Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(3), 943-952
Published Online May 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.153086

Hans汉斯

大模型时代的多模态融合：方法、评测与前沿挑战综述

蒋松冬

广西民族师范学院，数学与计算机科学学院，广西 崇左

收稿日期：2026年3月30日；录用日期：2026年5月21日；发布日期：2026年5月29日

摘　要

目的：系统梳理大模型时代多模态融合的方法演进与评测难点，说明如何在提高能力的同时，让实验结果可重复、评测过程可核查。方法：建立对齐、桥接、深度交互与统一建模四层分析框架，对照代表模型路线、训练数据的组织方式以及分阶段训练策略，并提出MMP-Next评测草案，包括模型与数据说明清单、评测运行配置表、推理过程是否稳定等指标，以及鲁棒性与安全方面的最小测试集合。结果：多模态大模型在通用理解、任务迁移和日常交互上进步明显，但长文本下的信息压缩损失、多模态幻觉、解码与提示词带来的分数波动，以及换场景或遇对抗样本时的安全与稳定性不足，在短期内仍难以单靠扩大模型规模根除。结论：后续研究宜在扩大参数之外，同步改进融合方式与评测规范，推动能对照证据的推理方式，以及有统一格式、可复核的评测流程。

关键词

多模态融合，多模态大模型，跨模态对齐，评测体系，多模态幻觉

Multimodal Fusion in the Era of Large Models: Methods, Evaluation, and Frontier Challenges

Songdong Jiang

School of Mathematics and Computer Science, Guangxi Minzu Normal University, Chongzuo Guangxi

Received: March 30, 2026; accepted: May 21, 2026; published: May 29, 2026

Abstract

Objective: To review methodological trends and evaluation difficulties of multimodal fusion in the

文章引用: 蒋松冬. 大模型时代的多模态融合: 方法、评测与前沿挑战综述[J]. 人工智能与机器人研究, 2026, 15(3): 943-952. DOI: 10.12677/airr.2026.153086

foundation-model era, and to clarify how stronger capability can coexist with reproducible experiments and auditable evaluation. Methods: We build a four-layer framework spanning alignment, bridging, deep interaction, and unified modeling; we relate representative model routes to how training data are organized and to staged training strategies, and we outline the MMP-Next evaluation draft, including model/data disclosure checklists, evaluation run sheets, indicators of whether inference is stable across settings, and a minimal test set for robustness and safety. Results: Multimodal large models improve markedly in general understanding, task transfer, and everyday interactive use, yet information loss under long-context compression, multimodal hallucination, score volatility from decoding and prompting, and gaps in safety and stability under domain shift or adversarial conditions are unlikely to be removed in the short term by parameter scaling alone. Conclusion: Beyond enlarging model size, research should jointly refine fusion mechanisms and evaluation norms, advancing evidence-grounded reasoning together with standardized, reviewable evaluation workflows.

Keywords

Multimodal Fusion, Multimodal Large Language Models, Cross-Modal Alignment, Evaluation Systems, Multimodal Hallucination

1. 引言

过去十年，多模态学习从围绕检索、视觉问答(VQA)等任务的专用模型，转向把图像、音频、视频与文档接到大语言模型上的统一接口。大语言模型擅长组织知识、听从指令并做多步推理，使多种感知信息与语言推理在开放场景中成为主流做法；其中，融合层决定各模态的信息能否在同一语义空间里稳定互通，从而直接影响复杂推理、可控生成以及输出能否对照证据核实[1]。

本文将多模态融合界定为：在同一任务目标下，通过可学习的模块把各模态编码后的结果映射到便于联合推理的表示，并支持长文本输入与高风险场景下的决策。与简单拼接特征相比，融合强调对齐、注意力、门控或路由等明确的跨模态联系，能减轻只看某一模态或名不副实的假融合问题，但在开放环境中仍要面对噪声被放大和幻觉等风险。

针对概念分散、不同论文之间评测难以直接对比等问题，本文贡献可概括为三点：一是用四层框架梳理主流融合路线及其常见工程组合；二是概括数据来源、质量控制和分阶段训练与各类测试集(benchmark)，并讨论评测是否可信；三是归纳幻觉、长序列、鲁棒性与安全等方面的典型失败表现，并给出 MMP-Next 式的协议化评测思路。

