

# 地面化自适应模型：面向多样化视觉任务的开放世界模型组装

国际数字经济研究院与社区

代码和演示：<https://github.com/IDEA-Research/Grounded-Segment-Anything>



图1：Grounded SAM能够根据用户提供的任意文本输入，同时检测并分割图像中的对应区域。该系统还能与其他开放世界模型无缝集成，以完成更复杂的视觉任务。

## 摘要

我们提出**Grounded SAM**，该方法将Grounding DINO[38]作为开放集目标检测器，与Segment Anything模型(SAM)[25]进行集成。这种集成

该系统能够基于任意文本输入实现区域检测与分割，并为连接多种视觉模型开辟了新途径。如图1所示，通过灵活运用Grounded SAM流程，可完成多种视觉任务。例如，该流程可实现自动标注流程。

仅基于输入图像的实现可通过整合BLIP[31]和RecognizeAnything[83]等模型来完成。此外，整合Stable-Diffusion[52]可实现可控的图像编辑，而OSX[33]的集成则有助于实现可提示的3D人体运动分析。Grounded SAM在开放词汇基准测试中也表现出色，在结合Grounding DINO-Base和SAM-Huge模型后，于SegInW（野外分割）零样本基准测试中取得48.7的平均AP值。

## 1. 介绍

在开放世界场景中，视觉感知与理解任务对自动驾驶、机器人导航及智能安防监控等应用的发展至关重要。这些应用需要具备强大且多功能的视觉感知模型，以实现对开放世界环境的解读与交互。

目前，解决开放世界视觉感知挑战主要有三种方法。第一种是**统一模型**方法，通过在多个数据集上训练uninext[66]和OFA[59]等模型来支持各类视觉任务。该方法还包括在不同视觉问答数据集上训练大型语言模型以统一任务，例如LLaVA[34]、InstructBLIP[12]、Qwen-VL[3]以及其他MLLMs[60, 40, 80]。然而，这种方法的一个显著局限在于其数据范围有限，尤其在开放集分割等复杂任务中。第二种是**LLM作为控制器**方法，试图在视觉专家与语言模型之间架起桥梁。例如HuggingGPT[55]、Visual ChatGPT[62]和LLaVA-Plus[35]。这些方法利用大型语言模型的语言理解能力来指导各类视觉任务，但该方法高度依赖于大型语言模型的功能与局限性。第三，**集成基础模型**方法旨在通过协作整合针对特定场景设计的专家模型，以完成复杂场景中的开放世界任务。该方法通过结合各类专业模型的优势，提供了灵活性。

尽管通过这些方法论在开放世界任务领域已取得进展，但市场上仍缺乏能够支持复杂基础开放世界任务（如开放集分割）的稳健流程。Grounded SAM从集成基础模型方法论视角出发，创新性地整合了开放集检测模型（如Grounding DINO[38]）与可提示分割模型（如SAM[25]），通过将开放集分割挑战分解为开放集检测与可提示分割两大核心模块，有效攻克该难题。基于此方法，Grounded SAM构建了一个强大而全面的平台，进一步推动了

高效融合不同专家模型以应对更复杂的开放世界任务。

基于Grounded SAM作为基础架构，并利用其强大的开放集分割能力，我们可以轻松整合其他开放世界模型。例如，当与“识别万物”（RAM）[83]结合时，RAM-Grounded-SAM模型无需任何文本输入即可自动识别并分割图像中的物体，从而实现自动图像标注任务。类似的功能通过与BLIP[31]的集成也能实现。此外，当Grounded SAM与Stable Diffusion的图像修复能力结合时（如Grounded-SAM-SD模型所示），可执行高度精确的图像编辑任务。我们将在第3节通过整合其他开放世界模型，对Grounded SAM及其增强功能进行更详细的讨论。

## 2. 相关工作

### 2.1. 任务特异性视觉模型

在计算机视觉领域，已在多种任务中取得重大进展，包括图像识别[47, 31, 18, 83, 17]、通用目标检测[49, 87, 43, 27, 77, 36, 19, 38, 51, 50, 20, 30]、通用图像分割[9, 8, 26, 29, 78, 88, 79, 25, 79, 16, 28]、目标检测与分割[41, 37, 86]、目标跟踪[67, 84]、图像生成[75, 54, 48, 45, 23, 14, 52, 22, 82, 44]、图像编辑[42, 1, 2, 53, 21]，以人为中心的感知与理解[72, 71, 73, 70, 69, 4, 33, 74]，以及以人为中心的运动生成[39, 6, 32, 61, 5]。然而，尽管取得了这些进展，当前模型大多具有任务特异性，通常难以应对更广泛的任务范围。

