MedTrinity-25M
发布时间：2024-08-06
发布机构：华中科技大学、加州大学、哈佛大学、斯坦福大学
项目主页: https://yunfeixie233.github.io/MedTrinity-25M/
数据集说明：MedTrinity-25M 是一个大规模多模态医学数据集，包含超过 2500 万张图像，涉及 10 种模态和 65 种疾病。数据集通过自动化的数据构建流程生成，不依赖于配对的文本描述，而是通过专家模型和知识库增强的多模态大型语言模型生成多粒度视觉和文本注释。数据集的创建过程包括从 90 多个在线资源收集数据，应用专家模型识别感兴趣区域（ROIs），并构建知识库以生成详细的文本描述。MedTrinity-25M 旨在支持广泛的医学多模态任务，如图像标注和报告生成，以及视觉中心的任务如分类和分割，推动医学领域基础模型的发展。

MMSci
发布时间：2024-07-06
发布机构：加利福尼亚大学等
地址：https://github.com/Leezekun/MMSci
数据集说明：MMSci 数据集是一个多模态、多学科的高质量学术文章和图表集合，涵盖 72 个科学领域。数据集包含 131,393 篇文章和 742,273 个图表，主要来源于 Nature Communications 期刊。创建过程中，数据集通过爬取开放获取的文章和图表，确保了数据的真实性和高质量。该数据集主要用于评估和提升大型多模态模型在科学领域的理解和应用能力，特别是在理解和生成科学图表方面。

OBIMD
发布时间：2024-07-04
发布机构：安阳师范学院、腾讯等
项目主页:https://www.jgwlbq.org.cn/dt/oracleFragment
数据集说明：甲骨文多模态数据集（Oracle Bone Inscriptions Multi-modal Dataset - OBIMD）共包含一万片甲骨的拓片、摹本，甲骨单字对应位置、对应字头、对应释文以及辞例分组、释读顺序等数据。据介绍，所有研究者都能基于该数据集研发甲骨文检测、识别、摹本生成、字形匹配和释读等算法，加速甲骨文研究智能化进程。

OpenVid-1M
发布时间：2024-07-02
发布机构：南京大学、字节跳动、南开大学
项目主页: https://nju-pcalab.github.io/projects/openvid/
数据集说明：OpenVid-1M 是由南京大学、字节跳动和南开大学联合创建的一个大规模高质量文本到视频生成数据集。该数据集包含超过 100 万个视频片段，每个视频具有至少 512×512 的高分辨率，并配有详细的字幕。数据集的创建过程严格筛选了美学、时间一致性、运动差异和清晰度等方面，确保了视频的高质量。OpenVid-1M 主要应用于文本到视频生成领域，旨在解决现有数据集质量不高或过于庞大的问题，推动高清晰度视频生成技术的发展。

MINT-1T
发布时间：2024-06-17
发布机构：华盛顿大学、Salesforce Research、斯坦福大学
项目主页: https://github.com/mlfoundations/MINT-1T
数据集说明：MINT-1T 是由华盛顿大学和 Salesforce Research 合作创建的开放源代码多模态交错数据集，包含一万亿文本令牌和三十亿图像，是目前最大且最多样化的开放源代码多模态交错数据集。数据集内容丰富，涵盖 HTML、PDF 和 ArXiv 等多种来源，旨在通过提供大规

模、多样化的训练数据，推动前沿大型多模态模型（LMMs）的发展，解决现有开放源代码多模态数据集规模和多样性不足的问题。

Touch100k
发布时间：2024-06-06
发布机构：北京交通大学、腾讯微信 AI 团队、北京邮电大学
项目主页: https://cocacola-lab.github.io/Touch100k/
数据集说明：Touch100k 数据集是北京交通大学联合腾讯微信 AI 团队及北京邮电大学构建的一个大规模触觉-语言-视觉多模态数据集。该数据集包含了 10 万个与触觉、视觉和语言描述相关联的样本，这些样本描述了不同粒度的触觉感受，比如句子级别的自然表达和短语级别的关键特征描述。研究人员首先从公开的触觉数据集中收集和整理了视觉-触觉观察结果，然后使用 GPT-4V 生成了多粒度的文本描述，并通过多步骤的质量增强过程确保了数据的准确性和实用性。Touch100k 数据集以其丰富的触觉感知描述，为机器人学和人工智能领域提供了宝贵的资源。

