Soft Modality-Guided Expert Specialization in MoE-VLMs

Zi-Hao Bo Yaqian Li Anzhou Hou Rinyoichi Takezoe Ertao Zhao
Tianxiang Pan Jiale Yan Mo Guang Kaiwen Long
Li Auto Inc.
{bozihao1, liyaqian, houanzhou, takezoerinyoichi, zhaoertao,
pantianxiang, yanjiale, guangmo, longkaiwen}@lixiang.com

Abstract

Mixture-of-Experts (MoE) has become a prevalent backbone for large vision-language models (VLMs), yet how modality-specific signals should guide expert routing remains under-explored. Existing routing strategies are either hand-crafted or modality-agnostic, relying on idealized priors that ignore the layer-dependent modality fusion patterns in MoE-VLMs and provide little guidance for expert specialization. We propose Soft Modality-guided Expert Specialization (SMoES), which consists of dynamic soft modality scores that capture layer-dependent fusion patterns, an expert binning mechanism aligned with expert-parallel deployment, and an inter-bin mutual information regularization that encourages coherent modality specialization. Our method leverages attention-based or Gaussian-statistics modality scores to optimize mutual information regularization. Experiments across four MoE-based VLMs and 16 benchmarks demonstrate improvement on both effectiveness and efficiency: 0.9% and 4.2% average gain on multimodal and language-only tasks, 56.1% reduction in EP communication overhead, and 12.3% throughput improvement under realistic deployment. These results validate that aligning routing with modality-aware expert specialization unlocks MoE-VLM capacity and efficiency.

1. Introduction

The Mixture-of-Experts (MoE) architecture [14, 19, 27, 45] has emerged as a cornerstone for modern large vision-language models (VLMs). Its principle of conditional computation allows for a massive expansion of model capacity with only a modest increase in the per-token computational budget by routing inputs to specialized expert subnetworks [26, 45]. This paradigm is particularly well-suited for fusing heterogeneous modalities. Consequently, leading systems like DeepSeek-VL2 [60], Kimi-VL [51], GLM-4.5V [22], and InternVL-3.5 [56] have adopted MoE to achieve better performance while maintaining tractable inference costs. Despite this widespread adoption, a fundamental question remains under-explored: how do modality-specific signals (vision vs. text) interact with the expert routing mechanism, and can it be optimized to enhance both effectiveness and efficiency?

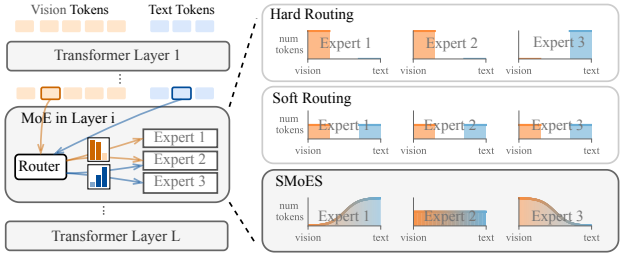

Figure 1. Comparison of routing strategies in MoE-VLMs. Hard routing enforces strict modality-expert separation, while soft routing allows arbitrary, uncontrolled mixing. SMoES uses soft modality scores to encourage dynamic expert specialization.

Mainstream MoE-VLMs route tokens using either hard or soft routing paradigms (Fig. 1). Hard routing [2, 55] pre-assigns experts to a specific modality, creating sharp specialization at the cost of rigid boundaries that hinder adaptation to cross-modal features and ignore the natural blending of representations across layers. In contrast, soft routing [22, 42, 51, 60], the prevailing paradigm, allows experts to process any token. Yet, these methods often rely on heuristic priors or auxiliary losses disconnected from the evolving modality distributions, resulting in either over-mixing or under-specialization. Hybrid approaches [2, 53, 55] partition experts into modality-specific (hard-routed) and shared (soft-routed) groups, but this partitioning is typically hand-crafted and layer-agnostic, failing to align with feature evolution. Fundamentally, these paradigms are guided by idealized assumptions or uncontrolled mixing rather than the dynamic, data-driven geometry of modality distributions, offering little guidance on how expert specialization should evolve with network depth.

We analyze modality fusion patterns of token features across layers in LLaVA-1.5 [35] and a DeepSeekMoE-based [11] VLM, as shown in Fig. 2. The analysis shows that modality fusion in MoE-VLMs is highly heteroge-

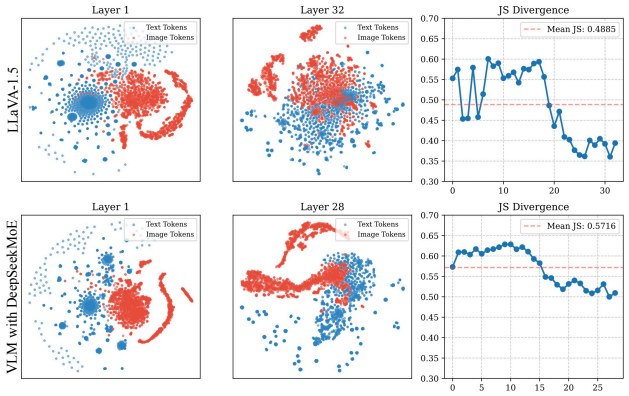

Figure 2. Modality fusion pattern. Token feature distributions vary across models and layers. Text tokens are divided into several semantic clusters, while vision tokens are roughly separated into content and padding clusters. The last column shows the JS divergence between vision and text tokens across all layers.

neous, with no clear boundary between modality-specific and cross-modal states. At a macro level, fusion patterns vary across models and layers, as reflected by distinct Jensen–Shannon (JS) divergence trajectories. At a micro level, even within the same layer and modality, some tokens remain modality-specific while others become cross-modal. This multi-scale heterogeneity indicates that rigid routing assumptions—either enforcing hard separation or imposing uniform mixing—are misaligned with how modalities actually interact across depth. Therefore, guided by signals that respect the evolving modality structure, the expert pool in MoE has the potential to naturally enable dynamic modality specialization.

Beyond effectiveness, modality specialization is also closely tied to the inference efficiency of MoE-VLMs. Vision and text tokens differ sharply in both quantity and information density. Vision tokens often dominate the sequence length but typically carry lower information density due to spatial redundancy, whereas text tokens are fewer yet more semantically concentrated [5, 24, 40]. In real-world scenarios, inputs may even be text-only. This asymmetry poses two efficiency challenges. First, standard routing objectives coupled with vanilla load balancing tend to allocate most experts to the dominant yet low-information-density vision modality, hindering optimal specialization. Second, under expert parallelism (EP) [34]—a common deployment strategy—modality-agnostic routing scatters tokens across devices, inflating inter-communication overhead. Carefully-designed modality specialization addresses both issues: by establishing clear expert–modality affinities and scheduling tokens accordingly, aligned experts can be co-located on the same device, reducing communication while preserving balanced computation and capacity.

