arXiv:2305.05880v2 [cs.MM] 6 Aug 2023

ChinaOpen：开放世界多模态学习数据集

陈澳竹 中国人民大学信息学院，中国北京

王子元 中国人民大学信息学院，中国北京

董成博 信息学院，中国人民大学，中国北京

田凯斌，中国人民大学信息学院，中国北京

赵瑞祥 信息学院，中国人民大学，中国北京

梁勋 信息学院，中国人民大学，北京，中国

康展辉 腾讯 深圳，中国

李熹荣* DEKE 教育部重点实验室，中国人民大学，北京，中国

摘要

本文介绍了*ChinaOpen*，这是一个来自哔哩哔哩（一家颇受欢迎的中国视频分享网站）的数据集，旨在支持开放世界多模态学习。尽管当前最先进的多模态学习网络在自动视频标注和跨模态视频检索方面表现出色，但其训练和评估主要基于带有英文文本的 YouTube 视频进行。它们在中文数据上的有效性仍有待验证。为支持新情境下的多模态学习，我们构建了*ChinaOpen-50k*，这是一个包含 5 万个哔哩哔哩视频的网络标注训练集，每个视频都附有用户生成标题和标签。我们通过基于文本和基于内容的数据清理，预先移除了低质量视频。为进行多方面评估，我们构建了*ChinaOpen-1k*，这是一个包含 1,000 个视频的人工标注测试集。每个测试视频都配有经人工核查的用户标题和人工撰写的视频描述。此外，我们还为每个视频人工添加标签，以描述视觉内容中呈现的物体、动作和场景。原始用户标签也经过了人工核查。此外，将所有中文文本译成英语后，ChinaOpen-1k 也适用于评估使用英文数据训练的模型。除了 ChinaOpen，我们还提出了生成式视频到文本转换器（GVT），用于中文视频描述。我们在新数据集上对当前最先进的单任务和多任务模型进行了广泛评估，并得出了一系列新发现和见解。

*通讯作者：李熹荣 (xirong@ruc.edu.cn)

允许免费制作本作品的全部或部分内容的数字副本或纸质副本，供个人或课堂使用，前提是副本不得用于营利或商业利益，也不得为此而制作或分发，并且副本首页须保留本声明及完整引文。必须尊重本作品中由作者以外的其他人拥有版权的部分。注明出处后允许摘录。若要以其他方式复制、重新发布、张贴到服务器或分发至列表，须事先获得具体许可和/或支付费用。请联系 permissions@acm.org 申请许可。*MM '23，2023年10月29日至11月3日*，加拿大安大略省渥太华© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 979-8-4007-0108-5/23/10...$15.00 https://doi.org/10.1145/3581783.3612156

CCS 概念

• 信息系统 → 多媒体数据库; **• 计算方法 → 基于内容的视觉内容索引与检索**。

关键词

中文视频数据集，多模态学习，多任务评估

ACM 参考格式：

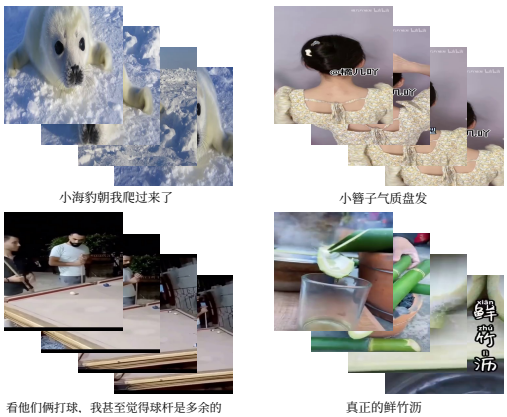
陈澳竹、王子元、董成博、田凯斌、赵瑞祥、梁勋、康展辉和李熹荣。2023年。ChinaOpen：一个用于开放世界多模态学习的数据集。载于第*31*届 *ACM* 国际多媒体会议论文集（*MM '23*），*2023*年*10*月*29*日至*11*月*3*日，加拿大安大略省渥太华。ACM，纽约，美国纽约州，9页。https://doi.org/10. 1145/3581783.3612156

1 引言

在线视频通常时长从几十秒到几分钟不等，在各种社交媒体平台和视频分享网站的信息传播中发挥着日益重要的作用。这些视频由业余创作者或专业团队创作或重新剪辑，涵盖范围广泛的主题，从娱乐和幽默到教育和信息类内容，涉及的领域十分广泛。对于基于视觉内容的索引与检索，一个内容不受控制的开放世界已经形成。

尽管最先进的多模态学习网络在自动视频标注[1, 35] 和跨模态视频检索 [10, 26, 27], 方面表现出色，但其训练和评估主要基于带有英文文本的 YouTube 视频进行。此类模型在多大程度上能够泛化到中文数据，仍是一个悬而未决的问题。我们旨在通过*ChinaOpen*填补这一空白，这是一个面向开放世界多模态学习与评估的新视频数据集。

如图 1 所示，ChinaOpen 由两个子集组成：*ChinaOpen-50k* 和 *ChinaOpen-1k*。前者包含 5 万个经过精心筛选的视频及其用户生成标题（和其他元数据）。后者包含 1000 个视频，这些视频的用户生成标题/标签经过人工审核，字幕由人工撰写，且标签由人工标注，用于描述视频内容中出现的视觉对象/动作/场景。



(a) 用于训练的 ChinaOpen-50k



(b) 用于评估的 ChinaOpen-1k

图1：ChinaOpen 中的视觉示例和标注：(a) ChinaOpen-50k，一个精选的、由用户命名的视频集，用于多模态学习；(b) ChinaOpen-1k，一个经人工标注的视频集，用于评估多模态模型。

由于训练大型视频-语言模型的需求迅速增长，已经开发出多个网络标注视频数据集，例如 HowTo100M [28] 和 WebVid [2]。HowTo100M 由从 YouTube 收集的叙事性视频片段组成，并以转录文本作为其标注。WebVid 的数据来源于 Shutterstock，其描述经过专业编辑。这两个数据集均来自使用英语文本的英语网站。相比之下，ChinaOpen 的数据来源于哔哩哔哩1，这是中国领先的视频分享网站，拥有约 9000 万日活跃用户。因此，ChinaOpen 具有独特性。

现有的人工标注视频数据集要么基于标签，最初用于视频分类 [6, 12]，，要么基于字幕 [37, 38]，没有人工标签。ChinaOpen-1k 独具特色，每个视频都带有人工标注的中文标签，明确描述视频内容中呈现的物体、动作和场景。与现有数据集相比，ChinaOpen-1k 包含许多与物体、动作和场景相关的新标签，见表1。此外，每段视频还附有一条用户生成标题（经人工核验）和一段人工撰写的基于内容的字幕。我们认为，给定视频的标题应由上传者提供，因为这位用户了解视频的背景，因此更适合撰写引人注目的标题。

表1：ChinaOpen-1k 与公开数据集的标签集对比。新颖标签是 ChinaOpen-1k 独有的标签。

数据集	模态	常见标签	新颖标签
物体			
VidOR [32]	vid	43	537
MSCOCO [24]	img	45	535
物体365 [33]	img	131	449
OpenImages [16]	img	188	392
HVU [6]	vid	226	354
LVIS [8]	img	229	351
VisualGenome [14]	img	231	349
COCO-CN [22]	img	362	218
动作			
UCF-101 [34]	vid	9	457
Sports-1M [11]	vid	16	450
Kinetics-600 [3]	vid	130	336
Kinetics-400 [12]	vid	139	327
Kinetics-700 [4]	vid	139	327
HVU	vid	146	320
场景			
HVU	vid	44	86
Places365 [41]	img	78	52

