arXiv:2602.01037v1 [cs.CV] 1 Feb 2026

VEQ：模态自适应量化：面向混合专家视觉语言模型

广硕 Qin *1 Zhiteng Li *1 Zheng Chen [1] Weihang Zhang [2] Linghe Kong [1] Yulun Zhang †1

摘要

混合专家（MoE）视觉语言模型（VLMs）性能卓越，但会产生高昂的内存和计算成本，因此压缩至关重要。训练后量化（PTQ）是一种有效的免训练技术，可用于解决巨大的内存和计算开销。现有量化范式存在不足，因为它们忽视了两种关键的异质性：视觉和语言词元之间固有的差异，以及不同专家贡献的不均衡性。为弥补这一不足，我们提出视觉专家量化（VEQ），这是一种双重感知量化框架，旨在同时适应跨模态差异和专家间的异质性。具体而言，VEQ 包含 1）**模态-专家感知量化**，该方法利用专家激活频率，优先对关键专家进行误差最小化；以及 2）**模态-亲和度感知量化**，该方法将 token-专家亲和度与模态信息相结合，构建增强海森矩阵，以指导校准过程。在多个不同基准上的广泛实验表明，VEQ 的表现始终优于当前最先进的基线方法。具体而言，在 W3A16 配置下，与此前当前最优的量化方法相比，我们的方法在 Kimi-VL 和 Qwen3-VL 上的平均准确率分别提升了 2.04% 和 3.09%，展现出在各种多模态任务中的卓越鲁棒性。我们的代码将在以下地址开放：https://github.com/guangshuoqin/VEQ。

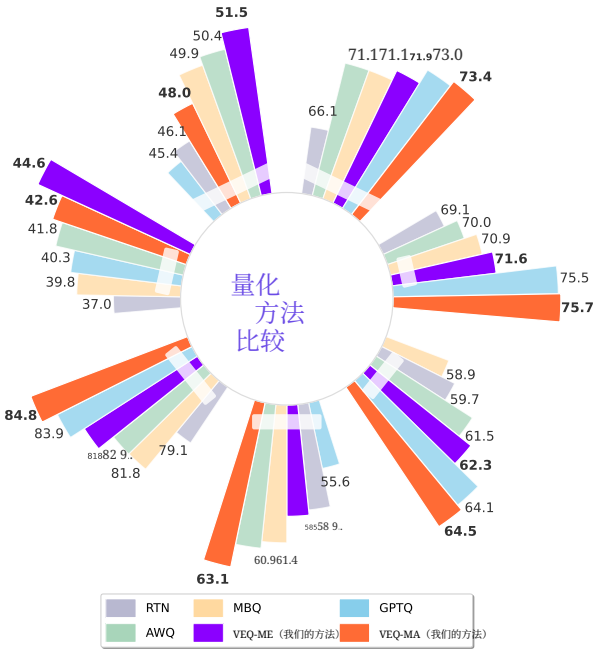


图 *1.* Kimi-VL-Instruct 在 3 位权重量化（W3A16）下的零样本性能。我们的方法始终优于现有基线方法，展现出更强的稳健性。

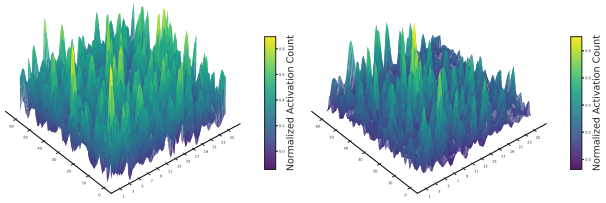
1. 引言

近年来，视觉语言模型（VLMs）在广泛的多模态任务领域展现出前所未有的能力，涵盖基础的视觉问答和图像描述，以及复杂的视觉推理（Radford 等人，2021年；Alayrac 等人，2022年；Li 等人，2023年）。这些模型将视觉感知与语言语义无缝对齐，有效弥合了跨模态鸿沟，使智能体能够以接近人类的多样化能力感知、推理并与物理环境交互（Lu 等人，2019年；Hurst 等人，2024年）。随着对强健多模态理解的需求日益增长，视觉语言模型在实际应用中的重要性也与日俱增。

为了在保持计算效率的同时进一步扩大模型容量，混合专家（MoE）架构已被最先进水平的视觉语言模型广泛采用。与对每个令牌都激活全部参数的稠密模型不同，混合专家模型利用稀疏路由机制，仅激活一部分专家，从而在保留庞大参数空间的同时有效降低推理成本（Fedus等人，2022年）。知名开源视觉语言模型，如DeepSeek-VL2（Wu等人，2024年）、Kimi-VL（Team等人，2025年）、Qwen3-VL（Bai等人，2025年）以及ERNIE-4.5-VL

*贡献相同 [1]上海交通大学 [2]清华大学。通讯作者：Yulun Zhang <yu- lun100@gmail.com>。

预印本。*2026*年*2*月*3*日。



(a) 文本令牌。 *(b)*视觉令牌。

图*2.* 不同模态间激活特征的比较分析。峰值表示高激活频率。

（Baidu-ERNIE-Team，2025年）已成功利用该架构，以可控的资源消耗实现更优性能。尽管相较对应的稠密模型具有效率优势，混合专家视觉语言模型在推理过程中仍会产生显著的内存占用与延迟，因此需要有效的模型压缩技术。训练后量化（PTQ）已成为一种实用的解决方案。主流仅权重量化方法，如AWQ（Lin 等人，2024年）和GPTQ（Frantar 等人，2022年），主要侧重于最小化权重重构误差，以保持模型性能。与此同时，SmoothQuant（Xiao等人，2024年）和SpinQuant（Liu等人，2024b年）等激活感知方法通过平滑或旋转特征空间来应对激活离群值带来的挑战。最近，MBQ（Li等人，2025年）通过探索多模态令牌的敏感性提出了一种新视角，以进一步最小化量化误差。

