一项关于视频大语言模型如何回答视频问题的经验研究

Chenhui Gou[1], Ziyu Ma[2, 1], Zicheng Duan[3], Haoyu He[1], Feng Chen[3], Akide Liu[1], Bohan Zhuang[4], Jianfei Cai[1], Hamid Rezatofighi[1]

[1]Monash University [2]Hunan University [3]The University of Adelaide [4]Zhejiang University

Abstract

利用大规模数据和预训练语言模型，视频大型语言模型（Video-LLMs）在回答视频问题方面表现出了强大的能力。然而，大多数现有的研究集中在提高性能上，对理解其内部机制的注意力有限。本文旨在通过系统性实证研究来弥补这一差距。为了解释现有的 VideoLLMs，我们采用注意力敲除作为主要分析工具，并设计了三种变体：视频时间敲除、视频空间敲除和语言到视频敲除。然后，我们在不同数量的层（窗口）上应用这三种敲除方法。通过仔细控制层窗口类型和类型的敲除方式，我们提供了两种设置：全局设置和细粒度设置。我们的研究揭示了三个关键发现：(1) 全局设置表明视频信息提取主要发生在早期层中，形成一个清晰的两阶段过程——较低层关注感知编码，而较高层处理抽象推理；(2) 在细粒度设置下，某些中间层对视频问题回答的影响尤为巨大，作为关键异常值发挥作用，而大多数其他层贡献较小；(3) 在这两种设置中，我们观察到时空建模更多依赖于语言引导的检索而非视频令牌之间的帧内和帧间自注意力，尽管后者计算成本较高。最后，我们展示了这些见解可以用来减少 Video-LLMs 中的注意力计算。据我们所知，这是首次系统地揭示 Video-LLMs 如何内部处理和理解视频内容的工作，为未来的研究提供了可解释性和效率视角。

介绍

视频大型语言模型（视频-LLMs）最近展示了它们强大的理解视频内容和回答各种类型问题的能力 (Zhang et al. 2024c,d; Yao et al. 2024; Bai et al. 2025; Chen et al. 2024b; Cheng et al. 2024)。最近关于视频-大语言模型的研究主要集中在提高模型性能上，例如，提升视频指令数据集的质量和规模 (Zhang et al. 2024d; Yao et al. 2024; Bai et al. 2025; Chen et al. 2024b; Cheng et al. 2024)，延长输入帧长度 (Zhang et al. 2024a; Xue et al. 2024; Liu et al. 2024b)，以及优化视频标记的位置编码 (Liu et al. 2025; Ge et al. 2024; Wei et al. 2025)。然而，对于它们内部机制的理解仍然有限——特别是它们如何处理和推理视频内容。深入了解这些机制对于增强解释性、提高模型效率和促进未来模型发展至关重要。

在图像领域，许多研究人员试图提高这些大型多模态模型的可解释性以避免纯粹的黑盒使用。这些工作研究了 MLLMs 的内部状态如何与其外部行为对应 (Zhang et al. 2025; Kaduri, Bagon, and Dekel 2024; Basu et al. 2024; Zhao et al. 2024; Zhang et al. 2024b)。这包括从图像到不同阶段模式形成的信源流分析 (Zhang et al. 2025; Chen et al. 2024a; Lin et al. 2025)、逻辑分布中不希望的内容生成模式 (Zhao et al. 2024)、两阶段模式和安全机制精细化 (Xu et al. 2024)、与对象相关视觉线索的基础和演变 (Neo et al. 2024; Schwettmann et al. 2023; Ma et al. 2024)、信息在模型参数中的存储 (Basu et al. 2024)，以及冗余视觉标记的减少 (Zhang et al. 2024b)。相比图像领域 MLLMs 可解释性的丰富研究，在高维视频领域仍很大程度未被探索。最近的一项工作 (Xiao et al. 2025) 进行了全面的实验来分析各种视频-大语言模型的行为，并报告了几项有趣的观察结果：这些模型在视频问答方面表现出色，但在时间定位上表现不佳；它们对语言变化非常敏感，而对视频扰动则不太敏感。虽然这项工作主要侧重于分析模型的外部行为，但其内部透明度仍然很大程度上未被探索。

我们的工作朝着揭示现有视频大语言模型内部模式的方向迈进，并理解这些模式如何与其在视频问答中的出色表现相关联。现代视频大语言模型 (Zhang et al. 2024a; Wang et al. 2025; Li et al. 2024a; Zhang et al. 2024c) 通常遵循类似的架构：预训练的视觉编码器将视频转换为标记，投影层将这些标记映射到语言空间中，以及一个大型语言模型（LLM），该模型接收视频和语言标记以生成响应。每个视频标记由 LLM 中的每一层进行处理，并通过注意力机制与其他视频标记和问题标

