/TemplateVersion (2026.1)

当语言占主导地位时：揭示多模态大语言模型中的文本主导性

Huyu Wu[1],
Meng Tang[2], Xinhan Zheng[3],
Haiyun Jiang[4*]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]Department of Computer Science, Aberystwyth University
[3]Beijing University of Posts and Telecommunications
[4]Shanghai Jiao Tong University
huyu-wu@outlook.com, Met57@aber.ac.uk, chengfengke@bupt.edu.cn, haiyunjiangnlp@gmail.com

Abstract

多模态大型语言模型（MLLMs）在各种多模态任务中展示了卓越的能力。然而，这些模型存在一个核心问题，称为文本主导性：它们在推理时严重依赖文本，而对其他模态的利用不足。虽然已有研究在视觉-语言任务中承认了这一现象，通常将其归因于数据偏差或模型架构。在本文中，我们首次系统地调查了跨各种数据模态的文本主导性，包括图像、视频、音频、时间序列和图表。为了衡量这种不平衡，我们提出了两个评价指标：模态主导指数（MDI）和注意力效率指数（AEI）。我们的综合分析表明，文本主导性在所有测试的模态中既显著又普遍。我们的深入分析识别出了三个根本原因：非文本模态中的严重令牌冗余导致的注意力稀释、融合架构设计的影响以及隐含偏向文本输入的任务设定。此外，我们提出了一种简单的令牌压缩方法，有效地重新平衡了模型的注意力。例如，将该方法应用于 LLaVA-7B，可以将其 MDI 从 10.23 大幅降低到一个平衡良好的值 0.86。我们的分析和方法框架为开发更公平、更全面的多模态语言模型提供了基础。

最近的多模态大语言模型（MLLMs）在包括图像、视频、音频以及图形数据等多种模态下的理解和生成方面取得了令人瞩目的成功。然而，这些模型的一个关键弱点是其模态不平衡。一个主要限制是 MLLMs 经常忽略非文本输入，即使在存在丰富视觉信息的情况下，生成的输出也主要基于文本上下文。

这种模态不平衡此前已在视觉问答（VQA）等任务中被观察到。例如，一些研究 (Liu et al. 2024b) 显示，即使在缺少图像的情况下，VQA 模型也常常能够正确回答问题，这揭示了对语言先验的高度依赖。近日，Leng 等人 (Leng et al. 2024) 提出了模态重要性评分（MIS），作为一个定量指标来评估视频问答基准测试中的模态不平衡。然而，先前的工作主要将这种偏差归因于数据伪影 (Wang et al. 2024) 或编码器设计 (Liu et al. 2024b; Luo et al. 2025)，主要是在图像-文本模态对中。内部注意力机制的作用——它是 Transformer 架构的核心——在导致这种不平衡，尤其是在更广泛的模态范围内，仍然没有得到充分探索。这一差距引发了一个关键研究问题：文本主导是否是 Transformer 架构在 MLLMs 中的基本缺陷，超越视觉延伸到音频、时间序列和图等模态?

为研究这一点，我们对领先的多模态大语言模型在这五种模态中的跨模态注意力进行了首次系统分析。我们引入了两个新的指标，模态主导指数（MDI）和注意力效率指数（AEI），以量化这种行为。我们的研究发现突出了一种显著的不平衡：在 VideoLLaMA-7B 中，MDI 达到 157，这表明输出标记对文本标记的关注度是视觉标记的 157 倍。

通过综合分析，我们确定了导致文本占主导地位的三个主要因素。首先，非文本模态通常包含过多的冗余标记，这极大地削弱了模型的注意力。其次，复杂的多模态融合架构往往会放大这种不平衡，而更简单的融合设计则有助于更均衡的注意力分配。第三，许多多模态任务的制定自然倾向于优先考虑文本输入，自然引导模型更加关注文本模态。

受到我们对注意力稀释的发现的启发，我们提出了一种简单而有效的解决方案：标记压缩。通过在非文本模式中有策略地减少冗余标记，这种方法实质上重新平衡了跨模态注意力分布。该方法增强了每个标记的信息密度，并有效地缓解了文本主导的问题。

在此基础上，我们的贡献如下：

- 我们首次提供证据表明文本主导性是 Transformer 架构的多模态语言模型中一个根本且普遍的偏差，这种偏差延伸到广泛的多样性模态中。
- 我们对基本原因进行了全面分析，包括非文本模态中的标记冗余、融合架构设计的影响，以及隐含地偏向文本输入的任务形式。
- 我们提出并验证了令牌压缩，这是一种直接且有效的方法以减少文本主导性。

本文的主要内容在以下部分中呈现。第 3 部分　详细介绍了评估框架并形式化了我们的核心指标。第 4 部分　提供了关于文本主导性的全面分析，考虑了不同模型架构和跨多模态任务设计的影响。第 5 部分　描述了令牌压缩方法，并检验其在解决模态不平衡问题上的有效性。