然而，将这些传统量化方法直接应用于混合专家视觉语言模型，效果并不理想。现有方法通常将模型视为单一的稠密结构，因此忽视了混合专家架构固有的结构稀疏性。具体而言，这些方法未能考虑不同专家的重要性差异。如图2所示，少数热门专家被更频繁地调用，并主导模型输出，而其他专家则处于休眠状态（Fedus等人，2022年；Chi等人，2022年）。此外，当前方法往往忽视视觉令牌和文本令牌之间显著不同的统计分布异质性（Liang等人，2022年）。对这些模态采用统一的量化策略会导致性能显著下降，因为连续视觉嵌入与离散文本表示对量化噪声的敏感度存在很大差异。

为应对这些挑战，我们提出了视觉专家量化（**VEQ**），这是一种专为混合专家视觉语言模型设计的新型框架。我们在领先的混合专家架构上进行了全面评估，包括 Kimi-VL（Team等人，2025年）和 Qwen3-VL（Bai等人，2025年），并使用 MMMU（Yue 等人，2024年）、MME-RealWorld（Zhang 等人，2024b年）、MMBench（Liu 等人，2024a年）和 InfoVQA（Mathew 等人，2022年）等具有挑战性的多模态基准测试进行测试。实证结果表明，VEQ 始终优于已有基线方法（例如，AWQ（Lin 等人，2024年）、GPTQ（Frantar 等人，2022年）、MBQ（Li 等人，2025年）），为量化混合专家视觉语言模型确立了新的最先进水平。值得注意的是，VEQ 在低比特设置下展现出更强的鲁棒性。据我们所知，这是首次尝试压缩大规模混合专家视觉语言模型。

我们的主要贡献总结如下：

- **混合专家视觉语言模型的开创性框架：**据我们所知，本研究首次尝试在量化领域同时应对多模态异质性与基于混合专家（MoE）的架构所特有的结构属性这两大挑战。

- **模态专家感知量化：**我们提出了一种新颖策略，根据视觉和文本词元的路由频率为专家分配重要性分数。通过显式建模这些路由模式，我们利用这些分数有效降低关键专家的量化误差。

- **模态亲和性感知量化：**我们引入了一种改进方法，利用路由器亲和度逻辑值和输入词元模态对海森矩阵重新加权。该方法将语义亲和度纳入优化过程，进一步提升量化模型的精度。

2. 相关工作

2.1. 视觉语言模型量化

视觉语言模型（VLMs）的训练后量化（PTQ）仍是一个具有挑战性且相对探索不足的领域，主要原因在于视觉和文本模态之间的分布异质性。VLMQ（Xue 等人，2025年）通过提出重要性感知目标函数来应对视觉词元冗余。它借助轻量级分块反向传播，生成纳入词元级重要性因子的增强海森矩阵。Q-VLM（Wang等人，2024年）利用激活熵作为代理来识别跨层依赖关系，从而实现高效分块。它进一步优化视觉编码器以解耦这些依赖关系，从而在保持准确率的同时降低搜索开销。MBQ（Li等人，2025年）考虑到视觉和语言词元具有不同的敏感度水平，将基于梯度的敏感度指标纳入校准过程，旨在平衡各模态之间的重构损失。Bi-VLM（Wang等人，2025年）实现了一种显著性感知混合量化算法，该算法基于高斯分位数对权重进行非均匀划分。该方法为显著离群值分配更高精度，同时对其余参数进行二值化。MQuant（Yu等人，2025年）引入模态专属静态量化和注意力不变切换机制，以应对分布差异。此外，它采用旋转幅度抑制（RMS）来缓解在线哈达玛变换所引起的离群值。

