arXiv:2305.02677v3 [cs.CV] 6 Jul 2023

Caption Anything：通过多样化多模态控制实现交互式图像描述

Teng Wang[*†]、**Jinrui Zhang**[*]、**JunjieFei**[*]、**Hao Zheng**、**Yunlong Tang**、**ZheLi**、**Mingqi Gao**、**Shanshan Zhao**南方科技大学 VIP 实验室



图 1：Caption Anything 支持多种视觉和语言控制，可轻松适用于以物体为中心的聊天和图像段落描述。

摘要

可控图像描述是一个新兴的多模态主题，旨在根据人类意图用自然语言描述图像，例如，关注指定区域，或采用特定的文本风格进行描述。当前最先进的方法都基于输入控制与输出描述的标注对进行训练。然而，这类高质量标注多模态数据的匮乏，在很大程度上限制了这些方法在交互式人工智能系统中的实用性和可扩展性。利用单模态指令跟随基础模型是一种很有前景的替代方案，它能够受益于更广泛的数据来源。本文提出 Caption AnyThing（CAT），一种基础模型增强型图像描述框架，支持多种多模态控制：1）视觉控制，包括点、框和轨迹；2）语言控制，例如情感、长度、语言和事实性。借助 Segment Anything Model（SAM）和 ChatGPT，我们将视觉提示和语言提示整合到一个模块化框架中，从而能够灵活组合不同控制。大量案例研究展示了该框架对用户意图的对齐能力，为视觉语言应用中的有效用户交互建模提供了启示。我们的代码已公开发布于https://github.com/ttengwang/Caption-Anything。

[*] 贡献相同。 [†] 本工作在腾讯PCG的ARC实验室实习期间完成。感谢 Yixiao Geand Ying Shan 的支持和富有建设性的讨论。

技术报告。

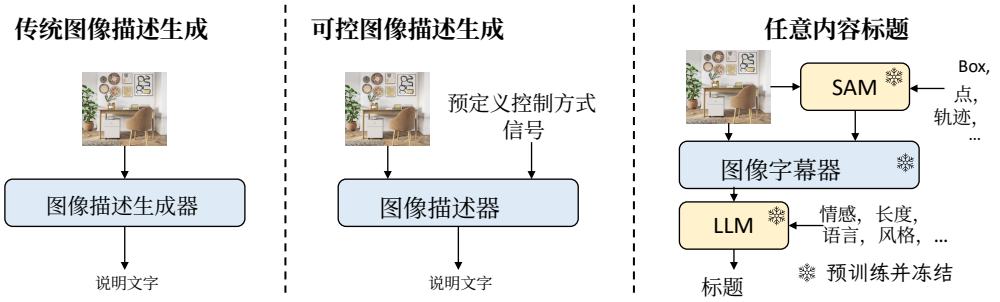

图 2：图像描述流程对比。传统图像描述方法缺乏明确的控制能力，因此不适合与用户交互。以往的可控图像描述方法主要依赖规模有限、由人工标注且带有特定控制信号的数据，并且仅支持预定义的控制信号。本文提出的 CAT 无需训练，并支持多样的视觉控制和语言控制。

1 引言

用自然语言描述图像是视觉语言学习中的一个关键问题。图2展示了不同图像描述系统之间的差异。基础图像描述 [11, 8, 20,9, 40] 和密集图像描述 [13, 47, 43] 通常会以客观且实用的语言呈现图像的显著特征，可以是一个句子，也可以是一组句子。然而，由于缺乏与用户意图相匹配的明确控制，这类方法生成的内容可能过于简短或过于复杂，不适合用于用户交互。可控图像描述（CIC）是一个很有前景的研究方向，旨在使语言输出与用户意图相匹配。随后有研究 [6, 3, 5, 29]提出将多种控制信号融入图像描述模型。然而，这些方法的适用性受到两个主要因素的限制：（1）现有 CIC 模型通常依赖于使用人工标注的（图像、文本、控制信号）元组进行训练 [10, 25]。数据集规模有限，制约了这些模型理解控制信号的能力；（2）这些模型仅支持预定义的一种或几种控制信号，因此难以灵活组合不同控制，也难以引入新的可控维度。