相关工作

多模态大型语言模型的扩展前沿

大型语言模型（LLM）(Yin et al. 2024; Kumar 2024) 的显著成功催化了向多模态大型语言模型（MLLM）(Yin et al. 2024; Qin et al. 2025) 的范式转变，它整合了多种数据模态。规范的 MLLM 架构包括一个预训练的模态特定编码器，一个作为认知核心的强大 LLM，以及一个精心设计的接口以对齐跨模态的表示 (Liang et al. 2024)。

*Corresponding author

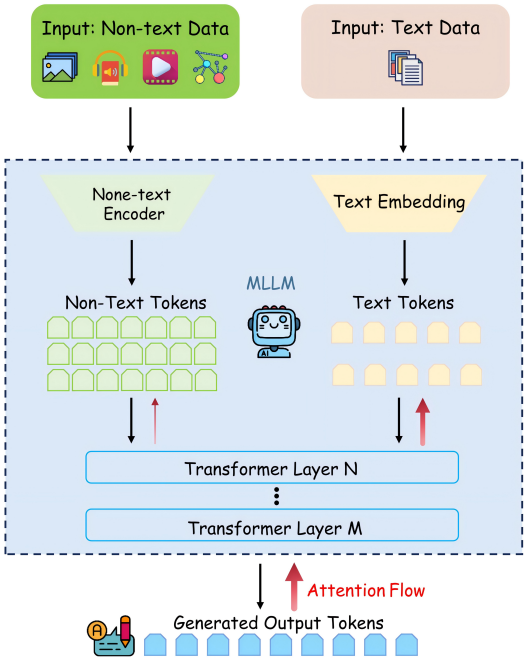

Figure 1: 每种模态都被分词并由 MLLM 联合处理。红色箭头说明了非文本、文本和生成输出标记之间的注意力机制。与非文本模态关联的较细箭头反映了它们较大的标记数量，从而导致每个标记的注意力权重较低。

基于这一基础，研究人员迅速扩展 MLLMs 的能力，从超越图像到捕捉视频中的时空动态。像 Video-LLaMA (Zhang et al. 2025) 及其后继者这样的模型整合了专门的组件，用于显式建模时间依赖性并融合视听信号。对于音频，像 Qwen-Audio (Chu et al. 2024) 这样的模型通过向量量化（VQ）进行标记化，将连续波形转换为与 LLMs 兼容的离散序列。这一探索进一步深入到序列和结构化数据。例如，像 Chat-TS (Xie et al. 2024) 这样的模型已经被开发出来，以通过将时间模式编码到 LLM 的潜在空间中来处理复杂的时间序列数据。在图结构数据领域，GraphGPT (Tang et al. 2024) 展示了 LLMs 通过将图结构翻译成 LLMs 可以处理的格式，从而理解和推理关系信息的潜力。

多模态大型语言模型中的模态失衡

模态失衡现象 (Prabhu 2025) 指的是模型倾向于过度依赖文本，同时未充分利用或完全忽略来自另一种模态的信息，如视觉模态。

模态不平衡的根源可以追溯到数据和模型架构。首先，固有的数据偏差是主要因素，因为与复杂的图像数据相比，文本的信息密度更高，为模型 (Park et al. 2025) 提供了一个可利用的捷径。其次，多模态语言模型 (MLLMs) 的架构设计系统地加剧了这种不平衡。大多数 MLLMs 展现出不对称的模态主干能力，将预训练于数万亿文本标记上的极其强大的语言模型与训练于相对较小规模数据上的视觉编码器相结合 (Li et al. 2023; Liu et al. 2023) 。

为了解决模态失衡，研究界从多个角度提出了缓解策略。一项研究工作集中在在数据层面对训练过程进行重新设计，以主动防止失衡。Data Remixing 框架 (Ma, Chen, and Deng 2025) 引入了一种两阶段的训练策略。它通过屏蔽更强的模态来进行样本级的解耦，从而迫使模型依赖较弱的模态，以对抗模态惯性。一种最近的方法，MBPO 框架 (Liu et al. 2025) ，直接针对模型对文本的过度依赖。它在对抗生成的“困难负样本”上使用直接偏好优化（DPO），迫使模型更倾向于视觉证据而非语言驱动的幻觉。

多模态大型语言模型中的文本主导性

概述

多模态大语言模型（MLLMs）的快速发展展示了它们在多模态理解和推理方面的显著能力。尽管理论上这些模型能够集成如文本、图像、视频、音频、时间序列数据和图表等模态的信息，但一个持续的挑战已经出现：在生成过程中，MLLMs 通常给予文本比非文本模态更大的权重。这一现象，被称为文本模态主导，表现为模型分配给文本内容的注意资源远远超过其他模态。