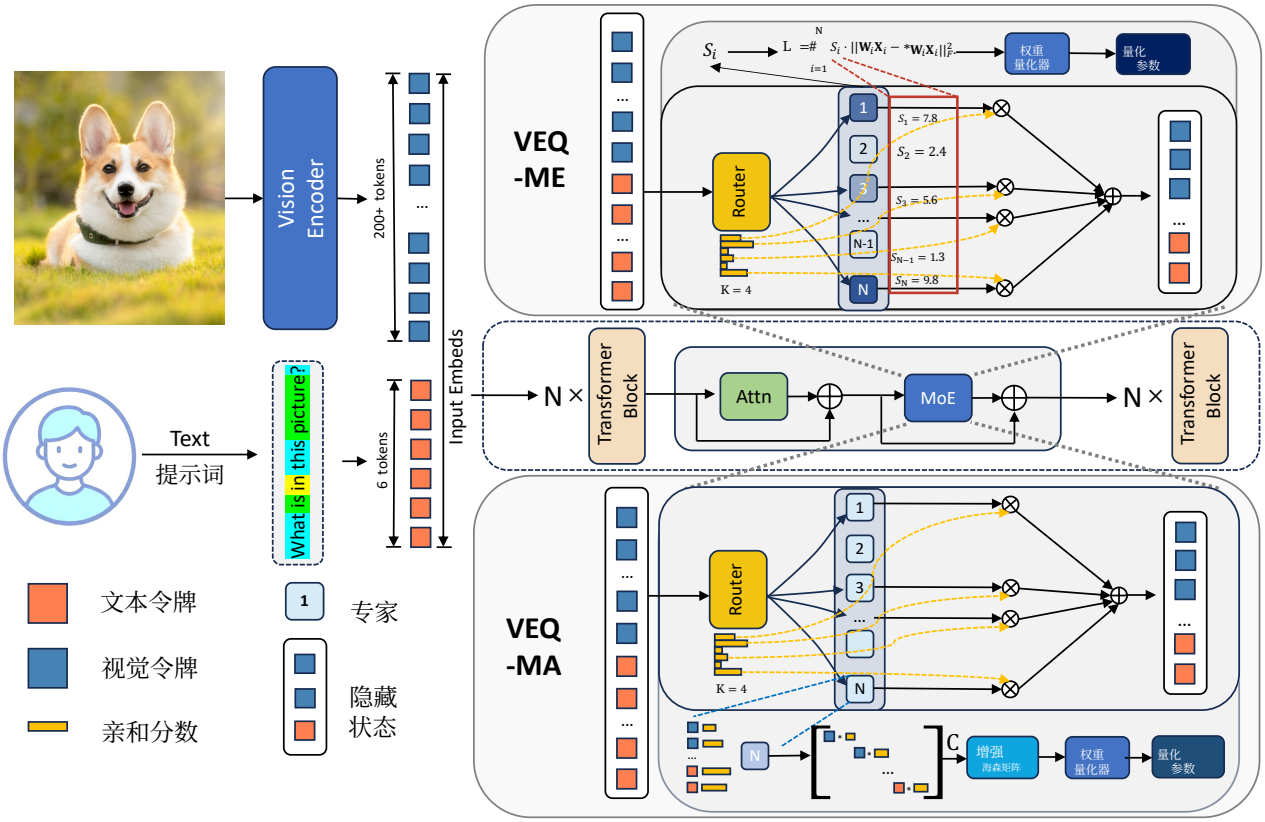


图3. 所提出的VEQ框架概览。我们的方法由两个核心组件组成：（1）**VEQ-ME**，根据专家的激活频率动态地为专家分配重要性分数 S_i，从而在重构损失中优先最小化关键专家的误差；以及（2）**VEQ-MA**，通过整合令牌-专家亲和度分数与模态敏感性构建增强海森矩阵，使校准过程能够适应多模态令牌的不同敏感性。

2.2. 混合专家大语言模型量化

训练后量化同样给混合专家（MoE）架构带来了显著困难，这源于固有的专家稀疏性以及令牌与专家之间复杂的亲和度。近期多项研究致力于解决这些问题。MoEQuant（Hu等人，2025年）采用专家均衡自采样方法进行校准，以应对专家间与专家内的不平衡。它进一步采用亲和度引导量化策略，根据令牌与专家的相关性对误差进行加权。MoQa（Zheng等人，2025年）通过多阶段数据-模型分布分析实现专家级混合精度基础量化，并辅以通道级动态调整机制，以适应新的数据分布。MoQE（Zhang等人，2025年）利用混合专家架构实现推理加速：将单一模型的多种量化变体视为专家，并使用轻量级路由器将输入数据动态分配给最优量化专家。MxMoE（Duanmu等人，2025年）提出了一种准确率-性能协同设计框架，以线性块为粒度分配位宽。该方法通过分析参数敏感度与专家激活频率，优化用于量化的混合精度配置。

3. 方法

本节提出**视觉专家量化（VEQ）**，这是一种专为混合专家视觉语言模型设计的新型训练后量化框架。我们首先在第3.1节分析混合专家视觉语言模型固有的异质性。基于这些观察，我们在第3.2节构建模态感知专家重要性指标。最后，我们在第3.3节介绍一种亲和度感知量化算法，以最小化重构误差。

3.1. 混合专家视觉语言模型量化中的异质性

3.1.1. 模态异质性

不同模态在混合专家层中呈现出独特的路由模式和敏感性水平。为量化这种异质性，我们使用监督微调（SFT）损失的梯度作为指标，衡量视觉和文本令牌的误差敏感性。

视觉词元具有空间冗余的特点，通常表现出较低的梯度幅值，这意味着它们对推理结果的影响敏感度较低。相比之下，信息密集型文本令牌主导着输出分布。如图4所示，我们对 COCO（Lin 等人，2014年）数据集中的 128 个样本进行了分析

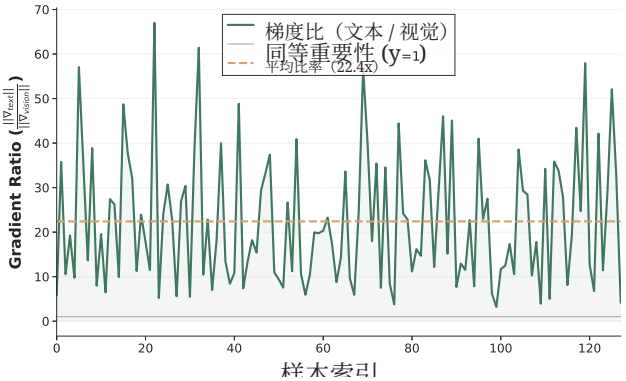


图 *4.* 对 COCO（Lin 等人，2014年）数据集中 128 个样本的梯度幅值进行分析。与视觉token 相比，文本词元的梯度范数明显更高，平均比率为 22.4。

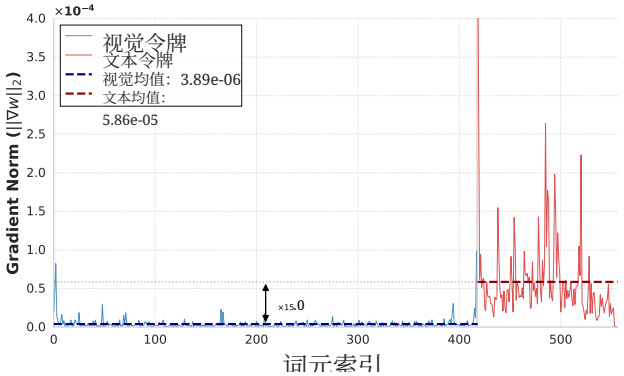


图*5.* 代表性样本（样本88）的梯度分析细节。该可视化结果显示，文本与视觉梯度比约为 15，证实了文本信息在推理过程中占据主导地位。

结果表明，文本词元的平均梯度幅值是视觉词元的**22.4**倍。在模态极度不平衡的样本中（例如，只有 5 个文本令牌，却有 200 多个视觉令牌），这种差异尤为明显，梯度比会急剧上升。这表明，尽管文本令牌数量稀少，但它们对最终生成结果的影响却大得不成比例。进一步考察一个具体案例（样本88，详见图5）也印证了这一观察：文本与视觉的平均梯度比为**15**。不同场景下如此显著的梯度差距，凸显了模态间根本性的敏感度不平衡。

这一差异表明，对所有专家采用统一的量化策略会忽略模态特有的敏感性。对高影响力的文本词元采用与冗余视觉令牌相同的粒度，可能会损害性能，尤其是在精确语言生成至关重要的跨模态任务中。

