frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 01 September 2025
DOI 10.3389/fcomp.2025.1626346

OPEN ACCESS

EDITED BY
J. de Curtò,
Barcelona Supercomputing Center, Spain

REVIEWED BY
Lei Chen,
Nanjing Forestry University, China
Rohit Shukla,
University of Wisconsin-Madison,
United States

*CORRESPONDENCE
Juan Hao
hj2008@hebiace.edu.cn

RECEIVED 10 May 2025
ACCEPTED 13 August 2025
PUBLISHED 01 September 2025

CITATION
Han Z, Liu X and Hao J (2025) LLaVA-GM: lightweight LLaVA multimodal architecture. *Front. Comput. Sci.* 7:1626346. doi: 10.3389/fcomp.2025.1626346

LLaVA-GM: lightweight LLaVA multimodal architecture

Zhiyin Han, Xiaoqun Liu and Juan Hao*

College of Information Engineering, Hebei University of Architecture, Zhangjiakou, China

Multimodal large-scale language modeling has become the mainstream approach in natural language processing tasks and has been applied to various cross-modal fields such as image description and visual question answering. However, large-scale language modeling has high computational complexity and a large operational scale, which presents significant challenges for deployment in many resource-constrained scenarios. To address such problems, a lightweight multimodal framework, LLaVA-GM, is proposed, based on LLaVA, which can be deployed on devices with low resource requirements and has greatly reduced model parameters. It can also be tested on common VQA tasks and achieves good performance. The main contributions and work are as follows: First, it is found that the backbone of the Vicuna language model in LLaVA is too redundant. When fine-tuning downstream tasks, a very small amount of data sets is difficult to affect the language model. It is replaced with a new Gemma language model, thereby achieving fast task-specific adaptation with fewer parameters and data. Second, in response to the problem of information redundancy, the MoE mixed expert model is introduced. This model can be used in combination with itself, combining the MoE mixed expert model with Gemma to reduce the amount of computation while maintaining performance. Directly training the entire model will lead to a decline in performance. A multi-stage training strategy is adopted to maintain performance. First, the MLP layer is trained for visual adaptation, then the entire Gemma model is trained to improve multimodal capabilities, and finally only the MoE layer is trained for sparsification to ensure a smooth transition from dense models to sparse models. The experiment was tested on multiple VQA datasets and achieved good performance, confirming the potential of this compact model in downstream multimodal applications.

KEYWORDS

lightweight, LLaVA, Gemma, sparse expert, deep learning

1 Introduction

Large language models have shown excellent performance in various natural languages processing tasks, such as text generation (Stiennon et al., 2020), machine translation (Zhao et al., 2023), and question-answering systems (Kolomiyets and Moens, 2011). With technological advancements, their applications have expanded into cross-modal fields, including image captioning (Dong et al., 2021) and visual question answering (VQA). To further enhance the performance of LLMs, researchers have adopted multiple strategies. On one hand, increasing the model's parameter scale and training data volume can enhance its expressive and generalization abilities. On the other hand, leveraging techniques such as image encoders (Alsayed et al., 2023) and visual projection layers strengthens the visual perception of language models, enabling more effective processing of visual-language fusion tasks. However, large-scale models bring about issues such as high computational complexity and resource demands. Training and deploying these models require substantial computing resources, resulting in high hardware costs and the need for specialized parallel computing equipment and optimization techniques. This poses a

significant challenge for applications requiring rapid iteration and updates. For specific downstream tasks like specialized Q&A or image captioning, a smaller, specialized model may be more suitable. It can be optimized for particular tasks through training on relevant data, achieving efficient and precise processing. This approach not only reduces computational costs but also enhances the model's relevance and practicality, offering more effective solutions for specific domains. We chose the LLaVA large model as our research entry point. It consists of a pre-trained visual encoder and a large language model, connected by a simple linear layer that maps to the language embedding space. This modular design reduces architectural and training complexity, making the model easy to implement and extend. To optimize LLaVA for VQA tasks and achieve a lightweight model for easy deployment in downstream tasks, we made the following improvements:

1. We found that Vicuna, the language model in LLaVA, has a large architecture that is not conducive to specialized task improvement. Through experiments, we replaced it with the smaller Gemma speech model and trained it. We discovered that this smaller model responds more quickly to fine-tuning for specific tasks.
2. We gradually replaced FFNN layers with MoE layers in LLaVA and combined them with the Gemma language model to introduce sparsity, exploring the potential of small models in multimodal tasks, especially in resource-constrained scenarios. This approach maintains most of the model's performance while improving efficiency.
3. We used a multi-stage training strategy. First, we trained only the MLP layers to adapt to visual inputs. Then, we trained the entire Gemma model to build multimodal capabilities. Finally, we trained just the MoE layers to achieve sparsification, ensuring a smooth transition from a dense to a sparse model and leveraging Gemma's compactness to boost training efficiency. We've developed a lightweight LLaVA multimodal architecture for VQA tasks.

As shown in Figure 1, the LLaVA-GM-2B model achieves a high object hallucination benchmark score with few active parameters, indicating better performance with images closer to the top-left corner of the coordinate axis.