2. 相关概念与评述标准

各模态差异很大：文本是离散符号且语言习惯强，图像和视频带空间结构和时间顺序，文档则混合版式与文字符号。在大模型里，每种模态先由各自的编码器压缩成向量，融合模块则在尽量保留区分内容所需信息的前提下，建立模态之间的接口；评价一种方法时，需要同时看编码是否失真、跨模态对齐是否可靠。

形式上，给定各模态输入集合 $X = \{x_i\}$，经编码得 $Z = \{z_i\}$，融合表示 $h = f_{fusion}(Z)$，任务输出 $\hat{y} = g(h)$。训练时往往把多种损失加在一起：对比学习把语义一致的样本拉近；生成损失支撑开放式回答；

匹配或定位损失强化细粒度对齐；偏好优化则改善安全性和是否听从指令。各项权重大小会直接影响检索准不准、生成流不流畅以及事实是否一致，需要在中间做权衡。

本文从三方面描述融合发生的位置：按阶段可分为早期、中间与后期融合(文献中常写作 early、intermediate、late)；按粒度可分为词元级、特征级与决策级(token, feature, decision)；按更新方式可分为更新全部参数，以及 Adapter、LoRA、提示微调等只训练少量参数的做法。全文沿用四项评价标准：对齐质量、推理是否有效、效率与规模是否可扩展、可靠性与安全性。

3. 发展脉络：从双塔对比到多模态大模型

3.1. 双塔对齐阶段

双塔结构分别把图像和文本编成向量，用对比学习在批次内拉近配对的样本、推开不配对的样本；CLIP 表明，大量带噪声的图文对配合 InfoNCE 等对比学习目标，可以得到便于零样本迁移的共享语义空间[2]。ALIGN 及后续工作进一步说明，超大规模噪声图文仍可训练，且先对齐再融合的多目标联合策略可行。这一阶段为跨模态检索和分类打下基础，但训练目标偏重是否匹配，对细粒度物体关系和开放式生成的约束仍弱，多步推理和长文本场景下的稳定性也有限。

3.2. 统一预训练与生成衔接

编码器与解码器结构以及弱监督条件生成等路线把视觉信息更直接送进语言解码端，描述和问答类任务往往更好，但层与层之间交互变多，训练和推理成本上升，且对提示词分布更敏感。这一阶段的意义是从可对齐的向量表示过渡到按图像等条件生成文字，为后来把多模态词元接入自回归大语言模型铺路。

3.3. MLLM 阶段

随着大语言模型能力快速提升，主流做法变成：先用较强的视觉编码器，经连接器接到冻结或半冻结的大语言模型上，再做视觉指令微调，整条链路分步搭建。LLaVA 用简单投影和指令数据实现了便于复现的多任务接口[3]；Qwen-VL 在多语言和文字识别(OCR)上做了加强[4]。总体来看，开放式问答、文档理解和复杂指令执行进步明显，但连接器可能挡住部分视觉信息，长视频和高分辨率输入又受词元长度限制，再加上各篇论文数据清洗和推理设置不同，分数往往不宜直接横向对比。

从发展顺序看，三个阶段并不是谁完全取代谁：双塔对齐仍是许多系统冷启动和检索增强的组件；统一预训练阶段形成的生成式目标与数据用法，仍体现在指令微调数据里；多模态大模型阶段在好用程度上从能用走向好用，却把幻觉、评测是否透明、部署成本等问题一并摆到台前。ALBEF 等强调先把跨模态对齐做好再进入融合训练，并结合动量蒸馏等技巧，为大规模数据上稳定对齐提供了可反复借用的范式[5]。

4. 大模型多模态融合方法分类

本文用对齐、桥接、深度交互与统一建模四个层次来归类：对齐指跨模态语义上是否靠近；桥接指如何映射到大语言模型的词元空间；深度交互指网络中层持续的跨模态信息交换；统一建模指把各模态都变成同类词元并联合训练目标。该框架用于比较方法和做选型，不能代替数据与安全方面的管理，也不必与某一个 benchmark 逐条一一对应。

上述四层可理解为总览，说明各模态编码器输出之后，在何处、以多强的力度发生跨模态结合：对齐层多出现在编码器之后、进入大语言模型之前的大规模对比或匹配；桥接层多是一次性或较浅的映射；

深度交互层多是在语言模型主干里插入跨注意力等模块；统一建模层则希望在词元序列这一层面弱化模态边界。实际论文里常见多层混合，例如先做对齐预训练，再用连接器接冻结的大语言模型，又在较高层加少量跨注意力块，因此讨论时宜拆到子模块，不宜只凭模型名字归类。多模态融合方法总览框架图如图 1 所示。

