DenseVLM：面向开放词汇密集预测的无偏区域语言对齐

李运恒 [1], 李宇轩 [1], 曾泉胜 [1], 王文海 [3,4], 侯淇彬 [1,2,†], 程明明 [1,2]

[1] VCIP, CS, 南开大学　　[2]NKIARI, 深圳福田

[3] OpenGVLab, 上海人工智能实验室　　[4] 香港中文大学

yunhengli@mail.nankai.edu.cn, yuxuan.li.17@ucl.ac.uk

Abstract

尽管如 *CLIP* 等预训练视觉语言模型（*vision-language models*，*VLMs*）在零样本识别任务中展现出了卓越的能力，但在对局部区域识别的密集预测任务中仍表现不尽如人意。近年来，自蒸馏（*self-distillation*）逐渐成为一种有前景的微调策略，能够在无需大量标注的前提下，使 *VLM* 更好地适应局部区域。然而，现有的方法普遍存在显著的“前景偏置”（*foreground bias*）问题，即模型容易将背景区域误识为前景目标。为缓解这一问题，我们提出 *DenseVLM* 框架，旨在从强大的预训练 *VLM* 表征中学习无偏区域-语言对齐，缓解前景偏置问题。*DenseVLM* 利用预训练的 *VLM* 为无标注区域检索潜在类别，并对前景与背景特征之间的干扰进行解耦，从而在训练过程中既保证区域-类别对齐的准确性，又保留语义区分性。实验表明，*DenseVLM* 可以作为开放词汇目标检测与图像分割方法中原始 *VLM* 的直接替代模块，带来显著性能提升。此外，在更大规模、更具多样性的数据集上进行训练时，*DenseVLM* 也展现出良好的零样本扩展能力。

1. Introduction

开放词汇的密集预测任务，主要包括开放词汇的目标检测 [13, 23, 38, 57] 和语义分割 [4, 11, 28, 29, 52, 53]，旨在根据文本描述识别图像中任意类别的区域级或像素级的密集视觉概念。得益于强大的预训练视觉语言模型（vision-language models, VLMs）的发展，近年来面向密集预测的开放词汇方法 [21, 53, 63] 取得了显著进展。

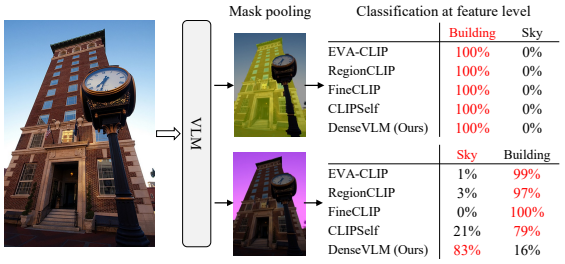


图 1. 前景偏差示例. 以往的方法将背景区域误判为相似的前景类别，而 DenseVLM 有效缓解了这一问题。

主流的视觉语言模型，如 CLIP [42] 和 EVA-CLIP [45]，在图像级语义理解方面展现出卓越的零样本识别能力。然而，这类模型在局部视觉语义建模上仍存在显著的局限性，尤其在小目标定位与背景区域识别方面表现不佳 [60, 63]。这一局限主要源于视觉语言模型的训练范式，即模型倾向于对整幅图像与文本进行全局对齐，而忽略了图像中局部区域与其对应文本描述之间的细粒度关联。为缓解这一问题，一些研究尝试构建区域级视觉-文本对齐的标签或伪标签进行视觉语言模型的训练 [27, 32, 59, 60]，但这类方法受限于高昂的人工标注成本，且在开放世界场景下缺乏良好的可扩展性。相比之下，近年来的自监督方法如 CLIPSelf [48] 和 MaskEmbed [5]，分别通过裁剪图像区域或重建被掩蔽的图像块特征编码，实现了区域语义和文本的对齐。该类方法不依赖人工标注，具有更高的灵活性与跨数据集的可扩展性。

然而，尽管已有方法在视觉语言模型的区域级视觉-文本配对预训练方面取得了一定进展，但主流的视觉语言模型 [42, 45, 60] 仍普遍存在“前景偏置”（foreground bias）问题。具体而言，在密集预测任务中，此类视觉语言模型往往会将背景区域误分类为一同出现的前景目标。这是由于在训练过程中，模型对前景目标过度关注从而忽视对

†通讯作者，邮箱：houqb@nankai.edu.cn

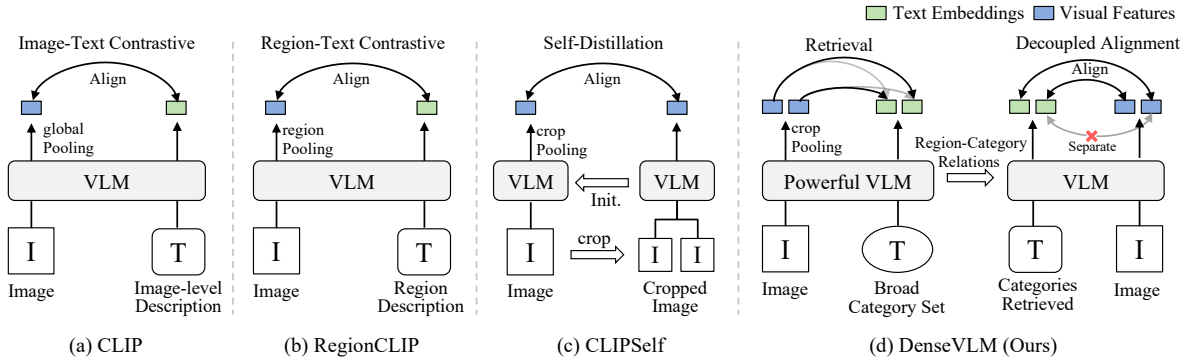


图 2. 不同视觉语言模型训练框架的对比. 不同于使用 (a) 图像-文本对比学习 [42], (b) 图像局部区域-文本对比学习 [60] 或 (c) 自蒸馏 [48] 的方法, DenseVLM 使用强教师模型的表征进行区域级视觉-文本的对齐。

背景语义的学习。这种关注比例失衡导致模型在识别过程中对前景目标产生过强偏好，进而错误地将背景区域误判为前景类别。为进一步揭示这一问题，我们对多个主流视觉语言模型 [45, 48, 60] 进行了比较实验，使用真实标注掩码从图像区域中提取特征，并进行分类。如图1所示，这些模型容易将“天空”（背景类）误判为“建筑物”（前景目标），反映出前景偏置在实际任务中的广泛存在。

为缓解上述前景偏置问题，我们提出将前景与背景区域进行显式解耦对齐，通过构建彼此独立的类别集合，实现语义层面的清晰分离。为此，我们设计了 DenseVLM，一个面向无偏区域级文本对齐的端到端框架。具体而言，对于无标注的图像区域，DenseVLM 借助一个从大规模多样化数据中预训练而来的强大视觉语言模型，无需依赖显式配对数据或自蒸馏方法 [45, 48, 60]，即可检索出语义相关的候选类别。为了提升语义覆盖能力，DenseVLM 进一步引入了来自大规模开放数据集或生成式模型 [56] 预测的广义类别集合，以增强类别多样性与泛化能力。DenseVLM 的关键设计之一是基于预定义类别集合将图像区域判定为前景或背景区域。这种判别机制使得前景与背景的区域特征得以有效解耦，从而减少二者之间的语义干扰。在训练阶段，通过分别对齐前景与背景区域的类别标签，DenseVLM 实现了更具判别性的语义分离。此外，DenseVLM 直接从预训练视觉语言模型的密集特征中提取区域语义信息，避免了传统方法中对图像块的裁剪操作 [48, 64]，在实现端到端训练从而提升建模效率的同时，也带来了更优的性能。