总而言之，本文的贡献如下：

- **数据**。我们构建了*ChinaOpen*，其中 ChinaOpen-50k 用于多模态学习，ChinaOpen-1k 用于多模态模型评估。据我们所知，ChinaOpen 是同类中的首个数据集[2]。
- **模型**。我们提出了用于中文视频描述的生成式视频到文本转换器（GVT）。GVT 在 GIT [35] 的基础上，采用简单的视觉令牌缩减层，有效增加了输入视频帧的数量，从而取得更好的性能。
- **评估**。我们评估了多达 15 个最先进模型（11 个英语模型和 4 个中文模型），详见表2。我们的评估涵盖了最新进展，例如 用于开放集视频标注的 ERNIE-ViL2 [31], CN-CLIP [39] 和 Taiyi [40] 、用于文本到视频检索的 X-CLIP [27]，以及用于视频描述的 Flamingo [1], GIT、mPLUG [17] 和 BLIP-2 [18]。

本文其余部分安排如下。我们将在第2节讨论相关工作。第3节详细介绍 ChinaOpen，随后在第4节介绍 GVT，并在第5节进行评估。第6节给出结论。

2 相关工作

网络标注视频数据对于开发大型多模态学习模型至关重要。同时，人工标注视频数据也是正确评估所开发模型的必要条件。我们简要回顾这两个方面的进展，并相应说明 ChinaOpen 如何以独特方式为该领域作出贡献。

网络标注视频数据集的进展。随着视频语言相关任务对训练大型多模态模型的需求日益增长，人们已积极尝试从网络上收集弱标注视频 [2, 28]。HowTo100M [28], 由数亿个带有旁白的视频片段组成，

[1]https://www.bilibili.com/

[2]ChinaOpen 可访问：https://ruc-aimc-lab.github.io/ChinaOpen/

已被用于学习文本-视频嵌入。尽管规模极其庞大，旁白视频片段的转录文本噪声过多，因此必须采用专门设计的抗噪学习算法。较新的数据集 WebVid 包含 1000 万个短视频及其文本描述，这些内容来自素材视频网站 [2]。具体而言， [2] 的作者使用了 WebVid 中一个包含 250 万个样本的子集来训练深度视频-文本匹配模型。我们发现，这些描述往往措辞精炼，能够简洁地描述视频内容，因此与常见的用户生成标题存在显著差异。还需注意，HowTo100M 和 WebVid 均来自英语网站。在此背景下，我们的 ChinaOpen-50k 由一家热门中国视频分享网站上的用户标题视频组成，独具特色。此外，我们开发了自动化数据清洗方法，以剔除标注质量低或缺乏有意义视觉元素的视频。因此，尽管 ChinaOpen-50k 的规模相对较小，它在我们的实验中已展现出良好的潜力。

人工标注视频数据集的进展。现有的人工标注视频数据集大多侧重于特定任务，包括人体动作识别（HMDB [15] 和 UCF-101 [34]）、体育相关活动（Sports-1M [11]）或范围更广的动作（Kinetics-400 [12]）。整体视频理解（HVU）数据集将标签范围从动作扩展到场景、物体、属性和概念 [6]。然而，上述数据集的标注形式都是标签，因此不适用于视频-语言任务，例如匹配视频与自然语言文本的文本到视频检索，以及生成视频内容文本描述的视频描述。与此同时，目前的 MSVD [5], 、MSR-VTT [38], 和 VaTeX [37] 等视频-文本数据集均没有人工标签。我们的 ChinaOpen-1k 数据集独具特色，每段视频都带有人工标注的中文标签，明确描述视频内容中呈现的物体、动作和场景。此外，每段视频还附有一个（经人工核查的）用户生成标题，以及一段人工撰写的基于内容的字幕。

表2：在 ChinaOpen-1k 上评估的最先进模型。

模型	#参数	视觉	语言	标注	检索	图像描述
ResNet-P365 [41]	2400 万张图像		英文	✓	✗	✗
SwinB-K400 [25]	88M	vid	EN	✓	✗	✗
CLIP4Clip [26]	164M	视频	EN	✗	✓	✗
X-CLIP [27]	164M	视频	EN	✗	✓	✗
CLIP-32/B [30]	151M	img	EN	✓	✓	✗
CN-CLIP [39]	1.88亿	图像	CN	✓	✓	✗
ERNIE-ViL2 [31]	2.04亿	图像	CN	✓	✓	✗
Taiyi [40]	2.54亿	图像	CN	✓	✓	✗
CLIP-L/14@336px [30]	4.27亿	图像	EN	✓	✓	✗
OFA-Chinese [36]	1.6亿	图像	CN	✗	✗	✓
GIT [35]	1.61亿	图像/视频	EN	✗	✗	✓
BLIP [19]	247M	图像	EN	✗	✗	✓
mPLUG [17]	574M	img	EN	✗	✗	✓
Flamingo [1]	1,138M	图像/视频	EN	✗	✗	✓
BLIP-2 [18]	3,745M	img	EN	✗	✗	✓
GVT（本文）	146M	vid	CN	✗	✗	✓

3 THE *ChinaOpen* 数据集

如图 2 所示，ChinaOpen 数据集分三个阶段构建：首先从哔哩哔哩收集原始数据，然后通过自动化数据清洗获得用于多模态学习的 ChinaOpen-50k，最后进行人工视频标注，生成用于多任务评估的 ChinaOpen-1k。下面将逐一介绍各个阶段。

3.1 原始数据收集

为了获得具有代表性且多样化的哔哩哔哩短视频子集，我们随机生成了大量视频 ID 作为候选项。本研究收集了近 10 万个于 2010 年 5 月至 2021 年 9 月上传的视频。除了 MP4 视频文件外，我们还下载了各种元数据，包括标题、标签、描述、评论和弹幕（即飞屏评论，如有）。下载视频的时长从 2 秒到 608 秒不等，平均值为 28.4 秒，中位数为 25 秒。每个视频标题中的中文字数从 1 到 80 不等，平均值为 16.4，中位数为 14。哔哩哔哩平台将用户上传的视频归入不同频道，这些频道是由哔哩哔哩定义的关键词，用于从较高层面描述视频。ChinaOpen 中的视频采集自近 100 个频道，其中排名前 10 的频道为日常、搞笑、名人、社会、电影和电视剪辑、综合、美容护肤、健身、猫和穿搭。

3.2 自动化数据清洗

由于原始数据种类繁多，标注质量参差不齐，因此有必要进行自动化数据清洗，移除标注质量低或缺乏面向广大受众的有意义视觉内容的视频。通过对原始数据进行人工筛选，我们发现，不需要的视频大致可分为以下四类：空标题、仅含人脸、文字密集和无内容。为此，我们设计了一种多模态方法，按步骤识别这四类视频，并相应地将其移除。

3.2.1 空标题视频移除。如果视频标题中不含动词-名词短语（VNP），因而无法提供多少有关视频内容的信息，我们就将其视为空标题。为确定是否存在 VNP，我们使用 HanLP [9], 这一开源中文自然语言处理工具箱[3] 对给定标题进行解析，以获得标题的句法表示。随后，通过句法模式匹配，在该表示中查找动宾结构、含修饰语-中心语结构的名词短语，以及主谓宾结构。

