arXiv:2305.03048v2 [cs.CV] 4 Oct 2023

单样本个性化分割一切模型

Renrui Zhang[1,2], **Zhengkai Jiang**[3]*郭子瑜[2]***Shilin Yan**[2],潘俊廷[1],马显征[2]董豪[4], **Yu Qiao**[2], **Peng Gao**[2],**Hongsheng Li**[1†]

[1]香港中文大学多媒体实验室 [2]上海人工智能实验室[3]中国科学院自动化研究所[4]北京大学计算机科学学院计算机科学前沿研究中心{zhangrenrui, gaopeng, guoziyu, qiaoyu}@pjlab.org.cn, kaikaijiang.jzk@gmail.com hsli@ee.cuhk.edu.hk

摘要

在大规模数据预训练的推动下，Segment Anything Model（SAM）已被证明是一个强大的可提示框架，彻底改变了分割领域。尽管 SAM 具有通用性，但如何在无需人工提示的情况下针对特定视觉概念定制 SAM，仍鲜有研究，例如自动在大量图像中分割出你的宠物狗。本文提出一种无需训练的**PerSAM** 个性化方法，称为**PerSAM**。仅给定单样本数据，即一张带有参考掩码的图像，我们首先为新图像中的目标概念获取正负位置先验。随后，借助目标视觉语义，我们通过两项提出的技术赋予 SAM 个性化对象分割能力：目标引导注意力和目标语义提示。如此一来，我们便能在无需任何训练的情况下，有效地定制通用 SAM 供个人使用。为进一步消除分割尺度上的歧义，我们提出一种高效的单样本微调变体 **PerSAM-F**。在冻结整个 SAM 的同时，我们引入尺度感知微调来聚合多尺度掩码，仅需在 ***10*** 秒内调整 ***2*** 个参数，即可提升性能。为展示我们方法的有效性，我们构建了一个用于评估个性化对象分割的新数据集 PerSeg，并在多个单样本图像和视频分割基准测试上对方法进行测试。此外，我们提出利用 PerSAM 改进 DreamBooth，以实现个性化文生图合成。通过减轻训练集背景的干扰，我们的方法展现出更出色的目标外观生成效果，并能更忠实地遵循输入文本提示。代码已发布于 https://github.com/ZrrSkywalker/Personalize-SAM。



图1：**SegmentAnything 模型的个性化**。我们针对特定视觉概念（例如您的宠物狗）定制 SegmentAnything 模型（SAM）（Kirillov 等人，2023年）。仅凭单样本数据，我们提出两种高效方案：无需训练的 PerSAM，以及微调版 PerSAM-F。

* 共同贡献。 † 通讯作者。



图2：**个性化分割示例**。 我们的 PerSAM（左）能够在任何场景中以良好的性能分割个性化物体，而 PerSAM-F（右）通过尺度感知微调进一步缓解了歧义问题。

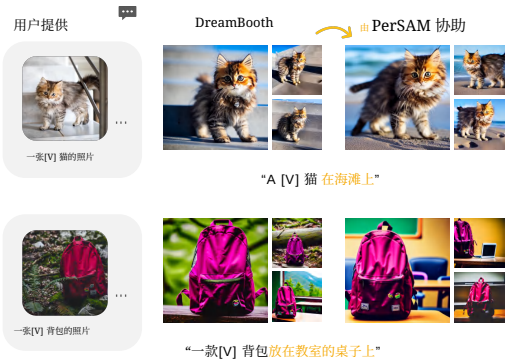

图 3：**改进 DreamBooth（Ruiz 等人，2022年）与 PerSAM**。通过减轻训练过程中背景的干扰，我们的方法有助于实现更高质量的个性化文本到图像生成。

1 引言

视觉（Li 等人，2022年；Zou 等人，2023年；Wang 等人，2022年）、语言（Brown 等人，2020年；Touvron 等人，2023年；Radford 等人，2019年）和多模态（Radford 等人，2021年；Jia 等人，2021年；Li 等人，2023年）领域的基础模型日益普及，这得益于大规模数据集和计算资源的日益充足。它们在零样本场景中展现出卓越的泛化能力，并能够灵活地融入人类反馈进行交互。受此启发，SegmentAnything（Kirillov 等人，2023年）开发了一个精巧的数据引擎，用于收集 1100 万份图像掩码数据，并在此基础上训练了一个名为 SAM 的分割基础模型。它定义了一种新颖的可提示分割框架：输入手工设计的提示，返回预期的掩码，从而能够分割视觉场景中的任意物体。

然而，SAM 本质上失去了分割特定视觉概念的能力。想象一下，你想从一本厚厚的相册中裁剪出心爱的宠物狗，或是从卧室的照片中找到丢失的时钟。使用原版 SAM 会非常耗费人力和时间。对于每张图像，你都必须在复杂的场景中精确找到目标对象，然后用恰当的提示启动 SAM 进行分割。鉴于此，我们提出一个问题：我们能否对 *SAM* 进行个性化设置，使其以简单高效的方式自动分割用户指定的视觉概念？

为此，我们提出 **PerSAM**，一种面向 Segment Anything Model 的免训练个性化方法。如图 1所示，我们的方法仅使用单样本数据（即用户提供的参考图像以及个人概念的粗略掩码），即可高效地定制 SAM。具体而言，我们首先通过特征相似度，为测试图像中的目标对象生成位置置信度图，该图综合考虑每个前景像素的外观。随后，根据置信度分数选取两个点作为正负位置先验，并将其编码为提示令牌，输入 SAM 的解码器进行分割。在解码器中，我们提出两种技术，注入目标对象的视觉语义，以充分释放 SAM 的个性化分割能力：

- **目标引导注意力**。 我们利用位置置信度图引导 SAM 解码器中的每个令牌到图像交叉注意力层。这会明确促使提示令牌主要聚焦于前景目标区域，以进行密集的特征聚合。

- **目标语义提示**。 为了向 SAM 明确提供高层目标语义，我们将原始提示令牌与目标对象的嵌入融合，为低层位置提示补充额外的视觉线索，从而实现个性化分割。

通过上述设计以及级联后处理优化，PerSAM 在各种姿势或场景下对独特主体均表现出良好的个性化分割性能。值得注意的是，我们的方法能够很好地应对以下场景：需要从多个相似物体中分割出一个物体、同时分割同一图像中的多个相同物体，或沿视频追踪不同物体。然而，如图 2所示，仍可能偶尔出现一些失败案例，

当物体包含需要分割的视觉上不同的子部件或层次结构时，例如泰迪熊头顶的帽子，或机器人玩具的头部。这种歧义给 PerSAM 带来了挑战，因为它难以确定输出掩码的适当尺度；对于 SAM 而言，局部部件和整体形状都可以被视为有效掩码。

为缓解这一问题，我们进一步提出了该方法的一种微调变体：**PerSAM-F**。我们冻结整个 SAM，以保留其多样化的预训练知识，仅在***2*** 个参数上进行微调，并在单个 ***A100 GPU*** 上用时 ***10*** 秒。具体而言，我们使 SAM 能够生成具有不同掩码尺度的多个潜在分割结果。为了针对不同物体自适应地选择最佳尺度，我们为每个掩码尺度设置一个可学习的相对权重，并通过加权求和得到最终输出。通过这种高效的尺度感知训练，PerSAM-F 避免了在单样本数据上的过拟合，并展现出更高的分割准确率，如图 2 （右）所示。