3.1.2. 专家的异质性

为了解跨模态处理的内部机制，我们研究了混合专家层中的路由行为。我们的分析聚焦于不同模态下的激活模式和专家偏好，并揭示了有关专家亲和性的三个关键发现。

内在稀疏性与负载不平衡。 首先，我们分析专家激活固有的稀疏性。无论校准数据集的组成如何，专家群体中始终存在负载不平衡。值得注意的是，即使将校准批次大小增加到 64（约 30,000 个令牌），某些专家接收到的输入词元数量仍为零。这一现象表明，专家利用率分布不均是预训练权重的内在属性，而非数据采样造成的假象。因此，鉴于模型输出主要由少数专家主导，对所有专家采用统一指标来衡量量化误差并非最优选择。

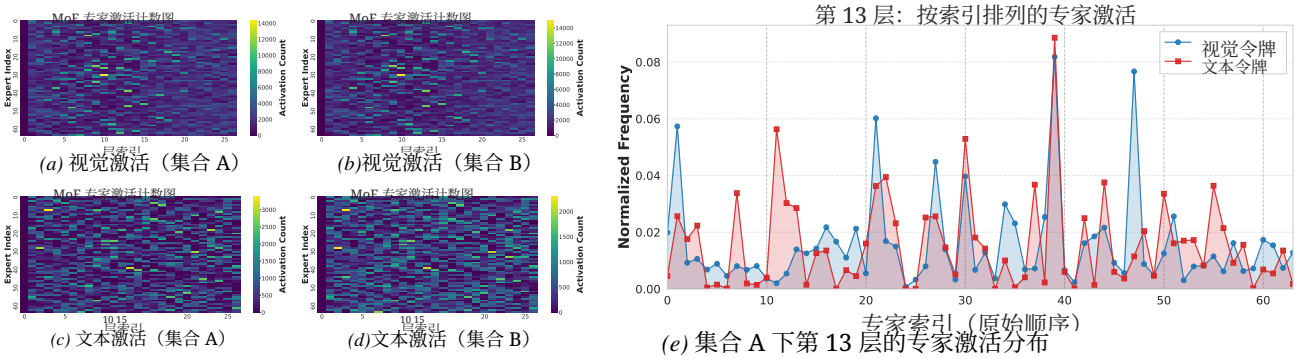
令牌分布异质性。 其次，对路由路径的分析揭示了微妙的功能分工，其特点是通用专家与模态专属专家并存。一部分专家充当通用处理器，可由视觉和文本输入共同激活；而其他专家则表现出不同的路由偏好。

具体而言，如图 6(e) 所示，我们观察到专精集群自发涌现：其中一些专门处理空间冗余视觉特征，另一些则只关注文本信息的语义密度。这种分布证实，模型既为跨模态对齐分配了容量，也为模态专属处理分配了容量。这些集群在激活频率和幅值上的差异，凸显了不同专家贡献的不均衡。因此，必须根据这些模态依赖的激活特性分配量化误差权重，确保优化策略能够适应每个专家的特定功能角色。

路由偏差与关键专家。 最后，路由器输出表现出显著偏差，少数专家对模型输出起着关键作用。路由概率分布高度倾斜；对于任意给定的令牌，路由器只会为少数专家（top-k 选择）分配较高的置信度分数，而其余专家的亲和度则趋近于零。梯度幅度分析表明，这些高亲和度专家主导着推理结果，而未被选中专家的影响微乎其微。这意味着，保持这些关键专家的精度对于维持模型性能至关重要。

3.2. 模态专家感知量化

基于第3.1节的分析，我们已确定，混合专家层中的专家在激活频率和模态特异性方面都存在显著差异。具有代表性的训练后量化框架，例如



(a) 视觉激活（集合 A） (b)视觉激活（集合 B）

(c) 文本激活（集合 A） (d)文本激活（集合 B）

(e) 集合 A 下第 13 层的专家激活分布

图6. 专家亲和模式的可视化。子图(a)-(d)展示了视觉和文本令牌在不同输入范围内截然不同的激活分布，这些范围来自**COCO 数据集**（Lin 等人，2014年），突出了专家激活如何随不同输入样本而变化，同时保持稀疏性与模态特定聚类。子图(e)可视化了Kimi-VL-Instruct（Team等人，2025年）在集合A下第13层的激活特征，突出了固有的负载不平衡。

AWQ（Lin 等人，2024年）通常通过网格搜索确定最优量化参数，以最小化通用重构损失。然而，这种搜索过程对所有专家一视同仁，未能考虑它们在推理过程中的重要性差异。

为解决这一问题，我们提出一种模态-专家感知量化方法，将专家重要性纳入误差最小化目标。

量化专家重要性。首先，我们将第 i个专家的重要性分数 S_i 定义为其跨模态贡献的均衡度量。由于混合专家视觉语言模型中固有的模态不平衡，视觉token的数量显著多于文本词元，原始频率计数会不成比例地偏向以视觉为主的专家。

因此，我们引入一种全面的归一化机制，以更好地平衡不同模态的贡献幅度。重要性分数 S_i 定义为加权求和：

$$S_i = \gamma \cdot N_i^{\text{text}} + \beta \cdot N_i^{\text{vis}}, \tag{1}$$

其中，N_i^{text}和 N_i^{vis}分别表示路由至第 i 个专家的文本令牌数和视觉令牌数。设 T_{text} 和 T_{vis}分别表示校准集中文本令牌和视觉令牌的总数。系数 $\beta = T_{\text{text}}/T_{\text{vis}}$用作**数量归一化因子**，用于缩减频繁出现的视觉激活，使其与文本计数相当。系数 $\gamma = \|\nabla_{\text{text}}\|/\|\nabla_{\text{vis}}\|$ 则用作**质量敏感因子**，体现文本词元更高的梯度影响。这一公式确保最终得分由令牌的语义重要性而非其原始频率决定，从而保护负责逻辑推理的关键专家。

重要性感知优化目标。如上所述，AWQ 等成熟的训练后量化策略（Lin 等人，2024年）通过基于网格的搜索来优化量化尺度，并将混合专家层视为一个整体。这些方法对模型采用统一的优化目标。具体而言，在混合专家模块中，这一搜索过程等价于最小化所有专家重建误差的无权重总和：

$$\mathcal{L}_{\text{Standard}} = \sum_{i=1}^{M} \|\mathbf{W}_i\mathbf{X}_i - \hat{\mathbf{W}}_i\mathbf{X}_i\|_F^2, \tag{2}$$

