Eve: Efficient Multimodal Vision Language Models with Elastic Visual Experts

Miao Rang, Zhenni Bi, Chuanjian Liu, Yehui Tang, Kai Han*, Yunhe Wang*

Huawei Noah's Ark Lab
{rangmiao1,bizhenni, liuchuanjian, yehui.tang, kai.han,yunhe.wang}@huawei.com

Abstract

Multimodal vision language models (VLMs) have made significant progress with the support of continuously increasing model sizes and data volumes. Running VLMs on edge devices has become a challenge for their widespread application. There are several efficient VLM efforts, but they often sacrifice linguistic capabilities to enhance multimodal abilities, or require extensive training. To address this quandary, we introduce the innovative framework of Efficient Vision Language Models with Elastic Visual Experts (**Eve**). By strategically incorporating adaptable visual expertise at multiple stages of training, Eve strikes a balance between preserving linguistic abilities and augmenting multimodal capabilities. This balanced approach results in a versatile model with only **1.8B** parameters that delivers significant improvements in both multimodal and linguistic tasks. Notably, in configurations below 3B parameters, Eve distinctly outperforms in language benchmarks and achieves state-of-the-art results **68.87**% in VLM Benchmarks. Additionally, its multimodal accuracy outstrips that of the larger 7B LLaVA-1.5 model.

Introduction

As the swiftly evolving of artificial intelligence, the understanding of vision and language has gained significant attention, becoming a prominent research focus. Multi-modal models, such as Vision-Language Models (VLMs), are designed to combine visual information and textual descriptions, aiming to enhance semantic comprehension. These models, including GPT4V (Liu et al. 2023b) and Gemini (Team et al. 2023), have shown substantial potential in various applications, such as visual reasoning, visual question answering, and multi-modal retrieval.

Most of the existing VLMs primarily enhance multimodal capabilities by expanding data volumes or enlarging the model sizes. Consequently, numerous high-quality visual-textual datasets (Zhang et al. 2024; Chen et al. 2023; Zhao et al. 2023; Chen et al. 2024) have been developed alongside a suite of large model tuning techniques (Team et al. 2023; Bai et al. 2023; Liu et al. 2023a,c). These approaches not only boost the model's generalization capabilities, enabling it to adeptly handle a diverse range of visual and textual inputs, but also enhance its ability to recognize and comprehend complex real-world scenarios and relationships. However, these models are usually large in size, making it difficult to deploy and perform efficient inference on the devices, hindering their practical applications.

To develop efficient VLMs, several methods are proposed to maintain multimodal capabilities while reducing the model size (Chu et al. 2023; Lin et al. 2024; Yuan et al. 2024). However, these methods often focus on augmenting multimodal capabilities at the expense of linguistic proficiency. MoE-LLAVA (Lin et al. 2024), for instance, significantly enhances multimodal capacities by integrating multiple experts. However, a notable degradation in the precision of language tasks. Conversely, DeepSeek-VL (Lu et al. 2024) maintains linguistic abilities during multimodal training by incorporating substantial amounts of language data. Although effective, this strategy heavily enlarges the training cost.

In this paper, we propose an efficient VLM framework to build multimodal and language capabilities under relatively small model size and low training cost. Based on the existing powerful LLMs, we introduce elastic vision experts to process visual inputs and enhance multimodal capabilities. The proposed Eve framework consists of three training stages and strategically embeds elastic vision experts at each stage. In the initial two stages, we can leverage the well pretrained vision encoders such as ResNet and Vision Transformer (ViT), in the public community which is elastic for building strong visual capability. In the third stage, we integrate an elastic vision feed forward network (FFN) in the LLM transformer while keeping the original LLM part frozen. This structured integration allows each expert to concentrate on distinct visual tasks, thereby enhancing multimodal capabilities without compromising the intrinsic linguistic abilities. Furthermore, this approach obviates the need for substantial textual data during training, thus significantly expediting the model training process. Within a model size of 3B parameters, Eve achieves state-of-the-art performance, and compared to other multimodal models, our language capabilities are better preserved.

In summary, the contributions of this paper are as follows:

- We present the Elastic Visual Expert (**Eve**) framework, meticulously structured across three training stages, and

*Corresponding Author.