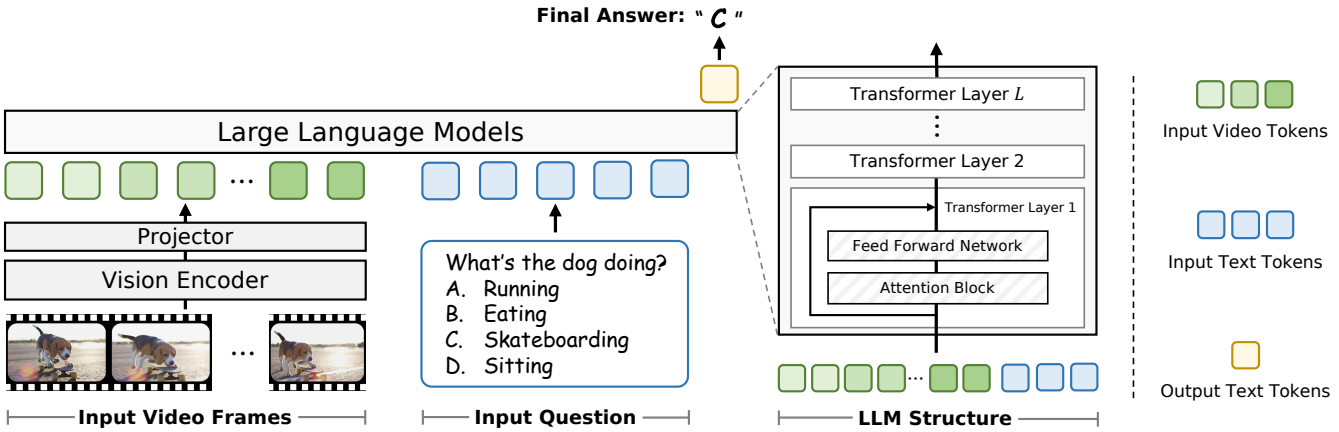

图 1: 典型的 Video-LLMs 架构设计。

记相互作用。这有助于信息在各层之间的提取和传播，最终对最终答案有所贡献。在每一层内部，注意力机制可以分解为三种类型：时间注意力（跨视频帧），空间注意力（每个帧内）和语言到视频的注意力。我们设计了三种对应的关注点敲除方法，选择性地禁用一种特定类型的注意力，使我们能够分析其单独的影响。为了获得不同粒度级别的见解，我们在两种设置下研究了视频大语言模型的内部模式：*全局设置*和*细粒度设置*，通过仔细控制敲除层范围和注意力类型来实现。在全局设置中，我们探讨两个问题：（1）*视频 LLM 是否表现出与图像级 VLM 类似的两阶段行为——即早期层的感知编码和后期层的语义推理——还是遵循一个不同的范式？*（2）*在全球范围内，每种注意力类型如何贡献于视频问答性能？*在细粒度设置中，我们探索：（3）*每种注意力类型如何影响不同层的视频问答（VideoQA）？*

我们分析了一系列具有代表性的视频-大语言模型，包括 LongVA(Zhang et al. 2024a)、InternVideo2.5(Wang et al. 2025)、LLaVA-OneVision(Li et al. 2024a) 和 LLaVA-Video(Zhang et al. 2024c)，在主流的视频问答基准测试上进行评估，这些基准测试涵盖了多种任务类型和视频长度。这包括短时多任务视频问答：MVBench(Li et al. 2024b)、中时自我中心视角视频问答：EgoSchema(Mangalam, Akshulakov, and Malik 2023) 以及长时开放领域视频问答：Video-MME(Fu et al. 2024)。通过包含超过 300 个数据点的广泛实验，我们总结了关键发现：(1) 全局设置下的观察：从一定比例的层（例如，整个模型的 60%）开始应用语言到视频注意力剔除，并未显示出对各种基准和模型性能有显著影响。(2) 全局设置下的观察：应用完全的语言到视频注意力剔除导致了明显的性能下降，这种下降幅度远大于时间或空间剔除所造成的，如 fig. 5 所示。(3) 细粒度设置下的观察：对于大多数单独的层而言，语言到视频注意力剔除的影响强于时间或空间注意力剔除，如 fig. 6 所示。(4) 细粒度设置下的观察：剔除特定层（例如，第 12 至 16 层）导致了显著的性能下降，而大多数其他层则影响甚微，如 fig. 6 所示。

这些结果揭示了关于 VideoQA 的以下见解：(a) 视频大语言模型表现出明显的两阶段处理模式，其中早期层主要关注提取视频信息。(b) 当前的视频大语言模型严重依赖于语言到视频的关注机制来检索和建模视频内容，而不是依赖计算成本更高的时间和空间上的视频注意力。(c) 少数几层在 VideoQA 中发挥着关键作用，并作为注意力路径中的重要异常值出现。最终，我们应用了早期退出的视觉令牌策略，在达到特定比例的层数后丢弃所有视觉令牌，这直接基于我们的两阶段发现。我们发现这种简单的策略可以显著减少计算开销，仅对性能产生极小的影响。据我们所知，这是首次研究在 VideoQA 背景下视频大语言模型处理和理解视频的内部模式的工作。我们的发现增强了视频大语言模型的可解释性，并为它们未来的发展提供了有价值的见解。

相关工作

视频大型语言模型（Video-LLMs）。现有的基于 LLM 的视频理解方法可以分为两大类：(1) **专业视频** LLM。这些方法利用冻结的 LLMs（*例如.*，LLaMA3(Dubey et al. 2024)，Mistral(Jiang et al. 2023)，和 Qwen(Yang et al. 2024)）或通用图像 MLLMs（*例如.*，BLIP-2(Li et al. 2023)，LLaVA-Next(Liu et al. 2024a)，LLaMA-Adapter(Zhang et al. 2023)），同时在特定的视频理解数据集上微调适配器或 LoRA(Hu et al. 2022) 模块。代表

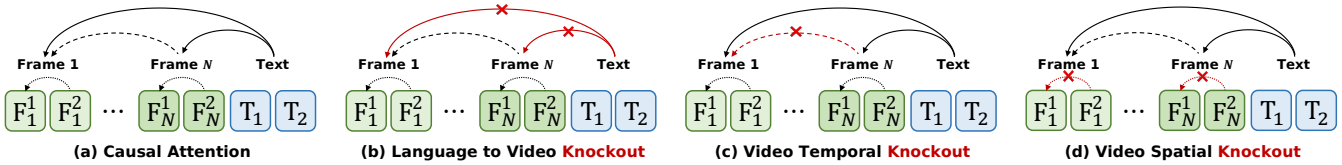

图 2: 默认因果注意力和三种类型的注意力剔除机制。为了清晰起见，我们使用两个标记（例如.,F_n^1, F_n^2）来可视化每个帧 $F_{1...N}$ 和所有文本提示 T（实际上每个帧的标记数量大于 100，而所有文本提示包含少于 100 个标记）。(a) 原始因果注意力：大型语言模型中的原始因果注意力机制。(b) 文本到视频剔除：该机制移除了从文本提示到视频帧中所有标记的注意力。(c) 视频时间剔除：此方法阻止了帧之间的时序注意力，同时保留了文本到视频和帧内空间注意力。(d) 视频空间剔除：这种设置禁用了视频帧内的空间注意力。