请注意，由于中国视频分享文化，许多标题主要由俚语构成，例如“名场面”、“打卡挑战”和“跟着UP主创作吧”。这些词语使用频率极高，几乎无法提供有关视频内容的信息，因此构成了我们的停用词表基础。我们进一步扩充词表，纳入“觉得”“知道”和“建议”等心理动词，以及标点符号、表情符号和日文／韩文字符等非中文字符。我们先用停用词表过滤给定视频的标题，再进行句法模式匹配。如果未找到 VNP，则移除相应视频。

[3]https://github.com/hankcs/HanLP

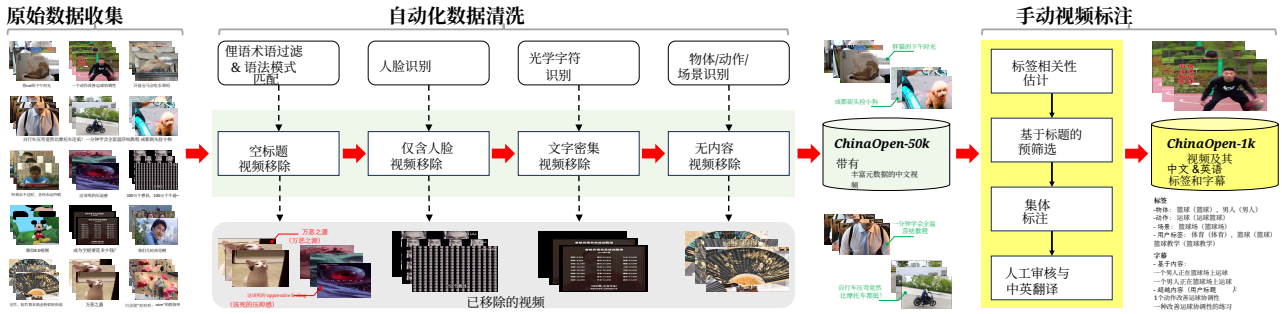


图2：所提出的*ChinaOpen*数据集构建概念图。给定从哔哩哔哩随机收集的10万条中文视频（第3.1节），我们进行自动化数据清洗，移除标注质量低或缺乏有意义视觉内容的视频（第3.2节）。由此得到*ChinaOpen-50k*，这是一个包含5万条视频、用于多模态学习的网络标注数据集。为进行多方面评估，我们构建了一个真实标注测试集*ChinaOpen-1k*（第3.3节）。每个测试视频都配有经人工核查的用户标题、人工撰写的基于内容的字幕、经人工核查的用户标签，以及若干描述视频内容中物体／动作／场景的标签。将所有中文文本翻译成英语后，ChinaOpen-1k也适用于直接评估在英语数据上训练的多模态模型。

3.2.2 纯人脸视频移除。纯人脸视频中主要呈现人脸，几乎没有背景或其他视觉元素。具体而言，我们观察到两种主要模式：即说话人头部和人脸拼图。说话人头部视频只展示一个人的头部（和肩膀），而人脸拼图视频则包含展示许多（小）人脸拼贴画的帧。因此，我们采用逐帧人脸检测。我们使用开源深度人脸分析库 InsightFace4，并采用其默认检测模型（SCRFD-10GF）。对于给定视频，我们均匀采样其中的帧。逐帧执行人脸检测。如果检测到的人脸区域超过图像面积的 50 %，则将该帧归类为说话人头部帧。因此，如果给定视频中超过 75% 的帧属于说话人头部帧，则将其视为说话人头部视频。为判断视频是否为人脸拼图视频，我们统计每帧检测到的人脸数量的最大值。如果该数量超过指定阈值（根据经验设为 8），则将视频标记为人脸拼图视频。由此，我们移除纯人脸视频。

3.2.3 文字密集型视频移除。帧中含有大量文字的视频通常缺少常见的视觉元素。为过滤掉这类文字密集型视频，我们对帧进行光学字符识别（OCR）。具体而言，我们采用 PaddleOCR5，这是一款公开且领先的 OCT 工具，可识别图像中的多语言文本。如果 OCR 检测到的字符数超过 50，则将该帧视为文字密集帧。同样，如果视频中超过 75% 的帧属于文字密集帧，我们就将该视频视为文字密集型视频。

3.2.4 无内容视频移除。 如果给定视频缺乏可识别的物体、动作或场景，我们就将其视为无内容视频。为此，我们采用现有的视觉识别模型来判断给定视频中是否存在任何物体、动作或场景。对于场景识别，我们采用在 Places-365 图像数据集上训练的 ResNet-152 网络 [41]（ResNet-P365）。由于 ResNet-P365 的输入应为图像，我们简单地提取给定视频的中间帧。对于动作识别，我们使用在 Kinetics-400 视频数据集上训练的视频 Swin Transformer（以 Swin-B 为其骨干网络） [25] [12], 我们将其称为 SwinB-K400。请注意，Places365 和 Kinetics-400 所定义的 365 个场景类别和 400 个动作类别不足以覆盖中文视频的丰富内容。ResNet-P365 往往会将狗或猫的视频错误地归类为“兽医诊所”，而 SwinB-K400 则倾向于将烹饪视频误标为“烹饪鸡肉”。尽管存在这些偏差，它们的预测仍有助于筛除无内容视频。

对于物体识别，我们通过合并 MSCOCO [24], VisualGenome [14], Objects365 [33],LVIS [8]和 OpenImages [16] 中的物体类别，整理出包含 3,841 个物体标签的集合。为了预测这些物体与视频的相关性，即 视频中是否出现这些物体，我们使用预训练的 CLIP 模型（ViT-B/32） [30] 在中间帧上进行零样本标注。我们先将英文标签手动翻译成中文，再使用 CN-CLIP [39], （CLIP 的中文版本）为视频标注中文物体标签。

在物体方面，如果视频被预测为至少包含一个置信度较高的标签（阈值设为第 75 百分位数），或两个置信度中等的标签（阈值设为第 50 百分位数），我们就认为该视频并非无内容。动作和场景标签也采用类似的后处理方式。如果未输出任何标签，则将视频视为无内容。

通过上述视频清理流程，我们获得了一个包含 5 万个视频的清理后数据集，称为 ChinaOpen-50k。视频时长介于 3 秒至 608 秒之间，平均时长为 29.8 秒，中位数为 27 秒。每个视频的文件大小介于 81.9 KB 至 20.9 MB 之间，平均为 1.4 MB，中位数为 1.2 MB。ChinaOpen-50k 的视频总时长为 431.2 小时，总大小为 69.1 GB。视频标题的平均长度为 19.2 个字符，长于原始数据中的标题（平均 16.4 个字符）。更重要的是，与原始数据相比，ChinaOpen-50k 为多模态学习和细粒度人工标注提供了更好的起点。

除了用户生成标题外，我们还通过自动生成视频描述来丰富 ChinaOpen-50k 的标注。我们采用了一个在 MSR-VTT [38], VaTeX [37], TGIF [23] 和 Action-GIF [29]联合数据集上训练的现有模型[7], 。由于生成的视频描述为英语，我们使用机器翻译[6] 将其转换为中文。