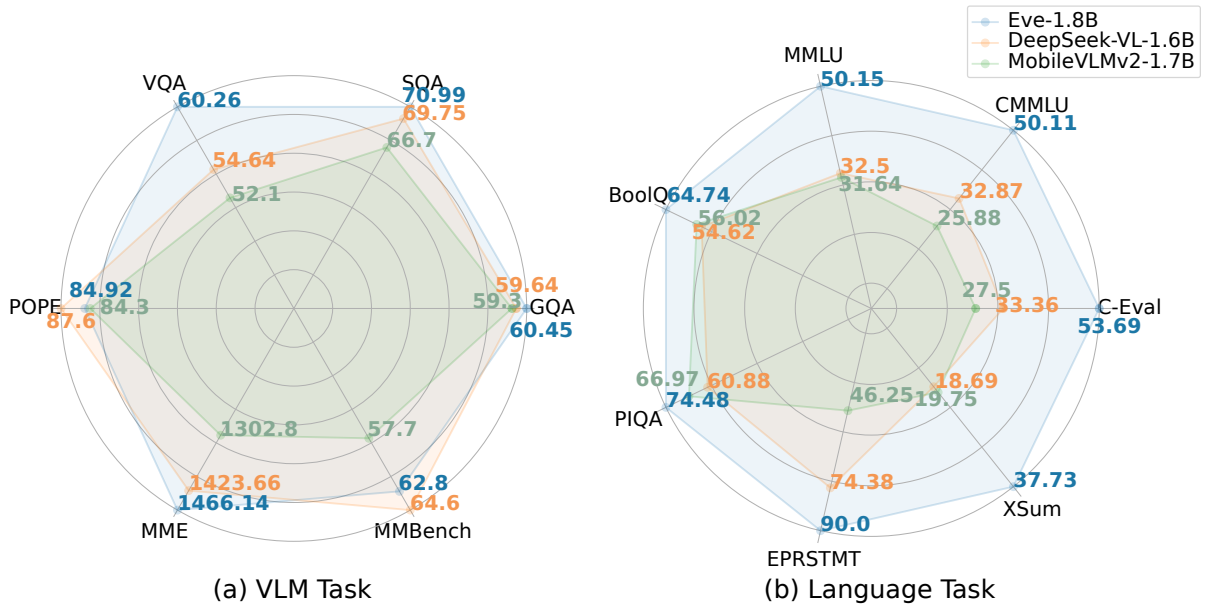

Figure 1: Comparison with SOTA methods with 1B scale across VLM and language benchmarks.

ingeniously incorporates dynamically adaptive visual experts in each phase, enabling each expert to concentrate on distinct domain-specific tasks. Throughout the training process, we strategically amalgamate the peak performance of these experts to bolster multimodal capabilities, all while maintaining the inherent linguistic proficiency;

- The Elastic Visual Experts, featuring the Elastic Vision Encoder (EVE) and Elastic Vision Feed-Forward Network (EVF), are engineered with remarkable adaptability. During the first two stages of training, the visual encoder remains frozen, facilitating seamless integration with various visual encoders, while preserving the language model's performance. In the third stage, the EVF is introduced, unifying with the model's linguistic capabilities to create a powerful synergy. This fusion significantly elevates the model's ability to process and merge visual and textual data, thereby markedly enhancing its multi-modal performance;
- Eve stands out in multimodal tasks with less than 3 billion parameters, achieving top performance in VLM and language benchmarks, and is on par with the larger 7B LLaVA-1.5 in terms of multimodal accuracy.

Related Work

Large Vision Language Models. As the capabilities of Large Language Models (LLMs) have significantly increased in tasks such as reasoning, comprehension, and question answering, Large Vision Language Models (LVLMs) are integrating powerful large language models with visual branches to expand the reasoning abilities of LLMs for processing multimodal data, thereby achieving more comprehensive and in-depth understanding and generation capabilities. In the field of visual-language learning, a notable example is CLIP (Radford et al. 2021), which employs a large number of image-text pairs for contrastive learning to align images and language in a semantic space. Building upon CLIP, BLIP (Li et al. 2022) utilizes single-modal encoders for image and text encoding, with the text encoder, similar to BERT (Devlin et al. 2019), adding a new token [CLS] to represent the entire sentence in the input. BLIP-2 (Li et al. 2023b) introduces Q-Former to align the frozen visual base model and LLM. Additionally, MiniGPT-4 (Zhu et al. 2023) introduce visual instruction fine-tuning through a projection layer, aligning a frozen visual encoder with an advanced frozen LLM Vicuna to enhance instruction following capabilities. ShareGPT4V (Chen et al. 2023) has generated a high-quality image-text description dataset covering a wide range of domains, including world knowledge, object properties, spatial relationships, and aesthetic evaluation, significantly improving the model's accuracy in multimodal benchmark testing. Qwen-VL (Bai et al. 2023) integrates training data from tasks such as image captioning, visual question answering, OCR, and document understanding, incorporating visual foundational capabilities into Qwen-VL. The generated model demonstrates outstanding performance across these diverse tasks.

Efficient Vision Language Models. The practical application of multimodal large language models has been limited by the computational costs and memory requirements during both training and inference stages. Recently, several studies have delved into the exploration of Small Vision Language Models (SVLMs) are characterized by a parameter spectrum that encompasses a range from 1 billion to 3 billion from different perspectives. For instance, LLaVA-Phi (Zhu et al. 2024) leverages the pre-trained small language model Phi-2 (2.7B) as the core of the multimodal