2 Related work

In recent years, large vision-language models (LVLMs) have made significant progress in visual-language tasks by integrating powerful language models with visual encoders. Models like OpenAI's GPT series (Roumeliotis and Tselikas, 2023), DeepSeek (Guo et al., 2024) from Hugging Face, and Google's Gemini (Menger and Keiper, 2000) series have been widely used in image captioning (Bernardi et al., 2016), visual question answering (VQA) (Lan et al., 2023), and cross-modal reasoning (Guan et al., 2023). CLIP (Radford et al., 2021) laid the foundation for multimodal tasks through contrastive learning on large-scale image-text pairs, but focuses on global feature alignment. BLIP (Li et al., 2022) improved this by jointly optimizing image and text encoders, enhancing performance in image captioning and VQA. BLIP-2 (Li et al., 2023) further reduced computational costs by using a pre - trained image encoder [e.g., ViT (Khan et al., 2022)] and a few visual mapping layers to inject visual features into large language models, while LLaVA (Liu et al., 2024) achieved efficient collaboration in visual - language tasks using a pre - trained visual encoder (CLIP-ViT) (Yang et al., 2024) and a large language model (e.g., LLaMA) (Touvron et al., 2023), projecting image features into the language model's embedding space. However, LLaVA-OneVision (Li et al., 2024), though demonstrating strong video understanding and cross-scene capabilities, comes with high costs. The performance improvements of these models often depend on expanding model size and dataset scale. For example, increasing the parameters of the visual encoder or using a larger language model [like GPT-4 (Achiam et al., 2023)] can achieve higher accuracy in downstream tasks, but also increases computational costs. These dense models require full forward propagation for each token, leading to high inference costs and making deployment difficult in resource-constrained scenarios. Moreover, the computational complexity grows exponentially in per-pixel tasks. The introduction of MoE sparse matrix models offers a solution. MoE's (Hwang et al., 2023) core idea is to divide the model into multiple expert subnetworks and dynamically select active experts via a routing mechanism, achieving computational sparsity. For instance, the Switch Transformer (Fedus et al., 2022) activates only the top 1 expert for each input, successfully scaling the model to trillion-parameter levels while maintaining inference efficiency. GShard and GLaM further optimized the MoE architecture by introducing expert parallelism (Zhou et al., 2022) and load balancing strategies, enabling sparse models to perform well in large-scale language tasks. LLaVA-Gemma (Hinck et al., 2024) was the first to integrate Gemma into the LLaVA multimodal model, reducing the model size but not the total parameters. MoE-LLaVA (Lin et al., 2024) fine-tuned the LLaVA model with MoE, and Deepseekmoe (Dai et al., 2024) also adopted this architecture, laying the foundation for model lightweight. According to Jin et al. (2024), various algorithms for hardware-efficient multimodal LLM were presented. At the same time, the keyword enhancement and self-supervised contrastive learning techniques of Chen et al. (2024), Chen and Zhu (2025), and Chen et al. (2025) have a certain promoting effect on the research. Based on this, we propose the LLaVA-GM model architecture, which retains most of the model's performance while significantly reducing parameters.

3 Methods

3.1 Language models

Our work aims to boost model performance in diverse tasks, especially VQA and multimodal benchmarks, through lightweight design and sparse computation. To lighten the model, we compared major multimodal architectures like FLAVA (Singh et al., 2022), LAMM (Yin et al., 2023), CLIP (Radford et al., 2021), and BLIP (Li et al., 2022). CLIP excels in image-text matching but is large. FLAVA offers multimodal pre-training advantages, yet it is computationally intensive. LAMM integrates image and text info effectively, but underperforms in specific tasks. Consequently, we chose LLaVA as

Abbreviations: LLaVA, Large Language and Vision Assistant; VQA, Visual Question Answering; MoE, Mixture of Experts; LVLM, Large Visual Language Model; MHSA, Multi-Head Self-Attention; FFNN, Feed-Forward Neural Network.

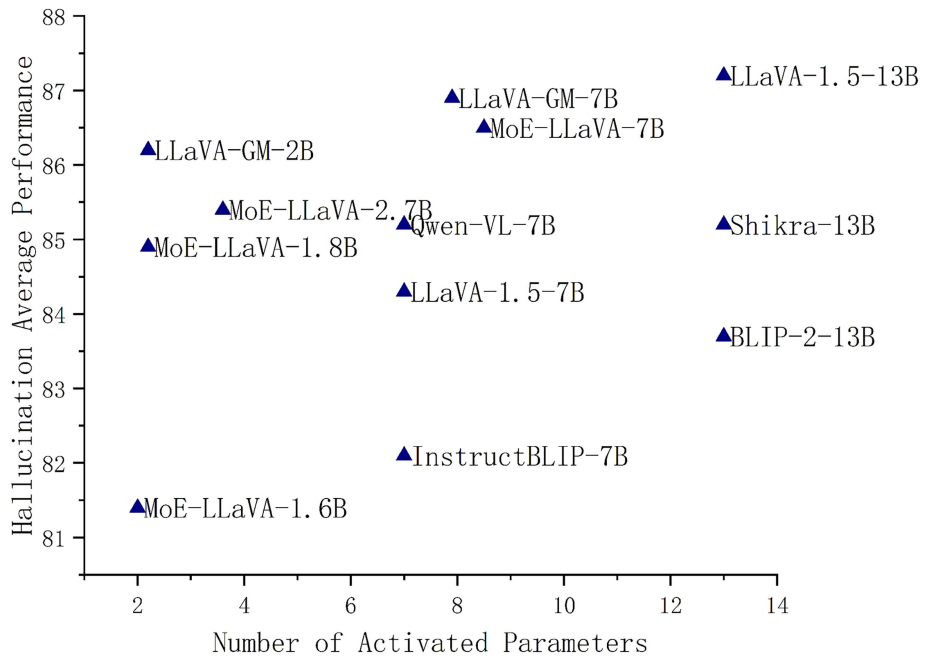

FIGURE 1
Compare LLaVA-GM with open-source LVLMs on the object hallucination benchmark and active parameter size.

the main framework due to its efficient and simple mapping mechanism that reduces model complexity while maintaining performance. After determining the main model framework, we focus on streamlining the language model to achieve further model lightweight while preserving its original architecture as much as possible. We conduct a comprehensive evaluation of several candidate models, including Llama (Touvron et al., 2023) and the GPT (Achiam et al., 2023) series, and finally select Google's Gemma (Kolomiyets and Moens, 2011), which offers a smaller size and decent performance with flexible options of 2b and 7b scales. To assess Gemma's performance across tasks, we design experiments covering GQA (Ainslie et al., 2023), VQAv2.0 (Goyal et al., 2017), ScienceQA-IMG (Lu et al., 2022), MMBench (Liu et al., 2024), and MME (Liang et al., 2024). By experimenting with and evaluating Gemma-2b and Gemma-7b, we analyze the impact of model size on performance. As shown in Figure 2, the new language model Gemma is integrated into LLaVA. Image inputs are processed by the CLIP ViT-L/336px visual encoder. Text inputs are tokenized, fed into Gemma, and mapped to a 2-layer MLP for processing.

