoder with a Liquid Foundation Model. To further enhance visual grounding, we propose a Token-Grid Correlation Module, which computes lightweight correlations between text tokens and image patches and modulates via the state-space model with FiLM conditioning. This enables the model to selectively emphasize visual regions relevant to the textual prompt while maintaining linear-time inference. Experimental results across multiple benchmarks demonstrate that Firebolt-VL achieves accurate, fine-grained understanding with significantly improved efficiency.

1. Introduction

Multimodal Large Language Models (MLLMs) have emerged as a prominent research direction due to their remarkable performance across a wide range of tasks. These include image captioning, visual question answering (VQA), and optical character recognition (OCR). The rapid progress of state-of-the-art models such as LLaVA [24], IDEFICS [20], OpenFlamingo v2 [2], MiniGPT-4-v1 [42], MiniGPT-4-v2 [4], LLaVA-1.5 [25], LLaVA-Next [21], Chameleon [33], InternVL [8], Qwen-VL [8], and FastV [5] underscores the growing importance of MLLMs in real-world applications. Nevertheless, one of the primary limitations of current Vision-Language Models (VLMs) lies in their substantial computational requirements, which hinder their deployment in resource-constrained environments. As MLLMs continue to advance, they hold significant potential to enable real-time interaction in dynamic environments, enhance cross-modal retrieval, and achieve seamless integration of linguistic and visual reasoning in practical technologies.

Most recent MLLMs are built upon Transformer-based Large Language Models (LLMs), which exhibit quadratic computational complexity with respect to sequence length, as discussed in Mamba [14]. This computational overhead makes such models inefficient for inference on resource-limited devices and poses challenges for fast processing of long-context inputs. Therefore, improving the efficiency of LLMs is crucial to enable faster inference and facilitate the deployment of MLLMs in low-resource environments.

To alleviate previous limitations, several efficient architectures have been proposed, including Kosmos-2 [30], MobileVLM [37], MobileVLM V2 [9], MoE-LLaVA [23], LLaVA-Phi [43], and SmolVLM 2 [28]. These models leverage lightweight language backbones or incorporate mixture-of-expert mechanisms [10] to reduce model size and computational cost. Although such approaches have shown promising results on relatively simple benchmarks such as image captioning and VQA, they still face two major challenges. The first challenge is the quadratic computational complexity and limited ability to model long-range dependencies inherent to small Transformer-based architectures. The second challenge lies in the lack of precision when attending to task-relevant visual regions, which often leads to failures in handling fine-grained or detail-oriented questions that require rich visual representations.

In this work, we explore the integration of the Liquid Foundation Model (LFM) [1, 18] with visual information to design an efficient MLLM. To address the challenges faced by previous efficient MLLMs, we introduce a Token-Grids Correlation Modulation mechanism, resulting in the CMM module, where visual grid representations are fused with in-

struction text tokens to emphasize task-relevant visual cues. This mechanism enables the model to attend more effectively to fine-grained and detail-oriented information, enhancing its ability to reason over complex visual inputs. Based on this design, we propose **Firebolt-VL**, an efficient MLLM capable of handling a wide range of tasks, including image captioning, VQA, chart understanding, and other fine-grained visual reasoning benchmarks.

In summary, our main contributions are fourfold:

(1) We introduce **Firebolt-VL**, a novel MLLM that integrates the Liquid Foundation Model (LFM) [1, 18] for efficient sequence modeling, significantly reducing computational cost while maintaining strong multimodal reasoning performance.

(2) We propose a **Cross-Modal Modulator** mechanism that fuses visual grid representations with instruction text tokens, enabling more precise attention to task-relevant regions and improving fine-grained visual understanding.

(3) We conduct extensive experiments on multiple benchmarks, including image captioning, VQA, and OCR. Results demonstrate that Firebolt-VL achieves competitive or superior performance compared to existing efficient MLLMs, while substantially improving inference efficiency and scalability.

(4) We release the source code and pretrained model to promote transparency and encourage further research in the development of efficient MLLMs.

2. Related Work

2.1. Multimodal Large Language Model

Multimodal large language models (MLLMs) have become a central research direction in generative AI due to their wide applicability in document understanding, smart cameras, and virtual assistants. Recent works, such as LLaVA [24], IDEFICS [20], OpenFlamingo v2 [2], MiniGPT-4-v1 [42], MiniGPT-4-v2 [4], LLaVA-1.5 [25], LLaVA-Next [21], Chameleon [33], InternVL [8], Qwen-VL [8], and FastV [5] have achieved remarkable progress in visual understanding and text generation, pushing MLLMs closer to real-world deployment. Despite these advancements, the computational demands of modern VLMs remain a major barrier. High inference cost and memory overhead significantly hinder scalability to millions of users and limit practicality on resource-constrained devices.

Consequently, designing efficient vision–language models (VLMs) has emerged as a critical challenge for the MLLM community, as efficiency directly governs deployability across diverse hardware platforms. Early efforts such as MobileVLM [37] and MobileVLM V2 [9] reduce computational burden by employing lightweight MobileLLaMA backbones for text generation. Subsequent approaches, including MoE-LLaVA [23] and LLaVA-Phi [43], adopt Mixture-of-Experts (MoE) architectures [10] to activate only a fraction of parameters during inference, thereby eliminating redundant computation. More recently, SmolVLM 2 [28] introduced a compact language backbone combined with pixel-shuffle and inner-patching strategies to reduce the number of visual tokens and further improve efficiency.