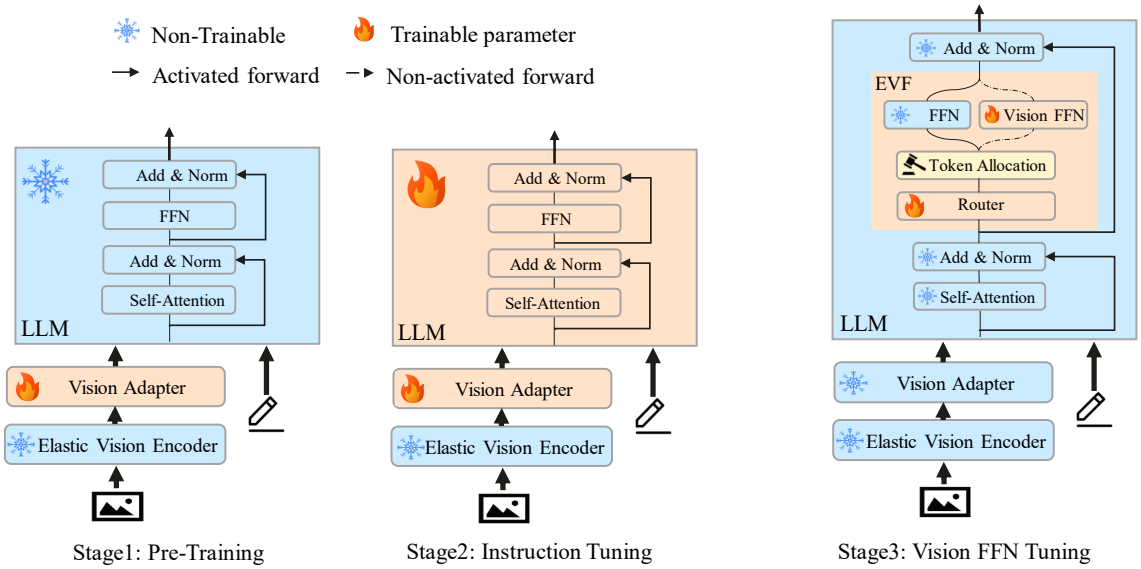

Figure 2: The Eve training framework and strategy. The Eve employs a meticulously structured three-stage training approach. **Stage 1**: Training is dedicated to the vision adapter to adapt the LLM specifically for processing visual inputs. **Stage 2**: To enhance multimodal capabilities, training vision adapter and LMM with LoRA. **Stage 3**: We introduce a new EVF layer, consisting of an elastic vision FFN and a fixed language FFN. The weights from the original FFN are duplicated to initialize the two FFNs in alternating half-layers of the LLM. This stage involves isolated training of the vision FFN, aimed at significantly enhancing the model's proficiency in visual information comprehension.

model, incorporates CLIP ViT-L/14 as the visual encoder, and employs two layers of MLP to connect the visual encoder, demonstrating outstanding performance in visual understanding, reasoning, and multimodal perception. MobileVLM (Chu et al. 2023, 2024) provides an open-source approach for 1B/3B visual language models, enhancing the performance of SVLM through innovative adapter designs and high-quality data.

Mixture of Experts in Multi-modal Learning. The concept of Mixture of Experts (MoE) was first introduced in (Jacobs et al. 1991) as a novel supervised learning process, which employs multiple models (or "experts") to learn and utilizes a gating network to determine which model is best suited to train on each data point. This approach reduces interference between different types of samples, enabling each expert to focus on processing a single task more effectively. In V-MoE (Riquelme et al. 2021), the authors introduced the first large-scale application of MoE to the Vision Transformer (ViT), significantly increasing accuracy while reducing computational costs. The VL-MoE (Shen et al. 2023) is the first work to apply MoE in the fusion of image and text modalities, demonstrating outstanding performance across multiple tasks. The VLMo (Bao et al. 2022) model employs three experts, specializing in visual, linguistic, and visual-linguistic tasks, using the MoE framework to balance the depth encoding of each modality and the fusion of multi-modal information. This design enables flexible handling of both single-modal and multimodal data pairs. Building upon the VLMo model structure, BEiT-3 (Wang et al. 2022) further expands the model's scale and simplifies the pre-training loss function. MoE-LLaVA (Lin et al. 2024) proposes a sparse model architecture based on MoE, featuring a soft router, which requires fewer activation parameters to achieve or even surpass the performance of dense models.

Methods

Overview

Our proposed model, **Eve**, incorporates a sophisticated three-stage framework, strategically integrating elastic vision experts at each stage, as depicted in Fig. 2. A key focus of our approach is the preservation of linguistic capabilities throughout the training process. Notably, the linguistic proficiency of the model remains unaffected by the variations in pre-training data used for the visual encoder during the first two stages of training. This stability in linguistic performance is a significant accomplishment, as it ensures that the model's ability to process and comprehend language is not compromised by changes in the visual encoder's pre-training. Further details on this aspect are provided in Elastic Vision Encoder.

In the later stages of the training strategy, we introduce a novel elastic visual FFN in the third phase to enhance the model's capacity for multimodal data processing. This component is specifically designed to complement the model's linguistic capabilities, allowing it to effectively analyze visual data while preserving its proficiency in language tasks.

The integration of the elastic visual FFN not only strengthens the model's multimodal abilities but also ensures the retention of its inherent linguistic competencies. As a result, the model is capable of handling complex multimodal inputs while preserving its linguistic processing and comprehension capabilities. A detailed discussion of the design and implementation of the elastic visual expert is provided in Elastic Vision FFN.

Elastic Vision Encoder

Over the past decade, the research community has developed numerous foundational visual models, such as ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2020), which exhibit exceptional capabilities in visual signal processing and are capable of extracting rich visual representations. Building upon these advancements, we propose an elastic vision encoder that harnesses the strengths of existing visual models, effectively standing on the shoulders of giants to construct a more robust vision-language model (VLM).

