World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

ViMoE: Vision Mixture of Experts with Multimodal Context Awareness

Adele Chinda *

Computer Science, Georgia State University, USA.

World Journal of Advanced Research and Reviews, 2026, 29(01), 1886-1901

Publication history: Received on 22 December 2025; revised on 28 January 2026; accepted on 31 January 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.29.1.0242

Abstract

Multimodal large language models (MLLMs) rely heavily on vision encoders to understand diverse image content. While recent approaches have explored combining multiple vision experts to address the limitations of single encoders, they typically perform image-level expert selection and fusion, ignoring the spatial heterogeneity within images where different regions may benefit from different experts. In this paper, we propose ViMoE (Vision Mixture of Experts with Multimodal Context Awareness), a novel MLLM that introduces three key innovations: (1) Token-Level Sparse Expert Activation (TLSEA) that enables different spatial tokens to utilize different expert combinations, allowing fine-grained, content-aware feature extraction; (2) Hierarchical Context Aggregation (HCA) that captures multi-scale visual context to guide expert routing at different granularities; and (3) Expert Confidence Calibration (ECC) that learns to estimate and calibrate expert contribution confidence to reduce noise from unreliable features. Through these innovations, ViMoE achieves more precise expert utilization by recognizing that a single image often contains diverse content requiring different visual expertise. Extensive experiments demonstrate that ViMoE achieves significant improvements over state-of-the-art methods across challenging multimodal benchmarks including MME, MMBench, and various VQA tasks, while maintaining computational efficiency through sparse activation patterns. Code is available at: https://arrel.github.io/vimoe/

Keywords: Vision Mixture of Experts; Token-level routing; Multimodal large language mode; Hierarchical context aggregation; Confidence calibration; Sparse expert activation

1. Introduction

Multimodal large language models (MLLMs) [33, 41, 3, 50] have demonstrated remarkable capabilities in understanding and reasoning about visual content. These models typically combine pre-trained vision encoders with large language models (LLMs) to enable sophisticated visual understanding. The CLIP vision encoder, trained on billions of image-text pairs, has become the de facto choice for most leading MLLMs due to its strong semantic understanding capabilities.

However, a single vision encoder cannot excel at all visual tasks. CLIP, while powerful for general image understanding, often struggles with fine-grained tasks such as document parsing, chart understanding, and precise object localization [66, 36]. This observation has motivated recent works to incorporate multiple task-specific vision experts into MLLMs. For instance, SPHINX integrates DINOv2 [51] for improved grounding, while vary [66] introduces specialized encoders for document understanding.

MoVA [77] represents a significant advancement by proposing a coarse-to-fine framework that first uses an LLM to select relevant vision experts based on the input image and instruction, then fuses selected expert features through a mixture-of-vision-expert adapter (MoV-Adapter). While effective, MoVA operates at the *image level*—all spatial tokens in an image utilize the same set of experts with identical weights. This design overlooks a crucial observation: different regions within a single image often contain diverse content that would benefit from different expert combinations.

* Corresponding author: Adele Chinda

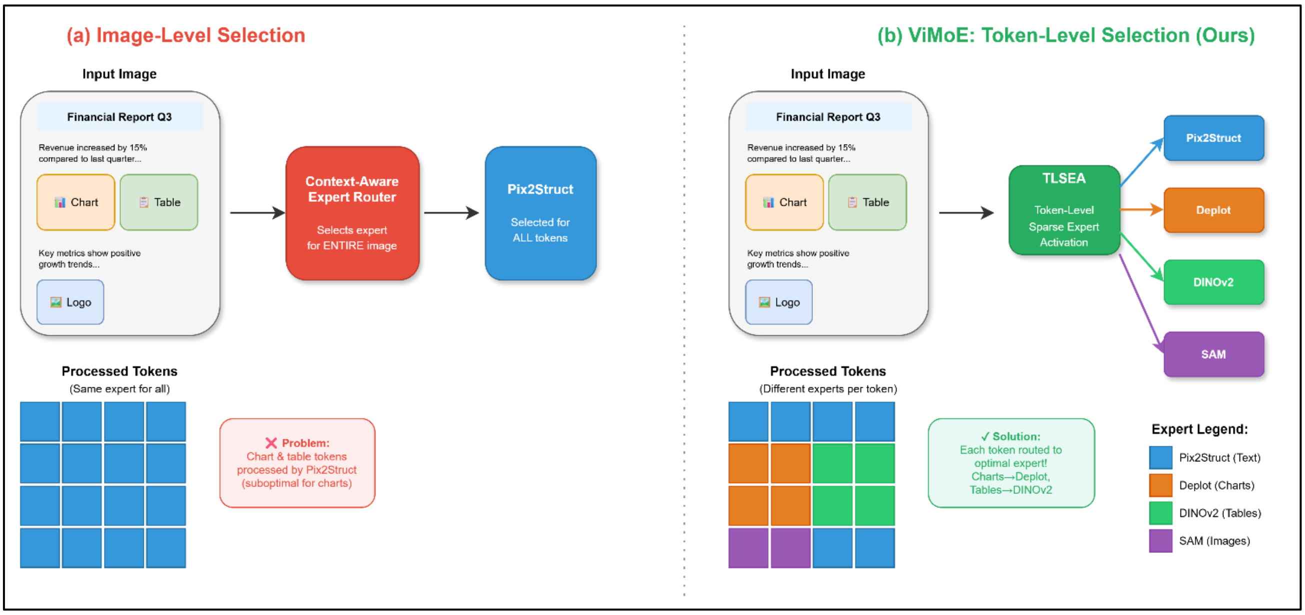

(a) Image-Level Selection
Input Image
Financial Report Q3
Revenue increased by 15% compared to last quarter...
Chart
Table
Key metrics show positive growth trends...
Logo
Context-Aware Expert Router
Selects expert for ENTIRE image
Pix2Struct
Selected for ALL tokens
Processed Tokens
(Same expert for all)
✕ Problem:
Chart & table tokens processed by Pix2Struct (suboptimal for charts)
(b) ViMoE: Token-Level Selection (Ours)
Input Image
Financial Report Q3
Revenue increased by 15% compared to last quarter...
Chart
Table
Key metrics show positive growth trends...
Logo
TLSEA
Token-Level Sparse Expert Activation
Pix2Struct
Deplot
DINOv2
SAM
Processed Tokens
(Different experts per token)
✓ Solution:
Each token routed to optimal expert!
Charts→Deplot,
Tables→DINOv2
Expert Legend:
Pix2Struct (Text)
Deplot (Charts)
DINOv2 (Tables)
SAM (Images)

