arXiv:2304.03022v1 [cs.IR] 6 Apr 2023

TagGPT：大语言模型是零样本多模态标注器

陈黎[1] **葛逸潇**[1] **毛嘉勇**[2] **李典**[2] **单颖**[1]

[1]ARC实验室，腾讯PCG[2]基础技术中心，腾讯PCG

摘要

在当代互联网时代，标签对于促进多媒体内容在各种应用中的有效分发至关重要，例如搜索引擎和推荐系统。近年来，大语言模型（LLM）在广泛的任务中展现出了令人印象深刻的能力。在这项工作中，我们提出了TagGPT，这是一种能够以完全零样本方式进行标签提取和多模态标注的全自动系统。我们的核心洞见是，通过精心设计的提示工程，大语言模型能够根据多模态数据的文本线索提取并推理出恰当的标签，例如，光学字符识别、自动语音识别、标题、等等。具体而言，为了针对特定应用自动构建一个能够反映用户意图和兴趣的高质量标签集，TagGPT 通过提示大语言模型，从一系列原始数据中预测大规模候选标签，并依据频率和语义对其进行筛选。对于一个需要进行标签标注以便分发的新实体，TagGPT 提供了两种可选的零样本标注方案，即，一种是结合标签集进行后期语义匹配的生成式方法，另一种是在提示中进行前期匹配的选择式方法。值得注意的是，TagGPT 提供了一种基于模块化框架的系统级解决方案，该框架配备了一个预训练大语言模型（此处使用GPT-3.5）和一个句子嵌入模型（此处使用 SimCSE），并且可以无缝替换为你所需的任何更先进的模型。TagGPT适用于现代社交媒体中的各种数据模态，并展现出对广泛应用的强大泛化能力。我们在公开可用的数据集上评估了TagGPT，即，快手和 Food.com，并证明了与现有主题标签和现成标签器相比，TagGPT 的有效性。[1]

1 引言

标签作为语义内容的简洁描述，已被证明在众多下游任务 [8] 中发挥着极其重要的作用。它们有助于系统快速、准确地理解用户兴趣和搜索意图，从而简化搜索 [9； 6] 以及推荐 [16； 1； 15] 流程。例如，如图 1 所示，在包含多模态内容的社交媒体平台上，用户可以使用自定义标签轻松描述自己发布的内容，吸引兴趣相似的用户。因此，获取优质标签、构建全面的标签系统及相应的标签生成器，一直是提升此类应用服务质量的重要途径。

获取标签的传统来源主要是人工标注和语言模型提取。然而，人工标注得到的标签往往带有主观性且成本高昂 [16]，，而语言模型提取的标签则受限于其理解能力有限

[1]项目页面：https://github.com/TencentARC/TagGPT

技术报告，进行中。

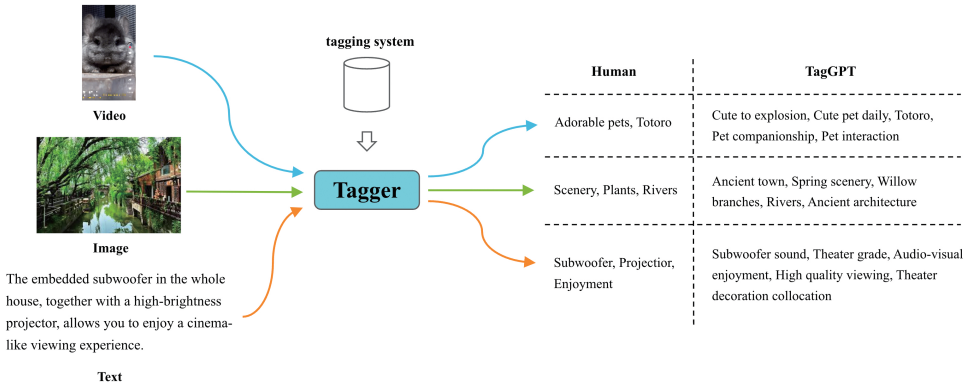


图1：给定来自社交媒体（例如 Twitter、微博、等）的多模态内容，标签生成器旨在生成若干能够恰当地描述该内容并反映用户兴趣的短语。

并且可能缺乏可靠性 [8]。尤其是随着当今社交网络中语义空间的迅速扩展，以及数据模态日益多样化，如何从海量多模态数据中快速获取高质量标签，已成为亟待解决的研究与应用问题。

为构建高质量的面向内容的标签系统及相应的零样本标签器，我们在本文中提出了一种生成式解决方案TagGPT。我们的方法受到大语言模型（大语言模型）令人印象深刻的推理能力及其在零样本设置中的卓越性能的启发 [3; 13;19; 17; 5]。我们首先借助成熟的单模态模型，将多模态数据转换为统一的文本线索。随后，我们将这些线索插入离散的指令提示中，并将其输入大语言模型以生成潜在标签。通过从大规模输入中收集标签，我们可以整理出一组内容标签。接着，我们利用无监督语义相似度融合和频率分布过滤等后处理技术，获得高质量的面向内容的标签系统。最后，我们设计了两种不同的零样本标签器，它们依靠大语言模型的生成能力或上下文推理能力来进行零样本数据标注。