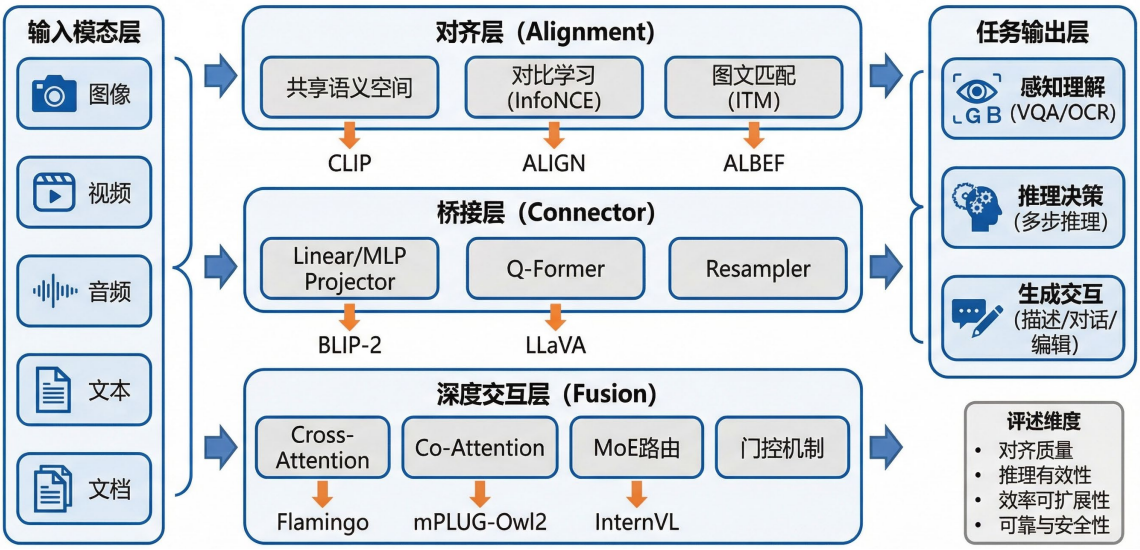

Figure 1. Overview framework of multimodal fusion methods
图 1. 多模态融合方法总览框架图

对齐融合以对比学习、图文匹配和区域级对齐为主，易扩展到大数据，适合作为系统底座和零样本迁移的来源，但对复杂生成式推理帮助有限，细粒度物体等仍常要靠后面的交互模块补足[5]。

桥接(连接器)通过线性或多层感知机投影、Q-Former、Perceiver Resampler 等，把视觉或其他模态的序列映射到大语言模型输入空间；在冻结较强编码器和大语言模型时，只训少量参数就能扩展多模态能力[6]。主要矛盾在压缩程度：压得太狠会丢细节，压得太松会占满上下文并抬高推理成本；分步训练和适度解冻往往更稳。

深度交互在 Transformer 中层加入跨注意力等机制，使视觉与语言表示多轮互相更新，有利于文档理解、细粒度定位(grounding)和长视频推理[7]；代价是显存和延迟明显增加，常需配合分块、稀疏化或共享参数。

MoE 与专家路由通过按条件激活部分专家来扩大有效容量，适合多任务和长尾分布，但要防止路由塌缩、专家忙闲不均以及某一模态被系统性偏重，实践中常与负载均衡正则和蒸馏一起用。

指令微调把融合后的能力变成能交互执行的任务：视觉问答、OCR、对话和多步推理样本都写成统一指令格式训练，系统可用性明显提高[8]。若数据模板过于单一或夹杂错误事实，容易出现图文不符，需要配合偏好对齐、拒答规则等。

统一词元与原生多模态主张各模态用同一种序列化方式并联合建模，理论上能减少模块接口带来的误差；对比与生成联合训练也为这一路提供了经验。现实中序列长度和各模态噪声分布差别大，优化和工程难度高。若将来出现更强的统一分词与分层压缩，有望在保持感知分辨率的同时控制上下文长度，使原生统一路线与模块化堆叠路线的差距重新拉开；在那之前，混合架构仍是较稳妥的权衡。

