

视频对于训练视频语言模型有多重要？

George Lydakis¹ Alexander Hermans¹ Ali Athar² Daan de Geus^{1,3} Bastian Leibe¹

¹RWTH Aachen University ²ByteDance Seed ³Eindhoven University of Technology

{ lydakis,hermans,leibe } @vision.rwth-aachen.de ali.athar@bytedance.com d.c.d.geus@tue.nl

Abstract

视频大语言模型 (Video LLMs) 的研究进展迅速, 仅仅几年内就出现了大量模型和基准测试。通常, 这些模型以一个预训练的仅限文本的 LLM 初始化, 然后在图像和视频字幕数据集上进行微调。在本文中, 我们展示了研究结果, 表明视频 LLMs 在仅图像训练后具备比预期更强的时序推理能力, 而视频特定训练带来的改进却出人意料地小。具体来说, 我们展示了两个使用最近的 LongVU 算法训练的 LLM 的图像训练版本, 在一个时序推理基准 TVBench 上表现明显高于偶然水平。此外, 我们引入了一种简单的微调方案, 该方案涉及带有注释图像的序列和针对时序能力的问题。这一基线在时序推理性能上接近于使用视频训练的 LLM, 有时甚至更好。这表明当前模型没有充分利用真实视频中的丰富时序特征。我们的分析激发了进一步研究机制以了解图像训练的 LLM 如何进行时序推理, 以及导致当前视频训练方案低效的瓶颈。

1. 引言

大型语言模型 (LLM) 最初是作为面向文本的架构开发的, 如今已被证明在多模态推理方面也是具有竞争力的选择。视频语言任务也不例外: 近年来出现了大量的视频 LLM [1, 3, 6, 12–14, 20, 22, 24]。这里的典型设置是查询模型以获取视频的描述, 或请求回答与之相关的特定问题。

在该领域的大多数工作中, 使用的基本道路相同: 视频帧通过如 DINOv2 [18] 或 CLIP [19] 这样强大的预训练模型被编码成标记, 并将这些标记的一个子集与提示文本的嵌入进行连接。然后训练大型语言模型 (LLM)——要么作为一个整体训练, 要么通过诸如 LoRA [7] 这样的自适应方法——以预测所需的输出, 无论是视频的描述还是问题的答案。视频标记化的具体方案一直是广泛研究的主题 [3, 6, 13, 20, 22], 训练数据的各个方面也是如此, 从视频语料库的规模和多样性, 到注释的质量和细节程度 [1, 2, 11, 21, 24]。

我们的研究动机来源于观察当前的 Video LLMs 训练过程通常涉及三个阶段, 而只有最后一个阶段涉及实际视频数据: 首先是基于文本的 LLM 训练, 然后是

对图文数据集的多模态微调, 最后是对视频文本数据集的多模态微调。这让我们提出一个关键问题: Video LLMs 的时间推理能力中, 哪一部分是来自于视频训练的, 哪一部分是已经从文本和图像预训练中获得的? 这可以为现有架构从不同类型数据中学习的有效性提供见解, 尤其是当前视频数据集中的视频文本对。

为了阐明这一点, 我们在几个训练阶段后评估视频语言模型 (Video LLMs) 的时间推理能力。对于这些实验, 我们使用 TVBench [4], 这是一个评估视频理解的时间方面的多项选择问答 (MCQA) 基准测试。令人惊讶的是, 我们发现最近使用 LongVU [20] 方法训练的两个视频语言模型的图像训练版本在这个基准测试的一些任务中几乎立即显著高于随机水平表现, 而未曾真实视频上训练。此外, 我们发现将这些模型微调在通过拼接不相关的带字幕图像形成的伪视频上, 其表现接近并偶尔高于那些经过视频训练的模型。这表明当前视频语言模型在利用视频数据集方面存在潜在瓶颈, 可能是数据质量、架构或两者的结合问题。

我们的分析提出了一些有趣的研究问题。首先, 哪些机制使得图像训练的 LLM 具备时间推理能力? 其次, 哪些因素导致基于视频的训练效果显著低于预期? 由于训练视频 LLM 的计算成本很高, 我们希望激励研究人员努力识别这些因素并提出改进和建议。

2. 相关工作

视频 LLM 的研究进展迅速, 导致在短时间内推出了大量模型。诸如 Video-ChatGPT [17]、Video-LLaMA [25] 和 Video-LLaVA [14] 等作品引入了一个基本的架构, 包括一个文本预训练的 LLM、图像和/或视频编码器, 用于将视频输入标记化, 以及简单的模块——通常是线性层或小型 MLP——将这些视频标记映射到 LLM 可解释的特征空间。虽然这一架构基本保持不变, 但已经提出了几种视频标记化方案, 通常旨在增加处理帧的数量。这些方案可能涉及池化 [3, 22]、静态 [13] 或自适应压缩方案 [20], 以及记忆库 [6]。

