VideoMind: 用于深度认知视频理解的全方位模态视频数据集，包含意图定位

Baoyao Yang
Guangdong University of Technology, China

Wanyun Li
Wechat, Tencent, China

Dixin Chen
Guangdong University of Technology, China

Junxiang Chen *
Wechat, Tencent, China

Wenbin Yao*
Wechat, Tencent, China

Haifeng Lin
Guangdong University of Technology, China

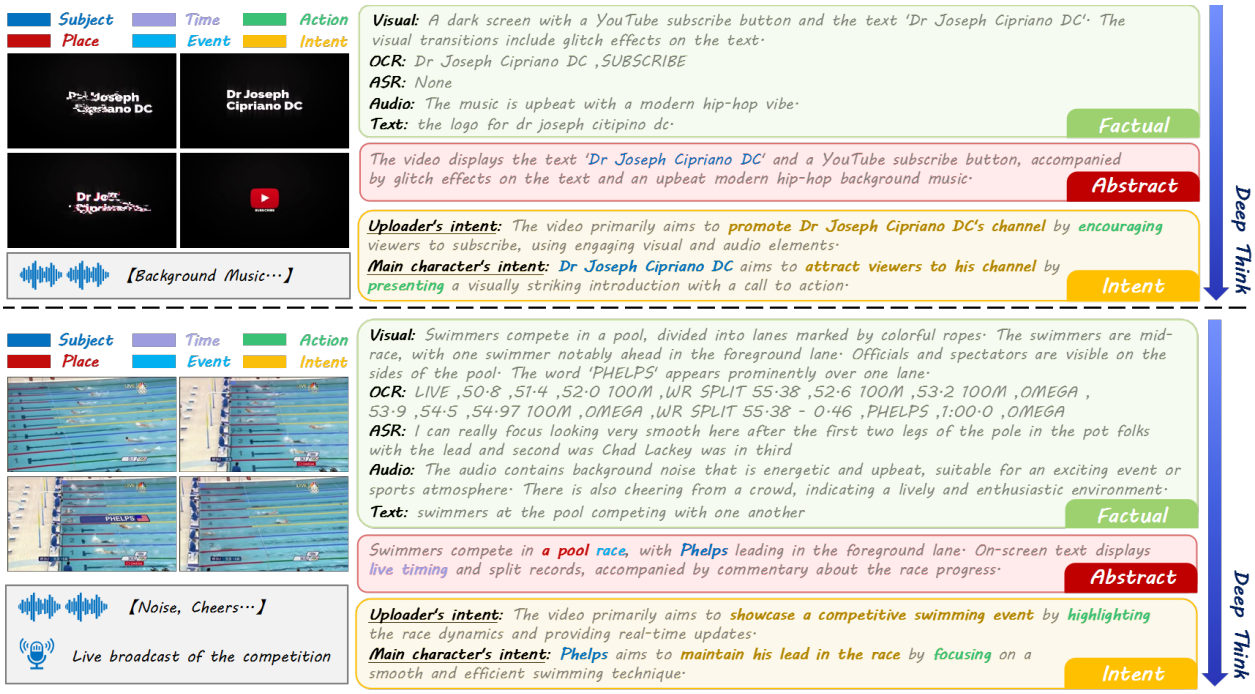


Figure 1: VideoMind 中的视频片段示例（更多示例在 https://opendatalab.com/Dixin/VideoMind 中展示）。

Abstract

本文介绍了 VideoMind，一个以视频为中心的全模态数据集，它能够深入认知视频内容并增强多模态数据的特征表示。VideoMind 数据集包含 103K 个视频样本（其中 3K 仅用于测试），每个样本都伴随着音频以及系统和详细的文本描述。具体来说，每个视频样本及其音频数据在三个层次（事实、抽象和意图）进行描述，从表面到深度逐层推进。总共有超过 2200 万字，包括每个样本平均约 225 个单词。与现有以视频为中心的数据集相比，VideoMind 的显著特征在于提供了直观不可获取、必须通过整合整个视频上下文进行推测的意图表达。引入了链式思维（COT）的文本生成方式，其中通过逐步指导，提示大语言模型（mLLM）生成深度认知表达。在详细描述的基础上，标注了包括主体、地点、时间、事件、动作和意图在内的各种注释，为一系列下游识别任务服务。更重要的是，我们建立了一个黄金标准基准，包括 3000 个经过精心手动验证的样本，用于评估深度认知视频理解。为了更适当地评估模型对视频的深度理解，设计了混合认知检索实验，通过多级检索指标评分。多种标准基础模型（InternVideo、VAST、UMT-L 等）的评估结果已经发布。被认为 VideoMind 是一个强大的基准测试，不仅促进了细粒度的跨模态对齐，还推动了需要深入理解视频的领域，如情感和意图识别。数据可在三个托管平台（GitHub、Huggingface 和 Opendatalab）上公开获取。https://github.com/cdx-cindy/VideoMind 。

