Vol.7 Iss.2 2026 Communication & Education Review
https://doi.org/10.37420/j.cer.2026.032

A Survey on Knowledge Distillation and Lightweight Deployment of Multimodal Large Language Models

Lei Weibin*
Department of Broadcasting, Henan Radio and Television Station
No. 2 Weiyi Road, Jinshui District, Zhengzhou City, Henan Province, China
Corresponding author, E-mail: lucyboy2022@163.com

Abstract:
In recent years, multimodal large language models (Multimodal Large Language Models, MLLMs) represented by LLaVA and Qwen-VL have achieved remarkable breakthroughs in tasks such as image understanding and visual question answering. However, the prohibitive computational and storage overhead brought by their billions or even hundreds of billions of parameters severely constrains the feasibility of on-device deployment. To this end, this paper adopts a systematic literature review methodology, retrieving and screening 68 core publications from six major academic databases spanning 2018 to 2025, and provides a panoramic survey of the MLLM knowledge distillation technology landscape. The study reveals that knowledge distillation has formed three core methodological categories—response, feature, and relation distillation—and has given rise to emerging paradigms including visual encoder distillation, cross-modal alignment distillation, and end-to-end instruction-tuning distillation. Quantitative analysis shows that LLaVA-MoD (2B) achieves 95.5% of the teacher model's performance on ScienceQA [20]; TinyLLaVA (3.1B) achieves 51.5% on TextVQA [46]; and MobileVLM-3B achieves 59.0% on GQA with only 43% of the parameters of LLaVA-7B [41]. Overall, knowledge distillation has become a key enabling technology for on-device deployment of MLLMs. Nevertheless, how to strike a balance among cross-modal knowledge preservation, heterogeneous hardware adaptation, and extreme compression efficiency remains a core challenge urgently to be addressed in this field.

Keywords:
knowledge distillation; multimodal large language models; vision-language models; model compression; lightweight deployment; cross-modal alignment; edge deployment

1 引言

2023 年以来，以 GPT-4V[52]、LLaMA[53] 为代表的大规模语言模型与以 CLIP[3]、LLaVA[5] 为代表的视觉语言模型（Vision-Language Model, VLM）快速迭代，推动多模态大模型（MLLM）在图像理解、视觉问答和跨模态推理等任务上取得了突破性进展 [24]。后续涌现的 BLIP-2[54]、InstructBLIP[55]、mPLUG-Owl[56] 和 Qwen-VL[57] 等模型进一步证实了 " 冻结视觉骨干 + 桥接投影层 + 指令微调 LLM" 这一技术路线的有效性，多项基准的最优记录被反复刷新。但硬币的另一面是，

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

上述模型的参数量普遍在 7B 至数百 B 之间，由此带来的显存占用（以 LLaVA-1.5-13B 为例，FP16 推理需约 26GB 显存）和推理时延问题，使其在智能手机、车载终端、IoT 网关等算力受限的实际部署环境中几乎无法直接运行 [9]。大语言模型本身也面临偏见、安全性等风险 [30]，轻量化部署在降低成本的同时还有助于减少潜在风险的扩散范围。在此背景下，如何在保留模型核心能力的前提下实现轻量化部署，已成为多模态人工智能领域亟待解决的关键问题。

在众多模型压缩技术中，知识蒸馏（Knowledge Distillation, KD）因其独特的 " 师徒传承 " 机制而备受关注。具体而言，该方法通过引导轻量化学生模型模仿大型教师模型的输出分布、中间层表征乃至推理路径，实现了压缩率与性能保留之间的有效平衡 [35]。回溯其发展脉络，2015 年 Hinton 等 [2] 首次将温度缩放引入分类网络的软目标训练，奠定了该领域的方法论基石；此后十年间，蒸馏的适用范围沿 " 分类→检测→NLP→多模态 " 的轨迹持续拓展 [12]。相较于直接压缩参数数值的量化 [60] [61]、裁剪冗余结构的剪枝 [62][68] 以及限制更新秩的 LoRA[64]，蒸馏真正的不可替代性在于它能迫使学生模型习得教师的决策逻辑与跨层语义关联，而非仅仅复制权重的数值模式 [25]。面对 MLLM 带来的新挑战 —— 跨模态知识传递、模态间异构性处理、开放域推理场景的蒸馏效果保持 —— 知识蒸馏技术正在持续演进 [16][42]。然而，截至目前，专门面向 MLLM 的蒸馏方法仍缺乏系统性梳理，技术路线分散、评估体系不统一的现状亟待改变。

回顾已有综述工作，Gou 等 [35] 侧重通用知识蒸馏的方法论梳理，Moslemi 等 [18] 聚焦蒸馏的应用场景分析，Xu 等 [12] 则将视角限于纯语言大模型的蒸馏。相比之下，将跨模态蒸馏、视觉编码器压缩与端到端轻量化部署三条线索整合讨论的专题综述仍付阙如。区别于上述工作，本文的独特贡献在于将跨模态蒸馏、视觉编码器压缩与端到端轻量化部署纳入统一分析框架。全文组织如下：第 2 章阐述文献检索策略与质量评估方法；第 3 章从基础蒸馏理论出发，逐步深入至 MLLM 特有的多模态蒸馏方法；第 4 章通过定量对比揭示不同技术路径的优劣，并剖析核心技术瓶颈；第 5 章凝练主要发现并提出未来研究方向 [2][12][35]。