为解决这些问题，我们提出 Caption AnyThing（CAT），一种借助预训练基础模型增强的零样本可控图像描述框架。具体而言，CAT 将预训练图像描述器 [18, 17, 38] 与 SAM [14] 以及经过指令微调的大语言模型相结合。首先，SAM 对图像和视觉控制进行处理，生成与选定区域对应的像素级掩码，从而便于以用户感兴趣的对象为中心进行感知。得益于 SAM 训练期间使用的多种视觉提示（例如，点、边界框），CAT 在与用户交互时支持灵活的视觉控制。随后，经过指令微调的大语言模型会进一步润色输出语句。GPT-4 [27] 和 FLAN-T5 [4]，等知名大语言模型经过人类反馈调优，使 CAT 能够适应各种语言控制，并更有效地契合用户意图。

与现有的可控图像描述方法不同，CAT 借助基础模型实现可控性，而非仅依靠从训练数据中学习预定义的控制信号。这种方法不仅降低了对人工标注数据的依赖，使模型无需训练，还通过利用海量预训练数据中蕴含的知识，增强了模型的迁移能力。此外，CAT 支持多种控制信号，因此具有很强的交互适应性和可扩展性。目前，它支持 3 种视觉控制信号（点击、框、轨迹）和 4 种语言控制信号（情感、长度、语言、事实性），这些信号可以灵活组合，以生成多样化且个性化的图像描述。此外，我们的模型为这两类控制提供了统一的表示。具体而言，视觉控制和语言控制分别统一表示为像素级掩码和文本提示。基于这一设计，CAT 可以轻松扩展，以支持任何能够转换为这些统一表示的控制信号，从而增强其灵活性和可扩展性。

我们通过一系列全面的定性示例，展示 CAT 强大的用户交互能力。如图 1 所示，用户可以通过各种视觉控制选择感兴趣的物体，并以偏好的风格生成图像描述。此外，通过集成额外的光学字符识别和视觉问答

工具，CAT 可以轻松扩展为两种多模态应用，即以物体为中心的聊天和图像段落描述。前者允许用户围绕特定物体进行聊天，从而加深对感兴趣物体的理解。后者有效整合了来自不同基础模型、涵盖多个领域的知识，能够生成详尽且逻辑连贯的描述。总之，这些丰富的案例表明，CAT 是一个交互性很强的多模态系统，在实际应用中展现出巨大潜力。

总而言之，本文的贡献主要体现在以下三个方面：1）我们提出了一种基于基础模型构建的免训练可控图像描述（CIC）框架，从而降低了对人工标注数据的依赖。2）我们的方法支持更加多样化的控制方式，并为视觉控制和语言控制提供了统一的表示形式，从而便于扩展并纳入新的可控性维度。3）实验表明，CAT 具有强大的用户交互能力。

2 相关工作

图像描述与密集图像描述。图像描述是一项多模态任务，旨在为给定图像生成描述。当前主流方法主要采用编码器-解码器范式 [36, 44, 1, 12, 18, 17, 38]来解决这项任务。为加强对复杂场景的细致理解，密集图像描述 [13, 47, 19, 48, 31, 43] 被提出，用于为图像中的所有显著物体生成局部描述。Yin 等人 [48] 提出了一种多尺度上下文信息共享技术，以捕捉细粒度的全局上下文信息。Wu 等人 [43] 提出了一种用于密集图像描述和目标检测的统一框架，从而提升了物体理解能力。尽管如此，这些方法仅根据图像的显著特征生成描述，因此不适合用户交互，因为它们缺乏与用户意图相匹配的显式控制机制。与这些方法相比，我们的方法专注于交互式图像描述，并支持多样化的多模态控制，从而更好地契合用户需求。

可控图像描述。 生成与用户指定的图像中物体相关的期望描述的可控能力在实际应用中非常有用。一方面，可控图像描述中的视觉控制通常涉及边界框 [5], 鼠标行为 [29, 46]。Cornia 等人 [5] 提出根据一系列或一组图像区域生成相应的图像描述。Pont-Tuset 等人 [29] 发布了一个交互式图像描述数据集，要求标注者描述鼠标轨迹覆盖的图像区域。LoopCAG [46]进一步提升了图像描述的生成质量和交互式可控性。另一方面，生成的图像描述的文本风格可以根据不同的应用场景灵活改变。通常，可控的文本风格可以概括为：长度可控性 [6],情感可控性 [25], 想象力可控性（例如，浪漫或幽默的描述） [51,10]。一些工作 [49, 39] 将这些可控风格统一到单一架构中。BLIP2 [17], 这一视觉语言预训练模型遵循自然语言提示执行零样本图像到文本生成。在本文中，我们进一步扩展了可控图像描述的灵活性：交互可以是用户指定的点、框或轨迹，并且生成描述的语言风格能够最大程度地满足用户的要求。