Training schemes. 一项重要的研究方向, 特别是与我们的研究密切相关的部分, 集中在训练数据的数量和质量上。例如, Tarsier [21] 证明, 在大型且经过适当筛选的数据集上训练时, 从专有和公共来源获得的数据集上, 现代视频大型语言模型中使用的简单架构可以

实现强大的性能。在此之前, ShareGPT4Video [2] 和 VideoChat2 [11] 等工作也强调了在视频内容和问题形式方面包括各种数据源的重要性。

在文献中, 一个普遍趋势是将视频大语言模型 (Video LLMs) 训练在图片和视频的组上, 或者是混合训练 [14, 16], 或者是分两阶段训练 [20, 22]。特别是对于视频训练, 一些方法侧重于创建具有理想属性的文本注释。例如, 将模型的预测与人类偏好的输出对齐 [1], 目标既包括总结, 也包括时间定位 [12], 增加带有定位事件的帧范围的字幕 [24], 并自动生成不同时间粒度的字幕 [26]。然而, 到目前为止, 视频训练相比于文本和图像训练在启用时间理解能力方面的相对重要性还没有被广泛研究。

为了评估视频大语言模型 (Video LLMs) 的时序推理能力, 需要一个适当的基准。在模型开发的同时, 已经提出了几种评估数据集和方法论。根据问题的格式, 这些可以分为两大类。第一类基于“自由形式”的问题, 模型会得到一个视频和一个问题, 并需要在没有其他提示的情况下生成答案。例如包括 MSRVTT、TGIF 和 ActivityNet-QA。答案质量的评估通常使用另一个大语言模型来完成, 该模型的任务是根据给定的标准答案对视频大语言模型的输出评分。第二类基于多项选择题, 模型的输出需要与给定选项之一相匹配。在这里, 我们发现了像 MVBench、Video-MME、TVBench 和 LongVideoBench 这样的基准。在这种情况下, 指标是标准的准确度。

第一种范式的主要优点是, 设计不容易被排除的错误选项无需费力, 同时也没有给视频 LLM 提供提示。然而, 这种度量标准的可解释性较差, 并且强烈依赖于用于评估的 LLM。由于 (1) 我们研究的重点是视频 LLM 的时间推理能力, 并且 (2) 我们希望我们的结果独立于 LLM 的选择, 因此我们选择了多项选择题的 TVBench 基准 [4]。TVBench 专门设计用于解决以往视频 LLM 基准测试中存在的不足——即很多问题可以通过单帧甚至仅从文本查询中得到答案。通过专门设计题目和诱饵选项以要求时间理解, TVBench 缓解了这一问题, 使其非常适合我们的研究。

3. 方法

在训练视频大型语言模型时, 一个显著的挑战在于这些模型必须具备三种不同类型的推理能力, 并能够将它们结合起来。首先, 它们必须能够有效地解析文本并提取所问内容的信息。其次, 所提出的问题所需的视觉特征方面非常多样, 可能需要编码 e.g. 物体类型、颜色、形状、数量和空间位置。第三, 对视频的推理需要理解这些特征随时间的演变。

任务的高复杂性和组合性使得显而易见这些模型为何以文本预训练的 LLM 进行初始化, 以及为何在训练过程中使用标注的图像数据集。在这项工作中, 我们感兴趣调查视频数据集相较于文本-图像预训练的相对影响。换句话说, 视频 LLM 在多大程度上通过视频微调获得了时间推理能力?

为此, 我们提出在不涉及真实视频训练的两种设置

下研究其时间推理行为。首先, 我们评估仅在文本和图像上进行训练的模型。其次, 我们通过以下程序使用图像-字幕数据集替代视频训练阶段以完成时间任务。假设一个带字幕的图像数据集 $D = \{d_1, d_2, \dots, d_N\}$ 有 N 个样本, 其中 $d_j = (i_j, c_j)$ 是由图像 i_j 和对应字幕 c_j 组成的对, 我们按如下方式构建伪视频。首先, 从完整数据集 D 中随机选择 S 个图像-字幕样本 d_j , 每一个样本构成单个视频“场景”的基础。随后, 为这些场景中的每一个选择一个持续时间 F_j , 并重复对应的图像 i_j 共 F_j 次。对每一帧施加相对温和的仿射变换, 以在非常粗略的水平上模拟视频。最终, 生成的伪视频长度为 $\sum_{j=1}^S F_j$ 帧。

这种相对简单的方法的优点在于伪视频的时间演化是完全已知的。假设图像字幕中噪声较小, 我们因此可以生成各种需要模型进行时间推理的问题。图 1 (底部) 展示了与标准视频训练 (图 1 (顶部)) 相比的训练过程。我们的意图是将其作为一个基线: 伪视频的信息量明显少于真实视频, 并且与推理期间看到的真实数据不太相似。因此, 一个运行良好的视频训练方案预计将比这个基线表现显著更好。

学习表示的质量可能受到一些因素的影响, 例如单个场景的持续时间、它们的数量以及可能更重要的是, 向模型提出的问题。在这项工作中, 我们研究了多项选择题, 其中只有一个答案是正确的。