Figure 1 Motivation of VIMOE. Unlike prior methods that perform image-level expert selection, VIMOE enables token-level sparse expert activation. In this example, a document image contains both text regions (better served by Pix2Struct) and chart regions (better served by Deplot). VIMOE routes different tokens to appropriate experts based on local content, achieving more precise knowledge extraction

Consider the example in Figure 1: a document page containing both text paragraphs and embedded charts. Image-level approaches would select experts based on global image characteristics, potentially choosing either document-focused experts (missing chart details) or chart-focused experts (degrading text recognition). The optimal strategy is to route text regions to document experts like Pix2Struct while routing chart regions to visualization experts like Deplot.

In this paper, we propose ViMoE (Vision Mixture of Experts with Multimodal Context Awareness), which introduces three novel components to address these limitations:

1.1. Token-Level Sparse Expert Activation (TLSEA)

Unlike image-level expert selection, TLSEA enables each spatial token to independently select and weight its expert contributions. This allows different image regions to utilize different expert combinations based on their local content, achieving fine-grained, content-aware feature extraction while maintaining computational efficiency through sparsity.

1.2. Hierarchical Context Aggregation (HCA)

Expert routing decisions should consider both local details and global semantics. HCA aggregates visual context at multiple scales and fuses it with textual context to provide rich, multi-granular information for routing decisions. This contrasts with MoVA's single-scale global average pooling approach.

1.3. Expert Confidence Calibration (ECC)

Not all expert contributions are equally reliable. ECC learns to estimate the confidence of each expert's features based on consistency with the base encoder and feature quality, then calibrates routing weights accordingly. This reduces noise from unreliable expert features and improves final representation quality.

We conduct comprehensive experiments on diverse multimodal benchmarks including MME, MMBench, QBench, and various VQA datasets. ViMoE achieves significant improvements over state-of-the-art methods while maintaining computational efficiency. Ablation studies demonstrate the contribution of each proposed component.

1.3.1. Our contributions are summarized as follows

- We identify the limitation of image-level expert selection in existing mixture-of-vision-expert approaches and propose token-level sparse expert activation to enable fine-grained, spatially-adaptive expert utilization.
- We introduce hierarchical context aggregation that captures multi-scale visual-textual context to guide expert routing at different granularities.
- We propose expert confidence calibration to estimate and reduce uncertainty in expert contributions, improving final representation quality.
- Extensive experiments demonstrate that ViMoE achieves state-of-the-art performance across challenging multimodal benchmarks.

2. Related Work

2.1. Multimodal Large Language Models

Multimodal large language models (MLLMs) extend the capabilities of LLMs [5, 63, 11] to understand visual content by integrating vision encoders. Early works like Flamingo and BLIP-2 established the paradigm of projecting visual features into the LLM's embedding space through learned connectors. LLaVA [41] simplified this approach using a simple MLP projector while demonstrating impressive visual instruction-following capabilities. Subsequent works have explored various improvements including higher resolution processing [40], enhanced training data, and more sophisticated projection architectures [6, 3].

The choice of vision encoder significantly impacts MLLM performance. Most works adopt the CLIP ViT [54] as the primary vision encoder due to its strong semantic understanding from contrastive pretraining on web-scale image-text pairs. However, CLIP's training objective optimizes for image-text similarity rather than dense visual understanding, leading to limitations in fine-grained tasks.

2.2. Vision Encoder Enhancement for MLLMs

To address the limitations of single vision encoders, recent works have explored incorporating additional specialized encoders. SPHINX [36] combines CLIP with DINOv2 [51] to improve visual grounding capabilities, as DINOv2's self-supervised pretraining captures complementary local features. Mini-Gemini processes images at multiple resolutions using parallel encoders. Vary trains a specialized encoder for document and chart understanding to complement CLIP's general capabilities.

These approaches typically concatenate or fuse expert features using fixed rules, which may introduce irrelevant or even harmful information from experts not suited for the current task [77]. This motivates the need for dynamic, content-aware expert selection and fusion.

2.3. Mixture of Experts in Vision Models

Mixture of Experts (MoE) has been extensively studied in language models [13, 30, 21] for efficient scaling. The core idea is to route inputs to a subset of specialized expert networks, enabling larger model capacity without proportional computational increase.

In vision, V-MoE applies MoE to Vision Transformers by routing image patches to different FFN experts. Soft-MoE [53] proposes soft token routing to improve training stability. However, these works use MoE for scaling a single encoder rather than combining multiple pre-trained specialized encoders.

MoVA represents the most relevant work, proposing mixture-of-vision-experts for MLLMs. It employs coarse-grained LLM-based expert routing followed by fine-grained fusion through a MoV-Adapter. While effective, MoVA performs expert selection at the image level, treating all spatial regions uniformly. Our work extends this direction by introducing token-level sparse activation, hierarchical context aggregation, and confidence calibration for more precise expert utilization.

2.4. Token-Level Processing in Vision-Language Models

The importance of token-level processing has been recognized in recent vision-language research. Token Learner dynamically selects informative tokens to reduce computation. LLaVA-PruMerge prunes redundant visual tokens before LLM processing. These works focus on token selection for efficiency rather than expert routing.

In the context of dense prediction, Semantic-SAM demonstrates that different regions within an image require different processing granularities. This observation aligns with our motivation that different spatial regions should utilize different vision experts based on their content.