ˆ其中，$\mathbf{W}_i$ 和 $\mathbf{W}_i$ 分别表示第 i 个专家的全精度权重和量化权重，$\mathbf{X}_i$ 表示路由至该专家的输入令牌，M 为专家总数。这一公式将不常用的专家与关键专家一视同仁，未能考虑它们对模型最终推理结果的不同贡献。

相比之下，我们提出的方法通过引入重要性分数 S_i（定义见公式1）作为加权因子，重新定义了优化目标。加权量化误差表示如下：

$$\mathcal{L}_{\text{Weighted}} = \sum_{i=1}^{M} S_i \cdot \|\mathbf{W}_i\mathbf{X}_i - \hat{\mathbf{W}}_i\mathbf{X}_i\|_F^2. \tag{3}$$

在搜索最优量化参数时，通过最小化 $\mathcal{L}_{\text{Weighted}}$，可以优先抑制激活频率高且模态敏感度高的专家中的量化噪声。这确保量化参数经过校准后，能够保留最具影响力的专家的功能，从而准确反映量化对模型下游任务整体性能的真实影响。

3.3. 模态亲和性感知量化

其他一些具有代表性的训练后量化框架，例如GPTQ（Frantar 等人，2022年），通常依赖二阶信息来确定最优量化参数。具体而言，它们利用海森矩阵 H，并通过输入校准数据计算其近似值 X，如 $H = 2XX^\top$。这种形式隐含地假设所有输入词元对重建误差的贡献相同，因而将序列维度上的优化空间视为均匀分布。

亲和度与模态不平衡。然而，正如第3.1节所分析的，这种均匀性假设并不适用于混合专家视觉语言模型。输入词元与专家的交互存在显著差异：（1）路由多样性：词元与特定专家的亲和度各不相同，这由路由器输出的概率决定；（2）模态敏感性：文本词元虽然数量较少，但与空间上冗余的视觉token相比，具有更高的梯度密度和信息价值。

将统一的海森矩阵计算直接应用于混合专家层，无法考虑这些细微差异，可能会导致关键语义信息被忽略。

增强海森矩阵。为解决这一问题，我们提出了一种有效的海森矩阵模态-亲和度感知表述。我们引入了一种逐token重要性加权机制。令 $X \in \mathbb{R}^{d\times N}$ 表示专门路由至当前专家的输入词元，其中 N 为此类词元的数量。我们根据每个词元的模态特定亲和度缩放其贡献，从而重构海森矩阵 $\tilde{H}$：

$$\tilde{H} = (X \cdot \sqrt{\mathbf{C}})(X \cdot \sqrt{\mathbf{C}})^\top = X\mathbf{C}X^\top, \quad (4)$$

其中$\mathbf{C} \in \mathbb{R}^{N\times N}$ 是一个对角矩阵，表示每个词元的重要性权重。对于第 j个词元 x_j，相应的对角元素 c_j 定义如下：

$$c_j = p_j \cdot \alpha_j, \quad \text{where } \alpha_j = \begin{cases} \gamma & x_j \text{ is text token,} \\ 1 & x_j \text{ is vision token.} \end{cases} \quad (5)$$

此处， p_j 表示词元 j 与目标专家之间的亲和度。项 γ 表示梯度缩放因子，其形式化定义为文本与视觉模态之间梯度幅值的比值： $\gamma = \|\nabla_{\text{text}}\|/\|\nabla_{\text{vis}}\|$。

将 $\mathbf{C}$ 纳入海森矩阵计算后，量化目标会动态地重新加权。路由器亲和度高且模态敏感性强的令牌，对 $\tilde{H}$的结构产生更大的影响。因此，在应用逆海森矩阵更新权重时，算法会优先保留这些高影响力令牌的精度，从而最大限度地维持模型在低比特设置下的性能。

4. 实验

4.1. 实验设置

我们开展了全面的实验，以评估所提出量化方法的有效性。

模型与基准测试。 我们选择使用 Kimi- VL-Instruct（Team等人，2025年）和 Qwen3-VL-30B-A3B- Instruct（Bai等人，2025年）进行实验。为确保对多模态能力进行稳健评估，我们采用了一组广受认可且具有多样性的基准测试，包括用于多学科推理的 MMMU（Yue等人，2024年）和 MMBench（Liu等人，2024a年），用于图表理解的 AI2D（Kembhavi等人，2016年）、用于图表理解的 InfoVQA（Mathew等人，2022年）和 TextVQA（Singh等人，2019年），用于光学字符识别相关视觉问答；此外，还包括 MME-RealWorld（Zhang等人，2024b年）、RealWorldQA（Team，2024年）、ScienceQA（Lu等人，2022年）和 VizWiz-VQA（Gurari等人，2019年），以涵盖现实世界和科学场景。

基线方法。 我们将所提出的方法与全精度版本及多种成熟的量化基线进行比较，以展示其优越性：1）BF16：采用 Bfloat16 精度的原始模型，作为性能上限；2）RTN（舍入到最近值）：一种朴素的基线方法，通过将权重舍入到最近的网格点进行量化；3）先进的训练后量化框架：我们纳入 GPTQ（Frantar 等人，2022年）和 AWQ（Lin 等人，2024年），它们被广泛用作大语言模型压缩的标准基线方法；此外还包括 MBQ（Li 等人，2025年），这是一种专为视觉语言模型量化设计的当前最优方法。

实现细节。我们的评估流程基于开源的 Lmms-eval（Zhang 等人，2024a年）框架。该框架是一款标准化工具包，专为严格评估大型多模态模型而设计。推理后端采用 SGLang（Zheng 等人，2024年），这是一款针对混合专家架构优化的高吞吐量服务引擎，可提升推理效率。在评估过程中，系统会生成模型输出，并将其与各基准测试对应的标准答案进行比较，以计算最终准确率分数。

4.2. 主要结果

我们在 Kimi-VL-Instruct（Team等人，2025年）和 Qwen3-VL-30B-A3B-Instruct（Bai等人，2025年）模型上，将所提出的方法与最先进的基线方法进行全面评估。为便于下文分析，我们将基于 AWQ（**Lin 等人**，2024年）实现的模态专家感知量化称为 VEQ-ME，而将基于 GPTQ 实现的**模态亲和性感知量化**称为 VEQ-MA。为全面评估量化的鲁棒性，我们在 4 位（W4）和 3 位（W3）权重量化设置下进行了实验。七项多模态基准测试的详细对比结果见表1。