我们使用来自两款热门应用的数据构建了基准测试集，并开展了实验。实验结果表明，TagGPT 能够构建一流的标签系统，并在零样本场景下为输入的多模态数据分配合适的标签。如需获取完整项目代码和更多示例，请参阅 https://github.com/TencentARC/TagGPT。

2 TagGPT

TagGPT 是一个用于处理多模态内容的框架。该框架包含一条可自动构建标签系统的流水线，以及一个可完成零样本标注的标签器。该框架的核心能力在于以低成本借助大语言模型的能力，获得高性能、高质量的标签系统。接下来的章节将介绍组成 TagGPT 的两个主要模块。

2.1 标签系统构建

构建面向内容的标签系统，旨在从大规模数据中提取有价值的语义信息，并形成完整的标签系统。在 TagGPT 中，为了能够从大规模多模态数据中挖掘高质量的标签，如图 2 所示，我们提出了一条由三个子模块组成的流水线。

2.1.1 文本线索转换器

随着移动网络和多媒体技术的发展，社交网络、电子商务平台和媒体信息平台等热门应用越来越多地将多模态数据作为主要的信息分发载体。一个健全的快速发展

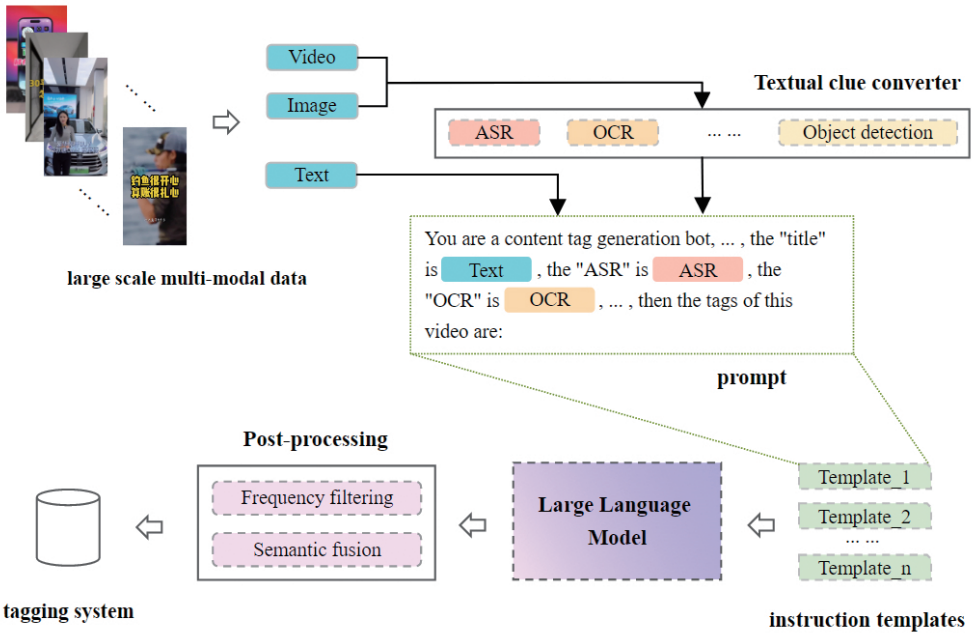


图 2：给定来自特定应用的一系列原始数据，TagGPT 无需额外知识或人工标注，即可完全以零样本方式构建高质量的标注系统。这种范式能够以零人力成本为新应用即时添加标签。

基于此类多模态数据的标签系统可以为搜索和推荐等服务提供基础支持。

近年来，大语言模型在内容理解、推理和生成任务方面表现出色，因此我们尝试将大语言模型应用于标注系统的构建。为最大限度地发挥大语言模型的优势，受 [22 启发的 TagGPT； 18] 首先借助各种单模态基础模型处理多模态数据，并从中提取丰富的文本线索，作为下游模块的输入。例如，在处理视频数据时，TagGPT 会采用强大的自动语音识别（ASR） [21] 和光学字符识别（OCR） [4] 模型，从视频帧和音频中提取文本信息。同样，为满足不同领域标注系统的构建需求，可以利用目标检测 [2] 和动作识别 [10] 等处理技术，从不同角度为 TagGPT 提供文本线索。

2.1.2 基于大语言模型的标签生成

随着大语言模型在训练数据规模、参数规模和训练策略方面不断发展，GPT [3; 13] 和 PaLM [5] 等大语言模型展现出了卓越的理解、推理和生成能力。因此，许多下游任务都从这些模型中获益，性能显著提升。近期，InstructGPT [14] 的问世进一步增强了大语言模型理解各种任务指令的能力，从而提升了它们在不同任务间切换的能力。在 TagGPT 中，我们致力于将大语言模型强大的指令理解能力应用于新任务，并发挥其理解、推理和生成内容的潜力。这将有助于构建一个标签系统。具体而言，我们创建了多个用于获取标签的指令模板，并在相应位置填入与上游提供的单条多模态数据对应的文本线索。随后，将输入文本送入大语言模型，即可获得相应的标签结果。经过大批量生成操作后，便可获得大语言模型生成的大规模标签，从而形成完整且全面的候选标签集。