[4]https://insightface.ai/ [5] https://github.com/PaddlePaddle/PaddleOCR

[6]https://fanyi-api.baidu.com/

3.3 人工视频标注

如前所述，我们旨在构建一个经过真实值标注的中文视频数据集，以支持多任务评估。任务包括通用视频内容识别（物体、动作和场景）、辅助用户标注/图像描述，以及通过自然语言文本进行视频检索。由于众所周知，人工标注成本高昂且规模受到很大限制，因此有必要预先筛选视频，以使人工标注过程获得充分的投入回报。

3.3.1 视频预筛选。众所周知，用户标签具有主观性和个性化特征 [20]。为了从 ChinaOpen-50k 中找出可能带有内容相关标签的视频，我们采用经典的邻居投票算法 [20]。对于每个视频，我们根据视频级 CLIP 特征之间的余弦相似度，从数据集中检索其 200 个邻居。如果某个与查询视频关联的用户标签也出现在邻居视频的用户标签列表中，则认为该标签在视觉上相关。接下来，我们从至少带有一个内容相关标签的视频中随机抽取 10k 个视频，按如下方式进行基于标题的预筛选。

与机器生成的视频描述旨在客观描述画面中可见内容不同，用户生成标题往往更引人注目，能够为读者提供对视频的超越内容解读。为了在相关性和吸引力之间取得适当平衡，我们倾向于选择相关标题，使普通用户能够轻松地将标题与视频内容联系起来。遵循这一标准，由三位经验丰富的标注员组成的评审小组阅读了抽样的 10k 个视频的标题，并据此选出 3k 个视频进行集体标注。为减少标注工作量，事先排除时长超过 60 秒的视频。

3.3.2 集体标注。 我们的标注团队由16名成员组成，他们均为实验室的工作人员和学生。每位标注人员都接受了培训，按照由粗到细的方式对指定视频进行标注。首先，标注人员需要再次检查用户标题是否确实与视频内容相关。如果答案是否定的，则跳过该视频。其次，标注人员撰写视频描述，忠实地概述视频内容的主旨。接下来，标注人员使用中文标签描述视频中呈现的物体、动作和场景。最后，标注人员检查用户提供的标签是否与视频内容相关。为确保标注的丰富性，如果某个视频在特定方面（物体、动作、场景或用户标签）没有任何标签，则该视频将被舍弃。

3.3.3 人工审核。 集体标注阶段结束后，评审委员会会对标注进行二次核查，目的有二：修正个别标注人员偶尔出现的标签问题，并借助机器翻译将中文字幕和标签译成英文。我们共对 1,092 个视频进行了人工标注，得到 1,092 个用户生成标题、1,092 条基于内容的字幕、7,910 个中文标签和 7,856 个英文标签[7]。中文／英文标签的不同项分别有 2,100 个和 2,030 个。与现有的（中文）视频描述数据集（例如只有基于内容的字幕的 VaTeX-CN）相比，用户标题的加入使我们能够在一种全新的超越内容赛道中评估模型。尽管该数据集面向中文模型，但英文标注也使我们能够评估英文模型。我们将这个新的测试集称为 ChinaOpen-1k。

测试集总共包含 9.9 小时、1.6 GB 的视频。每段视频的播放时长介于 5 秒至 60 秒之间，平均值为 32.5 秒，中位数为 30 秒。文件大小介于 0.2MB 至 3.4MB 之间，平均值为 1.5MB，中位数为 1.4 MB。中文物体/动作/场景/用户标签词汇表的规模分别为 580 / 466 / 130 / 924。表3详细展示了标注统计信息。与现有的基于标签的数据集相比，ChinaOpen-1k 包含数百个独特标签，详见表1。ChinaOpen-50k 与 ChinaOpen-1k 的视频之间不存在重叠。

表3：ChinaOpen-1K 的标注统计。

标注	最小值	最大值	Mean	Median
每个视频的中文标签数量：				
物体	1	9	2.45	2
动作	1	6	1.37	1
场景	1	3	1.10	1
已验证的用户标签	1	9	2.33	2
每条中文字幕的字符数：				
用户生成标题	6	79	18.88	17
基于内容的字幕	5	38	14.23	14

4 中国ChinaOpen上的多模态学习

作为在 ChinaOpen-50k 上进行多模态学习的展示，本节介绍如何在这个通过网络标注的数据集上训练基于 Transformer 的中文视频描述模型。

我们以生成式图像到文本转换器（GIT）[35], 为基础。它是多个视觉到语言生成任务中的先进模型，涵盖图像描述、视频描述和视觉问答（VQA）。以图像描述为例。在训练阶段，给定一张尺寸为 256×256 的输入图像和一条包含 m 个词的参考描述，GIT 使用视觉Transformer（ViT）对输入图像进行编码，生成一组 $14\times14{=}196$ 个视觉词元以及一个特殊的 [CLS] 词元，每个词元都带有一个 768 维的嵌入向量。同时，描述也会被分词，并表示为一组由 $(m+2)$ 个嵌入构成的序列，其中额外的两个词元分别表示句子的开头和结尾，即 [BOS] 和 [EOS]。随后，视觉词元和文本令牌会被拼接起来，并输入语言 Transformer 以生成文本。请注意，为防止解码过程中发生信息泄漏，语言 Transformer 使用因果自注意力，使每个文本令牌只能“看到”它之前的令牌，也就是说，所有视觉词元以及当前词元之前的文本令牌。对于视频描述，GIT 只需拼接输入视频帧的视觉词元。若输入 k 帧，这种策略会产生大量 $k\times197$ 个视觉词元。因此，GIT 必须将 k 设置为较小的数值（即 6），以确保计算可行。然而，较小的 k 意味着对视频帧进行稀疏采样，必然会造成大量视觉信息损失。

请注意，对于特定帧而言，其 [CLS] 词元已在 ViT 中由该帧的其他词元进行更新。因此， [CLS] 词元在很大程度上代表了该帧。还请注意，对于给定视频，其中间帧通常比其他帧更具代表性。因此，有选择地结合中间帧中的所有词元以及其他帧的 [CLS] 词元似乎

[7]“喵星人”“cat”“猫”等中文标签会被翻译为“cat”，因此英文标签的数量相对较少。

是合理的。我们通过一个简单的视觉词元缩减（VTR）层实现了这一想法，见图 3。借助 VTR，输入语言 Transformer 的视觉词元数量从 $k \times 197$ 大幅减少至 $k + 196$。这一小幅调整[8] 使我们能够有效地将输入视频帧的数量从 6 扩展到 16。借助 VTR 层，GIT 针对视频描述进行了定制。我们将改进后的模型命名为生成式视频到文本转换器（GVT）。

GVT 的视觉与语言 Transformer 由预训练的 GIT_BASE[9] 初始化。为了适应中文，语言专用层（例如 文本分词器、词元嵌入和最后一个全连接层）在适用时会从头开始重新训练。