我们可以根据回答问题所需的时间推理类型, 将被检查的问题分为两大类:

- 需要理解相对时间概念的问题。示例包括要求模型选择描述视频中场景顺序的选项, 或者确定某一描述的场景是否出现在另一场景之前或之后。
- 模型必须能够对绝对数量进行推理的问题。这可能是以绝对时间定位的形式, e.g. 为第 i 个场景提供字幕, 或者以计数的形式, e.g. 计算视频中出现的不同场景的数量。

原则上, 这两种问题类型都针对视频语言模型中理想的能力。表 1 列出了我们实验的六个不同问题及其对应的缩写。在这些问题上, R_1, R_2, R_3, R_4 可以被描述为需要相对时间推理, 而 A_1, A_2 需要绝对时间推理。

4. 实验

我们使用的视频 LLM 基于近期的 LongVU [20] 架构。LongVU 利用 DINOv2 [18] 特征空间相似性来选择输入帧序列的一个子集, 然后应用额外的启发式方法以高或低分辨率表示每个剩余的帧。我们选择这些模型是基于它们在各种基准测试中的强大性能, 以及支持 Llama3.2 [5] 和 Qwen2 [1] 骨架的图像和视频训练模型的训练代码和检查点的可用性。

除非另有说明, 我们使用的批量大小为 64, 与原始论文中的一致, 并对学习率调度进行了轻微的修改, 这在初步实验中使收敛速度有所加快。学习率在总训练步骤的前 3% 中线性增加到 $5 \cdot 10^{-6}$, 然后在剩余的训练过程中遵循余弦衰减计划。LongVU 算法特定的所有超参数与原始工作 [20] 中的保持一致。

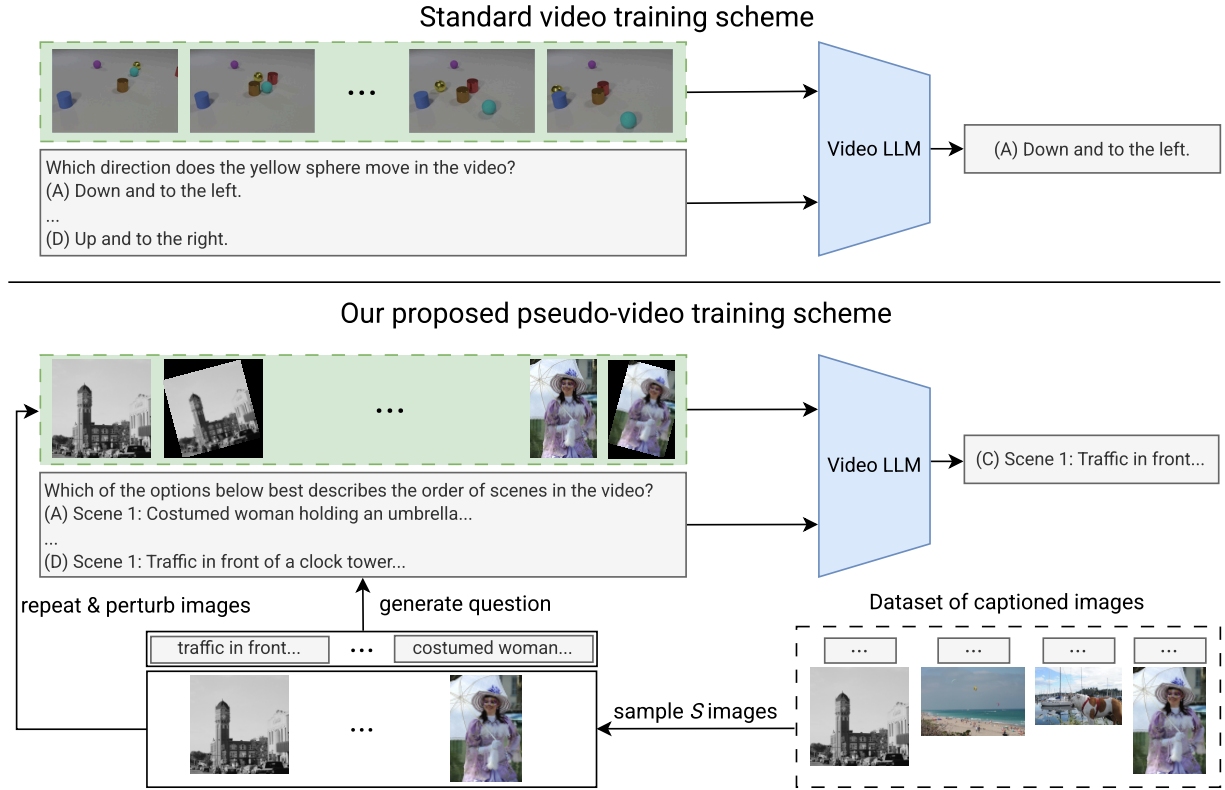


Figure 1. 标准视频训练方案进行视频 LLMs 的比较（上图），以及我们提出的伪视频训练方案（下图）。我们利用标注图片数据集自动生成短伪视频和问题以进行训练。