我们在多个开放词汇基准数据集 [48] 上系统评估了 DenseVLM 的有效性，涵盖目标检测与图像分割两大类任务。实验表明，DenseVLM 具有良好的架构兼容性，适用于多种主干网络结构，包括 Vision Transformer（ViT） [8] 与卷积神经网络（CNN） [26]，并在各项评测中均优于现有主流方法 [42, 45, 48]。此外，DenseVLM 具备良好的可扩展性。在基于大规模 SA-1B [22] 数据集进行扩展训练时，其性能仍能稳步提升，展现出强大的迁移泛化能力。在下游任务方面，DenseVLM 显著提升了开放词汇目标检测中双阶段方法 [23] 在 OV-COCO [2] 和 OV-LVIS [15] 基准上的表现。同时其在开放词汇语义分割任务中，相较于当前最先进的 SAN [53] 和 Cat-Seg [4] 方法也取得了显著的性能增益。我们的主要贡献总结如下：

- 我们系统分析了现有视觉语言模型中普遍存在的前景偏置问题，并提出一种基于解耦类别标签引导的区域级文本对齐策略，通过显式的语义结构设计实现区域级语义解耦。
- 我们提出了 DenseVLM 框架，一种面向区域级文本对齐的训练范式。该方法利用强大的预训练视觉语言模型为无标注区域自动检索类别，同时对前景与背景特征进行有效解耦，从而显著缓解前景偏置。
- 我们在多个密集预测基准数据集上进行验证，结果表明 DenseVLM 不仅显著优于现有方法，还展现出良好的可扩展性与泛化能力。

2. Related work

开放词汇密集预测： 开放词汇的密集预测方法旨在突破预定义类别的限制，从而提升其在目标检测 [13, 23, 38, 57] 和图像分割 [11, 28, 29, 52, 53] 等密集预测任务中的应用能力。预训练视觉语言模型（vision-language models，VLMs）在开放词汇中展现的零样本泛化能力进一步推动了该方向的发展。在开放词汇目标检测任务中，已有研究 [13, 47] 利

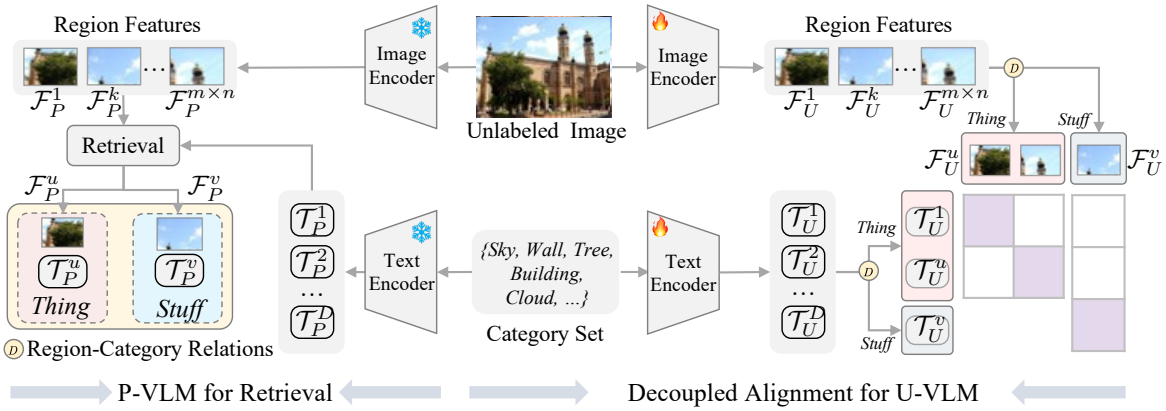


图 3. DenseVLM 的网络架构. DenseVLM 使用强视觉语言模型（Powerful VLM，P-VLM）检索文本类别，并对前景和背景的类别分别划分。在训练期间, DenseVLM 对前景和背景类别的标签进行解藕，分别与各自的文本编码进行对比学习。

用 CLIP [42] 模型实现了对新类别目标的有效识别。此外, 一些方法 [23, 49] 构建了基于冻结 CLIP 编码器的检测器，在保证性能的同时显著降低了计算开销。对于开放词汇语义分割 [6, 7, 34, 52] ，常见的方法采用双阶段流程，即先通过类别无关的掩码生成器预测候选区域，再借助冻结的 CLIP 编码器进行跨模态对齐与掩码分类。同时，一些方法 [54] 也在探索在共享冻结的 CLIP 编码器的基础上引入侧向适配网络，或设计端到端的单阶段框架 [55]。然而，由于 CLIP 仅在图像文本对上进行训练，缺乏对局部视觉与语言之间精细对应关系的建模，使其在密集预测任务中的表现较差。尽管已有研究尝试对 CLIP 进行区域级 [19] 或像素级的微调 [4, 51] 以增强区域对齐能力，但这类方法普遍受到高质量密集标注数据稀缺的限制。

图像级与区域级的视觉文本对齐：通过使用大规模图文对作为训练数据 [44]，训练出的视觉语言模型如 CLIP [42] 与 ALIGN [17] 等方法展现出了强大的零样本分类能力 [17, 37, 42]。为提升视觉语言模型在密集特征层面的视觉语言对齐效果，一些研究 [24, 25, 46, 50] 尝试对 CLIP 的输出层结构进行修改，以无需训练的方式对齐区域级视觉和文本表征。另有一些方法 [40, 41, 43] 则通过将视觉切块与图像级文本对齐，或通过自蒸馏学习局部和全局视觉的语义对应关系，但这些方法通常依赖大量图文配对数据。为实现精确的局部区域对齐，有研究 [32] 利用视觉定位（Visual Grounding）数据集中的区域标注来训练区域级视觉文本对齐模型。例如,GLIP [27, 59] 和 Grounding DINO [32] 分别通过真实区域级视觉文本对标签或伪标签（如 RegionCLIP [60]）学习区域级语言定位能力。在开放词汇目标检测与图像分割任务中，一些方法采用掩码注意力机制 [7, 21, 53, 63] 或对 CLIP [4, 18, 51] 进行微调，以实现密集的视觉与语言对齐。但这类方法高度依赖精细标注，标注成本高昂，难以在大规模场景中部署。一些方法利用自监督训练的方式缓解标注数据稀缺的问题，例如 CLIPSelf [48] 通过图像裁剪实现区域语义对齐，MaskEmbed [5] 则采用生成式思路，对掩码掩盖的视觉区域进行重建帮助模型学习区域级视觉表征。然而，这些自蒸馏方法的效果仍受限于教师模型本身的性能，并易受到前景偏置问题的干扰。为克服上述限制，我们提出利用强大的视觉语言模型为无标注区域检索类别，并借助文本类别引导对前景与背景特征进行显式解耦，以实现更稳健的区域级语义对齐。

3. Method

我们的目标是构建一种区域级视觉语言对齐模型，能够有效对齐局部视觉特征与语义信息，从而提升视觉语言模型在开放词汇密集预测任务中的表现。为实现这一目标，关键在于缓解现有视觉语言模型 [42, 45] 由于训练于图像文本对而普遍存在的前景偏置问题。此外，我们的方法旨在突破传统自蒸馏策略 [48] 所带来的性能瓶颈，实现更优的模型效果。

3.1. 视觉语言模型的公式表示

视觉语言模型通常旨在在一个共享的语义空间中同时学习全局视觉特征与文本表征。给定输入图像 I，在基于 Vision Transformer（ViT）的视觉语言模型 [3, 45] 中，密集的视觉特征通过残差注

意力模块进行提取。为了实现区域级视觉特征与文本的对齐，我们将最后一层残差注意力模块进行修改，去掉 SoftMax 操作并添加映射层，将提取到的特征表示为 $\mathcal{F}$。对于一个类别集合 $\{c_1, c_2, ...c_D\}$，其文本描述统一用一个提示模版 "*This is a photo of the c in the scene.*" 进行封装，并被文本编码器编码为文本嵌入特征集合 $\mathcal{T} = \{\mathcal{T}^1, \ldots, \mathcal{T}^D\}$。其中的 D 为类别的总体数量。