While these models show promising performance and increasing adoption, they still rely heavily on Transformer-based architectures whose attention mechanism incurs quadratic time and memory complexity. This fundamental limitation restricts their ability to scale to long-context inputs and prevents truly lightweight, real-time deployment. To address this limitation, Firebolt-VL leverages the Liquid Foundation Model (LFM) Decoder [1, 18], achieving linear-time complexity and significantly improving the overall efficiency of vision-language modeling.

2.2. Cross-modal Integration

In recent works, most Vision–Language Models (VLMs) introduce cross-modal alignment through a simple linear projection layer, which maps visual features into a joint embedding space shared with the language encoder. While effective for large-scale models with strong visual backbones, this strategy becomes problematic for compact VLMs, whose vision encoders possess limited representational capacity, often resulting in weak or unstable alignment.

To improve alignment quality in smaller models, several enhanced strategies have been proposed. Dense Connector [39] enriches the visual representation by aggregating multi-level features from earlier layers. Align-KD [11] leverages knowledge distillation to transfer cross-modal alignment cues from larger teacher models, thereby strengthening the alignment of compact VLMs. Building on this direction, Align-GPT [41] introduces a hierarchical alignment scheme that learns multiple alignment levels during pre-training and adaptively fuses them during instruction tuning to support diverse task requirements.

Despite their effectiveness, these approaches still exhibit limited interactive fusion between visual and textual cues, often failing to direct the model's attention toward the most relevant visual regions for a given instruction. Qwen-VL [3] addresses this issue by incorporating cross-attention between image and text tokens, enabling richer cross-modal interaction. However, cross-attention incurs quadratic computational complexity, making it unsuitable for lightweight or latency-constrained deployment.

To overcome these limitations, Firebolt-VL introduces the Cross-Modal Modulator (CMM), which leverages a state-space model (SSM) [14] to efficiently encode and fuse grid-level visual tokens with textual representations. By computing lightweight token–grid correlations and apply-

ing FiLM-based modulation within an SSM framework, CMM allows the model to dynamically emphasize the most informative visual elements while maintaining near-linear complexity. This design enables stronger fine-grained grounding and contextually accurate multimodal reasoning without the computational overhead of cross-attention.

3. Method

3.1. Preliminaries

State-Space Models (SSM): A SSM [17], is a sequence model that modifies the hidden state $\mathbf{h}(t)$ over time through a linear dynamical system presented in Equation 1.

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{y}(t) = \mathbf{C}\mathbf{h}(t). \tag{1}$$

where $\mathbf{u}(t)$, $\mathbf{y}(t)$, and $\mathbf{h}(t)$ denote the input, output, and hidden state, respectively, and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are learnable matrices governing system dynamics. After discretization, the recurrence can be expressed as $\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{u}_t$ and $\mathbf{y}_t = \bar{\mathbf{C}}\mathbf{h}_t$, which can be viewed as a learnable 1-D convolution $\mathbf{y} = \mathbf{K} * \mathbf{u}$ with kernel $\mathbf{K}_t = \mathbf{C}\mathbf{A}^t\mathbf{B}$. This formulation allows the model to capture both short- and long-range temporal dependencies through continuous dynamics.

Following this concept, the Structured State-Space Model (S4) [15] and Mamba [14] introduce a parameterization of $\mathbf{A}$ that guarantees stability and expressiveness, allowing efficient training on long sequences while preserving global dependencies. S4 thus bridges the gap between the dynamical-system view of sequence modeling and the content-based attention mechanism [35] of Transformers.

Feature-wise Linear Modulation (FiLM): Feature-wise Linear Modulation (FiLM) [31] is a lightweight yet effective conditioning mechanism that modulates one modality's representation based on another by applying learned, feature-wise affine transformations. Given a visual feature vector $\mathbf{x} \in \mathbb{R}^D$ and a conditioning signal $\mathbf{c}$ (e.g., a text embedding), FiLM generates two modulation parameters, $\gamma(\mathbf{c})$ and $\beta(\mathbf{c})$, through a learnable function such as a linear projection:

$$\mathrm{FiLM}(\mathbf{x}, \mathbf{c}) = \gamma(\mathbf{c}) \odot \mathbf{x} + \beta(\mathbf{c}), \tag{2}$$

where $\odot$ denotes element-wise multiplication. Through this formulation, FiLM enables the conditioning signal to adaptively scale and shift visual features, integrating semantic cues into the representation without requiring explicit token-level attention.

FiLM is particularly effective in cross-modal architectures because it provides efficient and interpretable feature conditioning with linear computational complexity. By dynamically modulating feature channels, FiLM emphasizes task-relevant visual attributes (e.g., color, shape, or spatial relationships) while suppressing irrelevant ones, thereby aligning the visual representation with the contextual semantics of the conditioning signal. Unlike cross-attention, which computes dense pairwise interactions across all tokens, FiLM achieves context-aware modulation through direct channel-wise transformation, making it suitable for lightweight or real-time applications. When combined with the sequence model, FiLM acts as an efficient fusion layer that injects semantic information into the dynamic state representation, allowing long-range temporal dependencies to be propagated in a contextually grounded manner.