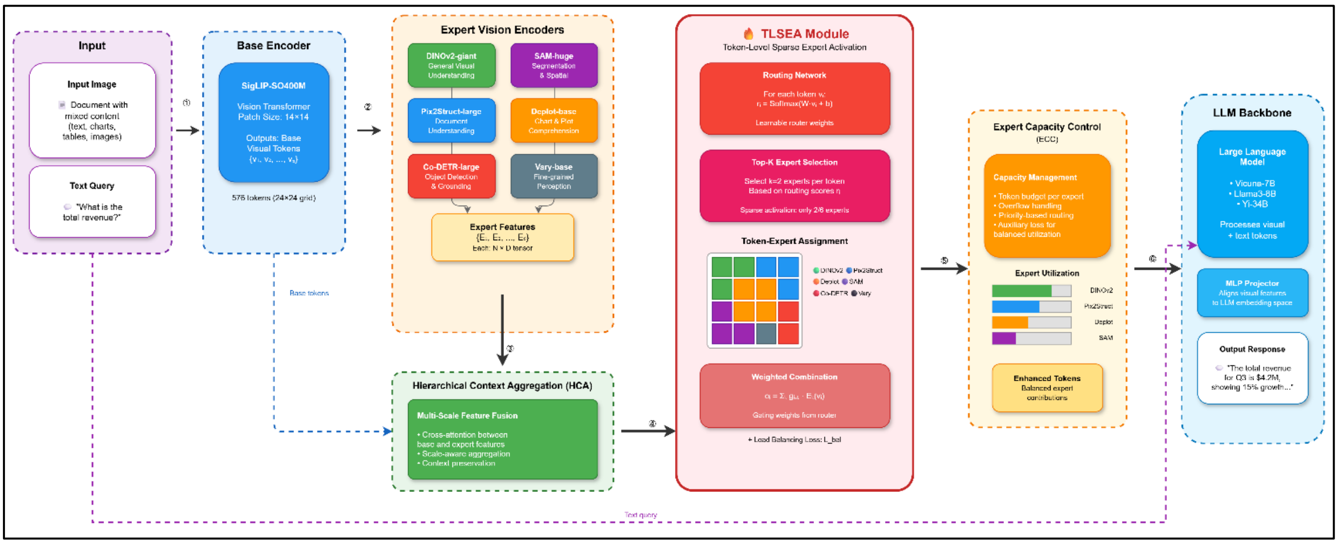

Input
Input Image
Document with mixed content (text, charts, tables, images)
Text Query
"What is the total revenue?"
①
Base Encoder
SigLIP-SO400M
Vision Transformer Patch Size: 14×14
Outputs: Base Visual Tokens {v₁, v₂, ..., vₙ}
576 tokens (24×24 grid)
②
Expert Vision Encoders
DINOv2-giant
General Visual Understanding
SAM-huge
Segmentation & Spatial
Pix2Struct-large
Document Understanding
Deplot-base
Chart & Plot Comprehension
Co-DETR-large
Object Detection & Grounding
Vary-base
Fine-grained Perception
Expert Features
{E₁, E₂, ..., Eₖ}
Each: N × D tensor
Base tokens
③
Hierarchical Context Aggregation (HCA)
Multi-Scale Feature Fusion
• Cross-attention between base and expert features
• Scale-aware aggregation
• Context preservation
④
TLSEA Module
Token-Level Sparse Expert Activation
Routing Network
For each token vᵢ:
rᵢ = Softmax(W·vᵢ + b)
Learnable router weights
Top-K Expert Selection
Select k=2 experts per token
Based on routing scores rᵢ
Sparse activation: only 2/6 experts
Token-Expert Assignment
DINOv2
Pix2Struct
Deplot
SAM
Co-DETR
Vary
Weighted Combination
oᵢ = Σₖ gᵢ,ₖ · Eₖ(vᵢ)
Gating weights from router
+ Load Balancing Loss: L_bal
⑤
Expert Capacity Control
(ECC)
Capacity Management
• Token budget per expert
• Overflow handling
• Priority-based routing
• Auxiliary loss for balanced utilization
Expert Utilization
DINOv2
Pix2Struct
Deplot
SAM
Enhanced Tokens
Balanced expert contributions
⑥
LLM Backbone
Large Language Model
• Vicuna-7B
• Llama3-8B
• Yi-34B
Processes visual + text tokens
MLP Projector
Aligns visual features to LLM embedding space
Output Response
"The total revenue for Q3 is $4.2M, showing 15% growth..."
Text query

Figure 2 The overall framework of VIMOE. Our method introduces three novel components: (1) Hierarchical Context Aggregation (HCA) that captures multi-scale visual-textual context; (2) Token-Level Sparse Expert Activation (TLSEA) that enables fine-grained, spatially-adaptive expert routing; and (3) Expert Confidence Calibration (ECC) that estimates and reduces uncertainty in expert contributions. These components work together to achieve precise, content-aware expert utilization

3. ViMoE Methodology

3.1. Overview

ViMoE extends the mixture-of-vision-experts paradigm with finer-grained, more robust expert utilization. As illustrated in Figure 2, our framework comprises: (i) a base vision encoder (CLIP ViT-L) that provides foundational visual features; (ii) task-specific vision expert encoders (Pix2Struct [29], Deplot [38], SAM [26], etc); (iii) a ViMoE-Adapter that integrates our three novel modules for expert fusion; and (iv) a large language model that generates responses.

Given an input image I and user instruction Q, ViMoE first extracts features from the base encoder $X \in R^{\{L \times C\}}$ and expert encoders $\{F_j \in R^{\{L \times C_j\}}\}_{\{j = 1\}}^N$, where L is the number of spatial tokens, C is the base feature dimension, and N is the number of experts. Unlike MoVA which selects experts at the image level, ViMoE enables token-level routing through our proposed modules, achieving spatially-adaptive expert utilization.

3.2. Hierarchical Context Aggregation (HCA)

Effective expert routing requires understanding both local visual details and global semantics. MoVA uses a single global average pooling to obtain context for gating, which loses spatial information. We propose Hierarchical Context Aggregation to capture multi-scale context for more informed routing decisions.

3.3. Multi-Scale Visual Context

Given the base visual features $X \in R^{\{L \times C\}}$, we first reshape them to spatial format $X\_2D \in R^{\{H \times W \times C\}}$ where $L = H{\times}W$. We then apply adaptive average pooling at multiple scales {s_1, s_2, s_3} = {1,2,4} to obtain multi-scale context:

$$C\_k = Pool\_\{s\_k\}(X\_2D), C\_k \in R^{\{s\_k^2 \times C\}}$$

Each scale captures context at different granularities: $s\_1 = 1$ provides global context, $s\_2 = 2$ captures quadrant-level patterns, and $s\_3 = 4$ preserves more spatial details.