性作品包括 FrozenBiLM(Yang et al. 2022)、SeViLA(Yu et al. 2023)、LLaMA-VQA(Ko et al. 2023) 和 Video-LLaMA(Zhang, Li, and Bing 2023)。(2) **通用视频大语言模型**。这些方法没有在特定的视频理解数据集上进行微调，而是预先在大规模视频数据集上进行了训练，从而使它们能够处理诸如图像/视频问答、检索和字幕等多模态任务。它们通常增强视频指令数据集的质量和规模（例如.，LLaVA-NeXT-Video(Zhang et al. 2024c)，LLaVA-Video(Zhang et al. 2024d)，MiniCPM-V(Yao et al. 2024)，Qwen2.5-VL(Bai et al. 2025)，InternVL2.5(Chen et al. 2024b)，InternVideo2.5(Wang et al. 2025)，LLaVA-OneVision(Li et al. 2024a)，和 VideoLLaMA2(Cheng et al. 2024)），延长输入帧的长度（例如.，LongVA(Zhang et al. 2024a)，LongVILA(Xue et al. 2024)，和 Kangaroo(Liu et al. 2024b)），并优化视频标记的位置编码（例如.，VRoPE(Liu et al. 2025)，V2PE(Ge et al. 2024)，和 VideoRoPE(Wei et al. 2025)）以提高视频理解能力并取得显著性能。在本文中，我们考察了最先进的广泛使用的开源视频大语言模型，包括 LongVA、InternVideo2.5、LLaVA-OneVision (Li et al. 2024a) 和 LLaVA-Video (Zhang et al. 2024d)，以确保不同视频大语言模型的代表性生成能力。

多模态模型的可解释性。多模态模型的可解释性已成为关键的研究重点。现有方法可以大致分为三类：(1) 黑盒分析 (Cao et al. 2020; Frank, Bugliarello, and Elliott 2021)，通过分析输入-输出关系来理解模型行为，包括评估各种模态 (Cao et al. 2020) 的重要性及其对任务的贡献 (Frank, Bugliarello, and Elliott 2021)；(2) 单样本归因 (Aflalo et al. 2022; Chefer, Gur, and Wolf 2021; Lyu et al. 2022; Stan et al. 2024)，使用注意力得分聚合 (Aflalo et al. 2022; Stan et al. 2024)、基于梯度的方法 (Chefer, Gur, and Wolf 2021) 或模型解缠 (Lyu et al. 2022) 将预测追溯到特定输入；(3) 自上而下的表征探测 (Lindström et al. 2021; Hendricks and Nematzadeh 2021; Salin et al. 2022)，通过调查学习到的表征来揭示高级概念，如视觉语义 (Lindström et al. 2021)、动词理解 (Hendricks and Nematzadeh 2021) 和形状或大小 (Salin et al. 2022)。与这些方法不同，我们的研究调查了 Video-LLMs 在处理视频问答任务中的内部处理机制。

MLLMs **的机制可解释性**。在图像层面的理解中，一些早期研究已经开始通过将外部行为与特定机制联系起来来探讨 MLLMs 的内部状态。这些方面包括模型参数内的信息保留 (Basu et al. 2024)，反映在初始标记 logits 分布中的无意内容生成 (Zhao et al. 2024)，对象相关视觉信息的跟踪和转换 (Neo et al. 2024; Schwettmann et al. 2023)，安全机制的检测 (Xu et al. 2024)，以及冗余视觉标记的最小化 (Zhang et al. 2024b)。最近的一项研究 (Xiao et al. 2025) 进行了全面的实验，以分析各种视频 LLM 的行为，并表明这些模型在视频问答方面表现出色，但在时间定位上有所欠缺。此外，它们对语言变化敏感，而受视频扰动的影响较小。虽然这项工作主要集中在分析模型的外部行为上，但其内部透明度仍很大程度上未被探索。我们的工作提供了一项重要的初步努力来填补这一空白，并作为补充研究 (Xiao et al. 2025)。

调查设计

初步的

一个 Video-LLM(Liu et al. 2024a) 通常由预训练的视觉编码器、投影层和仅解码器的语言模型组成，如图 1所示。视觉编码器从视频输入中提取视觉标记，投影层将它们映射到语言空间。然后仅解码器 LLM(Yang et al. 2024) 采用视频和文本标记，并以自回归方式输出生成的响应标记。具体而言，一段视频序列以 N 帧进行采样，其中每一帧通过视觉编码器（例如., CLIP-L-14(Sun et al. 2023)）被编码为一串视觉标记 $\boldsymbol{F}_i$。这

些视觉标记随后通过一个投影层映射到文本空间。数学上，这个过程可以表示为

$$\boldsymbol{V} = [\boldsymbol{F}_i]_{i=1}^{N}, \quad \boldsymbol{F}_i = \text{Proj}(\text{Enc}_v(\boldsymbol{x}_i)) \in \mathbb{R}^d, \quad (1)$$