Motivated by these observations, we propose Soft Modality-guided Expert Specialization (SMoES), which addresses these challenges through three key components. First, we introduce dynamic soft modality scores through two complementary estimators—attention-accumulated and Gaussian-statistics—that enable adaptive expert selection aligned with actual fusion states rather than hard boundaries. Second, we introduce expert binning—partitioning experts into coherent groups—which serves as the structural foundation for modality specialization and enables natural device placement units for modality-aware expert parallelism deployment. Third, to drive effective modality specialization, we introduce a mutual information (MI) objective that encourages different expert bins to specialize on distinct modality patterns.

In summary, we make three key contributions: (1) Dynamic soft modality scores—Gaussian-statistics and attention-accumulated estimators—that align routing with the evolving modality fusion patterns across layers. (2) An expert binning mechanism coupled with an MI-driven specialization objective that promotes coherent modality grouping for both effectiveness and deployment efficiency. (3) Extensive experiments across four MoE-VLM backbones and 16 benchmarks demonstrating SOTA performance in accuracy and efficiency, including reduced EP communication overhead in realistic deployment.

2. Related Works

2.1. Mixture-of-Experts

Sparsely-gated MoE architectures enable conditional computation through sparse activation and efficient gating [14, 27, 45]. Recent works like DeepSeek-V3 further improve routing efficiency, load balancing, and communication performance [34, 43, 68]. Mutual-information (MI) constraints have been explored to regulate expert selection. For expert-task alignment, Mod-Squad and MTHL maximize MI between tasks and experts [7, 8]. For expert-token alignment, ModuleFormer and CoMoE maximize MI between tokens and modules through entropy balancing or contrastive learning [15, 48]. However, these methods overlook modality differentiation in multimodal models. Recent VLMs increasingly adopt MoE backbones [18, 22, 51, 52, 56, 60]. Related approaches also include LoRA-based expert expansion [6, 46, 59] and converting dense VLMs to sparse MoE architectures [25, 31, 67].

2.2. Modality Specialization in MoE-based VLMs

For modality-aware routing, existing methods can be categorized by their degree of expert-modality coupling. **Hard routing** assigns experts exclusively to single modalities [1, 13, 29, 30, 33, 39, 47], achieving strong specialization but sacrificing flexibility across modalities. **Hybrid routing** allows certain experts to handle multiple modalities while others remain modality-specific [2, 53, 55], yet relies on manual partitioning misaligned with actual fusion dynam-

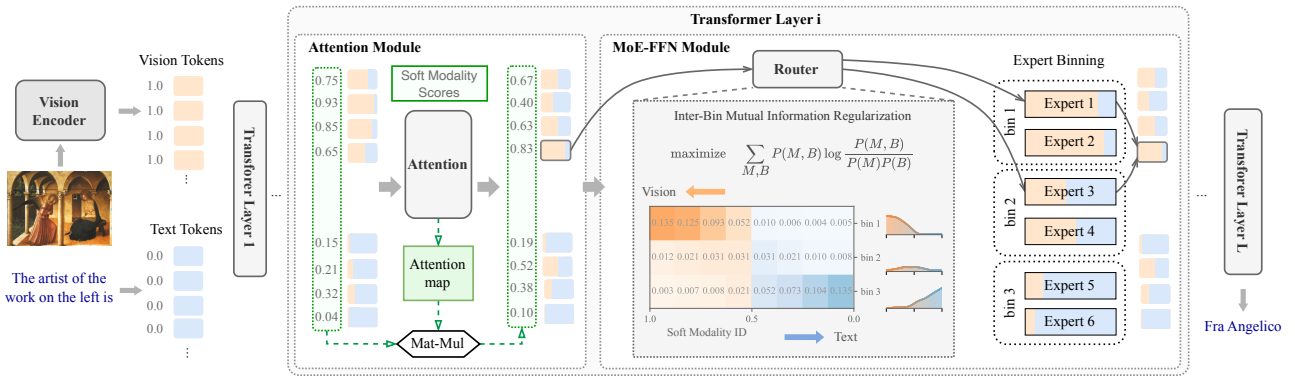

Figure 3. Overview of SMoES. **Soft Modality Scores**: Tokens start with hard modality IDs (0/1 for vision/text) and are progressively refined into layer-wise soft scores $M \in [0, 1]$. We estimate them via (i) attention-accumulated (depicted here) or (ii) Gaussian-statistics estimation (detailed in Fig. 4). **Expert Binning**: experts learn bin-level modality preference for deployment efficiency. **Mutual Information** Regularization: maximizes mutual information between soft modality scores (M) and bins (B), inducing modality preference.

ics. **Naive soft routing** [4, 42] permits flexible expert mixing without explicit modality guidance, leaving specialization to implicit optimization. **Smarter soft routing** introduces auxiliary objectives to shape routing: LTDR balances expert load on long-tailed vision distributions [3], and STGC mitigates gradient conflicts within experts [63]. SMAR achieves modality differentiation by using KL divergence to regularize routing distributions toward modality-specific patterns [61]. In contrast to KL-based regularization, our approach employs mutual information constraints between modalities and experts to guide specialization. Moreover, we introduce layer-adaptive soft modality scores that capture dynamic fusion patterns across layers, and expert binning for efficient distributed deployment.

2.3. Efficient Deployment of MoE Models

As MoE models scale, deployment efficiency becomes increasingly critical, with load imbalance and All-to-All communication overhead as key challenges in distributed expert-parallel (EP) settings [66]. To mitigate load imbalance, MoGE enforces balanced expert activation through predefined groups [50], while Grove-MoE explores heterogeneous expert sizes that dynamically adapt to token complexity [58]. Complementary strategies address the straggler effect: Expand Drop employs capacity-aware token dropping to limit expert overload [20], and AEP decouples layer execution with asynchronous queuing and adaptive re-batching [54]. However, for MoE-VLMs, how to leverage modality fusion characteristics to guide expert partitioning and communication strategies under EP deployment remains largely underexplored.

3. Method

Modality fusion in MoE-VLMs exhibits layer-varying patterns with smooth token-level transitions, while vision/text asymmetry creates load imbalance and inflates EP communication overhead. To address these challenges, we introduce SMoES, which integrates three components: soft modality scores capturing continuous token-level modality patterns, expert binning creating modality-aligned partitions, and inter-bin mutual information regularizing coherent specialization, as illustrated in Fig. 3.