*Corresponding author: JX Chen and WB Yao ({ caryjxchen,wenbinyao } @tencent.com)

1 介绍

随着社交媒体的发展，视频已经成为信息传播的主要媒介。精确理解视频内容，尤其是其潜在目的和意图，对于促进智能通信和保障互联网的合法性与安全性至关重要。近年来，数十种多模态基础模型（如 Video-LLaVA [9] 、SoRA 和 Qwen-vl [15] ）已经发布。据报道，它们在生成和对话任务中表现出色。这些

Table 1: 提出的 VideoMind 与当前视频为中心的多模态数据集之间的比较

Dataset	Domain	# Clips	Text Source	Len $_{text}$	Image	Video	Audio	ASR	OCR	Tag	Intent
MSR-VTT [20]	Open	10K	Manual	9.3		✓					
MSVD [2]	Open	1970	Manual	8.7		✓					
LSMDC [12]	movie	118K	Manual	7.0		✓					
ANet Caption [6]	Action	20K	Manual	13.5		✓					
VaTex [16]	Open	41.3K	Manual	15.2		✓					
HT100M [11]	Instructional	136M	ASR	4		✓		✓			
HD-VILA-100M [21]	Open	103M	ASR	32.5		✓		✓			
WebVid-10M [1]	Open	10.7M	Alt-texts	12		✓					
How2 [13]	instructional	80K	Manual	20		✓					
VALOR [10]	Open	1.1M	Manual	16.4		✓	✓				
VAST [3]	Open	27M	Generated	32.4		✓	✓				
InternVid [17]	Open	234M	Generated	17.6		✓		✓		✓	
GME [24]	Open	1.1M	Generated	–	✓						
VideoMind (Ours)	Open	103K	COT-Generated	225	✓	✓	✓	✓	✓	✓	✓

成功很大程度上依赖于使用大规模多模态数据的预训练。尽管许多大型企业已将预训练数据私有化，但仍有不少以视频为中心的数据库被发布。例如，马克斯 · 普朗克信息学研究所与加州大学伯克利分校合作，发布了 LSMDC [12] 数据集，该数据集包含超过 10 万段电影片段，并附有对视频内容的手动描述。Max Brain 随后提出了一个 1000 万条数据集 WebVid [1]，其中视频的文本描述来自于网络。HD-VILA-100M [21] 进一步将数据规模扩展至 1 亿，视频片段通过其 ASR 结果自动注释。不满意于匆忙的文本提取，上海 AI 实验室采用基于 mLLM 的文本生成技术来提高文本表达的质量 [17] 。后续研究如 VAST [3] 和 VALOR [10] 进一步整合音频信息以丰富数据模态的范围并增强数据多样性。这种整合拓宽了多模态大语言模型（mLLM）的接收领域，并扩展了它们的应用范围。然而，尽管为模型训练提供了丰富的信息，现有的数据集仍表现出以下局限性：

1) 过于简洁的文本表达：尽管现有数据集中涉及多模态元素，但视频内容通常只用一个大约 20 个词的简短句子来描述。这种文本描述通常仅对应于视频帧或特定区域的一个子集，导致跨模态信息存在显著差异。

2) 缺乏深入的解读：视频的描述仅停留在纯粹的视觉观察水平，没有提供需要深入思考和推理的潜在信息。这导致模型无法理解视频的潜在意图。

3）严重的任务偏倚：大多数现有的数据集是专为特定任务而设计的，比如描述或视频 Q & A。这类数据可能会引导模型提取有助于文本生成的嵌入，但忽视了普遍的代表性。

上述问题限制了基础模型的泛化，更重要的是，阻碍了对视频内容更深层次的认知理解。对视频内容缺乏深入理解将妨碍用户需求的准确对接，最终导致商业价值的损失。此外，深入的视频理解对于维护社交平台的秩序至关重要，它能够有效识别和抑制低质量内容，同时维护版权完整性。为此，本文提出了一个新数据集，VideoMind，它提供了对视频内容全面（广度和深度）文本解释。如图 1 所示，每个视频样本在三个层次上进行了系统性描述——事实层、抽象层和意图层，从表面到深入解读。这些描述是通过应用于 mLLM 的逐步思维链（COT）提示方法生成的，其总体流程在第 3.1 节详细介绍。在意图层的核心中，我们建立了明确的意图表达规则，并设计了两个角色扮演任务，旨在最小化不明确的表达并确保对视频潜在意图的准确推测。此外，VideoMind 包含全模态数据（图像、视频、音频和文本），以及许多相关信息，如自动语音识别（ASR）、光学字符识别（OCR）和各种语义标签，如表 1 所总结。