Leveraging Vision Encoders in the Wild. The vision encoder extracts features from an RGB image, fundamentally transforming this input into a sequence of visual embeddings that capture critical visual feature details. To enhance multimodal capabilities while preserving the inherent abilities of the language model, the vision adapter is continuously trained to align vision features with the language model's feature space in the first two stages, while the language model undergoes only light LoRA-based (Hu et al. 2021) fine-tuning in the second stage, as shown in Fig 2. The vision encoder remains frozen to elastically support any vision backbone models. This design ensures that the language model's performance remains largely unaffected even when the vision encoder is flexibly replaced. Consequently, our architecture can incorporate elastic vision encoders, including the use of various pre-training data. There are a large number of pretrained vision backbones in the opensource community, which elastically provide a rich source of vision encoders. Therefore our approach can maximize the use of existing industry capabilities to enhance multimodal capabilities while preserving the inherent linguistic abilities of the language model.

Pre-training Data in Vision Encoder. To minimize the temporal costs associated with individual trials, our study employs smaller-scale visual models in conjunction with the PanGu-π-1.5B (Wang et al. 2023) language model. Specifically, we utilize the ResNet-50 architecture (He et al. 2016) as the vision encoder, evaluating its performance across different datasets: ImageNet-1K (Russakovsky et al. 2015), ImageNet-22K (Ridnik et al. 2021), and LAION-400M (Radford et al. 2021). The experimental results are summarized in Table 1. We observe that the model trained on ImageNet-22K achieves the highest precision, with a value of 53.36% on the VLM benchmark. Although LAION-400M, being a larger dataset, contains more diverse internet-sourced data, its quality is compromised by the presence of non-ideal samples. In contrast, ImageNet-22K offers higher-quality data, demonstrating that superior data quality can outweigh the benefits of larger, lower-quality datasets in training visual models. Notably, the accuracy difference between the best-performing model (trained on ImageNet-22K) and the lowest-performing one (trained on LAION-400M) is within a narrow 0.3% range, as shown in the final column of Table 1. Furthermore, pre-training the vision encoder with elastic pre-trained datasets effectively preserves the linguistic capabilities of the model.

Vision Encoder	P-Dataset	Data Size	VLM AVG	L-AVG
ResNet50	ImageNet-1K	1.2M	49.06	51.73
ResNet50	ImageNet-22K	14M	**53.36**	**51.98**
ResNet50	LAION-400M	413M	52.68	51.68

Table 1: Vision encoder pre-training dataset impact on VLM and language tasks. P-Dataset denotes the pre-training dataset utilized for the vision encoder, and L-AVG is short for Language Task Average Accuracy.

Elastic Vision FFN

In the third phase, inspired by MoE-LLaVA (Jiang et al. 2024), we introduce an Elastic Vision Feed-Forward Network (EVF) to enhance the multimodal capabilities of the model. This addition is designed to improve the processing of visual information within the large language model (LLM). To preserve the model's existing linguistic proficiency, we freeze the majority of the parameters in the language model, allowing only the parameters of the EVF layer to be updated during training.

Elastic Vision FFN Layer (EVF). As depicted on the right side of Fig. 2, each EVF layer incorporates a sophisticated routing mechanism and a dedicated token allocation strategy, alongside two distinct Feedforward Networks (FFNs): one specialized for linguistic processing and the other for visual information. This dual-FFN design significantly enhances the model's multimodal capabilities by efficiently processing both language and visual data.

During forward propagation in the large language model (LLM), image tokens processed through the vision adapter and text tokens are concatenated and jointly fed into the LLM. Initially, the routing layer assigns each token to a recommended Feedforward Network (FFN). The token allocation mechanism then considers both the routing layer's recommendation and the current capacity of the FFN to determine whether the token should be assigned to that specific FFN. The routing mechanism employs a linear layer to compute the probabilities of assigning each token to its respective FFN, selecting the one with the highest probability as the preferred FFN. The routing mechanism can be formalized as follows:

$$\mathrm{P}(\mathbf{x})_i = \frac{e^{f(\mathbf{x})_i}}{e^{f(\mathbf{x})_l} + e^{f(\mathbf{x})_v}}, \tag{1}$$

where the router generates weight logits $f(x) = W \cdot x$, which are then normalized using the softmax function. Here, W represents the lightweight trainable parameters. The logit

$f(x)_l$ corresponds to the language FFN, while $f(x)_v$ corresponds to the vision FFN.

In the initialization phase of Stage 3, we duplicate the FFN parameters from Stage 2 into both the language and vision FFNs. During the training phase of Stage 3, we freeze all parameters of the vision encoder and vision adapter, as well as the majority of the parameters in the language model, restricting training to the vision FFN and the routing layer within the language model. In the inference phase for multimodal tasks, both FFNs within the EVF layer are activated. The routing and token allocation mechanisms collaborate to assign tokens to the appropriate FFN. For language-only tasks, however, the EVF layer excludes both the routing layer and the vision FFN, retaining only the untrained language FFN. As a result, the language model operates in its standard configuration, with the language FFN remaining untrained and its linguistic capabilities fully preserved.

Token Allocation. The token allocation mechanism plays a pivotal role in determining which Feedforward Network (FFN) each token is assigned to. In the EVF layer, each FFN e_i has a predefined capacity C, which limits the number of tokens it can process. In conventional token allocation mechanisms, if the number of tokens M recommended by the routing mechanism for an FFN e_i exceeds its capacity C, only a random subset of C tokens is selected from M for allocation to FFN e_i, and the excess M-C tokens are discarded. This indiscriminate discarding of tokens can significantly degrade the model's accuracy.