3.1. Preliminaries: MoE in VLM

A typical MoE-based VLM consists of a vision encoder, projection layer, and an MoE-augmented LLM backbone. Within each MoE layer l, token features $\mathbf{x}_{ij} \in \mathbb{R}^D$ (where i indexes the sample and j indexes the token) are routed by a gating network to a pool of N_e experts. The router computes gating scores $g_{ij,e} = \text{softmax}(\mathbf{W}_{\text{gate}}\mathbf{x}_{ij})_e$, and standard top-$k$ routing selects the k highest-scoring experts. To prevent routing collapse, a load-balancing auxiliary loss [14, 45] is typically applied:

$$\mathcal{L}_{\text{bal}} = \sum_{l} N_e \sum_{e=1}^{N_e} f_e \, P_e \tag{1}$$

where f_e is the fraction of tokens routed to expert e, and P_e is the average gating score.

3.2. Soft Modality Scores

The multi-scale heterogeneity of modality fusion (Fig. 2) reveals smooth, layer-dependent transitions in token representations rather than fixed modality identities. Hard modality indicators—binary labels assigned at the input—fail to capture these continuous fusion dynamics. To guide routing aligned with evolving modality patterns, we introduce *soft modality scores* $M^{(l)}_{ij,m} \in [0, 1]$ for each token and modality $m \in \{\text{text}, \text{vision}\}$ at layer l, where $\sum_m M^{(l)}_{ij,m} = 1$. We develop two complementary estimators: an *attention-accumulated score* that captures local cross-token interactions; and a *Gaussian-statistics score* that captures global statistical regularities across the dataset.

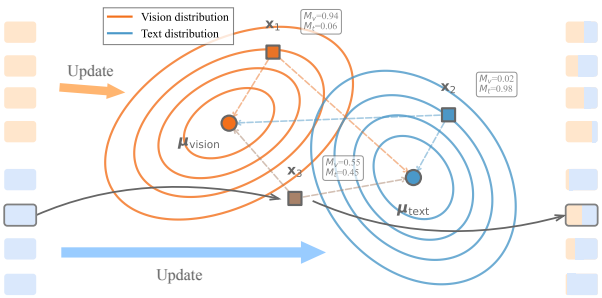

Figure 4. **Gaussian-statistics score estimation.** Per-modality Gaussian distributions are continuously updated from token batches. For each token, we compute modality affinity as soft scores based on log-likelihood under each distribution.

3.2.1. Attention-Accumulated Score

The attention mechanism provides a natural pathway for modality signal propagation: when a token attends to others, it absorbs their modality characteristics proportionally to attention weights. We initialize scores at layer 0 with hard modality indicators:

$$M_{ij,m}^{\text{attn},(0)} = \mathbf{1}\{m = m(\mathbf{x}_{ij})\}. \tag{2}$$

For subsequent layers, we update in two steps. First, aggregate attended tokens' scores using attention weights:

$$\tilde{M}_{ij,m}^{\text{attn},(l)} = \sum_{j'=1}^{J} \text{Attn}_{j,j'}^{(l)} \cdot M_{ij',m}^{\text{attn},(l)} \tag{3}$$

where $\text{Attn}^{(l)}$ is the layer l attention matrix averaged across heads, and J is the sequence length. Second, combine aggregated and input scores via residual-weighted update:

$$M_{ij,m}^{\text{attn},(l+1)} = \frac{\|\mathbf{x}_{\text{attn},ij}^{(l)}\| \cdot \tilde{M}_{ij,m}^{\text{attn},(l)} + \|\mathbf{x}_{ij}^{(l)}\| \cdot M_{ij,m}^{\text{attn},(l)}}{\|\mathbf{x}_{\text{attn},ij}^{(l)}\| + \|\mathbf{x}_{ij}^{(l)}\|} \tag{4}$$

This mirrors the residual structure of Transformers ($\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} + \mathbf{x}_{\text{attn}}^{(l)}$), where feature norms weight the contributions of attention and residual paths.

3.2.2. Gaussian-Statistics Score

Complementary to the attention-accumulated score's local, sequence-specific perspective, we propose the *Gaussian-statistics score*—a global, distribution-based estimator. The key insight is that tokens from different modalities exhibit distinct feature distributions in the embedding space. We exploit these distributional differences to instantaneously infer modality affiliation at each layer, independent of Layer 0 initialization.

For each layer l and modality m, we maintain a Gaussian distribution with diagonal covariance: mean $\boldsymbol{\mu}_m \in \mathbb{R}^D$ and variance $\boldsymbol{\sigma}_m^2 \in \mathbb{R}^D$. Statistics are updated online via exponential moving average (EMA) variant of Welford's algorithm [57]:

$$N_{m,t} = \beta N_{m,t-1} + |\mathcal{T}_m| \tag{5}$$

$$\mathbf{S}_{\mu,m,t} = \beta \mathbf{S}_{\mu,m,t-1} + |\mathcal{T}_m| \cdot \boldsymbol{\mu}_{\mathcal{T}_m} \tag{6}$$

$$\mathbf{S}_{\sigma^2,m,t} = \beta \mathbf{S}_{\sigma^2,m,t-1} + \sum_{\mathbf{x}\in\mathcal{T}_m} (\mathbf{x} - \boldsymbol{\mu}_{\mathcal{T}_m})^2 + (\boldsymbol{\mu}_{\mathcal{T}_m} - \tfrac{\mathbf{S}_{\mu,m,t-1}}{N_{m,t-1}})^2 \cdot \tfrac{\beta N_{m,t-1}|\mathcal{T}_m|}{\beta N_{m,t-1}+|\mathcal{T}_m|} \tag{7}$$

where $\mathcal{T}_m$ is the set of modality-m tokens in the current batch, $\boldsymbol{\mu}_{\mathcal{T}_m}$ is the batch mean, and β is the EMA decay factor. The distribution parameters at time t are: $\boldsymbol{\mu}_m = \mathbf{S}_{\mu,m,t}/N_{m,t}$ and $\boldsymbol{\sigma}_m^2 = \mathbf{S}_{\sigma^2,m,t}/N_{m,t}$.