总之，本文介绍了首个深度认知全模态视频数据集 VideoMind，该数据集涵盖了视频内容的全面描述和潜在意图的详细解释。通过增强数据嵌入的意图感知能力，VideoMind 使全模态基础模型能够全面探索样本之间的内在关系，从而提高精确匹配的能力，并促进社交网络中的有效质量监控。

2 相关工作

视频文本对是训练强大多模态模型的基础，增强了视频理解并推导出跨模态表示，以支持各种下游任务。多个中心在数据采集方面投入了大量精力，建立了依赖于人工标注的视频中心数据集 [2, 6, 10, 12, 13, 16, 20] 。作为劳动密集型项目，这些数据集通常在数据量上表现出限制，或者受到过于简化的文本表示的困扰。例如，MSVD [20] 仅包含 1,970 个视频文本对，而 LSMDC [6] 中的样本平均用七个字表达。为了降低撰写文字的劳动成本，其他工作利用网页图像中的 Alt-text [1, 5, 14] 和视频片段中的 ASR [11, 21, 23] 作为视频内容的文本描述。这一策略为数据收集提供了极大的便利，因此数据集规模已扩大到超过 1 亿。然而，Alt-text 和 ASR 输出的代表性具有争议，因为有些视频可能根本没有音频讲话，而 Alt-text 的质量无法始终得到保证。

当今，mLLM 的进步已经催化了一个向生成方法的范式转变，以用于数据集的构建。鉴于 mLLM 强大的数据解析能力，上海人工智能实验室公布了一个名为 InternVid 的规模庞大的视频-文本数据集，在这个数据集中，要求 mLLM 为每个视频片段生成一个简单的文本描述。随后，mLLM 的衍生策略扩展到三模态，发布了一个名为 VAST 的大规模视觉-音频-语言数据集。这开启了全模态表示学习的过程。然而，这些数据集中数据集规模的不均衡优先考虑，导致包含了低质量的视频和模糊的字幕。南京大学不只是追求规模化，而是强调增强视频的质量和美学价值。为此，他们提出了 OpenViD-1M，一个包含 40 万条 1080P 分辨率视频的数据集。

尽管已经提出了许多以视频为中心的数据集，但很少有研究考虑文本表达的全面性。不同模态之间的信息内容仍然存在显著差距，这导致了全模态学习的自然障碍。最近的一些工作，如 Video-MME [4] 和 IntentQA [7] ，尝试通过 QA 方式补充文本内容。然而，这些描述往往是片面的，特别关注视频中的某些位置或帧。因此，本工作旨在解决这一差距，为视频提

供更全面和深入的文本注释，从而促进视频理解并增强全模态表示。

3 VideoMind: 一个具有广泛和深入基础的全模态数据集

全面而深入的全模态数据集是弥合多模态语言模型理解隐含信息能力差距的重要基础，如意图、情感和动机。为了对视频进行全面理解，VideoMind 提供了全面的模态（包括关键帧、视频、音频和文本），以及从您所见的事实到深入意图推测的深度思考文本表达。具体来说，每个视频的文本描述包括三层：事实层、抽象层和意图层。事实层从各种模态中对内容进行全面详细的描述，而抽象层则基于事实描述提供视频样本的总结。在利用上述两层信息后，意图层进一步推测并记录视频的目的。此外，视频信息被分解为不重叠的元素：视觉（不考虑 OCR）、音频（背景音）、OCR、ASR 和文本。各种标签，如主题、地点和时间，也被标注，以支持广泛的下游任务，如事件识别、身份识别和场景识别。

总之，VideoMind 包含 103,000 个视频样本，这些样本配有大约 2200 万条文本注释。关于数据量、文本长度、信息标签等的统计比较见表 1 。为了为涉及深入理解视频的任务提供高质量和标准化的评估，从 VideoMind 数据集中精心挑选了 3000 个样本，建立了一个基准验证集。验证集中所有样本的合理性已由三名专业标注员独立验证。