3.2. Overview

Figure 1 presents the overall architecture of our framework, which comprises three main modules: the Vision Encoder $Vis(\cdot)$, the Large Language Model $LLM(\cdot)$, and the proposed Cross-Modal Modulator $CMM(\cdot)$.

Vision Encoder. The vision encoder transforms an input image $X_I \in \mathbb{R}^{3\times H\times W}$ into grid-level visual representations. Following the SigLIP [34] framework, the image is partitioned into G grid regions, each of which is further divided into patches and processed by convolutional layers to produce patch tokens. The number of grids G can be adjusted to control the granularity of visual grounding with respect to the textual cues; in our experiments, we set $G = 5$ to balance fine-grained alignment and computational cost. These patch tokens are then propagated through L Transformer layers of a Vision Transformer (ViT), yielding grid-level visual representations $X_v \in \mathbb{R}^{G\times D_v}$. An optional global pooling operation can be applied to aggregate these representations into a compact visual embedding $V \in \mathbb{R}^{1\times D_v}$ for downstream tasks.

Cross-Modal Modulator (CMM). Given the text embedding $X_t \in \mathbb{R}^{T\times D_t}$ from the tokenizer and the visual embeddings $X_v \in \mathbb{R}^{G\times D_v}$ from the vision encoder, the CMM module aligns and fuses both modalities. Specifically, X_v is projected into a shared latent space with X_t, where a token–grid correlation is computed to identify the most relevant visual patches for each text token. With each token, a weighted visual context vector is generated and modulates the text representation via the Feature-wise Linear Modulation (FiLM) [31]. The modulated sequence X_f is then processed by a Structured State-Space Model (SSM) to capture long-range and cross-modal dependencies efficiently. Finally, a second FiLM modulation and a feed-forward refinement yield the multimodal output $X_{mm} \in \mathbb{R}^{T\times D_t}$, representing visually grounded text features.

Large Language Model. The Large Language Model receives the text embeddings and the multimodal features X_{mm} from CMM. These representations are concatenated or integrated as inputs to the LLM for multimodal reasoning and response generation. This design enables the LLM to leverage fine-grained visual cues while maintaining efficient, linear-time processing through the CMM module.

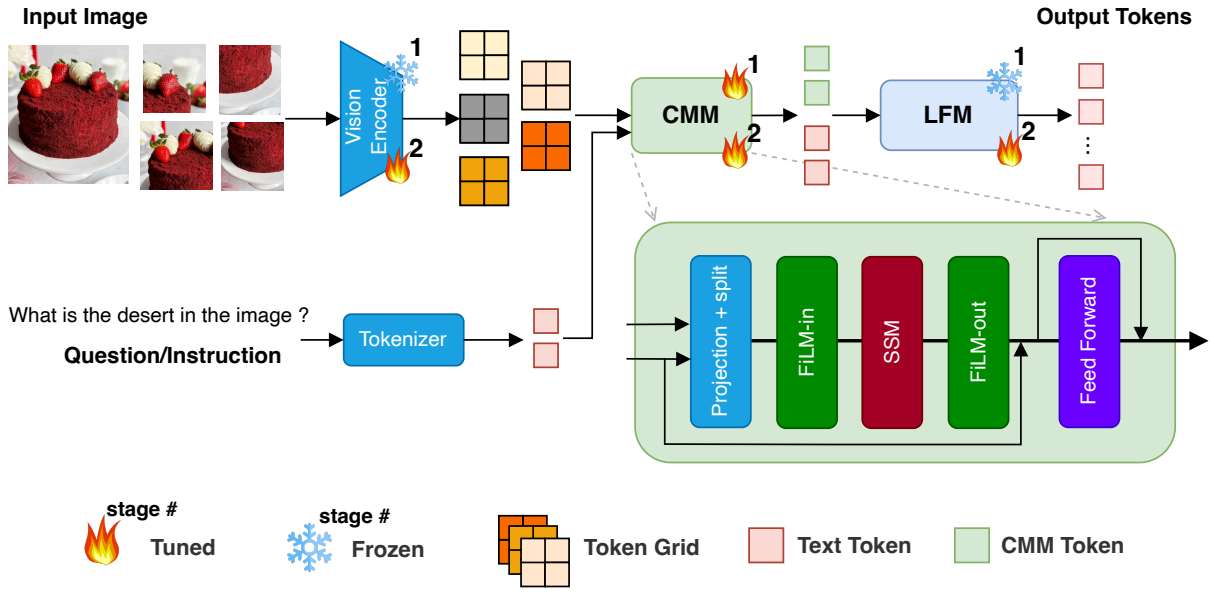

Figure 1. Overview of the Firebolt-VL architecture. The Cross-Modal Modulator (CMM) fuses textual instructions with the visual representations of the query image to produce conditioned tokens, which are then processed by the Liquid Foundation Model (LFM). The model is trained in two stages: (1) CMM pre-training to initialize modulation parameters, and (2) end-to-end training of the full framework.