交互式图像分割。 交互式图像分割 [22, 37, 45, 32] 是一项重要的研究课题，要求模型根据用户的意图，通过交互对图像进行分割（例如，点、轨迹）。Xu 等人 [45] 首次提出基于深度学习的交互式图像分割方法。Li 等人提出了一种端到端的交互式图像分割方法 [21]。整个架构分为两个卷积网络，第一个网络根据用户输入生成各种掩码，第二个网络则经过训练，从这些掩码中选出一个解。FCA-Net [23] 更好地利用首次点击来提升交互式分割效果。SAM [14] 构建了一个可通过提示进行分割的基础模型，其中提示可以是任何能够指示图像中待分割内容的信息，例如一组点、粗略的框或掩码，或自由形式的文本。

大语言模型。 大语言模型因其在多种语言处理任务中展现出的强大迁移能力，以及与人类交互的新兴能力而受到广泛关注。近期，最重要的突破来自拥有 175B 参数的 GPT-3 [2],a 模型，它揭示了少样本学习技术的巨大潜力。这一卓越表现推动了众多后续大语言模型的发展 [50, 4, 35, 30]。为了增强

大语言模型与人类用户的交互能力，当前最先进的模型通常会使用人类反馈进行微调 [28, 33, 26, 27, 34]。这种方法使大语言模型能够有效适应多样的语言指令，从而更贴近用户意图。受大语言模型用户对齐能力的启发，我们将语言控制统一为文本提示，并应用大语言模型生成符合用户偏好的图像描述。

3 Caption Anything

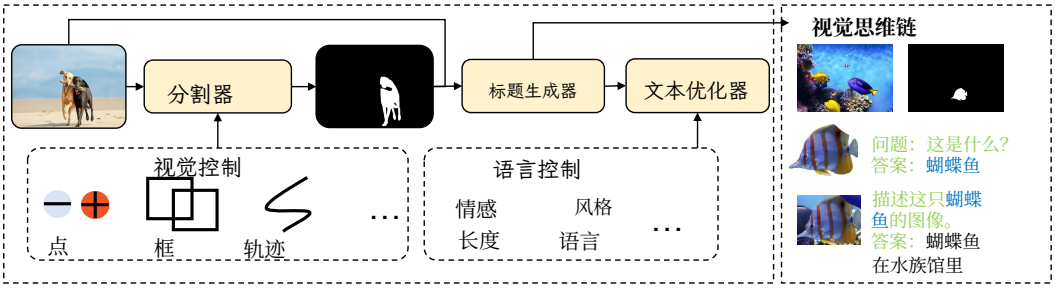

图3：Caption Anything 的整体框架。它将多模态控制引入图像描述，使生成的视觉焦点和语言风格能够与人类意图相契合。分割器首先将视觉提示转换为掩码提示。随后，图像描述器根据掩码所界定的区域生成原始描述。为了让图像描述器聚焦于用户感兴趣的对象，我们采用一种简单的视觉思维链技术进行逐步推理。最后，将文本提示和原始描述一并输入文本优化器，由其根据所需体裁生成用户偏好的描述。

为了增强当前图像描述系统以用户为中心的交互性，我们提出一种基础模型增强策略，使图像描述器能够适配多种多模态控制。具体而言，我们的方法可以表述为一个三元组求解器 {分割器, 图像描述器, 文本优化器}。如图3所示，分割器首先接收交互式视觉控制（例如点、框、轨迹），并通过像素级掩码表示用户感兴趣的区域。随后，图像描述器根据原始图像和提供的掩码，生成与指定区域相关的原始描述。为了帮助图像描述器聚焦于用户感兴趣的对象，我们设计了一种采用逐步推理的视觉思维链技术。最后，文本优化器通过融入用户定义的语言控制来优化原始描述，从而根据用户偏好调整语言风格。

分割器。理想的分割器能够根据视觉控制分割图像的任意部分。SAM [14] 表现出色，满足各项要求，并凭借可提示预训练和 SA-1B 数据集，在新的图像领域展现出令人瞩目的零样本迁移能力 [1]（该数据集是规模最大的分割数据集，包含 1,100 万张图像上的 10 亿个掩码）。SAM 将交互式分割 [24]扩展为可提示能力：提示是指用于指示图像中要分割内容的任意交互（例如点、框），SAM 根据提示生成有效的分割掩码。获得用户指定的分割掩码后，就可以根据原始图像和掩码提示轻松生成所需的图像描述。