为了确保视频意图的多样性，我们选择在社交网络上发布的视频作为我们的原始资源。根据 InternVid 的方法，VideoMind 数据集中的片段来自公开可获得的 YouTube 视频，并涵盖广泛的领域，包括游戏、新闻、娱乐、体育等。（有关类别统计，请参见第 3.3 节）。涉及敏感信息的样本，例如政治倾向和色情内容，被排除在外。为了确保样本包含足够的意图信息，所有视频片段的时长至少为 5 秒。对于每个视频片段，我们随机选择 12 帧作为图像模态的代表。对应于视频片段的原始音频和文本被同时提取，形成每个样本的四元表达式（视频、图像、音频、文本）。选择的四元组将经过文本生成过程（图 2 ），并且只有通过双重验证的那些才会被保留。选择的四元组将经过文本生成过程（图 2 ），对文本内容进行全面而深入的重写，只有那些成功通过双重验证程序的才会被保留在 VideoMind 数据集中。

3.1 基于 COT 的文本生成

提供视频的广泛和深入的文本描述是 VideoMind 的主要特征。该功能促进了对视频内容的更深层次理解，并增强了与认知和情感相关的各种下游任务。为实现这一目标，引入了一个 mLLM（Qwen2.5-Omni），并通过三个阶段的解析以 COT 方式逐步生成视频内容的多层次表达，如图 2 所示。较深层次的表达始终基于对前一层分析和推测的基础上产生。因此，视频文本变得更丰富，从表面呈现演变到更深的内涵：1）事实层，描述可观察和可听见的元素；2）抽象层，将多模态信息综合成一个连贯的摘要；3）意图层，推测视频创作者和视频中主要主题的动机。我们将在下文中详细阐述生成过程的每一步。步骤 1：多视角描述和高级总结。与以前的视频为中心的数据集不同，VideoMind 希望能在不丢失任何信息的情况下全面反映来自各种来源的信息。因此，要求多模态大语言模型 (mLLM) 在第一阶段分别描述多模态数据。结果，在事实层中记录了五个不重叠的元素。即：1）视觉：无论图像文本如何看到的内容描述；2）音频：无论人类语言如何听到的内容描述；3）OCR：视频的光学字符识别结果；4）ASR：音频的自动语音识别结果；5）文本：视频的原始文本，由人工编写。在事实层中，多模态数据按类别进行详细描述。这有助于弥合跨模态的信息鸿沟，因为视觉和音频数据通常比视频部分帧对应的原始文本提供更多信息。然后，要求 mLLM 对覆盖事实层所有信息的内容进行总结，在抽象层形成简要概述。这一层还标注了核心元素，如主体、地点、时间和事件（见第 3.2 节）。第 2 步: 意图推测。根据事实层和抽象层的描述，我们进一步提示 mLLM 推断每个视频的意图。为了防止可能干扰分析的模糊或混乱表达，我们已建立了明确的意图表达规则。遵循意图始终伴随行动的自然法则，意图推测的表达规则制定为：[主体] 通过 [行动] 旨在 [意图]。这种规定的表达方式有助于后续意图的提取和定位。在这里，我们更进一步促进意图解释：设计了两个角色扮演任务，并分别进行意图推测。具体来说，角色 A 是视频的上传者——mLLM 需要推测上传视频的目的。而角色 B 则是视频中的主角。mLLM 将自己想象为主角，解释所描绘行为背后的动机。通过这种角色扮演模式，思维链的发展方向得以明确，从多个角度呈现视频意图，同时有效降低模型幻想的概率。
第三步：验证：质量控制。验证过程用于评估视频上传者和主要角色的意图推测结果（意图层的描述）的可靠性。每个意图表达必须经过预验证和后验证过程才能在 VideoMind 数据集中合格为有效样本。在预验证阶段，添加一个额外的 mLLM 来执行相同的意图推测任务。这将产生两个固定格式的意图表达。按照指定的格式，文本中代表意图的术语可以轻松提取，通过评估这两个意图表述术语的嵌入相似性来评估意图推测结果。只有当两个模型的意图表达词的语义意义表现出相似性时，意图才被认为是正确的。对于后验证，采用文本到视频生成技术。使用 Wan2.1 为每个意图表达生成一个 10 秒的视频，并由两位专家注释者评估视频内容的合理性，从而间接评估文本写作的质量。

此外，我们精心挑选了 3,000 个样本，涵盖了广泛的类别范围。意图层的质量由专业标注员进行严格评估。随后，我们设立了一个基准，作为深入理解视频的第一个标准化标准。