W4 设置下的性能。如表格上半部分所示，1，大多数方法在 4 位量化设置下都保持了稳健的性能。具体而言，对于 Kimi-VL-Instruct 模型，成熟的基线方法，如 AWQ（Lin 等人，2024年）、MBQ（Li et al.，2025年）和 GPTQ（Frantar 等人，2022年），以及我们提出的 VEQ，恢复了 BF16 平均准确率的近 98%。不同量化策略之间的性能差距相对较小，这表明 4 位精度足以表示混合专家权重，而不会造成灾难性的信息损失。

表1。 Kimi-VL-Instruct 和 Qwen3-VL-30B-A3B-Instruct 在 3 位（W3）和 4 位（W4）仅权重量化下的主要对比结果。我们报告所有任务的零样本准确率（%）。每种位宽下的最佳结果以**粗体**标出。VEQ-MA 相较于最佳基线方法的提升幅度以括号标注。缩写：InfoV：InfoVQA，TextV：TextVQA，RWQA：RealWorldQA，SciQA：ScienceQA，Viz：VizWiz，MMB：MMBench，MME-R：MME-RealWorld。

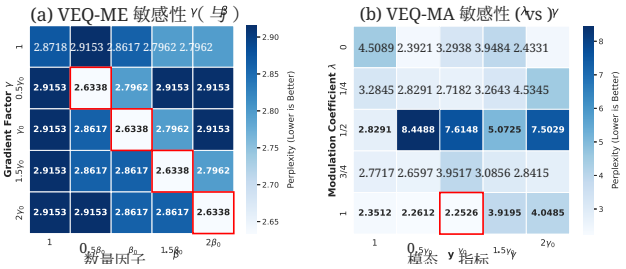
模型	Bit	方法	基准测试									Avg.
			MMMU						Viz	MMB		
Kimi-VL-指令微调	**BF16**	-	51.11	83.65	83.38	85.93	66.41	92.10	69.00	82.73	58.22	74.73
	W3	RTN	37.00	69.14	59.68	74.58	56.47	79.06	58.51	66.06	46.13	60.74
		AWQ	41.78	70.05	61.49	74.04	58.17	81.84	61.37	71.13	50.44	63.37
		MBQ	39.76	70.89	58.86	74.35	58.74	81.82	60.87	71.13	49.91	62.93
		GPTQ	40.33	75.49	64.14	64.30	58.82	83.87	55.61	73.02	45.41	62.33
		VEQ-ME（我们的方法）	**44.56**	71.57	62.33	73.36	**58.95**	82.93	58.90	71.91	**51.46**	64.00
		VEQ-MA（我们的方法）	42.56	**75.65**	**64.48**	**78.30**	58.30	**84.85**	**63.10**	**73.42**	48.03	**65.41** (+2.04)
	W4	RTN	48.44	82.38	80.16	84.60	**66.80**	90.83	**69.17**	80.67	52.48	72.84
		AWQ	49.00	81.74	79.01	83.37	66.27	90.92	69.07	80.58	53.70	72.63
		MBQ	48.89	81.70	78.49	83.18	64.58	90.29	69.03	80.93	55.08	72.46
		GPTQ	50.00	82.32	80.04	84.63	64.84	91.39	68.86	81.44	53.14	72.96
		VEQ-ME（我们的方法）	49.22	81.41	79.75	83.69	66.36	91.06	69.17	80.84	54.20	72.86
		VEQ-MA（我们的方法）	**50.30**	**82.55**	**80.37**	**84.71**	66.41	**91.56**	68.91	**81.67**	**55.61**	**73.57** （+0.61）
Qwen3-VL-30B-A3B-指令版	**BF16**	-	73.67	86.27	81.43	81.31	65.49	93.52	83.27	85.91	59.92	78.98
	W3	RTN	57.33	77.10	44.96	68.64	43.01	89.08	64.07	78.60	**45.73**	63.17
		AWQ	58.89	73.61	44.62	67.93	45.36	88.40	61.75	77.06	42.38	62.22
		MBQ	50.56	71.44	40.15	64.23	53.82	87.31	59.48	76.46	45.27	60.97
		GPTQ	64.89	74.11	**47.27**	71.82	54.25	88.23	63.39	69.39	44.08	64.16
		VEQ-ME（我们的方法）	60.56	73.38	44.95	69.91	53.73	88.40	62.47	77.75	44.24	63.93
		VEQ-MA（我们的方法）	**65.89**	**79.15**	47.14	**72.74**	**54.77**	**89.55**	**69.46**	**82.30**	43.25	**67.14** (+3.09)
	W4	RTN	63.11	**83.33**	62.62	79.77	45.23	91.51	71.08	85.13	57.31	70.89
		AWQ	65.20	81.22	58.32	77.40	63.14	91.06	69.73	85.22	57.31	72.07
		MBQ	69.67	81.41	59.36	78.99	64.31	91.51	70.78	83.68	56.10	72.87
		GPTQ	**72.67**	82.93	62.54	80.18	53.46	**93.23**	71.32	84.53	57.96	73.20
		VEQ-ME（我们的方法）	68.78	82.47	59.53	78.57	**65.32**	91.61	70.97	83.56	56.79	73.07
		VEQ-MA（我们的方法）	71.56	82.95	**62.86**	**81.10**	61.70	92.64	**72.42**	**85.88**	**58.12**	**74.36** (+1.16)

W3 设置下的性能。在激进的 3 位设置下，不同方法之间的差异显著扩大。如表格下半部分所示，1，RTN 和 AWQ（Lin 等人，2024年）等传统基线方法性能严重下降，尤其是在 MMMU（Yue et al.，2024年）等推理密集型任务，以及 InfoVQA（Mathew et al.，2022年）等细粒度视觉任务上。这表明，当位宽极其有限时，均匀量化假设会失效。这也证明，忽略异质性会在极端压缩下导致误差。相比之下，VEQ 展现出非凡的鲁棒性。通过显式建模专家重要性和模态亲和性，VEQ 的表现显著优于基线方法。例如，在 TextVQA（Singh et al.，2019年）上，VEQ 相较于 Kimi-VL-Instruct 上的原始量化方法，性能提升了 **21.4%**。这些结果验证了，在低比特场景下，保护关键专家并区分不同模态的敏感度至关重要。