如图 2 所示，可对代表多模态大模型里视觉编码器、连接器、语言主干和指令适配等典型部件的信息走向与接口关系作概括。

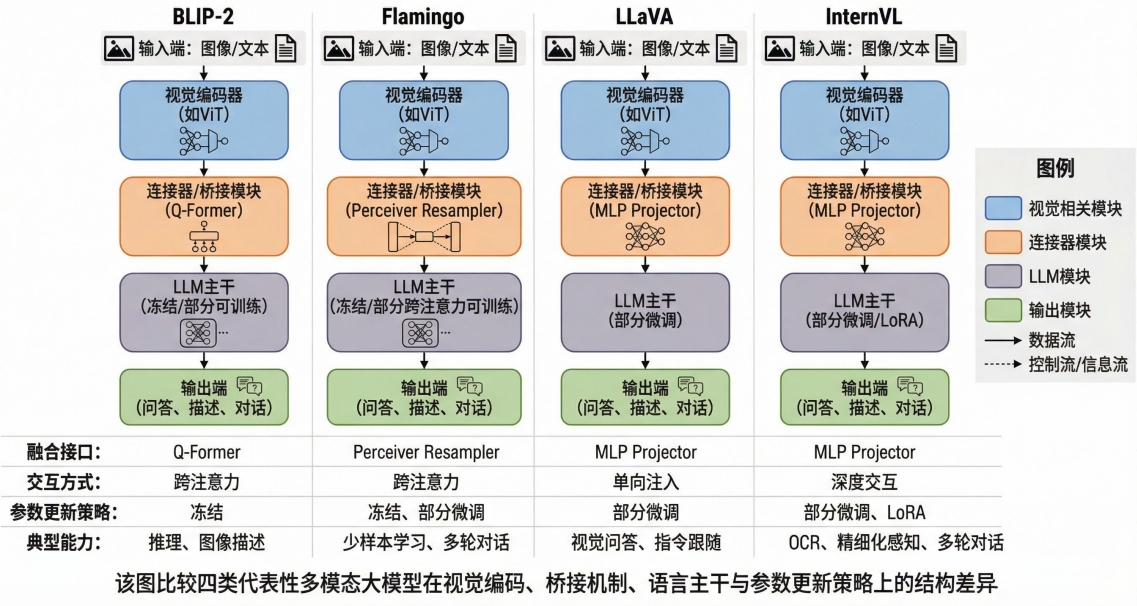

Figure 2. Comparison of representative MLLM architectures
图 2. 代表性 MLLM 架构对比图

综上，各路线可以组合使用：工业界常见流程是先做对齐预训练，再接连接器和指令微调，难样本上再叠加跨注意力或专家分支，在算力有限的前提下换取更好的细粒度和稳定性。

如表 1 所示从若干代表模型概括其支持的模态、融合要点、优势和不足。

Table 1. Cross-model comparison of representative systems
表 1. 代表模型横向比较

模型	年份	模态	融合要点	优势	局限
CLIP	2021	图文	双塔对比(InfoNCE)	零样本迁移强	弱生成与复杂推理
ALBEF	2021	图文	对齐先于融合 (对比、匹配与生成联合)	对齐与下游协同	超大规模扩展成本高
Flamingo	2022	图像或视频与文本	重采样器与交叉注意力	Few-shot 与跨任务稳	训练推理成本高
BLIP-2	2023	图文	Q-Former 桥接冻结 编码器与 LLM	参数高效、易部署	连接器瓶颈
InstructBLIP	2023	图文	指令感知 Q-Former 与指令微调	指令跟随好	依赖指令数据质量
LLaVA-1.5	2023	图文	MLP 投影与视觉指令微调	结构简洁、生态成熟	长上下文与 细粒度受限
Qwen-VL	2023	图文	跨模态连接模块	多语与 OCR	推理稳定性依赖提示
CogVLM	2023	图文	Visual Expert 与 LLM	视觉细节利用强	部署资源要求高
InternVL 1.5	2024	图文	强视觉底座与多阶段对齐	通用榜单与 OCR 均衡	数据与流程工程复杂

章内小结

本章在统一四层框架下回顾了对齐、桥接、深度交互、专家路由、指令微调与统一建模六条技术脉络。总体看：对齐与大规模预训练提供可扩展的语义基础，连接器与参数高效策略降低实际部署门槛，

跨注意力与 MoE 有助于难样本上的细粒度和时序建模，指令与偏好对齐决定交互形态和安全边界，原生统一路线则代表更长期的架构走向。工程上更现实的仍是模块化加分阶段训练，重要的是按任务风险和算力预算，让模型里的融合深度与评测设计相匹配，避免只靠堆结构复杂度。

5. 数据引擎与训练策略

多模态大模型的能力往往来自多个训练阶段和不同参数策略的叠加；各阶段在对齐、桥接和指令融合上的侧重点并不相同。如图 3 所示，可对常见训练流水线作总览，依次包括对齐预训练、适配、指令微调。