3.2 标记

VideoMind 进一步提供 6W 元素标签，包括主体（谁）、地点（哪里）、时间（何时）、意图（为什么）、动作（如何）和事件（什么），以扩大其应用范围，支持各种下游任务。具体而言，一个 LLM（Qwen2.5-vl）负责从抽象层中的表达识别并突出显示与地点、时间和事件相关的名词。此外，对于意图的固定表达规则，即 [主体] 旨在通过 [动作][意图]，主体、动作和意图在意图层中自动标注。标记结果的示例在图 1 中呈现，其中 6W 元素通过不同颜色的标签区分开来。

3.3 统计和特征

VideoMind 的关键统计数据和特征在表 1 中总结，并提供了与现有视频为中心的多模态数据集的比较。
数据多样性。为了广泛反映公众的行为意图，VideoMind 包括各种在社交媒体上发布的样本，涵盖 45 个国家/地区和 24 个类别。最常见的语言是英语（EN，39.7 %），韩语（KO，9.0 %），中文（CN，7.5 %），日语（JA，7.5 %）和俄语（RU，3.8 %）。相当一部分视频是无语言的，对应于 14.7 % 的样本没有 OCR 或 ASR。图 3 的上部子图显示，相当比例的视频没有 OCR（33.0 %）或 ASR（41.0 %）输出。对于包含 OCR/ASR 的视频，大多数 OCR/ASR 输出长度少于 100 个字。仅有 5.2 % 和 4.5 % 的视频分别包含长 OCR 和 ASR 输出（>100 字）。流行的视频类别