### 2.2. 统一模型

为解决多种任务，已开发出统一模型。在语言领域，大型语言模型（LLMs）如GPT-3[13]、LaMDA[57]和PaLM[11]是通用统一模型的范例，它们通过自回归和生成方法处理语言任务。与依赖统一结构化标记表示的语言任务不同，视觉任务涵盖多种数据格式，包括像素、空间（如框坐标、关键点）、时间等。近期研究尝试从两个角度开发统一视觉模型以适应这些多样化模态：首先，部分模型致力于将不同视觉模态统一为单一模态，例如Pix2Seq[7]和OFA[59]尝试将空间模态（如框坐标）融合为语言；其次，部分模型寻求兼容不同模态输出的统一模型，uninext[66]即是一个典型范例。

不同端口输出的实例级任务结果存在差异。尽管这些统一视觉模型正在推动通用智能的发展，但现有模型仅能处理有限数量的任务，且其性能常不及任务特定模型。

### 2.3. 带控制器系统的模块化装配

与我们的研究方向不同，Visual ChatGPT[62]和HuggingGPT[55]提出利用大语言模型（LLM）来控制不同AI模型以解决不同任务。相较于这些模型，基础模型组装方法未采用LLM作为控制器，这使得整个流程更加高效灵活。我们证明复杂任务可以解耦，且无需训练即可通过模型组装方式实现逐步视觉推理。

## 3. 地基式地对空导弹试验场

在本章中，我们以基于现实的SAM（情境适应性模型）为基础，展示了将来自不同领域的专家模型融合的方法，以促进更全面视觉任务的完成。

### 3.1. 初步的

本文讨论了Grounded SAM及其他领域专家模型的基本组成部分。

**Segment Anything Model (SAM)** [25] 是一款开放世界分割模型，通过点、框或文本等提示即可“裁剪”任意图像中的目标。该模型基于超过1100万张图像和11亿个掩码进行训练，虽然零样本性能表现优异，但无法通过任意文本输入识别被遮挡目标，通常需要点或框提示才能运行。

**Grounding DINO**[38]是一款开放集目标检测器，能够根据任意自由形式的文本提示检测任意物体。该模型基于超过1000万张图像进行训练，包含检测数据、视觉标注数据及图像-文本配对数据，展现出强大的零样本检测能力。但需注意的是，该模型需要文本作为输入，且仅能检测与对应短语匹配的框。

**OSX**[33]是用于表达性全身网格恢复的最先进模型，旨在从单目图像中联合估计人体三维姿态、手势和面部表情。该模型首先需要检测人体框，裁剪并调整人体框大小，然后进行单人网格恢复。

**BLIP**[31]是一个视觉语言模型，它统一了视觉语言的理解和生成任务。我们在实验中使用了BLIP的图像描述模型。该描述模型能够根据任何图像生成描述。然而，该模型无法执行对象级任务，例如检测或分割对象。

**识别任何事物模型 (RAM)** [83]是一种强大的图像标注模型，能够对输入图像进行高精度的各类常见类别识别。然而，RAM仅能生成标签，无法为识别出的类别生成精确的框选和掩码。

**Stable Diffusion**[52]是一种从训练数据的已学分布中采样图像的图像生成模型。其最广泛的应用是通过文本提示生成图像。我们在实验中使用了其图像修复变体。尽管该模型生成效果出色，但无法执行感知或理解任务。

**ChatGPT与GPT-4**[15, 46]是基于GPT（生成式预训练变换器）架构开发的大型语言模型，该架构用于构建对话式AI代理。它们通过海量文本数据训练，能够生成与人类相似的用户输入响应。该模型可理解对话上下文，并生成恰当回应，其效果常与人类无异。