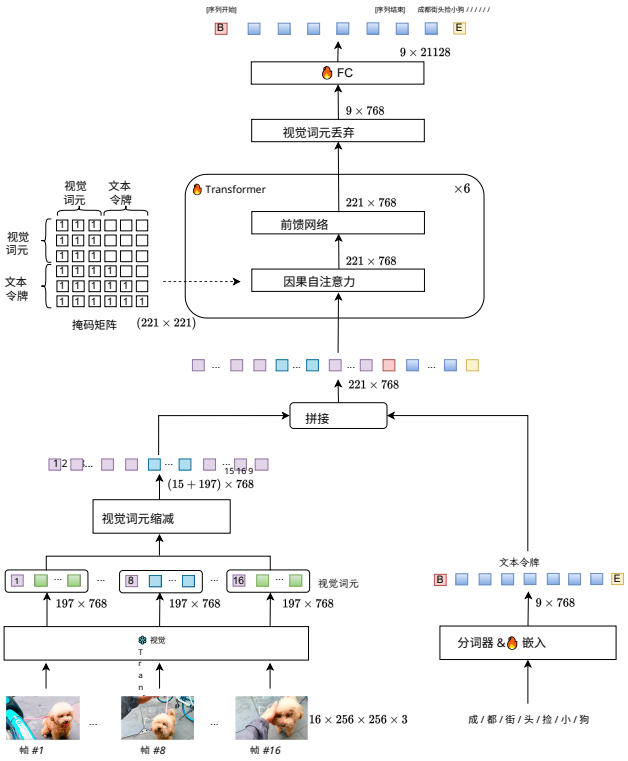


图 3：用于视频描述的生成式视频到文本转换器（GVT）。GVT 在 GIT 的基础上，增加了一个简单的视觉标记缩减层，从而有效地将输入视频帧数从 GIT 的 6 帧提升至 16 帧。

5 评估

5.1 通用设置

考虑到模型的 PyTorch 训练/推理代码是否可用，以及我们的计算能力（8×NVIDIA RTX 3090 GPU），我们收集了以下最先进模型：

• **视频标签**（2）：*ResNet-P365*（在 Places365 上训练的 ResNet-152）[41] 和*SwinB-K400* [25]（以 Swin-B 为骨干网络、在 Kinetics-400 视频动作数据集上训练的视频 Swin Transformer）。

• **文本到视频检索**（2）：*CLIP4Clip* [26]（将 CLIP 迁移到视频领域，并使用序列 Transformer 进行时序建模）和*X-CLIP* [27]（在 CLIP4Clip 的基础上改进，实现细粒度的跨模态匹配）。

• **标签与检索** (5)：*CLIP-B/32* [30]（以 ViT-B/32 作为视觉编码器、GPT-2 作为文本编码器）、*CLIP-L/14@336px* [30]（以 ViT-L/14-336px 作为视觉编码器、GPT-2 作为文本编码器）、*Taiyi* [40]（以 ViT-B/32 作为视觉编码器、Chinese-Roberta-wwm-base 作为文本编码器）、*CN-CLIP* [39]（以 ViT-B/16 作为视觉编码器、Chinese-Roberta-wwm-base 作为文本编码器），以及*ERNIE-ViL2* [31]（以 ViT-B/16 作为视觉编码器、ERNIE-3.0-base 作为文本编码器）。

• **视频描述**(6)：*OFA-Chinese* [36] (以 ResNet-101 为视觉编码器、以 Transformer 为文本编码器)、*GIT* [35] (以 ViT-B/16 为视觉编码器)、*BLIP* [19] (以 ViT-B/16 为视觉编码器、以 BERT-base 为文本编码器)、*BLIP-2* [18] (以 ViT-L/14 为视觉编码器、以查询 Transformer 为文本编码器)、*mPLUG* [17] (以 ViT-B/16 为视觉编码器、以 6 层 Transformer 为文本编码器)和 *Flamingo* [1] (以 ViT-L/14 为视觉编码器、以 OPT-350m 为文本编码器)。

默认情况下，除非另有说明，否则每个模型均使用其对应语言的真实标注进行评估。

5.2 任务I：开放集视频标注

5.2.1 任务设置。 在开放集视频标注中，模型需要使用一个临时词汇表为给定视频添加标签，而该词汇表并非模型专门针对其进行调优的。回顾一下，ChinaOpen-1k 包含四个维度的标签，即即 物体、动作、场景和用户标签。按维度评估视频标注，可以揭示模型在识别物体、动作和场景，以及辅助用户标注方面的能力。

通过基于提示词的标签嵌入，大型多模态模型（如 CLIP、CN-CLIP、Taiyi 和 ERNIE-ViL2）天然适用于开放集设定。然而，对于使用固定词汇表训练的 ResNet-P365 和 SwinB-K400 而言，这一设定颇具挑战性。例如，与 Places365 相比，ChinaOpen-1k 包含52个新场景标签，详见表1。为解决标签不匹配问题，我们对 ResNet-P365 的预测结果进行如下转换。对于每个测试视频，我们使用 ResNet-P365 预测其中间帧的前5个标签。对于 ChineseOpen-1k 场景词汇表中的每个（英语）标签，通过将预测标签的分数乘以其与该标签之间的 BERT相似度分数并求和，计算该标签与给定视频的相关性分数。我们以类似的方式处理 SwinB-K400。

在这三个预训练中文模型（CN-CLIP、ERNIE-ViL2 和 Taiyi）中，只有 CN-CLIP 发布了 PyTorch 训练代码。因此，我们选择使用 ChinaOpen-50k 对该模型进行微调，学习率设为 1e-5。由于 CN-CLIP 是一种图像模型，我们简单地提取每个视频的中间帧。我们还尝试使用相同的微调策略处理 VaTeX-CN（带有中文字幕的 VaTeX）。为便于参考，我们使用 ${model} [${数据集}]这一记法来表示在 ${dataset} 上训练的 ${model}。

5.2.2 性能指标。我们按照评估多标签分类的常用方法，计算每张测试图像的平均精度（AP）[21]。整体性能以平均AP衡量。

5.2.3 结果。 表4报告了不同模型的性能。在四种英文模型中，CLIP系列明显优于ResNet-P365和SwinB-K400，展现了大型多模态模型在开放集标注场景中的优势。

[8]我们还尝试添加首帧和末帧中的所有词元。因此，视觉词元的数量从 k+196 增加到 k+588。增加更多帧只能略微提升性能，却会显著增加计算开销。 [9] https://publicgit.blob.core.windows.net/data/output/GIT_BASE/snapshot/model.pt

同时，CLIP-L/14@336px明显优于CLIP-B/32。请注意，这两种模型的主要区别在于，前者配备了大得多的视觉Transformer。中文模型中也观察到了类似结果：使用ViT-B/16的CN-CLIP优于使用较小ViT-B/32的Taiyi。在ChinaOpen-50k上微调的CN-CLIP优于原始版本。

表4：开放集视频标注的性能。指标：平均精度。

模型	物体	动作	场景	用户标签	Mean
English models:					
SwinB-K400	-	7.7	-	-	-
ResNet-P365	-	-	8.7	-	-
CLIP-B/32	34.6	32.8	32.4	27.9	31.9
CLIP-L/14@336px	**44.0**	**42.7**	**36.7**	**35.7**	**39.8**
中文模型:					
Taiyi	38.8	38.2	45.7	35.7	39.6
ERNIE-ViL2	40.8	40.0	47.4	35.3	40.9
CN-CLIP [*VaTeX-CN*]	40.7	41.7	**48.3**	34.4	41.3
CN-CLIP	39.2	**43.0**	47.1	36.4	41.4
CN-CLIP [*ChinaOpen-50k*]	**42.6**	42.2	45.8	**37.3**	**42.0**