3.2. DenseVLM 框架

DenseVLM 是一个端到端的区域级视觉文本对齐框架，通过将无标注图像区域与其对应的语义类别进行精确对齐实现对偏置的抑制，旨在有效缓解前景偏置问题。如图 3 所示，该框架主要包含两个关键模块：首先，利用冻结参数的强大视觉语言模型（Powerful VLM，P-VLM）从区域特征中检索出对应的类别语义，该阶段不依赖任何人工标注，用于检索的类别语义可直接来源于大规模数据集 [1, 62]，或借助如 NXTP [56] 等方法从图像中自动生成类别集合；其次，将这些区域-语言对齐关系按照前景与背景进行解耦，并用于训练无偏视觉语言模型（Ubaised VLM，U-VLM），以显式抑制前景偏置的干扰。

为了实现区域级视觉语义和文本特征对齐，首先需要提取区域级的视觉特征。我们采用和 CLIP-Self [48] 一样的策略，将密集特征图 $\mathcal{F}$ 划分为一个 $m \times n$ 的网格。但不同于以往方法的是，我们不对原始图像进行裁剪，从而提高了计算效率并增强了特征表征能力。在每轮训练中，m 和 n 值从集合 $2, \cdots, M$ 中随机采样，默认 $M = 6$，以实现可变尺寸的网格划分。随后，通过池化操作（RoIAlign）[16] 从密集特征图 F 中提取语义区域表示 $\mathcal{F}_1, \cdots, \mathcal{F}_{m \times n}$。这种网格划分策略能够有效覆盖前景目标与背景区域的语义特征。然而，受限于现有视觉语言模型 [42, 45, 48] 的前景偏置问题，我们观察到尽管部分区域仅包含极少量与前景相关的视觉区域，但在训练中仍经常被误分类为前景类别。

在获得区域级的视觉特征后，需要对其进行文本匹配以获得用于下游模型训练的数据。我们利用 P-VLM 实现从图像网格视觉特征到文本语义的映射。该模型可同时提取区域级视觉特征 $\mathcal{F}^P = \mathcal{F}_1^P, \cdots, \mathcal{F}_{m \times n}^P$ 与文本嵌入特征 $\mathcal{T}^P = \mathcal{T}_1^P, \cdots, \mathcal{T}_D^P$。随后，通过计算区域特征与文本嵌入之间的余弦相似度，对未标注区域进行类别检索与匹配。具体地，对于某一区域 k，我们将其视觉特征表示为 $\mathcal{F}_k^P = \mathcal{F}^P[k, :]$，并计算其与所有类别文本嵌入的相似度：

$$\cos(\mathcal{F}_P^k, \mathcal{T}_P^i) = \frac{\mathcal{F}_P^k \cdot \mathcal{T}_P^i}{\|\mathcal{F}_P^k\| \|\mathcal{T}_P^i\|}, \quad \forall i = 1, 2, \ldots, D \tag{1}$$

其中 $\cdot$ 表示点积运算，$|\cdot|$ 表示欧几里得范数。接着，我们使用 SoftMax 函数将相似度归一化为类别概率分布，以确定该区域最可能对应的类别 c：

$$p^k(y = c|\mathcal{F}_P^k, \mathcal{T}_P) = \frac{\exp(\cos(\mathcal{F}_P^k, \mathcal{T}_P^c)/\tau)}{\sum_{j=1}^{D} \exp(\cos(\mathcal{F}_P^k, \mathcal{T}_P^j)/\tau)}, \tag{2}$$

其中，$\tau = 0.01$ 是温度超参数，用于调节分布的平滑程度。由于区域特征是通过随机网格划分方式提取的，这种策略在多目标场景下往往存在覆盖不完整单个目标的情况，从而显著影响了对齐的准确性。为缓解这一问题，我们引入了一种区域去噪机制，即对于匹配概率低于阈值 θ 的区域，我们直接将其丢弃，默认设定 $\theta = 0.3$。这一检索与去噪过程能够有效剔除低置信度区域，从而获得更为精确且可靠的区域视觉-文本对齐结果。对于每个区域 k，我们选择具有最大匹配概率的类别作为其对应类别，记为 $c_k = \arg\max(p_k)$。最终，每一个区域与类别之间的关系可表示为 (k, c_k)，这一对齐结果将作为后续前景与背景解耦训练的基础。

在获得可靠的区域视觉-文本对应关系之后，我们进一步对区域特征与文本表示进行对齐，以训练 U-VLM。一种直接的思路是，将每个区域的视觉特征与其对应类别的文本嵌入进行匹配，同时最大化其与无关类别之间的语义间隔。然而，由于 U-VLM 是直接继承自 P-VLM 的预训练权重，其内部仍保留了显著的前景偏置特性。若不加处理地进行区域对齐，容易导致模型在训练过程中继续强化这种偏向前景目标的趋势，从而影响整体的泛化性能。

为缓解上述问题，我们提出了一种区域解耦对齐策略，分别对前景与背景区域进行独立对齐。具体而言，训练阶段中 U-VLM 的区域特征表示记作 $\mathcal{F}^U$，其划分网格方式与 P-VLM 保持一致，对应的文本嵌入表示为 $\mathcal{T}^U$。借助 P-VLM 检索得到的区域-类别对应关系 (k, c_k)，我们在 U-VLM 中建立了区域特征与其对应类别文本嵌入的映射关系。为了区分前景与背景的语义区域，我们将区域-类别对齐关系按照预定义的两类语义集合进行

Method	COCO						ADE20K					
	Boxes		**Masks-T**		**Masks-S**		**Boxes**		**Masks-T**		**Masks-S**	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
OpenCLIP [3]	49.8	74.3	51.9	72.2	29.2	54.9	28.4	54.1	29.6	53.4	37.9	66.6
DFN [10]	38.3	65.0	31.0	57.0	26.4	54.9	30.6	57.9	24.2	49.9	32.2	57.7
SigLIP [58]	39.9	61.4	40.4	60.1	30.3	56.4	25.9	49.2	27.3	47.6	34.5	57.3
EVA-CLIP [45]	44.3	68.7	44.7	66.0	26.2	51.9	33.0	57.6	33.9	56.2	36.2	62.3
RegionCLIP† [60]	68.5	89.5	60.7	84.3	22.0	53.5	43.2	72.2	34.0	62.6	37.7	68.6
FineCLIP† [20]	64.7	86.1	62.5	80.9	36.9	70.3	43.9	71.2	45.5	68.6	46.0	74.8
CLIPSelf† [48]	69.1	88.2	66.7	83.0	41.7	75.2	48.1	77.7	47.5	74.2	53.7	82.8
DenseVLM† (Ours)	**72.3**	**89.9**	**70.1**	**84.4**	**44.9**	**76.4**	**51.0**	**81.8**	**49.3**	**76.5**	**57.0**	**84.0**

表 1. 各模型在密集表示学习上的性能比较。我们报告了在分类框和全景分割掩码任务上的 Top-1 与 Top-5 平均准确率 (thing 和 stuff)。† 表示模型在 COCO 数据集上训练，并在 ADE20K 数据集上以零样本（zero-shot）设置进行评估。

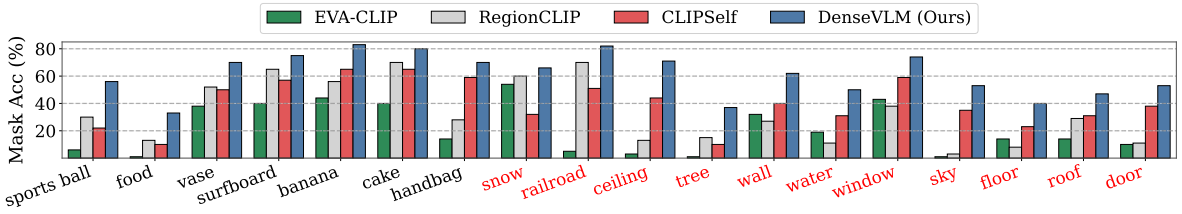


图 4. 在 COCO 数据集的各类别掩码准确率对比中，我们的方法展现出显著优势，特别是在前景偏置问题上的改进尤为明显。前景类别以黑色显示，背景类别以红色显示。

解耦：前景目标集（Thing，记作 $\mathcal{U}$）与背景区域集（Stuff，记作 $\mathcal{V}$）。在训练中，我们采用选择性对比机制，即对于前景（或背景）区域，仅与语义上无关的类别进行对比，从而引导模型更加关注语义相关的背景区域，减少无关前景类别的干扰。这种选择性对比策略有助于模型更好地捕捉背景区域的判别特征，进而实现前景与背景之间更准确的语义区分。根据公式1，当 $c_k \in \mathcal{V}$（即该区域属于背景类别）时，区域 k 的概率 q_k 可按如下方式计算：

$$q^k = \frac{\exp\left(\cos\left(\mathcal{F}_U^k, \mathcal{T}_U^c\right)/\tau\right)}{\sum_{j=1}^{\mathcal{U}\cup\mathcal{V}} \exp\left(\cos\left(\mathcal{F}_U^k, \mathcal{T}_U^j\right)/\tau\right)} \tag{3}$$

