Vision Experiments

- 图像数据处理的四个关键阶段：质量过滤、感知去重、重采样和光学字符识别（OCR），并强调了安全措施的应用。

- 质量过滤：移除非英语和低质量字幕的图像-文本对，使用CLIP分数（由Radford等人提出的一种衡量图像和文本对之间语义对齐程度的分数）进行筛选，确保数据质量。
- 去重：通过使用先进的SSCD复制检测模型对图像进行去重，以减少冗余数据，提高训练效率，并保护隐私。重复图像通过连接组件算法分组，仅保留每组中的一个图像-文本对。
- 重采样：通过对n-gram词汇（n-gram：一个文本片段中连续出现的n个词语或字符的序列。）进行频率分析，确保数据集的多样性。对低频类别和细粒度识别任务有帮助，通过重采样保留稀有n-gram对应的图像-文本对。
- 光学字符识别（OCR）：通过提取图像中的文本并将其与字幕结合，提升对需要OCR能力的任务的性能，如文档理解。使用专有的OCR管道进行文本提取。

SSCD去重

- SSCD（Self-Supervised Contrastive Deduplication）复制检测模型是一个先进的工具，用于在大规模数据集中识别重复的图像。其工作原理是首先将每个图像转化为一个512维的特征向量，然后通过最近邻搜索和余弦相似度计算来查找相似的图像。如果两个图像的相似度高于某个阈值，它们被认为是重复的。随后，使用连接组件算法将这些重复的图像分组，每组只保留一个图像-文本对。

重采样

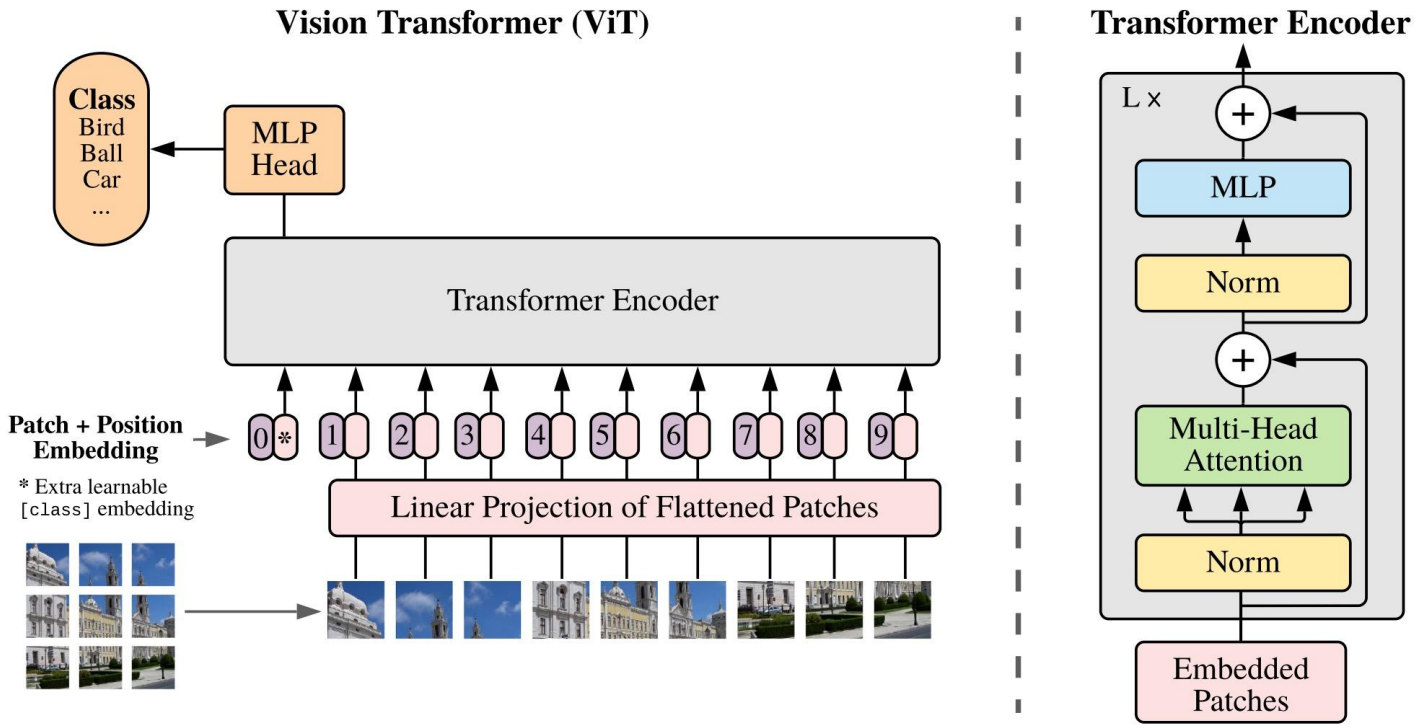
- 首先建立一个包含n-gram的词汇表，然后计算每个n-gram在数据集中的出现频率。如果某个图像-文本对中的n-gram出现频率较低，则保留该对；反之，如果n-gram频率较高，则根据一个概率模型（由n-gram的频率决定）来决定是否保留该图像-文本对。