从这四个维度来看，结果表明，表现最佳的英语模型（CLIP-L/14@336px）对物体的识别效果最好，而三个中文模型的表现相对一致，都最擅长识别场景。对于英语和中文模型而言，它们在用户标签维度上的表现都相对最差，这表明辅助用户标注更具挑战性。

5.3 任务二：文本到视频检索

5.3.1 任务设置。 文本到视频检索旨在根据视频与给定文本查询之间的跨模态相似度对视频进行排序。请注意，每个测试视频都关联一个用户生成标题和一条人工撰写的描述。这使我们能够设置两个评测赛道：基于内容的赛道使用人工描述作为测试查询，超越内容赛道则使用用户标题作为测试查询。

对于大型图像文本模型（CLIP模型、CN-CLIP、Taiyi 和 ERNIE-ViL2），其视频级特征通过对相应的帧级特征进行均值池化获得。至于最初为文本到视频检索开发的 X-CLIP 和 CLIP4Clip，我们使用 ViT-B/32 作为它们的视觉编码器，并在四个常用的英语视频-文本数据集上对其进行训练：即 MSVD [5], MSR-VTT [38], VaTeX [37] 和 ActivityNet-Caption [13], 。

5.3.2 评估指标。我们报告常用的第N位召回率（R@N，N=1、5、10）及其总和，记为 SumR。

5.3.3 结果。各个模型的文本到视频检索性能汇总于表5。在英文模型中，CLIP-L/14@336px 再次表现最佳，凸显了使用更大型视觉Transformer的重要性。尽管如此，通过在大量视频文本对上训练特定任务网络，可以有效缩小 CLIP-L/14@336px 与 CLIP-B/32 之间的性能差距（基于内容的赛道中，SumR 分别为 233.2 和 210.3）；参见 SumR 为 229.7 的 X-CLIP [*Va- TeX*] 。与中文模型相比，英文模型表现更优，主要是因为采用了大得多的视觉Transformer，例如 CLIP-L/14@366px。在模型规模（参数数量）相当的情况下，参见表*2*。CN-CLIP（含 1.88亿个参数）和 CLIP B/32（含 151M 个参数）的表现大致相当（SumR 分别为 213.2 和 210.3）。在 ChinaOpen-50k 上微调后，CN-CLIP 的性能有所提升。

比较这两个赛道，我们发现，X-CLIP [*VaTeX*] 在超越内容赛道相较于 CLIP-B/32 的性能提升远低于其在基于内容的赛道上的提升（SumR 中为 7.4 而非 19.4）。这一结果表明，人工撰写的视频描述与用户生成标题之间存在明显差异。对于所有模型而言，它们在超越内容赛道上的表现始终低于基于内容的赛道。由此我们得出结论：使用用户标题进行查询更加困难。

比较这三个中文模型，尽管 CN-CLIP 在视频标注任务中表现最佳，但 ERNIE-ViL2 此时的表现优于 CN-CLIP 和 Taiyi。鉴于 ERNIE-ViL2 和 CN-CLIP 使用相同的视觉编码器（ViT-B/16），但使用不同的文本编码器（ERNIE-3.0 而非 Chinese-Roberta-wwm），这一结果表明 ERNIE-3.0 能提供更好的文本查询表示。

表 5：文本到视频检索性能。各语言的模型按总体性能排序。

模型	基于内容的赛道				超越内容赛道			
	R@1	R@5	R@10	SumR	R@1	R@5	R@10	SumR
英文模型:								
CLIP-B/32	49.5	75.9	84.9	210.3	32.3	59.0	68.3	159.6
X-CLIP [*MSR-VTT*]	50.7	79.9	86.9	217.5	32.5	58.8	69.9	161.2
CLIP4CLIP [*MSVD*]	51.3	77.3	86.6	215.2	34.6	60.2	70.6	165.4
CLIP4CLIP [*ActivityNet*]	53.5	79.9	87.7	221.1	34.1	58.3	68.8	161.2
CLIP4CLIP [*MSR-VTT*]	52.3	79.9	88.5	220.7	33.9	59.9	68.9	162.7
X-CLIP [*MSVD*]	53.7	79.3	86.6	219.6	34.1	61.8	71.7	167.6
CLIP4CLIP [*VaTeX*]	56.0	82.5	88.9	227.4	34.0	60.3	70.6	164.9
X-CLIP [*ActivityNet*]	55.0	81.3	88.6	224.9	35.6	62.5	71.2	169.3
X-CLIP [*VaTeX*]	56.9	**83.3**	89.5	229.7	35.0	61.0	71.0	167.0
CLIP-L/14@336px	**59.5**	83.2	**90.5**	**233.2**	**45.2**	**69.4**	**78.8**	**193.4**
中文模型:								
Taiyi	48.4	77.5	85.8	211.7	41.1	68.2	79.8	189.2
CN-CLIP	48.2	77.6	87.5	213.2	43.8	72.2	80.5	196.4
ERNIE-ViL2	53.6	81.8	89.4	224.7	46.1	72.5	80.8	199.4
CN-CLIP [*VaTeX-CN*]	59.3	**87.2**	92.2	238.7	42.0	70.1	78.3	190.4
CN-CLIP [*ChinaOpen-50k*]	**62.5**	86.4	**92.7**	**241.5**	**52.2**	**79.2**	**88.0**	**219.4**

5.4 任务III：视频描述

5.4.1 任务设置。 视频描述旨在生成自然语言句子来描述视频内容。与文本到视频检索任务类似，我们也设置了两个赛道。基于内容的赛道使用人工撰写的描述作为真实标注，而超越内容赛道则使用用户生成标题。

对于四种图像描述模型，即 mPLUG、BLIP、BLIP-2 和 OFA-Chinese，我们使用中间帧作为视觉输入。对于 Flamingo，我们遵循原论文 [1], ，对每个视频均匀采样 8 帧作为视觉输入。

由于 GVT 源自 GIT，因此后者是前者的直接基线。我们分别在 ChinaOpen-50k 和 VaTeX-CN 上对这两个模型进行微调。为了验证我们提出的数据清洗流程的必要性，我们从原始数据中随机选取了一个规模相同的数据集，并将其称为 Bilibili-50k。此外，我们尝试通过添加一个基于 token 的二元分类头来预测给定的视频-文本对是否相关，从而实现 GVT 的多任务版本。该分类头以 [EOS] token 作为输入，该 token 能够获取所有先前 token 的信息。通过基于提示的标签嵌入，可以实现开放集视频标注。鉴于视频描述生成。

损失（ℓ_g）和视频文本匹配损失（ℓ_m），组合形成的损失函数为 $w \times \ell_g + (1-w) \times \ell_m$，权重为 $w \in [0, 1]$。

5.4.2 评估指标。我们采用三种常见指标，即 BLEU-4、METEOR 和 CIDEr，并以其平均值进行整体比较。中文视频描述的相关计算以词为单位进行，使用结巴[10] 进行中文分词。

5.4.3 结果。如表6所示，BLIP-2 能够以冻结权重的方式利用预训练的大语言模型，表现明显优于其他英文模型。在基于内容的赛道中，GIT [*ChinaOpen-50k*] 优于 GIT [*VaTeX-CN*]（42.2 对比 23.1），而 GVT [*ChinaOpen-50k*] 优于 GVT [*VaTeX-CN*]（44.4 对比 31.3），这表明 ChinaOpen-50k 有助于打造更出色的中文视频描述模型。超越内容赛道也得出了相同结论。鉴于 ChinaOpen-50k 是自动构建的，这些结果令人鼓舞。