4.3. 消融研究

为验证所提出方法的有效性和稳健性，我们进行了两方面的消融研究。首先，我们评估了各组件对下游任务整体性能的贡献。其次，我们在随机抽取的验证集上进行超参数敏感性分析，以论证参数选择的合理性。此外，我们在相同的量化配置下报告所有消融结果，以确保比较公平。为隔离各组件的影响，我们还在不同设置下固定校准数据和评估协议。

组件对下游任务的影响。 我们重点关注两种核心公式：VEQ-ME 中的专家重要性分数，以及 VEQ-MA 中的模态亲和度感知海森矩阵。我们将超参数设为默认最优值，并测量禁用特定组件后性能的下降幅度。



(a) VEQ-ME 敏感性。 *(b)*VEQ-MA 敏感度。

图7。 验证集 PPL 的参数敏感度可视化分析。(a) VEQ-ME：结果证实了我们方法的尺度不变性，即保持相对比例可确保持续降低量化误差。(b) VEQ-MA：结果表明，降低 λ 通常会导致 PPL 上升，验证了路由器置信度对于实现准确量化至关重要。

表2。 模态-专家重要性（VEQ-ME）在下游任务上的消融研究。 γ：梯度因子； β：数量因子。

配置	MMMU]			Avg.
不含 γ （$\gamma = 1$）	43.44	60.39	82.24	62.02
w/o β ($\beta = 1$)	43.89	60.34	82.67	62.30
VEQ-ME（完整）	**44.56**	**62.33**	**82.93**	**63.27**

1) 模态-专家重要性（*VEQ-ME*）的影响。重要性分数的公式为 $S_i = \gamma N_i^{text} + \beta N_i^{vision}$。我们研究梯度缩放因子 γ和数量归一化因子 β的必要性：

- **w/o γ ($\gamma = 1$):** 移除梯度缩放后，视觉和文本的重要性完全根据令牌数量来衡量。如表 2所示，这会导致文本密集型推理任务的性能明显下降，证实文本词元需要更高的敏感度权重。

- **w/o β ($\beta = 1$):** 忽略数量差异会使数量庞大的视觉令牌主导评分。结果表明，这种变化会降低性能，因为路由器会偏向空间冗余视觉特征。

2）模态亲和性感知的影响（*VEQ-MA*）。对于亲和性感知量化，令牌权重定义为$c_j = p_j \cdot \alpha_j$。我们考察亲和度 p_j 和模态指示器 α_j的作用：

- **w/o p ($p = 1$):** 忽略路由器亲和度逻辑值，会将路由至同一专家的所有令牌视为同等重要。表 3 显示，这会导致次优结果，证明路由置信度更高的令牌更具代表性。

- **无 α （$\alpha = 1$）：** 在海森校准过程中移除模态特定的重新加权会导致性能下降，证实区分信息密集型文本令牌与冗余视觉令牌对于准确的误差最小化至关重要。

表3. 亲和性感知海森（VEQ-MA）在下游任务上的消融研究。 p：路由器置信度； α：模态指示器。

配置	MMMU	InfoVQA	ScienceQA	Avg.
不含 p ($p = 1$)	42.33	64.30	83.35	63.63
不含 α ($\alpha = 1$)	41.33	63.72	82.34	62.46
VEQ-MA（完整配置）	**42.56**	**64.48**	**84.85**	**63.96**

参数敏感性分析。为进一步验证我们方法的稳健性，我们在不同参数配置下分析了64个样本的平均困惑度（PPL）。这些样本随机抽取自MMMU（Yue等人，2024年）验证数据集。我们对VEQ-ME（γ， β）和VEQ-MA（λ， γ）中的变量对进行网格搜索。值得注意的是，尽管VEQ-MA的公式在理论上由亲和度 p 和敏感度比 α决定，直接调节 p 和 α 缺乏直观的可解释性。因此，我们选择采用（λ， γ）作为本次消融研究的变量对。

1) VEQ-ME（γ 与 β）的敏感性。在我们的公式中， γ 表示文本词元相对于视觉词元的敏感性，而 β 表示两个模态之间的数量比。这些超参数决定了专家重要性，并指导对最优量化参数的搜索。如图7所示，我们观察到，按比例缩放 γ 和 β 会产生一致的 PPL 值。这表明，量化参数的搜索取决于专家重要性的相对比值。

2) VEQ-MA（λ 与 γ）的敏感性。我们通过改变模态敏感度比 α 和亲和强度 p来分析海森加权。具体而言，我们使用调制系数p 来处理 λ的变化，该系数控制路由器置信度的强度。我们将有效亲和度定义为加权插值：$(1-\lambda) \cdot \mathbf{1} + \lambda \cdot \mathbf{p}$。在此公式中， $\lambda = 1$ 表示原始路由器置信度，而 $\lambda = 0$表示均匀亲和度。如图7所示，模型在完整配置（$\lambda = 1$， $\gamma = \gamma_0$）下实现了最佳稳定性（最低 PPL 为 2.2526）。

5. 结论

在本研究中，我们提出了**视**觉**专**家**量**化（**VEQ**），这是一种专门的训练后量化框架，旨在应对压缩混合专家视觉语言模型（混合专家视觉语言模型）所面临的独特挑战。VEQ突破了将混合专家前馈神经网络视为单一稠密结构的局限，有效应对了专家激活固有的稀疏性，以及视觉和文本模态之间的统计异质性。在一系列多样化的多模态基准测试中，VEQ始终优于既有基线方法。通过使量化策略与混合专家视觉语言模型的结构和模态属性相契合，VEQ确立了新的最先进水平，为在资源受限环境中高效部署大规模多模态智能体铺平了道路。

参考文献

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., 等。Flamingo：一种用于少样本学习的视觉语言模型。载于*NeurIPS*，2022年。