其中 $\boldsymbol{x}_i$ 代表第 i 帧视频，$\text{Enc}_v(\cdot)$ 表示视觉编码器，$\text{Proj}(\cdot)$ 是投影层，而 d 是在大语言模型中的隐藏维度的数量。类似地，文本输入通过预训练的嵌入代码簿映射到文本嵌入标记 $\boldsymbol{T}$。生成的文本标记与视频标记连接起来形成多模态输入序列：$\text{MMs} = [\boldsymbol{F}_1, \ldots, \boldsymbol{F}_N, \boldsymbol{T}]$，其中 $\boldsymbol{T}$ 表示文本标记。请注意，此序列为有序排列，先以视频标记开始，然后是文本标记。我们忽略前缀系统标记，这些标记用于控制 LLM 的输出行为。**隐藏表示和注意力**。多模态输入序列随后通过 L 层变换器块 (Vaswani et al. 2017) 来获取隐藏表示。每个变换器块由一个注意力块和一个前馈网络（FFN）组成。每层的隐藏表示 ℓ 可以写成

$$MMs^{(\ell)} = \boldsymbol{FFN}(\boldsymbol{Attn}(MMs^{(\ell-1)})), \quad (2)$$

其中 Attn 表示注意力块。这里为了简化符号，忽略了残差连接。LLM 最终层的最后一个标记用于解码输出标记。要理解视频，文本标记需要从视频中提取时空信息。此外，通过注意力机制，视频标记会在每个帧内以及跨帧进行相互交流。

注意. 在每个注意力块中，多模态序列（MMs）被投影到查询、键和值空间以获得 Q、K 和 V 矩阵。MMs 中的每个标记通过因果注意力 (Yang et al. 2021) 与其他所有标记交换信息。

$$\text{CausalAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V \quad (3)$$

这里，$\sqrt{d_k}$ 是缩放因子，M 是一个强制因果关系的掩码矩阵，确保 token 只能关注之前的或当前的 token。具体而言，M 定义为

$$M_{ij} = \begin{cases} 0, & \text{if } j \le i \\ -\infty, & \text{if } j > i. \end{cases} \quad (4)$$

这种因果注意力如图 fig. 2(a) 所示。

三种注意力敲除类型。

注意力敲除方法 (Geva et al. 2023) 是大型语言模型可解释性领域中广泛使用的一种技术。它通过阻断特定标记之间的注意力连接来研究这些特定信息流的影响。在详细说明当前 Video-LLMs 中的注意力计算之后，我们将整体因果注意力分解为三个部分：语言到视频、视频时间注意力和视频空间注意力。相应地，我们引入了三种类型的注意力敲除来阻止特定的注意力流动，即.，语言到视频敲除（LV-K）、视频时间敲除（VT-K）和视频空间敲除（VS-K），分别如 fig. 2（b）、（c）和（d）所示。语言到视频敲除禁止信息从视频流向语言。视频时间敲除阻止视频帧之间的信息交换，而视频空间敲除则防止在每一帧内的注意力流动。

调查设置

我们采用了不同的设置来进行研究。具体来说，我们考虑两个自由度：应用注意力剔除的层和相应的剔除类型（KT）。对于一个 L 层的 Transformer，每个层的剔除配置 L_i^{KT} 有四种可能的选择：

$$L_i^{\text{KT}} \in \{\text{no knockout, LV-K, VT-K, VS-K}\} \quad (5)$$

其中 $i \in \{1, \ldots, L\}$。这里，不进行剔除表示原始的注意力操作。通过仔细控制这两个变量，我们定义了以下三种研究设置来探讨关于 Video-LLMs 内部机制的三个问题。

全局设置 1. 我们将语言到视频的注意力在某个截止深度 $i \in \{1, 3, 5, \ldots, L\}$ 之外进行阻断。每层 j 的注意力配置定义为：

$$L_j^{\text{KT}} = \begin{cases} \text{no knockout}, & j \le i \\ \text{LV-K}, & j > i \end{cases} \quad (6)$$

这意味着模型只能通过前 i 层访问视频信息。当 $i = L$ 时，不应用任何阻断，作为基线。我们以步长为 2 变化 i 来追踪性能作为阻塞深度的函数。此设置用于探索视频-大语言模型是否表现出明显的分阶段模式？如 figs. 3 and 4 所示。

全局设置 2 为了评估不同注意力类型的整体贡献，我们选择一种类型并将其对应的去除应用于所有层。我们逐一迭代所有去除类型：

$$L_j^{\text{KT}} = \text{KT}, \quad \forall j \in \{1, \ldots, L\} \quad (7)$$

其中 $\text{KT} \in \{\text{LV-K, VT-K, VS-K}\}$。此设置用于探索在全球范围内，每种注意力类型如何贡献于视频问答性能？如图 fig. 5 所示。

细粒度设置。为了细粒度地检查每种注意力类型的影响，我们应用一个特定的击除 $\text{KT} \in \{\text{LV-K, VT-K, VS-K}\}$ 在以第 x 层（$x \ge 4$）结束的 4 层滑动窗口内，使窗口外的层不进行击除。我们定义受影响的层为：

$$L_p^{\text{KT}} = \text{KT}, \quad \forall p \in \{x-3, x-2, x-1, x\} \quad (8)$$