表6：视频描述性能。

模型	基于内容的赛道				超越内容赛道			
	BLEU-4	METEOR	CIDEr	平均值	BLEU-4	METEOR	CIDEr	平均值
英文模型：								
Flamingo	8.7	7.9	27.3	14.6	3.2	3.5	8.8	5.2
mPLUG	10.9	12.8	33.7	19.1	**3.9**	**4.6**	10.0	6.2
GIT	9.7	9.2	43.0	20.6	2.2	3.1	9.2	4.8
BLIP	17.9	13.3	62.0	31.1	2.8	3.5	10.0	5.4
BLIP-2	**19.5**	**15.2**	**79.3**	**38.0**	3.6	4.2	**14.3**	**7.4**
中文模型：								
OFA-Chinese	3.8	6.3	13.6	7.9	1.1	3.0	4.3	2.8
GIT[*VaTeX-CN*]	10.7	18.4	40.1	23.1	**1.7**	4.3	4.9	3.6
GVT[*VaTeX-CN*]	**18.5**	18.4	56.9	31.3	1.6	4.5	4.8	3.6
GIT[*Bilibili-50k*]	14.9	18.4	67.8	33.7	0.9	3.9	6.5	3.8
GIT[*ChinaOpen-50k*]	17.0	**19.1**	90.1	42.1	1.2	4.5	9.2	5.0
GVT[*ChinaOpen-50k*]	17.7	**19.1**	96.3	44.4	1.5	4.6	9.1	5.1

在基于内容的赛道中，GIT [*ChinaOpen-50k*] 的表现优于在 Bilibili-50k 上训练的对应模型，平均精度为 42.1，而 33.7；在超越内容赛道中则为 5.0，而3.8。该结果证明了数据清洗的必要性。多任务 GVT 的性能如表 7 所示。对于标注/检索，我们可以清楚地看到多任务 GVT 与最先进水平之间存在性能差距。开发统一模型并非易事，还需要进一步研究。

对于英语和中文模型而言，它们在超越内容赛道上的表现得分都远低于内容赛道上的对应模型。显然，最先进水平的视频描述模型生成的内容与真实用户希望为自己的视频拟定的标题之间存在很大差距。定性结果见图 4。



图 4：领先模型的定性结果，即用于标注的 CLIP-L/14@336px 和 CN-CLIP，以及用于图像描述的 BLIP-2、GIT 和 GVT。中间帧以红色边框标出。

6 总结与结论

我们构建了 ChinaOpen，这是一个面向开放世界多模态学习的新型视频数据集。该数据集由 ChinaOpen-50k（一个网络标注的视频集，用于训练）以及 ChinaOpen-1k（一个人工标注的双语视频集，用于测试）组成。我们对 15 个最先进模型和提出的 GVT 进行了评估，得出以下结论。对于开放集视频标注，表现最佳的英文／中文模型分别是 CLIP-L/14@336px 和 CN-CLIP。预测用户标签比识别物体、动作和场景更困难。对于文本到视频检索，CLIP-L/14@336px 再次成为表现最佳的英文模型，而 EARNIE-ViL2 是表现最佳的中文模型。对于视频描述，BLIP-2 生成的英文字幕质量最佳，而在 ChinaOpen-50k 上训练的 GVT 生成的中文字幕质量最佳。在检索和视频描述任务中，超越内容赛道似乎都比基于内容的赛道更具挑战性。ChinaOpen 为未来研究展现了新的机遇。

致谢。本研究得到国家自然科学基金（NSFC，62172420）、腾讯营销解决方案犀牛鸟重点研究计划以及中国人民大学公共计算云的支持。第一作者感谢阿姆斯特丹大学（UvA）的 H. Doughty 和 C. Snoek 就该主题进行的有益讨论。

[10]https://github.com/fxsjy/jieba

表 7：多任务 GVT 在视频标签、检索和图像描述方面的性能。

	开放集视频标注					文本到视频检索								视频描述							
						基于内容的赛道				超越内容赛道				基于内容的赛道				超越内容赛道			
w	物体	动作	场景	用户标签	Mean	R@1	R@5	R@10	SumR	R@1	R@5	R@10	SumR	BLEU-4	METEOR	CIDEr	Mean	BLEU-4	METEOR	CIDEr	Mean
0.8	16.0	17.1	17.1	11.7	15.5	35.3	65.6	79.9	180.8	15.8	41.4	54.9	112.1	**16.6**	**18.8**	**87.5**	**41.0**	**1.5**	**4.5**	**9.1**	**5.0**
0.5	16.8	18.5	**20.8**	12.8	17.2	39.0	70.3	81.9	191.2	19.4	47.3	61.3	128.0	16.2	18.5	83.9	39.5	1.4	4.4	8.8	4.9
0.2	**20.0**	**21.6**	19.3	**16.1**	**19.3**	**40.2**	**71.3**	**83.3**	**194.8**	**22.3**	**51.6**	**64.0**	**137.9**	15.6	18.4	77.6	37.2	1.2	4.2	8.4	4.6

参考文献

[1] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katherine Millican、Malcolm Reynolds 等。2022。Flamingo：一种用于小样本学习的视觉语言模型。载于*NeurIPS*。[2] Max Bain、Arsha Nagrani、Gül Varol 和 Andrew Zisserman。2021。Frozen in Time：一种用于端到端检索的联合视频图像编码器。载于*ICCV*。[3] Joao Carreira、Eric Noland、Andras Banki-Horvath、Chloe Hillier 和 Andrew Zisserman。2018。关于 Kinetics-600 的简短说明。arXiv:1808.01340[4] Joao Carreira、Eric Noland、Chloe Hillier 和 Andrew Zisserman。2022。关于Kinetics-700人类动作数据集的简短说明。arXiv:1907.06987[5] David Chen 和 William Dolan。2011。收集用于释义评估的高度平行数据。载于*CVPR*。[6] Ali Diba、Mohsen Fayyaz、Vivek Sharma、Manohar Paluri、Jürgen Gall、Rainer Stiefelhagen 和 Luc Van Gool。2020。大规模整体视频理解。载于*ECCV*。[7] 董成博、Xinru Chen、陈澳竹、Fan Hu、Zihan Wang 和李熹荣。2021。通过语义增强学习实现用于视频描述的多层次视觉表示。载于*ACMMM*。[8] Agrim Gupta、Piotr Dollar 和 Ross Girshick。2019。LVIS：一个用于大词汇实例分割的数据集。载于*CVPR*。[9] Han He 和 Jinho D. Choi。2021。干细胞假说：Transformer编码器多任务学习背后的困境。载于*EMNLP*。[10] Fan Hu、陈澳竹、Ziyue Wang、Fangming Zhou、Jianfeng Dong 和李熹荣。2022。轻量级注意力特征融合：文本到视频检索的新基线。载于*ECCV*。[11] Andrej Karpathy、George Toderici、Sanketh Shetty、Thomas Leung、Rahul Suk- thankar 和 Li Fei-Fei。2014年。使用卷积神经网络进行大规模视频分类。载于*CVPR*。[12] Will Kay、Joao Carreira、Karen Simonyan、Brian Zhang、Chloe Hillier、Sudheendra Vijayanarasimhan、Fabio Viola、Tim Green、Trevor Back、Paul Natsev 等。2017。Kinetics人类动作视频数据集。arXiv:1705.06950[13] Ranjay Krishna、Kenji Hata、Frederic Ren、Li Fei-Fei 和 Juan Carlos Niebles。2017。视频中的事件密集描述。载于*ICCV*。[14] Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalantidis、Li-Jia Li、David A. Shamma、Michael S. Bernstein 和 Li Fei-Fei。2017。Visual Genome：使用众包密集图像标注连接语言与视觉。*IJCV* 123，1（2017），32–73。[15] Hildegard Kuehne、Hueihan Jhuang、Estíbaliz Garrote、Tomaso Poggio 和 Thomas Serre。2011。HMDB：一个用于人体运动识别的大型视频数据库。载于*ICCV*。[16] Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Uijlings、Ivan Krasin、Jordi Pont-Tuset、Shahab Kamali、Stefan Popov、Matteo Malloci、Alexander Kolesnikov、Tom Duerig 和 Vittorio Ferrari。2020。Open Images数据集V4：大规模统一图像分类、目标检测和视觉关系检测。*IJCV*（2020）。[17] Chenliang Li、Haiyang Xu、Junfeng Tian、Wei Wang、Ming Yan、Bin Bi、Jiabo Ye、He Chen、Guohai Xu、Zheng Cao、Ji Zhang、Songfang Huang、Fei Huang、Jingren Zhou 和 Luo Si。2022。mPLUG：通过跨模态跳跃连接实现有效且高效的视觉语言学习。载于*EMNLP*。[18] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。2023。BLIP-2：借助冻结图像编码器和大语言模型引导语言图像预训练。arXiv:2301.12597[19] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。2022。BLIP：为统一的视觉语言理解与生成引导语言图像预训练。载于*ICML*。[20] 李熹荣、Cees G. M. Snoek 和 Marcel Worring。2009。通过邻居投票学习社交标签相关性。*TMM* 11，7（2009），1310–1322。[21] 李熹荣、Tiberio Uricchio、Lamberto Ballan、Marco Bertini、Cees G. M. Snoek 和 Alberto Del Bimbo。2016。社交化语义鸿沟：图像标签分配、优化与检索的比较综述。*CSUR* 49，1（2016），14:1–14:39。