虽然这一现象主要在视觉-语言领域中被记录，但我们提出这种支配性也存在于视频、音频、时间序列和图形模态中。然而，缺乏系统的跨模态实证验证。

为了解决这个问题，我们提出了一系列利用生成式多模态语言模型（MLLMs）固有的交叉注意力机制的词元级别分析。具体来说，我们利用 MLLMs 在生成过程中的交叉注意力机制，定量分析输出词元和输入词元之间在不同模态下的注意力分布。

这使得直接的统计测量成为可能：我们比较分配给文本输入的注意力比例与分配给非文本输入的注意力比例。得到的指标提供了对文本模态主导性进行定量且可解释的评估。

为了构建一个全面和强健的评估框架，我们为五个关键模态选择了具有代表性的数据集和最新的模型，包括图像、视频、音频、时间序列和图形。对于图像模态，我们采用了 MMMU-Pro 基准测试 (Yue et al. 2024) ，它排除了仅通过文本即可回答的问题，以评估视觉-文本融合。我们在该任务上评估了三种最新的视觉-语言模型：Qwen2.5-VL-7B (Bai et al. 2025) 、LLaVA-1.5-7B (Liu et al. 2024a) 和 Kimi-VL-A3B-Instruct (Team et al. 2025) ，每个模型代表不同的多模态架构。

对于视频分析，MMBench-Video 基准 (Fang et al. 2024) 通过开放性问题评估 YouTube 长视频内容的时间推理能力。我们在这个基准上的评估包括两个不同的模型：Qwen2.5-VL-7B (Bai et al. 2025) ，这是一个从图像-文本基础上改编的通用模型，以及 VideoLLaMA3-7B (Zhang et al. 2025) ，这是一个专门优化用于视频中心任务的专业模型。

对于音频数据，IEMOCAP 数据集 (Busso et al. 2008) ，其包含多轮注释对话，用于测试 Qwen2-Audio-7B-Instruct (Chu et al. 2024) ，这是一种集成了语音编码的语言模型。

在时间序列中，我们评估了为多变量时间推理设计的 ChatTS-14B (Xie et al. 2024) ，在合成任务中，重点关注文本和时间序列数据之间的注意力平衡。

对于图数据，我们采用 GraphGPT-7B (Tang et al. 2024) 及其相应的基准测试 GraphGPT-eval-instruction。该框架通过一个两阶段的指令微调范式，将大型语言模型与图知识对齐。我们使用其指令集进行推理测试，以测量模型在图信息上的注意力分配。

评估指标

为了描述一个 MLLM 是如何在不同模态间分配其计算资源的，我们采用了两个互补的指标：模态优势指数（MDI）和注意力效率指数（AEI）。MDI 捕捉生成中的整体模态优势，而 AEI 则衡量每种模态相对于其令牌比例的注意力效率。

模态主导指数。MDI 量化了在自回归生成过程中多模态模型对文本输入与非文本输入的相对依赖性。对于包含一组文本标记 $\mathcal{T}$ 和一组非文本标记 $\mathcal{O}$ 的输入序列，我们首先计算针对每种模态的总注意力分数。设 A_T 和 A_O 为在生成 N 输出标记期间针对 $\mathcal{T}$ 和 $\mathcal{O}$ 中所有标记分别聚合的注意力分数，并进行了归一化，使得 $A_T + A_O = 1$ 。然后 MDI 被定义为两种模态之间平均每标记注意力的比率：

$$\mathrm{MDI} = \left(\frac{A_T}{|\mathcal{T}|}\right) \cdot \left(\frac{A_O}{|\mathcal{O}|}\right)^{-1} \tag{1}$$

因此，MDI 值大于 1 表示文本占优势；值小于 1 表示非文本占优势；而值接近 1 表示两者的影响平衡。

注意力效率指数。为了补充 MDI，我们引入了注意力效率指数（AEI），它考虑了每种模态所消耗的计算资源。尽管大多数现有指标侧重于绝对注意力主导性，但它们常常忽略了跨模态的令牌分配等成本。AEI 衡量了某个模态将其令牌表示转换为注意力的效率，为多模态生成中的资源使用提供了一种归一化的评估。

令 A_T 为文本标记的总注意力得分，A_O 为非文本标记的总注意力得分。文本模态所占注意力的比例 P_T 为：

$$P_T = \frac{A_T}{A_T + A_O} \tag{2}$$

给定 $|\mathcal{T}|$ 个文本标记和 $|\mathcal{O}|$ 个非文本标记，输入中的文本模态的比例大小 Q_T 为：

$$Q_T = \frac{|\mathcal{T}|}{|\mathcal{T}| + |\mathcal{O}|} \tag{3}$$

然后，文本模态的 AEI 被定义为其注意力份额与其标记份额的比率：

$$\mathrm{AEI}_T = \frac{P_T}{Q_T} = \frac{A_T/(A_T + A_O)}{|\mathcal{T}|/(|\mathcal{T}| + |\mathcal{O}|)} \tag{4}$$