此设置用于探索在细粒度层面，每种注意力类型如何影响不同层的视频问答？如 fig. 6 所示。

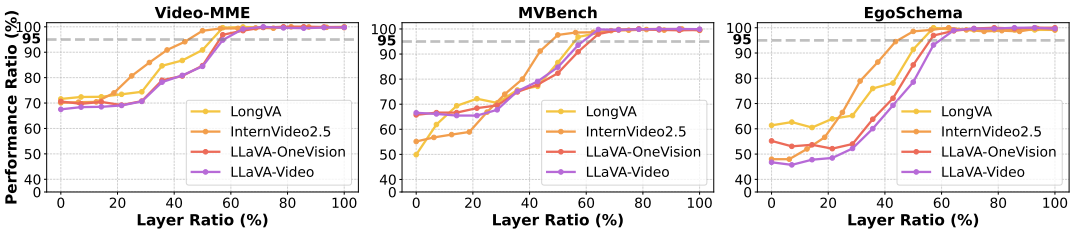

图 3: 不同基准上不同模型的性能变化归一化。设置：将语言到视频淘汰（LV-K）应用于超过特定截止深度的情况。例如，60%表示 LV-K 应用于前 60%层之后的所有层。归一化是相对于整个模型的性能而言的。

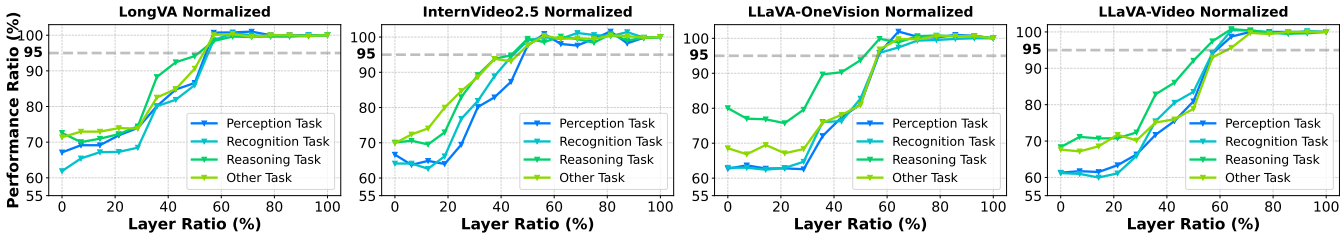

图 4: 不同模型在 Video-MME 上执行不同任务的标准化性能。设置：将语言到视频淘汰（LV-K）应用于超过某个截止深度的情况。该标准化是相对于完整模型性能而言的。

实验和观察

本节我们将首先概述数据集和模型，然后展示实验结果并进行详细分析。

实验详情

数据集。我们专注于视频问答任务，并使用三个已建立的主流数据集（即.，Video-MME(Fu et al. 2024)，MVBench(Li et al. 2024b) 和 EgoSchema(Mangalam, Akshulakov, and Malik 2023)）来探索 Video-LLMs 的内部机制。这些数据集在持续时间（从 3 分钟到 1 小时不等）、视角（第一人称和第三人称视角）、场景背景（例如.、家庭环境、电影）以及问题类型（例如.、时间推理和空间推理）方面有所不同。Video-MME(Fu et al. 2024) 包含总计 254 小时的 900 个视频，有 2,700 个人类标注的问题答案对。这些视频涵盖了六大主要领域——知识、影视、体育、表演艺术、日常生活以及多语言内容，并且视频长度各不相同。MVBench(Li et al. 2024b) 是一个短时长的多模态基准测试，旨在评估 MLLMs 在动态任务中的时间推理能力。与静态图像基准不同的是，它包括 20 个时间相关的任务，如动作序列、预测、物体恒存性分析以及运动分析，每个任务有 200 个样本，共计 4,000 个问题答案对。EgoSchema 数据集 (Mangalam, Akshulakov, and Malik 2023) 包含从 5,000 个三分钟的自我中心视频中衍生出的 5,000 个多选题。它包括一个公开可用的 500 个问题子集，而完整的评估则在服务器上进行。由于问题数量庞大，我们采用了广泛使用的子集来进行评估。

模型。我们研究了广泛使用的开源视频大型语言模型（Video-LLMs）(Zhang et al. 2024a; Wang et al. 2025; Zhang et al. 2024c; Li et al. 2024a)，这些模型在多样化的视频理解任务中实现了前沿性能。这些模型采用相似的架构，但训练数据集不同且公开可用，这使我们能够系统地探索 Video-LLMs 中的空间-时间建模，同时尽量减少与模型架构相关的混淆因素。我们测试了四个 7B 模型，包括 LongVA(Zhang et al. 2024a)、InternVideo2.5(Wang et al. 2025)、LLaVA-Video-7B(Zhang et al. 2024c) 和 LLaVA-OneVision-7B(Li et al. 2024a)。我们还在附录的实验部分提供了来自一个更大 34B 模型的结果，而大多数实验由于计算成本限制都是在 7B 模型上进行的。对于所有测试的模型，我们都使用相同的均匀采样策略从每个视频中抽取 32 帧。其他细节见附录中的实现细节部分。

视频-大语言模型是否表现出明显的阶段模式？

为了调查 Video-LLMs 中的视频处理是否也遵循一个明确的分阶段模式，我们将语言到视频注意力剔除应用于某个层之后的所有层，对应于全局设置 1。不同的 Video-LLMs 在基准测试上的性能可能有所不同，并且层数也可能不同。为了更方便地可视化和比较，我们引入了两个指标—**层比**和**性能比**—来评估移除特定层以