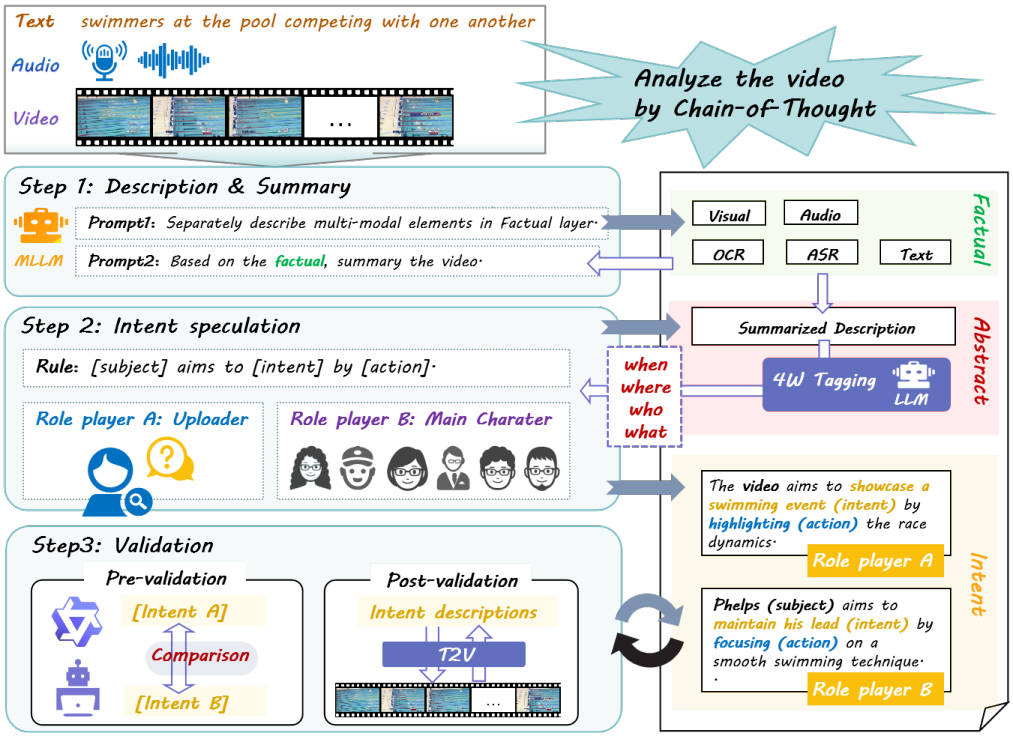


Figure 2: VideoMind 的文本生成流程。

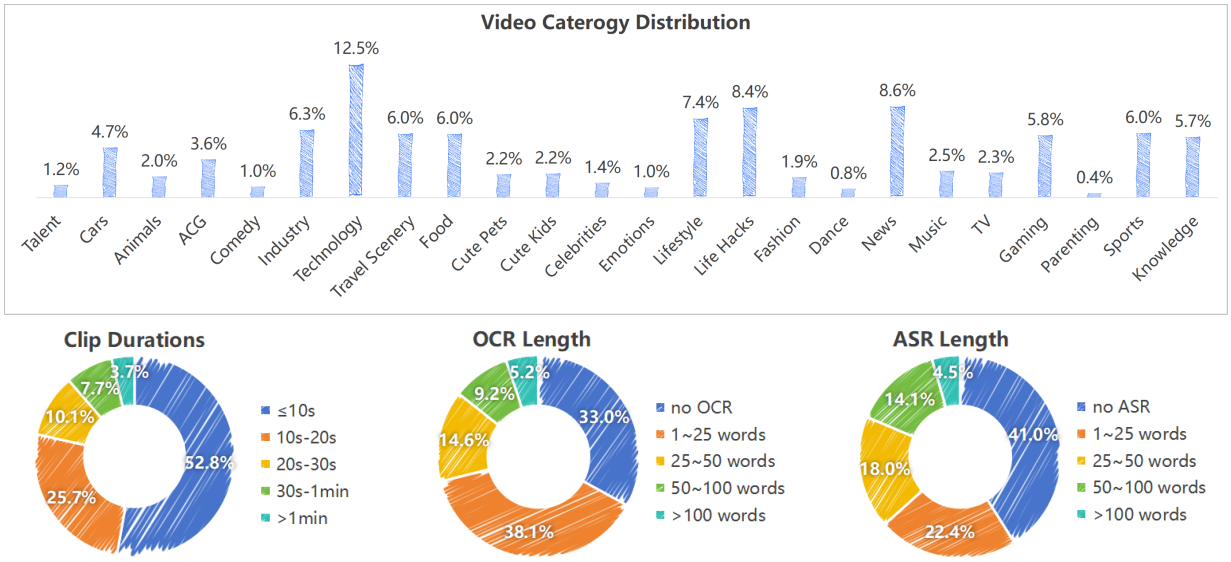
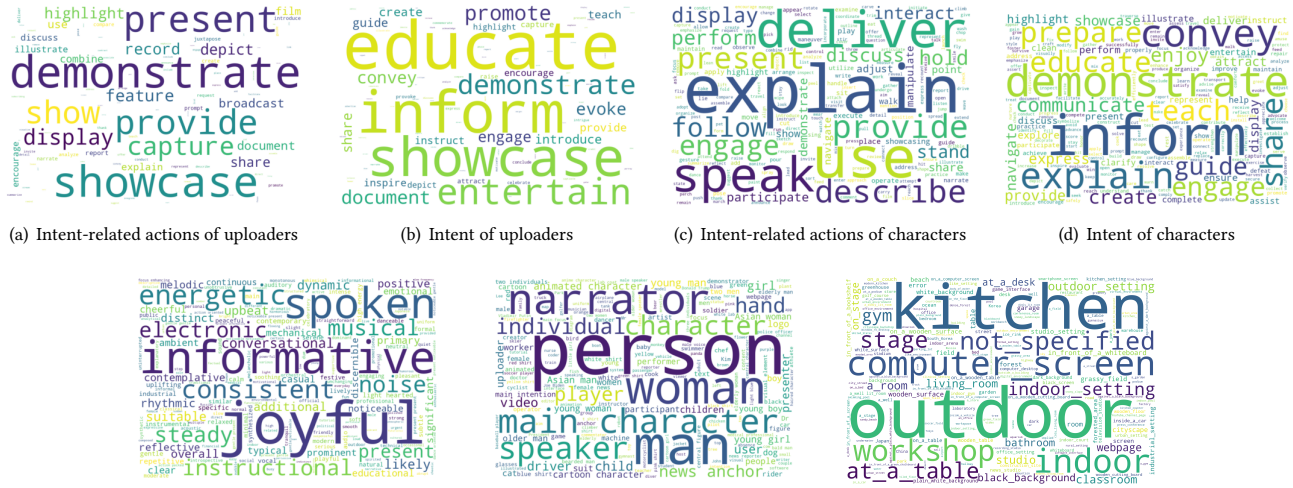


Figure 3: VideoMind 中的视频统计数据。

包括技术（12.5 %）、新闻（8.6 %）、生活妙招（8.4 %）、生活方式（7.4 %）、行业（6.3 %），如图 3 所示。在时长方面，时长不足 5 秒的剪辑被视为无效样本，并从 VideoMind 中排除，因为这些过于简短的视频通常无法体现有意义的意图。平均时长为每段视频 16.23 秒。52.8 % 的视频长度在 5 到 10 秒之间，而 25.7 % 的视频时长为 10 到 20 秒。仅有少量样本（3.7 %）超过 1 分钟。

全面性和丰富性（广泛而深入）。VideoMind 是一个全模态的数据集，每个样本包含来自各种模态的展示，包括关键帧、音频和文本。更重要的是，每种模态的信息都经过系统分类，并随后详细描述。VideoMind 的主要特征是能够提供视频的广泛和深入描述。VideoMind 以 COT 方式为视频生成全面的文本描述，每个样本平均包含 225 个描述词，是现有视频为中心数据集的约 10 倍。具体来说，在事实层、摘要层和意图层中，平