Abbreviation	Question description
R_1	“ Which of the following options best describes the order of scenes in the video? ”, followed by a number of different permutations of the scene captions, each of which has the form “ Scene 1: [caption] Scene 2: [caption] ... Scene [N]: [caption]”
R_2	“ In the given video, does the scene that can be captioned as ” [caption] ” happen before or after the scene that can be captioned as ” [caption] ”? ”, always followed by exactly two answer options, “ before ” and “ after ”.
R_3	“ The following scenes appear in the video, not necessarily in this order: [comma-separated caption list] . Of those scenes, which occurs [either first or last] ? ”, followed by a number of different scene captions from the pseudovideo.
R_4	“ One of the scenes in the video can be described as ” [caption] ”. Describe the scene immediately [either “ before ” or “ after ”] it. ”, followed by a number of different scene captions from the pseudo video, or an option stating that the given scene is the first/last one and hence has no scene before/after it respectively. This option may either be correct or wrong.
A_1	“ How many different scenes appear in the video? ”, followed by a number of numerical answers.
A_2	“ There are [number of scenes] in the video. What does scene [number between 1 and the number of scenes] depict? ”, followed by a number of different scene captions from the pseudo video.

Table 1. 我们实验使用的所有问题的描述和缩写。 R_i 问题侧重于询问相对场景排序，而 A_i 变体则侧重于在伪视频内以绝对方式定位场景。

如前所述，我们选择的基准是 TVBench，因为它侧 重于视频理解的时间方面。接下来，我们将使用表格 2

Task name	Abbreviation
Action Count	AC
Object Count	OC
Action Sequence	AS
Object Shuffle	OS
Scene Transition	ST
Action Localization	AL
Action Antonym	AA
Unexpected Action	UA
Egocentric Sequence	ES
Moving Direction	MD

Table 2. TVBench [4] 任务名称缩写。

中的缩写来指代 TVBench 任务。有关这些任务性质的更多详细信息，请参阅原始出版物 [4]。

4.1. 图像训练的 LLMs 中的时间推理

我们首先进行实验，以确定仅在图像上训练的两个 LongVU 模型在 TVBench 上的表现有多好。理想情况下，这可以简单地通过在基准上评估提供的检查点来完成。然而，当从图像输入转换为视频输入时，图像被标记的方式会略有变化。因此，我们发现直接在 TVBench 上评估图像预训练的检查点会导致模型输出不符合所需多项选择格式的文本。这个问题可以通过在由多帧组成的样本上训练模型仅一步来解决。因此，我们展示了在伪视频上训练一步，并带有问题 R_1 的情况下的结果。考虑到训练时间短且这些模型从未见过真实视频，我们认为这些配置与仅在图像上训练的大型语言模型足够接近，因此使用它们作为评估的代理。

表格 3 展示了图像训练模型的性能，并将其与完全视频训练的 LongVU 模型进行比较。尽管这些模型从未经过视频训练，但当仅在文本和图像上进行训练时，Llama 和 Qwen 的得分均大大高于偶然水平。考虑到训练所需的额外数据和计算量，图像训练模型与视频训练模型之间的差距（~ 5%）令人惊讶地小。另外，提供乱序帧会导致两种模型的性能下降。尽管最终的准确性不是在偶然水平上，但这与原始工作中在相同实验设置下对几种模型的报告结果一致。这表明这些模型确实在一定程度上进行时间推理，而不仅仅是利用数据集的潜在弱点或“捷径”。

在关注模型表现最佳的一些子任务时，可以做出进一步有趣的观察。其中两个任务，物体计数（OC）和运动方向（MD），基于来自 CLEVRER [23] 数据集的视频。虽然模型没有看过该数据集的视频，但训练过程确实包括来自 CLEVR 数据集 [8] 的图像，这些图像在视觉上与 CLEVRER 相似。这些结果表明，LLM 在一定程度上可以通过在视觉上相似的图像-文本数据上进行训练来解决时间任务。为了更深入地了解这个问题，可以重新训练图像-文本模型，同时排除 CLEVR 来源的图像和问题，并评估对 OC 和 MD 子集的影响。然而，此实验的规模超出了我们的计算资源。

总体而言，结果表明视频大型语言模型能够仅从文本和图像中学习简单的时间推理，而无需经过视频训练。理解其中的机制将为这些模型如何处理视频提供有价值的见解，我们鼓励未来在这一方向开展研究。此外，这些结果提出了一个问题，即是否可以使用文本-图像数据来进一步提高模型的时间推理能力。我们在下一节中实验的伪视频设置是实现这一目标的一个例子，但我们预期还有各种其他有效的实验设置。