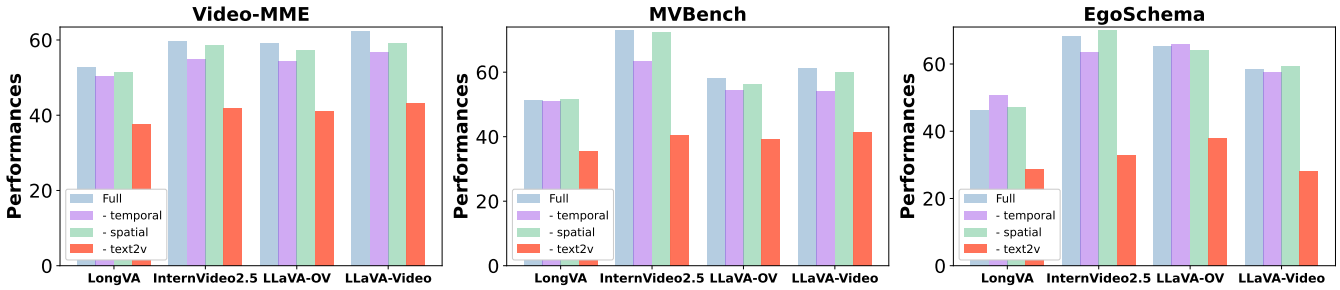

图 5: 不同基准测试的绝对性能。设置：对所有层应用不同的剔除。

外的视觉标记对模型性能的影响。**层比**表示相对于总层数而言保持语言到视频注意力的比例。例如，60% 的层比率意味着剩余 40%的层的语言到视频注意力被关闭，从而阻止任何进一步的视觉信息流向文本标记。**性能比**定量了在当前击溃设置下模型相对于其原始（无击溃）100% 性能的表现，使用 100% 层进行衡量。图 3显示了视频-LLMs 在基准测试中的表现。

我们得出以下观察结果：(i) 当所有层都无法访问视觉标记时，三个 Video-LLMs 的性能最多保留其原始性能的 70%（例如 LongVA 在 Video-MME 上）和最少 48%（例如 LLaVA-Video 在 EgoSchema 上）。这表明视觉信息在准确回答问题中起着关键作用，同时也展示了嵌入语言模型中的广泛世界知识。(ii) 随着层比例从 0% 增加到 60%，模型性能逐渐提高，在所有视频理解基准测试中的 50% – 60%范围内提升最为显著。这表明中间层对视觉信息处理的贡献最大，这一点还得到了我们精细设置实验的支持。(iii) 在不同视频理解基准上，阻塞超过 60%层数深度的信息对模型性能几乎没有影响。这表明视觉信息主要在早期层进行处理，而剩余层则主要负责高级推理—显示出一个清晰的阶段化处理模式。此外，图 4展示了不同 Video-LLMs 在 Video-MME 基准测试中的各种任务（包括感知、识别、推理和其他类型的任务）上的性能表现情况。我们观察到，在不同的任务类别中，随着视觉标记逐渐被阻塞，模型的性能表现出的趋势与图 3所示相似。这进一步验证了我们的发现的有效性。不同 Video-LLMs 在各种基准上进行附加任务级别的性能结果、定性结果和一个拥有 320 亿参数的更大规模模型实验也都观察到了类似的模式，这进一步证实了这一趋势的存在。所有这些都可以在附录中的实验部分找到。

在全球范围内，每种注意力类型如何贡献于视频问答性能？

我们进一步探讨了每种注意力机制如何从全局视角对视频问答性能做出贡献。为了回答这个问题，我们在全局设置 2 下进行了实验。具体来说，我们系统地将每种注意力机制的剔除应用到每个模型的所有层，并在各种基准测试中评估其表现。如图 5所示，在所有层上应用视频时间剔除和视频空间剔除导致不同基准上的性能下降最小。然而，跨所有层应用语言到视频的剔除则导致显著的性能下降。这表明在 Video-LLMs 中，时空建模主要通过语言令牌与视频令牌之间的交互进行，而时间和空间自注意力贡献较少。值得注意的是，在视频问答任务中，时间注意力和空间注意力的计算成本通常远高于语言到视频的注意力，当前依赖更多的是语言到视频的注意力而不是其他两者。

在细粒度层面，每种注意力类型如何影响不同层的视频问答？

我们在细粒度设置下进行了实验，采用每种剔除类型作用于一个滑动窗口中的少量层（我们使用的是 4），然后考察在这种特定窗口内每种注意机制如何影响最终答案。这里报告了这些测试模型在每个任务数据集上的绝对性能变化。如图 6所示，我们观察到以下几点：(i) 对于大多数滑动窗口层，语言到视频的注意力剔除导致显著更大的性能下降，相比之下时间或空间注意的剔除较少，如图 fig. 6 所示。(ii) 对于一部分层，应用剔除会导致显著的性能下降，而对剩余层的应用影响较小。(iii) 对于大多数单个层，语言到视频的注意力剔除的影响比时间和空间注意的剔除要强。(iv) 在某些情况下，应用剔除甚至会改善性能。例如，在长 VA 的情况下，当某些层被剔除时，在 EgoSchema 上的表现得到了提升。

潜在应用

我们之前的观察揭示了现有视频 LLMs 在视频问答任务上的低效。例如，在全局设置中，我们在视频 LLM 中识别出一个两阶段处理模式。具体来说，第二阶段封锁所有视觉令牌可以几乎保留原始性能的同时显著减少计算成本。以 LLaVA-Video 为例，每个视频