AEI 值大于 1 表示高效率，表明该模式在其令牌分配中获得了不成比例的注意力突出。通过区分绝对支配性和资源效率，AEI 量化了某种模式如何有效利用其令牌表示来影响模型的注意力机制。

MDI 和 AEI 一起使我们能够将支配性和效率分开：MDI 评估哪种模式最终支配生成过程，而 AEI 评估一种模式如何有效地利用其有限的令牌预算来捕捉模型的关注点。

实验结果

为了量化在多模态大模型中的注意力分配，我们分析了不同模型层中的模态优势指数（MDI）和注意力效率指数（AEI）。如表 1 详细所示，我们的测量揭示了跨越各种模型、模态和基准的明显且一致的模式：无论层深如何，文本优势显而易见，尽管其程度有所不同，通常在某些任务的较深层中加强，而在其他任务中保持稳定或适中。

这种以文本为主导的层级趋势在图像和视频的主流模式中尤为明显。对于图像模式下的 Qwen2.5-VL-7B 模型，MDI 从早期层的 2.26 上升到后期层的 33.10。这表明在处理的后期阶段，分配给每个文本标记的平均注意力是图像标记的 33 倍以上。同时，AEI 从 14.24 下降到 1.42，说明了注意力分配的变化。在视频任务中，VideoLLaMA3-7B 在 MMBench-Video 基准测试中达到后期层的 MDI 为 157.53，表明文本标记吸引的注意力比视频帧标记多出两个数量级以上。

我们通过控制实验进一步调查了非文本信息量对注意力分配的影响。在音频和时间序列任务中，我们保持文本输入不变，同时将非文本标记序列复制了五倍和十倍。数据显示，这种输入规模的变化系统性地加剧了文本的主导性。对于 Qwen2-Audio-7B-Instruct，当复制因子增加时，晚层 MDI 从最初的 1.16 增加到 6.73 和 8.70。同样地，ChatTS-14B 的晚层 MDI 从 3.52 上升到 9.28 和 16.25。这些结果表明，随着输入中非文本标记的比例增加，模型对文本的相对关注也不成比例地增长。

相比之下，涉及图 modality 的任务则呈现出初步的例外情况。在标准条件下，对于 GraphGPT-7B 的晚期层 MDI 为 0.20，表明倾向于非文本的图 modality。然而，当非文本标记重复 10 倍时，MDI 上升到 1.35，超过了平衡阈值 1.0，表明转向文本 modality 主导。这表明即使在最初不倾向于文本的模型中，改变输入比例也可以产生这样的主导效应。

总之，我们对 MDI 和 AEI 的逐层评估证实了 MLLM 中文本模态主导现象的普遍存在。在许多任务中，这种主导性往往在更深层中得到加强，尽管这种模式因模态和输入条件而异。这种现象出现在诸如图像和视频等模态中，随着非文本令牌比例的增加而增强，甚至可能在最初偏向非文本模态的任务中出现，如在令牌复制下基于图的例子中所观察到的。这些发现为下一节中探索因果机制奠定了基础。

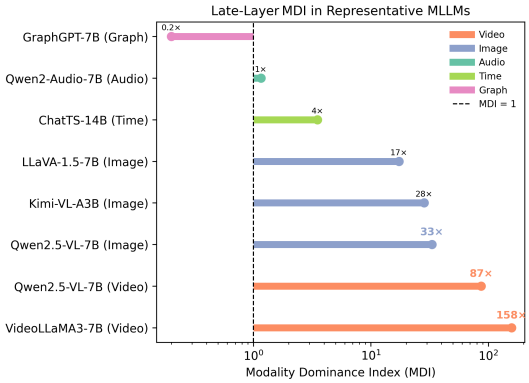

Figure 2: 跨不同模态 MLLMs 的文本主导现象。图中的虚线标记了 MDI = 1；位于其右侧的点展示了文本主导模式。图形模态落在此阈值的左侧，我们在分析部分提供了这一现象的专门解释。