To infer each token's modality affiliation, we compute its log-likelihood under each distribution:

$$\text{LL}_{ij,m} = -\frac{1}{2}\sum_{d=1}^{D}\left(\log \sigma_{m,d}^2 + \frac{(x_{ij,d} - \mu_{m,d})^2}{\sigma_{m,d}^2}\right) \tag{8}$$

The soft modality score is obtained via temperature-scaled softmax:

$$M_{ij,m}^{\text{gauss}} = \frac{\exp(\text{LL}_{ij,m}/\tau)}{\sum_{m'\in\{\text{text,vision}\}} \exp(\text{LL}_{ij,m'}/\tau)} \tag{9}$$

where temperature τ controls score sharpness.

3.3. Expert Binning

In expert-parallel (EP) deployments, modality-agnostic routing scatters tokens across devices, amplifying communication overhead. To reduce this, we co-locate experts with similar modality preferences on the same device. Moreover, real deployments vary in device count, requiring a flexible binning mechanism.

We introduce *expert binning*: at each layer, partition experts into N_{bins} groups $\mathbf{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_{N_{\text{bins}}}\}$ with $N_B = N_e/N_{\text{bins}}$ experts each, where N_{bins} can match device count. By encouraging bins to develop distinct modality preferences through inter-bin MI (detailed below), we enable modality-aligned placement: bins with similar preferences co-locate, reducing communication while balancing load.

To form bins reflecting modality preferences, we adopt *momentum-adaptive binning*. We track each expert's load from each modality using EMA:

$$\bar{C}_{m,e,t} = \beta \bar{C}_{m,e,t-1} + (1-\beta) C_{m,e} \tag{10}$$

where $C_{m,e} = |\{(i,j) \in \mathcal{T}_m \mid e \in \text{TopK}(g_{ij})\}|$ is the count of modality-m tokens routed to expert e. We compute each expert's text-bias score:

$$f_{\text{spec}}(e) = \frac{\bar{C}_{\text{text},e}}{\bar{C}_{\text{text},e} + \bar{C}_{\text{vision},e}} \tag{11}$$

Experts are sorted by f_{spec} and partitioned into N_{bins} consecutive bins, grouping experts with similar modality preferences together.

3.4. Inter-Bin Mutual Information Regularization

We introduce an *inter-bin MI objective* to drive modality differentiation among expert bins. The intuition is that high MI between modality M and bin $\mathbf{B}$ means knowing the selected bin provides strong information about token modality, thereby encouraging specialized bins. Mathematically, we maximize $I(M;\mathbf{B})$.

To compute MI, we first compute the average gating score for each sample i, modality m, and bin $\mathbf{B}_k$:

$$\bar{S}_{i,m,\mathbf{B}_k} = \frac{\sum_{e\in\mathbf{B}_k}\sum_j M_{ij,m}\cdot g_{ij,e}}{N_B\sum_j M_{ij,m}} \tag{12}$$

where $M_{ij,m}$ is the soft modality score and $g_{ij,e}$ is the gating score. We treat this as an unnormalized distribution and compute the normalized joint probability:

$$P_i(m,\mathbf{B}_k) = \frac{\bar{S}_{i,m,\mathbf{B}_k}}{\sum_{m'}\sum_{k'}\bar{S}_{i,m',\mathbf{B}_{k'}}} \tag{13}$$

Marginal probabilities are derived from joint probability, and the MI is:

$$I_i(M;\mathbf{B}) = \sum_m\sum_k P_i(m,\mathbf{B}_k)\log\frac{P_i(m,\mathbf{B}_k)}{P_i(m)\cdot P_i(\mathbf{B}_k)} \tag{14}$$

The loss over all layers encourages specialization:

$$\mathcal{L}_{\text{MI}} = -\sum_l\frac{1}{N_{\text{batch}}}\sum_{i=1}^{N_{\text{batch}}} I_i(M;\mathbf{B}) \tag{15}$$

Operating at the bin level (rather than expert level) naturally aligns with device placement granularity in EP: bins with distinct modality affinities can be co-located on devices, reducing cross-device communication.

3.5. Training Objective and Implementation Details

In EP deployments, balanced load within each device is essential. We adopt bin-level load-balancing loss:

$$\mathcal{L}_{\text{bal}} = \sum_l\sum_{k=1}^{N_{\text{bins}}} N_B\sum_{e\in\mathbf{B}_k} f_e\,P_e \tag{16}$$

where f_e and P_e are computed for tokens in bin $\mathbf{B}_k$ (as defined in Sec. 3.1).

The full objective combines task loss, per-bin load-balancing, and inter-bin MI:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha_{\text{bal}}\mathcal{L}_{\text{bal}} + \alpha_{\text{MI}}\mathcal{L}_{\text{MI}} \tag{17}$$

where $\mathcal{L}_{\text{task}}$ is language modeling loss.

We train on 8 NVIDIA A800 GPUs with $N_{\text{bins}} = 8$. For Gaussian-soft modality scores, the temperature is $\tau = 0.5\cdot D$. EMA decay is $\beta = 0.99$ for both Gaussian updates and momentum-adaptive binning. Loss weights are $\alpha_{\text{bal}} = 0.001$ and $\alpha_{\text{MI}} = 0.0001$. Further implementation details and ablation studies are in supplementary material.

4. Results

We conduct extensive experiments to evaluate SMoES across multiple dimensions: effectiveness on diverse VLM and language-only tasks, ablation studies to validate each design choice, visualization of modality specialization, and computational efficiency under expert-parallel deployment.

4.1. Experimental Setup

Model Setup. We build VLMs with CLIP ViT-L/14 vision encoder [44], a 2-layer MLP projector, and MoE language backbones. We evaluate on four architectures spanning different scales and designs: DeepSeekMoE [11], OLMoE [41], Moonlight-MoE [36], and Qwen3-MoE [62]. Models are initialized from public checkpoints and fine-tuned with the same protocol. We compare against four routing strategies: Soft Routing [42], Hard Routing [39], MoIIE [53] (hybrid-routing), and SMAR [61] (KL-divergence).

Training Data and Benchmarks. We follow LLaVA's two-stage training protocol using LLaVA-Pretrain-558K and LLaVA-Instruct-665K datasets respectively [35]. Evaluation covers 10 multimodal benchmarks: MMMU-val/test [64], GQA [23], POPE [28], SQA-img [38], TextVQA [49], MME [16], MMBench/MMBench-CN [37], VQA-v2 [17]; and 6 language benchmarks: MMLU [21], HellaSwag [65], ARC-C/E [9], GSM8k [10], TruthfulQA [32].