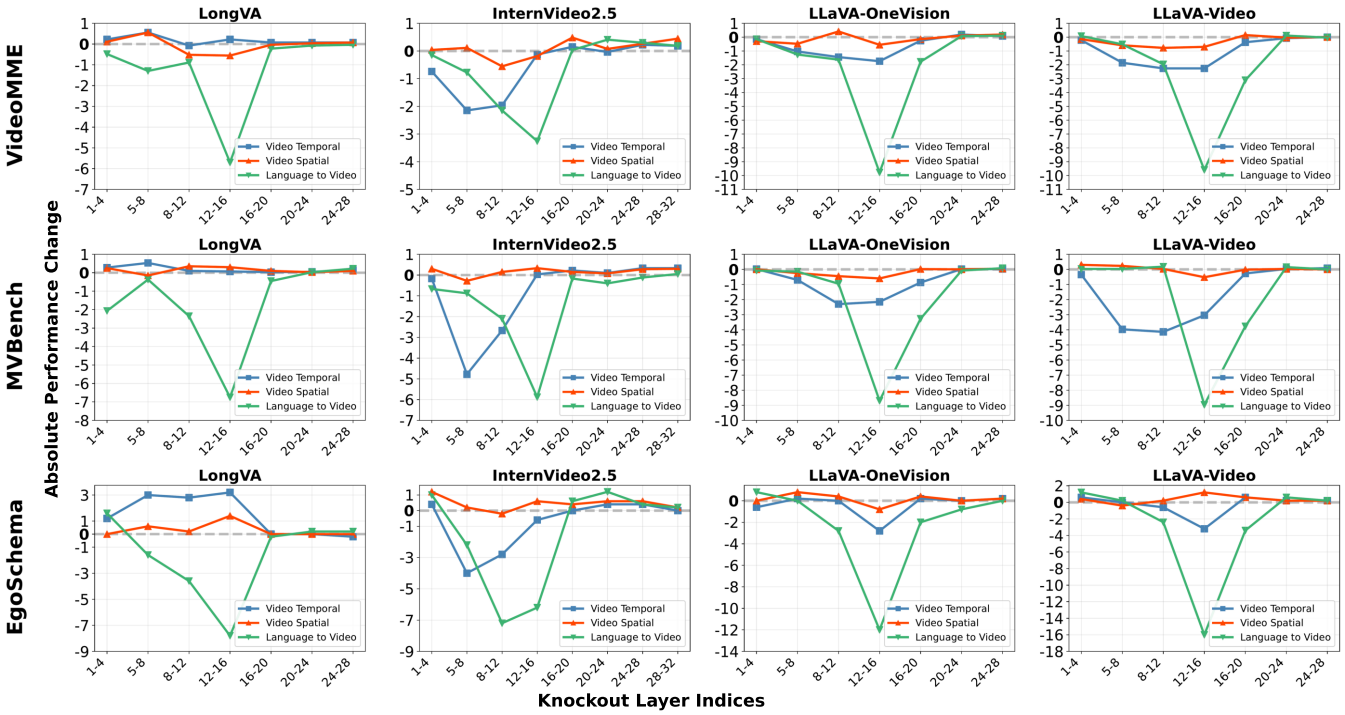

图 6: 不同基准下的绝对性能变化。设置：在一个滑动窗口内对各层应用不同的剔除。

帧包含 196 个令牌，并且每段视频有 32 帧，完整的视频序列长度达到 196×32 个令牌。相比之下，文本序列通常包含的令牌少于一个单帧（196 个令牌）。因此，每一层的视频注意力计算比语言自注意力高出 32^2 倍以上。因此，通过在第二阶段省略视频令牌可以实现显著的计算节省。在我们的细粒度设置中，我们发现了不同的层级级异常值——这个属性可以被用来进一步减少第一阶段的计算量。具体来说，时间注意力带来的计算成本是空间注意力的 31 倍。通过限制每个帧内的注意力为空间注意力，以这些识别出的异常值为前提，我们大幅度减少了总体计算负载。我们提出了一种简单策略：对于 LLaVA-OneVision 和 LLaVA-Video，我们在第 1 层到第 8 层封锁时间注意力（第 8 层被识别为一个异常值），随后从第 18 层开始移除视觉令牌。table 1 中的结果表明，这种策略在显著减少计算开销的同时保持了与基准方法相当的性能。LongVA 和 InternVideo2.5 的详细浮点运算次数和结果提供于补充材料的实验部分。

结论

本文揭示了 Video-LLMs 在处理视频问答任务时的内部工作机制。我们的实验表明，不同的 Video-LLMs 在各种基准测试和任务中表现出类似的处理模式。具体而言，视频信息提取在早期层完成，而视频推理则发生在后期层。此外，空间-时间建模主要由语言引导的检索驱动，而不是通过 Video Temporal 或 Video Spatial 注意力机制驱动。另外，一小部分层在视频问答中起着关键作用。最后，我们展示了可以通过减少非关键层中的计算并在第二阶段退出视频标记来降低计算成本。这些发现增强了对 Video-LLMs 的理解性，为深入研究这些模型如何处理和理解视频内容提供了新的研究方向。此外，它们还提供了改进下游任务的有效性和效率以及优化模型设计的见解。

参考文献

Aflalo, E.; Du, M.; Tseng, S.-Y.; Liu, Y.; Wu, C.; Duan, N.; and Lal, V. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 21406–21415.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923.

Basu, S.; Grayson, M.; Morrison, C.; Nushi, B.; Feizi, S.; and Massiceti, D. 2024. Understanding information storage and transfer in multi-modal large language models. arXiv preprint arXiv:2406.04236.

Model	Settings	Attention Flops	MME	MVBench	EgoSchema
Llava-Video	Baseline	100%	62.4	61.1	58.4
	Exit only	64.3%	62.0	61.1	58.0
	Exit + window	37.1%	60.0	60.8	58.2
Llava-OneVision	Baseline	100%	59.1	58.3	65.2
	Exit only	64.3%	58.2	57.3	64.4
	Exit + window	37.1%	58.0	57.6	65.2

表 1: 不同模型在三个基准测试中的性能表现。仅退出表示视频标记在经过某一层后离开模型。退出 + 窗口表示除了退出外，我们还控制了时间注意力范围——即对于某些层，仅允许视频帧执行空间注意力。对于 LLaVA-OneVision 和 LLaVA-Video，我们在第 18 层之后让视频标记退出，并限制前 8 层仅执行空间注意力，因为它们作为非关键层。

Cao, J.; Gan, Z.; Cheng, Y.; Yu, L.; Chen, Y.-C.; and Liu, J. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, 565–580. Springer.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In Proceedings of the IEEE/CVF international conference on computer vision, 397–406.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024a. 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, 19–35. Springer.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024b. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv preprint arXiv:2412.05271.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Frank, S.; Bugliarello, E.; and Elliott, D. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. arXiv preprint arXiv:2109.04448.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075.