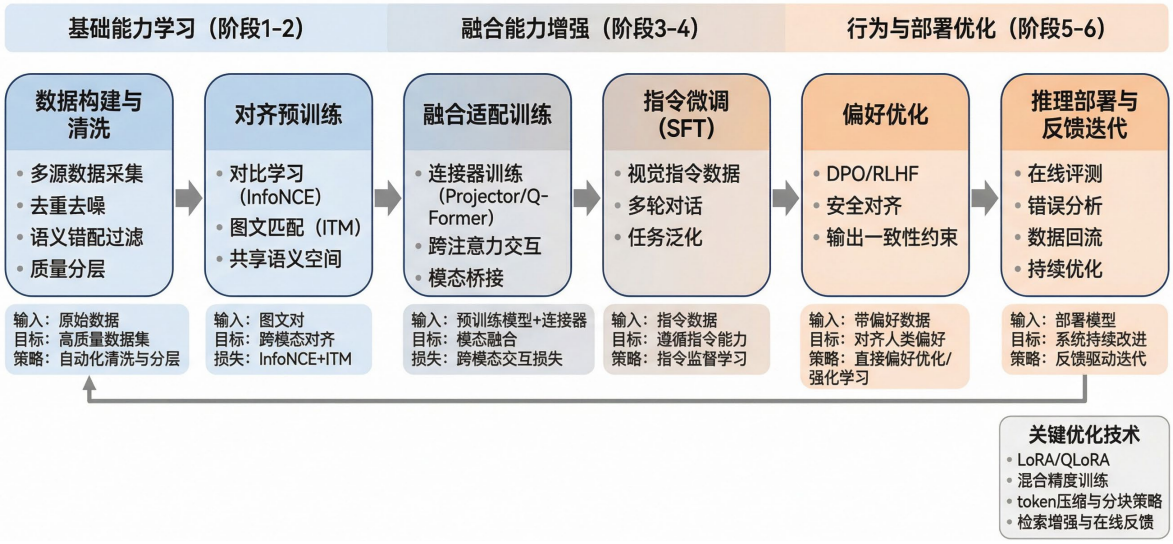

Figure 3. Staged training pipeline for multimodal large models
图 3. 阶段化训练流程图

5.1. 数据来源与构成

训练语料通常包括图文对话、视频配文字、语音配文字、带版式的文档图以及多轮指令对话等；其中大量弱标注图文仍是对齐预训练的主力，视频和文档类数据补全时间顺序和版面理解，指令类数据则决定模型能执行哪些交互、覆盖哪些任务。合成和蒸馏数据能较快补全长尾指令和难例，但若占比过高，容易出现套话和风格偏移，在真实分布上反而泛化变差。实践中更强调在真实度、覆盖面和难度梯度之间配平，而不是一味堆样本量。

5.2. 质量控制

去重、去噪和有害内容过滤可以减轻死记硬背和不公平关联；多模态里特别要注意的是语义错配，即图和文或视频与解说对不上，会直接破坏对齐稳定性。工程上常用规则加模型打分先筛一遍，再抽一部分人工核对；也要避免过度过滤，使数据分布过窄、换场景时鲁棒性变差。

5.3. 训练流程与参数策略

常见流程依次是跨模态对齐预训练、连接器或交互模块适配、视觉或多模态指令微调，需要时再做人反馈与偏好对齐。LoRA、QLoRA、Adapter 等只训少量参数的做法更适合行业快速适配。面对长序列和高分辨率输入，常配合词元压缩、分块前向和混合精度，在效果、速度、稳定性和实验记录是否完整

之间做权衡。若以多轮对话为主，还要注意样本怎么构造：轮次之间信息是否重复、是否人为截短证据链，都会影响模型有多依赖融合后的记忆。

6. 评测任务、基准与评价指标

6.1. 任务体系

可按能力大致分成三类：一是感知理解，如视觉问答(VQA)、文字识别(OCR)、文档问答(DocVQA)、图表问答(ChartQA)、视频问答(VideoQA)等；二是推理决策，如多步看图推理、读科学图表等；三是生成与交互，如图像描述、视觉对话和编辑类任务。除单项分数外，还应关注多轮对话、长文本和跨场景组合下成绩是否突然大幅下滑，以判断融合是否真的支撑复杂交互。同一类任务里也有浅层和深层之分：例如文档问答有时只需找到关键词，有时却要跨表格格子做数值推导。