Modality Specialization Index (MSI). We introduce MSI to quantify modality-based expert specialization. Modality affiliation probability for expert e at layer l:

$$\tilde{\mathbf{C}}^{(l)}_{m,e} = \frac{\mathbf{C}^{(l)}_{m,e}/\sum_{e'}\mathbf{C}^{(l)}_{m,e'}}{\sum_{m'}\left(\mathbf{C}^{(l)}_{m',e}/\sum_{e'}\mathbf{C}^{(l)}_{m',e'}\right)} \tag{18}$$

MSI measures average deviation from uniform distribution across experts and layers:

$$\text{MSI} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{E}\sum_{e=1}^{E} 2\cdot\left|\tilde{\mathbf{C}}^{(l)}_{\text{text},e} - 0.5\right| \tag{19}$$

where MSI $\in[0,1]$: 0 = no specialization, 1 = perfect specialization.

4.2. Main Results

We present comprehensive results across all four backbones and 16 benchmarks. Results on DeepSeekMoE and OLMoE are shown in Tab. 1. Results on Moonlight-MoE and Qwen3-MoE are shown in supplementary material.

Across four backbones, SMoES improves over soft routing baseline by 2.2% average (0.9% multimodal, 4.2% language-only). Soft routing lacks modality-specialization guidance and suffers from data imbalance and load-balancing constraints that constrain expert capability.

Table 1. Multimodal and language-only results on VLMs based on DeepSeekMoE and OLMoE. **Bold** and underline: first and second best performance. MSI: Modality Specialization Index. †t/v/s: number of text/vision/shared experts. *[a, b]: KL divergence threshold range.

Method	MSI	Multimodal Tasks (10)											Language-Only Tasks (6)							Overall
		MMMU^{val}	*MMMU^{test}*	*GQA*	*POPE*	*SQA-IMG*	*TextVQA*	*MME*	*MMB*	*MMB-CN*	*VQAv2*	**Avg**	*MMLU*	*HellaSwag*	*ARC-C*	*ARC-E*	*GSM8k*	*TruthfulQA*	**Avg**	**Avg**
VLM based on DeepSeek-MoE (A3B/16B, top-6/64 experts)																				
No Specialization	.177	31.9	30.5	59.5	84.9	68.0	56.5	1718	62.1	61.7	77.4	100%	46.2	49.1	52.3	75.4	10.5	43.8	100%	100%
Hard Routing [39]																				
t32-v32†	1.	30.2	29.2	58.2	**85.6**	63.3	56.3	1555	59.9	57.6	76.2	-3.9%	37.4	42.4	40.4	60.7	2.5	41.2	-26.2%	-12.3%
t48-v16†	1.	30.3	29.7	58.6	84.6	66.0	55.4	1657	62.5	61.5	76.9	-1.8%	41.9	46.6	50.2	72.4	5.3	37.5	-14.5%	-6.6%
MoIIE [53]																				
t16-v16-s32†	.504	32.1	29.9	58.4	85.2	65.1	55.9	1656	62.2	60.1	76.9	-1.5%	41.4	45.6	49.6	72.2	5.8	40.9	-13.1%	-5.8%
t24-v24-s16†	.752	31.9	29.8	58.5	84.6	63.0	55.9	1594	61.1	59.9	76.6	-2.6%	41.2	42.8	47.9	68.2	2.1	42.8	-20.7%	-9.3%
t32-v16-s16†	.800	30.4	29.7	58.5	84.8	64.7	55.8	1667	62.3	61.1	76.9	-1.9%	42.3	50.0	50.9	72.2	6.7	40.2	-9.6%	-4.8%
SMAR [61]																				
d_{KL}*-[0.5, 1.0]**	.543	31.2	**31.5**	58.7	85.5	69.3	57.3	1714	63.6	61.7	77.3	+0.6%	42.1	43.1	47.3	67.3	7.4	45.2	-11.3%	-3.9%
d_{KL}*-[1.5, 2.0]**	.648	32.9	30.2	59.1	85.2	68.3	57.6	**1730**	60.9	58.9	77.3	-0.2%	41.7	44.3	40.6	61.7	6.3	47.7	-15.3%	-5.8%
d_{KL}*-[2.5, 3.0]**	.743	32.2	29.9	58.3	84.9	66.8	57.1	1678	60.6	59.1	77.1	-1.3%	40.7	39.2	32.3	47.5	5.5	46.5	-25.0%	-10.2%
SMoES (ours)																				
attention-soft	.487	**34.7**	31.0	59.6	85.1	69.0	**58.3**	1706	63.5	62.3	77.4	**+1.8%**	**46.5**	**56.4**	**56.0**	**79.0**	**10.9**	46.6	**+6.2%**	**+3.5%**
gaussian-soft	.440	32.4	30.8	**59.9**	85.6	**69.6**	57.5	1689	**65.1**	**62.5**	**77.5**	+1.3%	46.4	53.6	53.9	78.0	10.2	**49.0**	+4.2%	+2.4%
VLM based on OLMoE (A1B/7B, top-8/64 experts)																				
No Specialization	.205	30.0	29.4	58.0	84.9	66.8	**56.6**	**1667**	**62.4**	49.4	75.7	100%	49.9	48.3	59.0	79.1	**32.6**	44.2	100%	100%
Hard Routing [39]																				
t32-v32†	1.	28.4	27.4	57.2	84.8	61.2	50.2	1595	57.3	42.5	74.6	-6.1%	36.3	35.4	40.1	59.6	3.5	42.1	-34.1%	-16.6%
t48-v16†	1.	31.3	29.0	58.2	85.6	65.3	53.3	1603	59.7	45.5	75.6	-2.0%	43.7	42.4	51.4	73.8	13.4	47.0	-16.2%	-7.3%
MoIIE [53]																				
t16-v16-s32†	.509	28.9	28.9	57.6	**85.7**	65.5	53.6	1542	59.4	47.2	75.1	-3.0%	43.6	44.7	51.7	74.1	14.1	47.5	-14.7%	-7.4%
t24-v24-s16†	.754	31.0	28.8	57.6	85.2	62.7	52.3	1507	57.5	44.2	75.0	-4.2%	41.4	44.9	45.8	71.1	3.8	46.2	-23.4%	-11.4%
t32-v16-s16†	.800	32.1	29.3	57.4	85.0	63.8	53.0	1586	60.2	47.5	75.2	-1.8%	41.8	37.2	49.6	74.7	13.9	46.8	-18.7%	-8.2%
SMAR [61]																				
d_{KL}*-[0.5, 1.0]**	.381	32.9	29.6	58.0	84.8	67.6	54.5	1640	60.7	46.3	75.8	-0.3%	48.1	50.8	58.5	77.8	25.0	43.1	-4.5%	-1.9%
d_{KL}*-[1.5, 2.0]**	.485	**33.1**	29.6	57.8	84.8	65.5	54.0	1620	61.5	47.7	75.7	-0.4%	49.4	55.1	58.7	80.2	26.5	46.0	-0.1%	-0.3%
d_{KL}*-[2.5, 3.0]**	.645	32.4	**29.8**	58.0	84.3	66.2	55.6	1650	60.6	47.8	75.8	-0.1%	48.0	50.5	56.5	79.6	26.6	**50.9**	-1.1%	-0.5%
SMoES (ours)																				
attention-soft	.620	31.5	29.7	**58.3**	84.7	66.5	55.8	1644	62.3	50.5	**75.9**	+0.5%	**50.8**	**62.2**	**60.3**	80.8	31.8	47.7	**+6.7%**	**+2.9%**
gaussian-soft	.754	31.4	29.6	58.1	85.1	**67.9**	55.4	1643	62.2	**50.6**	75.7	**+0.6%**	50.1	55.6	58.7	**81.1**	31.4	49.5	+4.3%	+2.0%