To overcome the limitations of conventional token allocation, we introduce GBPR, a novel strategy that prioritizes token distribution within a complete batch based on token importance, as determined by the routing score $P(x)$. Specifically, when the number of tokens M exceeds the capacity C of an FFN e_i, GBPR prioritizes the allocation of the C most important tokens to e_i, while discarding the remaining less important $M - C$ tokens.

Further improving this approach, we propose Img-GBPR, a mechanism for distinctly managing vision and text tokens. This mechanism assigns a default recommended FFN for each token type to ensure that visual and linguistic FFNs can focus on their respective tasks. Image tokens are initially assigned a score $S_i \in \mathbb{R}^{P\times 2}$, with values in the visual FFN column approaching 1, thus directing them primarily to the visual FFN. In contrast, values in the linguistic FFN column are near 0. Text tokens are scored differently, with $S_t \in \mathbb{R}^{N\times 2}$, where the score for the visual FFN approaches 0 and the score for the linguistic FFN approaches 1, facilitating their allocation to the linguistic FFN. The final priority score for each token is calculated by summing the routing score $P(x)$ and the initial score S_i or S_t, optimizing token allocation based on their modality. The score can be represented as follows:

$$S(x) = \begin{cases} P(x) + S_i, & \text{if } x \in \text{image token} \\ P(x) + S_t, & \text{if } x \in \text{text token,} \end{cases} \tag{2}$$

Furthermore, when the number of tokens M assigned to a FFN e_i exceeds its capacity C, we prioritize the selection of the most important C tokens based on $S(x)$ for allocation to FNN e_i. The remaining $M - C$ unallocated tokens are reintroduced to the candidate pool for redistribution. A certain proportion W is randomly selected and allocated to another FFN to minimize the loss of tokens. This scoring strategy ensures that image tokens are preferentially assigned to the visual FFN, optimizing their processing for visual tasks, while text tokens are directed to the linguistic FFN, thereby enhancing the model's capabilities for textual interpretation. This dual-token allocation approach maximizes the efficiency of each FFN, enabling them to operate optimally within their respective modalities.

Balanced Loss. We draw upon MoE-LLaVA, where the overall loss function consists of both the regression loss $\mathcal{L}_{\text{regressive}}$ and the auxiliary loss $\mathcal{L}_{\text{aux}}$. The regression loss optimizes the model's performance, while the auxiliary loss encourages balanced load distribution across the FFNs. Given our unique token allocation mechanism, we have adjusted the coefficient α of the auxiliary loss to 0.001.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{regressive}} + \alpha \cdot \mathcal{L}_{\text{aux}} \tag{3}$$

Experiments

Settings

Model Details. Eve is meticulously designed around three core components: a vision adapter, a visual encoder, and a language model. The vision adapter, empowered by the Lightweight Downsample Projector (LDP)(Chu et al. 2023), serves as an innovative bridge between the visual encoder and the language model, enabling seamless integration and alignment of multimodal features. The visual encoder leverages SigLip-L(Zhai et al. 2023), built upon the robust VIT-L backbone and utilizing a patch size of 576, which is renowned for its exceptional ability to capture intricate details from visual data. Complementarily, the linguistic backbone is formed by is PanGu-π-1.5B-Pro (Tang et al. 2024), a powerful architecture featuring 22 layers, a width of 2048 dimensions, and a vocabulary size of 48,000 entries. This high-capacity design significantly enhances Eve's capacity to comprehend nuanced language structures and generate sophisticated text, thereby strengthening its overall competence in cross-modal understanding and expression.

Implementation Details. In Stage 1, we freeze both the vision encoder and the LLM, focusing exclusively on training the efficient vision adapter. In Stage 2, we fine-tune the vision adapter in conjunction with the LLM, utilizing the LoRA technique. Finally, in Stage 3, we train only the vision FFN, assigning each FFN a capacity of C=1.5. Detailed training settings are provided in Table 2.

Training Dataset. Our dataset has been carefully refined and expanded to create high-quality datasets that enhance cross-modal understanding. In the first two phases, we utilize the CC-595K and LLaVA-mixed-665 datasets to develop foundational multimodal capabilities. In the third phase, we curate a diverse collection of datasets across several domains, including General Multi-modality, Visual

Configuration	Stage 1	Stage 2	Stage 3
Vision encoder init	SigLip-L	SigLip-L	SigLip-L
LLM init	PanGu-π-1.5B Pro	PanGu-π-1.5B Pro	Eve Stage2
Vision adapter init	Random	Eve Stage1	Eve Stage2
Image resolution	384x384	384x384	384x384
Learning rate	1e-3	2e-5	2e-5
LR schedule	Cosine decay	Cosine decay	Cosine decay
Weight decay	0	0	0
Optimizer	AdamW(b_1=0.9, b_2=0.95)		
Warmup ratio	0.03	0.03	0.03
Global batch size	256	128	128
Training steps	2181	5197	25510
Training hours	0.8	7	34
Epoch	1	1	1
GPU	8xV100-32G	8xV100-32G	8xV100-32G

Table 2: Training hyperparameters of Eve.