2.1.3 后处理

当然，以这种方式构建的标签集必然包含大量噪声和冗余语义。因此，在标签系统构建的最后阶段，TagGPT 加入了一个后处理模块，希望通过一系列操作提升标签系统的质量。首先，该模块根据标签频率，截除频率过高或过低的标签；频率较高的标签可能源于流行度偏差，且缺乏样本区分度 [23], ；频率较低的标签则可能仅对应极少数样本，缺乏显著的标签价值 [20]。随后，该模块会检查剩余标签的语义相似度，并合并语义相近的标签，以进一步缩小标签系统的规模。具体而言，系统基于无监督预训练文本编码器对标签系统中的所有标签进行编码，得到相应的表示。然后，该模块计算余弦相似度，以确定是否应合并这些标签。

到目前为止，我们可以根据给定的大规模多模态数据自动构建相应的标签系统。

2.2 零样本多模态标注器

除了自动构建面向内容的标签系统外，TagGPT 还提供相应的零样本标签器。如图 3 所示，根据标注逻辑，我们将分别介绍两种不同的标注器。

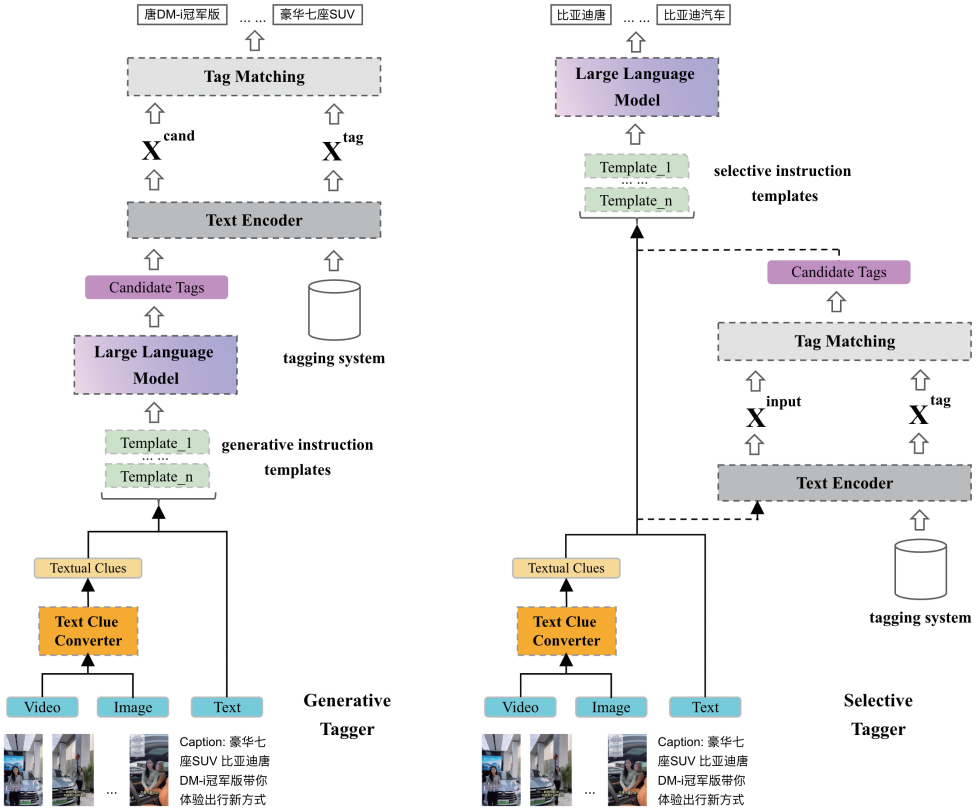


图 3：给定图 2 中建立的标签系统，TagGPT 可以通过两种备选范式对新数据进行零样本标注。

2.2.1 生成式标注器

由于大语言模型具备强大的理解和推理能力，因此是出色的零样本标注器。此前的标签系统正是基于这一能力构建的。因此，如果能将大语言模型推断出的标签快速与现有标签系统进行语义相似度匹配，就能实现零样本标注。基于这一思路，我们提出了一种生成式标注器，它

将输入的多模态样本转换为文本线索，填入候选指令模板的相应位置，并使用大语言模型提供候选标签。候选标签使用预训练文本编码器编码为 $\mathbf{X}^{cand} \in \mathbf{R}^{n \times m}$，而现有标注系统中的标签编码为 $\mathbf{X}^{tag} \in \mathbf{R}^{N \times m}$，$n$，$N$，$m$ 分别表示候选标签的数量、标注系统中的标签数量和表示维度。随后将匹配分数矩阵 $\mathbf{S} \in \mathbf{R}^{N \times n}$ 计算为 $\mathbf{S}_{i,\ j} = \mathbf{X}_i^{tag} \cdot \mathbf{X}_j^{cand\top} / (|\mathbf{X}_i^{tag}| \times |\mathbf{X}_j^{cand}|)$。最后，可通过预设阈值直接选出匹配结果。图 3 左侧展示了这一过程。