此外，我们观察到，该方法还可以帮助 DreamBooth（Ruiz 等人，2022年）更好地微调扩散模型，以实现个性化文生图，如图3所示。给定一些包含特定视觉概念的图像，例如你的宠物猫或背包，DreamBooth 会学习将这些图像转换为词嵌入空间中的标识符 [V]，然而该标识符也可能同时包含背景信息，例如楼梯或森林。这会覆盖新提示中的背景，并干扰目标外观生成。因此，我们提出利用 PerSAM 对训练图像中的目标对象进行分割，仅使用前景区域监督 DreamBooth，从而实现质量更高的文生图合成。

我们将本文的贡献总结如下：

- **个性化对象分割**。我们首先探究如何以最小成本将通用分割模型（SAM）定制到个性化场景中。为此，我们提出了两种高效且有效的方法，并构建了一个新的分割数据集 PerSeg，用于评估个性化对象分割。

- **PerSAM 和 PerSAM-F**。在 PerSAM 中，我们提出了三种无需训练的技术，利用目标对象的高层语义来引导 SAM。在 PerSAM-F 中，我们设计了一种仅需 10 秒、使用 2 个参数的尺度感知微调方法，可有效缓解掩码歧义问题。
- 我们的方法在多种任务上取得了有竞争力的结果，包括 PerSeg 基准测试、单样本部件分割与语义分割，以及视频目标分割。此外，PerSAM 还可以增强 DreamBooth，从而实现效果更好的个性化文生图合成。

2 相关工作

基础模型。凭借强大的泛化能力，预训练基础模型可以适配各种下游场景，并取得出色的性能。在自然语言处理领域，BERT（Devlin 等人，2018；Lu 等人，2019）、GPT 系列（Brown 等人，2020；OpenAI，2023；Radford & Narasimhan，2018；Radford 等人，2019）和 LLaMA（Zhang 等人，2023c）展现了卓越的上下文学习能力，并可通过特定领域的提示迁移到新任务中。同样，在图文对上进行对比学习的 CLIP（Radford 等人，2021）和 ALIGN（Jia 等人，2021），在零样本视觉识别中表现出极高的准确率。Painter（Wang 等人，2022）提出了一种视觉模型，通过统一网络架构和上下文提示来完成各种视觉任务，无需下游微调。CaFo（Zhang 等人，2023d）串联不同的基础模型，并协同利用其预训练知识，以实现稳健的小样本图像分类。SAM（Kirillov 等人，2023）提出了一种用于图像分割的基础模型，该模型通过 10 亿个掩码进行预训练，并执行基于提示的分割。与此同时，一些研究将 SAM 扩展至高质量分割（Ke 等人，2023）、更快的推理速度（Zhao 等人，2023；Zhang 等人，2023a）、通用匹配（Liu 等人，2023）、三维重建（Cen 等人，2023）、物体跟踪（Yang 等人，2023）以及医学图像处理（Ma & Wang，2023；Huang 等人，2023）。从另一个角度出发，我们提出针对特定视觉概念个性化分割基础模型 SAM，仅需一个样本，即可将通用模型适配为专用模型。我们的方法还可辅助文本到图像基础模型的个性化，例如 Stable Diffusion（Rombach 等人，2022）和 Imagen（Saharia 等人，2022），通过从背景干扰中分割出前景目标物体，提升生成质量。

分割领域的大模型。作为计算机视觉中的一项基础任务，分割（Long 等人，2015；Jiang 等人，2022；Zhao 等人，2017；Xu 等人，2021；Jiang 等人，2023；Lin 等人，2022）要求对图像进行像素级理解。研究者探索了各种与分割相关的任务，例如语义分割，即将每个像素归入预定义类别（Badrinarayanan 等人，2017；Chen 等人，2017；Zheng 等人，2021；Cheng 等人，2022；Xie 等人，2021；Song 等人，2020）；实例分割，侧重识别单个物体实例（He 等人，2017；Wang 等人，2020；Tian 等人，2020）；全景分割，同时分配类别标签和实例标识（Kirillov 等人，2019；Li 等人，2019）；以及交互式分割，通过人工干预进行细化（Hao 等人，2021；Chen 等人，2021）。近期，受语言基础模型启发（Zhang 等人，2023c；Brown 等人，2020），多项同期研究提出了用于图像分割的大规模视觉模型。这些模型通过大量掩码数据进行预训练，在多种图像分布上展现出强大的泛化能力。Segment Anything Model（SAM）（Kirillov 等人，2023）利用带有模型参与标注的数据引擎，学习可提示分割框架，并以零样本方式泛化至下游场景。Painter（Wang 等人，2022）和 SegGPT（Wang 等人，2023）提出了稳健的上下文学习范式，只需给定图像－掩码提示即可分割任意图像。SEEM（Zou 等人，2023）进一步提出了一种由多模态参考（如语言和音频）提供提示的通用分割模型，并融合了多样的语义知识。本研究提出一项名为个性化对象分割的新任务，并标注了一个用于评估的新数据集 PerSeg。我们的目标并非开发大型分割模型，而是对其进行个性化，使其能够分割用户提供的、处于任意姿势或场景中的物体。我们提出了两种方法 PerSAM 和 PerSAM-F，能够高效地定制 SAM，以实现个性化分割。

参数高效微调。直接在下游任务上调优整个基础模型，可能会带来高昂的计算成本和巨大的内存开销，从而给资源受限的应用带来挑战。为了解决这一问题，近期研究致力于开发参数高效的方法（Sung 等人，2022；He 等人，2022；Rebuffi 等人，2017；Qin & Eisner，2021），冻结基础模型的权重，并添加小规模模块进行微调。提示调优（Lester 等人，2021；Zhou 等人，2022；Jia 等人，2022；Liu 等人，2021）提出，在冻结模型的同时使用可学习的软提示来执行特定的下游任务。与完整模型调优相比，该方法能够随着模型规模扩大实现更具竞争力的性能，并具备稳健的领域迁移能力。低秩适配（LoRA）（Hu 等人，2021；Cuenca & Paul，2023；Zhang 等人，2023b；Hedegaard 等人，2022）将可训练的秩分解矩阵同时注入每个预训练权重中，从而显著减少下游任务所需学习的参数数量。Adapter（Houlsby 等人，2019；Pfeiffer 等人，2020；Lin 等人，2020；Chen 等人，2022）被设计为插入原始 Transformer 的层与层之间，引入轻量级 MLP 进行特征变换。不同于现有工作，我们采用了一种专为 SAM 精心设计的更高效适配方法，即 PerSAM-F 的尺度感知微调，仅需 2 个参数和 10 秒即可完成。这一方法有效避免了单样本数据上的过拟合问题，并缓解了分割尺度的歧义，同时取得了卓越的性能。