最终的区域对齐通过最大化区域特征与文本嵌入之间的余弦相似度来实现。同样地，当 $c_k \in \mathcal{U}$（即该区域属于前景类别）时，其概率 $\tilde{q}_k$ 的计算方式如下：

$$\tilde{q}^k = \begin{cases} \frac{\exp\left(\cos\left(\mathcal{F}_U^k, \mathcal{T}_U^c\right)/\tau\right)}{\sum_{j=1}^{u} \exp\left(\cos\left(\mathcal{F}_U^k, \mathcal{T}_U^j\right)/\tau\right)} & \text{if } c \in \mathcal{U} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

同样依据公式3和公式4计算 P-VLM 的预测概率 p_k 和 $\tilde{p}_k$。DenseVLM 支持通过 KL 散度优化实现端到端训练，具体过程如下所述：

$$\mathcal{L}^k = \begin{cases} \mathrm{KL}\left(p^k \| q^k\right) & \text{if } c^k \in \mathcal{U} \\ \mathrm{KL}\left(\tilde{p}^k \| \tilde{q}^k\right) & \text{otherwise} \end{cases} \tag{5}$$

每张图像的总体损失定义为 $L = \frac{1}{m' \times n'} \sum_{k=1}^{m' \times n'} L_k$，其中，求和范围覆盖所有未被区域去噪机制剔除的区域。$m' \times n'$ 表示当前图像中保留的有效区域总数。

4. Experiments

4.1. Benchmarks

（1）数据集介绍：1）COCO [31] 是一个大规模的全景分割数据集，涵盖 80 个 Thing 类和 53 个 Stuff 类。训练集包含 118,000 张图像，验证集包含 5,000 张图像，广泛用于目标检测与语义分割研究；2）ADE20K [61] 覆盖了丰富的室内与室外场景，验证集包含 2,000 张图像，包含 100 个 Thing 类与 50 个 Stuff 类。我们分别使用其两种版本进行开放词汇评估：完整版包含 847 个语义类别（记作 A-847），以及更常见的 150 类子集版本（记作 A-150）；3）Pascal-Context [9] 是基于

VLMs Frozen & Training	Region Cropping	Alignment Strategy	GPU Memory (per card)	Time Overhead (per epoch)	**Boxes**		**Masks-T**		**Masks-S**	
					Top1	Top5	Top1	Top5	Top1	Top5
ViT-B/16 & ViT-B/16	Images	Features KD	37G	25min	69.1	88.2	66.7	83.4	41.7	75.2
ViT-L/14 & ViT-B/16	Images	Features KD	39G	37min	24.2	52.4	23.4	51.1	10.1	39.1
ViT-L/14 & ViT-B/16	Images	Logics KD	39G	55min	72.2	89.8	68.8	83.8	42.6	75.2
ViT-L/14 & ViT-B/16	Features	DenseVLM	39G	**23min**	**73.4**	**90.5**	**71.0**	**84.8**	**45.6**	**77.8**

表 2. 各类框架的结果与对比，对比参数包括用于特征对齐的视觉语言模型、区域裁剪策略，以及对齐策略。此外，还报告了各方法在 GPU 显存利用效率和时间开销方面的表现。所有模型均在四块 A40 GPU 上训练，每个训练轮次包含 11.8 万张图像。

item	value
image size	512 × 512
optimizer	AdamW [36]
learning rate	0.0001
β_1	0.9
β_2	0.98
weight decay	0.1
batch size (per card)	48
warmup steps [12]	1000
epochs	6
learning rate scheduler	cosine decay [35]
number of GPUs	4
automatic mixed precision[1]	True

表 3. 实验设置细节。

Pascal VOC 2010 构建的扩展数据集，涵盖更细致的语义标注。我们采用其完整版本，共 459 个语义类别（记作 PC-459），用于开放词汇语义分割评估；4）OV-COCO 是开放词汇目标检测数据集，由 OV-RCNN [57] 提出，将 COCO 中的 65 个对象类别划分为 48 个基础类别与 17 个新类别，用于评估模型在识别未见类别方面的泛化能力；5）OV-LVIS 是由 ViLD [13] 引入的开放词汇检测基准，将 LVIS v1.0 [15] 数据集中 337 个稀有类别定义为 novel 类别。该基准强调模型在长尾类别检测中的表现。

（2）实验设置：为了验证所提出的 DenseVLM 的有效性,我们在密集预测任务上进行了实验,使用 COCO Panoptic 的 val2017 和 ADE20K Panoptic 的验证集进行评估。参照 CLIPSelf [48]，我们分别使用从根据目标框池化的特征进行框分类评估，使用根据掩码池化的特征进行掩码分类评估，并区分前景目标（Masks-T）与背景内容（Masks-S）。该流程与图 1所示类似，即利用真实标注提取局部特征，并对局部区域进行分类评估。所有实验结果均以 Top-1 与 Top-5 的平均分类准确率进行报告。

（3）实验细节：我们采用 CLIPSelf 中的 ViT-L/14 作为强表征视觉语言模型（P-VLM），并使用 EVA-CLIP 中的 ViT-B/16 作为待训练的去偏视觉语言模型（U-VLM）。为提高计算效率，P-VLM 的参数保持冻结，仅对 U-VLM 的图像编码器进行训练，配合预提取的文本嵌入进行优化。考虑下游任务的实际应用场景以及性能与效率之间的权衡，我们将输入图像统一调整为 512×512 的分辨率。模型训练使用 AdamW [36] 优化器，权重衰减系数设为 0.1，训练周期为 6 轮。初始学习率设为 0.00001，并采用余弦退火策略 [35] 进行调度。我们在 NVIDIA A40 GPU 上训练所有模型，以确保各项实验具有公平的对比性。详细的配置见表 3。对于 SA-1B [22] 数据集，我们使用 8 张 A40 GPU 以实现高效且可扩展的训练。在开放词汇语义分割任务中，我们在 COCO-Stuff [1] 数据集上训练如 SAN [53] 和 CAT-Seg [4] 等模型，迭代次数为 80k。在开放词汇目标检测任务中，模型分别在 OV-COCO [2] 基准上训练 3 轮，在 OV-LVIS [15] 基准上训练 48 轮。

4.2. 对比实验

（1）定量评估：我们在 COCO Panoptic [31] 和 ADE20K Panoptic [61] 数据集上，对多种视觉语言模型（VLM）的密集表征能力进行了全面的定量评估。如表1所示，尽管已有方法 [3, 10, 45, 58] 在零样本图像分类任务中表现优异，但在区域识别任务中的表现仍明显不足。例如，EVA-CLIP [45] 在 COCO 上的框分类 Top-1 准确率仅为 44.3%，在 ADE20K 上更是降至 33.0%。虽然 RegionCLIP [60] 通过区域-文本对进行训练，在 COCO 上取得了一定提升，但在如 ADE20K 等其他数据集上泛化能力较弱。此外，FineCLIP [20] 与 CLIPSelf [48] 借助自蒸馏机制在前景掩码（Mask-T）分类上取得了较好成绩，但在背景掩码