3.3. Vision Encoder

For the Vision Encoder, in this work, we leverage the SigLIP [34] model, a multilingual vision–language encoder that replaces the traditional softmax contrastive objective with a sigmoid-based loss for image–text alignment. Given an input image $X_I \in \mathbb{R}^{3\times H\times W}$, the encoder first divides the image into patches and processes them through a Vision Transformer (ViT) backbone to obtain patch embeddings $X_v \in \mathbb{R}^{G\times D_v}$.

In our framework, we employ the SigLIP encoder to extract grid-level visual embeddings X_v from the input image while preserving native aspect ratios. These embeddings are then projected into the shared latent space for multimodal fusion via the proposed Cross-Modal Modulator (CMM), ensuring fine-grained correspondence between textual cues and spatial visual features.

3.4. Large Language Model

To generate the output text tokens, which are detokenized to the text answer, the Large Language Model based on LFM 2, text only model [1, 18]. Given the concatenated multimodal representation from CMM and the textual embeddings, denoted as $H = [X_t; X_{mm}] \in \mathbb{R}^{(L_t+L_{mm})\times D_t}$, the model autoregressively generates the target sequence $Y = \{y_i\}_{i=1}^{L_y}$ as depicted in Equation 3.

$$p_\theta(Y \mid H) = \prod_{i=1}^{L_y} p_{\theta_i}(y_i \mid H, y_{<i}). \tag{3}$$

where θ denotes all learnable parameters.

Lastly, the predicted tokens are detokenized into the final response in natural language. By unifying both modalities under a single autoregressive decoder, Liquid simplifies the architecture, reduces modality-specific alignment overhead, and enables scalable, efficient multimodal reasoning.

3.5. Cross-Modal Modulator (CMM)

In prior efficient Multimodal-LLM works, such as Mobile-VLM [37], Mobile-VLM V2 [9], and SmolVLM 2 [28], various projectors have been introduced to align vision-token embeddings with the text-token embedding space. However, these approaches exhibit limited interaction between text and image modalities, often causing text tokens to attend weakly to relevant visual regions—particularly in tasks requiring fine-grained reasoning. To address this limitation, we propose the **Cross-Modal Modulator (CMM)**, which enhances text–image interaction by dynamically conditioning text tokens on their associated visual context. Given the text embedding $X_t \in \mathbb{R}^{T\times D_t}$ and the grid-level visual representations $X_v \in \mathbb{R}^{G\times D_v}$, the CMM outputs a multimodal representation sequence $X_{mm} \in \mathbb{R}^{T\times D_t}$.

Overview. To identify the image patches most relevant to each text token, we first compute a correlation matrix between text and vision embeddings. Both modalities are projected into a shared latent space as shown in Equation 4:

$$X_t' = W_t X_t, \quad X_v' = W_v X_v \tag{4}$$

The projected embeddings are then split into multiple heads, $X_t^H \in \mathbb{R}^{H\times T\times D_H}$ and $X_v^H \in \mathbb{R}^{H\times G\times D_H}$. We

compute the scaled dot-product correlation matrix S_{mm} between text tokens and vision grids as:

$$S_{mm} = \sigma\left(\frac{X_t^H X_v^{H^\top}}{\sqrt{D_H}}\right), \tag{5}$$

where $\sigma(\cdot)$ denotes the softmax function applied along the grid dimension to normalize the correlation scores to the range $(0, 1)$. This operation yields attention weights indicating how strongly each text token attends to each visual grid. We further retain only the top-k grid locations per token to focus on the most relevant visual patches. The optimal k value is analyzed in Section 5.3.

After selecting the relevant patches, we average the correlation matrix across attention heads and compute a weighted sum over the visual features to obtain the per-token visual context embedding $c \in \mathbb{R}^{T\times D_t}$, as formulated in Equations 6 and 7:

$$\hat{S}_g = \frac{1}{H}\sum_{h=1}^{H} S_{mm}^{(h)}, \tag{6}$$

$$c = \sum_{g=1}^{G} \hat{S}_g X_v'. \tag{7}$$

We then apply **Feature-wise Linear Modulation (FiLM)** [31] to fuse the text embedding with its corresponding visual context:

$$X_f = \mathrm{LN}(X_t) \odot (1 + \alpha\,\gamma_{\text{in}}) + \alpha\,\beta_{\text{in}}, \tag{8}$$

$$\text{where } [\gamma_{\text{in}}, \beta_{\text{in}}] = W_f\, c. \tag{9}$$

Here, W_f is the FiLM projection weight mapping the context c into two modulation vectors—the *scale* γ_{in} and the *shift* β_{in}. The scalar parameter α is a learnable gate that controls the strength of cross-modal modulation. Intuitively, γ_{in} scales the feature channels of the text representation based on visual evidence, while β_{in} adds adaptive offsets to introduce new activations conditioned on the image. Together, these parameters reshape the text representation according to what each token "sees" before sequential modeling.