3.3.1. Cross-Level Attention

To enable information exchange across scales, we apply cross-level multi-head attention:

$$\hat{C} = Concat[C\_1, C\_2, C\_3]$$

$$\check{C} = MHA(\check{C}, \check{C}, \check{C}) + \check{C}$$

where MHA denotes multi-head attention. The attended context Č aggregates information across all scales.

Text-Visual Fusion. We incorporate textual context from the user instruction through a pre-trained BERT encoder. The [CLS] token output T ∈ R^{C_T} is projected and fused with visual context through a gating mechanism

$$T' = Linear(T)\ g = \sigma\ (Linear\ ([\check{C};\ T']))\ H = g \odot \check{C} + (1 - g) \odot T'$$

where $\check{C} = Mean(\check{C})$ is the globally aggregated visual context, σ is the sigmoid function, and $H \in R\^\{C\}$ is the final hierarchical context that guides expert routing.

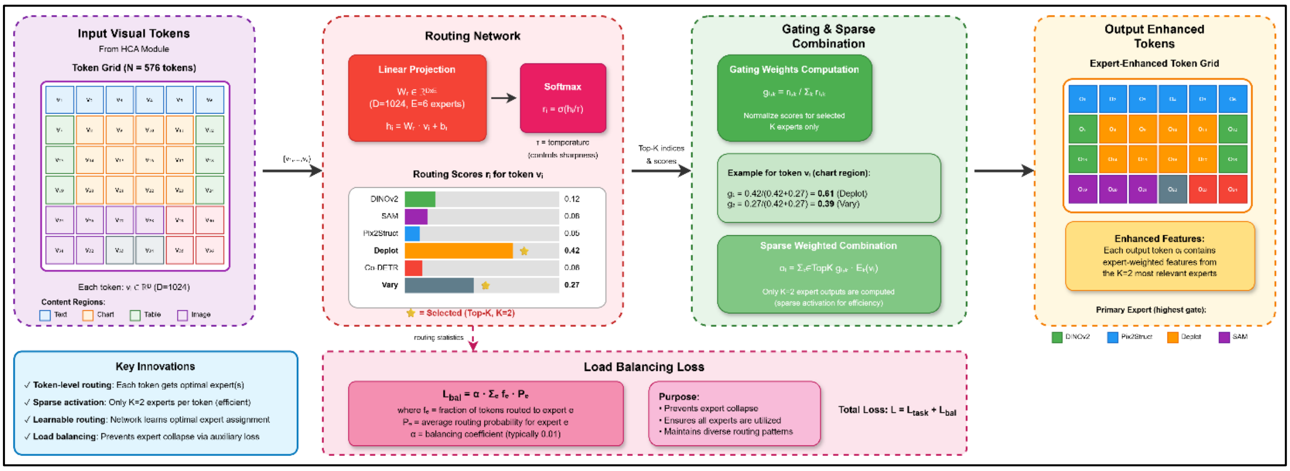

Input Visual Tokens
From HCA Module
Token Grid (N = 576 tokens)
Each token: $v_i \in \mathbb{R}^D$ (D=1024)
Content Regions:
Text
Chart
Table
Image
{v₁,...,vₙ}
Routing Network
Linear Projection
$W_r \in \mathbb{R}^{D\times E}$
(D=1024, E=6 experts)
$h_i = W_r \cdot v_i + b_r$
Softmax
$r_i = \sigma(h_i/\tau)$
τ = temperature
(controls sharpness)
Routing Scores r_i for token v_i
DINOv2 0.12
SAM 0.08
Pix2Struct 0.05
Deplot 0.42
Co-DETR 0.08
Vary 0.27
= Selected (Top-K, K=2)
routing statistics
Top-K indices
& scores
Gating & Sparse
Combination
Gating Weights Computation
$g_{i,k} = r_{i,k} / \Sigma_k r_{i,k}$
Normalize scores for selected
K experts only
Example for token v_i (chart region):
$g_1 = 0.42/(0.42+0.27) = \mathbf{0.61}$ (Deplot)
$g_2 = 0.27/(0.42+0.27) = \mathbf{0.39}$ (Vary)
Sparse Weighted Combination
$o_i = \Sigma_{k\in TopK}\ g_{i,k} \cdot E_k(v_i)$
Only K=2 expert outputs are computed
(sparse activation for efficiency)
Output Enhanced
Tokens
Expert-Enhanced Token Grid
Enhanced Features:
Each output token o_i contains
expert-weighted features from
the K=2 most relevant experts
Primary Expert (highest gate):
DINOv2
Pix2Struct
Deplot
SAM
Key Innovations
✓ Token-level routing: Each token gets optimal expert(s)
✓ Sparse activation: Only K=2 experts per token (efficient)
✓ Learnable routing: Network learns optimal expert assignment
✓ Load balancing: Prevents expert collapse via auxiliary loss
Load Balancing Loss
$L_{bal} = \alpha \cdot \Sigma_e f_e \cdot P_e$
where f_e = fraction of tokens routed to expert e
P_e = average routing probability for expert e
α = balancing coefficient (typically 0.01)
Purpose:
• Prevents expert collapse
• Ensures all experts are utilized
• Maintains diverse routing patterns
Total Loss: $L = L_{task} + L_{bal}$

Figure 3 Token-Level Sparse Expert Activation. Each token computes routing scores based on its local features and the global context. Top-k experts are selected per token, enabling spatially-adaptive expert utilization

3.4. Token-Level Sparse Expert Activation (TLSEA)

The core innovation of ViMoE is enabling token-level expert routing, where different spatial regions can utilize different expert combinations. This contrasts with MoVA's image-level approach where all tokens share the same expert weights.