文本主导的原因

多模态大型语言模型（MLLMs）在涉及图像、视频、音频和时间序列数据的任务中表现出色。然而，一种称为

Model	Modality	Dataset	Early		Middle		Late	
			MDI	AEI	MDI	AEI	MDI	AEI
Qwen2.5-VL-7B	Image	MMMU_Pro	2.26	14.24	21.12	10.86	33.10	1.42
Qwen2.5-VL-32B			3.84	2.82	54.96	21.88	26.03	13.95
Qwen2.5-VL-72B			9.33	6.15	92.21	60.43	24.46	14.60
LLaVA-1.5-7B			1.58	1.04	10.23	3.51	17.37	4.23
Kimi-VL-A3B-Instruct			2.27	3.91	3.78	2.99	28.39	2.59
Qwen2.5-VL-7B	Video	MMBench-Video	10.72	9.60	74.13	41.78	86.95	47.84
VideoLLaMA3-7B			19.14	17.90	140.10	73.75	157.53	76.26
Qwen2-Audio-7B-Instruct	Audio	IEMOCAP ×1	1.02	1.32	3.24	1.99	1.16	1.08
		IEMOCAP ×5	2.65	2.56	8.09	5.17	6.73	4.31
		IEMOCAP ×10	2.80	2.50	10.10	5.46	8.70	5.09
ChatTS-14B	Time-series	TimeSeries-Reasoning ×1	1.52	1.19	4.37	1.40	3.52	1.37
		TimeSeries-Reasoning ×5	2.08	1.95	10.72	3.15	9.28	3.03
		TimeSeries-Reasoning ×10	2.36	2.67	20.70	5.37	16.25	5.13
GraphGPT-7B	Graph	GraphGPT-Eval-Instruction ×1	0.14	0.84	0.14	0.84	0.20	0.90
		GraphGPT-Eval-Instruction ×5	0.20	0.69	0.35	0.83	0.69	0.98
		GraphGPT-Eval-Instruction ×10	0.31	0.71	0.68	0.97	1.35	1.14

Table 1: 在不同模型、模态和基准中对模态支配指数（MDI）和注意力效率指数（AEI）的比较分析。符号“× n ”代表应用于非文本模态中的标记的复制因子。“早期”、“中期”和“晚期”分别表示从前两层、中间两层和最后两层模型层中聚合的统计数据。

文本主导的现象不断出现：在推理过程中，MLLMs 倾向于过度强调文本标记，而未充分利用非文本模态。图 2 显示，这种模式在各种模型架构中出现。虽然以往的研究将这种现象归因于在预训练过程中引入的固有模态先验或对齐偏见，我们提出了一种不同的解释。我们的研究结果表明，文本主导不是静态模态偏好的反映，而是跨模态的标记级别不平衡导致的一种动态结果，我们称之为注意力稀释。

标记冗余导致注意力稀释

为了系统地调查 MLLM 中文本占优势的根本原因，我们对在广泛采用的多模态架构中，编码阶段出现的标记数量增加及其导致的注意力稀释现象进行了深入分析。我们的研究表明，非文本模态存在冗余标记，降低了它们在跨模态注意力计算中的有效性。

具体而言，视频输入被处理为扩展的帧序列，而音频和时间序列数据通常被划分为许多补丁或时间段。这样的预处理步骤不可避免地导致非文本模态的标记数量显著增加。因此，这些标记往往高度冗余，相对较低的语义密度。相比之下，文本标记在语义上是紧凑的，并包含集中的语义信息。

由于这种不平衡，注意力机制倾向于优先对文本标记赋予权重，导致非文本模态的注意力稀释现象明显。例如，在 MMBench-Video 基准上，Video-LLaMA3-7B 在模型的后期层显示出 157.53 的模态主导指数（MDI），这表明每个文本标记在生成过程中所得到的注意力权重平均超过 157 倍于单个视频帧标记所分配的权重。相应地，注意力效率指数（AEI）达到 76.26，突出显示文本标记虽然只占总输入的一小部分，却获得了模型注意力分配中不成比例的大份额。这揭示了多模态大语言模型中的注意力分配不平衡：即使非文本输入构成了标记的多数，模型在推理过程中仍主要依赖于文本信息。因此，视频帧和其他非文本标记在竞争性注意力机制下被有效边缘化，这可能限制模型充分利用多模态信息的能力。

除了输入模态的标记结构外，架构设计对注意力分布及哪个模态在推理时占主导地位具有关键影响。正如图 3 所示，我们对两种具有代表性的视觉-语言多模态模型之间的 MDI 和 AEI 进行了比较分析。

LLaVA-1.5 7b 使用一个浅层桥接架构，其中包含一个冻结的视觉编码器和线性投影模块，其中视觉任务的 MDI 从早期层的 1.58 上升到后期层的 17.37。相比之下，Qwen2.5-VL 采用了更为集成的融合机制，该机制结合了视觉 Transformer 编码器和基于 MLP 的视觉-语言合并模块，在相应阶段表现出了明显更高的模态优势指数，高达 33.1。这表明，较深的融合机制可以在一定程度上增强文本模态的主导性。

然而，从 AEI 的角度来看，LLaVA-1.5 保持了相对较高且不断增长的 AEI，从 1.03 上升到 4.23，而 Qwen2.5-VL 的 AEI 则表现出持续下降的趋势，从早期层的 14.24 下降到后期层的 1.42。该现象突出了一个值得注意的权衡：复杂的架构虽然可能增强文本控制，但可能损害整体注意力利用效率。相反，在资源分配受限的情况下，较简单的架构鼓励更有效地使用文本输入，从而在控制与注意力效率之间实现一种新颖的平衡。这些见解为未来的模型设计提供了宝贵的指导，强调了需要在增强模态表示能力与优化注意力资源分配之间寻找平衡。