3.2 Sparse architecture

In multimodal tasks, we aim to maintain performance while reducing computational costs through sparse activation for efficient inference, avoiding the full parameter activation of dense models. However, directly replacing the feedforward networks in the Transformer with MoE layers can significantly degrade model performance. Therefore, based on experimental findings, we gradually replace some FFNNs with MoE layers to reduce activated parameters. We also use CLIP-ViT-L-14 to process image inputs for tasks like VQA. The model's processing flow starts with the input layer. Given an RGB image $v \in \mathbb{R}^{H\times W\times 3}$, where H and W are the original resolutions (usually 336 × 336), the visual encoder processes the input image to obtain a sequence of visual tokens

$$Z = [z_1, z_2, \cdots, z_P] \in \mathbb{R}^{P\times C} \tag{1}$$

Subsequently, In Equation 1 the visual projection layer "f "maps $Z \in \mathbb{R}^{P\times C}$ to $V \in \mathbb{R}^{P\times D}$, where P is the visual token sequence length, C is the encoder output dimension, and D is the LLM's hidden dimension. Meanwhile, text inputs (e.g., "What's in the picture?") pass through the word embedding layer g to obtain projected sequence tokens

$$T = [t_1, t_2, \cdots, t_N] \in \mathbb{R}^{N\times D} \tag{2}$$

Here, In Equation 2 N denotes the sequence length of text tokens. Then, the visual tokens V and text tokens T are combined to form the input sequence.

$$[V; T] \in \mathbb{R}^{(P+N)\times D} \tag{3}$$

In Equation 3, We only train the visual projection layer f, while keeping the LLM and embedding layer g in their pre-trained states. The LLM is composed of stacked multi-head self-attention (MHSA) and feedforward network (FFN) blocks, each with layer normalization (LN) and residual connections. The formula is:

$$Layer(x) = LN(MHSA(x) + x) \rightarrow LN(\text{FFNN}(\ ) \text{ or } \text{MoE}(\ ) + \ldots) \tag{4}$$

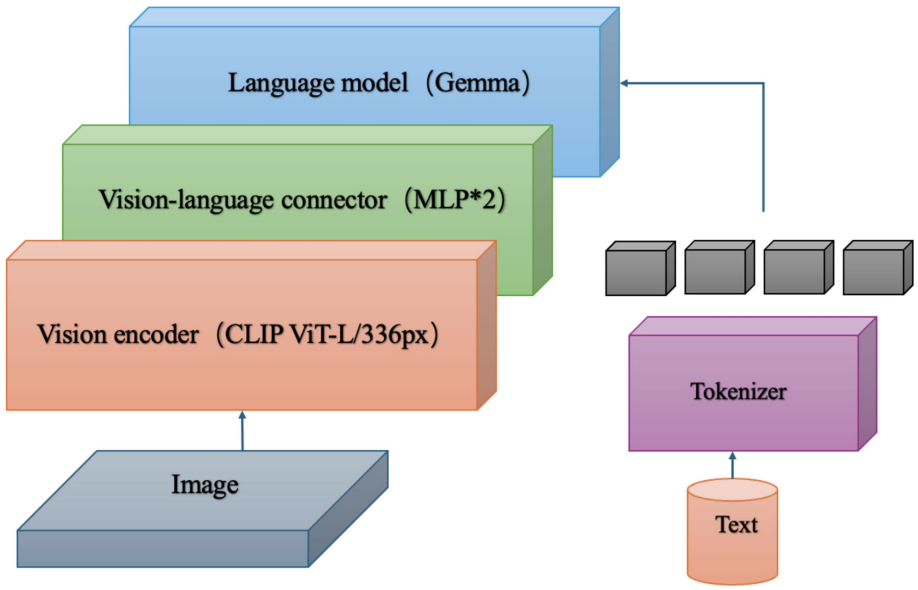

FIGURE 2
Overall sketch of the model.

We replace some FFNNs with MoE layers, each containing a router and four experts in Equation 4. The router, implemented as a linear layer, calculates expert scores based on the input x. Where … denotes omitted additional parameters, including visual features, language embeddings, attention masks, and model configuration parameters (such as the number of heads and hidden layer dimensions). Together, these parameters define the multimodal fusion process.

$$f(x) = W_{\text{router}} \cdot x \tag{5}$$

In Equation 5, The router calculates expert scores based on input × and uses softmax to obtain probabilities.

$$\mathcal{P}(x)_i = \frac{e^{f(x)_i}}{\sum_j e^{f(x)_j}} \tag{6}$$

In Equation 6, The router selects the top-k experts via softmax. Each expert is an FFNN, and the output is a weighted sum

$$MoE(x) = \sum_{i=1}^{2} \mathcal{P}(x)_i \cdot e_i(x) \tag{7}$$

In Equation 7, A linear transformation maps the output back to the vocabulary. Only the top-k experts are activated during inference to reduce parameters. The alternate replacement of FFNN and MoE balances generality and task specificity. As a submodule of LLN, the feedforward neural network (FNN) contains two layers of linear transformation and ReLU activation to enhance the expressiveness of unimodal features: the MoE module is integrated after LLN and dynamically selects the expert network to process the output of FNN through the gating network. Dynamic routing enables a flexible and efficient model, combining high performance with low cost. As shown in Figure 3, both image and text tokens are processed by MHSA and feedforward networks.