视频数据

- 文本清理和过滤：通过规则和语言识别模型确保文本的质量和语言一致性。（例如确保文本的最小长度并修正大写问题。）
- OCR过滤：移除那些带有过多叠加文本的视频。
- 对齐性检测：使用对比模型来确保视频和文本之间的内容匹配度，通过图像-文本和视频-文本相似度来筛选配对。
- 运动分数过滤：去除静态或低运动的视频，确保视频内容的动态性。
- 视频质量：虽然进行了上述过滤，但没有对视频的视觉美感或分辨率进行限制，确保多样性。
- 最终的数据集包含了时长主要在一分钟以内、分辨率从320p到4K不等的视频，并且视频的纵横比多样化但大多集中在1:2到2:1之间。
- #：没有对视频的视觉质量（如美学评分或分辨率过滤）应用任何过滤。

视觉识别模型架构-图像编码器

- 图像编码器是一个标准的视觉变换器(ViT；Dosovitskiy等人2020)，训练用于对齐图像和文本(Xu等人，2023)，具体而言
- 使用了ViT-H/14变体的图像编码器，它有630M——6.3亿参数，在25亿图像-文本对上训练了五个epoch



- 传统的对比性文本对齐目标训练方法难以保留细粒度的定位信息。为了解决这个问题，模型采用了多层特征提取，不仅使用最终层的特征，还额外使用了第4、8、16、24和31层的特征。
- 额外插入了8个门控自注意力层，以学习特定的对齐特征，这使得图像编码器最终具有850M个参数。编码器产生了7680维的特征表示，每个16 × 16的图像块都有对应的特征向量。
- 在随后的训练阶段，图像编码器的参数没有被冻结，这提高了性能，尤其是在文本识别领域。

#：图像适配器（Image Adapter）

- 1. 跨注意力层的引入

- 跨注意力层的主要目的是将视觉标记表示与语言模型生成的标记表示结合起来。这涉及到以下几个步骤：
- 视觉标记表示：图像编码器将输入的图像转换为一系列视觉标记（tokens），这些标记是图像内容的特征表示。
- 语言标记表示：语言模型将文本数据转换为一系列语言标记，这些标记表示文本的语义信息。
- 跨注意力机制：在跨注意力层中，模型通过关注（Attention）机制来学习视觉标记和语言标记之间的关系。具体来说，视觉标记和语言标记会相互作用，模型会计算它们之间的相关性，从而生成融合的多模态表示。这一过程通过跨注意力机制实现，其中视觉标记作为查询（query），语言标记作为键（key）和值（value）输入到注意力机制中。
- 应用位置：这些跨注意力层被放置在语言模型的自注意力层之后，每四个自注意力层后添加一个跨注意力层。这种设计有助于模型在处理文本的过程中逐渐融合图像信息，使得文本处理更具有上下文的视觉语义。