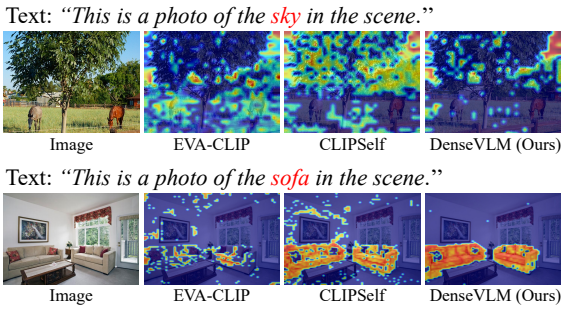


图 5. 利用视觉特征与文本嵌入之间余弦相似度图的可视化结果，我们对比了 DenseVLM 与现有方法的对齐性能。

P-Thing		P-Stuff				
Thing	*Stuff*	*Thing*	*Stuff*	**Boxes**	**Masks-T**	**Masks-S**
✓	✓	✓	✓	**74.3**	70.9	42.6
✓	✗	✗	✓	74.2	70.8	42.0
✓	✓	✗	✓	74.1	70.9	41.3
✓	✗	✓	✓	73.4	**71.0**	**45.6**

表 4. 针对解耦对齐策略的消融实验分析。✓表示该区域被判定为与该类别集合无关。

(Mask-S) 分类中表现不佳。相比之下，我们提出的方法 DenseVLM 在 COCO 上的 Top-1 Mask-T 分类准确率提升了 4.3%，Mask-S 分类准确率提升了 3.9%，显著优于上述方法，充分展示了其在密集预测任务中的有效性。

图1展示了不同类别下的掩码分类准确率对比，进一步验证了 DenseVLM 在区域视觉-文本对齐方面的有效性。值得注意的是，DenseVLM 在诸如“sky”(天空) 和“wall”(墙体) 等背景类别上显著提升了分类准确率，有效缓解了前景偏置问题。

(2) 定性评估：我们通过计算与文本描述的目标类别之间的余弦相似度，来可视化注意力图。如图5所示，DenseVLM 相较于 EVA-CLIP 和 CLIP-Self，能够实现更精确且更完整的目标定位。同时，该方法在保持语义分离方面表现更佳，有效减少了来自其他物体的干扰。

4.3. 消融实验

在消融实验中，DenseVLM 使用 COCO train2017 中的未标注图像进行训练，训练设备为 4 块 A40 GPU，并在 val2017 上进行评估。同时为了性能和效率的平衡，该章节的实验采用 EVA-CLIP [45] 中的 ViT-B/16 模型进行实验。

(1) 网络架构设计的实验：我们通过消融实验

	Boxes		**Masks-T**		**Masks-S**	
Categories	Top1	Top5	Top1	Top5	Top1	Top5
Category set from dataset annotation.						
133 (80)	71.1	88.5	68.7	83.0	44.7	75.2
171 (80)	72.3	89.8	69.4	85.8	44.2	76.0
273 (160)	72.3	89.9	70.1	84.4	44.9	76.4
316 (204)	**73.4**	**90.5**	**71.0**	**84.8**	**45.6**	**77.8**
Category set generated by NXTP [56].						
210 (133)	**72.6**	**90.2**	**70.4**	84.6	41.5	75.7
794 (484)	72.5	90.4	70.3	**84.8**	**44.1**	**76.2**

表 5. 不同类别集合设置下的消融实验。括号中所示数字为前景实物 (thing) 类别的个数。

	Boxes		**Masks-T**		**Masks-S**	
P-VLM	Top1	Top5	Top1	Top5	Top1	Top5
ViT-B/16	70.0	88.6	68.0	83.0	43.0	75.3
ViT-L/14	**73.4**	**90.5**	**71.0**	**84.8**	**45.6**	**77.8**

表 6. 针对用于检索未标注区域类别的不同 P-VLM 模型的消融实验分析。

系统探讨了 DenseVLM 中若干关键设计选择，包括用于表征对齐的目标视觉语言模型(VLM)、区域裁剪方式，以及优化策略，并重点分析它们在 GPU 显存使用与训练时间开销上的影响。以自蒸馏方法 [48] 为基线，当将目标模型替换为 CLIPSelf 中的 ViT-L/14 时，由于破坏了视觉与语言特征之间的对齐关系，模型性能出现明显下降。将特征蒸馏替换为 logit 蒸馏虽能带来性能提升，但由于需重复进行特征提取，训练效率显著下降，同时引入前景偏置问题。相比之下，DenseVLM 通过更高效的特征裁剪策略与前背景解耦对齐框架，不仅在性能上实现超越，同时大幅降低了训练时间与显存占用。

(2) 解耦对齐策略的消融实验：表4分析了解耦对齐策略对 DenseVLM 性能的影响。检索到的类别 c^k 可被划分为 Thing 类 (可数前景目标) 和 Stuff 类 (不可数背景区域)，分别记作“P-Thing”与“P-Stuff”。DenseVLM 利用每个区域对应的类别进行有选择性的对比学习。当所有区域统一与所有类别进行对比学习时，模型在识别前景区域方面表现较好，但在区分背景时存在困难。若采用完全解耦的设置，即每个区域仅与其所属类别集合进行对比学习时，性能反而下降很明显。当 P-Thing 区域对比 Stuff 类别，而 P-Stuff 区域不对比 Thing 类别时，模型前景偏置最为严重。在 DenseVLM

Input Image Size	GPU Memory (per card)	COCO						ADE20K					
		Boxes		**Masks-T**		**Masks-S**		**Boxes**		**Masks-T**		**Masks-S**	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
224	9G	60.1	79.9	49.4	62.4	35.3	64.2	40.0	70.0	36.3	56.4	50.3	77.0
320	11G	66.2	85.4	59.2	73.0	41.0	71.2	45.6	76.0	44.0	67.6	54.3	81.7
512	16G	73.4	90.5	71.0	84.8	45.6	77.8	51.3	82.2	52.1	78.0	57.8	85.5
768	27G	74.4	91.3	75.4	90.1	45.5	79.0	52.7	82.9	55.4	82.6	58.2	86.6
1024	39G	**76.6**	**93.1**	**78.7**	**93.6**	**46.5**	**79.8**	**53.2**	**83.6**	**56.8**	**83.2**	**58.6**	**86.8**

表 7. 不同输入图像尺寸设置下的消融实验。我们在 COCO Panoptic 与 ADE20K Panoptic 基准数据集上，对目标框与全景掩码的分类任务报告了 Top-1 和 Top-5 的平均准确率。显存使用统计基于 A40 显卡，在批量大小（batch size）为 12 的条件下获得。

Method	Region Proposals	COCO						ADE20K					
		Boxes		**Masks-T**		**Masks-S**		**Boxes**		**Masks-T**		**Masks-S**	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
CLIPSelf	✗	69.1	88.2	66.7	83.0	41.7	75.2	48.1	77.7	47.5	74.2	53.7	82.8
CLIPSelf	✓	70.2	89.2	68.1	83.5	35.7	71.8	49.8	79.7	51.5	76.0	50.9	80.7
DenseVLM	✗	73.4	90.5	71.0	84.8	45.6	77.8	51.3	82.2	52.1	78.0	57.8	85.5
DenseVLM	✓	74.4	91.3	75.4	90.1	45.9	79.0	52.7	82.9	55.4	82.6	58.2	86.6

表 8. 区域提议的消融实验。我们在 COCO Panoptic 与 ADE20K Panoptic 基准数据集上，评估了模型在分类目标框与全景掩码（涵盖 thing 类与 stuff 类）任务中的 Top-1 与 Top-5 平均准确率。

θ	**Boxes**		**Masks-T**		**Masks-S**	
	Top1	Top5	Top1	Top5	Top1	Top5
0.0	72.1	89.6	68.2	84.3	43.6	76.1
0.1	72.7	90.2	69.1	84.3	44.6	77.2
0.2	73.1	90.4	69.7	84.6	45.1	77.7
0.3	**73.4**	**90.5**	**71.0**	**84.8**	**45.6**	**77.8**
0.4	73.2	90.2	70.2	84.3	45.2	77.5
0.5	73.1	90.0	70.0	84.3	45.0	77.1
0.6	73.1	89.9	69.6	84.0	44.6	76.3