图像描述器。为了描述图像中用户指定的任意物体，图像描述器需要具备出色的零样本图像描述能力。换言之，理想的图像描述器应能针对各种新颖物体和不同的图像分布生成合理的描述。我们使用 BLIP2 作为图像描述器。它利用冻结的预训练图像编码器和冻结的大语言模型，并结合查询变换器来弥合模态差距，从而实现出色的零样本表现。

文本优化器。 在大多数情况下，与图像相关的描述应遵循用户的偏好。然而，根据用户指令优化由图像描述器 生成的原始描述是一项并非易事的任务。为实现这一目标，我们引入ChatGPT 作为 API，以生成更具表现力且更易控制的

[1]项目链接：https://segment-anything.c om

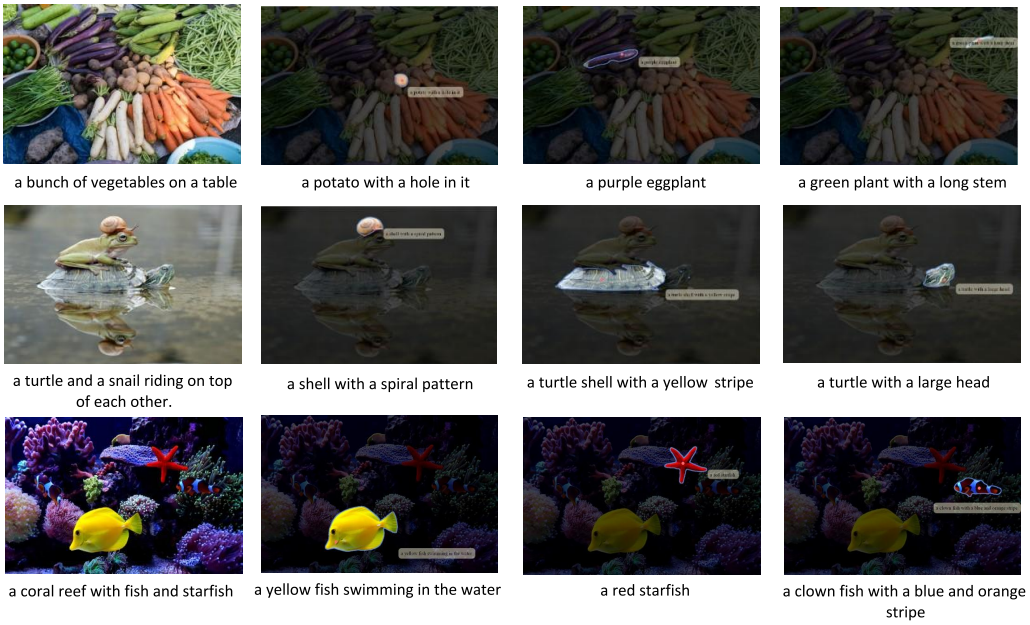

图4：使用基于点的视觉控制描述图像的可视化。

描述。 文本优化器 可以轻松替换为开源的大语言模型，例如 LLaMA [35], OPT-IML [50],BLOOM [30]。

视觉思维链。 我们通过实证发现，生成的物体描述很容易受到背景信息的影响。受源自自然语言处理的思维链（CoT）提示启发， [41, 15], 我们通过逐步生成文本来引导图像描述器 ，确保生成的描述聚焦于用户选定区域。如图 3 所示，我们首先保留用户选定物体，将背景替换为白色，并要求图像描述器 识别感兴趣物体的类别。随后，图像描述器 以生成的文本和带背景的裁剪图像作为提示，生成最终描述。通过这种方式，图像描述器 能够在生成图像描述的过程中聚焦于选定物体。

扩展至以物体为中心的聊天。 受到多模态对话系统近期进展的启发，我们探究了视觉对话的潜力，尤其是针对视觉提示所识别物体的对话。与旨在全面理解整张图像的主流聊天系统不同，局部区域对话在处理复杂、信息丰富且高分辨率的图像时具有更广泛的应用。这些应用包括视觉导航、具身智能和幼儿教育。给定物体的分割掩码和用户查询，我们使用现成的视觉问答模型 [17] 作为视觉 API，使 ChatGPT 能够通过提问来理解详细的视觉线索。具体而言，我们将生成的标题加入初始提示，并遵循 [42] 使用 LangChain [16] 作为控制中心来预测 API 调用链。