### 3.2. 离地弹道导弹的开放词汇检测与分割

在图像中定位与用户提供的文本描述相对应的区域并生成掩码，这对实现开放集分割等精细图像理解任务具有极大挑战。这主要源于野外场景分割任务中高质量数据的稀缺性，导致模型在数据匮乏的条件下难以完成精确的开放集分割。相比之下，开放集检测任务更具可操作性，主要基于以下两点原因：首先，检测数据的标注成本相对较低，能够收集更多高质量的标注数据；其次，开放集检测仅需根据文本描述识别图像中的对应物体坐标，无需精确的像素级掩码。类似地，基于框的条件预测（利用框位置的先验知识）比直接根据文本预测区域掩码更为高效。该方法已在先前研究（如OpenSeeD[79]）中得到验证，且通过利用SAM[25]开发的SAM-1B数据集，可有效解决数据稀缺这一重大问题。

因此，受Grounded Pre-training[81, 38]和SAM[25]等成功研究的启发，我们旨在通过整合强大的开放集基础模型来解决野外复杂分割任务。给定输入图像和文本提示后，我们首先利用Grounding DINO通过文本信息作为条件，生成图像中物体或区域的精确框。随后，通过Ground-

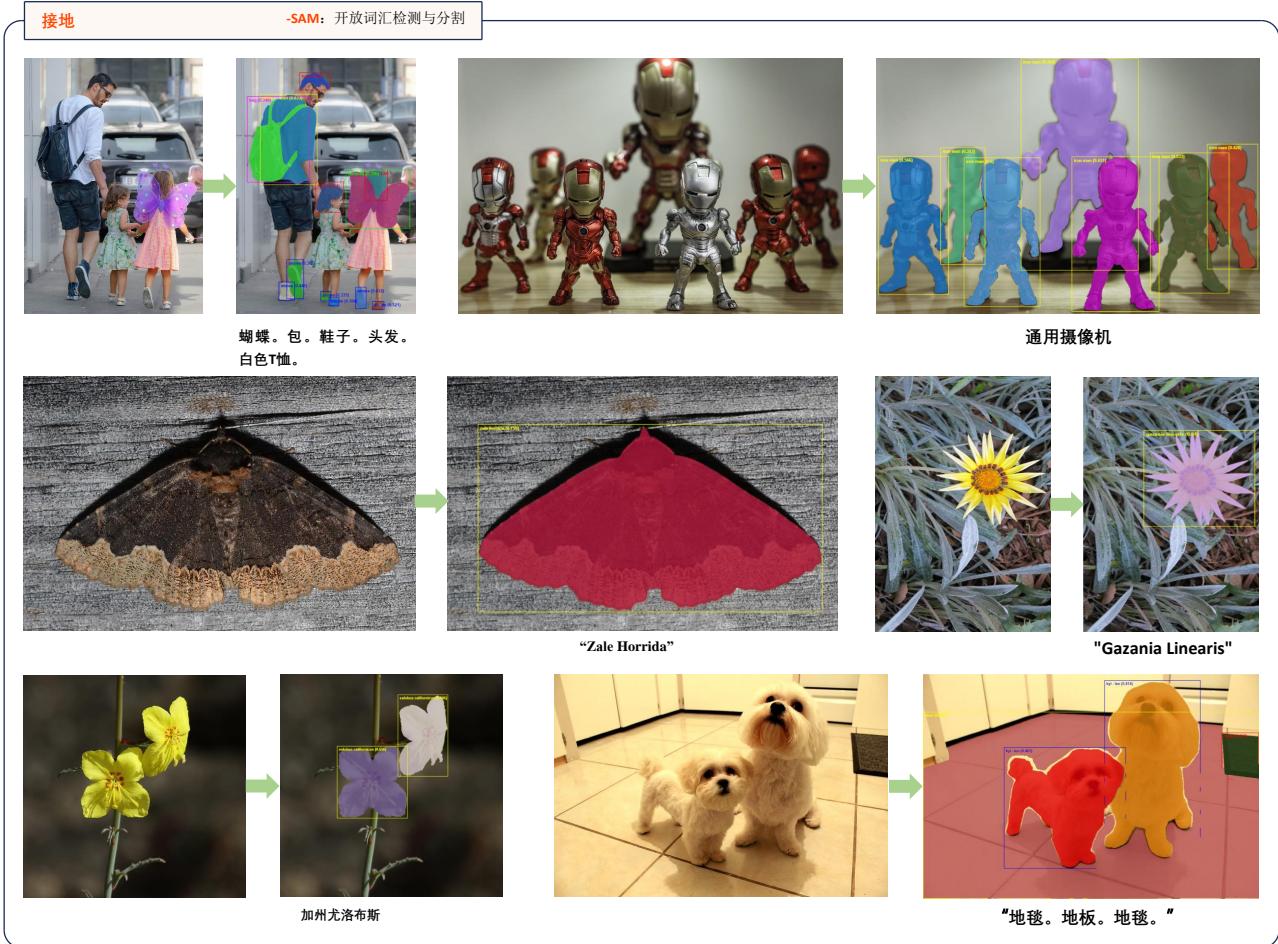