4.2. 伪视频训练

我们使用 COCO [15] 作为我们伪视频设置的标注图像来源。COCO 的描述相对较短，这在创建专注于时间推理而不是详细场景描述的问题时可能是有利的。此外，训练 LongVU 模型所用的 LLaVA-OneVision 数据 [10] 中包括 COCO 图像，这可能使模型更容易利用已获取的图像知识。

在创建伪视频时，如 3 第节所述，我们将把最大场景数（i.e.，采样的不同图像数）称为 S ，并将每个场景的最大帧数（i.e.，对图像应用仿射变换的重复次数）称为 F 。除非在某些情况下单个场景使问题变得简单（e.g. R_1 ），在这种情况下我们将从 $\{2, 3, \dots, S\}$ 中采样，否则场景数将从 $\{1, 2, \dots, S\}$ 中均匀随机采样。每个场景的帧数从 $\{1, 2, \dots, F\}$ 中均匀随机采样。

在问题生成中，我们通常使用 3 个错误选项，除非是 R_2 ，在这种情况下错误选项总是 1 个；在 R_3 中，我们将错误选项的数量设置为字幕列表的长度，该列表是从 $\{1, 2, 3\}$ 中取样的。除非另有说明，所有呈现的模型均在 100,000 个伪视频上训练了 2 个周期，对应 3,125 个训练步骤。

在我们中，我们展示了使用我们的伪视频方案微调图像和视频训练的 LongVU 检查点时取得的结果。在此处，我们报告了通过实验经验确定的最佳伪视频配置的结果。对于两个 Llama3.2 检查点，这些结果是在将 $(R_1 : S = 4, F = 20)$ 和 $(R_3 : S = 6, F = 5)$ 设定为问题生成设置时取得的。对于两个 Qwen2 模型，最佳结果是在 $(R_1 : S = 4, F = 5)$ 的情况下取得的，而在 $F = 10$ 的情况下，我们观察到略差的结果（-0.4）。由于我们使用的 A40s 与原始 LongVU 工作中使用的 H100s 相比，GPU 内存不足以用 $F > 10$ 训练 Qwen2，因此我们无法测试类似于用于 Llama3.2 的 20 帧设置。微调视频训练模型对伪视频的实验旨在确定这种类型的训练是否对同时在真实视频上训练的 LLM 有益。理想情况下，这可以通过在图像和真实视频训练之间包含伪视频培训作为中间阶段来完成。然而，这意味着要在最初使用的视频数据上训练 LongVU 检查点。这在收集使用的数据集方面具有挑战性，并且考虑到我们的资源在计算上不可行。因此，我们只能测试当这些伪视频在真实视频训练后用作训练阶段时的影响。

我们观察到，伪视频训练的 Llama 模型（53.5%）优于其视频训练的对模型（51.2%）。从这种类型的训练中特别受益的任务是“动作序列”（AS）、“场景过渡”（ST）和“移动方向”（MD）。这可能是直观的，因为这些任务在很大程度上依赖于模型关联时间帧的能

Configuration	LLM	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	Avg.
Chance level		25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0	33.3
Video-trained	Llama3.2-3B	31.0	59.4	70.2	33.3	76.2	44.4	59.1	32.9	27.0	75.9	51.2
↳ shuffled frame evaluation	Llama3.2-3B	30.4	32.4	54.5	33.8	54.6	28.7	51.9	31.7	25.0	28.4	38.8
Image-trained (1 pseudo video step)	Llama3.2-3B	27.1	45.3	60.2	32.4	54.6	36.2	55.0	35.4	30.0	77.6	45.6
↳ shuffled frame evaluation	Llama3.2-3B	27.6	33.1	56.3	32.9	51.3	29.4	52.5	36.6	31.5	26.3	38.8
Video-trained	Qwen2-7B	33.9	61.5	75.3	36.9	78.4	60.0	65.0	34.1	31.5	78.9	55.8
↳ shuffled frame evaluation	Qwen2-7B	31.7	37.2	54.2	38.2	58.4	30.0	50.0	28.0	30.0	24.1	39.7
Image-trained (1 pseudo video step)	Qwen2-7B	31.0	53.4	66.8	37.8	75.1	49.4	58.1	36.6	32.5	66.8	50.5
↳ shuffled frame evaluation	Qwen2-7B	31.1	33.8	55.6	37.3	51.9	32.5	51.2	36.6	30.5	27.6	40.0

Table 3. 对图像和视频训练的 LLM 在 TVBench [4] 上的表现进行比较。对于 Llama3.2 和 Qwen2，图像训练的 LLM 表现明显高于偶然水平。当视频帧被打乱时，性能的下降表明所有模型确实学习到了一种时间推理的形式，甚至包括那些仅在图像上训练的模型。