The visually modulated representation X_f is then processed by the **Structured State Space Model (SSM)** to capture long-range dependencies:

$$Y = \mathrm{SSM}(X_f). \tag{10}$$

The SSM efficiently models sequential interactions in $\mathcal{O}(T)$ time, allowing the text representation to evolve while preserving visual conditioning. Unlike self-attention, which explicitly computes pairwise interactions, the SSM propagates information implicitly through state transitions, capturing both local and global dependencies in a linear and memory-efficient manner. We evaluate alternative SSM variants in Section 5.3. After sequential modeling, we apply a second FiLM modulation (**FiLM-out**) to reintroduce the visual context:

$$Y_f = Y \odot (1 + \alpha\,\gamma_{\text{out}}) + \alpha\,\beta_{\text{out}}, \tag{11}$$

$$\text{where } [\gamma_{\text{out}}, \beta_{\text{out}}] = W_f'\, c. \tag{12}$$

Similar to FiLM-in, γ_{out} and β_{out} adjust the post-SSM features based on c, ensuring alignment between textual and visual features after temporal mixing.

Finally, residual connections and a feed-forward network (FFN) refine the fused representations:

$$X_{\text{out}} = \mathrm{LN}\big(X_t + Y_f + \mathrm{FFN}(X_t + Y_f)\big). \tag{13}$$

This step stabilizes the multimodal representation and enhances expressiveness through non-linear transformations in the FFN. The output $X_{\text{out}} \in \mathbb{R}^{B\times T\times D_t}$ thus contains text features that are visually modulated and temporally contextualized by the state-space model.

To obtain a global multimodal representation, we aggregate the token-level outputs X_{out} using mean pooling:

$$z = \frac{1}{T}\sum_{t=1}^{T} X_{\text{out},t}. \tag{14}$$

The resulting vector $X_{mm} \in \mathbb{R}^{B\times D_t}$ serves as a compact fused embedding that captures both linguistic and visual semantics. This embedding is subsequently passed to the language decoding stage for multimodal reasoning and response generation.

Complexity Analysis: CMM performs the token-grids correlation only once to derive a compact visual context and replaces the repeated cross-attention mechanism with a linear-time State-Space Model (SSM), whose complexity scales as $\mathcal{O}(TGD_t + TD_t^2 + TD_t f(T))$, where $f(T) \in \{1, \log T\}$. As a result, CMM achieves nearly linear scaling with respect to sequence length T and cost with respect to grid size G (for fixed or sparse top-k), substantially reducing both computational and memory overhead while preserving effective cross-modal alignment through FiLM-based modulation. Compared to multimodal fusion typically relies on cross-attention between text and vision tokens, which computes the attention matrix $S_{mm} = \mathrm{Softmax}(QK^\top/\sqrt{D_t})$ and the weighted aggregation AV, resulting in a computational complexity of $\mathcal{O}(TGD_t)$ per layer and a memory requirement proportional to $T \times G$. While effective, this operation becomes expensive as either the text length T or the number of visual grids G increases when compared with CMM.

4. Experimental Setup

4.1. Training Recipe

To train the Firebolt-VL model, we conduct the training in two stages. The first stage is the initialization stage for the

Method	LLM	Parameters	VQAv2	POPE	AI2D	MMMU_{Val}	MME^{p}	SQA^{I}	MMB_{dev}
IDEFICS [20]	LLaMA	9.0B	60.0	81.9	42.2	18.4	1177.3	53.5	45.3
OpenFlamingo v2 [2]	MPT	9.0B	60.4	52.6	31.7	<u>28.8</u>	607.2	44.8	–
MiniGPT-4-v1 [42]	Vicuna	8.0B	–	34.6	28.4	23.6	1047.4	39.6	–
MiniGPT-4-v2 [4]	LLaMA 2	8.0B	–	60.0	30.5	25.0	968.4	54.7	–
Chameleon [33]	LLaMA 2	7.0B	–	19.4	<u>46.0</u>	22.4	202.7	46.8	–
FastV [5]	Vicuna	7.0B	55.0	48.0	42.7	22.0	873.2	51.1	–
Kosmos-2 [30]	GPT 2	1.7B	45.6	66.3	25.6	23.7	721.1	32.7	–
MobileVLM [37]	Mobile-LLaMA	1.7B	–	84.5	36.6	25.3	1196.2	57.3	59.6
MobileVLM V2 [9]	Mobile-LLaMA	1.7B	–	84.3	38.1	19.0	1302.8	<u>66.7</u>	57.7
MoE-LLaVA [23]	Qwen	2.2B	<u>76.2</u>	**87.0**	42.1	26.6	1291.6	63.1	59.6
LLaVA-Phi [43]	Phi 2	2.7B	71.4	<u>85.0</u>	–	–	<u>1335.1</u>	**68.4**	<u>59.8</u>
SmolVLM 2 [28]	Mobile-LLaMA	0.3B	–	54.3	39.2	**28.9**	1236.5	58.8	–
Firebolt-VL (Ours)	**LFM-2**	0.8B	**76.6**	69.4	**46.2**	26.4	**1376.2**	56.7	**64.6**

Table 1. Quantitative comparison of the proposed **Firebolt-VL** model with existing MLLMs across seven benchmarks. The superscript p denotes the perception score on the MME benchmark, while T refers to the test set. The best results are shown in **bold**, and the second-best results are <u>underlined</u>. "–" indicates results not reported in the original papers.

cross-modal fusion, in which the vision encoder and the language model are frozen while the fuser is trained. In this stage, the CC3M dataset [29, 32] is leveraged to initialize the weight of the fuser. In the second stage, we conduct the training for the full model. To allow the reasoning ability of the model, we leverage the LLaVA-CoT dataset [38], and we processed the MMPR-v1.2 [6, 7, 36] dataset into the chain-of-thought format to make the model learn to do the reasoning.