Hand-crafted modality specialization (hard/hybrid routing) appears intuitive but inevitably mismatches the complex expert capacity and dynamic data distributions across layers. Hard routing achieves near-perfect modality specialization but incurs substantial performance degradation (-4.4% multimodal, -22.2% language-only). Hybrid routing (MoIIE) provides some recovery (-3.1% multimodal, -17.4% language-only) but remains well below the soft routing baseline. This reveals that rigid specialization cannot be blindly enforced; it must be learned and adaptively tailored to actual expert capacity and data dynamics.

Why do mainstream MoE-VLMs avoid explicit modality control? Our hard/hybrid routing results explain this: incorrectly-designed modality specialization schemes incur severe performance penalties. SMAR introduces an automatic approach through KL divergence regularization rather than explicit hand-crafted assignment. Since the original SMAR was designed for models with small expert counts (no more than eight), while ours has numerous small experts, we test various KL divergence regularization strengths in our setting. SMAR achieves significant MSI gain while maintaining more competitive performance: incurring only 1.0% and 15.1% performance loss on multimodal and language-only tasks. However, SMAR suffers from incompatibility between KL divergence regularization and load-balancing constraints (it disabled load-balancing in the final model), making it difficult to leverage specialization in MoE models with numerous small experts.

Our SMoES addresses this by unifying soft modality scores and inter-bin MI objectives, achieving strong modality specialization while simultaneously maximizing expert capacity utilization and maintaining load balance. Our attention-based soft modality estimation achieves average gains of 1.0% on multimodal tasks and 4.4% on language-only tasks, while our Gaussian-statistics variant achieves 0.9% and 4.1% respectively. Both significantly outperform

Table 2. Ablation on the modality score type (DeepSeekMoE).

Method	MSI	Multi-Modal	Language	Overall
No Specialization	.177	100%	100%	100%
SMAR (best)	.543	+0.6%	-11.3%	-3.9%
SMoES				
hard-score	.904	-0.8%	+0.5%	-0.3%
attention-soft	.487	**+1.8%**	**+6.2%**	**+3.5%**
gaussian-soft	.440	+1.3%	+4.2%	+2.4%

Table 3. Ablation on inter-bin specialization (DeepSeekMoE).

Method	MSI	Multi-Modal	Language	Overall
No Specialization	.177	100%	100%	100%
w/ binning	.415	+0.9%	+3.0%	+1.7%
w/ inter-bin				
KL	.724	-1.5%	-8.5%	-4.1%
MI-attention	.487	**+1.8%**	**+6.2%**	**+3.5%**
MI-gaussian	.440	+1.3%	+4.2%	+2.4%

all baseline methods, validating that well-designed learnable specialization schemes can unlock performance gains in large-scale MoE-VLMs while preserving load balance.

4.3. Ablation Studies

4.3.1. Modality Score and Specialization Objective

We first demonstrate the effectiveness of soft modality scores. Hard-score uses binary labels based on input source modality; although it achieves high MSI, it cannot improve model performance. Both attention-soft and gaussian-soft scores significantly outperform hard-score (Tab. 2). Notably, hard-score with MI-based specialization still exceeds the best SMAR configuration, validating the importance of our specialization objective.

We next evaluate our MI-based specialization with expert binning (Tab. 3). Expert-binning itself improves performance (+1.7%) by providing structure for modality-aware specialization. Our inter-bin MI objective further boosts gains (+3.5% for attention-soft, +2.4% for gaussian-soft), demonstrating that MI effectively guides experts toward coherent modality-specific patterns. In contrast, KL-divergence in SMAR fails to improve performance.

4.3.2. Expert Binning Strategy and Granularity

We evaluate our adaptive binning strategy (based on modality-aware EMA statistics) versus fixed binning (original expert order). Table Tab. 4 shows adaptive binning consistently outperforms fixed binning across both binning-only and MI specialization settings, confirming the effectiveness of modality-aware bin formation.

We next examine bin granularity by varying the number of bins (Fig. 5). Too many bins can increase deployment imbalance despite enabling finer specialization, while too few bins reduce specialization effectiveness.

Table 4. Ablation on expert binning strategies (DeepSeekMoE).

Method	MSI	Multi-Modal	Language	Overall
No Specialization	.177	100%	100%	100%
w/ binning				
fixed	.357	+0.9%	+2.9%	+1.6%
adaptive	.415	+0.9%	+3.0%	+1.7%
attention-soft				
fixed	.450	+2.0%	+0.2%	+1.3%
adaptive	.487	+1.8%	+6.2%	+3.5%
gaussian-soft				
fixed	.398	+1.9%	-1.0%	+0.8%
adaptive	.440	+1.3%	+4.2%	+2.4%

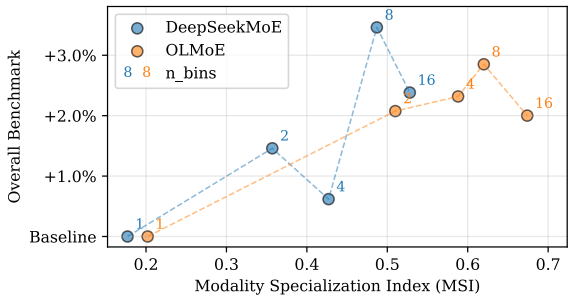

Figure 5. Ablation on number of expert bins ($\text{SMoES}_{\text{attention-soft}}$).

4.4. Visualization and Analysis

SMoES achieves clear bin-modality correspondences, as visualized in Fig. 6. Early layers exhibit higher MSI and sharper expert-modality separation, while deeper layers show balanced distributions with more modality fusion.