3 方法

在第 3.1 节中，我们首先简要回顾 Segment Anything Model (SAM)（Kirillov 等人，2023年），并介绍个性化物体分割的任务定义。随后，我们分别在第3.2 节和第 3.3 节阐述 PerSAM 和 PerSAM-F 的方法。最后，我们在第 Ruiz 等人，2022年）介绍利用我们的方法辅助 DreamBooth 进行更好的文本到图像生成3.4。

3.1 个性化物体分割

重新审视分割万物。SAM 由三个组件组成，分别是提示编码器、图像编码器和轻量级掩码解码器，分别记作 Enc_P、Enc_I 和 Dec_M。作为一个可提示框架，SAM 的输入为图像 I 和一组提示 P，提示可以是一个点、一个框或一个粗略掩码。具体而言，SAM 首先利用 Enc_I 获取输入图像特征，并采用 Enc_P 将长度为 k 的人为提示编码为提示词元，作为

$$F_I = \text{Enc}_I(I), \quad T_P = \text{Enc}_P(P), \tag{1}$$

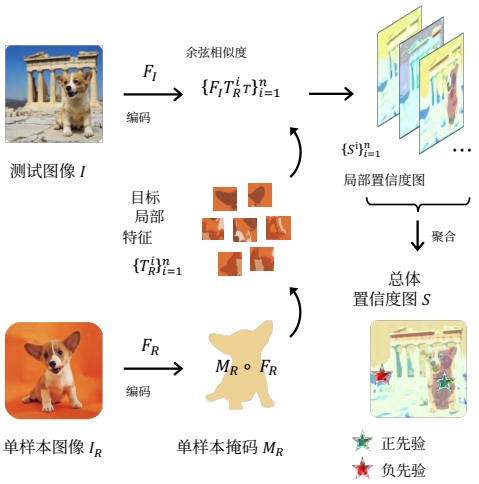

图 4：**正负位置先验**。我们根据所有局部部件的外观，计算目标对象在新测试图像中的位置置信度图。然后，选取该位置先验作为 PerSAM 的点提示。

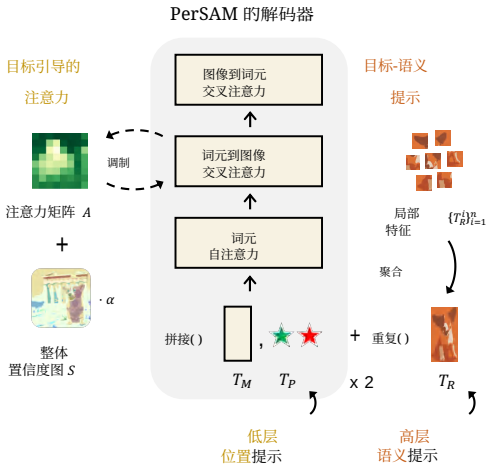

图 5：**目标引导注意力（左）与目标语义提示（右）**。为了向 SAM 注入目标语义，我们明确引导交叉注意力层，并提出使用高层线索进行额外提示。

其中 $F_I \in \mathbb{R}^{h\times w\times c}$ 和 $T_P \in \mathbb{R}^{k\times c}$，其中，$w$ 表示图像特征图的分辨率，c 表示特征维度。之后，将编码图像和提示输入解码器 Dec_M，进行基于注意力的特征交互。SAM 通过将若干可学习掩码标记 T_M 作为前缀，与提示词元 T_P拼接，构造解码器的输入词元。这些掩码标记负责生成掩码输出，其形式化表示为

$$M = \text{Dec}_M\left(F_I,\ \text{Concat}(T_M, T_P)\right), \tag{2}$$

其中，M 表示 SAM 预测的最终分割掩码。

任务定义。虽然 SAM 通过提示已具备对任意物体进行泛化分割的能力，但它无法自动分割特定主体实例。因此，我们为个性化对象分割定义了一项新任务。用户只需提供一张参考图像，以及一个标示目标视觉概念的掩码。该掩码既可以是精确分割结果，也可以是即时绘制的粗略草图。我们的目标是定制 SAM，使其无需额外的人工提示即可分割新图像或视频中的指定物体。为便于评估，我们标注了一个用于个性化分割的新数据集，名为 PerSeg。原始图像收集自主体驱动扩散模型相关研究（Gal 等人，2022年；Ruiz 等人，2022年；Kumari 等人，2022年），其中包含不同姿势或场景下的各类视觉概念。本文针对这一任务提出了两种高效的解决方案，具体如下。

3.2 无需训练的 PerSAM

位置置信度图。基于用户提供的图像 I_R 和掩码 M_R，PerSAM 首先获取一张置信度图，用于指示目标对象在新测试图像中的位置I。如图4所示，我们应用图像编码器提取 I_R 和 I的视觉特征。该编码器可以是 SAM 的冻结骨干网络或其他预训练视觉模型，默认情况下我们采用 SAM 的图像编码器Enc_I。我们将该过程表示为

$$F_I = \text{Enc}_I(I), \quad F_R = \text{Enc}_I(I_R), \tag{3}$$

其中 F_I，$F_R \in \mathbb{R}^{h\times w\times c}$。然后，我们利用参考掩码 $M_R \in \mathbb{R}^{h\times w\times 1}$ 从 F_R中裁剪视觉概念内前景像素的特征，从而得到一组 n 局部特征作为

$$\{T_R^i\}_{i=1}^n = M_R \circ F_R, \tag{4}$$

其中， $T_R^i \in \mathbb{R}^{1\times c}$ 和 $\circ$ 表示逐空间位置相乘。此后，我们通过 T_R^i 与测试图像特征 F_I 之间的余弦相似度，为每个前景像素 i 计算 n 置信度图。

$$\{S^i\}_{i=1}^n = \{F_I T_R^{i\,T}\}_{i=1}^n, \quad \text{where } S^i \in \mathbb{R}^{h\times w}. \tag{5}$$

请注意， F_I 和 T_R^i 均已按像素进行 L2 归一化。每个 S^i 都表示测试图像中物体不同局部部位的分布概率，例如狗的头部、身体或爪子。在此基础上，我们采用平均池化聚合所有 n 局部图，以获得目标对象的整体置信度图：

$$S = \frac{1}{n}\sum_{i=1}^{n} S^i \ \in \mathbb{R}^{h\times w}. \tag{6}$$

通过纳入每个前景像素的置信度， S 能够考虑物体不同部位的视觉外观，并获得相对全面的位置估计。

正负位置先验。为了为 PerSAM 提供测试图像上的位置先验，我们选取 S中置信度最高和最低的两个点，分别记为 P_h 和 P_l。前者表示目标对象最可能的中心位置，而后者则反向指示背景。随后，将它们视为正点提示和负点提示，并输入提示编码器：

$$T_P = \mathrm{Enc}_P(P_h, P_l) \ \in \mathbb{R}^{2\times c}, \tag{7}$$

它们表示 SAM 解码器的提示令牌。这样，SAM 会倾向于分割正点周围的连通区域，同时舍弃图像上的负点。

目标引导注意力。尽管已经获得正负点提示，我们仍进一步提出一种更明确的语义引导方式，作用于 SAM 解码器中的交叉注意力运算，使特征聚合集中在前景目标区域内。如图5所示，总体置信度图 S 在公式6 中能够清晰地指示测试图像中目标视觉概念的大致区域（颜色越暖表示得分越高）。基于这一特性，我们利用 S引导解码器中每个令牌到图像交叉注意力层的注意力图。具体而言，我们将经过 softmax 函数处理后的每个注意力图记为 $A \in \mathbb{R}^{h\times w}$，然后通过以下方式调制其注意力分布：

$$A^g = \mathrm{softmax}\Big(A + \alpha \cdot \mathrm{softmax}(S)\Big), \tag{8}$$