Video-MME
发布时间：2024-06-03
发布机构：北京大学、香港大学等
项目主页: https://video-mme.github.io/
数据集说明：Video-MME 是北京大学、香港大学等 6 所高校联手，发布的首个专为视频分析设计的多模态大模型评估基准。该数据集包含 900 个视频，总时长达 256 小时，研究人员通过反复观看视频内容，手动选择和注释共设计了 2,700 个高质量的多选题。数据集涵盖 6 大视觉领域，包括知识、电影与电视、体育竞赛、艺术表演、生活记录和多语言，并进一步细分为天文学、科技、纪录片等 30 个类别，视频长度从 11 秒到 1 小时不等。此外，Video-MME 还整合字幕和音频轨道，增强了对视频理解的多模态输入分析。更难能可贵的是，Video-MME 中所有数据，包括问答、视频、字幕和音频，都是手工收集和整理的，确保了该基准的高质量。该数据集的创建不仅为研究人员提供了一个富有挑战性的测试基准，也为研究外部信息对视频理解性能的影响提供了宝贵的资源。

MultiOOD
发布时间：2024-05-27
发布机构：苏黎世联邦理工学院、南加州大学、洛桑联邦理工学院
项目主页: https://github.com/donghao51/MultiOOD
数据集说明：MultiOOD 基准是由苏黎世联邦理工学院、南加州大学和洛桑联邦理工学院的研究人员联合创建的多模态异常检测数据集。该数据集数据源于五个公开的动作识别数据集（HMDB51、UCF101、EPIC-Kitchens、HAC 和 Kinetics-600），共计超过 85,000 个视频片段，这些数据集在类别数量和大小上各不相同，类别数从 7 到 229 不等，数据集大小从 3,000 到 57,000 不等。该数据集使用了视频、光流和音频作为不同的模态类型。MultiOOD 是一个创新的基准数据集，它通过结合多种类型的数据（视频、光流和音频），为研究人员提供了一个更为全面的数据资源来开发和测试异常检测类算法。

RLAIF-V-Dataset
发布时间：2024-05-19
发布机构：OpenBMB
项目主页: https://github.com/RLHF-V/RLAIF-V
数据集说明：RLAIF-V-Dataset 是 OpenBMB 构建的一个大规模多模态偏好数据集。该数据集是由 AI 生成的偏好数据集，涵盖各种任务和领域，包含 44,757 组高质量对比对。RLAIF-V-数

据集通过一个新颖的方法，采用开源大模型来对模型响应进行去混杂处理，并提供高质量的反馈。该数据集应用在了 MiniCPM-Llama3-V 2.5 模型的训练中，MiniCPM-Llama3-V 2.5 是第一个具有 GPT-4V 性能的端侧多模态大模型。RLAIF-V-Dataset 数据集可以有效减少多模态大模型的幻觉。

AnyWord-3M
发布时间：2024-04-18
发布机构：阿里巴巴
项目主页: https://github.com/tyxsspa/AnyText
数据集说明：目前，针对文字生成任务的公开数据集尤其是涉及非拉丁语系语言的，还相对缺乏。因此，我们提出了一个大规模多语言数据集 AnyWord-3M。数据集中的图片的来源包括 Noah-Wukong、LAION-400M 以及用于 OCR 识别任务的数据集，如 ArT、COCO-Text、RCTW、LSVT、MLT、MTWI、ReCTS 等。这些图片涵盖了包含文本的多种场景，包括街景、书籍封面、广告、海报、电影帧等。除了 OCR 数据集直接使用标注的信息外，所有其他图片都通过使用 PP-OCR 的检测和识别模型进行处理。然后，使用 BLIP-2 生成文本描述。通过严格的过滤规则和细致的后处理，我们共获得了 3,034,486 张图片，包含超过 900 万行文本和超过 2000 万个字符或拉丁文字。