4.2. Implementation Details

The model is trained using 2 NVIDIA H100 80GB GPUs with batch size of 128 in stage 1, and 8 in stage 2. For the optimizer, we employ the AdamW optimizer with a learning rate of 5e-4 for the first stage and 1e-4 for the second stage. The model is trained for 5 epochs in each of the two stages. The best model is collected after 2 epochs in each stage. We select the best model choose the best model using the perplexity metric on the validation set.

4.3. Comparison Baseline

To assess the generalization and reasoning ability of Firebolt-VL across diverse environments, we evaluate on several datasets such as VQAv2 [13], POPE [22], AI2D [19], MMMU [40] validation set, MME [12], SQA-Image [27], and MMB [26] development set with two settings for big models and small models. The first setting compares with the big models, which have more than 7 billion parameters, including IDEFICS (2023) [20], OpenFlamingo v2 (2023) [2], MiniGPT-4-v1 (2023) [42], MiniGPT-4-v2 [4], Chameleon (2024) [33], and FastV (2024) [5]. The second setting compares small models, which have fewer than 3 billion parameters, involving these methods: Kosmos-2 (2023) [30], MobileVLM (2024) [37], MobileVLM V2 (2024) [9], MoE-LLaVA (2024) [23], LLaVA-Phi (2024) [43], and SmolVLM 2 (2025) [28]. For the fair comparison, we conduct the benchmark on models, which are trained at same scale of the dataset.

5. Results

5.1. Quantitative Results

Comparison with previous works in image understanding: From the benchmark results in Table 1, despite having fewer than 1B parameters, Firebolt-VL achieves competitive or even superior performance compared to much larger models (over 7B parameters), and clearly outperforms compact models in the 0.3B–3B range. Notably, Firebolt-VL attains the highest scores on both VQAv2 and MME benchmarks, demonstrating strong visual reasoning and perceptual alignment capabilities. These findings highlight the effectiveness of our state-space–based architecture in capturing multimodal dependencies efficiently, without relying on heavy Transformer-based backbones. Results suggest that structured state-space modeling offers a promising alternative for scalable and efficient multimodal understanding.

Comparison of efficiency with efficient vision language models works To evaluate the computational efficiency of Firebolt-VL against MobileVLM [37], MobileVLM-V2 [9], MoE-LLaVA [23], and SmolVLM 2 [28]. We evaluate on the POPE [22] dataset using two key metrics: *latency* and *throughput* (tokens per second). All models were evaluated on a single NVIDIA H100 GPU with a maximum output length of 256 tokens for consistent comparison. As reported in Table 2, Firebolt-VL achieves the **lowest latency** among all lightweight multimodal baselines and gains the **highest throughput** at 46.67 tokens per second. These results demonstrate the strong computational efficiency of our Liquid-based architecture and the lightweight nature of our Cross-Modal Modulator. Overall, the integration of struc-