Configuration	LLM	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	Avg.
Chance level		25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0	33.3
Videos	Llama3.2-3B	31.0	59.4	70.2	33.3	76.2	44.4	59.1	32.9	27.0	75.9	51.2
Pseudo videos	Llama3.2-3B	31.7	54.1	75.1	38.7	84.9	41.9	57.2	34.1	28.5	83.6	53.5
Videos + pseudo videos	Llama3.2-3B	33.0	53.4	75.3	36.4	83.8	49.4	55.0	30.5	28.0	82.3	53.4
Videos	Qwen2-7B	33.9	61.5	75.3	36.9	78.4	60.0	65.0	34.1	31.5	78.9	55.8
Pseudo videos	Qwen2-7B	32.6	53.4	74.8	39.1	80.0	50.0	60.0	35.4	28.5	58.2	51.9
Videos + pseudo videos	Qwen2-7B	32.6	57.4	75.3	38.7	82.7	52.5	63.8	35.4	27.0	74.6	54.4

Table 4. 我们伪视频训练的 LLM 在最佳配置下取得的结果。令人惊讶的是，这些结果与视频训练的 LLM 相近，甚至更高，这引发了一个问题，即在训练过程中真实视频的使用效果如何。用伪视频微调视频训练的 LLM 产生了不一的结果：在某些任务上的性能有所提升，而在其他任务上则略有下降。

力。微调视频训练版本平均略有提高，结果 (53.4 %) 大致匹配通过微调图像训练模型所达到的性能。然而，我们也注意到，任务“对象计数”(OC)和“动作反义词”(AA)受到了这种额外微调的负面影响，这或许表明存在某种程度的任务特定遗忘。

对于 Qwen2 模型，结果表明伪视频训练效果较差。这部分原因可以归结为 Qwen2 的图像训练基线在 ST 和 AS 任务中已经相当强大，因此受益不大。此外，值得注意的是，MD 性能相比图像基线表现更差，而基于 Llama 的模型则有所改善，尽管这些模型都在相同的图像和伪视频数据上进行了训练。我们鼓励未来的研究进一步探索 LLM 之间的这种差异。对于视频训练的 Qwen2 模型，尽管“场景转换”性能有所提升，我们无法找到能够保持或提高平均准确率的伪视频设置。尽管结果并不意外，最佳伪视频训练的 Qwen2 模型平均准确率为 51.9 %，仅比视频训练的模型低 4 %。考虑到视频模型训练了大约 550,000 个真实视频 [20]，这样的性能差距并不像从如此简单的基线上预期的那么大。为了进一步探索该模型的伪视频训练，我们注意到数据生成过程是灵活的：例如，如果我们希望模型学习场景重现，这一约束可以轻松地合并到 3 节中描述的采样过程中。另一种方法是利用现有的遮罩注释：伪视频中各种模式运动态的物体开启了全新的问题领域，

这可能是该特定模型能够从这种时空推理中获得更多的益处。

虽然基准测试远未解决，但这些结果表明当前对视频大型语言模型 (Video LLM) 的训练程序在从真实视频中学习有用的时间特征方面可能没有我们预期的那样有效。具体而言，我们能够获得接近于——甚至在 Llama3.2 的情况下略好于——通过伪视频而不是捕捉真实视频复杂性的视频训练 LLM 取得的结果。因此，我们预期能够接触到真实视频动态的视频训练 LLM 表现应该会显著更好，但事实并非如此。考虑到在大型视频数据集上训练 LLM 所消耗的大量注释努力、存储需求和计算资源，我们认为查明和解决导致这个简单基线目前如此有竞争力的瓶颈很重要。

至于将伪视频用作额外的训练阶段，结果表明视频训练的模型未必会从中受益，甚至即使它们平均上受益，具体任务的表现仍可能下降。这种类型训练作为中间阶段而不是最后训练阶段的影响仍然是一个开放的问题。

接下来，我们研究伪视频生成的超参数以及训练中使用的问题类型。除非另有说明，这些实验是在 Llama3.2 模型上进行的，因为训练速度明显更快。

Questions posed to the model. 为了消除表 1 中每个问题的影响，我们对每个问题单独训练模型。我们设置

Configuration	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	Avg.
Chance level	25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0	33.3
$R_1, S = 4$	26.9	52.7	72.1	36.9	78.9	34.4	54.7	30.5	28.0	82.8	50.3
$R_2, S = 4$	28.0	53.4	72.5	38.7	62.2	35.6	56.2	30.5	30.0	84.1	50.1
$R_3, S = 6$	31.5	50.0	75.7	28.9	76.8	41.2	55.6	24.4	28.5	84.5	51.4
$R_4, S = 6$	29.7	52.7	74.1	39.1	64.3	39.4	55.9	42.7	30.0	85.8	51.6
$A_1, S = 6$	20.3	45.3	68.6	33.3	61.1	33.8	59.4	41.5	28.5	79.7	46.9
$A_2, S = 6$	28.2	44.6	73.7	32.9	64.9	39.4	56.2	42.7	28.0	78.9	49.5