Question Answering (VQA), Optical Character Recognition (OCR), Image Captioning, and Knowledge-intensive tasks. This comprehensive ensemble consists of over 3.2 million samples, all meticulously designed to significantly enhance the model's versatility and performance across a wide range of modal scenarios. Detailed descriptions of the various datasets are provided in Appendix B.

Evaluation. Our primary objective centers on rigorously evaluating the model's proficiency in both multimodal and linguistic tasks. Following the rigorous evaluation protocols established in prior works such as (Chu et al. 2023, 2024), we employ a comprehensive suite of VLM benchmarks for multimodal assessment, comprising GQA (Hudson and Manning 2019), SQA (Lu et al. 2022), TextVQA (Singh et al. 2019), MME (Guo et al. 2023), MMBench (Liu et al. 2023d) and POPE (Li et al. 2023c). Consistent with the approach outlined in (Tang et al. 2024), we employ a diverse array of benchmarks to evaluate linguistic competencies. These include C-Eval (Huang et al. 2024), CMMLU (Li et al. 2023a), MMLU (Hendrycks et al. 2020), BoolQ (Clark et al. 2019), PIQA (Bisk et al. 2020), EPRSTM (Xu et al. 2021) and XSum (Narayan, Cohen, and Lapata 2018).

Ablation Study

Effect of Elastic Vision FFN Layers. In the third stage, we compare the performance differences between the EVF and MoE layers on multimodal and language tasks, using a ResNet50 vision encoder and PanGu-π-1.5B as the language model. The MoE layers, introduced by MoE-LLAva, are adjusted to match the parameter count of the EVF layers by employing a "x2top1" strategy, which involves two FFNs without differentiation; both FFNs are activated and trained simultaneously. Detailed comparisons are presented in Table 3. Compared to Stage 2, the MoE layers improve multimodal task accuracy by 0.55%, but significantly reduce language task accuracy by 3%. In contrast, the EVF layer architecture not only enhances multimodal task accuracy by 0.47%, but also fully preserves language task accuracy.

Effect of Token Allocation. We visualized the token success rate across different EVF layers during training using

Stage	EVF	MoE	VLM AVG	Language AVG
Stage2			61.23	58.65
Stage3		✓	**62.23**	55.03
Stage3	✓		61.93	**58.65**

Table 3: Impact of EVF vs. MoE Layers on VLM and Language benchmark

Token Allocation	VLM AVG
Random	53.83
GBPR	54.37
Img-GBPR	**54.92**

Table 4: Imapct of different token allocation.

three distinct token allocation methods: Random, GBPR, and Img-GBPR. The analysis focused on layers 1, 11, and 21, with the results shown in Fig. 3. When employing the random token distribution mechanism, approximately 25% of tokens were discarded at layers 11 and 21. In contrast, GBPR improved the acceptance rates of tokens in the initial and final layers (Layer 1 and Layer 21) as training progressed, although a drop-off rate of approximately 25% persisted at layer 11 in the later stages. The introduction of Img-GBPR, with its redistribution strategy, resulted in a more substantial improvement in token success rates across all layers (initial, middle, and final), highlighting the effectiveness of token distribution strategies in optimizing model training.

Furthermore, we compared the impact of different token allocation methods on the accuracy of multimodal tasks, using a ResNet50 vision encoder and the PanGu-π-1.5B large language model. The experimental results are detailed in Table 4. Employing the GBPR method led to an improvement over the random allocation approach, with an average accuracy increase of 0.4 percentage points. When the Img-GBPR method was applied, model accuracy increased further by 0.5%.

Ablation Study of Best Results. To achieve optimal performance, we conduct a series of detailed ablation experiments across three dimensions: method, model, and training dataset. The specifics of these experiments are provided in Table 6. The baseline model is based on MobileVLM, utilizing a ResNet50 visual encoder, MobileLLaMA as the language model, and LDP as the vision adapter, all trained on the Stage 2 dataset.

Initially, we replace the language model with PanGu-π-1.5B, resulting in a significant increase of 1.4% in average accuracy. We then incorporate two effective schemes that we propose—the EVF layer and Img-GBPR—which further improve accuracy by 1.6%. To align with the current state-of-the-art model, DeepSeek, we replace both the visual and language models with stronger alternatives: the visual encoder is upgraded to SigLIP-L, which leads to a substantial 8% increase in multimodal accuracy. Additionally, replacing

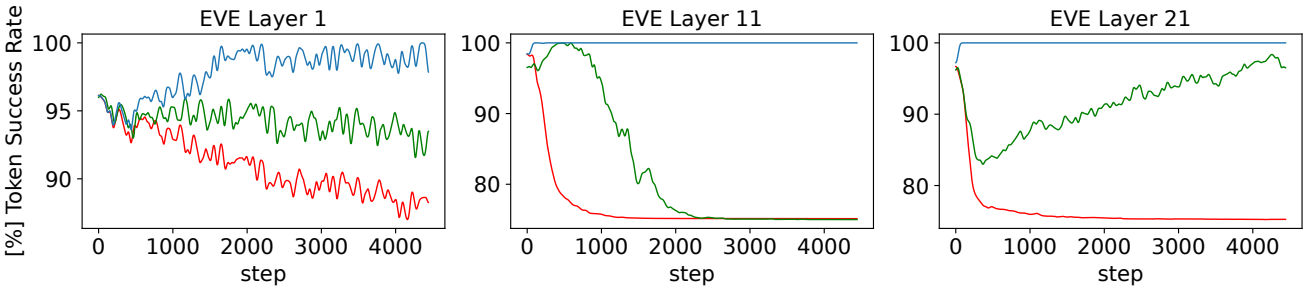

Figure 3: The impact of token allocation mechanisms on successful routing in Layer 1, 11 and 21. The red line represents "Random", the green line represents "GBPR", and the blue line represents "Img-GBPR".