图2: **Grounded-SAM**能根据用户输入有效检测并分割物体。其效果不仅适用于常见场景, 还涵盖长尾物体类别 (如 “Zale Horrida” 和 “Gazania Linearis” 等)。部分演示图像采自V3Det[58]数据集, 我们对其出色成果深表赞赏。

DINO作为SAM的框选提示, 可生成精准的掩码标注。通过整合这两个强大专家模型的性能, 开放集检测与分割任务得以更轻松完成。如图2所示, Grounded SAM在常规场景和长尾场景中, 都能根据用户输入准确实现文本检测与分割。

### 3.3. RAM-GROUND-SAM:自动密集图像标注

自动图像标注系统具有广泛的实际应用价值, 例如提升数据人工标注效率、降低人工标注成本, 或在自动驾驶领域提供实时场景标注与理解以增强行车安全。在Grounded SAM框架下, 该系统充分利用了Grounding DINO的技术优势。用户可灵活输入

通过自动匹配图像中的实体, 我们无需人工标注类别或标题。在此基础上, 我们可以采用图像-标题模型 (如BLIP[31]和Tag2Text[18]) 或图像标注模型 (如RAM[83]), 将它们的输出结果 (标题或标签) 作为Grounded SAM的输入, 为每个实例生成精确的框和掩码。这使得整个图像的自动标注成为可能, 实现了自动化标注系统。如图3所示, RAM-Grounded-SAM在不同场景下都能自动进行类别预测并为输入图像提供密集标注, 这不仅大幅降低了标注成本, 还极大提升了图像标注的灵活性。