(a) Intent-related actions of uploaders (b) Intent of uploaders (c) Intent-related actions of characters (d) Intent of characters

(e) Audio styple (f) Subject (g) Place

Figure 4: VideoMind 数据集中意图的词云、意图驱动的动作、音频风格、主题和地点。

均分别有 143、38 和 43 个词。VideoMind 也包含目前所知最广泛的模态信息范围，包括图像、视频、音频、ASR、OCR 以及文本。它还提供丰富的关于角色（谁）、地点（哪里）、时间（何时）、事件（做什么）、动作（如何）以及意图（为何）的标签，支持多种下游识别任务，如跨模态检索、意图 VQA、意图识别和情感识别。其中，角色、地点、时间和事件在摘要层标注，而动作和意图在意图层标注。

为了展示 VideoMind 中的意图表达，我们统计分析了视频上传者和视频中的主要角色所进行的意图和与意图相关的行为的词汇分布。从图 4 可以看出，视频上传者倾向于向观众展示、展示、呈现和提供信息，而主要角色通常解释、传达并阐述他们的目的。教育、通知、展示和娱乐是上传者最常见的意图；巧合的是，这些中的大多数也是视频中主要角色经常呈现的意图。意图重叠现象主要归因于某些视频的上传者与这些视频中的主要角色是相同的，特别是在生活方式、技术和才艺类别中。此外，主要角色展示的潜在意图范围比上传者更广，其中沟通、准备、解释和享受在代表主要角色意图的词云中也更加突出（图 4(d) ）。

我们还在图 4(e) 、4(f) 和 4(g) 中展示了与视频中出现的音频风格、主题和地点相关的词云。根据统计结果，视频中最常用的背景音乐风格是快乐的。此外，电子音乐、充满活力的音乐和器乐也是流行的背景音乐选项。其他视频，例如体育或游戏视频，可能充满了人群的噪音，展现出信息性、口述和噪音的音频风格。此外，在这些社交媒体中，女性主题比男性更多；像 “woman” 和 “girl” 这样的词的频率高于像 “man” 和 “boy” 这样的男性代词。就地点而言，室内/室外、厨房、电脑屏幕、车间、工作室和舞台是最常见的视频拍摄场地。

4 混合认知跨模态检索

VideoMind 提供了从浅层到深层的多层次视频文本描述，因此它是一个用于混合认知视频理解的综合数据集。特别地，它是第一个支持视频深层认知的数据集。跨模态检索被选为主要下游评估任务。正如在第 3 节所述，挑选了 3000 个数据样本构成深度认知跨模态检索的基准。这些数据的 COT 生成文本经过专家注释者的仔细验证。这些数据严格限制仅用于测试，决不可以涉及模型训练的任何阶段。在本节中，我们展示了几个标准视频为中心的基础模型的跨模态检索结果，包括 InternVideo [18] 、UMT-L [8] 、CLIP-VIP [22] 、mPLUG-2 [19] 和 VAST [3] 。

对于文本到视频的检索，我们通过利用每一层的文本来评估模型的检索能力，以及通过使用任意层的文本进行视频搜索所达到的整体性能。选择 Rank K（R@K, $K \in \{1, 5, 10\}$ ）和平均排名（MeanR）作为评估指标。正如预期的那样，表 2 显示，当基于事实描述性的文本查询检索视频时，大多数基础模型能够达到令人满意的性能——Rank-1 准确率达到 87 %。然而，随着表达深度的增加，检索能力逐渐降低。具体来说，使用抽象层次的文本作为查询时，Rank-1 检索准确率下降到大约 75 %，而当使用意图层进行检索时，性能进一步下降（∼ 35 %）。由于缺乏对视频内容的深刻理解，目前的基础模型在表示视频的潜在目的方面表现有限。因此，当使用强调潜在目的而非特定对象的意图层文本查询时，检索到准确和相关的视频变得极其困难。观察发现 MeanR 甚至可能超过 200，这表明对视频内容的重大误解。