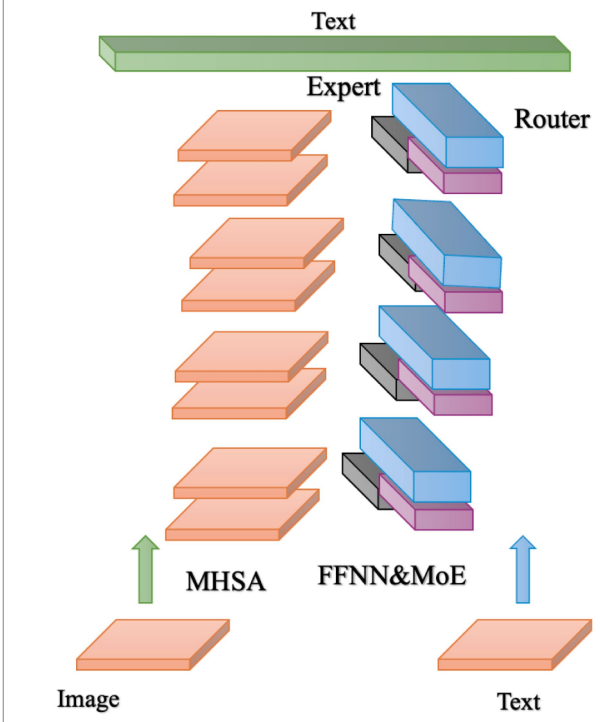

FIGURE 3
MoE mixing treatment stage flow.

However, half of the feedforward networks are replaced with MoE. After normalization, the gating network selects appropriate routes for processing, and finally, the top-k selection determines the corresponding expert outputs.

3.3 Training strategies

As mentioned in the previous chapter, we have designed a three-stage training process for our model to achieve lightweight and high-accuracy goals, as direct MoE layer replacement or model training fails to meet the requirements.

Stage 1: Focus on adapting visual inputs. Only the MLP visual mapping layer is trained, with Gemma and the word embedding layer frozen. The aim is to align image tokens V with the LLM's input space, treating them as pseudo-text tokens. Training data consists of image descriptions, and f is optimized to generate a compatible V.

Stage 2: Train the entire Gemma model to build multimodal capabilities. Unfreeze the MLP layer f and the word embedding layer g and use multimodal instruction data for training. Outputs are generated through a linear layer, transforming Gemma from a language model to a multimodal LVLM and establishing visual-language fusion.

Stage 3: Achieve sparsification by only training MoE layers. Replace half of the MoE layers, freeze half of the FFNN and MHSA, and only train the routers and experts of the MoE layers, with top-2 expert outputs selected. The remaining experts remain inactive, and this sparse path design improves efficiency.

As shown in Figure 4, it contains four stages: visual encoding, language encoding, language-visual fusion layer (LLN), and mixture of experts (MoE) module. Visual encoding uses pre-trained CLIP-ViT-L-336px, which has a high number of parameters but does not require retraining. The input image is 336X336px, and the language encoding is based on Gemma-2B/7B. The amount of calculation increases with the size of the model. The fusion module LLN module, fuses multi-modal features through multi-head attention, and the FNN submodule enhances feature expression. The MoE module reduces the amount of calculation through sparse activation and optimizes the gated network to ensure expert utilization. Make it load balanced. We conduct training in three stages. Image information is split and fed into the visual encoder in blocks. We simplify the text encoder to directly input tokens into the MLP, highlighting the alignment of image tokens and clarifying the structure. In the second stage, we unfreeze the MLP and train Gemma to build its multimodal capabilities. Finally, we use sparse expert selection to generate outputs.

4 Experiment

4.1 Data set and evaluation index

VQAv2.0 (Goyal et al., 2017) is a classic visual question answering dataset based on COCO and abstract scene images, featuring 265,016 images and over 1.1 million open-ended questions. Each question comes with 10 real answers and 3 plausible but potentially wrong options,

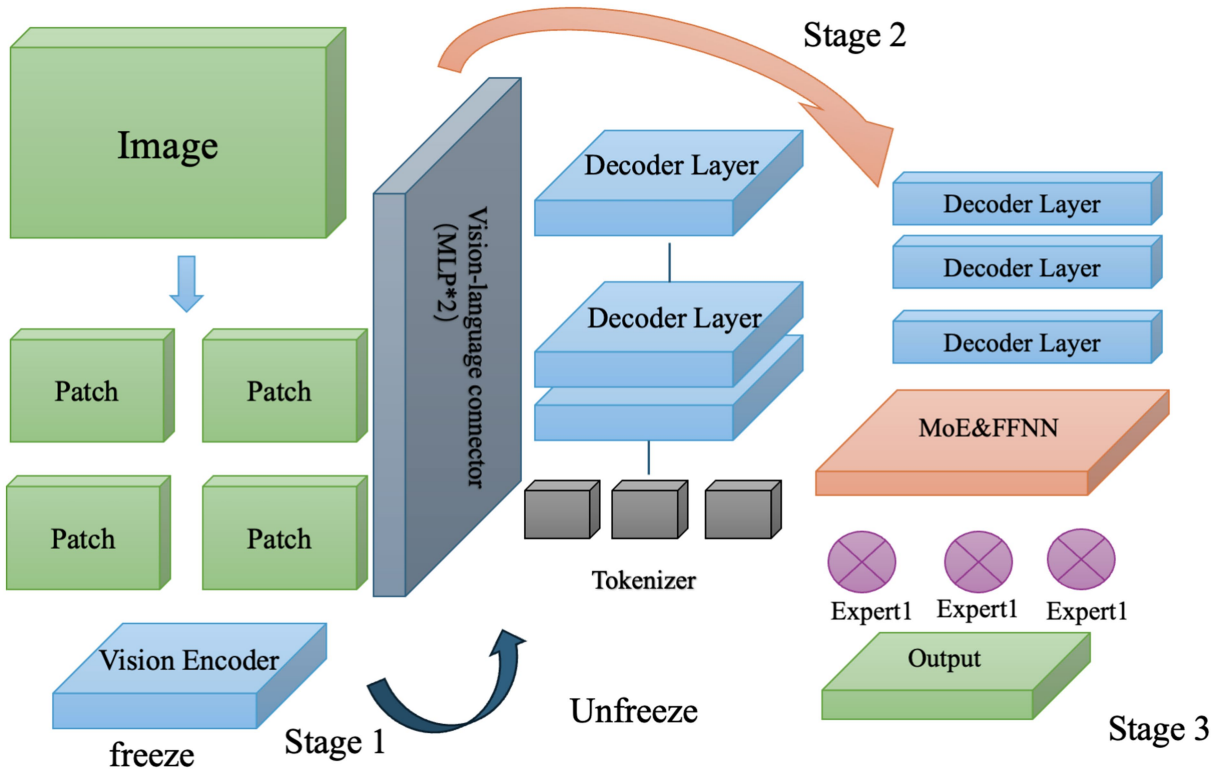

FIGURE 4
Three stages of training.

testing the model's understanding of image content, language, and common sense.