其中， α 表示一个平衡因子。借助注意力偏置，掩码和提示令牌会被促使捕获与目标主体相关的更多视觉语义，而非无关紧要的背景区域。这有助于在注意力机制中实现更有效的特征聚合，并以无需训练的方式提高 PerSAM 最终的分割精度。

目标语义提示。 原版 SAM 仅接收携带低层位置信息的提示，例如点或框的坐标。为了向 SAM 的解码器提供更多高层线索，我们提出利用目标概念的视觉特征作为额外的高层语义提示。首先，我们通过对不同局部特征进行平均池化，获取参考图像中物体的全局嵌入 T_R，具体如下

$$T_R = \frac{1}{n}\sum_{i=1}^{n} T_R^i \ \in \mathbb{R}^{1\times c}. \tag{9}$$

然后，我们对测试图像的所有输入词元逐元素加上 T_R，再将它们送入解码器块，如图2所示，具体如下5 所示

$$T^g = \mathrm{Repeat}(T_R) + \mathrm{Concat}(T_M, T_P), \tag{10}$$

其中， T^g表示由目标语义引导、用于解码器 Dec_M的输入词元，而重复操作会复制目标视觉嵌入。借助简单的词元融入，PerSAM 不仅受到低级位置点的提示，还受到高级目标视觉线索的提示。

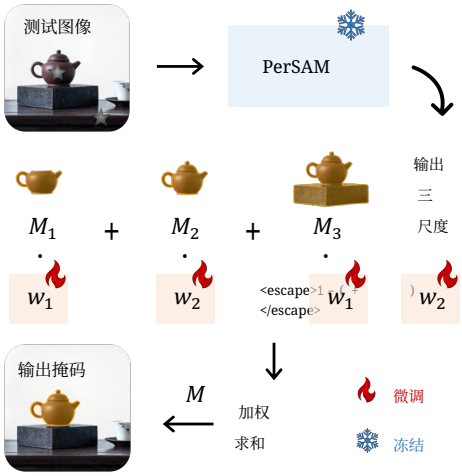

图 6：**PerSAM-F 中的尺度感知微调**。为消除尺度歧义，PerSAM-F采用两个可学习权重，自适应地聚合三个尺度的掩码。

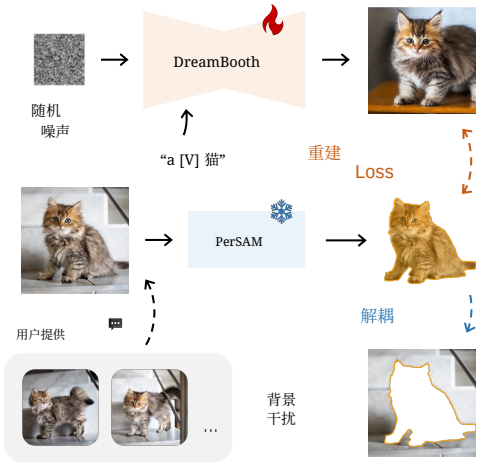

图 7：**PerSAM 辅助的 DreamBooth**。我们利用 PerSAM 将目标对象与背景解耦，以改进 DreamBooth 的生成效果。

级联后处理优化。 通过上述技术，我们从 SAM 的解码器获得测试图像的初始分割掩码，但其中可能包含粗糙边缘和孤立的背景噪声。为进一步细化，我们将掩码迭代地反馈给解码器 Dec_M，进行两步后处理。第一步，我们将当前预测的掩码与先前的正负点提示一同用于提示解码器。第二步，我们获取包围第一步所得掩码的边界框，并额外用此框提示解码器，以实现更准确的物体定位。由于我们只迭代运行轻量级解码器，而不运行大型图像编码器，因此后处理效率很高，额外耗时仅为 2%。

3.3 PerSAM-F 的微调

分割尺度的歧义。 无需训练的 PerSAM 能够以令人满意的分割精度处理大多数情况。然而，一些目标对象具有层级结构，因此会产生掩码尺度歧义。如图6所示，平台上的茶壶由两部分组成：壶盖和壶身。如果正点提示（以绿色五角星表示）位于壶身上，而负提示（以红色五角星表示）未能排除颜色相近的平台，PerSAM 就会受到误导，导致分割错误。SAM 中也讨论了这一问题（Kirillov 等人，2023年），其中提出了一种替代方案，可同时生成三个尺度的多个掩码，分别对应物体的整体、部分和子部分。随后，用户需要从三个掩码中手动选择一个。该方法虽然有效，却需要额外的人力。相比之下，我们的个性化任务旨在定制 SAM，以实现无需人工提示的自动物体分割。这促使我们通过参数高效微调，进一步开发具备尺度感知能力的 PerSAM 版本。

尺度感知微调。 为了以合适的尺度进行自适应分割，我们提出了一种微调变体 PerSAM-F。与只生成一个掩码的无需训练模型不同，PerSAM-F 首先按照 PerSAM 获取位置先验，再参考 SAM 的原始方案输出三个尺度的掩码，分别记为 M_1、M_2 和 M_3。在此基础上，我们采用两个可学习的掩码权重 w_1，w_2，并通过加权求和计算最终的掩码输出，具体如下

$$M = w_1 \cdot M_1 + w_2 \cdot M_2 + (1 - w_1 - w_2) \cdot M_3, \tag{11}$$

其中 w_1，w_2 均初始化为 $1/3$。为学习最优权重，我们在参考图像上进行单样本微调，并将给定掩码视为真实标注。请注意，我们冻结整个 SAM 模型以保留其预训练知识，仅在单张 A100 GPU 上用 ***10*** 秒 微调***2*** 个参数 （w_1，w_2）。通过这种方式，我们的 PerSAM-F 能够高效学习物体的尺度感知语义，并针对不同概念自适应地输出最佳分割尺度，从而提升 PerSAM 的泛化能力。

表 1： **PerSeg 数据集上的个性化对象分割**。我们比较了不同方法的整体平均交并比、边界交并比和可学习参数（Bar 等人，2022年；Wang 等人，2022年；2023年；Zou 等人，2023年），以及 PerSeg 中 10 个物体的平均交并比。“*”表示与我们的工作同期开展的研究。