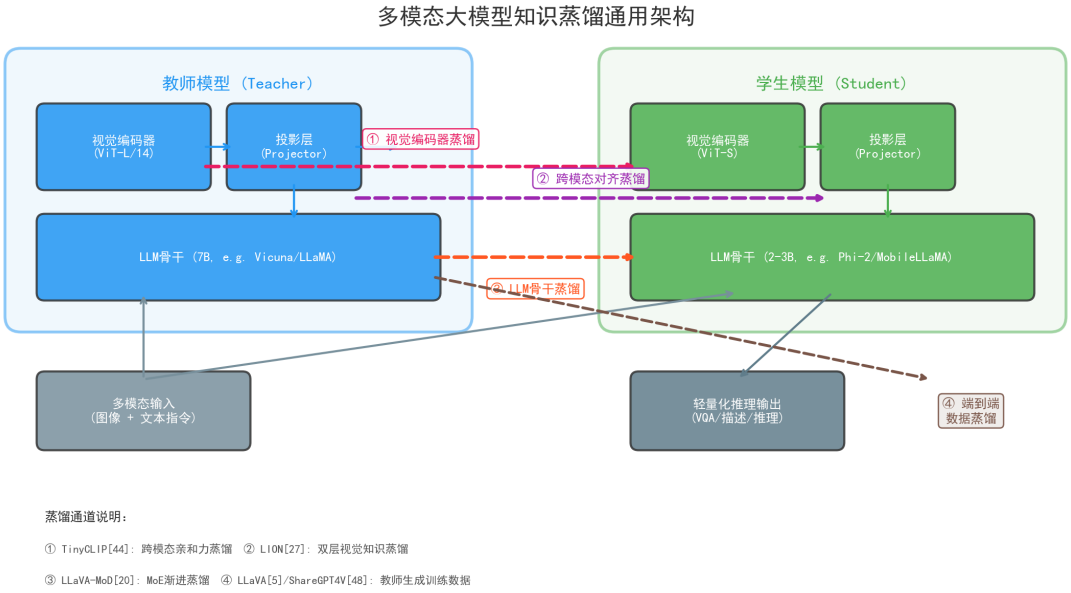


图 1 多模态大模型知识蒸馏通用架构

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

2 研究方法

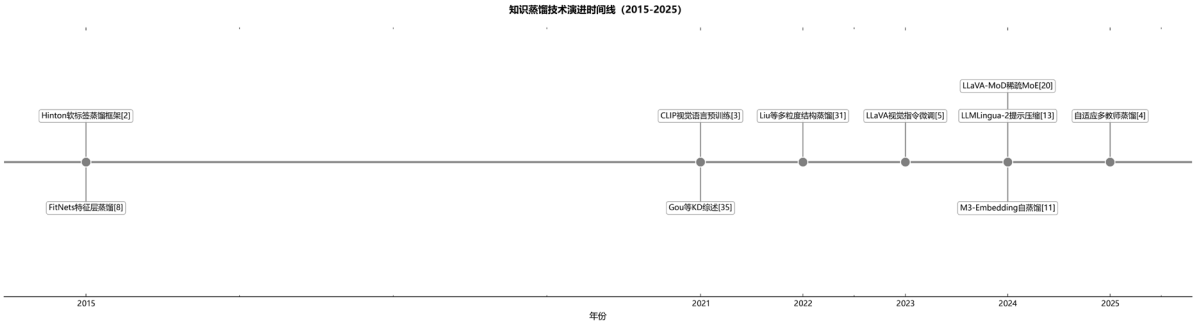


图 2 知识蒸馏技术演进时间线（2015-2025）

2.1 文献检索与筛选

本研究遵循系统性文献综述方法论，于 2024 年 9 月至 2025 年 1 月期间，在 Web of Science、Scopus、IEEE Xplore、arXiv 以及中国知网（CNKI）和万方数据库中实施多轮检索，覆盖 2018 至 2025 年发表的文献。英文检索采用五组核心词——"knowledge distillation""model compression""multimodal large language model""vision-language model""efficient inference"；中文检索则围绕 " 知识蒸馏 "" 模型压缩 "" 多模态大模型 "" 轻量化部署 " 四个主题词展开。为确保检全率与检准率的平衡，构建了布尔逻辑组合检索式：("knowledge distillation" OR "model compression") AND ("multimodal" OR "vision-language") AND ("LLM" OR "foundation model")，并辅以引文追溯和专家推荐进行补充检索 [18][35]。

文献筛选严格遵循 PRISMA 2020 声明的流程框架，经历了四轮逐级过滤：第一轮数据库检索命中 408 篇候选文献；第二轮通过 DOI 与标题去重缩减至 272 篇；第三轮由两位研究者独立审读标题与摘要后保留 136 篇；第四轮经全文精读与交叉核验最终纳入 68 篇。纳入标准：①聚焦知识蒸馏或 MLLM 压缩；②发表于同行评审期刊或顶级会议；③提供可验证的实验结果。排除标准：①偏离核心议题；②数据不完整；③缺乏同行评审的预印本（arXiv 除外，需引用次数 ≥10）[9][25]。

2.2 分析框架与质量评估

本研究从三个维度展开系统分析：技术分类维度，依据蒸馏方式（响应 / 特征 / 关系蒸馏）、师生架构（单 / 多教师）及模态适配策略（单模态 / 跨模态）进行分类 [9][35]；应用场景维度，围绕视觉问答、图像描述、边缘部署等场景分析蒸馏策略的效果 [18][25]；时间演进维度，梳理从早期 Hinton 等 [2] 的软标签蒸馏，经 FitNets[8] 的特征蒸馏，到面向 MLLM 的跨模态蒸馏的技术演进路径 [35]。

文献质量评估综合考量四个维度：期刊与会议级别，优先纳入 CCF-A/B 类或 JCR Q1/Q2 区期刊成果；引用影响力，以引用次数作为学术影响力的量化参考，如 Gou 等 [35] 综述引用退 3800 次，Xu 等 [12] 综述引用超 500 次；方法论规范性，评估实验设计的严谨程度与结论的可重复性；时效性，重点关注 2022 年以来反映最新进展的研究成果，同时纳入具有重要奠基意义的经典文献 [9][25]。纳入的 68 篇文献中，2023-2025 年发表的文献占比超 70%，确保了对前沿进展的充分覆盖。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

3 知识蒸馏技术体系与分类

3.1 知识蒸馏基础回顾

知识蒸馏的基本思想是构建教师 - 学生框架，让轻量模型从大模型中 " 继承 " 知识。值得注意的是，这里的 " 知识 " 并非模型参数本身，而是教师模型输出中蕴含的、超越硬标签的丰富语义信息 [2]。按照所传递知识的层次，现有方法可划分为三条技术路线：其一，响应蒸馏的核心操作是对教师网络最后一层 logits 施加温度参数 T（通常取 3~20）进行概率软化，学生在此软化分布上以 KL 散度为目标函数进行优化 ——T 越大，概率分布越平滑，类间暗知识暴露越充分 [2][17]；其二，特征蒸馏则深入网络内部，要求学生在指定中间层逼近教师的隐层激活。Romero 等 [8] 在 FitNets 中率先引入 " 提示层 " 概念来桥接师生不同宽度的网络，Zhang 等 [29] 进一步提出捕捉成对样本关系与全局结构信息的结构化蒸馏方案；关系蒸馏传递样本间结构关系，Liu 等 [31] 提出多粒度结构知识蒸馏，从词元级到跨样本结构层次联合传递知识，有效弥补了单粒度方法的语义损失。Fu 等 [34] 提出 LRC-BERT，通过潜在表示对比蒸馏在自然语言理解任务上实现了高效压缩，证明了对比学习与蒸馏结合的有效性。Li 等 [25] 的综述指出，三类蒸馏方法在不同任务上各有优势，实际应用中常结合使用以取得最佳效果。