ScienceQA-IMG (Lu et al., 2022) focuses on multimodal questions in scientific fields, containing 21,208 multiple-choice questions across 26 subjects, including natural, language, and social sciences. Each question is accompanied by detailed explanations, making it the first large-scale annotated dataset with lectures and explanations. It emphasizes scientific knowledge and reasoning.

GQA (Ainslie et al., 2023), based on Visual Genome, offers over 1 million questions and scene graph annotations with about 260,000 images. Questions are linked to objects, attributes, and relationships in images, with controlled and balanced question generation to reduce language ambiguity.

MMBench (Liu et al., 2024) is a newer benchmark with around 3,000 multiple-choice questions, covering 20 fine-grained ability dimensions (e.g., identity and attribute reasoning). Derived from an extension of ScienceQA, it innovatively assesses model predictions against options using ChatGPT for a more robust evaluation.

MME (Liang et al., 2024) aims to comprehensively evaluate multimodal models with various task types, such as perception and reasoning. Despite its small size, it features diverse tasks like "Does the object in the picture exist?" and "Answer questions based on charts," emphasizing objectivity and reproducibility.

Together, these datasets provide a comprehensive test of model performance, from basic understanding to complex reasoning, supporting the development of efficient and high-performing multimodal models.

4.2 Experimental settings and analysis

Training uses 3 V100 vGPUs - 32GB (96GB total memory) running on Ubuntu 22.04, PyTorch 2.1.0, Python 3.10, CUDA 12.1, The training power consumption of LLaVA-GM-2B is approximately 32 kWh, and that of LLaVA-GM-7B is approximately 69 kWh, which is a significant advantage over the original 7B and 78 kWh, verifying its deployment potential in mobile and embedded systems. During training, we used a learning rate of 2e-5, a batch size of 32, an AdamW ($\beta1 = 0.9$, $\beta2 = 0.999$, wd = 0.01) optimizer, and a cross-entropy loss. We loaded the Gemma-2B/7B pre-trained weights of Hugging Face and the CLIP-ViT-L-336px visual encoder and iterated for 3 epochs. Deployment on NVIDIA Orin NX shows that it benefits from its modular design and MoE sparse activation mechanism. Our method reduces memory access overhead and redundant computation. The model latency of LLaVA-GM 2B is further reduced to 0.5–0.6 s, with a power consumption of about 5.0 W and a video memory occupancy of only 1.9 GB. The 7B model occupies 4.2GB of video memory and consumes about 12 W of power. Even after distillation, the model has too many visual tokens and high inference latency. LLaVA-MoD (Shu et al., 2024) consumes 9.5 W of power and has 4.8GB of video memory. Our models have reduced resources to varying degrees. LLaVA-GM optimizes hardware efficiency while maintaining high accuracy, making it suitable for resource-constrained embedded platforms.

4.3 Experimental results

We found that the original language model, Vicuna in LLaVA is too bulky for deployment and fine-tuning in downstream tasks. In contrast, Gemma was trained with twice the data volume and incorporates knowledge distillation and architectural improvements, giving it an edge in handling complex tasks. Since our aim is to serve computation-constrained devices, we experimented as shown in Table 1. All our experiments are based on the average of 3 independent experiments and have been processed with a standard deviation.

We tested several LLaVA-based models using general datasets and GFLOPs as parameters. The visual framework was uniformly set as CLIP. As shown in Table 2, larger LLMs generally yield better performance. Most datasets show that the LLaVA model with Vicuna-13B performs the best. However, our goal is to achieve good performance with fewer parameters. Our LLaVA-GM performs comparably to the 7B model on various datasets while having less than half the GFLOPs, proving its superiority.

In Table 3, we compare expert-based models ranging from 1.6B to 7B. All models have 4 experts, activate Top-2, and half of their layers are replaced with MoE layers. LLaVA-GM-2B has only 2.2B active parameters, outperforming MoE-LLaVA-1.6B with similar sparsity.

Gemma-2b performs well on simple factual questions, showing competitiveness in specific scenarios. In the VQAv2 task, model size significantly impacts generalization and robustness. Gemma-7B offers more stable performance with diverse images and questions, while Gemma-2B may have larger errors with rare ones. Gemma-7B generally outperforms Gemma-2B, especially in complex reasoning and multimodal fusion tasks. For sentiment analysis, Gemma-7B captures emotional cues more finely and makes more accurate judgments. Considering model size, choose Gemma-7B for fine-grained tasks and Gemma-2B for daily tasks. LLaVA-Gemma (Hinck et al., 2024) only replaces the language model without introducing sparsity, resulting in high computational cost. The MoE of MoE-LLaVA (Lin et al., 2024) has low accuracy after sparsification. LLaVA-GM significantly improves the accuracy of the architecture and reduces the number of parameters required by the model through a more efficient MoE layer design (12–16 layers, 2.2B-7.9B activation parameters) and multi-stage

TABLE 1 Generic model comparison.