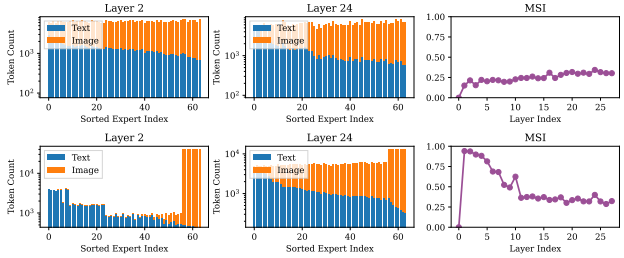

Figure 6. Routing distribution of tokens to experts in DeepSeek-MoE. First row: soft baseline; second row: $\text{SMoES}_{\text{attention-soft}}$.

Expert modality specialization evolves dynamically during training, as shown in Fig. 7. Soft routing without specialization results in experts handling both modalities simultaneously, constraining capacity. Our method shows sharper differentiation in shallow layers where tokens maintain clearer modality identity, and more fusion-aware adaptation in deeper layers.

Our soft modality scores exhibit smooth fusion transitions across layers, as illustrated in Fig. 8. Both attention-soft and Gaussian-soft estimators show increasing fusion intensity toward deeper layers. A slight difference is that attention-soft is biased toward text, while Gaussian-soft is

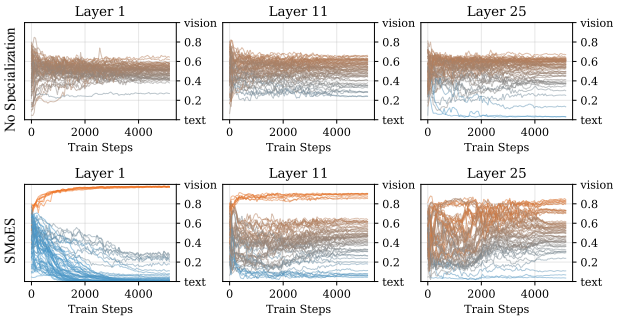

Figure 7. Evolution of expert specialization during training.

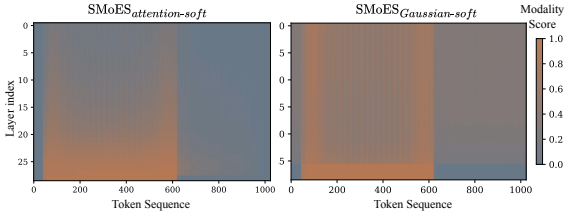

Figure 8. Soft modality score across layers in DeepSeekMoE.

biased toward vision in high-fusion regimes.

4.5. Efficiency Analysis

Beyond task accuracy, SMoES reduces expert-parallel (EP) communication overhead by aligning expert placement with modality preferences. We deploy on two NVIDIA Orin GPUs via 10Gb Ethernet using EP, representing a typical edge-side scenario in autonomous vehicles (Fig. 9). EP avoids weight and KV Cache redundancy compared to TP/DP, making it memory-efficient for edge resources. Baseline uses synchronous transmission since balanced experts provide no benefit from asynchrony, while SMoES's increased local expert concentration enables asynchronous transmission, overlapping communication and computation similar to PD separation.

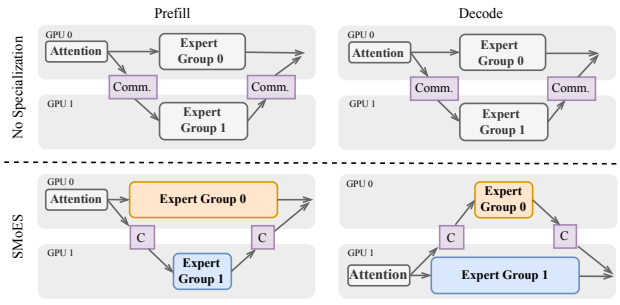

Figure 9. Expert-parallel (EP) deployment on two GPUs.

Tab. 5 shows cross-GPU EP transfer ratios for vision and text tokens at prefill and decode stages on OLMoE. Expert specialization routes tokens to local experts more frequently, reducing inter-device communication. We report separate ratios for vision and text tokens (quantities differ between phases) and the overall V+T ratio. DeDup (de-duplication) is employed to avoid duplicate transmission of tokens routed to experts on the same device in top-k routing.

Table 5. Cross-GPU EP transfer ratio at prefill/decode stages. V: Vision tokens; T: Text tokens.

Benchmark	Method	Prefill (P) V	Prefill (P) T	Prefill (P) V+T	Decode (D) T
MMMU (PV:PT:DT=32:8:1)	baseline	97.7%	99.5%	98.0%	86.5%
	SMoES	15.0%	99.3%	31.1%	43.3%
	Δ	↓84.6%	↓0.3%	↓68.3%	↓49.9%
SQA-IMG (PV:PT:DT=14:7:1)	baseline	97.5%	99.7%	98.2%	94.9%
	SMoES	13.0%	99.2%	40.6%	49.9%
	Δ	↓86.6%	↓0.5%	↓58.7%	↓47.4%

Table 6. TTFT and TPOT speed improvement of SMoES compared to soft-routing baseline. Δ: speedup percentage.

Benchmark	Method	Batch Size=1 TTFT(s)	Batch Size=1 TPOT(s)	Batch Size=8 TTFT(s)	Batch Size=8 TPOT(s)
MMMU	baseline	2.810	0.786	7.949	1.414
	SMoES	**2.519**	**0.703**	**6.203**	**1.287**
	Δ	↓10.3%	↓10.5%	↓22.0%	↓9.0%
SQA-IMG	baseline	1.493	0.766	5.824	1.278
	SMoES	**1.356**	**0.692**	**4.859**	**1.134**
	Δ	↓9.2%	↓9.7%	↓16.6%	↓11.3%

Tab. 6 shows TTFT for Prefill and TPOT for Decode. Performance improvement stems from reduced communication overhead and parallel execution of computation and communication. Larger batch sizes in Prefill increase communication proportion, yielding greater gains, while Decode maintains stable improvement ratios across batch sizes due to fewer activated experts.

5. Conclusion

In this paper, we addressed the challenge of modality-guided expert specialization in MoE-VLMs. We introduced SMoES, which consists of dynamic soft modality scores that capture layer-dependent fusion patterns, an expert binning mechanism aligned with expert-parallel deployment, and an inter-bin mutual information regularization that encourages coherent modality specialization. Extensive experiments across four MoE backbones and 16 benchmarks validate our approach, demonstrating consistent improvements in task accuracy while simultaneously reducing communication overhead and increasing throughput in expert-parallel deployments. Our work has some limitations, for example, the Gaussian-soft estimator currently uses a simple unimodal Gaussian with diagonal covariance for efficiency. Although we have explored GMM as a preliminary extension in the supplementary material, further refining richer density models remains an open problem. This work demonstrates the value of data-driven modality specialization and opens promising avenues for future exploration in expert specialization for MoE-based multimodal learning.