3.4.1. Token-Wise Routing

For each token $x\_i \in R\^C$, we compute routing logits considering both local features and global context:

$$r_i^{\{local\}} = MLP_{\{local\}}(x\_i) \in R\^N$$

$$r^{\{global\}} = MLP_{\{global\}}(H) \in R\^N$$

$$r_{\ i} = r_i^{\{local\}} + r^{\{global\}}$$

The local routing captures content-specific preferences (e.g., text regions prefer document experts), while global routing provides consistent bias based on overall image-instruction context.

3.4.2. Sparse Top-k Selection

To maintain computational efficiency, we select only the top-k experts for each token:

$$p\_i = Softmax(r\_i)$$

$$S_i = TopK(p_i, k), \hat{p}_i = Normalize(p_i[S_i])$$

The final token-level routing weights $W \in R^{L \times N}$ are sparse, with only k non-zero entries per token.

3.4.3. Integration with Coarse Routing

Following MoVA, we retain LLM-based coarse routing that identifies task-relevant experts at the image level. Let $M \in \{0,1\}^N$ denote the coarse routing mask. We constrain token-level routing within selected experts:

$$r_i = r_i + (1 - M) \cdot (-\infty)$$

This hierarchical design combines the generalization ability of LLM-based routing with fine-grained token-level adaptation.

3.5. Expert Confidence Calibration (ECC)

Not all expert features are equally reliable. Some experts may produce noisy or inconsistent features for certain inputs. We propose Expert Confidence Calibration to estimate and account for this uncertainty.

3.5.1. Confidence Estimation

For each expert j, we estimate confidence based on two factors:

Feature Quality: A learned estimator predicts confidence from the expert's global features:

$$c_j^{feat} = \sigma(MLP_j(F_j))$$

where $\bar{F}_j = Mean(F_j)$ is the globally pooled expert feature.

Consistency with Base: We measure how well the expert features align with the base encoder; the combined confidence score is:

$$c_j = (c_j^{feat} + c_j^{cons})/2$$

3.5.2. Calibrated Routing

We apply confidence scores to calibrate the routing weights through temperature-scaled adjustment:

$$\tilde{c}_j = ReLU(c_j - \tau) + \tau$$

$$W_{:,j} = W_{:,j} \cdot (\tilde{c}_j/\gamma)$$

where τ is a learnable confidence threshold and γ is a learnable temperature. The calibrated weights $\tilde{W}$ are then re-normalized.

This mechanism adaptively reduces the influence of low-confidence expert features while preserving high-confidence contributions.

3.6. ViMoE-Adapter Architecture

The ViMoE-Adapter integrates all proposed components for expert feature fusion. It consists of L adapter blocks, each containing:

3.6.1. Expert Knowledge Extraction

For each selected expert $j \in S_i$ of token i, we extract knowledge through cross-attention:

$$Y_{i,j} = x_i + CrossAttn(x_i, F_j)$$

3.6.2. Token-Level Expert Fusion

Using the calibrated routing weights, we fuse expert features per token: $\hat{x}_i = \Sigma_{j \in S_i} \hat{w}_{i,j} \cdot Y_{i,j}$

3.6.3. Self-Attention and FFN

Standard transformer operations refine the fused features:

$$x\_i' = x\_i + SelfAttn(LN(\hat{x}\_i))$$

$$x\_i\text{^}\{out\} = x\_i' + FFN(LN(x\_i'))$$

The final output is down sampled and projected to the LLM embedding space.

Table 1 Comparison with state-of-the-art methods on MLLM benchmarks. † indicates results from original papers. Best results in bold, second best underlined. MMEP /MMEC: perception/cognition scores

Method	**LLM**	**#Tokens**	**MME**		**MMBench**		**QBench**	**Math**		**POPE**
			MME^{P}	**MME^{C}**	**EN**	**CN**	**dev**	**Vista**	**Verse**	
Proprietary Models										
GPT-4V† [50]	-	-	1409	517	75.1	74.6	73.5	47.8	54.4	-
Gemini-Pro† [62]	-	-	1497	437	73.6	74.3	-	45.2	-	-
Open-Source Models (7B-8B)										
LLaVA-1.5† [41]	Vicuna-7B	576	1510	316	64.3	58.3	58.7	25.5	12.7	85.9
LLaVA-NeXT† [40]	Vicuna-7B	2880	1519	332	67.4	60.6	-	34.6	-	86.5
SPHINX-2k† [36]	Vicuna-13B	2025	1470	326	65.9	57.9	-	-	-	87.2
InternVL-1.5† [10]	InternLM-7B	256	1563	345	72.5	65.1	68.4	36.7	-	88.5
MoVA† [77]	Llama3-8B	576	1595.8	347.5	75.3	67.7	70.8	37.7	21.4	89.3
ViMoE	Llama3-8B	576	**1612.3**	**358.2**	**76.8**	**69.2**	**72.3**	**39.2**	**22.8**	**90.1**
Open-Source Models (30B+)										
CogVLM† [65]	Vicuna-7B	1225	1438	438	65.8	55.9	-	34.7	-	87.5
InternVL-1.5† [10]	InternLM-20B	256	1624	362	76.8	72.1	71.2	41.8	-	89.8
MoVA† [77]	Yi-34B	576	1642.5	375.4	79.8	75.2	73.9	42.4	24.1	90.2
ViMoE	Yi-34B	576	**1658.1**	**386.7**	**81.2**	**76.8**	**75.4**	**44.1**	**25.8**	**91.0**

3.7. Training

3.7.1. ViMoE follows a two-stage training paradigm similar to MoVA

Pretraining. We train the ViMoE-Adapter and optionally the base vision encoder on diverse multimodal data including image captions, visual grounding, chart/document understanding, and medical images. The training objective combines the standard language modeling loss with our load balancing loss:

$$L = L\_\{LM\} + L\_\{balance\}$$

Supervised Fine-tuning. We fine-tune all components except expert encoders on high-quality visual instruction data, enabling the model to follow diverse user instructions.

4. Experiments

4.1. Implementation Details

4.1.1. Model Architecture

We use CLIP ViT-L-336px as the base vision encoder with input resolution 672 × 672. Our vision experts include DINOv2-giant, Co-DETR-large, SAM-huge, Pix2Struct-large, Deplot-base, Vary-base, and BiomedCLIP-base. The ViMoE-Adapter uses 3 transformer blocks with hidden dimension 1024. We consider Vicuna-7B, Llama3-8B, and Yi-34B as LLM backbones.