6.2. 主流 benchmark 与评价指标

通用测试集如 MMBench [9]、MMMU [10]、MMVet [11]等覆盖多种题型，便于横向比较综合能力；医疗、遥感、工业等领域集则检验专业知识和面对风险时的决策是否可靠。两类测试互补，但不能把分数简单加总排名：题型是选择题还是开放生成、评分规则、语言和领域分布都会影响分数该怎么读。

准确率、F1、ANLS、BLEU、CIDEr 等自动指标对意思相近的结果和推理过程刻画不够；若用大模型评估，需要固定模型版本和提示词，并说明可能存在的偏向[12]。评测可信度常见风险包括：训练数据混进测试、改提示词分数就大幅波动、不报告方差或重复实验等。

如表 2 所示，给出部分代表性 benchmark 的主要指标。

Table 2. Benchmarks and evaluation metrics: a crosswalk
表 2. Benchmark 与指标对照

数据集	侧重	主要指标
MMBench	通用多模态理解	Accuracy
MMMU	跨学科专家级推理	Accuracy (分科)
MMVet	开放综合能力	GPT 评审辅以人工
SEED-Bench	理解与生成协同	Accuracy/生成指标
Video-MME	视频时序与事件	Accuracy/F1
DocVQA	文档版面问答	ANLS/F1
ChartQA	图表数值推理	Accuracy/EM
TextVQA	场景文字理解	Accuracy
VQAv2	基础视觉问答	Accuracy

7. 关键挑战

7.1. 多模态幻觉

模型说的话与画面、时间线或页面内容对不上，但行文通顺，用户往往一时难以发现。常见原因包括跨模态对齐做得不够、生成时更依赖语言习惯而非视觉内容，以及训练数据里图文对错配被放大。解决办法包括检索后再核对、生成时约束必须依据证据、细粒度对齐损失和可学习的拒答机制等，但在完全开放的场景里仍很难根除[13]。还要注意，幻觉不只出现在写答案时：在调用工具或输出结构化内容(如

JSON)时，模型也可能编造并不存在的图像区域编号或页码；因此评测除看文字是否答对外，还应逐步加入可自动核对的位置信息(grounding)和引用是否前后一致等检查。

7.2. 长视频、长文档与上下文瓶颈

词元长度上限迫使对长输入做分块、采样或分层摘要，容易打断跨片段的依赖关系，表现为视频前后顺序对不上或跨页文档无法推理。加记忆模块或用检索拼接上下文可以解决此类问题，但也可能让新的信息丢失，需要进行权衡。

7.3. 鲁棒性与泛化

在固定测试集上分数高，并不代表换到真实环境可以保持效果。分辨率变化、压缩产生的伪影、不同传感器和跨场景迁移都可能让表现明显下降；在分布外样本上，还常出现模型很自信却答错的情况。改进需要数据上加强扰动和长尾覆盖、模型上考虑不确定性和输出一致性、评测上补充跨场景和抗干扰测试集，通常很难靠一次精调就长期见效。

7.4. 安全、可信与部署成本

多模态比纯文本多了不少可被攻击的点：图或文档里故意写的字可能骗过 OCR，图文一起下的指令可能绕过只查文本的过滤；深度伪造和诱导泄露隐私也会增加合规压力。可信部署还要求能说清理由、便于追责，例如回答能否对应到具体图像区域或检索来源。另一方面，高分辨率、长序列和深层交互一起会占用显存、拉长延迟，边缘设备往往装不下；云端部署则要权衡吞吐量、延迟和费用。量化、蒸馏、动态路由和分级推理是常见做法，但可能掉精度或改变模型行为，需要按固定协议做回归测试。