扩展至段落描述生成。 我们的框架可适用于图像段落描述生成任务：借助 ChatGPT 将图像的密集描述和场景文本总结为一个段落。具体而言，我们首先使用 SAM 对图像中的所有内容进行分割，再通过 CAT 流程为每个物体生成描述。为将场景文字信息融入段落，我们还使用其他光学字符识别工具（例如，EasyOCR [7]）识别图像中的文字。随后，我们将密集描述和场景文本合并到预定义的提示模板中，并用它来指示 ChatGPT 将场景信息总结为连贯的段落。



图 5：使用视觉控制（轨迹和框）描述图像的可视化展示。



图 6：带有语言控制的生成图像描述可视化（第一行：多种语言；第二行：情感和事实性）。

4 实验

视觉控制。 如图所示。 4，我们展示了一些定性结果，展现了 CAT 卓越的视觉控制能力。通过在图像中的不同位置放置点击点提示，CAT 能够准确识别并描述相应物体，体现出其为任意给定图像中的各类物体生成描述的非凡能力。此外，如图所示。 5，CAT 的视觉控制可以基于轨迹，也可以基于边界框，进一步彰显了其为各种图像内容生成准确描述时的灵活性与适应性。

语言控制。 我们提供了更多定性结果，展示了模型出色的语言控制能力，如图 6所示。借助先进的语言控制（即情感和事实性），CAT 能够生成具有多种语言风格的图像描述，从轻松随意、对话式到正式且信息丰富。这些结果凸显了模型根据用户的具体需求和偏好调整语言输出的非凡能力。

以物体为中心的聊天。 在图 7中，我们进一步展示了 CAT 卓越的以物体为中心的聊天能力。具体而言，我们的模型能够执行视觉问答

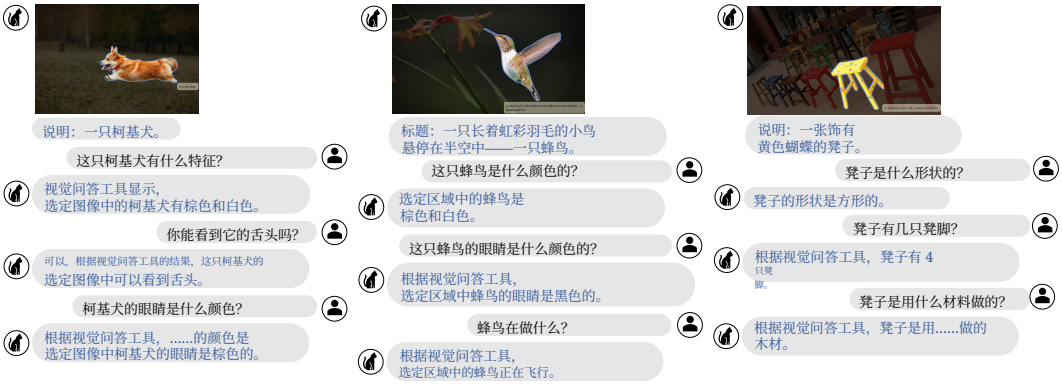

图 7：以物体为中心的聊天示例。



图 8：视觉思维链示例。



在这张图像中，我们看到一只熊猫和她的幼崽坐在地面上。熊猫妈妈体型较大，有黑白色皮毛，而她的幼崽体型较小，也有黑白色皮毛。它们都坐在地面上，熊猫妈妈稍微坐在幼崽后方。地面看起来像是覆盖着草，图像中没有其他可见物体。它们似乎身处一处天然的环境中，可能是在森林或动物园里。



这张图片中可以看到一只浣熊坐在岩石上。这只浣熊体型中等，有着独特的黑白色皮毛。岩石很大，看起来由粗糙的纹理材料制成。背景中有两处场景文字，一处标注为“Jannes”另一处标注为“Pockele”，两者都位于图像右侧。总的来说，这张图似乎描绘了一幅宁静的户外场景，其中浣熊是画面的主体。



图像中央是一张大型床，配有黑色床头板和床尾板。这张床看起来这张床是大号双人床尺寸，铺着白色床单和灰色被子。床头板.....呈矩形，质地光滑有光泽。床尾板同样呈矩形，并且有一个与床头板的纹理相似。床位于房间中央，但没有其他图像中可见的物体或细节。