对于视频到文本的检索，每层的文本表达被分解为三个独立的检索目标。我们分别比较两种检索性能：1）在前 K（$K \in \{1, 5, 10\}$ ）结果中检索到任意真实层的成功率；2）包含所有真实层在前 K （$K \in \{3, 5, 10\}$ ）检索结果中的案例比例。我们进一步记录三个真实文本的最早、最晚和平均被检索到的位置，分别缩写为 TopR、LowestR 和 AvgR。在表格 3 中，“命中任意层” 一栏的结果是相对可以接受的。对于大多数这些基础模型，Rank-1 检索准确率大约为 80 %。这是因为它们至少能够相对容易地搜索到相应的事实描述。因此，最高搜索排名（TopR）是相对可以接受的（∼ 10），这总是由事实层接触到。相比之下，在 “命中所有层” 指标中，所有测试模型表现不佳。它们几乎无法在前 10 个结果中检索到来自所有三个层的文本。因此，LowestR 可能超过 500，在 3000 个测试数据中排名接近六分之一。

表格 2 和 3 中的结果证实了现有基础模型在深入视频理解方面的不足。作为数据驱动的模型，它们的局限性很大程度上可以归因于缺乏可观测的样本，这些样本能够传达深刻的语

Table 2: 在 VideoMind-3K 上进行的混合认知文本到视频检索结果。

Model	factual → V				abstract → V				intent → V				any → V			
	R@1 ↑	R@5 ↑	R@10 ↑	MeanR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MeanR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MeanR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MeanR ↓
InternVideo [18]	82.10	95.50	97.50	2.40	77.60	92.50	95.50	4.10	50.80	72.10	78.30	49.70	70.13	86.69	90.42	18.74
UMT-L [8]	87.56	95.48	96.86	10.78	78.35	89.89	93.16	23.52	37.39	58.83	65.61	233.04	67.76	81.40	85.21	89.11
CLIP-VIP [22]	67.40	86.80	91.30	7.30	64.83	83.90	88.43	10.30	33.77	55.27	62.90	91.00	55.34	75.32	80.88	36.02
mPLUG-2 [19]	83.20	93.20	95.43	37.4	77.57	90.03	93.23	43.84	35.17	57.80	65.37	235.72	65.31	80.34	84.67	105.65
VAST [3]	83.93	95.60	97.20	3.98	74.90	90.37	92.87	9.69	43.83	66.07	72.87	73.05	67.55	84.01	87.64	28.90

Table 3: 在 VideoMind-3K 上的混合认知视频到文本检索结果。

Model	Hit any layer			Hit all layer					
	R@1 ↑	R@5 ↑	R@10 ↑	R@3 ↑	R@5 ↑	R@10 ↑	TopR ↓	LowestR ↓	AvgR ↓
InternVideo [18]	79.63	93.77	96.23	30.67	45.10	58.80	2.75	114.26	43.79
UMT-L [8]	82.70	93.00	95.60	29.73	43.23	54.23	10.15	523.49	201.01
CLIP-VIP [22]	72.13	88.97	93.01	19.23	32.03	44.40	7.21	205.22	79.99
mPLUG-2 [19]	72.97	88.13	91.67	26.26	38.60	51.27	23.51	621.33	254.96
VAST [3]	78.83	92.10	95.17	18.93	29.17	40.10	5.52	324.83	125.89

义内容。VideoMind 是一个包含 100,000 个样本的数据集，附有深刻而广泛的文本描述，非常适合有效地弥补这一空白。

5 声明

这是 VideoMind 的第一个版本。我们的目标是构建一个拥有百万规模的视频深度理解数据库，这将作为促进视频深度理解领域研究和发展的基础资源。我们相信 VideoMind 将加速视频嵌入和各种感知相关识别任务的进展。更多数据将在各个平台上不断更新。请继续关注后续发展。

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1728–1738.

[2] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Aannual meeting of the association for computational linguistics: human language technologies*. 190–200.

[3] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2024. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems* 36 (2024).

[4] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24108–24118.

[5] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17980–17989.

[6] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision*. 706–715.

[7] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11963–11974.

[8] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023. Unmasked teacher: Towards training-efficient video foundation models. In *IEEE/CVF International Conference on Computer Vision*. 19948–19960.

[9] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 5971–5984.

[10] Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2024. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[11] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2630–2640.

[12] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision* 123 (2017), 94–120.

[13] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Lo"ic Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS. http://arxiv.org/abs/1811.00347

[14] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[16] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *The IEEE International Conference on Computer Vision (ICCV)*.

[17] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2023. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. *arXiv preprint arXiv:2307.06942* (2023).

[18] Yi Wang, Kunchang Li, Yizhuo Li, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191* (2022).

[19] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*. 38728–38748.

[20] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5288–5296.

[21] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5036–5045.

[22] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2023. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In *International Conference on Learning Representation*.

[23] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems* 34 (2021), 23634–23651.

[24] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint*

arXiv:2412.16855 (2024).