2.2.2 选择式标注器

在将大语言模型投入实际应用时，人们发现它具有很强的上下文理解能力，一些方法甚至利用上下文学习策略，引导大语言模型生成更合理的结果 [11]。这启发我们开发了另一种零样本标注范式，如图 3 右侧所示。具体而言，我们首先将给定的多模态样本转换为文本线索。与生成式标注器不同，我们将其与标签系统中的标签一同输入预训练文本编码器，以获得相应的向量表示，即，$\mathbf{X}^{input}$ 和$\mathbf{X}^{tag}$。通过计算文本与标签之间的相似度，我们可以得到一系列相关的候选标签集。最后，将原始文本线索和候选标签集纳入选择式指令模板，并将其输入大语言模型，即可获得给定多模态样本的最终标注结果。

3 实验

本节将在两个真实世界的社交网络数据集上，测试 TagGPT 在零样本设置下构建标签系统并为给定的多模态数据分配标签的能力。

3.1 数据集与指标

数据集	# 项目	说明文字	类别	OCR	ASR
快手	222k	✔	✔	✔	✔
Food.com	5.4k	✔	✔	✗	✔

表 1：数据集统计信息。

我们选择了两个知名网站——快手[2] [12] 和 Food.com[3], ——作为实验的数据来源。详细统计信息见表 1。

在评估基于 TagGPT 的标签系统时，我们将使用来源自身的标签系统（即从快手和 Food.com 提取的标签）作为基线。为全面评估该标签系统的质量，我们将参考以往的研究设计，并考察以下指标:

- **流行度**：当大多数用户持续为一组相关对象分配同一个标签时，这表明该标签具有稳定的价值，也降低了它被视为垃圾标签的可能性。然而，如果标签的适用范围过于宽泛（即过于流行），其区分不同对象的能力就会大幅减弱，从而违背标签的初衷。为评估这一指标，我们计划使用相同的数据，追踪基线和基于 TagGPT 的标签系统中的标签分布。这样便能观察这两种标签系统之间的标签流行度差异。

- **实用性**：这一指标包含两个子含义，即所需付出的努力最少和对多个方面的高覆盖度。这两个指标彼此相关，第一个（即，所需付出的努力最少）关注分配给特定对象的标签数量，而第二个关注特定对象的方面覆盖度。事实上，从理论上讲，这两个指标呈正相关，即分配的标签越少，其能够覆盖的方面也越少。本文拟通过统计不同标注系统下每条给定数据中的标签数量，并评估

[2]https://www.kuaishou.com/ [3]
https://www.food.com/

每条给定数据中标签的语义冗余度来评估这两个指标。其中，标签的语义冗余度取决于每条给定数据对应标签中相似语义所占的比例。

- **一致性**：标签可能因个人习惯和应用场景而有所不同，从而导致对同一概念的不同描述。保留这些内容虽然可以提高标注的召回率并改善用户体验，但过多的标签会增加标签系统的规模并引入噪声。在评估时，我们将把整个标签系统中任意两个标签之间的语义冗余（即，语义相似度）作为计算方法。

3.2 设置

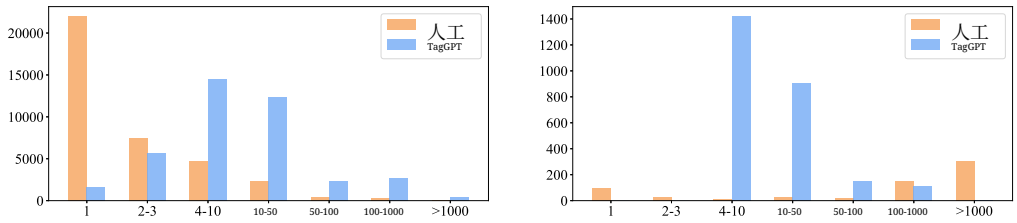
在大语言模型的选择上，我们选择了 GPT 3.5 [14]（API 接口，版本为 *gpt-3.5-turbo*）。无监督文本编码器选用 SimCSE [7]。在判断语义冗余时，我们将无监督语义向量表示的余弦相似度作为衡量指标。当相似度大于或等于 0.8 时，我们认为两个输入对象存在不可忽略的语义冗余。

为了能够为 TagGPT 设定合理的比较基线，我们将用户上传视频时提供的标签作为基线。具体而言，我们提取以“#”分隔的给定数据内容，作为用户标签器的结果，并将其收集为用户标注的标签系统。此外，我们还选择了两种知名的视觉[4] 和文本[5] 在线标签方法进行比较。

3.3 主要结果

在本节中，我们将分别从标签系统和标注器两个方面评估 TagGPT 中不同子模块的性能。

3.3.1 标签系统



(a) 快手　　(b) Food.com

图4：数据集中指标“流行度”的统计结果。横轴表示一个标签被分配给数据的次数，纵轴表示标签数量。

图4展示了 TagGPT 在“流行度”指标上的标签分布统计结果。显然，在这两个数据集中，用户标注的标签分布都并不理想。在快手数据集中，每个标签被用户使用的次数都少于四次，其中高达 60%的标签仅被使用过一次。另一方面，在 Food.com 数据集中，标签分布趋于两极分化：一些标签的使用频率过低，而另一些标签的使用频率过高。如前所述，如果标签流行度过低，系统将不得不面对大量无用标签，使用户难以高效匹配正确的信息。相反，如果标签流行度过高，系统将难以有效区分数据，从而导致用户体验不佳。TagGPT 在两个数据集中都展现出出色的标签覆盖率，其中有大量标签被重复使用超过 10 次，但并未被过度使用。这一特征