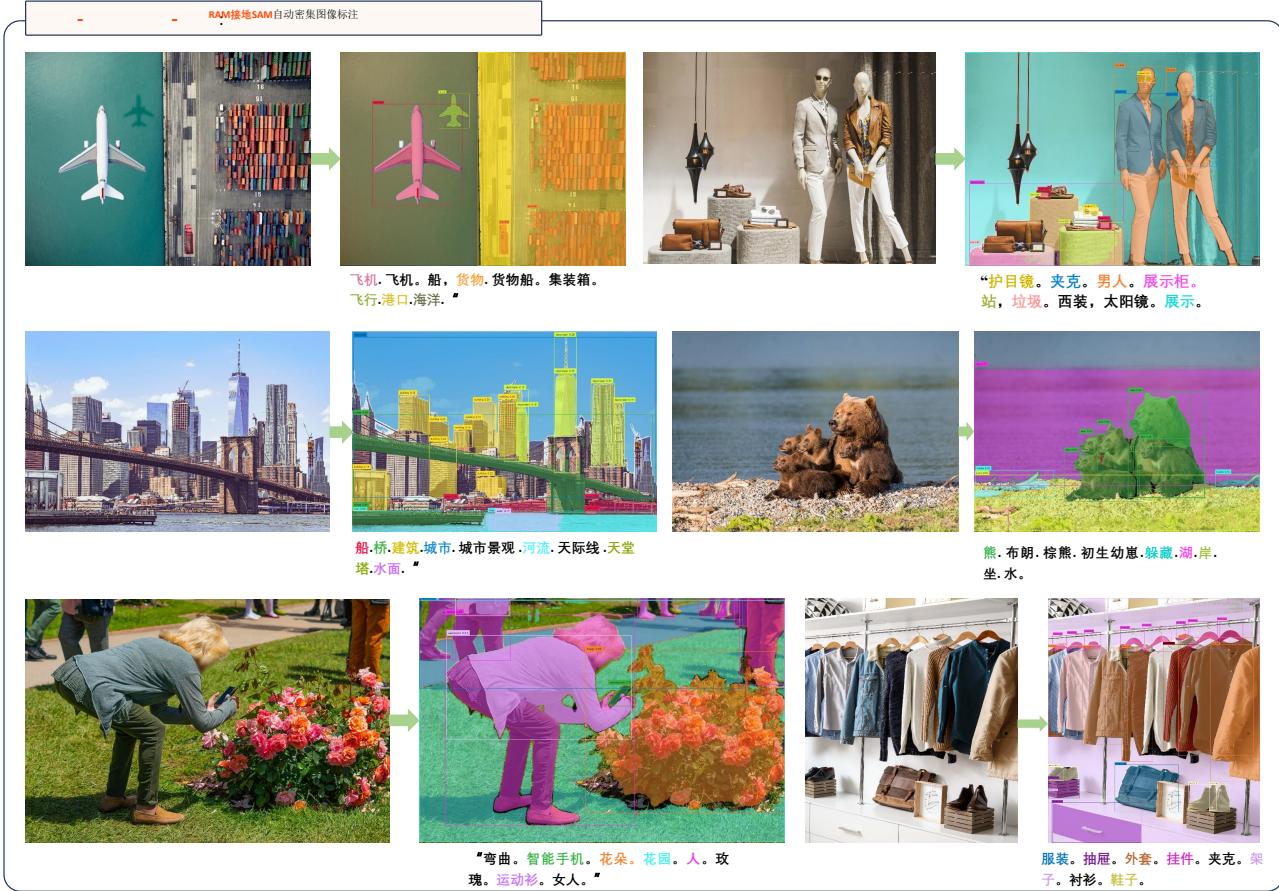


图3：RAM-Grounded-SAM结合了RAM[83]的稳健标记能力与Grounded SAM的开放集检测和分割能力，仅需图像输入即可实现自动密集图像标注（演示图像采自SA-1B[25]数据集）。

### 3.4. 高精度可控图像编辑系统——基于SAM-SD的地地面系统

通过将图像生成模型强大的文本转图像能力与Grounded SAM技术相结合，我们构建了一个完整的框架，能够创建一个强大的数据合成工厂，支持在部件级、实例级和语义级进行精细操作。如图4所示，用户可通过点击或绘制边界框等交互方式，在该流程中获取精确的掩码。此外，用户还能利用文本提示结合Grounding技术，自动定位目标区域。在此基础上，通过集成图像生成模型，我们实现了高度精准可控的图像操作，包括修改图像表示、替换对象、移除对应区域等。在数据稀缺的下游场景中，我们的系统可生成新数据，满足模型训练所需的数据需求。

### 3.5. 地面-空中导弹 OSX :可触发的人体运动分析

先前的表达式全身网格恢复方法首先检测所有（实例无关的）人体框，然后进行单人网格恢复。在许多实际应用中，我们需要指定要检测和分析的目标人物。然而，现有的人体检测器无法区分不同实例（例如，指定分析“穿着粉色衣服的人”），这使得细粒度的人体运动分析变得困难。如图5所示，我们可以通过整合Grounded SAM和OSX [33]模型，实现一种新型的可提示（实例特定）全身人体检测与网格恢复方法，从而构建可提示的人体运动分析系统。具体而言，给定一张图像和一个指向特定人物的提示时，我们首先使用Grounded SAM生成精确的特定人体框，然后利用OSX估计实例特定的人体网格以完成整个过程。

**接地-SAM-SD**: 高度可控的图像编辑



图4: 接地-SAM-SD将接地SAM的开放集能力与图像修复技术相结合

**地面SAM- OSX**: 可提示人体运动分析



图5: **Grounded-SAM- OSX**将GroundedSAM的文本提示功能与OSX [33]的全身网格恢复能力相结合，构建了一个精准的人体运动分析系统。

### 3.6. 基于地面的SAM的进一步扩展

除了上述主要应用外，Grounded SAM还可以通过集成更多模型来进一步扩展其使用范围。例如，在数据标注过程中，Grounded SAM可以与更快的推理SAM模型协作，如FastSAM[85]、MobileSAM[76]。