此外，除了架构和表示因素之外，任务设计本身也可以深刻影响注意力在不同模态之间的分配。在某些任务中，注意力向文本模态的转移不仅仅源于输入表示的差异，而更为根本地来自于任务构建中嵌入的文本提示的结构性依赖。例如，在时间序列任务中，关键的归一化因子和任务特定的元数据通常被编码在自然语言指令中，从输入阶段就确立了文本模态的逻辑主导地位。同样，在情绪识别或关键词配对等与音频相关的任务中，任务目标通常由文本提示引导，使文本模态成为语义和推理的核心。

为了进一步验证这一现象，我们分析了模型 ChatTS-14B 和 Qwen2-Audio-7B 在不同非文本标记复制（×1、×5、×10）水平下的表现，研究了它们的模态专注分布，如图 4 所示。值得注意的是，即使在没有扩展非文本标记的情况下（×1 配置），文本模态在注意力分配方面始终表现出明显的优势：ChatTS-14B 在后期层实现了

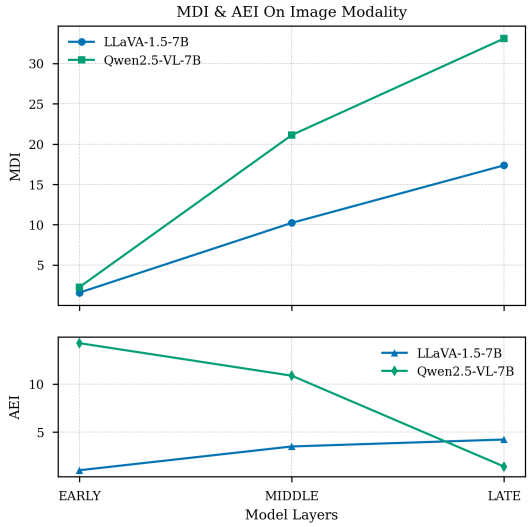

Figure 3: 不同阶段中 LLaVA-1.5-7B 和 Qwen2.5-VL-7B 在图像模态上的 MDI 和 AEI 比较。

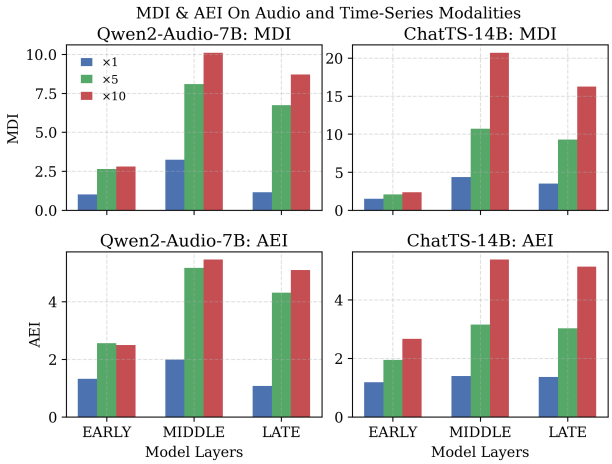

Figure 4: 音频和时间序列模型中的 MDI 和 AEI 与标记缩放。

1.37 的注意力效率指数（AEI），而 Qwen2-Audio-7B 在同一阶段达到 1.08 的 AEI。

随着非文本标记数量的增加，文本模态的主导地位不仅持续存在，且变得越来越明显。具体来说，对于 ChatTS-14B 而言，在单次复制设置下，中间层的 MDI 显著上升，从 4.37 增长到五倍复制下的中间层的 10.72，并在十倍复制的后期层中进一步激增至 20.70。相应地，其 AEI 在单次复制配置的后期层中从 1.37 增加到五倍和十倍复制的后期层分别达到 3.03 和 5.13。类似的模式也出现在 Qwen2-Audio-7B 中，其中 MDI 在单次复制的中间层从 3.24 上升到十倍复制的中间层的 10.10。同时，其 AEI 在单次复制设置的后期层从 1.08 上升到十倍复制的中间层的 5.09。

这些研究结果由观察到的趋势支持，提供了强有力的证据表明，在高度依赖文本提示的任务中，即使非文本模式更多，模型仍然始终优先将注意力分配给文本标记。这些结果表明，文本提示在引导注意力和推理方面在多模态语言模型中起着关键作用。

图形任务中的模态优势转移

与多模态模型中普遍观察到的文本模态主导趋势相反，GraphGPT 在与图相关的任务上表现出显著的例外。当图输入相对较小且图标记的数量远小于随附文本提示的数量时，模型的 MDI 初始测量值为 0.20。该低值表明，在这种配置中，平均来说，图标记比文本标记吸引了更多的注意力。同时，文本模态的 AEI 保持在 0.90，表明在这些条件下，文本输入既不占主导地位也不是特别有效地获得注意资源。