多模态融合挑战与对策关系如图 4 所示。

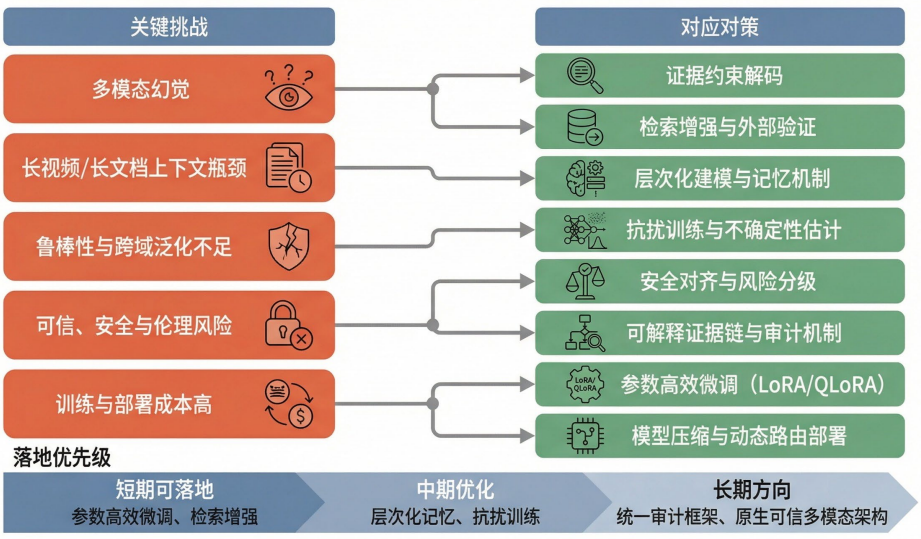

Figure 4. Mapping of key challenges to mitigation strategies
图 4. 挑战与对策映射图

8. 多模态融合研究前沿

8.1. 原生多模态基础模型与规模化路径

当前产业界仍以大语言模型为核心、外挂视觉等编码器的分层结构为主，做法清晰，但各模块接口

和训练阶段若切得太碎，深层协同会受限。前沿方向更强调在同一主干里联合建视觉、语言乃至音频和视频，减少信息反复投影带来的损耗，并让理解与生成共用更一致的训练目标。InternVL 1.5 等在扩大视觉基础模型并与语言对齐方面提供了可参照的大规模实践[14]；GPT-4V 系统说明则从能力和风险两方面指出：性能提升的同时，要把安全说明、拒答规则和已知失效场景一起迭代[15]。未来更现实的形态，将是原生多模态主干再搭配连接器、检索或 MoE 等部件的混合方案，在能力上限和落地成本之间取平衡。

8.2. 因果表示、可解释融合与多模态 Agent

只靠统计相关性做融合，在训练与使用场景不一致时容易失效。引入跨模态因果和反事实训练，可以减轻语言习惯带来的幻觉。可解释方面，除画注意力热图外，还可探索输出能否与检索到的证据、工具执行步骤逐条匹配。多模态智能体把融合从单次问答扩展成感知、规划、调用工具、反馈的闭环，评测就不能只看单轮对错，还要看多步任务成功率、工具是否被乱用、跨轮记忆是否一致，否则看不出系统级是否可靠。接上工具链后，融合模块还要和权限、参数检查和沙箱执行环境一起设计。

8.3. 低资源场景与评测体系升级

行业里长尾任务常缺标注，参数高效微调、迁移学习和合成数据能缩短适配时间，但要当心合成数据分布跑偏、模型背模板。更稳妥的是同时报告少样本下的表现和不确定性范围，并留好人机一起纠错的通道。只靠离线 benchmark 也难覆盖真实交互里动态出现的风险，因此需要定期更新测评体系，使评测场景尽量接近上线后会遇到的风险形态。

9. MMP-Next v0.1 评测草案

为提高不同论文、机构和模型版本之间的可比性，本文提出多模态模型评测协议草案(Multimodal Model Evaluation Protocol，下文简称 MMP-Next)作为一层最低限度的共同约定，强调可追溯、可复现、可比较和对风险敏感四条原则。

(1) Model/Data Card。用固定条目写清模型名称与版本、支持哪些模态、参数量和上下文长度、对齐与指令微调各做了什么、是否蒸馏或做人偏好对齐；数据侧列出用了哪些公开集、合成与蒸馏各占多少、如何去重和控质量、版权与许可；并写明已知的失效场景。未提供的信息统一标成 not disclosed，也避免与「不详」「未知」等笼统措辞混为一谈。。

(2) Evaluation Run Sheet。除最终分类外，还应公开硬件(GPU 型号、显存、驱动)、软件(框架与推理引擎版本)、解码参数(temperature、top-p、max tokens 等)、输入预处理(分辨率、视频怎么采样、OCR 怎么走)、提示模板、批量与重试策略、随机种子与重复次数；统计上给出均值、标准差或置信区间。建议正文主表在统一默认配置下给出比较指标，敏感性分析放到附录里每次只改一个变量，减轻隐性调参把排名抬高的现象。