#：图像适配器（Image Adapter）

2. 广义查询注意力（GQA）

- 广义查询注意力（GQA）是一种优化后的注意力机制，它的目的是提高模型的计算效率。在标准的注意力机制中，每个查询都会与所有键进行点积运算，计算复杂度很高。GQA通过对查询进行优化，减少了不必要的计算，从而提高了效率。
- 优化点：GQA通过将查询参数化为更具通用性的形式，减少了查询的维度或计算量。这使得模型在计算视觉和语言标记之间的注意力分数时更加高效，特别是在处理大规模数据时，能显著降低计算开销。

3. 大量可训练参数

- 跨注意力层的引入大大增加了模型的参数量。以Llama 3 405B模型为例，跨注意力层约有100B个参数。这些额外的参数使得模型能够更好地学习和表示视觉-语言之间的复杂关系，从而提高多模态任务的表现。

#：图像适配器（Image Adapter）

- 预训练：
- 图像适配器的预训练分为两个阶段，旨在确保模型能够有效学习视觉和语言之间的关联。

- 初始预训练：
- 数据集：使用大约6B图像-文本对的数据集进行预训练。为了提高计算效率，所有图像被调整为四块336 × 336像素的瓷砖大小，这样可以适应不同的纵横比。
- 目的：在这一阶段，模型主要学习图像和文本之间的初步对齐关系，建立基础的视觉-语言融合能力。

- 退火训练：
- 继续训练：在约500M张图像上继续训练，这一阶段被称为退火训练。
- 提高分辨率：退火训练时，提高每块图像的分辨率以更好地处理需要更高分辨率的任务，如信息图表理解。通过逐渐提高图像分辨率，模型能够更精确地捕捉细节，从而改善在需要高分辨率的任务中的表现。

#：视频适配器（Video Adapter）

- 原理与架构：

- 视频适配器处理视频输入，最多支持64帧视频（从整个视频中均匀采样）。每帧通过图像编码器处理后，再通过时间聚合器和额外的视频跨注意力层来建模视频的时间结构。
- 时间聚合器将连续32帧合并为一帧，并使用了一种称为Perceiver重采样的技术。视频跨注意力层则在每四个图像跨注意力层之前添加。
- 在预训练阶段，每个视频使用16帧进行处理，并聚合为1帧；在有监督微调时，输入帧数增加到64帧。视频适配器和跨注意力层的参数分别为0.6B和4.6B（针对Llama 3 7B和70B模型）。

Pre-training-图像预训练

- 初始化：
- 模型的初始化使用了预训练的文本模型和视觉编码器的权重。视觉编码器的参数是可训练的（即未冻结），而文本模型的权重则被冻结，不参与训练。

- 初始训练：
- 使用了6B（60亿）图像-文本对进行训练。每个图像被调整为四块336 × 336像素的瓷砖大小，以适应模型的输入。
- 训练采用的全局批次大小为16,384。
- 学习率采用余弦衰减策略，初始学习率为10×10^{-4}（0.001），权重衰减系数为0.01。这些学习率是在小规模实验的基础上确定的，但在更长的训练过程中，学习率在损失值停滞时被多次下调。
- 在基本预训练完成后，进一步提高图像分辨率，并在退火数据集（annealing dataset）上继续训练相同的权重。这一阶段的优化器重新初始化，学习率为2×10^{-5}（0.00002），并再次遵循余弦衰减策略。

视频预训练（Video Pre-Training）

- 初始化：
- 视频预训练从已经完成图像预训练和退火后的权重开始。这意味着在开始视频预训练时，模型已经具备了对图像的强大识别能力。
- 在视频架构中新增的视频聚合器和跨注意力层是随机初始化的，所有与视频相关的参数都是可训练的，而其他所有参数则被冻结。

- 训练策略：
- 视频预训练使用与图像退火阶段相同的超参数设置，唯一的差异在于学习率的微调。
- 从完整的视频中均匀采样16帧，每帧被表示为四块448 × 448像素的图像块。通过视频聚合器将16帧聚合成一个有效帧，然后文本标记通过跨注意力机制与这一帧交互。
- 训练时的全局批次大小为4,096，序列长度为190个标记，学习率设定为10^{-4}（0.0001）。