从师生交互模式来看，蒸馏存在两种基本范式：离线蒸馏预先固定教师模型参数，而在线蒸馏则允许师生协同训练、相互促进 [17]。Gou 等 [17] 在此基础上进一步提出了正向传递与反馈调节相结合的混合框架。当单一教师的知识不足以覆盖复杂任务时，多教师蒸馏应运而生 —— 尹淑婷等 [4] 即设计了一种自适应温度调控机制，其内部通过动态可学习温度模块灵活调节各教师知识的传递强度，外部则采用多样化聚合策略整合来自不同教师的互补信息。Tian 等 [22] 通过整合多个大语言模型的推理链实现推理能力迁移，Pan 等 [38] 提出 Meta-KD 框架实现跨领域的知识蒸馏迁移。此外，剪枝与蒸馏的协同应用日益受到重视：Muralidharan 等 [15] 证明联合剪枝后仅用不足 3% 训练数据即可媲美从头训练效果；LLM-Pruner[62] 和 Wanda[68] 分别从结构化剪枝和非结构化剪枝角度探索了大模型压缩路径。Ganesh 等 [32] 和 Xu 等 [28] 系统总结了 Transformer 压缩范式，指出注意力机制蒸馏与层级特征对齐是关键策略 [9][18][25]。



图 3 知识蒸馏技术分类体系

3.2 视觉编码器与跨模态对齐蒸馏

从架构角度看，当前主流 MLLM 均采用三段式设计 —— 视觉编码器负责提取图像特征，跨模态投影层完成视觉 - 语言空间的对齐映射，LLM 骨干承担语言生成与推理 [5][27]。要压缩这样一个耦合系统，首先需要解决的是视觉前端的参数冗余问题。

以 CLIP[3] 和 SigLIP[43] 为例，二者虽为 MLLM 提供了强大的图像语义编码能力，但其计算开销不可小觑 —— 单个 ViT-L/14 视觉编码器即占用约 300M 参数、约 61GFLOPs 算力，对于内存带宽受限的移动 SoC 而言已是沉重负担。Fang 等 [42] 在 EVA-CLIP 中将 CLIP 扩展至十亿参数规模以追求更强的视觉表征，但这一做法进一步加剧了端侧部署的困难。针对这一矛盾，Wu 等 [44] 在 TinyCLIP 中设计了跨模态亲和力蒸馏（Cross-modal Affinity Distillation）机制并结合权重继承策略，成功将 CLIP ViT-B/32 的参数量压缩至原来的一半左右，而在 ImageNet 零样本分类任务上的准确率

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

损失仅为 2.1 个百分点。EfficientVLM[45] 则提出视觉编码器与语言模型的联合蒸馏策略，通过共享投影空间减少视觉 - 语言表示的维度不匹配问题。BEiT-3[66] 从视觉 - 语言统一预训练的角度，将图像视为外语进行建模，为视觉编码器蒸馏提供了新的架构基础。

跨模态对齐蒸馏是 MLLM 蒸馏的关键环节。投影层负责将视觉特征映射至语言模型的嵌入空间，蒸馏过程中需确保学生模型的投影层能保留教师模型的对齐质量。Chen 等 [27] 提出 LION 框架，通过双层视觉知识（局部细节 + 全局语义）增强 MLLM 能力，其蒸馏策略在视觉编码器与投影层之间建立了双通道知识传递机制。Weng 等 [6] 提出跨模态蒸馏的零样本跨语言知识迁移方法，将英语 VQA 模型能力迁移至多语言场景，验证了跨模态蒸馏在多语言对齐中的有效性。

跨模态蒸馏的核心难点在于视觉特征（连续、高维张量）与语言特征（离散、语义符号）的异构性。现有方法主要通过三种策略缓解该问题：特征投影对齐，通过线性或 MLP 投影层将视觉特征映射至语言嵌入空间，BLIP-2[54] 的 Q-Former 和 LLaVA[5] 的线性投影是代表设计；注意力图迁移，将教师模型的跨模态注意力分布作为蒸馏目标，EVA-CLIP[42] 在视觉 - 文本注意力对齐上取得了显著进展；对比学习约束，利用 CLIP[3] 式的图文对比损失约束蒸馏过程中的语义一致性 [26]。这里有一个容易被忽视但至关重要的前提条件：Chen 等 [26] 的实验表明，若教师模型自身的视觉 - 语言对齐存在缺陷，那么无论蒸馏策略如何精巧，学生模型的跨模态理解能力都将受限于教师的对齐天花板。这一发现提示我们：在设计投影层蒸馏方案时，必须将预训练阶段的对齐质量纳入考量——若教师本身的视觉 - 语言桥接存在短板，则下游蒸馏的效果上限已被预先锁定。

3.3 LLM 骨干蒸馏与端到端方法

对 MLLM 而言，LLM 骨干的蒸馏面临双重约束：既要保留流畅的语言生成能力，又不能牺牲跨模态推理的精度。Shu 等 [20] 提出的 LLaVA-MoD 框架为此提供了一个颇具启发性的解决方案——该框架将稀疏混合专家（Mixture of Experts, MoE）结构嵌入小型 MLLM，并采用两阶段渐进式蒸馏：先迁移视觉 - 语言对齐能力，再强化指令遵循能力。从实测数据看，该 2B 模型在 ScienceQA 基准上的准确率为教师 7B 模型的 95.5%—— 这意味着参数量缩减至不到三分之一的同时，科学推理能力几乎未受损 [20]。