Table 5. 训练过程中使用的不同问题的效果。总体而言，相对问题比绝对问题表现更好。

Configuration	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	Avg.
Chance level	25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0	33.3
$R_1, F = 1$	25.9	50.0	68.6	35.1	63.2	32.5	59.1	24.4	29.0	73.3	47.4
$R_1, F = 5$	26.9	52.7	72.1	36.9	78.9	34.4	54.7	30.5	28.0	82.3	50.3
$R_1, F = 10$	27.8	50.7	73.2	36.4	78.4	40.0	55.9	35.4	27.5	84.1	51.2
$R_1, F = 20$	27.1	48.6	73.5	40.9	83.8	48.8	58.4	34.1	24.0	82.3	52.2
$R_1, F = 40$	30.0	54.1	73.9	40.4	82.2	41.9	56.2	29.3	24.0	81.9	52.1
$R_3, F = 5$	31.5	50.0	75.7	28.9	76.8	41.2	55.6	24.4	28.5	84.5	51.4
$R_3, F = 20$	31.9	57.4	77.3	36.0	71.9	47.5	54.4	28.0	28.5	75.0	52.0
$R_4, F = 5$	29.7	52.7	74.1	39.1	64.3	39.4	55.9	42.7	30.0	85.8	51.6
$R_4, F = 20$	27.6	50.0	71.8	31.5	64.3	43.7	57.5	35.4	28.0	71.1	48.7

Table 6. 改变每个场景最大帧数的影响。帧数的最佳值以及对此超参数的鲁棒性在一定程度上取决于问题。然而，令人惊讶的是，所有相对时间推理问题都能获得相当高的分数。

$F = 5$ 并根据问题变化场景的数量 S ，因为在某些情况下，这种变化会更加显著地影响可能答案的多样性。例如，问题 A_1 使用 $S = 4$ 和四个候选答案总是会给模型呈现 1, 2, 3, 4 的排列，从而限制了训练样本的多样性。因此，对某些问题我们使用稍高的 $S = 6$ 。

根据表格 5 中的相应结果，我们初步得出结论：对于需要某种绝对时间推理的问题，向视频大规模语言模型（Video LLM）查询的效果较差。这可能是伪视频训练特有的现象，也可能是所使用模型的一般特征。要确定是哪种情况，需要研究类似问题与真实视频的配对。进一步研究这一问题的另一种方法是构建一个视频基准测试，其中包含针对这些推理类型的任务。这将有助于确定视频大规模语言模型在推理期间是否也发现绝对时间推理更具挑战性。

每个场景的帧数部分决定了创建的伪视频的长度。在这个实验中，我们专注于 R_1 , R_3 和 R_4 ，因为在 Table 5 中取得了令人鼓舞的结果。从表 6 中可以观察到，使用 R_1 时，性能改善直到 $F = 20$ ，此时稍微下降。最大改善发生在 $F = 5$ ：与 $F = 1$ 相比，这使得一个场景平均持续时间超过一帧（平均 3 帧），这可能让我们的伪视频比单纯的完全不同图像序列更接近真实视频。考虑到其他问题类型， R_3 似乎也受益，尽管在“场景转换”和“移动方向”上获得的提升较小。然而， R_4 随着更多帧的增加表现出显著的性能下降。我们还报告了 ($R_1 : F = 20, S = 4$) 和 ($R_3 : F = 20, S = 6$)

的混合达到 51.5%，而表 4 中的最佳结果是 53.5%，其使用了 $F = 5$ 来处理 R_3 。由此我们总结出最佳的 F 值在不同问题和问题组合之间是变化的。

除了影响伪视频长度外，场景的数量对视频和问题的复杂性影响要超过 F 。当在问题 R_1 中设置 $S = 8$ 时，模型被要求将最多 8 个场景按正确的出现顺序排列。我们研究了通过固定 $F = 5$ 来替换 R_1 变更 S 的效果，除了在 $S = 2$ 中我们也尝试了 $F = 10$ 。我们这样做是为了排除由于 $S = 2, F = 5$ 导致的性能差异，相较于 $S = 4, F = 5$ 产生更小的平均值和最大帧数（最大为 10）。（最大为 20）。表 7 显示了从 $S = 2$ 到 $S = 4$ 的显著性能提升，特别是在“场景过渡”方面，而从 $S = 4$ 到 $S = 8$ 则略有下降。我们得出结论，该模型不仅仅学习了两个不同场景的简单排序，这些知识对于“场景过渡”任务特别有用。

在这里，我们使用 R_1 与 $S = 4, F = 5$ ，并在生成伪视频的数量变化的情况下训练等效于两个 epoch。从表格 8 的结果中，我们得出结论：在超过 100,000 个伪视频上训练可以提高“场景转换”任务的性能，这与训练期间提出的问题类似。然而，总体性能下降，这表明模型开始在特定任务上过拟合。除了这些结果之外，我们注意到，在较早的实验轮次中——尽管在不太严格的设置下——我们从未观察到超过大约 3,000 步/100,000 个伪视频的稳定性提升。