Post-Training

- 在预训练后如何通过一系列精细调整（fine-tuning）和优化步骤来提高其性能，特别是在多模态对话和推理任务中的表现。
- 1. 视觉适配器的后续微调和优化
- 在预训练之后，模型需要进一步微调，以便在特定的任务和数据集上表现更好。这个过程包括几个关键阶段：

- 微调多模态对话数据：模型首先在经过精挑细选的多模态对话数据集上进行微调，以增强其聊天能力。这些数据集涵盖了各种对话形式，包括开放式问答、描述性对话以及实际应用场景。

- 直接偏好优化（DPO）：为了提升模型在人类评估中的表现，使用了直接偏好优化技术。通过对比不同模型输出的优劣（由人类标注），模型能够学会更符合人类偏好的生成方式。

- 质量微调：在最终阶段，模型会在一个小而高质量的数据集上进一步微调。这个阶段旨在提高模型的响应质量，同时保持在各类基准测试中的表现。

2. 监督微调数据

- 图像数据的监督微调（SFT）
- 学术数据集：将现有的学术数据集转化为问答对。这些数据经过过滤和重写，以提升语言质量和指令多样性。
- 人工注释：通过人工注释获得多模态对话数据，涵盖了广泛的任务和领域。注释员被要求基于图像撰写对话，并使用中间的模型检查点生成的内容作为起点进行人类编辑。这种迭代式注释方法提高了数据的质量和效率。
- 合成数据：通过使用语言模型生成合成的多模态数据，如将文字数据转化为图像，或从现有图像中提取文本进行OCR（光学字符识别），生成额外的对话或问答数据。

- 视频数据的监督微调
- 类似于图像适配器的处理方式，使用带有预先注释的学术数据集并将其转化为合适的文本指令和目标响应。人类注释员被要求注释视频中的问题和答案，特别是那些需要时间理解的问题。

- 3. 监督微调方法（SFT Recipe）
- 图像微调
- 模型从预训练的图像适配器初始化，但使用指令调优后的语言模型权重替换预训练的语言模型权重。语言模型的权重被冻结，只更新视觉编码器和图像适配器的权重。
- 采用超参数搜索方法，通过多次实验确定最佳的学习率和权重衰减值，最后通过对多个顶级模型的权重进行平均，来获得最终的模型。这种方法能够减少对超参数的敏感性，并提升整体表现。
- 视频微调
- 视频聚合器和跨注意力层使用预训练的权重初始化，其他参数则从相应的微调模型中继承。然后仅在视频SFT数据上微调视频参数，同时增加视频长度和分辨率以与图像超参数保持一致。

- 4. 偏好数据
- 偏好数据用于奖励模型和直接偏好优化的训练，包括：

- 人类注释：人类注释员比较不同模型输出，并给出“选择”和“拒绝”的标签，同时有7级评分系统。注释员还可能提供编辑建议，以修正模型生成的错误。

- 合成数据：通过使用语言模型故意在SFT数据中引入错误，生成合成的偏好对。这些对比数据用于训练模型更好地区分正确和错误的生成内容。

- 5. 拒绝采样（Rejection Sampling）
- 为了增强模型的推理能力，特别是添加缺失的“思路链”（chain-of-thought）解释，使用拒绝采样技术。通过多次采样生成多个答案，并使用启发式方法或语言模型来判断最优答案，最后将高质量答案重新添加到训练数据中。

- 6. 奖励建模（Reward Modeling）
- 在视觉SFT模型的基础上训练视觉奖励模型（RM），以增强模型的判断能力，特别是在基于图像内容进行判断时。语言RM部分的参数被冻结，以保持准确性，同时加入正则化项来防止奖励分数的漂移。

- 7. 质量调优（Quality Tuning）
- 通过在高质量的小数据集上进一步训练DPO模型，提高模型的响应质量。这个阶段关注人类评估表现的提升，同时通过基准测试确保模型能力保持或增强。