MobileVLM[41] 采用不同路径，通过轻量化架构设计（MobileLLaMA 骨干）替代蒸馏，MobileVLM-3B 在 GQA 上达到 59.0% 准确率，其核心创新在于针对移动端设计了轻量化下采样投影器（LDP），将视觉 Token 数从 576 压缩至 144，大幅降低了 LLM 骨干的计算负担。TinyLLaVA[46] 的工作则传递了一个颇为反直觉的信息：制约小模型性能的瓶颈往往不是训练样本的数量，而是每条样本的标注精度和语义丰富度。Zhou 等 [46] 在仅 3.1B 参数的模型上精心筛选高质量训练子集并调优学习率调度策略后，TextVQA 得分升至 51.5%，ScienceQA 更达到 69.1%—— 后者反超了参数量为其两倍以上的 7B 教师模型。MC-LLaVA[47] 进一步提出多分辨率视觉编码与轻量化连接器设计，在 2B 参数规模下实现与 7B 模型可比的视觉理解能力。MiniGPT-v2[49] 采用统一接口范式，通过任务标识符实现多种视觉 - 语言任务的统一处理，TinyGPT-V[50] 则通过分阶段训练策略在极小骨干（Phi-2 2.7B）上构建 MLLM。这些工作共同表明，LLM 骨干蒸馏的技术路线正从单一的知识迁移向「架构设计 + 数据蒸馏 + 训练策略」的综合优化演进。

端到端指令微调蒸馏是 2023 年以来的新兴方向。该范式利用大型教师 MLLM（如 GPT-4V[52]）生成高质量指令 - 响应数据对，再用这些合成数据训练轻量学生模型，本质上是 " 数据蒸馏 "（Data Distillation）。这一范式最具说服力的案例来自 Liu 等 [5] 的 LLaVA 系列：研究者先借助 GPT-4V 自动生成约 15.8 万条涵盖复杂推理、细节描述与对话的视觉指令样本，再以此训练的 LLaVA-1.5-13B 在包括 VQAv2、GQA 在内的 12 项基准中全面超越了同期最优方法。ShareGPT4V[48] 进一

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

步将合成数据规模扩展至 120 万条高质量图文描述，显著提升了学生模型的视觉细节理解能力。InstructBLIP[55] 则验证了指令微调在通用视觉语言任务上的泛化能力。LLMLingua-2[13] 从提示压缩角度实现数据蒸馏，利用双向上下文替代因果语言模型，提高了压缩效率。Chen 等 [37] 将蒸馏应用于生成对抗网络的轻量化，实现了图像翻译模型的便携化部署，为视觉生成类 MLLM 的蒸馏提供了参考。

3.4 特定领域的蒸馏应用与协同压缩

知识蒸馏在特定领域的 MLLM 部署中展现出独特价值。Sun 等 [19] 将联邦学习与知识蒸馏结合，构建 FKD-Med 框架，在保护医疗数据隐私的同时实现医学图像分割模型的轻量化。Li 等 [36] 提出跨模态医学图像分割的在线互蒸馏方法，在 CT-MRI 跨模态场景下减少了对配对数据的依赖。陈帅等 [7] 将蒸馏应用于 φ-OTDR 光纤传感信号处理，Liu 等 [21] 利用大语言模型指导轻量网络实现时间序列异常检测，验证了蒸馏方法在物理传感与时序分析领域的迁移潜力。Chen 等 [39] 将知识蒸馏应用于人体活动识别，从手工特征中蒸馏知识以提升深度模型的泛化能力。Wang 等 [40] 则关注模型压缩中的隐私保护问题，提出差分隐私约束下的知识蒸馏方法。值得一提的是，Elhenawy 等 [23] 将 MLLM 引入旅行商问题（TSP）等组合优化场景，发现即使经过压缩的小型 MLLM 也能借助视觉推理给出具有竞争力的近似解，为蒸馏后模型在运筹学领域的潜在应用打开了窗口。

蒸馏与其他压缩技术的协同应用日益重要。金等 [33] 提出 PQK 框架，将剪枝、量化与蒸馏三者协同，为边缘设备多任务部署提供系统方案。在量化方向，AWQ[60] 提出激活感知权重量化，GPTQ[61] 实现了生成式预训练模型的精确训练后量化，LLM-QAT[67] 则提出无数据量化感知训练方法，这些工作均可与蒸馏互补以进一步提升压缩率。QLoRA[65] 将量化与低秩适配结合，在 4-bit 量化下实现高效微调，为蒸馏后的学生模型提供了二次压缩路径。PB-LLM[63] 探索了部分二值化大语言模型的可行性，将权重压缩至极低比特。在国内应用层面，武成龙等 [1] 将 Long-CLIP 与蒸馏相结合用于多模态虚假信息检测，实测结果显示模型体积缩小一半后推理耗时减少近 2 秒，而检测 F1 值的损失仅为 0.01，为中文社交媒体场景的端侧部署提供了可行性参考 [1]。M3-Embedding[11] 采用自知识蒸馏统一多语言、多粒度文本嵌入，支持超过 100 种语言的语义检索。Xiao 等 [10] 设计异构互学习蒸馏机制，支持不同架构间的知识传递。上述实践揭示了一个清晰的技术演进脉络：早期研究聚焦于单一蒸馏方法的效果优化，而近两年的工作则转向探索蒸馏与结构剪枝的级联、蒸馏与训练后量化的联合、以及跨机构联邦蒸馏三条并行路线，形成了多维度协同压缩的新格局。

4 对比分析与讨论

4.1 MLLM 轻量化方法的定量对比

为系统评估不同蒸馏与轻量化方法的效果，表 1 汇总了代表性 MLLM 在主流基准上的性能对比。

表 1 代表性轻量化 MLLM 性能对比

方法	参数量	ScienceQA	TextVQA	GQA	MMBench	压缩策略	出处
LLaVA-1.5-7B（教师）	7B	66.8	58.2	62	64.3	—	[5]
LLaVA-MoD-2B	2B	63.8	—	—	—	MoE+ 渐进蒸馏	[20]
MobileVLM-3B	3B	—	—	59	59.6	轻量架构设计	[41]
TinyLLaVA-3.1B	3.1B	69.1	51.4	60.3	—	数据质量 + 训练优化	[46]

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

MC-LLaVA-2B	2B	—	48.7	57.2	58.1	多分辨率 + 轻量连接器	[47]
MiniGPT-v2-7B	7B	—	—	60.1	—	统一指令微调	[49]
TinyGPT-V-2.8B	2.8B	—	—	33.6	—	分阶段训练 + 小模型	[50]