Light-HQ-SAM [24] 和 EfficientSAM [63]。这种协作可以显著减少整体推理时间并加快标注工作流程。Grounded SAM 还能利用 HQ-SAM [24] 模型生成更高质量的掩码，从而提升标注质量。在图像编辑领域，Grounded SAM 还能与新提出的生成式

表1：SGinW中Grounded-SAM的零样本基准测试结果。最佳和次优结果分别以粗体和下划线标出。\*表示结果由SAM-HQ[24]团队测试。我们非常感谢他们在测试中的协助，并高度赞赏他们的工作。

方法	平均	大象	手部	金属	西瓜	房屋部件	家庭物品	草莓	水果	Nutterfly-Squineel	手	战壕	鸡	飞机制件	脑肿瘤	板	电动剃须刀	瓶子	工具包	垃圾	鲑鱼片	幼犬	片剂	手机	奶牛	姜蒜
X-解码器-T[88]	22.6	65.6	22.4	16.2	5.5	50.6	41.6	66.5	62.1	0.6	28.7	12.0	0.7	10.5	1.1	3.6	1.2	19.0	9.5	19.3	15.0	48.9	15.2	29.9	12.0	7.9
X解码器-L-IN22K	26.6	63.9	20.3	13.5	4.9	50.5	74.4	79.1	58.8	0.0	24.3	3.5	1.3	12.3	0.5	13.4	18.8	43.2	14.6	20.1	12.3	57.3	6.9	<b>43.4</b>	12.3	15.6
X解码器	27.7	68.0	18.5	13.0	6.7	51.7	81.6	76.7	53.1	20.6	30.2	13.6	0.8	13.0	0.3	5.6	4.2	45.9	13.9	27.3	18.2	55.4	8.0	8.9	36.8	19.4
X解码器	32.2	66.0	42.1	13.8	7.0	53.0	67.1	79.2	68.4	75.9	33.0	8.6	2.3	13.1	2.2	20.1	7.5	42.1	9.9	22.3	19.0	59.0	22.5	15.6	44.9	11.6
OpenSeeD-L [79]	36.7	72.9	38.7	52.3	1.8	50.0	82.8	76.4	40.0	<u>92.7</u>	16.9	82.9	1.8	13.0	2.1	4.6	4.7	39.7	15.4	15.3	15.0	<b>74.6</b>	<b>47.4</b>	7.6	40.9	13.6
ODISE-L [64]	38.7	74.9	51.4	37.5	<b>9.3</b>	<b>60.4</b>	79.9	81.3	71.9	41.4	<u>39.8</u>	84.1	2.8	15.8	2.9	0.4	18.3	37.7	15.0	28.6	30.2	<u>65.4</u>	9.1	<b>43.8</b>	41.6	23.0
SAN-CLIP-ViT-L [65]	41.4	67.4	62.9	43.5	<u>9.0</u>	<u>60.1</u>	81.8	77.4	<u>82.2</u>	88.8	<b>46.5</b>	69.2	2.9	13.2	2.6	1.8	11.4	48.8	<b>31.2</b>	<b>41.4</b>	20.0	60.1	35.1	10.4	44.0	23.3
UNINEXT-H [66]	42.1	72.1	57.0	56.3	0.0	54.0	80.7	81.1	<b>84.1</b>	<b>93.7</b>	16.9	75.2	0.0	15.1	2.6	13.4	71.2	46.1	10.1	10.8	<b>44.4</b>	64.6	21.0	6.1	<b>52.7</b>	23.7
接地式总部-地面目标 (B+H) *	<b>49.6</b>	77.5	<b>81.2</b>	<b>65.6</b>	8.5	<u>60.1</u>	<b>85.6</b>	<u>82.3</u>	77.1	74.8	25.0	<u>84.5</u>	<u>7.7</u>	<u>37.6</u>	<b>12.0</b>	<u>20.1</u>	<b>72.1</b>	<b>66.3</b>	21.8	<u>30.0</u>	<u>42.2</u>	50.1	29.7	35.3	47.8	<u>45.6</u>
地基式导弹拦截系统 (B+H)	48.7	<u>77.9</u>	<b>81.2</b>	<u>64.2</u>	8.4	<u>60.1</u>	<u>83.5</u>	<u>82.3</u>	71.3	70.0	24.0	<u>84.5</u>	<b>8.7</b>	37.2	<u>11.9</u>	<b>23.3</b>	<u>71.7</u>	<u>65.4</u>	20.8	<u>30.0</u>	32.9	50.1	29.8	35.4	47.5	<b>45.8</b>
地基式导弹拦截系统	46.0	<b>78.6</b>	<u>75.2</u>	61.5	7.2	35.0	82.5	<b>86.9</b>	70.9	90.7	28.2	<b>84.6</b>	7.2	<b>38.4</b>	10.2	17.4	59.7	43.7	<u>26.9</u>	22.4	27.1	63.2	<u>38.6</u>	3.4	<u>49.4</u>	40.0