Methods	LLM	Finetune size	VQA2.0	GQA	SQA-IMG	MMbench	MME
BLIP-2 (Li et al., 2023)	Vicuna-13B	–	65.0	41.0	61.0	–	1,293.8
InstructBLIP (Huang et al., 2023)	Vicuna-7B	1.2 M	-	49.2	60.5	36.0	-
Shikra (Chen et al., 2023)	Vicuna-13B	5.5 M	77.4	–	–	58.8	–
Qwen-VL (Jiao et al., 2024)	LLaMA-7B	1.4B	**78.8**	59.2	**67.1**	38.2	–
LLaVA1.5 (Liu et al., 2024)	Vicuna-7B	**665 K**	78.5	**62.0**	66.8	**64.3**	**1,510.7**
LLaVA-GM	Gemma-2B	**665 K**	78.2	61.3	65.3	62.3	1,498.3

The bold parts represent the optimal values under this indicator.

fine-tuning training. These improvements make LLaVA-GM more suitable for application on resource-constrained devices.

The results generated by the model are shown below. Although the model size has been compressed to make it suitable for resource-constrained devices, it can be seen from the figure below that the quality of the generated text and the quality of the VQA command question and answer are not inferior at all.

As shown in Figure 5, LLaVA-GM can correctly identify the specific location, orientation, and color of objects in the image, introduce the scene, and reduce the description of non-important information and reduce redundancy.

As shown in Figure 6, we input a few simple math problems into the model and ask it to answer them. After quick thinking, the model answers the questions in sequential order without any redundant information.

As shown in Figure 7, this is a virtual driver's license photo, which contains various images of the driver. In the original LLaVA paper, it generates answers in JSON format. The two paragraphs on

TABLE 2 Compare similar models.

Methods	LLM	VQA2.0	GQA	SQA-IMG	MMbench	MME	GFLOPs
LLaVA-1.5 (Liu et al., 2024)	Vicuna-7B	78.5	62.0	66.8	64.3	1510.7	8,027
LLaVA-1.5 (Liu et al., 2024)	Vicuna-13B	**80.5**	63.3	**71.6**	**67.7**	**1531.3**	14,927
LLaVA-1.5 (Liu et al., 2024)	Phi-2.7B	67.5	–	68.4	58.8	1135.7	–
MoE-LLaVA (Lin et al., 2024)	Vicuna-7B	78.6	59.2	67.1	65.3	–	8,027
MoE-LLaVA (Lin et al., 2024)	Phi-2.7B	71.4	–	68.4	65.2	1423.5	3,754
TinyLLaVA (Zhou et al., 2024)	Phi-2B	72.4	58.4	67.2	66.1	1434.3	–
LLaVA-MoD (Shu et al., 2024)	Qwen1.5B	75.8	58.8	69.2	64.4	–	–
LLaVA-GM	Gemma-2B	76.4	61.3	69.3	62.3	1498.3	**3,623**
LLaVA-GM	Gemma-7B	76.8	**62.4**	70.1	63.1	1502.5	7,512

The bold parts represent the optimal values under this indicator.

TABLE 3 Effect of MoE parameter settings and adjustments on modeling.

Methods	Experts	Top-k	MoE layers	Layers	Total param	Activated param
MoE-LLaVA-1.6B (Lin et al., 2024)	4	2	16	32	2.9B	2.0B
MoE-LLaVA-1.8B (Lin et al., 2024)	4	2	12	24	3.1B	2.2B
MoE-LLaVA-2.7B (Lin et al., 2024)	4	2	16	32	5.3B	3.6B
MoE-LLaVA-7B (Lin et al., 2024)	4	2	16	32	14.3B	8.5B
LLaVA-GM-2B	4	2	12	24	3.3B	2.2B
LLaVA-GM-7B	4	2	16	32	13.4B	7.9B

Q: What's in the picture?

A: A rider rides a horse in an outdoor equestrian arena with a dirt surface surrounded by green trees and a wooden fence, the rider is wearing standard equestrian attire consisting of a black jacket, white pants, black boots and a helmet. She is seated on the back of a brown horse with white markings on its limbs.

FIGURE 5
Example of the model's ability to perceive the outdoor environment (reproduced from MS COCO datasets, https://cocodataset.org/, licensed under CC BY 4.0).

the left of Figure 7 are descriptions of a driver's license generated by LLaVA and LLaVA1.5, while the right side is a description of the driver's license information by LLaVA-GM. It can be seen that LLaVA and LLaVA1.5 have read errors in height, weight, birthday, and other information, but our model corrected the errors and gave the correct answer. The accuracy rate has increased by nearly 30%, which reflects the accuracy and sophistication of the model in image and text recognition.

5 Conclusion

5.1 Significance of work

Our work explores the trade-off between computational efficiency and multimodal understanding for small models and elucidates how model size affects performance on different tasks. This provides valuable references and benchmarks for future

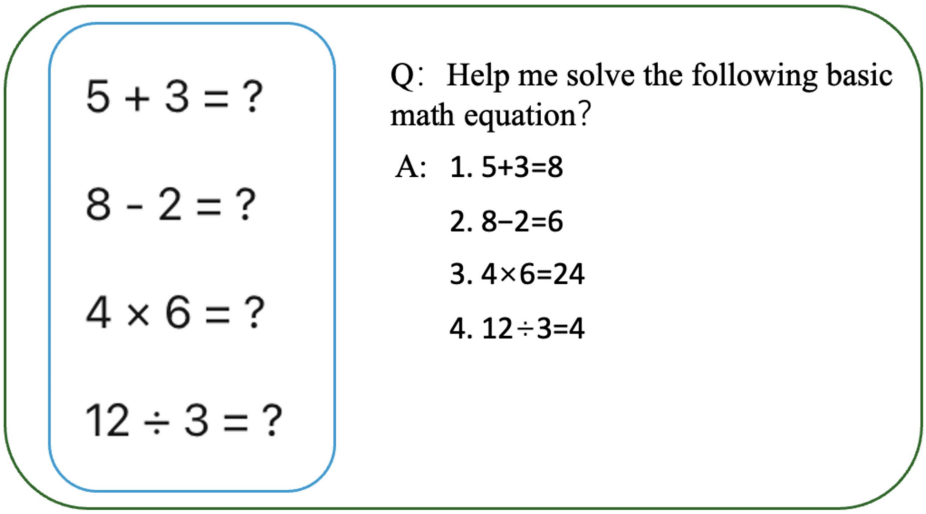

FIGURE 6
The model's answers to some simple math questions.