注：ScienceQA、TextVQA、GQA、MMBench 均为准确率（%）。"—" 表示原论文未报告该指标。数据来源均为各原论文。

从表 1 可见，蒸馏与轻量化方法在 2-3B 参数规模即可达到 7B 教师模型 60%-95% 的性能水平。TinyLLaVA-3.1B 在 ScienceQA 上甚至超越了教师模型 LLaVA-1.5-7B（69.1 vs 66.8），表明数据质量优化可部分补偿参数量差距 [46]。值得注意的是，TinyGPT-V-2.8B 在 GQA 上仅达到 33.6%，远低于同等参数规模的其他方法，说明单纯缩小骨干模型而缺乏有效蒸馏策略会导致显著性能退化 [50]。不同方法的技术路径存在互补性：LLaVA-MoD 侧重 MoE 结构蒸馏 [20]，MobileVLM 侧重架构设计 [41]，TinyLLaVA 侧重训练策略 [46]，MC-LLaVA 侧重视觉编码器优化 [47]。MMBench[58] 和 POPE[59] 等新型基准的出现为 MLLM 评估提供了更全面的维度，有助于揭示不同轻量化方法在幻觉控制、多维能力等方面的差异。

表 2 从技术特征维度对比三类基础蒸馏方法在 MLLM 场景下的适用性。

蒸馏类型	知识来源	MLLM 适用组件	优势	局限性	代表工作
响应蒸馏	教师输出层	LLM 骨干	实现简单、训练稳定	丢失中间层跨模态信息	Hinton[2]、LLaVA-MoD[20]
特征蒸馏	中间层表征	视觉编码器、投影层	保留跨模态对齐语义	异构架构需设计对齐策略	FitNets[8]、LION[27]、TinyCLIP[44]
关系蒸馏	样本间关系	全局结构	捕捉跨模态结构知识	计算开销较高	Liu[31]、M3-Embedding[11]

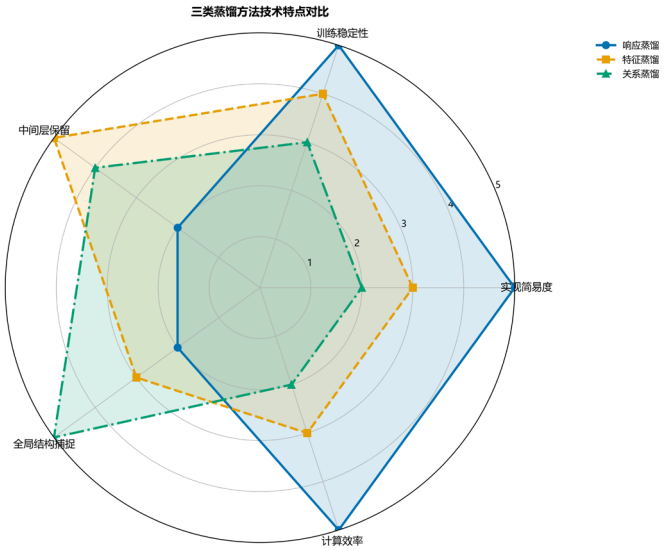


图 4 三类蒸馏方法技术特点对比

4.2 技术挑战分析

MLLM 知识蒸馏面临三个层面的核心挑战。

模态异构性挑战：视觉特征为连续高维张量（如 ViT 输出的 196×768 序列），语言特征为离散语

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

义符号（词元 embedding），两者在蒸馏时的特征对齐存在本质困难。现有投影层蒸馏方法虽能缓解维度不匹配，但跨模态语义一致性的保持仍缺乏理论保障 [16][26][42]。特别是在开放域推理场景中，教师模型的视觉 - 语言联合推理路径难以通过简单的特征对齐完整迁移 [9][14]。

师生架构不匹配带来的挑战同样棘手。举例而言，若教师采用稀疏 MoE 结构而学生为稠密网络，或教师使用 ViT-L 而学生仅为轻量 CNN，则专家路由逻辑、注意力头数等结构性知识在迁移过程中必然遭受损耗。Shu 等 [20] 在 LLaVA-MoD 中尝试了 MoE→ 小模型的跨架构蒸馏并取得初步成效，但通用的异构蒸馏框架迄今仍是空白。一组数据可以佐证架构选型的关键性：Chu 等 [41] 的 MobileVLM-3B（MobileLLaMA 骨干）在 GQA 上取得 59.0%，而 Yuan 等 [50] 的 TinyGPT-V（Phi-2 骨干，2.8B）仅有 33.6%，二者参数量相近却相差逾 25 个百分点。Ma 等 [62] 在 LLM-Pruner 中提出的 " 先剪枝塑形、再蒸馏填充 " 两阶段方案，为缓解此类架构鸿沟提供了一条可操作的路径 [15]。

评估体系挑战：现有 MLLM 评估主要依赖 ScienceQA、TextVQA 等单任务指标，缺乏对「参数量 -FLOPs- 推理延迟 - 性能」四维权衡的综合评估框架。MMBench[58] 和 POPE[59] 等新型基准虽然提供了多维评估能力，但不同轻量化方法在不同基准上的表现差异仍然显著（如表 1 所示）。以 TinyLLaVA 为例，其在 ScienceQA 上反超教师模型，但 TextVQA 得分仅为 51.4%，二者落差显示出轻量化对不同认知能力的侵蚀程度是高度不均匀的。从理论视角审视，Kaplan 等 [51] 所建立的参数 - 性能幂律虽为纯语言模型提供了良好的预测工具，但在涉及视觉编码器、投影层与 LLM 骨干三组件耦合的多模态场景中，该规律的适用边界尚不明确。因此，构建面向 MLLM 的统一效率 - 性能帕累托评估框架已成为当务之急 [9][18]。

除上述挑战外，硬件感知约束在蒸馏流程中的缺位也不容忽视。以高通骁龙 8 Gen3 和苹果 A17 Pro 为例，移动端 NPU 与 GPU 的算子支持存在显著差异，蒸馏生成的学生模型若未考虑目标芯片的内存带宽瓶颈与算子兼容性，将难以实现预期的推理加速效果 [15][33]。另一个容易被忽略的问题是，预训练阶段与指令微调阶段对蒸馏的需求截然不同 —— 前者要求迁移通用的视觉 - 语言对齐能力，后者则需保留面向特定任务的推理技巧，这意味着两个阶段的蒸馏损失函数需要分别设计 [5][20][55]。