表 9. 阈值 θ 的消融实验。

中，我们采用了一种平衡策略，即 P-Stuff 区域同时对比 Thing 和 Stuff 类别，而 P-Thing 区域仅对比 Thing 类别，从而有效缓解前景偏置，显著提升整体性能。

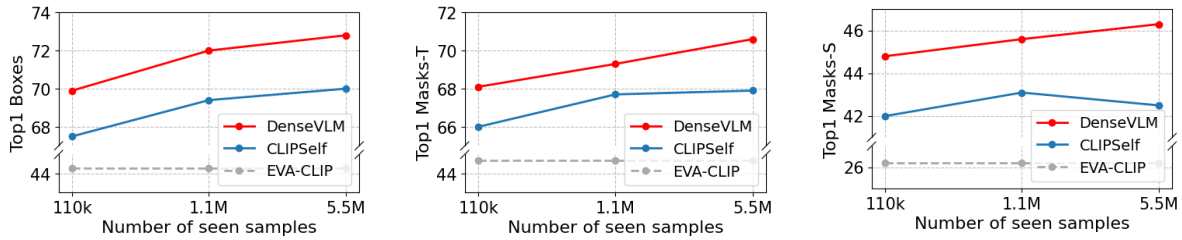
（3）不同类别集的消融实验：表5展示了在类别检索与对齐过程中，采用不同类别集对模型性能的影响。当使用来自训练数据集标注的类别集时，我们观察到，随着类别数量的增加模型性能持续提升。

仅增加背景类别数量有助于模型更清晰地区分前景特征，而进一步引入更多前景类别则带来了显著的性能增益。这些改进主要得益于更丰富的类别集合增强了模型的表征能力，使其在区域语义对齐任务中具备更强的区分性和泛化能力。

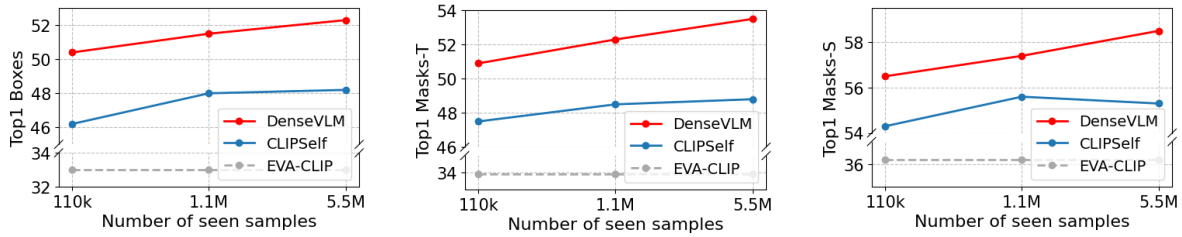
为了解决依赖类别标注集的局限性，我们借助生成模型 NXTP [56] 基于图像内容自动生成类别集合，并使用 DeepSeek-R1 [14] 对生成的类别进行前景与背景的分类。如表5所示，即使在无标注类别集的情况下，我们的方法仍展现出优越的性能。这表明 DenseVLM 具备良好的开放词汇泛化能力，能够适应更广泛的实际应用场景。

（4）不同强视觉语言模型的消融实验：表6展示了在使用不同检索模型（P-VLM）下，DenseVLM 的性能表现。当采用 ViT-B/16 作为检索模型时，我们的方法在掩码分类任务中表现优异，超过了基于相同结构（ViT-B/16）的 CLIPSelf 模型。这一性能提升表明，DenseVLM 有效缓解了前景偏置问题，增强了前景与背景区域的区分能力。进一步将 P-VLM 替换为更强大的 ViT-L/14 时，在所有评估指标上性能均有进一步提升，验证了强表征能力的预训练模型在区域语言对齐任务中的重要性。

（5）输入图像尺寸的消融实验：为了评估输入



(a) 各模型在 COCO 密集预测基准上的 Top-1 平均准确率。



(b) 各模型在 ADE20K 密集预测基准上的 Top-1 平均准确率。

图 6. 针对三种不同数据规模的预训练模型进行零样本性能比较。我们选取 SA-1B 数据集中 100K、1.1M 和 5.5M 的训练样本子集，针对 COCO 和 ADE20K 基准进行零样本性能评估。

图像分辨率对 DenseVLM 的影响，我们在训练和推理阶段分别使用 224、320、512、768 和 1024 像素的不同分辨率进行实验。如表7所示，随着图像分辨率从 224 提升至 1024 像素，模型在区域分类任务上的性能逐步提升。这一提升可归因于高分辨率图像能够提供更丰富的细节信息。然而，性能的提升也伴随着 GPU 显存占用的显著增长。综合计算资源与模型性能之间的权衡，我们最终将图像尺寸设定为 512×512 像素，以获得最佳平衡。

(6) 区域提议的消融实验：我们借鉴了 RegionCLIP [60] 的方法，使用伪标注的区域-文本对对 VLM 进行微调，并与 CLIPSelf [48] 在利用该类对的方式上进行了对比。如表 8所示，CLIPSelf 将原先的随机图像裁剪替换为伪区域-文本对，提升了前景目标的识别能力，但同时背景内容的识别准确率有所下降。相比之下，我们提出的 DenseVLM 不仅在前景目标的识别精度上取得显著提升，还进一步增强了对背景区域的辨别能力，展现出更全面的区域理解能力。

(7) 阈值 θ 的消融实验：我们通过消融实验评估了区域去噪中不同阈值 θ 的影响。如表 9所示，当 $\theta = 0$ 时，模型性能最差。较低的 θ 虽然提升了 Top-5 准确率，但整体性能反而下降，可能是由于低置信度的类别引入了对齐混淆，从而干扰模型判断。相反，当 θ 设置过高时，又会过滤掉过多的局部区域图像，导致有效信息的损失，进而影响最终性能。因此，我们在 DenseVLM 中默认选用 $\theta = 0.3$，以在信息保留与噪声抑制之间取得较好的平衡。

4.4. 不同数据规模下的零样本能力对比

为评估训练数据规模对视觉语言模型（VLM）性能的影响，我们从 SA-1B 数据集 [22] 中选取了三个不同规模的子集：100K、1.1M 和 5.5M 张图像。我们在这些数据子集上分别训练了本方法 DenseVLM 与对比方法 CLIPSelf，训练配置为使用 8 张 A40 GPU，每张卡的 batch size 为 48。

如图 6, 所示，在 COCO 与 ADE20K 两个开放词汇密集预测任务中，随着训练数据规模的扩大, CLIPSelf 的性能提升逐渐趋于饱和。相比之下, DenseVLM 展现出持续且稳定的性能提升趋势, 表现出更强的可扩展性。这一结果表明，DenseVLM 能够充分挖掘无标注大规模图像数据中的语义信息，具备更强的数据利用效率与泛化能力，适合在真实世界中进行大规模部署与应用。

4.5. 不同骨干网络的对比

为验证 DenseVLM 的通用性，我们将其应用于多种主流骨干网络架构，包括 OpenCLIP [3] 提供的 ViT-B/16、ResNet-50x4（R50x4）、ConvNeXt-B，以及 EVA-CLIP [45] 提供的 ViT-B/16，并在各架构下与 CLIPSelf [48] 进行对比。从表 10中可以观察到，CLIPSelf 在 ViT 系列架构上效果较好，但在 CNN 架构（如 R50x4）上表现明显退化，甚至低于基线方法。相比之下，DenseVLM 在所有架构下均实现了稳定且显著的性能提升，表现出强大的模型适应性。尤其值得注意的是，DenseVLM 有效