方法	mIoU	bIoU	参数	Can	Barn	时钟	Cat	反向-pack	泰迪熊 Bear	Duck Toy	Thin Bird	Red 卡通	机器人 Toy
Painter	56.4	42.0	354M	19.1	3.2	42.9	94.1	88.1	93.0	33.3	20.9	98.2	65.0
VP	65.9	25.5	383M	61.2	58.6	59.2	76.6	66.7	79.8	89.9	67.4	81.0	72.4
SEEM*	87.1	55.7	341M	65.4	82.5	72.4	91.1	94.1	95.2	98.0	71.3	97.0	95.8
SegGPT*	94.3	76.5	354M	96.6	63.8	92.6	94.1	94.4	93.7	97.2	92.6	97.3	96.2
PerSAM	89.3	71.7	0	96.2	38.9	96.2	90.70	95.39	94.6	97.3	93.7	97.0	60.6
PerSAM-F	95.3	77.9	2	96.7	97.5	96.1	92.3	95.5	95.2	97.3	94.0	97.1	96.7

表 2： **DAVIS 2017** 上的视频物体分割（Pont- Tuset 等人，2017）。我们使用灰色表示涉及域内训练的方法。

方法	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
AGSS	67.4	64.9	69.9
AFB-URR	74.6	73.0	76.1
Painter	34.6	28.5	40.8
SEEM	58.9	55.0	62.8
SegGPT	75.6	72.5	78.6
PerSAM	66.9	71.3	75.1
PerSAM-F	76.1	74.9	79.7

表 3：**FSS-1000 上的单样本语义和部件分割** (Li 等人，2020年）、LVIS-92^i (Gupta 等人，2019年）、PASCAL-Part（Morabia 等人，2020年），以及 PACO-Part（Ramanathan 等人，2023年）。我们报告平均交并比，并用灰色标示涉及域内训练的方法。

方法	单样本语义分割		单样本部件分割	
	FSS-1000	LVIS-92^i	PASCAL-Part	PACO-Part
HSNet	86.5	17.4	32.4	22.6
VAT	90.3	18.5	33.6	23.5
Painter	61.7	10.5	30.4	14.1
SegGPT	85.6	18.6	-	-
PerSAM	81.6	15.6	32.5	22.5
PerSAM-F	86.3	18.4	32.9	22.7

3.4 PerSAM 辅助的 DreamBooth

为了实现个性化文生图合成，DreamBooth（Ruiz 等人，2022年）通过给定的 3∼5 特定物体的照片（例如宠物猫）对预训练扩散模型进行微调。该模型学会根据文本提示生成所指代的猫：“一只 [V] 猫”，并在重建图像的整体上计算损失。这会将训练图像中冗余的背景信息注入标识符 [V]。因此，如图7所示，我们提出了一种策略，以减轻背景对 DreamBooth 的干扰。给定任意一张少样本图像的物体掩码，我们利用 PerSAM 分割出所有前景目标，并对属于背景区域的像素舍弃梯度反向传播。随后，仅微调 Stable Diffusion，使其记住目标对象的视觉外观。在没有背景监督的情况下，我们的 PerSAM 辅助 DreamBooth 不仅能够合成视觉对应更好的目标对象，还能根据输入文本提示增加新背景的多样性。

4 实验

我们首先在第 4.1 节中评估我们的方法在 PerSeg 上进行个性化分割的效果，并在第 4.2 节中评估其在多个现有单样本分割基准上的表现。随后，我们在第 4.3 节中展示 PerSAM 辅助 DreamBooth 的有效性。最后，我们在第 4.4 节中开展多项消融研究，以探究我们针对 PerSeg 的设计。

4.1 个性化评估

PerSeg 数据集。为了测试个性化能力，我们构建了一个新的分割数据集，称为 PerSeg。原始图像收集自主体驱动扩散研究的训练数据（Ruiz 等人，2022年；Gal 等人，2022年；Kumari 等人，2022年）。PerSeg 总共包含 40 个不同类别的物体，包括日用品、动物和建筑物。对于不同姿势或场景中的每个物体，都有 5∼7 张图像及其掩码，其中我们固定其中一组图像-掩码对作为用户提供的单样本数据。评估采用平均交并比和边界交并比（Cheng 等人，2021年）。有关实现细节及扩展后的 PerSeg 数据规模，请参阅附录。



图 8：**PerSAM-F 改进效果可视化**。我们的尺度感知微调能够很好地缓解 PerSAM 的尺度歧义。

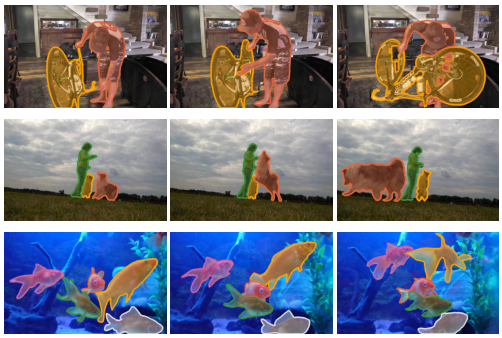

图 9：**视频物体分割可视化**。我们的方法在视频中分割多个物体方面表现良好。

性能。 在表1中，我们发现微调后的 PerSAM-F 取得了最佳结果，使 PerSAM 的整体平均交并比和边界交并比分别提升了 +2.7% 和 +5.9%。图8展示了更多 PerSAM-F 改进效果的可视化结果。视觉提示（VP）（Baretal.，2022年）、Painter（Wang 等人，2022年）、SEEM（Zou 等人，2023年）和 SegGPT（Wang 等人，2023年）都是上下文学习器，也能根据给定的单样本提示数据分割物体。如图所示，无需训练的 PerSAM 已经能够以不同幅度取得优于 Painter、VP 和 SEEM 的表现。通过高效的双参数微调，我们的 PerSAM-F 进一步超越了强大的 SegGPT，整体平均交并比和边界交并比分别高出+2.4% 和 +4.1%。与它们旨在开发通用分割模型的动机不同，我们的方法专为个性化对象分割而设计，在时间和计算资源方面都展现出高得多的效率。

4.2 现有分割基准测试

视频目标分割。给定第一帧图像和物体掩码，我们的 PerSAM 和 PerSAM-F 在 DAVIS 2017（Pont-Tuset 等人，2017）验证集上实现了具有竞争力的物体分割和跟踪性能。如表 2 所示，与未经视频训练的方法相比，无需训练的 PerSAM 大幅超越 Painter，提升了 +32.3% $\mathcal{J}\&\mathcal{F}$ 分数，而我们的 PerSAM-F 的性能可比 SegGPT 高出 +0.5%。值得注意的是，我们的一次性微调方法能够胜过使用大量视频数据进行完整训练的方法（Lin 等人，2019年；Liang 等人，2020年）。这些结果充分展现了我们对时序视频数据和复杂场景的强大泛化能力，这些场景中包含多个相似或被遮挡的物体，如图 9 所示。

单样本语义与部件分割。 在桌子3中，我们分别在四个数据集上评估了单样本图像分割方法：FSS-1000（Li 等人，2020年）、LVIS-92^i（Gupta 等人，2019年）、PASCAL-Part（Morabia 等人，2020年）和 PACO-Part（Ramanathan 等人，2023年），其中我们遵循 Matcher（Liu 等人，2023年）进行数据预处理和评估。如图所示，我们的 PerSAM-F 始终优于 Painter，且表现与 SegGPT 相当。对于采用域内训练的模型（Min 等人，2021年；Hong 等人，2022年），我们的方法得分高于 HSNet。这些实验充分证明，我们提出的方法不仅适用于物体级分割，也适用于 SAM 的类别级和部件级个性化。