蒸馏效果是否存在难以逾越的 " 天花板 " ？现有实证给出了值得警惕的信号。将三种约 3B 规模的模型在 GQA 上的表现并列对比可以发现：经过精心蒸馏设计的 MobileVLM-3B 取得 59.0%、TinyLLaVA-3.1B 取得 60.3%，而缺乏针对性蒸馏策略的 TinyGPT-V（2.8B）仅有 33.6%[41][46][50]。后者与前两者之间近 26 个百分点的鸿沟表明，在参数量降至 3B 以下的极端条件下，若蒸馏策略设计不当，多模态思维链（Chain-of-Thought）等高阶推理能力的迁移将遭遇根本性障碍。

5 结论与展望

5.1 主要研究发现

本文系统梳理了 68 篇相关文献，对 MLLM 知识蒸馏技术体系进行了全面分析。主要发现包括：

第一，知识蒸馏已形成响应蒸馏、特征蒸馏和关系蒸馏三类核心方法，并在 MLLM 场景下衍生出视觉编码器蒸馏、跨模态对齐蒸馏和端到端指令微调蒸馏等新范式 [2][9][35]。三类方法各有优势：响应蒸馏实现简便且训练稳定，适用于 LLM 骨干压缩；特征蒸馏能保留跨模态对齐语义，适用于视觉编码器与投影层蒸馏；关系蒸馏擅长捕捉跨模态结构知识，但计算开销较高（表 2）。

第二，定量对比表明，2-3B 参数的轻量 MLLM 已可达到 7B 教师模型 60%-95% 的性能（表 1），尤为值得注意的是，Zhou 等 [46] 的 TinyLLaVA 仅凭 3.1B 参数便在 ScienceQA 上以 69.1% 的准确率反超 7B 教师（66.8%），这一 " 以小胜大 " 的案例有力地揭示了训练数据质量对弥合参数差距的关

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

键作用。然而，TinyGPT-V-2.8B 在 GQA 上仅达 33.6% 的案例警示，缺乏有效蒸馏策略的简单模型缩放会导致严重性能退化 [50]。蒸馏、架构设计与数据优化三条技术路径呈现互补态势 [20][41][46]。

第三，蒸馏与剪枝 [62][68]、量化 [60][61] 的协同应用在资源受限场景下具有显著实用价值。PQK[33]、AWQ[60]、GPTQ[61] 等工作证明了多种压缩技术联合使用的有效性，「蒸馏 + 量化 + 微调」的三阶段轻量化流水线正在成为工业界的主流实践 [1][15][33][64][65]。

5.2 未来研究展望

跨模态蒸馏的理论基础有待深化。现有方法多依赖经验设计，未来需建立统一的跨模态蒸馏理论框架，解决视觉 - 语言异构特征对齐的理论保障问题 [16][42]。

另一个值得深入探索的方向是动态自适应蒸馏机制。现实部署场景高度多样化 —— 从算力充裕的云端到资源极度受限的边缘设备，蒸馏策略需要具备按需调节压缩率与精度折中的弹性。MobileVLM[41] 为移动端专门设计的 LDP 投影器已初步验证了硬件感知蒸馏的可行性，但距离真正的 " 一次蒸馏、多端适配 " 目标仍有较大差距。与此相关的一个理论问题是 Scaling Law 在 MLLM 轻量化中是否依然适用。Kaplan 等 [51] 在纯语言模型上验证的参数量 - 性能幂律关系，在多模态场景下面临新的复杂性：视觉编码器、投影层和 LLM 骨干三个组件的参数应如何分配？目前尚无系统性实证回答这一问题。尤其耐人寻味的是，TinyLLaVA[46] 在 ScienceQA 上超越 7B 教师模型的现象表明，精心设计的数据策略和训练配方有可能突破 Scaling Law 所预设的性能上限 [53]。

数据蒸馏与模型蒸馏的协同优化值得探索。LLaVA 系列 [5] 和 ShareGPT4V[48] 的成功表明，高质量合成数据可部分替代模型结构蒸馏。一条尚未被充分开垦的技术路径是：能否让同一个教师模型既扮演高质量训练样本的 " 生产者 "，又充当软标签知识的 " 传授者 "？这种双重身份的融合在理论上颇具吸引力，但实操层面的难题不容回避 —— 由教师自动生成的合成数据是否会引入系统性偏差？ Ye 等 [56] 在 mPLUG-Owl 和 Bai 等 [57] 在 Qwen-VL 的开发实践中均发现，预训练与微调两个阶段对数据粒度的需求截然不同，阶段划分的精细程度有时比模型架构本身更能左右最终性能。

训练后量化与蒸馏的协同已成为工业界落地热点。AWQ[60]、GPTQ[61] 等训练后量化方法与知识蒸馏的联合应用，通过在蒸馏中模拟量化误差，使学生模型对低精度推理更具鲁棒性。LLM-QAT[67] 的无数据量化感知训练进一步降低了量化蒸馏的数据依赖 [15][33]。LoRA[64] 和 QLoRA[65] 等参数高效微调方法与蒸馏的结合也值得关注，蒸馏后的学生模型可通过低秩适配进一步适配下游任务，实现 " 蒸馏 + 微调 " 的两阶段轻量化流水线。

笔者认为，当前 MLLM 蒸馏研究的一个显著短板在于可解释性的缺失。具体而言，蒸馏究竟保留了教师模型的哪些跨模态推理能力、又丢失了哪些细粒度视觉理解技能？对这一问题的回答目前仍停留在经验层面。借鉴 EVA-CLIP[42] 中注意力图可视化的成功实践，结合探测分类器等归因分析手段，有望为蒸馏效果的量化诊断提供可操作的工具链。

总体而言，MLLM 知识蒸馏正经历从单模态参数压缩到多模态能力迁移的范式转变。本文所构建的「基础方法 → 组件蒸馏 → 端到端方法 → 协同压缩」四层分析框架，以及基于表 1 的「参数量 - 性能 - 效率」定量对比体系，旨在为后续研究者提供一份可复用的技术地图。唯有持续突破跨模态蒸馏在理论与方法上的双重瓶颈，才能真正释放多模态智能在端侧场景中的应用潜力 [12][35]。