诸如Stable-Diffusion-XL[52]等模型可实现更高质量的图像编辑。此外，它还能与LaMa[56]和PaintByExample[68]等模型集成，以实现精确的图像擦除和定制化图像编辑。Grounded SAM还可与DEVA[10]等跟踪模型集成，基于特定文本提示执行目标跟踪。

#### 4. 接地式自适应调制解调器的有效性

为验证Grounded SAM的有效性，我们在包含25个零样本野外分割数据集的野外分割（SGinW）零样本基准上评估其性能。如表1所示，与先前统一的开放集分割模型（如uninext[66]和OpenSeeD[79]）相比，结合Grounding DINO Base和大型模型与SAM-Huge的组合在SGinW零样本设置中实现了显著性能提升。通过引入HQ-SAM[24]（该模型能生成比SAM更高质量的掩码），Grounded-HQ-SAM在SGinW上实现了更进一步的性能改进。

#### 5. 结论与展望

我们提出的基于专家模型的自适应模型（Grounded SAM）及其扩展方案，通过整合多种专家模型来完成不同视觉任务，其优势可归纳如下。首先，模型能力边界可通过整合各类专家模型实现无缝扩展。过去我们仅能用n个模型完成n项任务，如今考虑到所有可能的模型组合，最多可使用n个专家模型完成 $2^n - 1$ 项任务。我们能够将复杂任务拆解为若干子任务，由现有专家模型分别解决。其次，通过将任务分解为多个子任务，模型组装流程的可解释性得到显著提升。

通过观察各步骤的输出结果，我们可以追溯最终结论的推理过程。最终，通过整合多种专家模型，我们能够探索新的研究领域与应用方向，从而可能带来创新成果与技术突破。

**展望：**我们方法论的重要价值在于实现了标注数据与模型训练的闭环。通过整合专家模型，可大幅降低标注成本。此外，在不同阶段引入人工标注者有助于筛选或优化模型预测中的误差，从而提升标注质量。标注数据将被持续用于模型的进一步训练与优化。该方法的另一潜在应用是与大语言模型（LLMs）结合。由于我们构建的模型能处理几乎所有计算机视觉（CV）任务，且支持多种输入输出模态（尤其是语言模态），大语言模型通过语言提示调用我们的API即可高效执行CV任务。最后值得一提的是，该模型还能生成跨模态数据集，尤其与生成模型结合使用时效果更佳。

#### 6. 贡献与致谢

我们谨向研究界多位人士对Grounded SAM项目提供的大力支持表示最诚挚的谢意。下文列出了Grounded SAM项目的主要参与角色。每个角色内的贡献均等，且以随机顺序排列。各角色内部的排序顺序不代表贡献的先后顺序。

##### 导线

**天河人**，联合负责人，接地SAM与接地 SAMSD 流程。

**刘世龙**, 联合负责人, 负责接地SAM流程及在线演示。

**Ailing Zeng**, Grounded-SAM- OSX 管道与演示项目联合负责人。

**金玲**, Grounded-SAM- OSX 流程联合负责人及演示。

**何超**, Grounded-SAM-SD流程与交互式SAM编辑流程联合负责人。

**李坤昌**, BLIP-Grounded-SAM流水线与ChatBot联合负责人。

**陈佳宇**, 联合负责人, 负责Grounded SAM模型演示支持与代码优化。

**黄新宇**, RAM-Grounded-SAM演示支持联合负责人。

**Feng Yan**, 联合负责人, 基于 VISAM 追踪的地面上自适应移动设备 (SAM) 演示。

**陈玉康**, 3D-Box via Segment Anything联合负责人。

### 核心贡献者

**曾昭阳**

**张浩**

**冯丽**

**杨杰**

**李洪阳**

**清江**

**陈溪白屋**

**王振轩**

### 总体技术负责人

**张磊**

### 参考文献

[1] Omri Avrahami 、 Ohad Fried 和 Dani Lischinski 。 Blended Latent Diffusion, 2022年6月。 2