4.3 PerSAM 辅助 DreamBooth

我们遵循 DreamBooth（Ruiz 等人，2022年）中的所有超参数，对预训练的 Stable Diffusion（Rombach 等人，2022年）进行微调，以实现个性化图像合成。除了图 3，我们还在图 10中展示了更多 PerSAM 辅助 DreamBooth 的示例。对于躺在灰色沙发上的狗，DreamBooth 生成的“丛林”和“雪”场景仍然是带有绿色和白色装饰的沙发。在 PerSAM-F 的辅助下，新生成的背景与沙发完全解耦，并且与文本提示相符。对于山前的谷仓，我们的方法也减轻了背景干扰，从而正确生成“森林”和“蓝天”。



图10： **PerSAM引导的DreamBooth可视化结果**。 改进后的DreamBooth（Ruiz等人，2022年）能够更好地保留多样性，从而在新图像中合成各种场景。

表4：**我们所提出方法的主要组件消融实验**。

变体	mIoU	Gain
正先验	69.1	-
+ 负先验	72.5	+3.4
+ 后处理细化	83.9	+11.4
+ 引导注意力	85.8	+1.9
+ 语义提示	89.3	+3.5
+ 尺度调优	95.3	+6.0

表5： **不同方法的消融实验微调方法**。

方法	参数	mIoU
PerSAM	0	89.32
提示调优	12K	76.5
Adapter	196K	78.3
LoRA	293K	90.0
3 掩码权重	3	92.9
PerSAM-F	2	95.3

桌子 6： **使用框图像作为参考的消融实验**。

方法	Mask	Box
Painter	56.4	42.0
VP	65.9	38.1
SEEM	87.1	64.9
SegGPT	94.3	36.0
PerSAM	89.3	88.1
PerSAM-F	95.3	94.9

4.4 消融研究

主要组件。 在表 4中，我们从仅采用正位置先验的基线开始，研究各个组件的作用。随后，我们加入负点提示和级联后处理优化，分别提升 +3.6% 和 +11.4% 的平均交并比。在此基础上，我们将高层目标语义引入 SAM 的解码器，用于注意力引导和语义提示。由此带来的 +1.9% 和 +3.5% 提升充分体现了它们的重要性。最后，通过高效的尺度感知微调，PerSAM-F 将分数提升了 +6.0%，展现出卓越的准确性。

不同的微调方法。 在表格 5中，我们针对 PerSAM-F 尝试了其他参数高效微调（PEFT）方法，即提示调优（Liu 等人，2021年）、Adapter（Houlsby 等人，2019年）和 LoRA（Hu 等人，2021年）。我们冻结整个 SAM，仅调整注入 PerSAM 解码器中每个 Transformer 块的 PEFT 模块。如结果所示，提示调优和 Adapter 会对单样本数据产生过拟合，并严重降低准确率。相比之下，我们的尺度感知微调能够在调整最少可学习参数的同时，最大程度地提升 PerSAM 的性能。

使用框图像作为参考。 要求将准确的掩码作为单样本数据，对某些用户来说可能过于严格。在表 6中，我们放宽了输入限制，只需提供指定预期物体的边界框。对于我们的方法，可以将框视为提示，并利用现成的 SAM 生成单样本掩码。因此，框参考仅导致 PerSAM 和 PerSAM-F 的性能略有下降，却严重影响了其他方法。

5 结论

本文提出仅使用单样本数据，为特定视觉概念定制 Segment Anything Model（SAM）。首先，我们提出 PerSAM，通过无需训练的技术将高层目标语义注入 SAM。在此基础上，我们提出尺度感知微调变体 PerSAM-F。仅使用 2 个可学习参数，PerSAM-F 就能有效缓解掩码尺度的歧义，并在各种基准测试中取得领先性能。此外，我们还验证了该方法能够协助 DreamBooth 微调出更优的文生图扩散模型。我们希望本研究能够将 SAM 的适用性扩展到更广泛的场景。

参考文献

Vijay Badrinarayanan、Alex Kendall 和 Roberto Cipolla。SegNet：用于图像分割的深度卷积编码器-解码器架构。*IEEE*模式分析与机器智能汇刊，39(12)：2481–2495，2017。

Amir Bar、Yossi Gandelsman、Trevor Darrell、Amir Globerson 和 Alexei Efros。通过图像修复实现视觉提示。神经信息处理系统进展，35：25005–25017，2022年。

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器。神经信息处理系统进展，33:1877–1901，2020年。

Jiazhong Cen、Zanwei Zhou、Jiemin Fang、Wei Shen、Lingxi Xie、Xiaopeng Zhang 和 Qi Tian。使用神经辐射场在三维中分割万物。*arXiv*预印本 *arXiv:2304.12308*，2023年。

Liang-Chieh Chen、George Papandreou、Iasonas Kokkinos、Kevin Murphy 和 Alan L Yuille。Deeplab：使用深度卷积网络、空洞卷积和全连接条件随机场进行语义图像分割。*IEEE*模式分析与机器智能汇刊，40(4):834–848，2017。

Xi Chen、Zhiyan Zhao、Feiwu Yu、YileiZhang 和 ManniDuan。用于交互式分割的条件扩散。在*IEEE*国际计算机视觉会议论文集，第 7345–7354 页，2021年。

Zhe Chen、Yuchen Duan、Wenhai Wang、Junjun He、Tong Lu、Jifeng Dai 和 Yu Qiao。用于密集预测的视觉变换器适配器。*arXiv*预印本 *arXiv:2205.08534*，2022年。

Bowen Cheng、Ross Girshick、Piotr Doll´ar、Alexander C Berg 和 Alexander Kirillov。边界交并比：改进以对象为中心的图像分割评估。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第 15334–15342 页，2021年。

Bowen Cheng、Ishan Misra、Alexander G Schwing、Alexander Kirillov 和 Rohit Girdhar。用于通用图像分割的掩码注意力掩码变换器。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第 1290–1299 页，2022年。

Pedro Cuenca 和 Sayak Paul。使用 LoRA 高效微调 Stable Diffusion。https://huggingface.co/blog/lora，2023年1月。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018 年。

Rinon Gal、Yuval Alaluf、Yuval Atzmon、Or Patashnik、Amit H Bermano、Gal Chechik 和 Daniel Cohen-Or。图像胜过千言：使用文本反演个性化文本到图像生成。*arXiv* 预印本 *arXiv:2208.01618*，2022年。

Agrim Gupta、Piotr Dollar 和 Ross Girshick。Lvis：一个用于大词汇量实例分割的数据集。载于*IEEE*计算机视觉与模式识别会议论文集，第 5356–5364 页，2019年。

Yuying Hao、Yi Liu、Zewu Wu、Lin Han、Yizhou Chen、Guowei Chen、Lutao Chu、Shiyu Tang、Zhiliang Yu、Zeyu Chen 等。Edgeflow：通过边缘引导流实现实用的交互式分割。载于*IEEE*国际计算机视觉会议论文集，第1551–1560页，2021年。