[22] 李熹荣、Chaoxi Xu、Xiaoxu Wang、Weiyu Lan、Zhengxiong Jia、Gang Yang 和 Jieping Xu。2019。用于跨语言图像标注、图像描述和检索的 COCO-CN。*TMM* 21，9（2019），2347–2360。[23] Yuncheng Li、Yale Song、Liangliang Cao、Joel R. Tetreault、Larry Goldberg、Ale- jandro Jaimes 和 Jiebo Luo。2015。TGIF：一个用于动画 GIF 描述的新数据集和基准。载于*CVPR*。[24] Tsung-Yi Lin、Michael Maire、Serge Belongie、Lubomir Bourdev、Ross Girshick、James Hays、Pietro Perona、Deva Ramanan、C. Lawrence Zitnick 和 Piotr Dollár。2014。Microsoft COCO：上下文中的常见物体。载于*ECCV*。[25] Ze Liu、Jia Ning、Yue Cao、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Han Hu。2022。视频 Swin Transformer。载于*CVPR*。[26] Huaishao Luo、Lei Ji、Ming Zhong、Yang Chen、Wen Lei、Nan Duan 和 Tianrui Li。2022。CLIP4Clip：关于 CLIP 端到端视频片段检索和图像描述的实证研究。*Neurocomputing* 508（2022），293–304。[27] Yiwei Ma、Guohai Xu、Xiaoshuai Sun、Ming Yan、Ji Zhang 和 Rongrong Ji。2022。X-CLIP：用于视频文本检索的端到端多粒度对比学习。载于*ACMMM*。[28] Antoine Miech、Dimitri Zhukov、Jean-Baptiste Alayrac、Makarand Tapaswi、Ivan Laptev 和 Josef Sivic。2019。HowTo100M：通过观看数亿个带旁白的视频片段学习文本-视频嵌入。载于*ICCV*。[29] Yingwei Pan、Yehao Li、Jianjie Luo、Jun Xu、Ting Yao 和 Tao Mei。2022。GIF 自动字幕：用于视觉-语言预训练的大规模视频-句子数据集。载于*ACMMM*。[30] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。利用自然语言监督学习可迁移的视觉模型。载于*ICML*。[31] Bin Shan、Weichong Yin、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。2022。ERNIE-ViL 2.0：用于图像-文本预训练的多视图对比学习。arXiv:2209.15270[32] Xindi Shang、Donglin Di、Junbin Xiao、Yu Cao、Xun Yang 和 Tat-Seng Chua。2019。在用户生成视频中标注物体与关系。载于*ICMR*。[33] Shuai Shao、Zeming Li、Tianyuan Zhang、Chao Peng、Gang Yu、Xiangyu Zhang、Jing Li 和 Jian Sun。2019。Objects365：一个大规模、高质量的目标检测数据集。载于*ICCV*。[34] Khurram Soomro、Amir Roshan Zamir 和 Mubarak Shah。2012。UCF101：一个包含 101 个人类动作类别的野外视频数据集。arXiv:1212.0402[35] Jianfeng Wang、Zhengyuan Yang、Xiaowei Hu、Linjie Li、Kevin Lin、Zhe Gan、Zicheng Liu、Ce Liu 和 Lijuan Wang。2022。GIT：用于视觉与语言的生成式图像到文本转换器。arXiv:2205.14100[36] Peng Wang、An Yang、Rui Men、Junyang Lin、Shuai Bai、Zhikang Li、Jianxin Ma、Chang Zhou、Jingren Zhou 和 Hongxia Yang。2022。OFA：通过简单的序列到序列学习框架统一架构、任务和模态。载于*ICML*。[37] Xin Wang、Jiawei Wu、Junkun Chen、Lei Li、Yuan-Fang Wang 和 William Yang Wang。2019。VaTeX：用于视频与语言研究的大规模、高质量多语言数据集。载于*ICCV*。[38] Jun Xu、Tao Mei、Ting Yao 和 Yong Rui。2016。MSR-VTT：用于连接视频与语言的大型视频描述数据集。载于*CVPR*。[39] An Yang、Junshu Pan、Junyang Lin、Rui Men、Yichang Zhang、Jingren Zhou 和 Chang Zhou。2022。Chinese CLIP：中文视觉-语言对比预训练。arXiv:2211.01335[40] Jiaxing Zhang、Ruyi Gan、Junjie Wang、Yuxiang Zhang、Lin Zhang、Ping Yang、Xinyu Gao、Ziwei Wu、Xiaoqun Dong、Junqing He、Jianheng Zhuo、Qi Yang、Yongfeng Huang、Xiayu Li、Yanghan Wu、Junyu Lu、Xinyu Zhu、Weifeng Chen、Ting Han、Kunhao Pan、Rui Wang、Hao Wang、Xiaojun Wu、Zhongshen Zeng 和 Chongpei Chen。2022。封神榜 1.0：构建中文认知智能的基石。arXiv: 2209 02970 。[41] Bolei Zhou、Agata Lapedriza、Aditya Khosla、Aude Oliva 和 Antonio Torralba。2017。Places：一个包含一千万张图像的场景识别图像数据库。*TPAMI* 40，6（2017），1452–1464。