[4]Azure 视觉认知服务 (ACSV)：https://portal.vision.cognitive.azure.com/demo/generic-image-tagging. [5]Hanlp：https://www.hanlp.com/demonstrate.html。

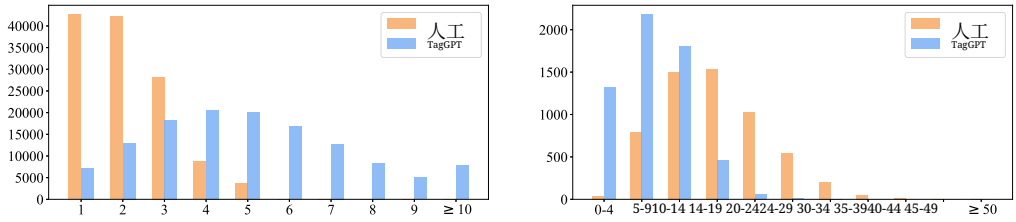
数据集	来源	# 标签	冗余率
快手	人工	37,320	2.85%
	TagGPT	40,238	0.0%
Food.com	人工	458	0.97%
	TagGPT	2,598	0.0%

表 2：两个数据集对应标注系统的“一致性”得分。

数据集	选择式标注器			生成式标注器		
	精确率	召回率	F1	精确率	召回率	F1
快手	78.3	69.4	73.6	80.1	72.1	75.9
Food.com	81.5	75.0	78.1	83.7	76.9	78.7

表 3：TagGPT 中两个替代标注器的定量结果。

促进下游推荐、搜索及其他算法，最终提升其性能。



(a) 快手　　(b) Food.com

图 5：数据集中“实用性”指标中“最少努力”的统计结果。横轴表示分配给单个数据样本的标签数量，纵轴表示样本数量。

为了评估标注系统的“实用性”指标，我们将从两个角度进行分析：分配给每个样本的标签数量（即“最少努力”）以及每个样本内标签的冗余程度（即“多个方面的高覆盖率”）。图 5 展示了数据集中的标签数量，结果表明，与基线方法相比，TagGPT 为每个视频分配的标签显著更多。此外，我们还计算了两种标注系统中每个样本内标签的平均冗余率。由于在标签统计过程中会过滤掉冗余标签，TagGPT 在两个数据集中的冗余率均为 0%。相比之下，基线方法将 0.8 设为统计时的相似度阈值，因此平均标签冗余率分别为 5.86% 和 3.77%。这表明，TagGPT 不仅为每个样本分配的标签多于基线方法，而且还能确保这些标签之间的语义一致性较低，从而符合每个样本的多样化视角。

在计算“一致性”得分时，我们将标签系统中的所有标签两两配对，以确定它们的语义相似度。随后，我们计算相似度超过 0.8 阈值的标签对占所有标签对的比例。结果见表 2，表明 TagGPT 构建的标签系统内部一致性显著优于基于人工标注构建的标签系统。这证明 TagGPT 的标签冗余度更低。此外，综合考虑标签数量及其冗余度，TagGPT 的标签能够覆盖更广泛的内容视角。



（a）根据快手数据集推断出的高频标签。



（b）根据 Food.com 数据集推断出的高频标签。

Fig图 6：TagGPT 构建的标签系统中排名前 150 的词语词云，展示于 bot 日博 h data集。

3.3.2 标签器

本文介绍了两种不同的零样本标签器，即生成式标注器和选择式标注器。为评估这些标签器的有效性，我们从数据集中随机留出 100 个样本作为测试集，这些样本未参与标签系统的开发。我们通过人工评估来衡量标签器的性能，结果见表 3。在零样本设置下，两种标签器都能够为给定数据分配高质量标签。对两种范式的结果进行更细致的比较后发现，选择式标注器的表现略逊于生成式标注器，这主要是由于标签匹配阶段不同。选择式标注器将所有文本线索作为输入，在送入大语言模型之前进行粗粒度标签匹配，这可能导致输入特征过多，并将一些正确的细粒度标签排除在候选集之外。然而，生成式标注器能够有效保留一些细节标签，因此其结果必然更贴近真实情况。

3.4 案例研究

为了更直观地了解 TagGPT 在零样本标注中的表现，本节将展示更多案例研究，以说明定性结果。具体结果见表 4。

对表 4 中来自不同来源的标注结果进行比较后可以发现，在处理多模态数据时，人工标注器通常赋予内容的标签语义较为模糊且有限。这通常是因为微视频创作者在上传作品时，无法快速准确地从海量语义标签中找到所需的语义标签