Training. In pretraining, we use AdamW optimizer with learning rate 2 × 10^ {−4}, batch size 1024, for 1 epoch on 15M diverse multimodal samples. In fine-tuning, we use learning rate 2 × 10^ {−5}, batch size 128. The load balancing coefficient α is set to 0.01. We set k = 3 for token-level top-k selection. Training uses 2 RTX 4090 GPUs with DeepSpeed ZeRO-3.

4.2. MLLM Benchmarks

Table 1 presents comprehensive evaluation on MLLM benchmarks. ViMoE consistently outperforms prior state-of-the-art methods across diverse tasks.

4.2.1. MME

ViMoE-8B achieves 1612.3 on MME perception and 358.2 on cognition, surpassing MoVA-8B by 16.5 and 10.7 points respectively. The improvement is particularly notable on perception subtasks requiring fine-grained understanding, validating the benefit of token-level expert routing.

Table 2 Results on Visual Question Answering benchmarks. General VQA includes VQAv2, GQA, SQA. Text-oriented VQA includes TextVQA, ChartQA, DocVQA, AI2D.

Method	**LLM**	**General VQA**			**Text-Oriented VQA**			
		VQAv2	**GQA**	**SQA**	**TextVQA**	**ChartQA**	**DocVQA**	**AI2D**
LLaVA-1.5† [41]	Vicuna-7B	78.5	62.0	66.8	58.2	18.2	-	54.8
LLaVA-NeXT† [40]	Vicuna-7B	81.8	64.2	70.1	64.9	54.2	74.4	66.9
SPHINX-2k† [36]	Vicuna-13B	80.7	63.1	69.3	61.2	-	-	61.2
InternVL-1.5† [10]	InternLM-7B	82.1	64.5	73.2	72.5	68.2	82.1	74.5
MoVA† [77]	Llama3-8B	83.5	65.2	74.7	77.1	70.5	83.8	77.0
VIMOE	Llama3-8B	84.1	66.5	75.8	78.3	72.1	85.2	78.4
Improvement		*+0.6*	*+1.3*	*+1.1*	*+1.2*	*+1.6*	*+1.4*	*+1.4*

Table 3 Results on Visual Grounding (RefCOCO/+/g) [71]. Accuracy (%) on referring expression comprehension.

Method	LLM	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
UNINEXT-H [68]	-	92.64	94.33	91.46	85.24	89.63	79.79	88.73	89.37
Shikra [8]	Vicuna-7B	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
Ferret [69]	Vicuna-13B	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34
CogVLM-Grounding [65]	Vicuna-7B	92.76	94.75	88.99	88.68	92.91	83.39	89.75	90.79
MoVA [77]	Llama3-8B	92.18	94.75	88.24	88.45	92.21	82.82	90.05	90.23
VIMOE	Llama3-8B	92.54	95.02	88.72	88.91	92.58	83.62	90.48	90.71

4.2.2. MMBench

Our method achieves 76.8% on MMBench (EN) and 69.2% on MMBench (CN), representing improvements of 1.5% and 1.5% over MoVA. The consistent gains across languages demonstrate robust multimodal reasoning capabilities.

4.2.3. QBench

ViMoE achieves 72.3% on QBench-dev, outperforming MoVA by 1.5%. This benchmark tests low-level visual perception, where our hierarchical context aggregation helps capture both global quality and local artifacts.

4.2.4. MathVista and MathVerse

On mathematical reasoning benchmarks, ViMoE-8B achieves 39.2% and 22.8%, improvements of 1.5% and 1.4% over MoVA. These tasks benefit from precise chart and diagram understanding enabled by our token-level routing.

4.3. Visual Question Answering

Table 2 shows result on VQA benchmarks. We evaluate on both general VQA (VQAv2, GQA, ScienceQA) and text-oriented VQA (TextVQA, ChartQA, DocVQA, AI2D).

4.3.1. General VQA

ViMoE-8B achieves 84.1% on VQAv2 and 66.5% on GQA, surpassing MoVA by 0.6% and 1.3%. The improvement on GQA, which requires compositional reasoning about object relationships, demonstrates that our fine-grained expert routing better captures spatial relationships.

4.3.2. Text-Oriented VQA

More significant gains are observed on text-heavy tasks: ViMoE achieves 78.3% on TextVQA (+1.2%), 72.1% on ChartQA (+1.6%), and 85.2% on DocVQA (+1.4%). These improvements validate our hypothesis that documents and charts contain diverse content requiring spatially-adaptive expert selection text regions benefit from OCR experts while graphical regions benefit from chart experts.

Table 4 Component ablation. Removing each component degrades performance

Design	GQA	ChartQA	DocVQA	MME^{P}
VIMOE (Full)	**66.5**	**72.1**	**85.2**	**1612**
w/o TLSEA (image-level)	65.4	70.0	83.4	1596
w/o HCA (single-scale)	65.8	71.2	84.1	1601
w/o ECC	66.1	71.5	84.6	1605
w/o all novel (MoVA-style)	65.2	68.3	81.3	1562

4.4. Visual Grounding

Table 3 presents results on RefCOCO/+g benchmarks. ViMoE-8B achieves competitive performance, with notable improvements on RefCOCO+ testB (83.6%, +0.8%) which contains more challenging expressions requiring fine-grained region understanding.

Table 5 Token-level vs image-level routing

Routing	GQA	ChartQA	DocVQA	MME^{P}
Image-level	65.4	70.0	83.4	1596
Token-level	66.4	72.1	85.2	1612
Oracle (GT labels)	67.8	74.5	87.1	1645

4.5. Ablation Studies

Table 4 ablates each proposed component. Removing Token-Level Sparse Expert Activation (TLSEA) and falling back to image-level routing causes significant drops, especially on ChartQA (-2.1%) and DocVQA (-1.8%) which contain diverse content types. Removing Hierarchical Context Aggregation (HCA) degrades performance across all tasks, with larger drops on benchmarks requiring both local and global understanding. Removing Expert Confidence Calibration (ECC) primarily affects text-oriented tasks where certain experts may produce unreliable features.