加上了视觉能力的llama3，其在图像理解层面的性能在与相关模型PK时，得到的结果如下

	Llama 3-V 8B	Llama 3-V 70B	Llama 3-V 405B	GPT-4V	GPT-4o	Gemini 1.5 Pro	Claude 3.5
MMMU (val, CoT)	49.6	60.6	64.5	56.4	**69.1**	62.2	68.3
VQAv2 (test-dev)	78.0	79.1	**80.2**	77.2	–	**80.2**	–
AI2 Diagram (test)	84.4	93.0	94.1	78.2	94.2	94.4	**94.7**
ChartQA (test, CoT)	78.7	83.2	85.8	78.4	85.7	87.2	**90.8**
TextVQA (val)	78.2	83.4	**84.8**	78.0	–	78.7	–
DocVQA (test)	84.4	92.2	92.6	88.4	92.8	93.1△	**95.2**

评估任务概述

- MMMU (Yue et al., 2024a): 多模态推理的挑战性数据集，要求模型理解图像并解决涵盖30个不同学科的大学水平问题。问题形式包括多项选择题和开放性问题。评估是在包含900张图片的验证集上进行的。
- VQAv2 (Antol et al., 2015): 这个数据集测试模型结合图像理解、语言理解和常识知识来回答关于自然图像的一般性问题的能力。VQAv2是视觉问答领域的标准基准之一。
- AI2 Diagram (Kembhavi et al., 2016): 评估模型解析科学图表并回答相关问题的能力。这个数据集使用与Gemini和x.ai相同的评估协议，并使用透明边界框报告分数。
- ChartQA (Masry et al., 2022): 要求模型能够视觉上理解不同类型的图表并回答关于图表的逻辑问题。
- TextVQA (Singh et al., 2019): 要求模型读取和推理图像中的文本，以回答相关问题，测试模型的OCR（光学字符识别）理解能力。
- DocVQA (Mathew et al., 2020): 专注于文档分析和识别的基准数据集，包含各种文档的图像，用于评估模型执行OCR理解和推理文档内容以回答相关问题的能力。

Video Recognition Results

	Llama 3-V 8B	Llama 3-V 70B	Gemini 1.0 Pro	Gemini 1.0 Ultra	Gemini 1.5 Pro	GPT-4V	GPT-4o
PerceptionTest (test)	53.8	**60.8**	51.1	54.7	–	–	–
TVQA (val)	82.5	**87.9**	–	–	–	87.3	–
NExT-QA (test)	27.3	**30.3**	28.0	29.9	–	–	–
ActivityNet-QA (test)	52.7	56.3	49.8	52.2	57.5	–	**61.9**

Table 30 Video understanding performance of our vision module attached to Llama 3. We find that across range of tasks covering long-form and temporal video understanding, our vision adapters for Llama3 8B and 70B parameters are competitive and sometimes even outperform alternative models.

评估任务概述

- PerceptionTest (Pătrăucean et al., 2023):
- 这个基准测试评估模型回答时间推理问题的能力，涵盖技能（如记忆、抽象、物理、语义）和不同类型的推理（描述性、解释性、预测性、反事实）。测试集包含11,600个问答对，每个视频平均时长为23秒。评估使用的是多项选择题，报告的性能基于提交至在线挑战服务器的测试集预测结果。
- NExT-QA (Xiao et al., 2021):
- 这个基准测试主要关注开放式问答，测试模型在时间和因果推理方面的能力。包含1,000个测试视频，每个视频平均时长为44秒，配有9,000个问题。评估通过Wu-Palmer相似度（WUPS）度量模型的回答与标准答案的相似度。
- TVQA (Lei et al., 2018):
- 这个基准测试评估模型的组合推理能力，要求模型进行时空定位、视觉概念识别，并结合字幕对话进行推理。数据集来自流行的电视剧，测试模型在利用这些电视剧外部知识回答问题的能力。数据集包含超过15,000个验证问答对，每个视频片段平均时长为76秒，采用多项选择题形式。
- ActivityNet-QA (Yu et al., 2019):
- 这个基准测试评估模型对长视频片段的推理能力，包括动作理解、空间关系、时间关系、计数等。包含8,000个测试问答对，来自800个视频，每个视频平均时长为3分钟。评估时，模型生成简短的一字或短语答案，正确性通过GPT-3.5 API与标准答案进行对比，报告平均准确率。