。此外，由于标签的流行度会直接影响微视频平台的推荐算法，创作者通常倾向于选择流行但含义模糊的标签，以提高作品的曝光度并触达更广泛的受众。

关键帧		人工标注器	选择式标注器	生成式标注器	ACSV	Hanlp
		iPhone小技巧， iCloud 空间不足， 苹果手机	存储空间管理， 苹果设备管理， iCloud 空间不足， 高新数码产品， iPhone 常见问题解决	iCloud 空间管理， 苹果手机技巧， 苹果设备管理， iCloud 存储空间， iCloud备份与恢复	文本， 手机， 屏幕截图， 小工具， 多媒体	云备份， 储存空间， 实用技巧
		iPhone 技巧， 不足 iCloud 空格，iPhone	存储空间管理， 苹果设备管理， iCloud 存储空间不足， 高科技数字产品， iPhone 常见问题解决	iCloud 空间管理， 苹果移动设备手机技巧， 苹果设备管理， iCloud存储空间， iCloud备份与恢复	文本， 移动设备， 屏幕截图， 小工具， 多媒体	云备份， 存储空间， 实用技能
		装修， 设计， 公寓装修	家居DIY ， 多功能房间设计， 单身公寓设计， 小户型设计， 实用家居	迷你公寓设计， 小户型设计， 装修技巧， 创意装修， 居家便捷设计	墙壁， 室内， 门， 瓷砖， 室内设计	鞋柜， 书柜， 创意设计， 公寓
		装饰 ， 设计， 公寓 装饰	家居 DIY， 多功能房间设计， 单身公寓设计， 小户型公寓设计， 实用家居布置	迷你公寓设计， 小型公寓设计， 装饰技巧， 创意装饰， 便捷家居设计	墙， 室内， 门， 瓷砖 室内设计	鞋柜 ， 书柜， 创意设计， 公寓
		比亚迪， 新能源汽车， 唐dmi冠军版	比亚迪唐， SUV汽车， 汽车销量分析， 豪华七座SUV ， 比亚迪汽车	唐DM-i冠军版， 新能源汽车， 比亚迪唐DM-i ， 豪华SUV ， 豪华七座SUV	车辆， 陆地车辆， 文本， 衣服， 汽车设计	老款， 冠军版， 快充， 尊贵， 比亚迪
		BYD 新能源 车辆,唐 dmi 冠军 版本	比亚迪唐， SUV汽车， 汽车销售分析， 豪华七座SUV， 比亚迪汽车	唐DM-i冠军版， 新能源汽车， 比亚迪唐 DM-i， 豪华SUV， 豪华七座SUV	车辆， 陆地车辆， 文本， 服装， 汽车设计	旧款， 冠军版， 快速充电， 尊敬的， BYD
		快来钓鱼， 快手野钓月， 哈利路亚， 快手钓鱼	休闲钓鱼， 钓鱼心得， 钓鱼乐趣， 钓鱼爱好者， 生活秀	钓鱼乐趣， 钓鱼放松， 娱乐休闲， 渔具装备， 钓技攻略	文本， 运动， 水， 海报， 钓鱼	钓鱼， 渔获， 算账， 快手
		快来钓鱼， 快手野钓 月， 哈利路亚， 快手钓鱼	休闲垂钓， 钓鱼体验， 钓鱼乐趣， 钓鱼爱好者， 生活秀	钓鱼乐趣， 钓鱼放松， 娱乐休闲， 钓鱼器材和装备， 钓鱼技巧攻略	文本， 体育， 水 海报 钓鱼	钓鱼， 渔业捕捞， 会计， 快手
		花卷， 花样美食制作， 面食	花卷制作， 自制面食技巧， 美食掌故， 家庭蒸菜， 花样面食	花卷制作， 食用小窍门， 精选主食做法， 葱油花卷制作技巧， 早餐美味面食	食品， 蔬菜	蒸葱， 拌均匀， 花卷， 面食
		蒸花卷 花式 食品制作， 熟制小麦制作的 food	蒸花卷制作， 自制意面的技巧， 美食故事， 家常蒸菜 ， 图案意大利面	蒸花卷 制作、食用建议、 选定的主食 做法，葱油 制作蒸花卷 技能，早餐吃的美味意大利面	食物， 蔬菜	蒸葱头， 搅拌均匀， 蒸花卷， 熟面食

表4：TagGPT 的定性结果。ACSV 和 Hanlp 分别是 https://portal.vision.cognitive.azure.com/demo/generic-image-tagging 和 https://www.hanlp.com/ demonstrate.html。

对 TagGPT 中两种不同范式的标注模型作进一步比较后发现，选择式标注器和生成式标注器都能比人工标注器生成更好的标注结果。选择式标注器会从大致匹配的标签子集中选择并分配标签，因此会在一定程度上造成语义粗糙度，但仍能匹配给定数据的语义。另一方面，生成式标注器能够理解给定的多模态数据，并以更精确、更准确的方式分配标签，使标签与内容细节更加贴合。选择式标注器与生成式标注器之间的比较，本质上是在标签分配过程中对通用性与精确率进行权衡。过度强调通用性可能会导致标签彼此难以区分、缺乏分析价值；而过度强调精确率则可能带来过度解读的风险，使相关内容难以得到正确关联。