RELI11D
发布时间：2024-03-28
发布机构：厦门大学、上海科技大学
项目主页: http://www.lidarhumanmotion.net/reli11d/
数据集说明：针对复杂且快速的全局人体动作捕捉问题，厦门大学联合上海科技大学基于激光雷达、IMU、RGB 相机和事件相机构建了多模态人体运动数据集-RELI11D。该数据集包含 10 名采集者在 7 个不同的真实体育场景中进行的 5 项体育运动（乒乓球、跆拳道、拳击、击剑和羽毛球）的 3.32 小时同步 LiDAR 点云、IMU 测量数据、RGB 视频和事件流。总计包含 199.26 分钟的视频数据，涵盖了 239k 帧的人体点云数据。RELI11D 数据集是一个高质量的多模态人体运动数据集，为人体运动估计任务提供了丰富的基准测试数据，它通过多模态数据的融合，为全面理解人体运动提供了新的视角。

NineRec
发布时间：2024-03-17
发布机构：西湖大学
项目主页: https://github.com/westlake-repl/NineRec
数据集说明：NineRec 是西湖大学提出的一个大规模、多样性的推荐系统评估基准数据集，旨在解决推荐系统领域迁移学习模型发展的瓶颈问题，尤其是缺乏大规模、高质量的迁移学习推荐数据集和基准测试套件。NineRec 包含一个大规模源域数据集和九个多样化的目标域数据集，涵盖短视频、新闻、图像等多种类型的原始内容。每条数据均配有描述性文本和高分辨率封面图像，使得模型能够通过学习原始多模态特征而非仅依赖预提取的特征来进行训练。NineRec 的丰富视觉与语义多样性，为推荐模型的可迁移性研究提供了宝贵的预训练资源，同时揭示了 TransRec 模型在跨界推荐任务中的潜力与挑战。

AlgoPuzzleVQA
发布时间：2024-03-13
发布机构：北新加坡科技设计大学
项目主页: https://algopuzzlevqa.github.io/
数据集说明：AlgoPuzzleVQA 是由新加坡科技设计大学构建的一个多模态推理数据集，旨在挑

战和评估多模态语言模型在解决需要视觉理解、语言理解和复杂算法推理的算法谜题方面的能力。数据集包含 18 种不同的谜题，涵盖了诸如布尔逻辑、组合学、图论、优化、搜索等多样化的数学和算法主题。该数据集通过自动化的方式从人类编写的代码生成谜题，确保了数据集可以任意扩展推理复杂性和数据集大小。这些谜题都是有确切解决方案的，可以通过算法找到，无需繁琐的人工计算。AlgoPuzzleVQA 可以作为多模态推理能力的基准测试，用于评估和推动多模态语言模型在解决结合视觉、语言理解和算法推理的复杂问题方面的能力。

MIntRec2.0
发布时间：2024-01-16 发布机构：清华大学等 项目主页：https://github.com/thuiar/MIntRec2.0
数据集说明：MIntRec2.0 是清华大学等提出的一个大规模多模态多方基准数据集，专门用于识别对话中的意图和检测非意图内容。相较于先前的 MIntRec，MIntRec2.0 的数据量增至 15K，涵盖 30 种意图类别，并包含约 9.3K 个意图内及 5.7K 个意图外的标注语句，涉及文本、视频和音频等多种模态。该数据集由 1,245 个对话组成，每个对话平均 12 个语句，每个语句均配有意图标签，且每个对话至少涉及两位发言者，所有语句均标记发言者身份。此外，针对开放世界场景的需求，MIntRec2.0 引入 OOS 标签，用于识别不属于已知意图类别的语句，以增强系统的鲁棒性。该数据集旨在促进多模态意图理解相关研究，为实现更自然的人机交互并通往 AGI 之路奠定坚实基础。