Junxian He、Chunting Zhou、Xuezhe Ma、Taylor Berg-Kirkpatrick 和 Graham Neubig。迈向参数高效迁移学习的统一视角。载于 国际学习表征会议，2022年。URL：https://openreview.net/forum?id=0RDcd5Axok。

Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Girshick。Mask r-cnn。载于*IEEE*国际计算机视觉会议论文集，第 2961–2969 页，2017。Lukas Hedegaard、Aman Alok、Juby Jose 和 Alexandros Iosifidis。结构化剪枝适配器。*arXiv*预印本 *arXiv:2211.10155*，2022。

Sunghwan Hong、Seokju Cho、Jisu Nam、Stephen Lin 和 Seungryong Kim。使用4D卷积Swin Transformer进行小样本分割的代价聚合。载于欧洲计算机视觉会议，第 108–126 页。Springer，2022。Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin De Laroussilhe、Andrea Gesmundo、Mona Attariyan 和 Sylvain Gelly。面向自然语言处理的参数高效迁移学习。载于国际机器学习会议，第 2790–2799 页。PMLR，2019年。Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。*arXiv*预印本 *arXiv:2106.09685*，2021年。

Yuhao Huang、Xin Yang、Lian Liu、Han Zhou、Ao Chang、Xinrui Zhou、Rusi Chen、Junxuan Yu、Jiongquan Chen、Chaoyu Chen 等。面向医学图像的 Segment Anything Model？*arXiv* 预印本 *arXiv:2304.14660*，2023年。

Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉与视觉语言表征学习。在国际机器学习会议上，第 4904–4916 页。PMLR，2021年。

Menglin Jia、Luming Tang、Bor-Chun Chen、Claire Cardie、Serge Belongie、Bharath Hariharan 和 Ser-Nam Lim。视觉提示调优。在欧洲计算机视觉会议上，第 709–727 页。Springer，2022年。

Zhengkai Jiang、Yuxi Li、Ceyuan Yang、Peng Gao、Yabiao Wang、Ying Tai 和 Chengjie Wang。面向域自适应语义分割的原型对比适配。在欧洲计算机视觉会议，第 36–54 页。Springer，2022年。

Zhengkai Jiang、Zhangxuan Gu、Jinlong Peng、Hang Zhou、Liang Liu、Yabiao Wang、Ying Tai、Chengjie Wang 和 Liqing Zhang。STC：用于视频实例分割的时空对比学习。在欧洲计算机视觉会议研讨会，第 539–556 页。Springer，2023年。

Lei Ke、Mingqiao Ye、Martin Danelljan、Yifan Liu、Yu-WingTai、Chi-Keung Tang 和 Fisher Yu。高质量地分割万物。*arXiv* 预印本 *arXiv:2306.01567*，2023年。

Alexander Kirillov、Kaiming He、Ross Girshick、Carsten Rother 和 Piotr Doll´ar。全景分割。在*IEEE*计算机视觉与模式识别会议论文集，第 9404–9413 页，2019年。

Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander CBerg、Wan-Yen Lo 等。分割万物。*arXiv* 预印本 *arXiv:2304.02643*，2023年。

Nupur Kumari、Bingliang Zhang、Richard Zhang、Eli Shechtman 和 Jun-Yan Zhu。文本到图像扩散模型的多概念定制。*arXiv* 预印本 *arXiv:2212.04488*，2022年。

Brian Lester、Rami Al-Rfou 和 Noah Constant。规模对参数高效提示调优的作用。*arXiv* 预印本 *arXiv:2104.08691*，2021年。

Hao Li、Jinguo Zhu、Xiaohu Jiang、Xizhou Zhu、Hongsheng Li、Chun Yuan、Xiaohua Wang、Yu Qiao、Xiaogang Wang、Wenhai Wang 等。Uni-perceiver v2：面向大规模视觉和视觉-语言任务的通用模型。*arXiv* 预印本 *arXiv:2211.09808*，2022年。

Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。Blip-2：利用冻结的图像编码器和大语言模型引导语言-图像预训练。*arXiv* 预印本 *arXiv:2301.12597*，2023年。

Xiang Li、Tianhan Wei、Yau Pun Chen、Yu-Wing Tai 和 Chi-Keung Tang。《FSS-1000：用于小样本分割的 1000 类数据集》。载于*IEEE*计算机视觉与模式识别会议论文集，第 2869–2878 页，2020年。

Yanwei Li、Xinze Chen、Zheng Zhu、Lingxi Xie、Guan Huang、Dalong Du 和 Xingang Wang。《用于全景分割的注意力引导统一网络》。载于*IEEE*计算机视觉与模式识别会议论文集，第 7026–7035 页，2019 年。Yongqing Liang 、 Xin Li 、 Navid Jafari 和 Jim Chen。《利用自适应特征库和不确定区域细化进行视频目标分割》。神经信息处理系统进展，33：3430–3441，2020年。Huaijia Lin、Xiaojuan Qi 和 Jiaya Jia。《AGSS-VOS：注意力引导的单次视频目标分割》。载于*IEEE*国际计算机视觉会议论文集，第 3949–3957 页，2019年。Zhaojiang Lin、Andrea Madotto 和 Pascale Fung。《通过参数高效迁移学习探索多功能生成式语言模型》。*arXiv*预印本 *arXiv:2004.03829*，2020年。Ziyi Lin、Shijie Geng、Renrui Zhang、Peng Gao、Gerard de Melo、Xiaogang Wang、Jifeng Dai、Yu Qiao 和 Hongsheng Li。《冻结的 CLIP 模型是高效的视频学习器》。载于欧洲计算机视觉会议，第 388–404 页。Springer，2022年。

Xiao Liu、Kaixuan Ji、Yicheng Fu、Weng Lam Tam、Zhengxiao Du、Zhilin Yang 和 Jie Tang。P-tuning v2：提示调优在各种规模和任务上都能普遍达到与微调相当的效果。*arXiv* 预印本 *arXiv:2110.07602*，2021年。

Yang Liu、Muzhi Zhu、Hengtao Li、Hao Chen、Xinlong Wang 和 Chunhua Shen。Matcher：利用通用特征匹配，通过单样本分割万物。*arXiv* 预印本 *arXiv:2305.13310*，2023年。

Jonathan Long、Evan Shelhamer 和 Trevor Darrell。用于语义分割的全卷积网络。发表于*IEEE*计算机视觉与模式识别会议论文集，第3431–3440页，2015年。

Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。Vilbert：面向视觉与语言任务的任务无关视觉语言表征预训练。发表于神经信息处理系统进展（*NeurIPS*），第13–23页，2019年。

Jun Ma 和 Bo Wang。医学图像中的分割万物。*arXiv* 预印本 *arXiv:2304.12306*，2023年。

Juhong Min、Dahyun Kang 和 Minsu Cho。用于小样本分割的超相关性压缩。载于*IEEE/CVF*计算机视觉国际会议论文集，第 6941–6952 页，2021年。

Keval Morabia、Jatin Arora 和 Tara Vijaykumar。基于注意力的物体及其语义部件联合检测。*arXiv* 预印本 *arXiv:2007.02419*，2020年。