在这张图片中，我们看到一排小提琴挂在墙上。这些小提琴的大小和形状都相似，它们有着弯曲的身体和修长的脖颈。它们看起来是用木材制成的，表面光滑，而且光滑的纹理。小提琴摆放得彼此靠近，琴弦朝外。大小规律他们身后的墙壁呈中性色调，为这些乐器提供了一个简洁的背景。总体而言，这个图像中描绘了一组制作精良的小提琴，可能陈列在与音乐相关的场所。

图 9：图像段落描述示例。

聚焦于图像中选定的物体，展现出其围绕特定视觉内容进行丰富且有意义对话的卓越能力。

视觉思维链。 图8展示了多个示例，说明视觉思维链的有效性。如图所示，对包含背景的图像进行直接推理往往会受到背景内容的不利影响，从而妨碍图像描述器专注于感兴趣的物体。通过引入逐步思考过程，生成的图像描述能够更加聚焦。同时，视觉思维链有助于揭示感兴趣物体的更多细节，使整个系统能够更有效地契合用户意图。

用一个段落描述所有内容。 图 9 展示了若干生成段落的示例。这些生成的段落涵盖了场景中的大多数物体和文本信息，甚至包含了一些推理信息。总体而言，段落提供的描述通常准确且逻辑连贯。

5 结论

本文提出了一种以基础模型为增强组件的可控图像描述框架 Caption AnyThing（CAT），旨在解决现有可控图像描述（CIC）方法的局限性。我们提出的框架利用预训练的图像描述器，并将其与 SAM 和经过指令微调的大语言模型（LLM）相结合，从而减少对人工标注数据的依赖，并扩大所支持的控制信号范围。CAT 对控制信号的统一表示提升了模型的灵活性和可扩展性，使其易于适配交互式应用，并能够扩展到新的可控维度。实验表明，CAT 为可控图像描述任务提供了一种无需训练且适应性强的解决方案，并具备出色的用户交互能力。

参考文献

[1] Peter Anderson、Xiaodong He、Chris Buehler、Damien Teney、Mark Johnson、Stephen Gould 和 Lei Zhang。用于图像描述和视觉问答的自下而上与自上而下注意力。载于《IEEE计算机视觉与模式识别会议论文集》，第6077–6086页，2018年。3[2] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。《神经信息处理系统进展》，33：1877–1901，2020年。3[3] Long Chen、Zhihong Jiang、Jun Xiao 和 Wei Liu。通过动词特定语义角色实现类人可控图像描述。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第16846–16856页，2021年。2[4] Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Eric Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma 等。扩展指令微调语言模型。arXiv预印本 arXiv:2210.11416，2022年。2, 3[5] Marcella Cornia、Lorenzo Baraldi 和 Rita Cucchiara。展示、控制与描述：一种生成可控且有依据的图像描述的框架。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8307–8316页，2019年。2, 3[6] Chaorui Deng、Ning Ding、Mingkui Tan 和 Qi Wu。长度可控的图像描述。载于《计算机视觉——ECCV 2020：第16届欧洲计算机视觉会议，英国格拉斯哥，2020年8月23日至28日，论文集，第XIII部分16》，第712–729页。Springer，2020年。2, 3[7] EasyOCR。https://www.jaided.ai/easyocr，2021年。5[8] Zhiyuan Fang、Jianfeng Wang、Xiaowei Hu、Lin Liang、Zhe Gan、Lijuan Wang、Yezhou Yang 和 Zicheng Liu。将语义概念注入端到端图像描述。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第18009–18019页，2022年。2