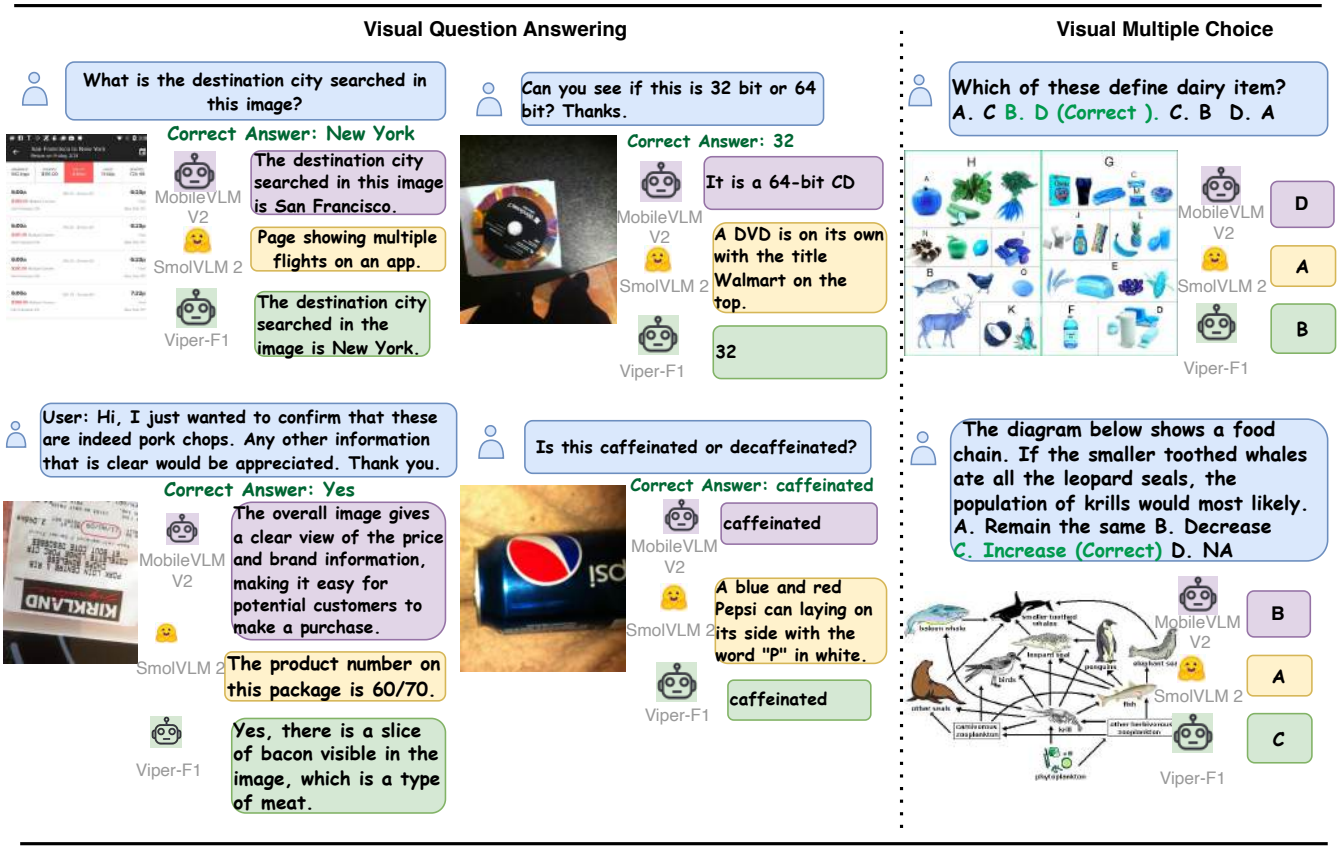

Figure 2. Qualitative comparison of responses from Firebolt-VL with recent efficient vision–language models, including MobileVLM-V2 [9] and SmolVLM-V2 [28], on detail-dependent question-answering tasks. Firebolt-VL demonstrates stronger fine-grained grounding and more accurate, instruction-aligned responses.

tured state-space modeling enables fast multimodal reasoning while maintaining low computational overhead, making Firebolt-VL suitable for real-time and resource-limited deployment scenarios.

Method	Latency (ms) ↓	Throughput (Tok/s)↑
MobileVLM [37]	50.00	39.99
MobileVLM-V2 [9]	46.04	42.86
MoE-LLaVA [9]	50.01	19.97
SmolVLM 2 [28]	65.72	37.80
Firebolt-VL (Ours)	**40.08**	**46.67**

Table 2. Efficiency benchmark comparing Firebolt-VL with state-of-the-art models in terms of latency and throughput.

5.2. Qualitative Results

In Figure 2, we present qualitative results that highlight Firebolt-VL's visual reasoning and text generation abilities across both visual question answering and visual multiple-choice tasks, compared against two efficient multimodal baselines—SmolVLM-2 [28] and MobileVLM-V2 [9]. Unlike prior lightweight models, which often produce generic descriptions, Firebolt-VL consistently generates precise, question-grounded answers. For example, when asked about the "destination city being searched," Firebolt-VL correctly attends to the search bar region in the image and extracts the appropriate answer, while baselines fail to localize this detail.

Similarly, in multiple-choice scenarios, Firebolt-VL demonstrates reliable fine-grained visual discrimination, such as identifying subtle differences among dairy product labels or tracking hierarchical relations in food chains. These examples collectively show that Firebolt-VL effectively attends to task-relevant visual grids and leverages localized visual cues to produce more accurate and context-aware responses than existing efficient VLMs.

5.3. Ablation Studies

State-space model (SSM) choice. To determine the most suitable state-space model for our Firebolt-VL framework, we conduct experiments to evaluate the impact of differ-

ent SSM variants on overall performance. Specifically, we compared three representative models—Mamba [14], S4D [16], and S4 [15], as depicted in Table 3.

Approach	POPE	AI2D	MMMU_{val}	Average
Mamba [14]	57.4	39.9	23.6	40.3
S4D [16]	**69.9**	45.6	23.7	46.4
S4 [15]	69.4	**46.2**	**26.4**	**47.3**

Table 3. Performance comparison of Mamba, S4D, and S4 state-space models on POPE [22], AI2D [19], and MMMU_{val} [40] benchmarks. Structured models (S4/S4D) outperform Mamba.

From the results, it can be observed that S4 and S4D yield higher performance compared to Mamba, indicating that structured state-space models are more effective in capturing the interactions between textual and visual embeddings. Since the visual embeddings represent information from five spatial grids of the image, the structured state-space architecture enables more accurate modeling of geometric relationships during the transition process, thereby achieving superior results compared to the Mamba model.

Cross-modal fusion mechanism: To evaluate the effectiveness of the CMM module, we conduct experiments comparing three fusion strategies: (1) Prepend, where the projected image features by a MLP layer are passed through the Language Model; (2) Cross-attend, where cross-attention follows Q-Former is applied to enable interaction between visual and textual features; and (3) CMM (ours), the proposed module that integrates state-space modeling for efficient and structured cross-modal fusion. The experimental results are depicted in Table 4.