References

[1] WU C L, HU M H, LIAO J Z, et al. Research on Fake Information Detection Technology in Re-

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

source-Constrained Scenarios[J]. Computer Science, 2024, 51(11): 15-22.
[2] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J]. arXiv preprint arXiv:1503.02531, 2015.
[3] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models From Natural Language Supervision[C]. International Conference on Machine Learning, 2021: 8748-8763.
[4] YIN S T, WANG X F, ZHANG W, et al. Multi-Teacher Knowledge Distillation Method Based on Adaptive Temperature[J]. Neurocomputing, 2025 (online preprint).
[5] LIU H, LI C, WU Q, et al. Visual Instruction Tuning[C]. Advances in Neural Information Processing Systems, 2023.
[6] WENG Y, DONG J, HE W, et al. Zero-Shot Cross-Lingual Knowledge Transfer in VQA via Multimodal Distillation[J]. IEEE Transactions on Computational Social Systems, 2025 (online preprint).
[7] CHEN S, LI X R, LI D M, et al. Pattern Recognition of φ-OTDR Based on Cross-Model Knowledge Distillation[J]. Laser & Optoelectronics Progress, 2024 (online preprint).
[8] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: Hints for Thin Deep Nets[C]. International Conference on Learning Representations, 2015.
[9] MANSOURIAN A M, AHMADI R, GHAFOURI M, et al. A Comprehensive Survey on Knowledge Distillation[J]. Trans. Mach. Learn. Res., 2025 (online preprint).
[10] XIAO Z, XING H, QU R, et al. Heterogeneous Mutual Knowledge Distillation for Wearable Human Activity Recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025 (online preprint).
[11] CHEN J, XIAO S, ZHANG P, et al. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation[C]. Annual Meeting of the Association for Computational Linguistics, 2024.
[12] XU X, LI M, TAO C, et al. A Survey on Knowledge Distillation of Large Language Models[J]. arXiv preprint arXiv:2402.13116, 2024.
[13] PAN Z, WU Q, JIANG H, et al. LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression[C]. Annual Meeting of the Association for Computational Linguistics, 2024.
[14] PATIL R, GUDIVADA V N. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)[J]. Applied Sciences, 2024, 14(10): 4307.
[15] MURALIDHARAN S, SREENIVAS S T, JOSHI R, et al. Compact Language Models via Pruning and Knowledge Distillation[C]. Neural Information Processing Systems, 2024.
[16] LIANG P P, ZADEH A, MORENCY L. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions[J]. ACM Computing Surveys, 2024, 56(10): 1-42.
[17] GOU J, CHEN Y, YU B, et al. Reciprocal Teacher-Student Learning via Forward and Feedback Knowledge Distillation[J]. IEEE Transactions on Multimedia, 2024, 26: 7598-7612.
[18] MOSLEMI A, BRISKINA A, DANG Z, et al. A survey on knowledge distillation: Recent advancements[J]. Machine Learning with Applications, 2024, 18: 100605.
[19] SUN G, SHU H, SHAO F, et al. FKD-Med: Privacy-Aware, Communication-Optimized Medical Image Segmentation via Federated Learning and Model Lightweighting Through Knowledge Distillation[J]. IEEE Access, 2024, 12: 89234-89248.
[20] SHU F, LIAO Y, ZHUO L, et al. LLaVA-MoD: Making LLaVA Tiny via MoE Knowledge Distillation[C]. International Conference on Learning Representations, 2024.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

[21] LIU C, HE S, ZHOU Q, et al. Large Language Model Guided Knowledge Distillation for Time Series Anomaly Detection[C]. International Joint Conference on Artificial Intelligence, 2024.
[22] TIAN Y, HAN Y, CHEN X, et al. Beyond Answers: Transferring Reasoning Capabilities to Smaller LLMs Using Multi-Teacher Knowledge Distillation[C]. Web Search and Data Mining, 2024.
[23] ELHENAWY M, ABUTAHOUN A, ALHADIDI T I, et al. Visual Reasoning and Multi-Agent Approach in Multimodal Large Language Models (MLLMs): Solving TSP and mTSP Combinatorial Challenges[J]. Machine Learning and Knowledge Extraction, 2024, 6(3): 1641-1657.
[24] WANG Y, PAN Y, YAN M, et al. A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions[J]. IEEE Open Journal of the Computer Society, 2023, 4: 280-302.
[25] LI Z, LI H, MENG L. Model Compression for Deep Neural Networks: A Survey[J]. Computers, 2023, 12(3): 60.
[26] CHEN F, ZHANG D, HAN M, et al. VLP: A Survey on Vision-language Pre-training[J]. Machine Intelligence Research, 2023, 20(1): 38-56.
[27] CHEN G, SHEN L, SHAO R, et al. LION: Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge[C]. Computer Vision and Pattern Recognition, 2023.
[28] XU C, MCAULEY J. A Survey on Model Compression and Acceleration for Pretrained Language Models[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(9): 10566-10575.
[29] ZHANG L, MA K. Structured Knowledge Distillation for Accurate and Efficient Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1513-1527.
[30] WEIDINGER L, UESATO J, RAUH M, et al. Taxonomy of Risks posed by Language Models[C]. 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022: 214-229.
[31] LIU C, TAO C, FENG J, et al. Multi-Granularity Structural Knowledge Distillation for Language Model Compression[C]. Annual Meeting of the Association for Computational Linguistics, 2022: 1001-1011.
[32] GANESH P, CHEN Y, LOU X, et al. Compressing Large-Scale Transformer-Based Models: A Case Study on BERT[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 1061-1080.
[33] KIM J, CHANG S, KWAK N. PQK: Model Compression via Pruning, Quantization, and Knowledge Distillation[C]. Interspeech, 2021: 3850-3854.
[34] FU H, ZHOU S, YANG Q, et al. LRC-BERT: Latent-representation Contrastive Knowledge Distillation for Natural Language Understanding[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(14): 12830-12838.
[35] GOU J, YU B, MAYBANK S, et al. Knowledge Distillation: A Survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.
[36] LI K, YU L, WANG S, et al. Towards Cross-Modality Medical Image Segmentation with Online Mutual Knowledge Distillation[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 775-783.
[37] CHEN H, WANG Y, SHU H, et al. Distilling Portable Generative Adversarial Networks for Image Translation[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 3585-3592.
[38] PAN H, WANG C, QIU M, et al. Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains[C]. Annual Meeting of the Association for Computational Linguistics, 2020: 3026-3036.
[39] CHEN Z, ZHANG L, CAO Z, et al. Distilling the Knowledge From Handcrafted Features for Human