[9] Zhengcong Fei、Xu Yan、Shuhui Wang 和 Qi Tian。Deecap：通过动态提前退出实现高效图像描述。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12216–12226页，2022年。2[10] Chuang Gan、Zhe Gan、Xiaodong He、Jianfeng Gao 和 Li Deng。Stylenet：生成具有风格的吸引人视觉描述。载于《IEEE计算机视觉与模式识别会议论文集》，第3137–3146页，2017年。2, 3[11] Xiaowei Hu、Zhe Gan、Jianfeng Wang、Zhengyuan Yang、Zicheng Liu、Yumao Lu 和 Lijuan Wang。扩展视觉-语言预训练以实现图像描述。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第17980–17989页，2022年。2[12] Lun Huang、Wenmin Wang、Jie Chen 和 Xiao-Yong Wei。用于图像描述的注意力之上的注意力。载于《IEEE/CVF国际计算机视觉会议论文集》，第4634–4643页，2019年。3[13] Justin Johnson、Andrej Karpathy 和 Li Fei-Fei。Densecap：用于密集图像描述的全卷积定位网络。载于《IEEE计算机视觉与模式识别会议论文集》，第4565–4574页，2016年。2, 3[14] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C. Berg、Wan-Yen Lo、Piotr Dollár 和 Ross Girshick。任意分割。arXiv预印本 arXiv:2304.02643，2023年。2, 3, 4[15] Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yutaka Matsuo 和 Yusuke Iwasawa。大语言模型是零样本推理器。arXiv预印本 arXiv:2205.11916，2022年。5[16] LangChain。https://github.com/hwchase17/langchain，2022年。5[17] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型实现语言-图像预训练。arXiv预印本 arXiv:2301.12597，2023年。2, 3, 5[18] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。BLIP：通过语言-图像预训练实现统一的视觉-语言理解与生成。国际机器学习会议，第12888–12900页，2022年。2, 3[19] Xiangyang Li、Shuqiang Jiang 和 Jungong Han。学习密集图像描述中的目标上下文。载于《AAAI人工智能会议论文集》，第33卷，第8650–8657页，2019年。3[20] Yehao Li、Yingwei Pan、Ting Yao 和 Tao Mei。理解并排序图像描述中的语义。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第17990–17999页，2022年。2[21] Zhuwen Li、Qifeng Chen 和 Vladlen Koltun。具有潜在多样性的交互式图像分割。《IEEE计算机视觉与模式识别会议论文集》，第577–585页，2018年。3[22] JunHao Liew、Yunchao Wei、Wei Xiong、Sim-Heng Ong 和 Jiashi Feng。区域交互式图像分割网络。2017 IEEE国际计算机视觉会议（ICCV），第2746–2754页，2017年。3[23] Zheng Lin、Zhao Zhang、Lin-Zhuo Chen、Ming-Ming Cheng 和 Shao-Ping Lu。利用首次点击注意力实现交互式图像分割。《IEEE/CVF计算机视觉与模式识别会议论文集》，第13339–13348页，2020年。3[24] Sabarinath Mahadevan、Paul Voigtlaender 和 Bastian Leibe。迭代训练的交互式分割。arXiv预印本 arXiv:1805.04398，2018年。4[25] Alexander Mathews、Lexing Xie 和 Xuming He。Senticap：生成带有情感的图像描述。载于《AAAI人工智能会议论文集》，第30卷，2016年。2, 3

[26] Niklas Muennighoff、Thomas Wang、Lintang Sutawika、Adam Roberts、Stella Biderman、Teven LeScao、M Saiful Bari、Sheng Shen、Zheng-Xin Yong、Hailey Schoelkopf 等人。通过多任务微调实现跨语言泛化。 *arXiv* 预印本 *arXiv:2211.01786*，2022年。4