Method	VM	LLM	Act.	Res.	Image Question Answering			Benchmark Toolkit			AVG	GPU-days
					GQA	SQA^I	VQA^T	POPE	MME^P	MMB^{dev}		
					12578	2017	5000	8910	2374	4329		
LLaVA-1.5 (Liu et al. 2023a)	ViT-L	V-7B	6.7B	336	62.00	66.80	58.20	85.90	1510.70	64.30	68.79	-
LLaVA-1.6 (Liu et al. 2024)	ViT-L	V-7B	6.7B	336	64.80	72.80	65.70	86.70	1498.00	68.70	72.27	-
TinyGPT-V (Yuan, Li, and Sun 2023)	ViT-L	P-2.7B	2.7B	448	33.60	41.22	11.40	50.56	507.80	35.55	33.85	-
Mini-Gemini (Li et al. 2024)	ConX-L	G-2B	2B	336	-	-	56.20	-	1341.00	59.80	-	-
MobileVLM (Chu et al. 2023)	ViT-L	M-1.4B	1.4B	336	56.10	57.30	41.50	84.50	1196.20	53.20	58.70	1.6
MobileVLM (Chu et al. 2023)	ViT-L	M-2.7B	2.7B	336	58.40	59.00	46.70	84.60	1296.40	57.00	61.75	2.6
MobileVLM v2 (Chu et al. 2024)	ViT-L	M-1.4B	1.4B	336	59.30	66.70	52.10	84.30	1302.80	57.70	64.20	9
MobileVLM v2 (Chu et al. 2024)	ViT-L	M-2.7B	2.7B	336	61.10	70.00	57.50	84.70	1440.50	63.20	68.10	-
LLaVA-Phi (Zhu et al. 2024)	ViT-L	P-2.7B	2.7B	336	**68.40**	66.39	48.60	85.00	1335.10	**66.70**	65.82	3.2
MoE-LLaVA-1.6Bx4-Top2 (Lin et al. 2024)	ViT-L	S-1.6B	2.0B	384	61.5	63.9	54.3	85.9	1335.7	63.3	65.95	8.6
DeepSeek-VL (Lu et al. 2024)	SigLIP	D-1.3B	1.3B	384	59.64	69.75	54.64	**87.60**	1423.66	64.60	67.90	896
Eve-VLM	ViT-L	PG-1.5B	1.5B	336	59.99	68.12	57.78	84.95	1292.35	60.82	66.05	15
Eve-VLM	SigLIP	PG-1.5B	1.5B	384	60.45	**71.49**	**60.26**	84.92	**1466.14**	62.80	**68.87**	15

Table 5: Comparison with SOTA methods across 6 VLM benchmarks. 'VM' signifies the vision model component utilized in the VLM, whereas 'LLM' indicates the language model component.'Act.' refers to the number of activated parameters within the models, 'Res.' denotes input image resolution. The models 'V', 'M', 'P', 'S', 'D', 'G', and 'PG' correspond to Vicuna (Chiang et al. 2023), Mobile LLaMA (Chu et al. 2023), Phi-2 (Li et al. 2023d), StableLM (Bellagente et al. 2024), DeepSeek-LM (Guo et al. 2024), Gemini (Li et al. 2024), and PanGu-π-1.5B-Pro (Tang et al. 2024), respectively. "AVG" stands for the weighted mean of 6 VLM benchmarks. 'GPU-days' quantifies the computational time required for model training.

the language model with PanGu-π-1.5B-Pro further boosts accuracy by 1.5%, reaching 64.52%. Finally, substituting the training data with our meticulously curated Stage 3 dataset results in an additional 4.3% increase in accuracy, achieving a peak accuracy of 68.87%.

Vision Encoder	LLM	EVF	Img GBPR	Stage3 Data	VLM AVG
ResNet50	MobileLLaMA				51.97
ResNet50	PanGu-π-1.5B				53.36
ResNet50	PanGu-π-1.5B	✓			53.83
ResNet50	PanGu-π-1.5B	✓	✓		54.92
SigLip-L	PanGu-π-1.5B	✓	✓		63.03
SigLip-L	PanGu-π-1.5B-Pro	✓	✓		64.52
SigLip-L	PanGu-π-1.5B-Pro	✓	✓	✓	**68.87**

Table 6: Ablation study for optimal results: effective methods, vision-language models, and training datasets."VLM AVG" represents the average accuracy of the VLM benchmarks.