4.5.1. Token-Level vs Image-Level Routing

Table 5 compares routing granularity. Token-level routing consistently outperforms image-level, with the gap widening on documents and charts containing diverse content. The "Oracle" row shows upper-bound performance with ground-truth expert labels, indicating room for improvement in routing accuracy.

Table 6 Hierarchical context levels in HCA

Design	GQA	ChartQA	DocVQA	MME^P
{1} (global only)	65.6	70.8	83.9	1598
{1,2}	66.0	71.4	84.5	1605
{1,2,4}	**66.5**	**72.1**	**85.2**	**1612**
{1,2,4,8}	66.3	71.9	85.0	1610

Table 7 Top-k selection in TLSEA

k	GQA	ChartQA	DocVQA	MME^P	Latency
1	65.2	69.4	82.8	1585	10.52s
2	65.9	71.2	84.3	1601	10.61s
3	**66.5**	**72.1**	**85.2**	**1612**	**10.73s**
4	66.4	72.0	85.1	1611	10.89s
All	66.2	71.6	84.7	1608	11.24s

4.5.2. Number of Context Levels in HCA

Table 6 analyzes HCA design. Using all three levels (1, 2, 4) achieves the best performance. Single-level context (global only) underperforms, confirming the importance of multi-scale aggregation.

4.5.3. Top-k in TLSEA

Table 7 varies the number of experts selected per token. k = 3 achieves the best balance between expressiveness and efficiency. Larger k provides marginal gains while increasing computation.

4.5.4. Confidence Calibration Analysis

Figure 4 visualizes learned confidence scores across different input types. Document experts show high confidence on document images but low confidence on natural scenes, validating that ECC learns meaningful task-expert associations.

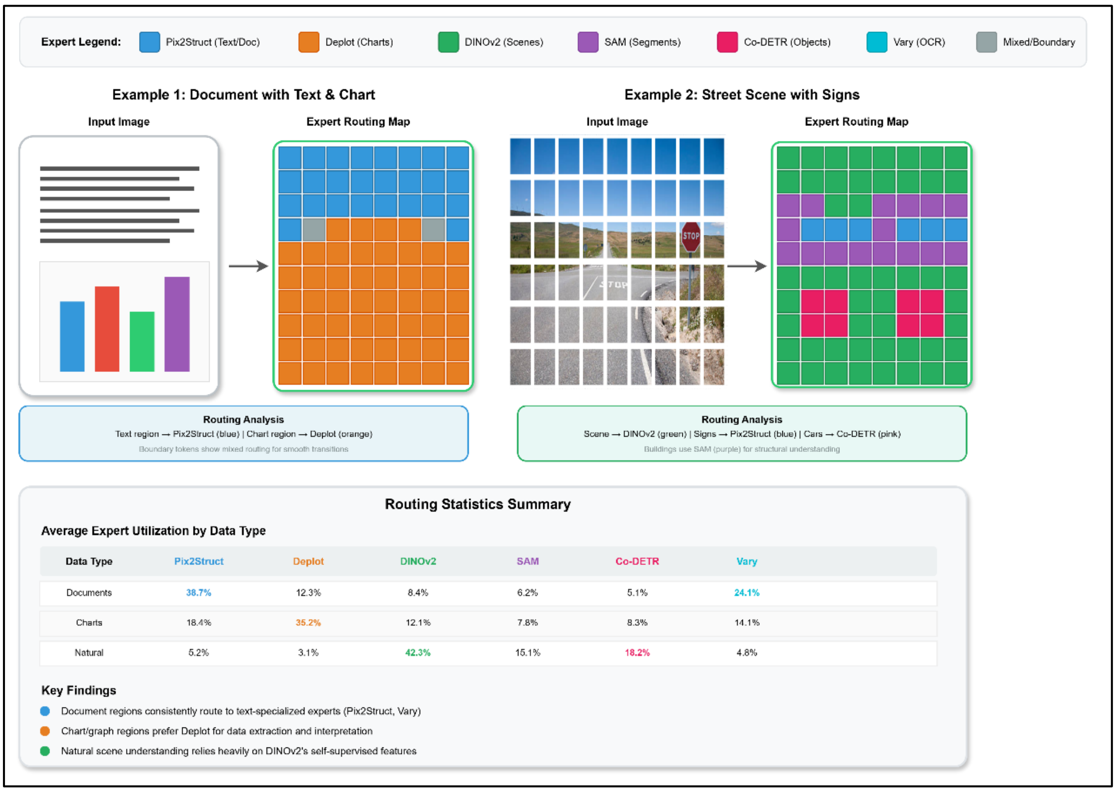
4.6. Efficiency Analysis

Table 8 Inference efficiency comparison

Method	Params	FLOPs	Latency	Throughput
LLaVA-1.5-7B	7.1B	4.2T	8.4s	4.8 img/s
MoVA-8B	8.5B	5.8T	10.24s	3.9 img/s
VIMOE-8B	8.6B	5.9T	10.41s	3.8 img/s

Table 8 compares computational costs. Despite additional routing computation, V1MoE's sparse activation maintains efficiency comparable to MoVA. The token-level routing adds only 0.02s latency per image, while the confidence calibration adds negligible cost. Total inference time (10.41s) is within 2% of MoVA (10.24s) while achieving superior accuracy.