在这种情况下，模型自然地将更多注意力分配给信息密集的图形标记，反映出对具有更高语义紧凑性的输入的内在偏好，而不论模态如何。为了进一步探究这一行为，我们通过复制系统地增加图形标记的数量——将其扩展到 5 倍和 10 倍，而不改变其语义内容。结果是，MDI 从 0.20 增加到 1.35，文本 AEI 从 0.90 增加到 1.14。这些变化表明模态主导性从图形向文本的逐渐转变，同时文本标记的注意力效率也相应提高——从低于基线到高于基线。

这种受控调制为我们的核心假设提供了令人信服的实证支持：模态优势不是由预训练编码的固定特征，而是由输入的结构和统计驱动的动态响应。模型对不同模态的注意力分配主要受令牌数量和信息密度的支配，而不是任何静态或特定模态的先验。在这种情况下，观察到的模态偏好是输入诱导和上下文敏感的结果，而不是不可变的结构偏差。

文本主导性的标记压缩

基于注意力稀释现象的发现，我们提出了当前架构的优化策略，以重新平衡模态整合。我们的结果表明，当多模态信息与文本输入结合时，文本的主导性往往会加强。例如，在 LLaVA-1.5-7B 中，在后期层，MDI 上升到 17.37，表明每个文本标记平均接收到的注意力是单个视觉标记的 17 倍以上。这突出了标记利用的不平衡：虽然文本输入在标记数量相对较少的情况下仍然保持语义密集，但一幅图像通常由数百个视觉标记表示，其中许多是冗余的或信息价值低的。

为了应对文本模态的主导性，我们在最新的研究基础上利用源自冻结视觉编码器的 [CLS] 标记注意力机制 (Zhang et al. 2024)，作为视觉标记修剪的更可靠指标。[CLS] 标记旨在通过自注意力捕捉图像的全局语义，并提供稳定的视觉标记显著性评估，这种评估在网络层中保持一致。形式上，给定 N 视觉标记 $V = \{v_1, \ldots, v_N\}$，由视觉变压器编码，我们计算每个标记 v_i 的重要性得分 s_i，算法如下

$$s_i = \mathrm{Attn}\big([\mathrm{CLS}], v_i\big). \tag{5}$$

然后，通过应用标记减少率 r，仅保留得分最高的顶

$$M = N(1 - r) \tag{6}$$

标记，形成压缩序列

$$V' = \{v'_1, \ldots, v'_M\}. \tag{7}$$

这种以 [CLS] 为导向的压缩策略通过减少非文本输入 $|\mathcal{O}|$ 的基数直接缓解了注意力稀释，从而重新平衡跨模

Method	Reduction	Early		Middle		Late	
		MDI	AEI	MDI	AEI	MDI	AEI
LLaVA-1.5-7B	0 %	1.58	1.04	10.23	3.51	17.37	4.23
FasterVLM	75 %	0.57	0.71	1.81	1.33	3.39	1.64
	90 %	0.57	0.80	1.10	1.03	1.84	1.17
	95 %	0.48	0.82	0.86	0.97	3.39	1.64

Table 2: 令牌缩减比对模态优势指数 (MDI) 和注意效率指数 (AEI) 的影响。统计数据报告了前两层（早期）、中间两层（中期）和最后两层（晚期）的结果。

态的注意力分配。根据给定的计算预算 R ，自适应地确定修剪阈值 τ ，如下所示：

$$\tau = \min\left\{\tau \mid \left|\left\{a \in a_{[\mathrm{CLS}]} \mid a \geq \tau\right\}\right| \leq N \times (1-R)\right\} \tag{8}$$

其中 $a_{[\mathrm{CLS}]}$ 表示来自 [CLS] 标记的注意力得分。

我们在 LLaVA-1.5-7B 模型上使用 MMMU Pro 基准进行了实验，评估了在不同压缩率：0 %，75 %，90 % 和 95 % 下的早期、中期和后期网络层中的 MDI 和 AEI。结果以方法名 FasterVLM 报告，该方法在融合之前应用 [CLS] 引导的 token 修剪以减少冗余可视 tokens。如表 2 所示，将压缩率从 0 % 增加到 90 % 导致后期层 MDI 显著下降，从 17.37 降至 1.84。这有效地缓解了文本模态主导性，并使注意力分布更趋于平衡，因为 MDI 接近于一。此结果表明，压缩非文本输入 tokens 使模型更好地利用视觉信息。

进一步分析表明，随着 MDI 的减少，文本模态的 AEI 也从 4.23 下降到 1.17。这表明从对文本输入的强烈依赖转向了对不同模态的更平衡整合。这些结果支持我们的主要假设，即文本的主导地位可以通过调整输入结构来影响。通过以适当的方式减少非文本标记的数量，可以重新分配模型的关注点，从而实现更平衡的多模态推理。