References

[1] Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025. 2

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in neural information processing systems*, 35:32897–32912, 2022. 1, 2

[3] Chaoxiang Cai, Longrong Yang, Kaibing Chen, Fan Yang, and Xi Li. Long-tailed distribution-aware router for mixture-of-experts in large vision-language model. *arXiv preprint arXiv:2507.01351*, 2025. 3

[4] Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1110–1119, 2024. 3

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

[6] Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024. 2

[7] Zitian Chen, Mingyu Ding, Yikang Shen, Wei Zhan, Masayoshi Tomizuka, Erik Learned-Miller, and Chuang Gan. An efficient general-purpose modular vision model via multi-task heterogeneous training. *arXiv preprint arXiv:2306.17165*, 2023. 2

[8] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023. 2

[9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 5

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 5

[11] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. 1, 5

[12] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022. 2

[13] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2

[14] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 1, 2, 3

[15] Jinyuan Feng, Chaopeng Wei, Tenghai Qiu, Tianyi Hu, and Zhiqiang Pu. Comoe: Contrastive representation for mixture-of-experts in parameter-efficient fine-tuning. *arXiv preprint arXiv:2505.17553*, 2025. 2

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 5

[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 conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5

[18] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 2

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 1

[20] Shwai He, Weilin Cai, Jiayi Huang, and Ang Li. Capacity-aware inference: Mitigating the straggler effect in mixture of experts. *arXiv preprint arXiv:2503.05066*, 2025. 3

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. 5

[22] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 1, 2

[23] 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*, pages 6700–6709, 2019. 5

[24] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 2

[25] Linglin Jing, Yuting Gao, Zhigang Wang, Wang Lan, Yiwen Tang, Wenhai Wang, Kaipeng Zhang, and Qingpei Guo. Evomoe: Expert evolution in mixture of experts for multimodal large language models. *arXiv preprint arXiv:2505.23830*, 2025. 2

[26] Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*, 2021. 1

[27] 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. *arXiv preprint arXiv:2006.16668*, 2020. 1, 2

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 5

[29] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Unimoe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[30] Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multimodal foundation models. *Transactions on Machine Learning Research*, 2025. 2

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

[32] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021. 5

[33] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024. 2

[34] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1, 5

[36] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025. 5, 1

[37] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 5

[38] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 5

[39] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24960–24971, 2025. 2, 5, 6, 3

[40] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024. 2

[41] Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 5, 1

[42] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022. 1, 3, 5

[43] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[44] 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*, pages 8748–8763. PmLR, 2021. 5, 1

[45] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 1, 2, 3

[46] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070, 2024. 2

[47] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models

with sparse mixture of experts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 2

[48] Yikang Shen, Zheyu Zhang, Tianyou Cao, Shawn Tan, Zhenfang Chen, and Chuang Gan. Moduleformer: Modularity emerges from mixture-of-experts. *arXiv preprint arXiv:2306.04640*, 2023. 2

[49] 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*, pages 8317–8326, 2019. 5

[50] Yehui Tang, Xiaosong Li, Fangcheng Liu, Wei Guo, Hang Zhou, Yaoyuan Wang, Kai Han, Xianzhi Yu, Jinpeng Li, Hui Zang, et al. Pangu pro moe: Mixture of grouped experts for efficient sparsity. *arXiv preprint arXiv:2505.21411*, 2025. 3

[51] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 1, 2

[52] Bin Wang, Bojun Wang, Changyi Wan, Guanzhe Huang, Hanpeng Hu, Haonan Jia, Hao Nie, Mingliang Li, Nuo Chen, Siyu Chen, et al. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding. *arXiv preprint arXiv:2507.19427*, 2025. 2

[53] Dianyi Wang, Siyuan Wang, Zejun Li, Yikun Wang, Yitong Li, Duyu Tang, Xiaoyu Shen, Xuanjing Huang, and Zhongyu Wei. Moiie: Mixture of intra-and inter-modality experts for large vision language models. *arXiv preprint arXiv:2508.09779*, 2025. 1, 2, 5, 6, 3

[54] Shaoyu Wang, Guangrong He, Geon-Woo Kim, Yanqi Zhou, and Seo Jin Park. Toward cost-efficient serving of mixture-of-experts with asynchrony. *arXiv preprint arXiv:2505.08944*, 2025. 3

[55] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 1, 2

[56] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 2

[57] Barry Payne Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3): 419–420, 1962. 4

[58] Haoyuan Wu, Haoxing Chen, Xiaodong Chen, Zhanchao Zhou, Tieyuan Chen, Yihong Zhuang, Guoshan Lu, Zenan Huang, Junbo Zhao, Lin Liu, et al. Grove moe: Towards efficient and superior moe llms with adjugate experts. *arXiv preprint arXiv:2508.07785*, 2025. 3

[59] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[60] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 1, 2

[61] Guoyang Xia, Yifeng Ding, Fengfa Li, Lei Ren, Wei Chen, Fangxiang Feng, and Xiaojie Wang. Smar: Soft modality-aware routing strategy for moe-based multimodal large language models preserving language capabilities. *arXiv preprint arXiv:2506.06406*, 2025. 3, 5, 6

[62] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 5, 1

[63] Longrong Yang, Dong Shen, Chaoxiang Cai, Fan Yang, Tingting Gao, Di Zhang, and Xi Li. Solving token gradient conflict in mixture-of-experts for large vision-language model. *arXiv preprint arXiv:2406.19905*, 2024. 3

[64] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5

[65] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019. 5

[66] Yan Zeng, Chengchuang Huang, Yipeng Mei, Lifu Zhang, Teng Su, Wei Ye, Wenqi Shi, and Shengnan Wang. Efficientmoe: Optimizing mixture-of-experts model training with adaptive load balance. *IEEE Transactions on Parallel and Distributed Systems*, 2025. 3

[67] Haizhong Zheng, Xiaoyan Bai, Xueshen Liu, Zhuoqing Morley Mao, Beidi Chen, Fan Lai, and Atul Prakash. Learn to be efficient: Build structured sparsity in large language models. *Advances in Neural Information Processing Systems*, 37:101969–101991, 2024. 2

[68] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022. 2