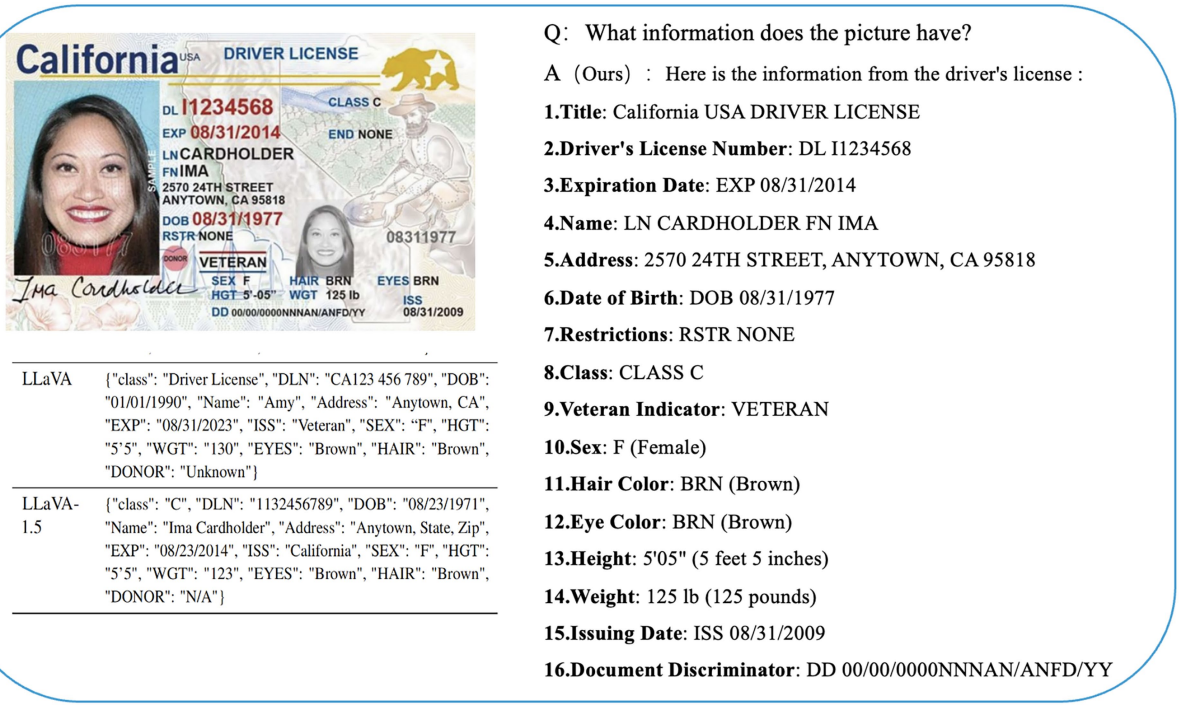

FIGURE 7
An example of a virtual driver's license recognition [reproduced from Yang et al. (2023), (2309.17421) The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), licensed under CC-BY 4.0].

research on small visual language models, contributing to further development and innovation in the field. Our research helps guide the choice of model size for optimal performance and efficiency in different scenarios. Compared to the dense fusion of LLaVA-1.5, LLaVA-GM achieves explicit "disentanglement" on a small model for the first time through a modular four-stage pipeline (visual encoding → language encoding → LLN fusion → MoE output). The model decomposes multimodal processing into clear stages (visual encoding, language encoding, language-visual fusion layer, and MoE module), and the functions and outputs of each stage can be analyzed independently, which enhances the transparency of model behavior. In the test of the dataset, an average of 2 to 3 experts are activated, reducing the amount of redundant calculation by about 70% while maintaining performance.

5.2 Limitation

a) Insufficient visual representation: the first stage freezes the weights of CLIP-ViT-L and only trains the MLP adapter, resulting in the loss of fine-grained visual details. It lags behind dynamic segmentation schemes in tasks that require high resolution or document understanding, such as MME and OCRBench.
b) Generalization bottleneck: the performance is significantly degraded on diverse and cross-domain datasets (such as DocVQA and Video-VQA), exposing the over-reliance of multi-stage training on early visual features.
c) Extremely low resource latency: although MoE has reduced GFLOPs by 70%, routing overhead still introduces additional latency on low-end mobile chips, affecting the real-time interactive experience.

5.3 Future work

We plan to make improvements in the following three aspects:

a) Optimize visual encoding - enhance detail perception through lightweight fine-tuning or introducing multi-granularity encoders such as DINOv2/SigLIP2;
b) Expand training data - use a unified protocol similar to LLaVA-MOE to continue fine-tuning on high-resolution, document, and video data to improve cross-domain robustness;
c) Upgrade gate efficiency - use low-rank routing + quantized expert weights to enable MoE to achieve zero latency penalty on extremely constrained devices.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

ZH: Conceptualization, Data curation, Resources, Validation, Writing – original draft. XL: Conceptualization, Funding acquisition, Project administration, Validation, Writing – review & editing. JH: Formal analysis, Investigation, Methodology, Software, Supervision, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by Hebei Provincial Science and Technology Program (grant number 20470302D).

Acknowledgments

The authors thank researchers from Hebei University of Architecture for their research support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., and Aleman, F. L. (2023). Gpt-4 technical report.

Ainslie, J., Lee-Thorp, J., De Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints.

Alsayed, A., Arif, M., Qadah, T. M., and Alotaibi, S. (2023). A systematic literature review on using the encoder-decoder models for image captioning in English and Arabic languages. *Appl. Sci.* 13:10894. doi: 10.3390/app131910894

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., et al. (2016). Automatic description generation from images: a survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.* 55, 409–442. doi: 10.1613/jair.4900

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. (2023). Shikra: Unleashing multimodal llm's referential dialogue magic.