Comparisons with State-of-the-art Methods

We compare Eve with current state-of-the-art models in Table 5. Among models with fewer than 3B activated parameters, Eve achieves the best accuracy 68.87%. When compared to models with similar parameter sizes, Eve outperforms DeepSeek-VL by 1.9% and offers significant advantages in training efficiency, requiring only 15 GPU-days. Eve even surpasses that of some 7B models, such as LLaVA-1.5. Additionally, as shown in Fig. 1, Eve notably outperforms existing VLM with fewer than 3B parameters, especially in maintaining full language task capabilities. Detailed results are provided in Appendix A.3.

Conclusion

In this work, we introduce efficient VLM framework of Eve with embedding elastic visual experts at various stages. Additionally, the adaptive token allocation mechanism facilitates the model's ability to process multimodal information more effectively. Therefore, the model not only retains its language capabilities but also significantly enhances its multimodal abilities.

Acknowledgments

We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; and Wei, F. 2022. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. arXiv:2111.02358.

Bellagente, M.; Tow, J.; Mahan, D.; Phung, D.; Zhuravinskyi, M.; Adithyan, R.; Baicoianu, J.; Brooks, B.; Cooper, N.; Datta, A.; et al. 2024. Stable lm 2 1.6 b technical report. *arXiv preprint arXiv:2402.17834*.

Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439.

Chen, G. H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; and Wang, B. 2024. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. arXiv:2402.11684.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. arXiv:2311.12793.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; and Shen, C. 2023. MobileVLM : A Fast, Strong and Open Vision Language Assistant for Mobile Devices. arXiv:2312.16886.

Chu, X.; Qiao, L.; Zhang, X.; Xu, S.; Wei, F.; Yang, Y.; Sun, X.; Hu, Y.; Lin, X.; Zhang, B.; and Shen, C. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. arXiv:2402.03766.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Guo, D.; Zhu, Q.; Yang, D.; Xie, Z.; Dong, K.; Zhang, W.; Chen, G.; Bi, X.; Wu, Y.; Li, Y.; et al. 2023. Mme:A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Guo, D.; Zhu, Q.; Yang, D.; Xie, Z.; Dong, K.; Zhang, W.; Chen, G.; Bi, X.; Wu, Y.; Li, Y.; et al. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming–The Rise of Code Intelligence. *arXiv preprint arXiv:2401.14196*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Hudson, D. A.; and Manning, C. D. 2019. 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*, 6700–6709.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1): 79–87.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Li, Y.; Hui, B.; Yin, Z.; Yang, M.; Huang, F.; and Li, Y. 2023d. Pace: Unified multi-modal dialogue pre-training with progressive and compositional experts. *arXiv preprint arXiv:2305.14839*.

Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; and Jia, J. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Huang, J.; Zhang, J.; Ning, M.; and Yuan, L. 2024. MoE-LLaVA: Mixture of Experts for Large Vision-Language Models. arXiv:2401.15947.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. OpenAI. Gpt-4v(ision) system card. 2023. 1, 2. *Advances in neural information processing systems*, 36.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Sun, Y.; et al. 2024. DeepSeek-VL: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ridnik, T.; Ben-Baruch, E.; Noy, A.; and Zelnik-Manor, L. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Pinto, A. S.; Keysers, D.; and Houlsby, N. 2021. Scaling Vision with Sparse Mixture of Experts. arXiv:2106.05974.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Shen, S.; Yao, Z.; Li, C.; Darrell, T.; Keutzer, K.; and He, Y. 2023. Scaling Vision-Language Models with Sparse Mixture of Experts. arXiv:2303.07226.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Tang, Y.; Liu, F.; Ni, Y.; Tian, Y.; Bai, Z.; Hu, Y.-Q.; Liu, S.; Jui, S.; Han, K.; and Wang, Y. 2024. Rethinking Optimization and Architecture for Tiny Language Models. *arXiv preprint arXiv:2402.02791*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; and Wei, F. 2022. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. arXiv:2208.10442.

Wang, Y.; Chen, H.; Tang, Y.; Guo, T.; Han, K.; Nie, Y.; Wang, X.; Hu, H.; Bai, Z.; Wang, Y.; et al. 2023. PanGu-π: Enhancing Language Model Architectures via Nonlinearity Compensation. *arXiv preprint arXiv:2312.17276*.

Xu, L.; Lu, X.; Yuan, C.; Zhang, X.; Xu, H.; Yuan, H.; Wei, G.; Pan, X.; Tian, X.; Qin, L.; et al. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.

Yuan, Z.; Li, Z.; Huang, W.; Ye, Y.; and Sun, L. 2024. TinyGPT-V: Efficient Multimodal Large Language Model via Small Backbones. arXiv:2312.16862.

Yuan, Z.; Li, Z.; and Sun, L. 2023. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang, Y.; Zhang, R.; Gu, J.; Zhou, Y.; Lipka, N.; Yang, D.; and Sun, T. 2024. LLaVAR: Enhanced Visual Instruction Tuning for Text-Rich Image Understanding. arXiv:2306.17107.

Zhao, B.; Wu, B.; He, M.; and Huang, T. 2023. SVIT: Scaling up Visual Instruction Tuning. arXiv:2307.04087.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.

Zhu, Y.; Zhu, M.; Liu, N.; Ou, Z.; Mou, X.; and Tang, J. 2024. LLaVA-Phi: Efficient Multi-Modal Assistant with Small Language Model. arXiv:2401.02330.