Configuration	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	Avg.
Chance level	25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0	33.3
$S = 2, F = 5$	28.5	49.3	72.3	34.2	60.0	35.0	57.5	28.0	33.5	80.2	49.3
$S = 2, F = 10$	26.9	49.3	72.3	31.5	60.0	35.6	57.5	25.6	31.0	81.5	48.6
$S = 4, F = 5$	26.9	52.7	72.1	36.9	78.9	34.4	54.7	30.5	28.0	82.3	50.3
$S = 8, F = 5$	27.1	49.3	73.2	35.5	77.3	36.2	57.8	25.6	23.5	82.7	50.0

Table 7. 对于问题 R_1 ，改变场景的最大数量的影响。对于“场景转换”任务，提供多于两个场景特别重要，但没有一个特定的设置能够在所有任务中表现同样良好。

Configuration	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	Avg.
Chance level	25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0	33.3
$N = 10,000$; 312 steps	26.5	48.6	72.1	32.9	68.1	32.5	57.8	30.5	28.5	84.1	49.2
$N = 100,000$; 3,125 steps	26.9	52.7	72.1	36.9	78.9	34.4	54.7	30.5	28.0	82.3	50.3
$N = 500,000$; 15,625 steps	28.4	50.7	72.3	36.9	82.2	34.4	55.6	31.7	21.5	64.7	48.7

Table 8. 训练时改变伪视频最大数量 N 对问题 R_1 的影响。再次看来，似乎有特定任务的最优值，但特别是在“自我中心序列”和“移动方向”任务中，使用更多数据进行训练明显效果更差。

5. 结论

我们展示了实验结果，这些结果对当前视频大语言模型训练方案的有效性提出了质疑。首先，我们展示了基于近期 LongVU 方法的两个模型仅使用图像和文本进行训练却在时间推理方面出乎意料地有效。这些模型从未在真实视频上进行训练，但在最近评估视频中各种时间推理方面的 TVBench 基准测试中，其准确率显著高于随机水平。其次，我们研究了一种简单的设置：将这些模型在通过连接和扰动 COCO 图像构造的伪视频上进行微调。当训练涉及需要推理伪视频场景相对顺序的某些问题时，我们观察到的性能提升与甚至高于在真实视频上进行微调所达到的效果。

视频训练的 Video LLM 与我们的基准之间的小差距表明，在训练期间更好地利用图像数据集可能降低对视频的需求。这也表明，虽然真实视频信息更丰富，但在 Video LLM 的时间理解上可能贡献小于预期。这指向了当前范式中的一个瓶颈，可能是由于视频和字幕数据的质量、训练设置或架构弱点。由于随着模型和数据集的增大，训练 Video LLM 变得越来越昂贵，研究这个瓶颈对于高效利用计算资源至关重要。

涉及更多视频基准和其他视频 LLM 架构的实验可能会进一步巩固我们的发现。对于特别需要理解复杂视频动态的任务，以及涉及潜在长达数小时的视频的长期设置，比较伪视频与视频训练将是特别有趣的。

Acknowledgments. 本项目由 BMBF 项目 NeuroSys-D (03ZU1106DA) 和 6GEM (16KISK036K) 部分资助。计算资源是通过约翰·冯·诺依曼计算研究所由高斯超级计算中心 e.V. 在利希超级计算中心的 GCS 超算机 JUWELS [9] 上提供的。

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923, 2025.
- [2] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. In NeurIPS, 2024.
- [3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. arXiv preprint arXiv:2406.07476, 2024.
- [4] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in Time: A New Temporal Benchmark for VideoLLMs. arXiv preprint arXiv:2410.07752, 2025.
- [5] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024.
- [6] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In CVPR, 2024.
- [7] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In ICLR, 2022.
- [8] Justin Johnson, Bharath Hariharan, Laurens Van

- Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In CVPR, 2017.
- [9] Jülich Supercomputing Centre. JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, 7 (A138), 2021.
- [10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy Visual Task Transfer. arXiv preprint arXiv:2408.03326, 2024.
- [11] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In CVPR, 2024.
- [12] Rui Li, Xiaohan Wang, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Temporal Preference Optimization for Long-Form Video Understanding. arXiv preprint arXiv:2501.13919, 2025.
- [13] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. In ECCV, 2024.
- [14] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In EMNLP, 2024.
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, 2014.
- [16] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution. arXiv preprint arXiv:2409.12961, 2024.
- [17] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In ACL, 2024.
- [18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. TMLR, 2024.
- [19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In ICLR, 2021.
- [20] Xiaoqian Shen, Yunsang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. arXiv preprint arXiv:2410.17434, 2024.
- [21] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for Training and Evaluating Large Video Description Models. arXiv preprint arXiv:2407.00634, 2024.
- [22] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. arXiv preprint arXiv:2404.16994, 2024.
- [23] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: CoLLision Events for Video REpresentation and Reasoning . In ICLR, 2019.
- [24] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. arXiv preprint arXiv:2501.07888, 2025.
- [25] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In EMNLP, 2023.
- [26] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video Instruction Tuning With Synthetic Data. arXiv preprint arXiv:2410.02713, 2024.