4 局限性

经验证，TagGPT 能够从大规模多模态数据中自动构建标签系统，并在零样本设置下为给定数据提供标签。然而，由于设计和技术方面的原因，它仍存在一些局限性：

大语言模型的局限性 直观而言，大语言模型的理解和生成能力与 TagGPT 的性能直接相关。因此，当大语言模型的性能下降时，TagGPT 的性能也会相应受到影响。因此，建议使用性能强劲的大语言模型，或使用经过提示调优的语言模型，以获得最佳效果。

输入长度的局限性 当提供的多模态数据包含过多信息时，可能会导致大语言模型的输入过长，使大多数大语言模型执行截断操作，从而使数据无法使用。这一现象使TagGPT在涉及较长视频或文本的情况下失效。

安全性和隐私的局限性由于大量大语言模型和无监督表征获取方法都源自网络中的API接口，未经谨慎考虑地上传大量用户数据及其他数据，可能会因数据泄露而引发重大的安全性和隐私问题。

5 结论

我们提出了 TagGPT，这是一个简洁且低耦合的框架，为标签系统构建和标注模型提供了一套完整工具。该框架使用户能够轻松获取大规模多模态数据的内容标签并执行零样本标注，可服务于搜索和推荐等各种下游应用。为了基于对多模态内容的理解构建高质量的标签系统，我们引入了一个大语言模型和一个无监督文本表示模型。我们还提出了两种不同的零样本标注范式，二者都能产生高质量的标注结果。未来，我们计划进一步优化 TagGPT 中的各个子模块，以提升其工作效率和准确性。

致谢

TagGPT 项目仍在开发中，旨在打造一个工业级的零样本标签器。TagGPT 是腾讯PCG旗下ARC-QQ联合实验室的合作成果，王泽坤、张云轩、易坤、苏树鹏和刘善松也参与了该项目。

A 指令模板

数据集	指令模板
快手	你是一个视频兴趣标签生成机器人，根据输入的视频标题、类别、光学字符识别、自动语音识别推理出合理的“兴趣标签”，以多个多于两字的标签形式进行表达，以顿号隔开。例如，给定一个视频，它的"标题“为”全屋嵌入式低音音响，主要是这个投影仪真的爱了，“类别"为"房产家居","光学字符识别"为"42平,一室一厅小户型","自动语音识别"为"看，远方灯火闪烁，发出光亮。你独自一人低着头走在路上。这座城市越大，越让人心慌。多向往，多漫长。祝一路行李太多伤。把最初笑容都淡忘。时光让我们变得脆弱，却坚强，让我在爱青青对你唱。我多想能“多陪你唱。把什么生的风景对你讲。”，兴趣标签生成机器人推断出合理的"兴趣标签"为"小户型装修、一室一厅装修、装修效果图”。那么，给定一个新的视频，它的“标题”为“{title}”，“类别"为"{category}","光学字符识别"为"{ocr}","自动语音识别"为"{asr}", 请推断该视频的"兴趣标签":
Food.com	你是一个用于食谱视频的标签生成机器人，能够推断合理的食谱基于输入视频的“菜名”、“描述”和“自动语音识别”生成标签，并表达以包含两个以上单词的标签形式呈现，并以逗号分隔。对于例如，给定一段“菜名”为“烤蒜香芝士”食谱视频。粗玉米糊", "description" 是 "我们喜欢粗玉米糊，这也是另一种不错的做法。搭配烤牛排或烤鸡食用时，是烤土豆的绝佳替代品。我相信这道食谱可以用即食玉米糁制作。 这 2 1/2 小时用于冷藏时间不包括在内。该食谱来自 Taste of Homes清淡又美味。", 以及"自动语音识别"是"在炖锅中将水煮沸；慢慢加入粗玉米糁和盐，不断搅拌；调小火：不加盖炖煮 40–45 分钟，或直至变稠，期间偶尔搅拌。加入奶酪和大蒜，搅拌至奶酪融化。用不粘烹饪喷雾喷洒一个 9 英寸的烤盘；盖上并冷藏 2 至 2 1/2 小时，或直至凝固。开始烧烤前，在烤架上喷上不粘烹饪喷雾；将玉米糊切成 3 英寸见方的小块；刷上橄榄油，确保两面都刷到。烧烤，涵盖了，用中火加热4至6分钟每面煎至浅金黄色，", 标签生成机器人推断出一个合理的将食谱标记为“制作时间、菜式、主要食材、准备方式、场合、配菜-菜肴、蛋类和乳制品、冰箱、糖尿病患者、素食者、谷物、奶酪、炉灶, 饮食类、低胆固醇、低卡路里、慰藉食物、低碳水、某种成分含量低、意大利面-米饭和谷物, 早午餐, 口味-情绪, 设备, 摆盘, 热食, 4 小时或更短”。然后，给定一个“菜名”为“{name}”的新视频，“描述”为“{caption}”，且“自动语音识别”为“{asr}”，请推断该视频的食谱标签为:

表 5：TagGPT 中用于标签生成的指令模板。

本节展示了用于不同语言和不同数据集的指令模板，如表 5 所示。

参考文献

[1] Sajad Ahmadian、MiladAhmadian 和Mahdi Jalili。基于深度学习的信任与标签感知推荐系统。神经计算，488：557–571，2022。[2] Yali Amit、Pedro Felzenszwalb 和 Ross Girshick。目标检测。计算机视觉：参考指南，第 1–9 页，2020。[3] Tom B. Brown、BenjaminMann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz

Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher B埃尔纳, Sam McCandlish, Alec Radford、IlyaSutskever 和Dario Amodei。《语言模型是少样本学习器》，2020。[4] Arindam Chaudhuri、Krupa Mandaviya、Pratixa Badelia、Soumya K Ghosh、Arindam Chaudhuri、KrupaMandaviya、Pratixa Badelia 和Soumya K Ghosh。《光学字符识别系统》。Springer，2017。[5] Aakanksha Chowdhery、SharanNarang 和等。《PaLM：利用路径扩展语言建模》，2022。[6] Maarten Clements、Arjen PdeVries 和 Marcel JT Reinders。《个性化对社交媒体搜索中标签查询长度的影响》。《信息处理与管理》，46(4)：403–412，2010。[7] Tianyu Gao、Xingcheng Yao 和 Danqi Chen。《SimCSE：句子嵌入的简单对比学习》。《*arXiv* 预印本*arXiv:2104.08821*》，2021。[8] Manish Gupta、Rui Li、Zhijun Yin 和 Jiawei Han。《社交标签技术综述》。《*ACM Sigkdd*探索通讯》，12(1)：58–72，2010。[9] Fouzia Jabeen、Shah Khusro、Amna Majid 和 Azhar Rauf。《社交标注系统中的语义发现：综述》。《多媒体工具与应用》，75：573–605，2016。[10] Yu Kong 和 Yun Fu。《人工动作识别与预测：综述》。《国际计算机视觉杂志》，130(5)：1366–1401，2022。[11] Sewon Min、Xinxi Lyu、Ari Holtzman、Mikel Artetxe、Mike Lewis、Hannaneh Hajishirzi 和Luke Zettlemoyer。《重新思考示例的作用：是什么让上下文学习发挥作用？》。《*arXiv* 预印本*arXiv:2202.12837*》，2022。[12] Liqiang Nie、Leigang Qu、Dai Meng、Min Zhang、QiTian 和 Alberto Del Bimbo。《面向搜索的微视频字幕生成》。载于《第*30*届 *ACM*国际多媒体会议论文集》，第3234–3243页，2022。[13] OpenAI。《GPT-4 技术报告》，2023。[14] Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、SandhiniAgarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、JanLeike 和RyanLowe。《训练语言模型遵循人类反馈指令》，2022。[15] Priyanka Panchal 和Dinesh J Prajapati。《使用深度学习方法进行图像和视频的社交标签推荐》。载于《情感分析与深度学习：*ICSADL 2022*论文集》，第241–261页。Springer，2023。[16] Filip Radlinski、Craig Boutilier、DeepakRamachandran 和 IvanVendrov。《对话式推荐系统中的主观属性：挑战与机遇》。载于《*AAAI* 人工智能会议论文集》，第36卷，第12287–12293页，2022。[17] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timo- thée Lacroix、BaptisteRozière、NamanGoyal、Eric Hambro、Faisal Azhar、Aurelien Rodriguez、Armand Joulin、EdouardGrave和 GuillaumeLample。《Llama：开放且高效的基础语言模型》，2023。[18] Zhenhailong Wang、Manling Li、Ruochen Xu、Luowei Zhou、Jie Lei、Xudong Lin、Shuohang Wang、Ziyi Yang、CheнguangZhu、Derek Hoiem 等。《带有图像描述符的语言模型是强大的少样本视频语言学习器》。《*arXiv* 预印本*arXiv:2205.10747*》，2022。[19] BigScience Workshop、：、Teven Le Scao、Angela Fan 和等。《BLOOM：一个拥有1760亿参数的开放获取多语言语言模型》，2023。[20] Zhichen Xu、YunFu、JianchangMao 和 Difu Su。《迈向语义网：协作式标签建议》。载于《*WWW2006*协作式网络标注研讨会，爱丁堡，苏格兰》，2006。

[21] DongYu 和 Li Deng。自动语音识别，第 1卷。Springer，2016 年。[22] Andy Zeng、Maria Attarian、Brian Ichter、Krzysztof Choromanski、Adrian Wong、Stefan Welker、Federico Tombari、Aveek Purohit、MichaelRyoo、Vikas Sindhwani、Johnny Lee、Vincent Vanhoucke和 Pete Florence。苏格拉底模型：将零样本多模态推理与语言结合，2022 年。[23] Yang Zhang、Fuli Feng、Xiangnan He、Tianxin Wei、Chonggang Song、GuohuiLing和Yongdong Zhang。在推荐中通过因果干预利用流行度偏差。载于第 *44* 届国际*ACM SIGIR* 信息检索研究与发展会议论文集，第 11–20 页，2021 年。