Backbones	VLMs	Boxes	Masks-T	Masks-S
ViT-L/14	CLIPSelf	75.2	73.1	44.5
ViT-B/16	OpenCLIP	49.8	51.9	29.2
ViT-B/16*	CLIPSelf	67.6	64.4	44.5
ViT-B/16*	DenseVLM	**71.9**	**70.0**	**47.8**
ViT-B/16	EVA-CLIP	44.3	44.7	26.2
ViT-B/16	CLIPSelf	69.1	66.7	41.7
ViT-B/16	DenseVLM	**73.4**	**71.0**	**45.6**
R50x4	OpenCLIP	59.2	50.5	39.1
R50x4	CLIPSelf	59.1	49.9	37.2
R50x4	DenseVLM	**65.6**	**55.2**	**40.9**
ConvNeXt-B	OpenCLIP	57.5	48.0	31.1
ConvNeXt-B	CLIPSelf	62.6	57.6	41.6
ConvNeXt-B	DenseVLM	**67.1**	**63.4**	**43.5**

表 10. 不同主干网络的性能对比。* 表示模型的初始化权重来源于 OpenCLIP [3]。

Method	A-150	A-847	PC-459
PACL [40]	31.4	-	-
OVSeg [29]	24.8	7.1	11.0
MAFT [18]	29.1	10.1	12.6
SED [51]	31.6	11.4	18.6
SCAN [33]	30.8	10.8	13.2
CAT-Seg+CLIPSelf [48]	29.7	10.1	-
CAT-Seg+FineCLIP [20]	32.4	12.2	-
SAN [53]	27.4	10.0	13.0
SAN+DenseVLM	**29.5**$_{+2.1}$	**10.4**$_{+0.4}$	**15.6**$_{+2.6}$
CAT-Seg [4]	31.4	11.7	18.4
CAT-Seg+DenseVLM	**34.1**$_{+2.7}$	**12.2**$_{+0.5}$	**18.7**$_{+0.3}$

表 11. 开放词汇语义分割性能对比。

缩小了 CLIPSelf 微调的大型模型（如 ViT-L/14）与轻量级模型之间的性能差距，部分轻量模型在背景分类任务（Masks-S）上甚至超过了 ViT-L/14。上述结果充分表明，DenseVLM 不仅适用于不同类型的主干网络，在资源受限的实际场景中也具备出色的适配能力与部署潜力。

4.6. 图像网格块分类可视化

我们使用 CLIPSelf [48] 中强大的 ViT-L/14 模型，对图像网格块的分类结果进行可视化。如图 7所示，该模型对前景目标的识别能力较强，但大量背景区域的网格块仍被误分类为前景对象，反映出模型存在显著的前景偏置。此外，训练过程中的 VLM 易于学习这些错误分类结果，进一步加剧问题。值得注意的是，那些被误分类的区域通常伴随着较低的置信度得分，这也强调了在训练中筛除这些低置信度区域的重要性，以提升模型对局部语义的辨别能力。

(a) OV-COCO benchmark

Method	Backbone	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
VLDet [30]	RN50	32.0	50.6	45.8
F-VLM [23]	RN50	28.0	-	39.6
BARON-Cap [47]	RN50	33.1	54.8	49.1
CORA [49]	RN50	35.1	35.5	35.4
RO-ViT [21]	ViT-B/16	30.2	-	41.5
RO-ViT [21]	ViT-L/16	33.0	-	47.7
F-ViT+CLIPSelf [48]	ViT-B/16	25.4	40.9	36.8
F-ViT+FineCLIP [20]	ViT-B/16	29.8	45.9	41.7
F-ViT [48]	ViT-B/16	17.5	41.0	34.9
F-ViT+DenseVLM	ViT-B/16	**33.1**	**52.5**	**47.4**

(b) OV-LVIS benchmark

Method	Backbone	mAP_r	mAP_c	mAP
VLDet [30]	RN50	21.7	29.8	30.1
BARON-Cap [47]	RN50	22.6	27.6	27.6
F-VLM [23]	RN50	18.6	-	24.2
COR [49]	RN50x4	22.2	-	-
RO-ViT [21]	ViT-B/16	28.0	-	30.2
F-ViT+CLIPSelf [48]	ViT-B/16	10.6	7.6	9.3
F-ViT+FineCLIP [20]	ViT-B/16	10.4	8.0	9.5
F-ViT [48]	ViT-B/16	11.5	12.3	15.4
F-ViT+DenseVLM	ViT-B/16	**23.9**	**18.4**	**21.4**

表 12. 开放词汇目标检测性能对比。

4.7. 在开放词汇密集预测任务中的应用

（1）实验设置：为了评估所提 DenseVLM 在下游任务中的表现，我们将其作为骨干网络应用于开放词汇的密集预测任务，包括目标检测与图像分割。为保证实验的公平性，所有 DenseVLM 模型均在 COCO train2017 数据集上进行训练，输入图像统一调整为 512×512 分辨率。

（2）开放词汇语义分割实验：我们将基于 OpenAI CLIP [42] 预训练权重初始化的 DenseVLM 应用于开放词汇语义分割任务，涵盖了使用冻结骨干网络的 SAN [53] 和使用微调骨干网络的 CatSeg [4] 两种方法。模型在 COCOStuff [1] 数据集上进行训练，并在 ADE20K [62]（ADE-150 与 ADE-847）及 PASCAL Context [39]（PC-459）数据集上进行评测，采用平均交并比（mIoU）作为评价指标。如表 11所示，DenseVLM 在所有评测数据

集上均实现了持续且显著的性能提升，进一步推动了该领域的最新性能水平。

（3）开放词汇目标检测实验：基于前人工作 [48]，我们采用 F-ViT 架构构建开放词汇目标检测器，该架构为基于 EVA-CLIP [45] 冻结 ViT 的两阶段检测器。如表 12所示，我们在 OV-COCO [2] 基准上，通过 IoU 阈值 0.5 下的框平均精度（AP）指标评估基础类（base）、新类（novel）及整体类别。在 OV-LVIS [15] 基准上，则采用稀有类、常见类及所有类别的均值平均精度指标进行评价。将 EVA-CLIP 的冻结 ViT 替换为 DenseVLM 模型后，两个基准测试均表现出明显性能提升，且整体表现与先前方法相比具有较强的竞争力。

5. Conclusions

本文提出了 DenseVLM 框架，旨在缓解区域级视觉语言对齐中的前景偏置问题。DenseVLM 可无缝应用于开放词汇的目标检测与图像分割任务，持续显著优于基线模型的性能表现。此外，我们通过高效的区域检索与解耦对齐策略，验证了 DenseVLM 在大规模数据集（如 SA-1B）上的良好扩展能力。总体而言，DenseVLM 为提升视觉语言模型中密集表征的泛化能力提供了一种通用且高效的解决方案，极大推动了开放词汇密集预测任务的发展。

6. Limitations and broader impact

局限性：我们的目标是构建一种区域视觉-文本对齐模型，能够有效整合局部视觉与语义特征，从而提升开放词汇密集预测任务的性能。与已有的预训练视觉语言模型（VLMs） [3, 45, 48, 60] 相比，我们提出的 DenseVLM 在多个任务中取得了更优的表现，并显著提升了下游任务的效果。我们相信 DenseVLM 仍具有更大的潜力，主要体现在以下几个方面：1）可扩展性：DenseVLM 基于一个高效的、无监督的区域-语言对齐框架设计，使其具备良好的跨数据集适应能力。然而，由于计算资源的限制，我们尚未能将该方法扩展至更大规模的数据集。2）模型容量：我们采用了 CLIPSelf [48] 中的 ViT-L/14 作为预训练的强大视觉语言模型（P-VLM）。若能利用更强的 VLM，将有望进一步提升性能，同时如何将其丰富的语义知识更有效地迁移到训练模型中，是一个值得探索的方向。3）细粒度语义能力：当前我们将物体划分为粗粒度的 *thing* 与 *stuff* 类别。更细粒度的语义分割能力，以及更精细的解耦对齐机制，将有助于模型更好地区分语义上相似的类别。我们计划在后续研究中进一步探索这些方向。