Chen, L., and Zhu, G. (2025). Self-supervised contrastive learning for itinerary recommendation. *Expert Syst. Appl.* 268:126246. doi: 10.1016/j.eswa.2024.126246

Chen, L., Zhu, G., Liang, W., Cao, J., and Chen, Y. (2024). Keywords-enhanced contrastive learning model for travel recommendation. *Inf. Process. Manag.* 61:103874. doi: 10.1016/j.ipm.2024.103874

Chen, L., Zhu, X., and Zhu, G. (2025). Exploiting attributes and keywords for session-based recommendation with multi-view graph neural network. *Expert Syst. Appl.* 13:128990. doi: 10.1016/j.eswa.2025.128990

Dai, D., Deng, C., Zhao, C., Xu, R. X., Gao, H., and Chen, D. (2024). Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models.

Dong, X., Long, C., Xu, W., and Xiao, C. (2021). "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2615–2624.

Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* 23, 1–39. doi: 10.48550/arXiv.2101.03961

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). "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* (6904–6913).

Guan, Q. L., Zheng, Y., Meng, L., Dong, L. Q., and Hao, Q. (2023). Improving the generalization of visual classification models across IoT cameras via cross-modal inference and fusion. *IEEE Internet Things J.* 10, 15835–15846. doi: 10.1109/JIOT.2023.3265645

Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., and Zhang, W. (2024). DeepSeek-coder: When the large language model meets programming--the rise of code intelligence.

Hinck, M., Olson, M. L., Cobbley, D., Tseng, S. Y., and Lal, V. (2024). Llava-gemma: accelerating multimodal foundation models with a compact language model.

Huang, J., Zhang, J., Jiang, K., Qiu, H., and Lu, S. (2023). Visual instruction tuning towards general-purpose multimodal model: A survey.

Hwang, C., Cui, W., Xiong, Y., Yang, Z., Liu, Z., Hu, H., et al. (2023). Tutel: adaptive mixture-of-experts at scale. *Proc. Mach. Learning Syst.* 5, 269–287. doi: 10.48550/arXiv.2206.03382

Jiao, Q., Chen, D., Huang, Y., Li, Y., and Shen, Y. (2024). Enhancing multimodal large language models with vision detection models: An empirical study.

Jin, Y., Li, J., Liu, Y., Gu, T., Wu, K., and Jiang, Z. (2024). Efficient multimodal large language models: A survey.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: a survey. *ACM Computing Surveys (CSUR)* 54, 1–41. doi: 10.1145/3505244

Kolomiyets, O., and Moens, M. F. (2011). A survey on question answering technology from an information retrieval perspective. *Inf. Sci.* 181, 5412–5434. doi: 10.1016/j.ins.2011.07.047

Lan, Y., Guo, Y., Chen, Q., Lin, S., Chen, Y., and Deng, X. (2023). Visual question answering model for fruit tree disease decision-making based on multimodal deep learning. *Front. Plant Sci.* 13:1064399. doi: 10.3389/fpls.2022.1064399

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). "Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning* (19730–19742). PMLR.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. *Int. Conf. n Mach. Learning* 2, 12888–12900.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., and Zhang, H. (2024). Llava-onevision: Easy visual task transfer.

Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., and Fu, Q. (2024). "A survey of multimodel large language models," in *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering* (405–409).

Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., and Jin, P. (2024). Moe-llava: Mixture of experts for large vision-language models.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., and Zhao, W. (2024). "Mmbench: is your multi-modal model an all-around player?," in *European Conference on Computer Vision* (216–233). Cham: Springer Nature Switzerland.

Liu, H., Li, C., Li, Y., and Lee, Y. J. (2024). "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 26296–26306).

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K. W., and Zhu, S. C. (2022). Learn to explain: multimodal reasoning via thought chains for science question answering. *Adv. Neural Inf. Proces. Syst.* 35, 2507–2521. doi: 10.48550/arXiv.2209.09513

Menger, F. M., and Keiper, J. S. (2000). Gemini surfactants. *Angew. Chem. Int. Ed.* 39, 1906–1920. doi: 10.1002/1521-3773(20000602)39:11<1906::AID-ANIE1906>3.0.CO;2-Q

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., and Agarwal, S. (2021). "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning* (8748–8763). PmLR.

Roumeliotis, K. I., and Tselikas, N. D. (2023). Chatgpt and open-ai models: a preliminary review. *Future Internet* 15:192. doi: 10.3390/fi15060192

Shu, F., Liao, Y., Zhuo, L., Xu, C., Zhang, L., and Zhang, G. (2024). Llava-mod: Making llava tiny via Moe knowledge distillation.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., and Rohrbach, M. (2022). "Flava: A foundational language and vision alignment model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (15638–15650).

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., et al. (2020). Learning to summarize with human feedback. *Adv. Neural Inf. Proces. Syst.* 33, 3008–3021. doi: 10.48550/arXiv.2009.01325

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., and Lacroix, T. (2023). Llama: Open and efficient foundation language models.

Yang, H., Xu, M., Sun, Z., Song, B., and Cheng, E. (2024). "CLIP-ViT detector: side adapter with prompt for vision transformer object detection," in *2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)* (pp. 1–8). IEEE.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C., Liu, L., et al. (2023). The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), [2309.17421] The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision).

Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., and Li, M. (2023). Lamm: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Adv. Neural Inf. Proces. Syst.* 36, 26650–26685. doi: 10.48550/arXiv.2306.06687

Zhao, Y., Zhang, J., and Zong, C. (2023). Transformer: a general framework from machine translation to others. *Mach. Int. Res.* 20, 514–538. doi: 10.1007/s11633-022-1393-5

Zhou, B., Hu, Y., Weng, X., Jia, J., Luo, J., and Liu, X. (2024). Tinyllava: A framework of small-scale large multimodal models.

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., and Zhao, V. (2022). Mixture-of-experts with expert choice routing. *Adv. Neural Inf. Proces. Syst.* 35, 7103–7114. doi: 10.48550/arXiv.2202.09368