[27] OpenAI。《GPT-4技术报告》，2023年。2，4[28] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray等。通过人类反馈训练语言模型遵循指令。神经信息处理系统进展，35：27730–27744，2022年。4[29] Jordi Pont-Tuset、Jasper Uijlings、Soravit Changpinyo、Radu Soricut和Vittorio Ferrari。通过局部化叙述连接视觉与语言。载于《计算机视觉——ECCV 2020：第16届欧洲会议》，英国格拉斯哥，2020年8月23日至28日，论文集，第V部分16，页码647–664，2020年。2，3[30] Teven Le Scao、Angela Fan、Christopher Akiki、Ellie Pavlick、Suzana Ili´c、Daniel Hesslow、Roman Castagné、Alexandra Sasha Luccioni、François Yvon、Matthias Gallé等。BLOOM：一个拥有1760亿参数的开放获取多语言语言模型。arXiv预印本 arXiv:2211.05100，2022年。3，5[31] Zhuang Shao、Jungong Han、Demetris Marnerides和Kurt Debattista。通过 Transformer 实现区域-物体关系感知的密集图像描述。IEEE神经网络与学习系统汇刊，2022年。3[32] Konstantin Sofiiuk、Ilia Petrov、Olga Barinova和Anton Konushin。f-BRS：重新思考交互式分割中的反向传播细化。IEEE/CVF计算机视觉与模式识别会议论文集，页码8623–8632，2020年。3[33] Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang和Tatsunori B. Hashimoto。斯坦福 Alpaca：一种指令跟随型 LLaMA 模型。https://github.com/tatsu-lab/stanford_alpaca，2023年。4[34] Ross Taylor、Marcin Kardas、Guillem Cucurull、Thomas Scialom、Anthony Hartshorn、Elvis Saravia、Andrew Poulton、Viktor Kerkez和Robert Stojnic。Galactica：科学领域大语言模型。arXiv预印本 arXiv:2211.09085，2022年。4[35] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timo-thée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar等。LLaMA：开放且高效的基础语言模型。arXiv预印本 arXiv:2302.13971，2023年。3，5[36] Oriol Vinyals、Alexander Toshev、Samy Bengio和Dumitru Erhan。Show and Tell：神经图像描述生成器。载于IEEE计算机视觉与模式识别会议论文集，页码3156–3164，2015年。3[37] Guotai Wang、Maria A Zuluaga、Wenqi Li、Rosalind Pratt、Premal A Patel、Michael Aertsen、Tom Doel、Anna L David、Jan Deprest、Sébastien Ourselin等。DeepIGeoS：用于医学图像分割的深度交互式测地线框架。IEEE模式分析与机器智能汇刊，41(7)：1559–1572，2018年。3[38] Jianfeng Wang、Zhengyuan Yang、Xiaowei Hu、Linjie Li、Kevin Lin、Zhe Gan、Zicheng Liu、Ce Liu和Lijuan Wang。GIT：用于视觉与语言的生成式图像到文本 Transformer。arXiv预印本 arXiv:2205.14100，2022年。2，3[39] Ning Wang、Jiahao Xie、Jihao Wu、Mingbo Jia和Linlin Li。通过提示实现可控图像描述。arXiv预印本 arXiv:2212.01803，2022年。3[40] Teng Wang、Yixiao Ge、Feng Zheng、Ran Cheng、Ying Shan、Xiaohu Qie和Ping Luo。通过自由语言建模加速视觉-语言预训练。arXiv预印本 arXiv:2303.14038，2023年。2

[41] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed Chi、Quoc Le和Denny Zhou。思维链提示能够激发大语言模型的推理能力。arXiv预印本 arXiv:2201.11903，2022年。5[42] Chenfei Wu、Shengming Yin、Weizhen Qi、Xiaodong Wang、Zecheng Tang和Nan Duan。Visual ChatGPT：借助视觉基础模型进行对话、绘图和编辑。arXiv预印本 arXiv:2303.04671，2023年。5[43] Jialian Wu、Jianfeng Wang、Zhengyuan Yang、Zhe Gan、Zicheng Liu、Junsong Yuan和Lijuan Wang。GRIT：用于物体理解的生成式区域到文本转换器。arXiv预印本 arXiv:2212.00280，2022年。2，3[44] Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel和Yoshua Bengio。Show, Attend and Tell：结合视觉注意力的神经图像描述生成。载于国际机器学习会议，页码2048–2057。PMLR，2015年。3[45] Ning Xu、Brian Price、Scott Cohen、Jimei Yang和Thomas S Huang。深度交互式物体选择。IEEE计算机视觉与模式识别会议论文集，页码373–381，2016年。3[46] Kun Yan、Lei Ji、Huaishao Luo、Ming Zhou、Nan Duan和Shuai Ma。对图像描述进行时空控制。载于计算语言学协会第59届年会暨国际自然语言处理联合会议第11届会议论文集（第1卷：长篇论文），页码2014–2025，2021年。3[47] Linjie Yang、Kevin Tang、Jianchao Yang和Li-Jia Li。通过联合推理和视觉上下文实现密集图像描述。载于IEEE计算机视觉与模式识别会议论文集，页码2193–2202，2017年。2，3[48] Guojun Yin、Lu Sheng、Bin Liu、Nenghai Yu、Xiaogang Wang和Jing Shao。上下文与属性锚定的密集图像描述。载于IEEE/CVF计算机视觉与模式识别会议论文集，页码6241–6250，2019年。3[49] Zequn Zeng、Hao Zhang、Zhengjue Wang、Ruiying Lu、Dongsheng Wang和Bo Chen。ConZIC：通过基于采样的润色实现可控零样本图像描述。arXiv预印本 arXiv:2303.02437，2023年。3[50] Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin等。OPT：开放的预训练 Transformer 语言模型。arXiv预印本 arXiv:2205.01068，2022年。3，5[51] Wentian Zhao、Xinxiao Wu和Xiaoxun Zhang。MemCap：记忆图像描述的风格知识。载于AAAI人工智能会议论文集，第34卷，页码12984–12992，2020年。3