此外，我们的工作扩展了先前研究的范围 (Zhang et al. 2024) ，表明分词压缩技术不仅提高了计算效率，还在减轻文本模态的主导性方面发挥了重要作用。总的来说，这些结果为平衡模态集成提供了实用策略，并为多模态大模型中注意力分布机制提供了更清晰的表征。

结论

在这项工作中，我们系统性地研究了多模态大型语言模型中的文本主导现象。我们引入了两个指标，模态主导指数（MDI）和注意力效率指数（AEI），用于测量和分析注意力在不同输入模态之间的分配情况。对图像、视频、音频、时间序列和图形数据的实验表明，文本模态主导在当前模型中很常见。我们还发现，压缩非文本标记可以缓解这种不平衡，并促进更公平的多模态整合。这些结果提供了宝贵的工具和指导，以构建更有效和平衡的多模态模型。

未来的工作将探索额外的策略，例如通过架构重组来促进更集成的模态融合，以及任务重构来减少对文本提示的过度依赖。这些方法将被系统地研究，以评估它们的有效性及其与标记压缩的潜在协同作用，旨在推进稳健和平衡的多模态基础模型的开发。通过这些方法，我们旨在减轻文本主导性，并最大化多模态信息的利用。

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4): 335–359.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759.

Fang, X.; Mao, K.; Duan, H.; Zhao, X.; Li, Y.; Lin, D.; and Chen, K. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. Advances in Neural Information Processing Systems, 37: 89098–89124.

Kumar, P. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. Artificial Intelligence Review, 57(10): 260.

Leng, S.; Xing, Y.; Cheng, Z.; Zhou, Y.; Zhang, H.; Li, X.; Zhao, D.; Lu, S.; Miao, C.; and Bing, L. 2024. The Curse of Multi-Modalities: Evaluating Hallucinations of Large Multimodal Models across Language, Visual, and Audio. arXiv.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, 19730–19742. PMLR.

Liang, Z.; Xu, Y.; Hong, Y.; Shang, P.; Wang, Q.; Fu, Q.; and Liu, K. 2024. A survey of multimodel large language models. In Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering, 405–409.

Liu, C.; Xiong, T.; Chen, R.; Wu, Y.; Guo, J.; Zhou, T.; and Huang, H. 2025. Modality-Balancing Preference Optimization of Large Multimodal Models by Adversarial Negative Mining. arXiv preprint arXiv:2506.08022.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In Proceedings

of the IEEE/CVF conference on computer vision and pattern recognition, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. Advances in neural information processing systems, 36: 34892–34916.

Liu, Y.; Bai, G.; Chenji, L.; Li, S.; Zhang, Z.; Liu, R.; and Guo, W. 2024b. Eliminating the Language Bias for Visual Question Answering with fine-grained Causal Intervention. In 2024 IEEE International Conference on Multimedia and Expo (ICME), 1–6. IEEE.

Luo, G.; Yang, X.; Dou, W.; Wang, Z.; Liu, J.; Dai, J.; Qiao, Y.; and Zhu, X. 2025. 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, 24960–24971.

Ma, X.; Chen, H.; and Deng, Y. 2025. Improving Multimodal Learning Balance and Sufficiency through Data Remixing. arXiv preprint arXiv:2506.11550.

Park, J.; Jang, K. J.; Alasaly, B.; Mopidevi, S.; Zolensky, A.; Eaton, E.; Lee, I.; and Johnson, K. 2025. Assessing modality bias in video question answering benchmarks with multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 19821–19829.

Prabhu, Y. 2025. Unveiling Bias in Multimodal Models. Ph.D. thesis.

Qin, L.; Chen, Q.; Zhou, Y.; Chen, Z.; Li, Y.; Liao, L.; Li, M.; Che, W.; and Yu, P. S. 2025. A survey of multilingual large language models. Patterns, 6(1).

Tang, J.; Yang, Y.; Wei, W.; Shi, L.; Su, L.; Cheng, S.; Yin, D.; and Huang, C. 2024. Graphgpt: Graph instruction tuning for large language models. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 491–500.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. arXiv preprint arXiv:2504.07491.

Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; et al. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. arXiv preprint arXiv:2411.10442.

Xie, Z.; Li, Z.; He, X.; Xu, L.; Wen, X.; Zhang, T.; Chen, J.; Shi, R.; and Pei, D. 2024. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. arXiv preprint arXiv:2412.03104.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. National Science Review, 11(12): nwae403.

Yue, X.; Zheng, T.; Ni, Y.; Wang, Y.; Zhang, K.; Tong, S.; Sun, Y.; Yu, B.; Zhang, G.; Sun, H.; et al. 2024. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. CoRR.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106.

Zhang, Q.; Cheng, A.; Lu, M.; Zhuo, Z.; Wang, M.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024. [CLS] Attention is All You Need for Training-Free Visual Token Pruning: Make VLM Inference Faster. arXiv e-prints, arXiv–2412.