更广泛的影响：DenseVLM 在开放词汇场景密集预测任务中展现出显著的潜力，这将有助于推动诸如机器人技术与环境监测等多个应用领域的发展。通过使系统具备在无需对特定类别进行预训练的情况下，识别并理解多种对象与语境的能力，DenseVLM 有望实现更具适应性与通用性的应用场景。鉴于其广泛的适用性与非特定性设计，我们的方法旨在支持多种技术层面的进步，而并非直接面向特定的社会问题。

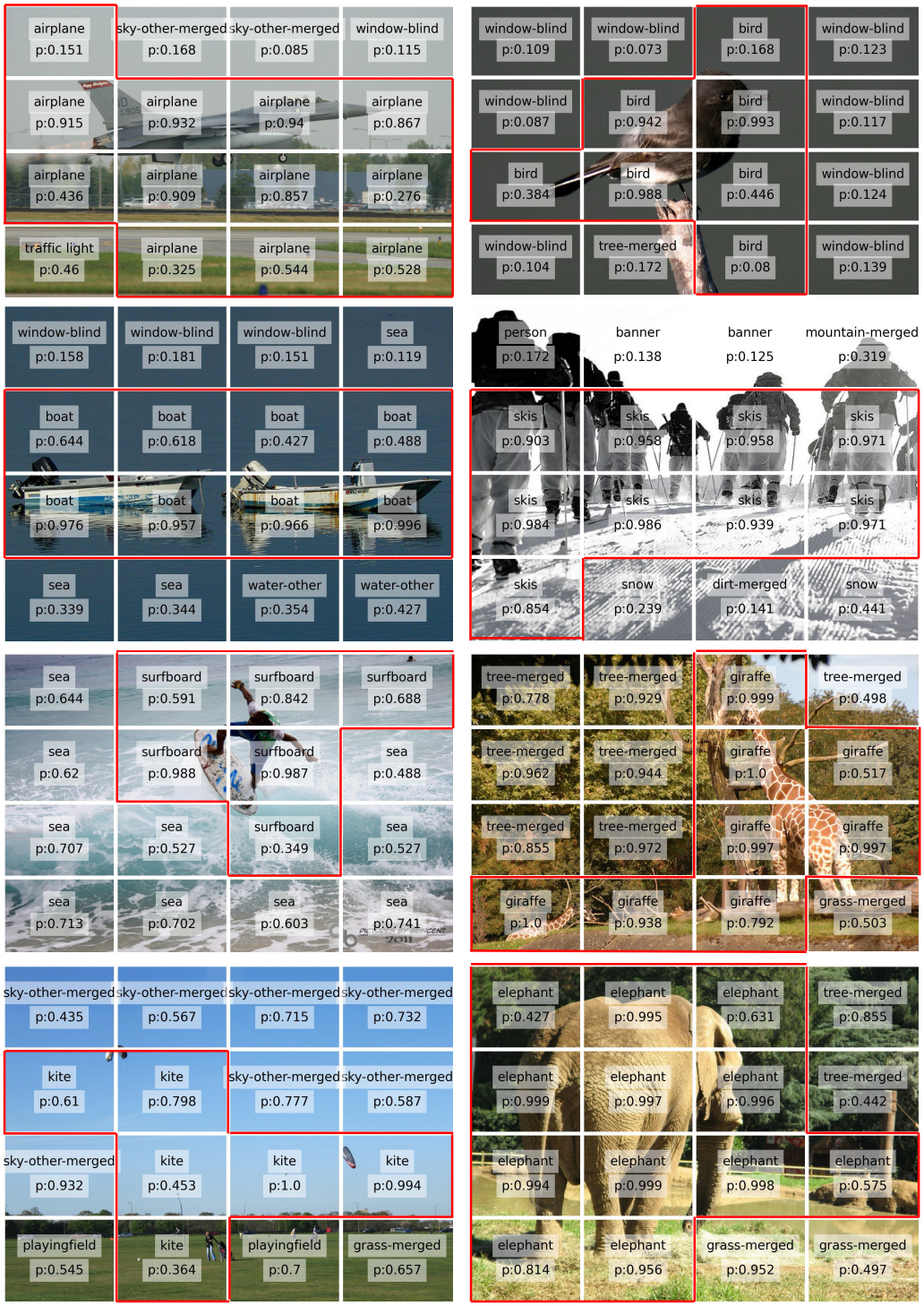


图 7. 图像中网格划分区域的分类可视化结果。即使存在大量背景 patch 被误判为前景目标，强大的 ViT-L/14 模型仍展现出对前景物体识别的明显偏重。

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 4, 6, 10

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 6, 11

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*,

pages 2818–2829, 2023. 3, 5, 6, 9, 10, 11

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, 2024. 1, 2, 3, 6, 10

[5] Ian Covert, Tony Sun, James Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. *arXiv preprint arXiv:2410.11087*, 2024. 1, 3

[6] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022. 3

[7] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 3

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 5

[10] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 5, 6

[11] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557, 2022. 1, 2

[12] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2018. 6

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, pages 1–20, 2022. 1, 2, 6

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 8

[15] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2, 6, 11

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 4

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 3

[18] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *NeurIPS*, 36:35631–35653, 2023. 3, 10

[19] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. *arXiv preprint arXiv:2408.00744*, 2024. 3

[20] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Wei Wei, Huiwen Zhao, Zhiwu Lu, et al. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *NeurIPS*, 37:27896–27918, 2024. 5, 6, 10

[21] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *CVPR*, pages 11144–11154, 2023. 1, 3, 10

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2, 6, 9

[23] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2022. 1, 2, 3, 10

[24] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, 2024. 3

[25] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, 2024. 3

[26] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 2

[27] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 1, 3

[28] Yunheng Li, ZhongYu Li, Quansheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *ICML*, pages 28243–28258, 2024. 1, 2

[29] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 1, 2, 10

[30] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *ICLR*, 2032. 10

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 6

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 3

[33] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3491–3500, 2024. 10

[34] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *CVPR*, 2024. 3

[35] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[37] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, pages 13–23, 2019. 3

[38] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, pages 728–755, 2022. 1, 2

[39] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014. 10

[40] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *CVPR*, pages 19413–19423, 2023. 3, 10

[41] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *ECCV*, pages 38–55. Springer, 2024. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 3, 4, 10

[43] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *ICCV*, pages 5571–5584, 2023. 3

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, pages 25278–25294, 2022. 3

[45] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong

Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1, 2, 3, 4, 5, 6, 7, 9, 11

[46] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, pages 315–332, 2025. 3

[47] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, pages 15254–15264, 2023. 2, 10

[48] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *ICLR*, 2024. 1, 2, 3, 4, 5, 6, 7, 9, 10, 11

[49] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, pages 7031–7040, 2023. 3, 10

[50] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *ECCV*, pages 320–337. Springer, 2024. 3

[51] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *CVPR*, 2024. 3, 10

[52] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, pages 2955–2966, 2023. 1, 2, 3

[53] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 1, 2, 3, 6, 10

[54] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE TPAMI*, pages 1–17, 2023. 3

[55] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *NeurIPS*, 2023. 3

[56] Kaiyu Yue, Bor-Chun Chen, Jonas Geiping, Hengduo Li, Tom Goldstein, and Ser-Nam Lim. Object recognition as next token prediction. In *CVPR*, pages 16645–16656, 2024. 2, 4, 7, 8

[57] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 1, 2, 6

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 5, 6

[59] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *NeurIPS*, 35:36067–36080, 2022. 1, 3

[60] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 1, 2, 3, 5, 6, 9, 11

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 5, 6

[62] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127:302–321, 2019. 4, 10

[63] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, page 696–712, 2022. 1, 3

[64] Yuanbing Zhu, Bingke Zhu, Zhen Chen, Huan Xu, Ming Tang, and Jinqiao Wang. Mrovseg: Breaking the resolution curse of vision-language models in open-vocabulary semantic segmentation. *arXiv preprint arXiv:2408.14776*, 2024. 2