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

Activity Recognition[J]. IEEE Transactions on Industrial Informatics, 2018, 14(10): 4334-4342.
[40] WANG J, BAO W, SUN L, et al. Private Model Compression via Knowledge Distillation[C]. AAAI Conference on Artificial Intelligence, 2018: 1756-1763.
[41] CHU X, QIN L, LIU H, et al. MobileVLM: A Fast, Strong and Open Vision Language Assistant for Mobile Devices[C]. Computer Vision and Pattern Recognition, 2024. DOI:10.48550/arXiv.2312.16886.
[42] FANG Y, LIAO B, WANG X, et al. EVA-CLIP: Improved Training Techniques for CLIP at Scale[C]. Computer Vision and Pattern Recognition, 2024. DOI:10.48550/arXiv.2303.15389.
[43] ZHAI X, MUSTAFA B, KOLESNIKOV A, et al. Sigmoid Loss for Language Image Pre-Training[C]. International Conference on Computer Vision, 2023. DOI:10.48550/arXiv.2303.15343.
[44] WU K, PENG H, ZHOU Z, et al. TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance[C]. International Conference on Computer Vision, 2023: 21970-21980. DOI:10.1109/ICCV51070.2023.02009.
[45] WANG T, ZHOU W, ZENG Y, et al. EfficientVLM: Fast and Accurate Vision-Language Models via Knowledge Distillation and Modal-adaptive Pruning[C]. Annual Meeting of the Association for Computational Linguistics, 2023: 8475-8492. DOI:10.18653/v1/2023.findings-acl.538.
[46] ZHOU B, HU Y, WENG X, et al. TinyLLaVA: A Framework of Small-scale Large Multimodal Models[C]. European Conference on Computer Vision, 2024. DOI:10.48550/arXiv.2402.14289.
[47] LI Y, WANG R, YANG J, et al. MC-LLaVA: Multi-Concept Personalized Vision-Language Model[C]. Computer Vision and Pattern Recognition, 2024. DOI:10.48550/arXiv.2404.17521.
[48] CHEN L, LI J, DONG X, et al. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions[C]. European Conference on Computer Vision, 2024. DOI:10.48550/arXiv.2311.12793.
[49] CHEN J, ZHU D, SHEN X, et al. MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning[C]. International Conference on Learning Representations, 2024. DOI:10.48550/arXiv.2310.09478.
[50] YUAN Z, LI C, SUN P, et al. TinyGPT-V: Efficient Multimodal Large Language Model via Small Backbones[C]. International Conference on Learning Representations, 2024. DOI:10.48550/arXiv.2312.16862.
[51] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling Laws for Neural Language Models[J]. arXiv preprint arXiv:2001.08361, 2020. DOI:10.48550/arXiv.2001.08361.
[52] OPENAI. GPT-4 Technical Report[J]. arXiv preprint arXiv:2303.08774, 2023. DOI:10.48550/arXiv.2303.08774.
[53] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and Efficient Foundation Language Models[C]. International Conference on Machine Learning, 2023. DOI:10.48550/arXiv.2302.13971.
[54] LI J, LI D, SAVARESE S, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models[C]. International Conference on Machine Learning, 2023: 19730-19742. DOI:10.48550/arXiv.2301.12597.
[55] DAI W, LI J, LI D, et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning[C]. Advances in Neural Information Processing Systems, 2023. DOI:10.48550/arXiv.2305.06500.
[56] YE Q, XU H, XU G, et al. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality[C]. Annual Meeting of the Association for Computational Linguistics, 2024. DOI:10.48550/arXiv.2304.14178.
[57] BAI J, BAI S, YANG S, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Lo-

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

calization, Text Reading, and Beyond[C]. Computer Vision and Pattern Recognition, 2024. DOI:10.48550/arXiv.2308.12966.
[58] LIU Y, DUAN H, ZHANG Y, et al. MMBench: Is Your Multi-modal Model an All-around Player?[C]. European Conference on Computer Vision, 2024. DOI:10.48550/arXiv.2307.06281.
[59] LI B, ZHANG Y, CHEN L, et al. POPE: Polling-based Object Probing Evaluation for Object Hallucination[C]. Empirical Methods in Natural Language Processing, 2023: 7134-7148. DOI:10.18653/v1/2023.emnlp-main.441.
[60] LIN J, TANG J, TANG H, et al. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration[C]. Machine Learning and Systems, 2024. DOI:10.48550/arXiv.2306.00978.
[61] FRANTAR E, ASHKBOOS S, HOEFLER T, et al. GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers[C]. International Conference on Learning Representations, 2023. DOI:10.48550/arXiv.2210.17323.
[62] MA X, FANG G, WANG X. LLM-Pruner: On the Structural Pruning of Large Language Models[C]. Advances in Neural Information Processing Systems, 2023. DOI:10.48550/arXiv.2305.11627.
[63] SHANG Y, YUAN Z, XIE B, et al. PB-LLM: Partially Binarized Large Language Models[C]. International Conference on Learning Representations, 2024. DOI:10.48550/arXiv.2310.00034.
[64] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-Rank Adaptation of Large Language Models[C]. International Conference on Learning Representations, 2022. DOI:10.48550/arXiv.2106.09685.
[65] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. QLoRA: Efficient Finetuning of Quantized LLMs[C]. Advances in Neural Information Processing Systems, 2023. DOI:10.48550/arXiv.2305.14314.
[66] WANG W, BAO H, DONG L, et al. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks[C]. Computer Vision and Pattern Recognition, 2023: 19175-19186. DOI:10.1109/CVPR52729.2023.01838.
[67] LIU Z, OGUZ B, ZHAO C, et al. LLM-QAT: Data-Free Quantization Aware Training for Large Language Models[C]. Annual Meeting of the Association for Computational Linguistics, 2024. DOI:10.48550/arXiv.2305.17888.
[68] SUN M, LIU Z, BAIR A, et al. A Simple and Effective Pruning Approach for Large Language Models[C]. International Conference on Learning Representations, 2024. DOI:10.48550/arXiv.2306.11695.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)