[2] Omri Avrahami 、 Dani Lischinski 和 Ohad Fried 。文本驱动 Natural Images 编辑 Blended Diffusion 。在 2022 IEEE/ CVF Computer Vision 与模式识别会议 (CVPR) , 2022年9月。 2

[3] 金泽柏、帅柏、杨书生、王世杰、谭思南、王鹏、林俊阳、周长洲、周景仁。 Qwen-VL: 一种用于理解、本地化、文本阅读等领域的多功能视觉-语言模型, 2023。 2

[4] 钟刚财、万奇银、艾玲曾、陈伟、孙庆平、王彦军、庞慧恩、梅海一、张明远、张磊等。 Smpler-x: 扩展表达能力

人体姿态与形状估计。在第三十七届神经信息处理系统会议数据集与基准赛道上, 2023。 2

[5] 陈俊明、刘云飞、王建安、曾爱玲、李宇和陈启峰。 Diffsheg: 一种基于扩散的实时语音驱动整体3D表情与手势生成方法。 arXiv 预印本 arXiv:2401.04747, 2024。 2

[6] 陈凌浩、张佳伟、李叶文、庞一仁、夏晓波、刘同亮。 HumanMAC: 用于人体运动预测的掩码运动补全。 2023。 2

[7] 陈婷、索拉布·萨克塞纳、李拉拉、大卫·J·弗利特和杰弗里·欣顿。 Pix2seq: 一种用于目标检测的语言模型框架。 arXiv 预印本 arXiv:2109.10852, 2021。 2

[8] Bowen Cheng、Anwesa Choudhuri、Ishan Misra、Alexander Kirillov、Rohit Girdhar 和 Alexander G. Schwing。 Mask2Former 用于视频实例分割。 2022。 2

[9] Bowen Cheng、Alexander G. Schwing 和 Alexander Kirillov。像素级分类并非语义分割的全部。 2021。 2

[10] 何启成、吴秀英、布莱恩·普莱斯、亚历山大·施温和李俊英。通过解耦视频分割追踪任何物体。在 ICCV , 2023。 7

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: 基于路径的规模语言建模. arXiv 预印本 arXiv:2204.02311, 2022. 2

[12] 戴文亮、李俊楠、李东旭、孟华通、赵俊奇、王伟生、李伯阳、冯帕斯卡尔和何史蒂文。 InstructBLIP: 迈向具有指令调优的通用视觉-语言模型, 2023。 2

[13] 卢西亚诺·弗洛里迪与马西莫·基里亚蒂。 GPT-3: 其本质、范围、局限与影响。《心智与机器》, 30:681-694, 2020。 2

[14] 奥兰·加夫尼、亚当·波利亚克、奥龙·阿舒阿尔、谢莉·谢宁、德维·帕里克和亚尼夫·泰格曼。 Make-A-Scene: 基于场景的人类先验文本生成图像。 2

[15] Roberto Gozalo-Brizuela 和 Eduardo C Garrido-Merchan。 ChatGPT并非全部。大型生成式AI模型的最新进展综述。 arXiv 预印本 arXiv:2301.04655, 2023. 3

[16] 胡杰、黄林燕、任天禾、张胜川、纪荣荣、曹柳娟。你只能分割一次: 迈向实时全景分割, 2023。 2

[17] 黄新宇、黄一杰、张友才、田伟伟、冯睿、张跃杰、谢彦春、李雅倩、张磊。基于多粒度文本监督的开放集图像标注, 2023。 2

[18] 黄新宇、张友才、马金字、田伟伟、冯睿、张跃杰、李雅倩、郭艳东、张磊。 Tag2Text: 通过图像标注引导视觉-语言模型, 2023。 2, 4

[19] 丁佳、于慧媛、何浩迪、吴晓佩、余浩军、林伟宏、孙磊、张超和韩虎。混合匹配的DETRs。 arXiv 预印本 arXiv: 2207.13080, 2022。 2