(3) 过程一致性指标。对答案对但推理过程来回变的现象，建议报告 Self-Consistency@N、换几种说法问同一件事是否稳定、证据对齐的精确率与召回，以及对关键证据做最小改动后的敏感程度，从正确率和一致性两个维度一起看，防止模型靠固定话术掩盖其实没看证据。

(4) 鲁棒与安全最小场景集。覆盖输入变差、场景偏移、长文本压力、缺某一模态以及模态互相矛盾等情况；安全侧覆盖恶意提示注入、越权指令、诱导泄露隐私、触发偏见和高风险领域里的错误建议等，并统一报告任务指标变化、校准情况、风险行为比例和防护拦截比例。建议在附录给出可复查的失败样例，方便社区对齐基线。

10. 结论

本文系统回顾了大模型时代多模态融合的方法脉络：从双塔对比与统一预训练，到以连接器和指令微调为核心的多模态大模型，再到跨注意力、专家路由与原生多模态等深化方向。用对齐、桥接、深度交互与统一建模四层框架组织全文，有助于在同一套坐标下比较各路线的瓶颈与常见组合，并与数据怎么匹配、训练分几步、benchmark 怎么设形成对照。总的来看，融合方式已成为影响能力上限和部署成本的重要因素，但长文本压缩导致证据丢失、多模态幻觉、评测分数随解码和提示词大幅波动，以及鲁棒性和安全表现不稳，仍是阻碍可信应用的主要短板。未来单靠扩大参数量很难解决这些问题；更可行的做法是在模型、数据和评测三方面一起推进。后续可在本文框架下，按具体模态(如高帧率视频、多页 PDF、多路麦克风)补充更详细的融合与评测案例库，并把 MMP-Next 草案与公开排行榜或开源评测工具对接。

基金项目

广西民族师范学院校级科研项目《基于大语言模型的多模态数据融合研究》(项目编号：2024QN055)。

参考文献

[1] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., *et al*. (2024) A Survey on Multimodal Large Language Models. *National Science Review*, **11**, nwae403. https://doi.org/10.1093/nsr/nwae403

[2] Radford, A., Kim, J.W., Hallacy, C., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. 2021 38*th International Conference on Machine Learning*, Online, 18-24 July 2021, 8748-8763.

[3] Liu, H.T., Li, C.Y., Wu, Q.Y., *et al*. (2023) Visual Instruction Tuning. https://arxiv.org/abs/2304.08485

[4] Bai, J., Bai, S., Yang, S., *et al*. (2023) Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities.

[5] Li, J., Selvaraju, R.R., Gotmare, A., *et al*. (2021) Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. 2021 *NeurIPS*, Online, 6-14 December 2021, 9694-9705.

[6] Li, J., Li, D., Savarese, S., *et al*. (2023) BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. *The* 2023 *International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 19730-19742.

[7] Alayrac, J., Barr, I., Barreira, R., Binkowski, M., Borgeaud, S., Brock, A., *et al*. (2022) Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems* 35, New Orleans, 28 November-9 December 2022, 23716-23736. https://doi.org/10.52202/068431-1723

[8] Dai, W., Fung, P.N., Hoi, S., Li, B., Li, J., Li, D., *et al*. (2023) Instructblip: Towards General-Purpose Vision-Language Models with Instruction Tuning. *Advances in Neural Information Processing Systems* 36, New Orleans, 10-16 December 2023, 49250-49267. https://doi.org/10.52202/075280-2142

[9] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., *et al*. (2024) Mmbench: Is Your Multi-Modal Model an All-Around Player? In: *Lecture Notes in Computer Science*, Springer, 216-233. https://doi.org/10.1007/978-3-031-72658-3_13

[10] Yue, X., Ni, Y., Zheng, T., Zhang, K., Liu, R., Zhang, G., *et al*. (2024) MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 9556-9567. https://doi.org/10.1109/cvpr52733.2024.00913

[11] Yu, W., Lu, J., Zhou, Y., *et al*. (2023) MMVet: Evaluating Large Multimodal Models for Integrated Capabilities.

[12] Gemini, T., Anil, R., Borgeaud, S., *et al*. (2023) Gemini: A Family of Highly Capable Multimodal Models.

[13] Liu, H., Li, C., Li, Y., *et al*. (2024) LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge.

[14] Wang, W., Chen, Z., Liu, Y., Cao, Y., Wang, W., Zhu, X., *et al*. (2024) InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In: *Advances in Computer Vision and Pattern Recognition*, Springer, 23-57. https://doi.org/10.1007/978-3-031-94969-2_2

[15] OpenAI (2023) GPT-4V(Vision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf