VIMOE: Vision Mixture of Experts with Multimodal Context Awareness

Adele Chinda
Georgia State University

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 (**Vi**sion **M**ixture **o**f **E**xperts 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://arrdel.github.io/vimoe/`

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 [54] 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 [36] 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**.

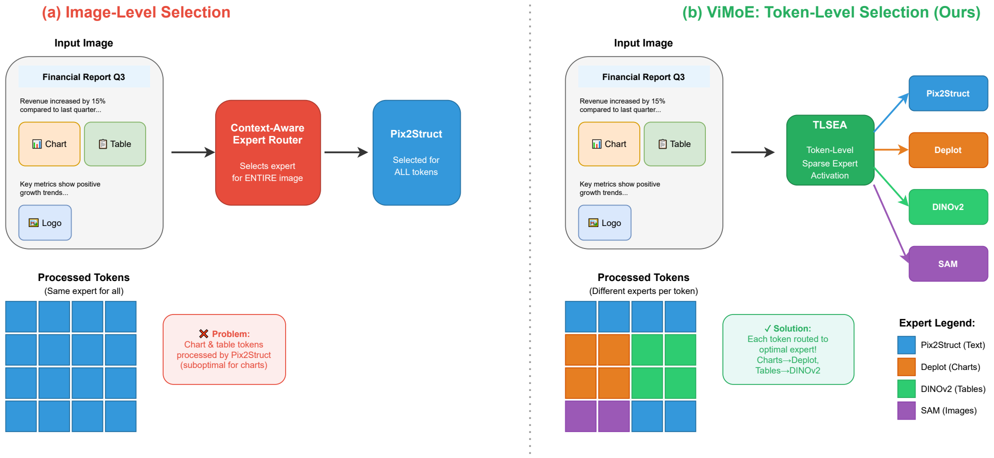

(a) Image-Level Selection
Input Image
Financial Report Q3
Revenue increased by 15% compared to last quarter...
Chart
Table
Key metrics show positive growth trends...
Logo
Context-Aware Expert Router
Selects expert for ENTIRE image
Pix2Struct
Selected for ALL tokens
Processed Tokens
(Same expert for all)
✗ 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 [29] while routing chart regions to visualization experts like Deplot [38].

In this paper, we propose VIMOE (**Vi**sion **M**ixture **o**f **E**xperts with Multimodal Context Awareness), which introduces three novel components to address these limitations:

(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.

(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.

(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 [14], MMBench [42], QBench [67], and various VQA datasets [17, 19, 61, 47, 48]. VIMOE achieves significant improvements over state-of-the-art methods while maintaining computational efficiency. Ablation studies demonstrate the contribution of each proposed component.

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 [2] and BLIP-2 [33] 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 [9, 7], 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 [60].

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 [35] processes images at multiple resolutions using parallel encoders. Vary [66] 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) [20] 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 [56] 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 [77] 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. TokenLearner [57] dynamically selects informative tokens to reduce computation. LLaVA-PruMerge [58] prunes redundant visual tokens before LLM processing. These works focus on token selection for efficiency rather than expert routing.

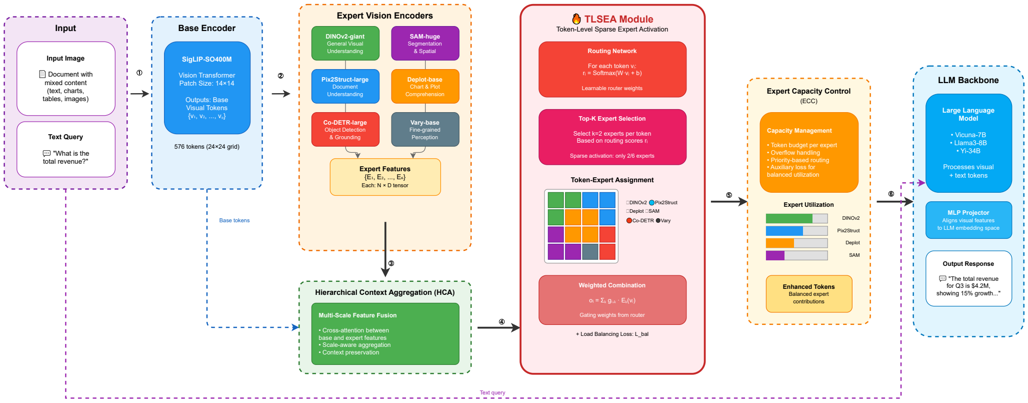

Input
Input Image
Document with mixed content (text, charts, tables, images)
Text Query
"What is the total revenue?"
①
Base Encoder
SigLIP-SO400M
Vision Transformer Patch Size: 14×14
Outputs: Base Visual Tokens {v₁, v₂, ..., vₙ}
576 tokens (24×24 grid)
②
Expert Vision Encoders
DINOv2-giant General Visual 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, ..., E₆} Each: N × D tensor
③
Base tokens
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.

In the context of dense prediction, Semantic-SAM [32] 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.

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 (DINOv2 [51], 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 $\mathcal{I}$ and user instruction $\mathcal{Q}$, VIMOE first extracts features from the base encoder $\mathbf{X} \in \mathbb{R}^{L \times C}$ and expert encoders $\{\mathcal{F}_j \in \mathbb{R}^{L \times C_j}\}_{j=1}^{N}$, 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.

Multi-Scale Visual Context. Given the base visual features $\mathbf{X} \in \mathbb{R}^{L \times C}$, we first reshape them to spatial format $\mathbf{X}_{2D} \in \mathbb{R}^{H \times W \times C}$ 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:

$$\mathbf{C}_k = \text{Pool}_{s_k}(\mathbf{X}_{2D}), \quad \mathbf{C}_k \in \mathbb{R}^{s_k^2 \times C} \tag{1}$$

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.

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

$$\hat{\mathbf{C}} = \text{Concat}[\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3] \tag{2}$$

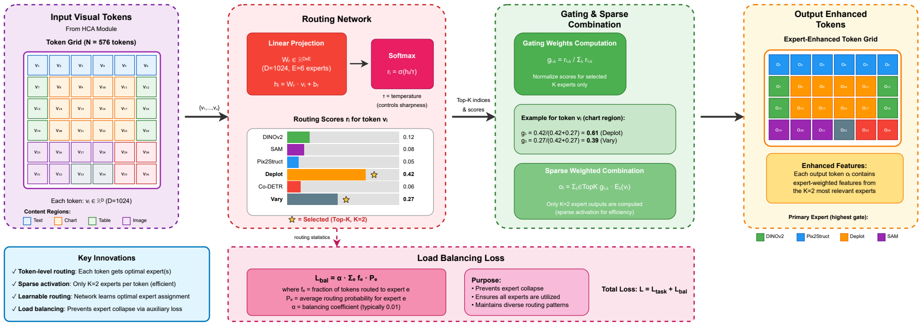

Input Visual Tokens
From HCA Module
Token Grid (N = 576 tokens)
Each token: vᵢ ∈ ℝᴰ (D=1024)
Content Regions:
Text
Chart
Table
Image
{v₁,...,vₙ}
Routing Network
Linear Projection
Wᵣ ∈ ℝ^{D×E}
(D=1024, E=6 experts)
hᵢ = Wᵣ · vᵢ + bᵣ
Softmax
rᵢ = σ(hᵢ/τ)
τ = temperature
(controls sharpness)
Routing Scores rᵢ for token vᵢ
DINOv2 0.12
SAM 0.08
Pix2Struct 0.05
Deplot 0.42
Co-DETR 0.06
Vary 0.27
☆ = Selected (Top-K, K=2)
routing statistics
Top-K indices
& scores
Gating & Sparse Combination
Gating Weights Computation
gᵢ,ₖ = rᵢ,ₖ / Σₖ rᵢ,ₖ
Normalize scores for selected K experts only
Example for token vᵢ (chart region):
g₁ = 0.42/(0.42+0.27) = 0.61 (Deplot)
g₂ = 0.27/(0.42+0.27) = 0.39 (Vary)
Sparse Weighted Combination
oᵢ = Σₖ∈TopK gᵢ,ₖ · Eₖ(vᵢ)
Only K=2 expert outputs are computed
(sparse activation for efficiency)
Output Enhanced Tokens
Expert-Enhanced Token Grid
Enhanced Features:
Each output token oᵢ 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 = α · Σₑ fₑ · Pₑ
where fₑ = fraction of tokens routed to expert e
Pₑ = 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.

$$\tilde{\mathbf{C}} = \text{MHA}(\hat{\mathbf{C}}, \hat{\mathbf{C}}, \hat{\mathbf{C}}) + \hat{\mathbf{C}} \tag{3}$$

where MHA denotes multi-head attention [64]. The attended context $\tilde{\mathbf{C}}$ aggregates information across all scales.

Text-Visual Fusion. We incorporate textual context from the user instruction through a pre-trained BERT encoder [12]. The [CLS] token output $\mathbf{T} \in \mathbb{R}^{C_T}$ is projected and fused with visual context through a gating mechanism:

$$\mathbf{T}' = \text{Linear}(\mathbf{T}) \tag{4}$$

$$\mathbf{g} = \sigma(\text{Linear}([\bar{\mathbf{C}}; \mathbf{T}'])) \tag{5}$$

$$\mathbf{H} = \mathbf{g} \odot \bar{\mathbf{C}} + (1 - \mathbf{g}) \odot \mathbf{T}' \tag{6}$$

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

3.3 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.

Token-Wise Routing. For each token $\mathbf{x}_i \in \mathbb{R}^C$, we compute routing logits considering both local features and global context:

$$\mathbf{r}_i^{local} = \text{MLP}_{local}(\mathbf{x}_i) \in \mathbb{R}^N \tag{7}$$

$$\mathbf{r}^{global} = \text{MLP}_{global}(\mathbf{H}) \in \mathbb{R}^N \tag{8}$$

$$\mathbf{r}_i = \mathbf{r}_i^{local} + \mathbf{r}^{global} \tag{9}$$

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.

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

$$\mathbf{p}_i = \text{Softmax}(\mathbf{r}_i) \tag{10}$$

$$\mathcal{S}_i = \text{TopK}(\mathbf{p}_i, k), \quad \hat{\mathbf{p}}_i = \text{Normalize}(\mathbf{p}_i[\mathcal{S}_i]) \tag{11}$$

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

Integration with Coarse Routing. Following MoVA [77], we retain LLM-based coarse routing that identifies task-relevant experts at the image level. Let $\mathbf{M} \in \{0, 1\}^N$ denote the coarse routing mask. We constrain token-level routing within selected experts:

$$\mathbf{r}_i = \mathbf{r}_i + (1 - \mathbf{M}) \cdot (-\infty) \tag{12}$$

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

Load Balancing Loss. To encourage balanced expert utilization and prevent routing collapse [13], we introduce an auxiliary load balancing loss:

$$\mathcal{L}_{balance} = \alpha \cdot \sum_{j=1}^{N} \left(\frac{1}{L} \sum_{i=1}^{L} p_{i,j} - \frac{1}{|\mathcal{A}|} \right)^2 \tag{13}$$

where $\mathcal{A}$ is the set of active experts (from coarse routing) and α is a balancing coefficient.

3.4 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.

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:

$$c_j^{feat} = \sigma(\text{MLP}_j(\bar{\mathcal{F}}_j)) \tag{14}$$

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

Consistency with Base: We measure how well the expert features align with the base encoder:

$$c_j^{cons} = \sigma(\text{MLP}_{cons}([\bar{\mathbf{X}}; \bar{\mathcal{F}}_j])) \tag{15}$$

The combined confidence score is:

$$c_j = \frac{c_j^{feat} + c_j^{cons}}{2} \tag{16}$$

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

$$\tilde{c}_j = \text{ReLU}(c_j - \tau) + \tau \tag{17}$$

$$\hat{\mathbf{W}}_{:,j} = \mathbf{W}_{:,j} \cdot \frac{\tilde{c}_j}{\gamma} \tag{18}$$

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

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

3.5 ViMoE-Adapter Architecture

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

Expert Knowledge Extraction. For each selected expert $j \in \mathcal{S}_i$ of token i, we extract knowledge through cross-attention:

$$\mathbf{Y}_{i,j} = \mathbf{x}_i + \text{CrossAttn}(\mathbf{x}_i, \mathcal{F}_j) \tag{19}$$

Token-Level Expert Fusion. Using the calibrated routing weights, we fuse expert features per token:

$$\hat{\mathbf{x}}_i = \sum_{j \in \mathcal{S}_i} \hat{w}_{i,j} \cdot \mathbf{Y}_{i,j} \tag{20}$$

Self-Attention and FFN. Standard transformer operations refine the fused features:

$$\mathbf{x}_i' = \mathbf{x}_i + \text{SelfAttn}(\text{LN}(\hat{\mathbf{x}}_i)) \tag{21}$$

$$\mathbf{x}_i^{out} = \mathbf{x}_i' + \text{FFN}(\text{LN}(\mathbf{x}_i')) \tag{22}$$

The final output is downsampled 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 <u>underlined</u>. MME^P/MME^C: 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**

3.6 Training

ViMoE follows a two-stage training paradigm similar to MoVA [77]:

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:

$$\mathcal{L} = \mathcal{L}_{LM} + \mathcal{L}_{balance} \tag{23}$$

Supervised Fine-tuning. We fine-tune all components except expert encoders on high-quality visual instruction data [41, 9], enabling the model to follow diverse user instructions.

4 Experiments

4.1 Implementation Details

Model Architecture. We use CLIP ViT-L-336px [54] as the base vision encoder with input resolution 672×672. Our vision experts include DINOv2-giant [51], Co-DETR-large [76], SAM-huge [26], Pix2Struct-large [29], Deplot-base [38], Vary-base [66], and BiomedCLIP-base [74]. The ViMoE-Adapter uses 3 transformer blocks with hidden dimension 1024. We consider Vicuna-7B [11], Llama3-8B [1], and Yi-34B [70] 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 [55].

4.2 MLLM Benchmarks

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

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
CogAgent [18]	Vicuna-7B	-	-	-	76.1	68.4	81.6	-
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**

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.

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.

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.

MathVista & MathVerse. On mathematical reasoning benchmarks [45, 73], 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 results on VQA benchmarks. We evaluate on both general VQA (VQAv2 [17], GQA [19], ScienceQA [44]) and text-oriented VQA (TextVQA [61], ChartQA [47], DocVQA [48], AI2D [24]).

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.

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

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	**66.5**	**72.1**	**85.2**	**1612**
Oracle (GT labels)	67.8	74.5	87.1	1645

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.

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.

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.

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.

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.

Confidence Calibration Analysis. Figure **??** 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 6: **Hierarchical context levels in HCA.**

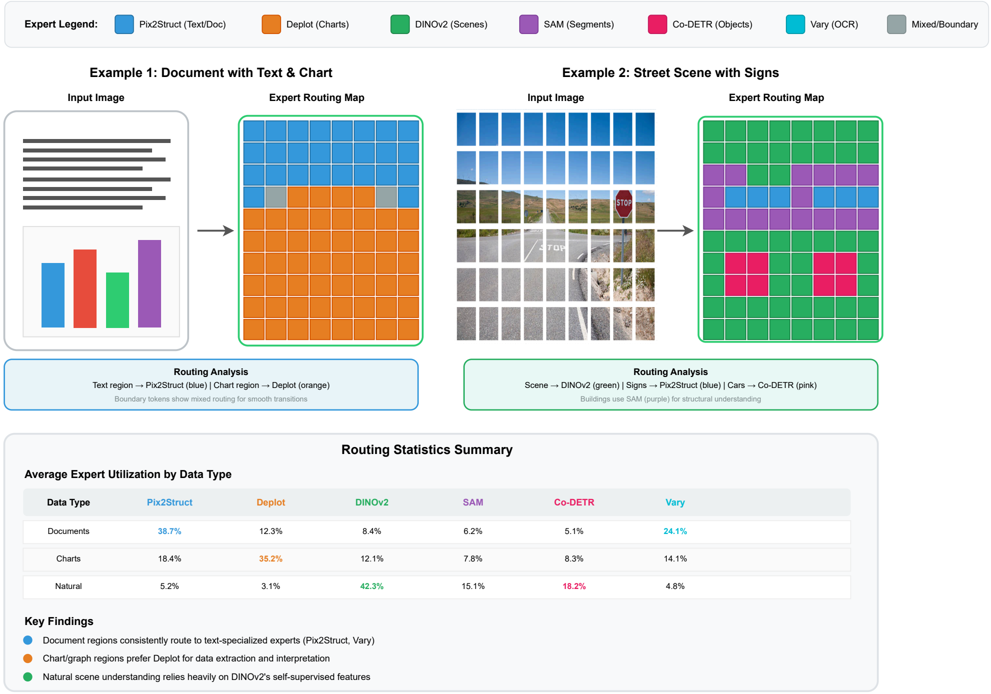
Context Levels	GQA	ChartQA	DocVQA	MME^P
{1} (global only)	65.6	70.8	83.9	1598
{1, 2}	66.0	71.4	84.5	1605
{1, 2, 4}	**66.5**	**72.1**	**85.2**	**1612**
{1, 2, 4, 8}	66.3	71.9	85.0	1610

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

Table 8 compares computational costs. Despite additional routing computation, ViMoE'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 [14], MMBench [42], 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.

References

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

[2] Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

[3] Jinze Bai et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Ali Furkan Biten et al. Scene text visual question answering. *ICCV*, 2019.

[5] Tom Brown et al. Language models are few-shot learners. *NeurIPS*, 2020.

[6] Junbum Cha et al. Honeybee: Locality-enhanced projector for multimodal llm. *CVPR*, 2024.

[7] Guiming Hardy Chen et al. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024.

[8] Keqin Chen et al. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[9] Lin Chen et al. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[10] Zhe Chen et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CVPR*, 2024.

[11] Wei-Lin Chiang et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 2023.

[12] Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[13] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 2022.

[14] Chaoyou Fu et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[15] Samir Yitzhak Gadre et al. Datacomp: In search of the next generation of multimodal datasets. *NeurIPS*, 2023.

[16] Jiahui Gao et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

[17] Yash Goyal et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *CVPR*, 2017.

[18] Wenyi Hong et al. Cogagent: A visual language model for gui agents. *CVPR*, 2024.

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*, 2019.

[20] Robert A Jacobs et al. Adaptive mixtures of local experts. *Neural computation*, 1991.

[21] Albert Q Jiang et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[22] Kushal Kafle et al. Dvqa: Understanding data visualizations via question answering. *CVPR*, 2018.

[23] Shankar Kantharaj et al. Chart-to-text: A large-scale benchmark for chart summarization. *ACL*, 2022.

[24] Aniruddha Kembhavi et al. A diagram is worth a dozen images. *ECCV*, 2016.

[25] Geewook Kim et al. Ocr-free document understanding transformer. *ECCV*, 2022.

[26] Alexander Kirillov et al. Segment anything. *ICCV*, 2023.

[27] Ranjay Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[28] Jason J Lau et al. Vqa-rad: A dataset of visual questions and answers in radiology. *Scientific Data*, 2018.

[29] Kenton Lee et al. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *ICML*, 2023.

[30] Dmitry Lepikhin et al. Gshard: Scaling giant models with conditional computation and automatic sharding. *ICLR*, 2021.

[31] Chunyuan Li et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 2023.

[32] Feng Li et al. Semantic-sam: Segment and recognize anything at any granularity. *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. *ICML*, 2023.

[34] Shuheng Li et al. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*, 2023.

[35] Yanwei Li et al. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[36] Ziyi Lin 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 et al. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *ISBI*, 2021.

[38] Fangyu Liu et al. Deplot: One-shot visual language reasoning by plot-to-table translation. *ACL*, 2023.

[39] Fuxiao Liu et al. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *NAACL*, 2024.

[40] Haotian Liu et al. Llava-next: Improved reasoning, ocr, and world knowledge. 2024.

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[42] Yuan Liu et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[43] Pan Lu et al. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *ACL*, 2021.

[44] Pan Lu et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

[45] Pan Lu et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[46] Sai Krishna Mani et al. Pointqa: Provably efficient point cloud question answering. *CVPR*, 2020.

[47] Ahmed Masry et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ACL*, 2022.

[48] Minesh Mathew et al. Docvqa: A dataset for vqa on document images. *WACV*, 2021.

[49] Minesh Mathew et al. Infographicvqa. *WACV*, 2022.

[50] OpenAI. Gpt-4v(ision) system card. 2023.

[51] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[52] Bryan A Plummer et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *ICCV*, 2015.

[53] Joan Puigcerver et al. From sparse to soft mixtures of experts. *ICLR*, 2024.

[54] Alec Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[55] Samyam Rajbhandari et al. Zero: Memory optimizations toward training trillion parameter models. *SC*, 2020.

[56] Carlos Riquelme et al. Scaling vision with sparse mixture of experts. *NeurIPS*, 2021.

[57] Michael S Ryoo et al. Tokenlearner: What can 8 learned tokens do for images and videos? *NeurIPS*, 2021.

[58] Yuzhang Shang et al. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[59] Shuai Shao et al. Objects365: A large-scale, high-quality dataset for object detection. *ICCV*, 2019.

[60] Aleksandar Shtedritski et al. What does clip know about a red circle? visual prompt engineering for vlms. *ICCV*, 2023.

[61] Amanpreet Singh et al. Towards vqa models that can read. *CVPR*, 2019.

[62] Gemini Team et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[63] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[64] Ashish Vaswani et al. Attention is all you need. *NeurIPS*, 2017.

[65] Weihan Wang et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[66] Haoran Wei et al. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.

[67] Haoning Wu et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.

[68] Bin Yan et al. Universal instance perception as object discovery and retrieval. *CVPR*, 2023.

[69] Haoxuan You et al. Ferret: Refer and ground anything anywhere at any granularity. *ICLR*, 2024.

[70] Alex Young et al. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*, 2024.

[71] Licheng Yu et al. Modeling context in referring expressions. *ECCV*, 2016.

[72] Ming-Liang Zhang et al. Plane geometry diagram parsing. *IJCAI*, 2022.

[73] Renrui Zhang et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

[74] Sheng Zhang 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 et al. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[76] Zhuofan Zong et al. Detrs with collaborative hybrid assignments training. *ICCV*, 2023.

[77] Zhuofan Zong et al. Mova: Adapting mixture of vision experts to multimodal context. *NeurIPS*, 2024.

A Appendix

A.1 Vision Expert Details

Table 9 provides detailed configurations of vision experts used in ViMoE.

Table 9: **Vision expert configurations.** We adopt official pretrained checkpoints for all experts.

Model	Params	Resolution	Width	Depth	Output Shape
DINOv2-giant [51]	1.1B	518×518	1536	40	1536×37×37
Co-DETR-large [76]	304M	1280×1280	1024	24	256×80×80
SAM-huge [26]	632M	1024×1024	1280	32	256×64×64
Pix2Struct-large [29]	513M	720×720	1536	18	1536×45×45
Deplot-base [38]	92M	720×720	768	12	768×45×45
Vary-base [66]	86M	1024×1024	768	12	512×32×32
BiomedCLIP-base [74]	86M	224×224	768	12	768×16×16

A.2 Training Data

Pretraining Data. We construct 15M samples covering:

- **Image Captions:** DataComp-1B [15] (4M samples), ShareGPT4V-PT [9], ALLaVA-4V [7]
- **Visual Grounding:** Objects365 [59], RefCOCO [71], VisualGenome [27], PointQA [46], Flickr30K [52]
- **Chart/Document:** MMC-Instruction [39], Chart2Text [23], DVQA [22], SciGraphQA [34], LLaVAR-PT [75], Common Crawl documents (3M)
- **Medical:** LLaVA-Med [31]

Fine-tuning Data. We use LLaVA-665K [41] as the base and add DocVQA [48], ChartQA [47], InfographicVQA [49], AI2D [24], ST-VQA [4], TextVQA [61], SynthDoG-en [25], Geometry3K [43], PGPS9K [72], Geo170K [16], RefCOCO [71], LLaVA-Med [31], VQA-RAD [28], and SLAKE [37].

A.3 Additional Ablations

Effect of Load Balancing Coefficient. Table 10 shows sensitivity to α:

Table 10: Effect of load balancing coefficient α.

α	MME^P	GQA	ChartQA	DocVQA
0	1598	65.8	70.4	83.6
0.001	1605	66.2	71.5	84.5
0.01	**1612**	**66.5**	**72.1**	**85.2**
0.1	1594	65.4	69.8	83.1

Expert Routing Statistics. We analyze routing patterns on different data types:

Table 11: Average expert utilization (%) per data type.

Data Type	DINO	CoDETR	SAM	Pix2St	Deplot	Vary	BioMed
Natural	42.3	18.2	15.1	5.2	3.1	4.8	11.3
Document	8.4	5.1	6.2	38.7	12.3	24.1	5.2
Chart	12.1	8.3	7.8	18.4	35.2	14.1	4.1
Medical	15.2	6.1	8.4	4.2	2.1	3.8	60.2

The routing patterns align with expert specializations: documents heavily use Pix2Struct and Vary, charts prefer Deplot, and medical images route to BiomedCLIP.

A.4 Computational Cost Breakdown

Table 12: Inference latency breakdown (seconds) on A100 GPU.

Component	MoVA	ViMoE
Preprocessing	0.19	0.19
Base encoder	0.05	0.05
LLM routing	0.14	0.14
HCA	-	**0.01**
TLSEA	-	**0.02**
ECC	-	**0.005**
Expert encoders + Adapter	0.07	0.075
LLM generation	10.24	10.24
Total	10.69	10.73

Our novel components (HCA, TLSEA, ECC) add only 0.035s total latency (<0.4% overhead).