OpenAI。GPT-4 技术报告。*ArXiv*，abs/2303.08774，2023年。

Jonas Pfeiffer、Aishwarya Kamath、Andreas R¨uckl´e、Kyunghyun Cho 和 Iryna Gurevych。Adapter 融合：用于迁移学习的非破坏性任务组合。*arXiv* 预印本 *arXiv:2005.00247*，2020年。

Jordi Pont-Tuset、Federico Perazzi、Sergi Caelles、PabloArbel´aez、Alex Sorkine-Hornung 和 Luc Van Gool。2017 年 DAVIS 视频目标分割挑战赛。*arXiv* 预印本 *arXiv:1704.00675*，2017 年。

Guanghui Qin 和 Jason Eisner。《学习如何提问：使用软提示混合向语言模型发起查询》。*arXiv* 预印本 *arXiv:2104.06599*，2021年。

Alec Radford 和 Karthik Narasimhan。通过生成式预训练改进语言理解。2018 年。

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等人。《语言模型是无监督的多任务学习者》。*OpenAI* 博客，1(8):9，2019年。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等人。《可迁移视觉模型的自然语言监督学习》。发表于国际机器学习会议，第 8748–8763 页。PMLR，2021年。

Vignesh Ramanathan、Anmol Kalia、Vladan Petrovic、Yi Wen、Baixue Zheng、Baishan Guo、Rui Wang、Aaron Marquez、Rama Kovvuri、Abhishek Kadian 等。Paco：常见物体的部件与属性。载于*IEEE*计算机视觉与模式识别会议论文集，第 7141–7151 页，2023年。

Sylvestre-Alvise Rebuffi、Hakan Bilen 和 Andrea Vedaldi。使用残差适配器学习多个视觉域。神经信息处理系统进展，第 30 卷，2017。

Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10684–10695 页，2022年。

Nataniel Ruiz、Yuanzhen Li、Varun Jampani、Yael Pritch、Michael Rubinstein 和 Kfir Aberman。DreamBooth：微调文生图扩散模型，以实现主体驱动生成。*arXiv* 预印本 *arXiv:2208.12242*，2022年。

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans 等。具备深度语言理解能力的照片级文生图扩散模型。神经信息处理系统进展，35：36479–36494，2022年。

Lin Song、Yanwei Li、Zhengkai Jiang、Zeming Li、Xiangyu Zhang、Hongbin Sun、Jian Sun 和 Nanning Zheng。重新审视通用特征变换中的可学习树形滤波器。神经信息处理系统进展，33：3991–4002，2020年。

Yi-Lin Sung、Jaemin Cho 和 Mohit Bansal。VL-Adapter：面向视觉与语言任务的参数高效迁移学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第5227–5237页，2022年。

Zhi Tian、Chunhua Shen 和 Hao Chen。用于实例分割的条件卷积。载于欧洲计算机视觉会议，第 282–298 页。Springer，2020 年。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste Rozi`ere、Naman Goyal、Eric Hambro、Faisal Azhar、Aurelien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lample。Llama：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*，2023 年。

Xinlong Wang、Rufeng Zhang、Tao Kong、Lei Li 和 Chunhua Shen。Solov2：动态快速的实例分割。神经信息处理系统进展，33：17721–17732，2020年。

Xinlong Wang、Wen Wang、Yue Cao、Chunhua Shen 和 Tiejun Huang。图像以图像言说：用于上下文视觉学习的通用绘画模型。*arXiv* 预印本 *arXiv:2212.02499*，2022年。

Xinlong Wang、Xiaosong Zhang、Yue Cao、Wen Wang、Chunhua Shen 和 Tiejun Huang。SegGPT：上下文中的万物分割。*arXiv* 预印本 *arXiv:2304.03284*，2023年。

Enze Xie、Wenhai Wang、Zhiding Yu、Anima Anandkumar、Jose MAlvarez 和 PingLuo。SegFormer：基于 Transformer 的语义分割简洁高效设计。神经信息处理系统进展，34：12077–12090，2021年。

Mutian Xu、Junhao Zhang、Zhipeng Zhou、Mingye Xu、Xiaojuan Qi 和 Yu Qiao。学习几何解耦表征，以实现对三维物体点云的互补理解。载于*AAAI*人工智能会议论文集，第35卷，第3056–3064页，2021年。

Jinyu Yang、Mingqi Gao、Zhe Li、Shang Gao、Fangjing Wang 和 Feng Zheng。《追踪万物：分割万物遇上视频》。*arXiv* 预印本 *arXiv:2304.11968*，2023年。

Chaoning Zhang、Dongshen Han、Yu Qiao、Jung Uk Kim、Sung-Ho Bae、Seungkyu Lee 和 Choong Seon Hong。《快速分割一切：迈向适用于移动应用的轻量级 SAM》。*arXiv* 预印本 *arXiv:2306.14289*，2023a。

Qingru Zhang、Minshuo Chen、Alexander Bukharin、Pengcheng He、Yu Cheng、Weizhu Chen 和 Tuo Zhao。用于参数高效微调的自适应预算分配。*arXiv* 预印本 *arXiv:2303.10512*，2023b。

Renrui Zhang、Jiaming Han、Aojun Zhou、Xiangfei Hu、Shilin Yan、Pan Lu、Hongsheng Li、Peng Gao 和 Yu Qiao。Llama-Adapter：通过零初始化注意力高效微调语言模型。*arXiv* 预印本 *arXiv:2303.16199*，2023c。

Renrui Zhang、Xiangfei Hu、Bohao Li、Siyuan Huang、Hanqiu Deng、Hongsheng Li、Yu Qiao 和 Peng Gao。提示、生成，然后缓存：基础模型级联造就强大的少样本学习器。*arXiv* 预印本 *arXiv:2303.02151*，2023d。

Hengshuang Zhao、Jianping Shi、Xiaojuan Qi、Xiaogang Wang 和 Jiaya Jia。金字塔场景解析网络。载于 *IEEE*计算机视觉与模式识别会议论文集，第 2881–2890 页，2017。

Xu Zhao、Wenchao Ding、Yongqi An、Yinglong Du、Tao Yu、Min Li、MingTang 和 Jinqiao Wang。快速分割万物。*arXiv* 预印本 *arXiv:2306.12156*，2023年。

Sixiao Zheng、Jiachen Lu、Hengshuang Zhao、Xiatian Zhu、Zekun Luo、Yabiao Wang、Yanwei Fu、Jianfeng Feng、Tao Xiang、Philip HS Torr 等。借助 Transformer 从序列到序列的视角重新思考语义分割。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 6881–6890 页，2021年。

Kaiyang Zhou、Jingkang Yang、Chen Change Loy 和 Ziwei Liu。学习为视觉-语言模型设计提示。《国际计算机视觉杂志》，130(9):2337–2348，2022年。

Xueyan Zou、JianweiYang、Hao Zhang、Feng Li、Linjie Li、Jianfeng Gao 和 Yong Jae Lee。随时随地分割一切。*arXiv* 预印本 *arXiv:2304.06718*，2023年。