Ge, J.; Chen, Z.; Lin, J.; Zhu, J.; Liu, X.; Dai, J.; and Zhu, X. 2024. V2PE: Improving Multimodal Long-Context Capability of Vision-Language Models with Variable Visual Position Encoding. arXiv preprint arXiv:2412.09616.

Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting recall of factual associations in auto-regressive language models. arXiv preprint arXiv:2304.14767.

Hendricks, L. A.; and Nematzadeh, A. 2021. Probing image-language transformers for verb understanding. arXiv preprint arXiv:2106.09141.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In International Conference on Learning Representations.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv preprint arXiv:2310.06825.

Kaduri, O.; Bagon, S.; and Dekel, T. 2024. What's in the Image? A Deep-Dive into the Vision of Vision Language Models. arXiv preprint arXiv:2411.17491.

Ko, D.; Lee, J. S.; Kang, W.; Roh, B.; and Kim, H. J. 2023. Large language models are temporal and causal reasoners for video question answering. arXiv preprint arXiv:2310.15747.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326.

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.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22195–22206.

Lin, Z.; Lin, M.; Lin, L.; and Ji, R. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 5334–5342.

Lindström, A. D.; Bensch, S.; Björklund, J.; and Drewes, F. 2021. Probing multimodal embeddings for linguistic properties: the visual-semantic case. arXiv preprint arXiv:2102.11115.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, J.; Wang, Y.; Ma, H.; Wu, X.; Ma, X.; Wei, X.; Jiao, J.; Wu, E.; and Hu, J. 2024b. Kangaroo: A powerful video-language model supporting long-context video input. arXiv preprint arXiv:2408.15542.

Liu, Z.; Guo, L.; Tang, Y.; Cai, J.; Ma, K.; Chen, X.; and Liu, J. 2025. VRoPE: Rotary Position Embedding for Video Large Language Models. arXiv preprint arXiv:2502.11664.

Lyu, Y.; Liang, P. P.; Deng, Z.; Salakhutdinov, R.; and Morency, L.-P. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 455–467.

Ma, Z.; Gou, C.; Shi, H.; Sun, B.; Li, S.; Rezatofighi, H.; and Cai, J. 2024. Drvideo: Document retrieval based long video understanding. arXiv preprint arXiv:2406.12846.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems, 36: 46212–46244.

Neo, C.; Ong, L.; Torr, P.; Geva, M.; Krueger, D.; and Barez, F. 2024. Towards interpreting visual information processing in vision-language models. arXiv preprint arXiv:2410.07149.

Salin, E.; Farah, B.; Ayache, S.; and Favre, B. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 11248–11257.

Schwettmann, S.; Chowdhury, N.; Klein, S.; Bau, D.; and Torralba, A. 2023. Multimodal neurons in pretrained text-only transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2862–2867.

Stan, G. B. M.; Aflalo, E.; Rohekar, R. Y.; Bhiwandiwalla, A.; Tseng, S.-Y.; Olson, M. L.; Gurwicz, Y.; Wu, C.; Duan, N.; and Lal, V. 2024. LVLM-Interpret: An Interpretability Tool for Large Vision-Language Models. arXiv preprint arXiv:2404.03118.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, Y.; Li, X.; Yan, Z.; He, Y.; Yu, J.; Zeng, X.; Wang, C.; Ma, C.; Huang, H.; Gao, J.; et al. 2025. Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386.

Wei, X.; Liu, X.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Tong, J.; Duan, H.; Guo, Q.; Wang, J.; et al. 2025. VideoRoPE: What Makes for Good Video Rotary Position Embedding? arXiv preprint arXiv:2502.05173.

Xiao, J.; Huang, N.; Qin, H.; Li, D.; Li, Y.; Zhu, F.; Tao, Z.; Yu, J.; Lin, L.; Chua, T.-S.; et al. 2025. Videoqa in the era of llms: An empirical study. International Journal of Computer Vision, 1–24.

Xu, S.; Pang, L.; Zhu, Y.; Shen, H.; and Cheng, X. 2024. Cross-modal safety mechanism transfer in large vision-language models. arXiv preprint arXiv:2410.12662.

Xue, F.; Chen, Y.; Li, D.; Hu, Q.; Zhu, L.; Li, X.; Fang, Y.; Tang, H.; Yang, S.; Liu, Z.; et al. 2024. Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems, 35: 124–141.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9847–9857.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V:

A GPT-4V Level MLLM on Your Phone. arXiv preprint arXiv:2408.01800.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2023. Self-chained image-language model for video localization and question answering. Advances in Neural Information Processing Systems, 36: 76749–76771.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024a. Long context transfer from language to vision. arXiv preprint arXiv:2406.16852.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199.

Zhang, X.; Shen, C.; Yuan, X.; Yan, S.; Xie, L.; Wang, W.; Gu, C.; Tang, H.; and Ye, J. 2024b. From redundancy to relevance: Enhancing explainability in multimodal large language models. arXiv e-prints, arXiv–2406.

Zhang, Y.; Li, B.; Liu, H.; Lee, Y. J.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024c. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024d. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713.

Zhang, Z.; Yadav, S.; Han, F.; and Shutova, E. 2025. Cross-modal information flow in multimodal large language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, 19781–19791.

Zhao, Q.; Xu, M.; Gupta, K.; Asthana, A.; Zheng, L.; and Gould, S. 2024. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In European Conference on Computer Vision, 127–142. Springer.