Approach	MME	AI2D	MMMU_{val}
Prepend	981.5	24.1	22.1
Cross-attended	1036.5	45.1	24.4
CMM (Ours)	**1376.2**	**46.2**	**26.4**

Table 4. Performance of different connector methods on POPE [22], AI2D [19], and MMMU_{val} [40]. The proposed CMM outperforms both Prepend and Cross-attention fusion across all benchmarks.

From the benchmark results, it can be observed that in the MME perception benchmark, the CMM approach significantly outperforms both the Cross-attend and Prepend methods, demonstrating that our fusion mechanism effectively enhances the model's perceptual ability. Moreover, in the AI2D and MMMU_{val} benchmarks, CMM also achieves higher scores—particularly on chart and diagram questions in AI2D and on mathematics and coding-related questions in MMMU_{val}. These results indicate that CMM enables more precise contextual alignment between modalities, leading to an overall improvement in the model's reasoning and understanding performance.

Top-K grid Assessment: To assess how many visual grids should be selected by the CMM module, we evaluate the model under different values of k, where k represents the number of top-ranked visual grids attended to by the fusion mechanism. Increasing k allows the model to aggregate information from a broader visual context, which may help capture additional details that are relevant to the question; however, selecting too many grids can also introduce noise or dilute the contribution of the most informative regions. Since the number of grids we extract from the visual encoder is 5, we evaluate Top-K values from 1 to 5.

top-k	MME	AI2D	MMMU_{val}
1	1039.8	43.9	22.2
2	1192.0	44.1	24.3
3	1258.6	45.4	25.8
4	**1376.2**	**46.2**	**26.4**
5			

Table 5. Evaluation of the number of top-k grid selections on POPE [22], AI2D [19], and MMMU_{val} [40]. Larger k values improve performance consistently.

From Table 5, we observe that performance improves consistently as k increases from 1 to 4 across all benchmarks (MME, AI2D, and MMMUval). Notably, the results on MMMUval increase along with the Top-K, suggesting that reasoning-intensive tasks benefit from integrating multiple visual cues. These results indicate that relying solely on the single most salient grid ($k = 1$) is insufficient for robust multimodal understanding, and that attending to multiple top-ranked grids allows the model to better capture fine-grained details and complementary visual signals. Overall, increasing k enhances the model's reasoning ability while maintaining stability across benchmarks.

6. Conclusion

We present Firebolt-VL, an efficient multimodal LLM that leverages the Liquid model as the language decoder and incorporates a lightweight fusion mechanism combining a state-space model with linear feature modulation. This design specifically addresses the challenge of fine-grained detail perception in efficient vision–language models. In this study, we successfully integrate the Liquid model within the VLM framework and demonstrate that our proposed Cross-Modal Modulator (CMM) enables the model to attend to visual details that are directly relevant to the input question or instruction. As a result, Firebolt-VL remains lightweight while achieving strong perceptual and reasoning performance.

Limitations. Our current model operates on single-image input, and extending CMM to support multiple images

or video sequences remains an open challenge. We have not yet developed an effective version of CMM for video input; however, future work will explore efficient cross-modal connectors capable of integrating temporal visual information with textual instructions.

References

[1] Liquid AI. Liquid Foundation Models: Our First Series of Generative AI Models | Liquid AI, 2024. 1, 2, 4

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1, 2, 6

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 2, 6

[5] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the European Conference on Computer Vision*, pages 19–35, 2024. 1, 2, 6

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 6

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1, 2

[9] 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. 1, 2, 4, 6, 7

[10] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 1, 2

[11] Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4178–4188, 2025. 2

[12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. 6

[13] 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*, pages 6904–6913, 2017. 6

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First conference on language modeling*, 2024. 1, 2, 3, 8

[15] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3, 8

[16] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022. 8

[17] James D Hamilton. State-space models. *Handbook of econometrics*, 4:3039–3080, 1994. 3

[18] Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023. 1, 2, 4

[19] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Proceedings of the European conference on computer vision*, pages 235–251, 2016. 6, 8

[20] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 1, 2, 6

[21] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1, 2

[22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6, 8

[23] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 1, 2, 6

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[25] 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*, pages 26296–26306, 2024. 1, 2

[26] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of the European conference on computer vision*, pages 216–233, 2024. 6

[27] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 6

[28] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 1, 2, 4, 6, 7

[29] Edwin G. Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. *arXiv preprint arXiv:2012.02339*, 2020. 6

[30] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 6

[31] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3, 5

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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*, 2018. 6

[33] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 2, 6

[34] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3, 4

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[36] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 6

[37] Qinzhuo Wu, Weikai Xu, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, Bin Wang, and Shuo Shang. Mobilevlm: A vision-language model for better intra-and inter-ui understanding. *arXiv preprint arXiv:2409.14818*, 2024. 1, 2, 4, 6, 7

[38] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. 6

[39] Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. *Advances in Neural Information Processing Systems*, 37:33108–33140, 2024. 2

[40] 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*, pages 9556–9567, 2024. 6, 8

[41] Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. Alignngpt: Multi-modal large language models with adaptive alignment capability. *arXiv preprint arXiv:2405.14129*, 2024. 2

[42] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 6

[43] 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*, pages 18–22, 2024. 1, 2, 6