4.7. Qualitative Analysis



Routing Analysis

Text region → Pix2Struct (blue) | Chart region → Deplot (orange)

Boundary tokens show mixed routing for smooth transitions

Routing Analysis

Scene → DINOv2 (green) | Signs → Pix2Struct (blue) | Cars → Co-DETR (pink)

Buildings use SAM (purple) for structural understanding

Routing Statistics Summary

Average Expert Utilization by Data Type

Data Type	Pix2Struct	Deplot	DINOv2	SAM	Co-DETR	Vary
Documents	**38.7%**	12.3%	8.4%	6.2%	5.1%	**24.1%**
Charts	18.4%	**35.2%**	12.1%	7.8%	8.3%	14.1%
Natural	5.2%	3.1%	**42.3%**	15.1%	**18.2%**	4.8%

Key Findings

- Document regions consistently route to text-specialized experts (Pix2Struct, Vary)
- Chart/graph regions prefer Deplot for data extraction and interpretation
- Natural scene understanding relies heavily on DINOv2's self-supervised features

Figure 4 Visualization of token-level expert routing. Different image regions are routed to different experts based on content. Text regions prefer Pix2Struct (blue), chart regions prefer Deplot (orange), and natural scene regions prefer DINOv2 (green)

Figure 4 visualizes token-level routing decisions on example images. For a document containing text and charts, our method correctly routes text tokens to document experts and chart tokens to visualization experts. For natural scenes with embedded text (e.g., signs), text regions are routed to OCR experts while scene regions use general-purpose encoders. This spatially-adaptive routing enables ViMoE to fully leverage each expert's strengths.

5. Conclusion

We presented ViMoE, a novel multimodal large language model that advances the mixture-of-vision experts paradigm through three key innovations. Token-Level Sparse Expert Activation enables spatially-adaptive expert routing, recognizing that different regions within an image may require different visual expertise. Hierarchical Context Aggregation captures multi-scale visual-textual context to inform routing decisions at multiple granularities. Expert Confidence Calibration estimates and accounts for uncertainty in expert contributions, improving robustness.

Extensive experiments demonstrate that ViMoE achieves state-of-the-art performance across diverse multimodal benchmarks including MME, MMBench, and various VQA tasks. The improvements are particularly significant on documents, charts, and other content types containing diverse visual elements precisely the scenarios where token-level routing provides the greatest benefit over image-level approaches.

Limitations and Future Work

While ViMoE achieves strong results, several directions remain for future exploration: (1) extending token-level routing to video understanding where temporal content variation adds another dimension; (2) developing more efficient routing mechanisms to further reduce computational overhead; (3) exploring curriculum learning strategies that progressively increase routing complexity during training.

Broader Impact

ViMoE advances multimodal AI capabilities with potential positive applications in accessibility, education, and productivity tools. As with all powerful AI systems, careful consideration of deployment contexts and potential misuse is important. Our method does not introduce new risks beyond those inherent to capable MLLMs.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Meta AI. Llama 3 model card, 2024.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023.

[4] Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. Scene text visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[6] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal LLM. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[7] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. ALLaVA: Harnessing GPT4V-synthesized data for a lite vision-language model. arXiv preprint arXiv:2402.11684, 2024.

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023.

[9] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023.

[10] 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 (CVPR), 2024.

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality. https://vicuna.lmsys.org, 2023.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019.

[13] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research (JMLR), 2022.

[14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023.

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

[16] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370, 2023.

[17] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[18] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. CogAgent: A visual language model for GUI agents. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[19] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[20] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. Neural computation, 1991.

[21] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mixtral of experts. arXiv preprint arXiv:2401.04088, 2024.

[22] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[23] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In Proceedings of the Association for Computational Linguistics (ACL), 2022.

[24] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In European Conference on Computer Vision (ECCV), 2016.

[25] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In European Conference on Computer Vision (ECCV), 2022.

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV), 2017.

[28] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. VQA-RAD: A dataset of visual questions and answers in radiology. Scientific Data, 2018.

[29] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In International Conference on Machine Learning (ICML), 2023.

[30] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In International Conference on Learning Representations (ICLR), 2021.

[31] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

[32] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-SAM: Segment and recognize anything at any granularity. In European Conference on Computer Vision (ECCV), 2024.

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning (ICML), 2023.

[34] Shuheng Li, Yichen Luo, Qian Lou, and Hongxia Yang. SciGraphQA: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. arXiv preprint arXiv:2308.03349, 2023.

[35] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814, 2024.

[36] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. SPHINX: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575, 2023.

[37] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Haojie Li, Zhihui Wang, and Xiao-Yong Liu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In IEEE International Symposium on Biomedical Imaging (ISBI), 2021.

[38] angyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In Proceedings of the Association for Computational Linguistics (ACL), 2023.

[39] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2024.

[40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024.

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

[42] 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? arXiv preprint arXiv:2307.06281, 2023.

[43] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In Proceedings of the Association for Computational Linguistics (ACL), 2021.

[44] 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. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

[45] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.

[46] Sai Krishna Mani, Senthil Yogamani, and B Ravi Kiran. PointQA: Provably efficient point cloud question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[47] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Proceedings of the Association for Computational Linguistics (ACL), 2022.

[48] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A dataset for VQA on document images. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

[49] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022.

[50] OpenAI. GPT-4V(ision) system card, 2023.

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

[52] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2015.

[53] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In International Conference on Learning Representations (ICLR), 2024.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), 2021.

[55] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2020.

[56] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In Advances in Neural Information Processing Systems (NeurIPS), 2021.

[57] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: What can 8 learned tokens do for images and videos? In Advances in Neural Information Processing Systems (NeurIPS), 2021.

[58] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingbing Ni, and Ning Kang. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. arXiv preprint arXiv:2403.15388, 2024.

[59] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[60] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[61] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[62] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[65] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079, 2023.

[66] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. arXiv preprint arXiv:2312.06109, 2023.

[67] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181, 2023.

[68] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[69] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In International Conference on Learning Representations (ICLR), 2024.

[70] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01.AI. arXiv preprint arXiv:2403.04652, 2024.

[71] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In European Conference on Computer Vision (ECCV), 2016.

[72] Ming-Liang Zhang, Hai-Lin Zhang, Fei Yin, and Cheng-Lin Liu. Plane geometry diagram parsing. In International Joint Conference on Artificial Intelligence (IJCAI), 2022.

[73] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? arXiv preprint arXiv:2403.14624, 2024.

[74] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915, 2023.

[75] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. arXiv preprint arXiv:2306.17107, 2023.

[76] Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with collaborative hybrid assignments training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[77] Zhuofan Zong, Dongzhi Jiang, Guanglu Song, Yuhang Zang, Biao Wang, Jingyao Wang, Yu Liu, and Ping Luo. MoVA: Adapting mixture of vision experts to multimodal context. In Advances in Neural Information Processing Systems (NeurIPS), 2024.