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C.,Ge,C.,Ge,W.,Guo, Z., 等。Qwen3-vl 技术报告。*arXiv* 预印本 *arXiv:2511.21631*，2025年。

Baidu-ERNIE-Team。文心4.5 技术报告，2025年。

Chi, Z., Dong, L., Huang, S., Dai, D., Ma, S., Patra, B., Singhal, S., Bajaj, P., Song, X., Mao, X.-L., 等。论稀疏混合专家的表示坍缩。载于*NeurIPS*，2022年。

Duanmu, H., Li, X., Yuan, Z., Zheng, S., Duan, J., Zhang, X., and Lin, D. Mxmoe：面向混合专家的混合精度量化，通过准确率与性能协同设计实现优化。*arXiv* 预印本 *arXiv:2505.05799*，2025年。

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers：使用简单高效的稀疏性扩展至万亿参数模型。*JMLR*，2022年。

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. GPTQ：面向生成式预训练 Transformer 的精准训练后量化。*arXiv* 预印本 *arXiv:2210.17323*，2022年。

Gurari, D., Li, Q., Lin, C., Zhao, Y., Guo, A., Stangl, A., and Bigham, J. P. VizWiz-Priv：一个用于识别盲人拍摄的图像中是否存在私有视觉信息及其用途的数据集。载于*CVPR*，2019年。

Hu, X., Chen, Z., Yang, D., Xu, Z., Xu, C., Yuan, Z., Zhou, S., and Yu, J. MoEQuant：通过专家均衡采样和亲和引导增强混合专家大型语言模型的量化。*arXiv* 预印本 *arXiv:2505.03804*，2025年。

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., 等。GPT-4o 系统卡。*arXiv* 预印本 *arXiv:2410.21276*，2024年。

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., 和 Farhadi, A. 一张图表胜过十几张图像。载于欧洲计算机视觉国际会议，2016年。

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2：通过冻结的图像编码器和大型语言模型引导语言-图像预训练。发表于*ICML*，2023年。

Li, S., Hu, Y., Ning, X., Liu, X., Hong, K., Jia, X., Li, X., Yan, Y., Ran, P., Dai, G., et al. MBQ：面向大型视觉语言模型的模态均衡量化。发表于*CVPR*，2025年。Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. 注意差距：理解多模态对比表示学习中的模态差距。发表于*NeurIPS*，2022年。Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. AWQ：面向端侧大语言模型压缩与加速的激活感知权重量化。发表于机器学习与系统会议，2024年。

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., 和 Zitnick, C. L. 微软COCO：上下文中的常见物体。载于 欧洲计算机视觉国际会议，2014年。

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., 等。MMBench：你的多模态模型是全能型选手吗？载于 欧洲计算机视觉国际会议，2024a年。

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. SpinQuant：使用学习到的旋转进行大语言模型量化。*arXiv* 预印本 *arXiv:2405.16406*，2024b年。

Lu, J., Batra, D., Parikh, D., and Lee, S. ViLBERT：为视觉与语言任务预训练与任务无关的视觉语言表征。发表于*NeurIPS*，2019年。

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. 学会解释：通过思维链进行多模态推理，以回答科学问题。发表于*NeurIPS*，2022年。

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., 和 Jawahar, C. 信息图表视觉问答。载于*WACV*，2022年。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等。通过自然语言监督学习可迁移视觉模型。载于*ICML*，2021年。

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., 和 Rohrbach, M. 面向能够阅读的视觉问答模型。载于*CVPR*，2019年。

Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., 等。Kimi-VL 技术报告。*arXiv* 预印本 *arXiv:2504.07491*，2025年。

Team, x. Realworldqa：真实世界多模态问答基准，2024年。网址 https://huggingface.co/datasets/xai-org/RealworldQA。

Wang, C.、Wang, Z.、Xu, X.、Tang, Y.、Zhou, J. 和 Lu, J. Q-vlm：面向大型视觉语言模型的训练后量化。发表于 *NeurIPS*，2024年。

Wang, X., Huang, J., Abdalla, R., Zhang, C., Xian, R., and Manocha, D. Bi-vlm：推动视觉语言模型超低精度训练后量化的边界。*arXiv* 预印本 *arXiv:2509.18763*，2025年。

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2：用于高级多模态理解的混合专家视觉语言模型。*arXiv* 预印本 *arXiv:2412.10302*，2024年。

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant：面向大型语言模型的精准高效训练后量化。发表于*ICML*，2024年。Xue, Y., Huang, Y., Shao, J., and Zhang, J. Vlmq：通过海森矩阵增强实现大型视觉语言模型的高效训练后量化。*arXiv* 预印本 *arXiv:2508.03351*，2025年。Yu, J., Zhou, S., Yang, D., Wang, S., Li, S., Hu, X., Xu, C., Xu, Z., Shu, C., and Yuan, Z. Mquant：通过全静态量化释放多模态大型语言模型的推理潜力。*arXiv* 预印本 *arXiv:2502.00425*，2025年。Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu：面向专家级通用人工智能的大规模跨学科多模态理解与推理基准测试。发表于*CVPR*，2024年。Zhang, J., Zhang, Y., Zhang, B., Liu, Z., and Cheng, D. Moqe：通过混合量化专家提升量化模型性能。*arXiv* 预印本 *arXiv:2508.09204*，2025年。Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J. A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., and Liu, Z. Lmms-eval：大型多模态模型评估的现实检验。*arXiv* 预印本 *arXiv:2407.12772*，2024a年。Zhang, Y.-F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al. Mme-realworld：多模态大语言模型能否挑战对人类而言困难的高分辨率真实场景？*arXiv* 预印本 *arXiv:2408.13257*，2024b年。Zheng, L., Yin, L., Xie, Z., Sun, C. L., Huang, J., Yu, C. H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J. E., et al. Sglang：高效执行结构化语言模型程序。发表于 *NeurIPS*，2024年。

Zheng, Z.、Cui, X.、Zheng, S.、Li, M.、Chen, J.、Liang, Y. 和 Chen, X. MoQa：重新思考混合专家量化，结合多阶段数据-模型分布感知。*arXiv* 预印本 *arXiv:2503.21135*，2025年。