在这里插入图片描述

CapsFusion-120M
发布时间：2024-01-08
发布机构：清华大学、北京智源人工智能研究院
项目主页: https://github.com/baaivision/CapsFusion
数据集说明：该数据集是清华大学和北京智源人工智能研究院于 2024 年推出的多模态图文数据集。该数据集可用于大规模多模态预训练的高质量资源。此版本包含来自 LAION-2B 和 LAION-COCO 数据集的相应字幕，便于进行比较分析和进一步深入研究图像文本数据的质量。每个数据条目有四个字段：图片网址、LAION-2B 标题（来自网络的原始替代文本）、LAION-COCO 字幕（由 BLIP 合成）、CapsFusion 标题（研究团队的）。

Multimodal C4 (mmc4)
发布时间：2023-04-14
发布机构：加州大学、华盛顿大学、艾伦人工智能研究所
项目主页: https://github.com/allenai/mmc4?tab=readme-ov-file
数据集说明：Multimodal C4 的数据集，语料库包含 103M 文档，其中包含了 585M 张图片和 43B 个英文单词，这些图片和文字相互交织。通过该数据集进行训练，可以更好地实现多模态的上下文学习，这对于未来更加丰富的多模态语言技术的发展非常重要。此外，还对数据集进行了详细的分析和筛选，确保了其中的图片和文字具有高度相关性。

ShareGPT4V
发布时间：2023-11-21
发布机构：中国科学技术大学、上海人工智能实验室
项目主页: https://sharegpt4v.github.io/
数据集说明：ShareGPT4V 数据集是一个由大量图像-文本对组成的高质量数据集，它被用于训练视觉-语言模型（VLM），以提高模型在图像理解和文本生成方面的能力。该数据集包含 120 万对图像-文本配对，这些数据有效地对齐了视觉和语言特征，增强了模型遵循指令的能力，并纳入了更多学术任务，例如 ScienceQA 、 TextVQA 、 SBU 等。通过引入这个数据集，模型在图像-文本对齐能力方面得到了显著提升，这对于多模态表示学习是一个关键方面。

LAION-5B
发布时间：2022-03-31
发布机构：LAION
项目主页: https://laion.ai/blog/laion-5b/
数据集说明：LAION 5B 是一个用于研究目的的大规模图文数据集。由 58.5 亿个 CLIP 过滤的图像-文本对组成，其中包含 23.2 亿的英语，22.6 亿的样本来自 100 多种其他语言，及 12.7 亿的未知样本。此外，发布方提供了几个最近邻索引、用于探索和子集创建的改进 Web 界面以及水印和 NSFW 的检测分数。

WuDaoMM
发布时间：2022-03-22
发布机构：北京智源人工智能研究院
项目主页: https://github.com/BAAI-WuDao/WuDaoMM
数据集说明：WuDaoMM 属于北京智源人工智能研究院 WuDaoCorpora 开源数据集的一部分。WuDaoMM 是图文多模态预训练数据，全量数据集包含 6.5 亿图文对，为 Wenlan、Cogview 等大规模中文多模态预训练模型提供了数据支撑，数据集包含强相关数据 5 千万对和弱相关数据 6 亿对。目前智源开放了基础版 WuDaoMM-base，该数据集是由强相关数据按照类别均衡抽取组成的，包含 19 个大类，分别为:能源、表情、工业、医疗、风景、动物、新闻、花卉、教育、艺术、人物、科学、大海、树木、汽车、社交、科技、运动等，单类别数据约 7 万~40 万左右。

Wukong
发布时间：2022-01-30
发布机构：华为诺亚方舟实验室
项目主页: https://wukong-dataset.github.io/wukong-dataset/
数据集说明：Wukong 是由华为诺亚方舟实验室创建的大规模中文跨模态预训练数据集，包含 1 亿对中文图像-文本对，用于推动视觉-语言预训练研究。数据集通过高频中文词汇列表收集，覆盖广泛视觉和文本概念，适用于多种下游任务，如零样本图像分类和图像-文本检索，旨在解决中文环境下跨模态学习的挑战。