

我们介绍Florence-2，这是一种新颖的视觉基础模型，它采用统一的、基于提示的表示方法，适用于多种计算机视觉和视觉-语言任务。尽管现有的大型视觉模型在迁移学习方面表现出色，但它们难以通过简单指令执行多样化任务，这种能力意味着需要处理各种空间层次和语义粒度的复杂性。Florence-2旨在将文本提示作为任务指令，并以文本形式生成理想结果，无论是图像描述、目标检测、定位还是分割。这种多任务学习设置需要大规模、高质量的标注数据。为此，我们共同开发了FLD-5B，该数据集包含在1.26亿张图像上的54亿个综合视觉标注，采用了自动化图像标注和模型优化的迭代策略。我们采用序列到序列的结构来训练Florence-2，以执行多功能且全面的视觉任务。在众多任务上的广泛评估表明，Florence-2是一个强大的视觉基础模型竞争者，具有前所未有的零样本和微调能力。

Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks

Bin Xiao[†] Haiping Wu^{*} Weijian Xu^{*} Xiyang Dai Houdong Hu
Yumao Lu Michael Zeng Ce Liu[‡] Lu Yuan[‡]

[†]project lead ^{*}equal contribution [‡]direcional lead

Azure AI, Microsoft

Abstract

We introduce Florence-2, a novel vision foundation model with a unified, prompt-based representation for a variety of computer vision and vision-language tasks. While existing large vision models excel in transfer learning, they struggle to perform a diversity of tasks with simple instructions, a capability that implies handling the complexity of various spatial hierarchy and semantic granularity. Florence-2 was designed to take text-prompt as task instructions and generate desirable results in text forms, whether it be captioning, object detection, grounding or segmentation. This multi-task learning setup demands large-scale, high-quality annotated data. To this end, we co-developed FLD-5B that consists of 5.4 billion comprehensive visual annotations on 126 million images, using an iterative strategy of automated image annotation and model refinement. We adopted a sequence-to-sequence structure to train Florence-2 to perform versatile and comprehensive vision tasks. Extensive evaluations on numerous tasks demonstrated Florence-2 to be a strong vision foundation model contender with unprecedented zero-shot and fine-tuning capabilities.

在人工通用智能(AGI)系统领域，已显著转向利用预训练的多功能表示，这些表示因其在多样化应用中的任务无关优势而受到认可。这一趋势在自然语言处理(NLP)中尤为明显，其中先进模型[5, 6, 19, 43, 65, 66]展现了通过简单指令即可适应跨多个领域和任务的广泛知识。NLP的成功激励了计算机视觉领域采取类似的方法。

1. Introduction

In the realm of Artificial General Intelligence (AGI) systems, there has been a notable shift towards utilizing pre-trained, versatile representations, acknowledged for task-agnostic benefits across diverse applications. This trend is evident in natural language processing (NLP), where advanced models [5, 6, 19, 43, 65, 66] show adaptability with comprehensive knowledge spanning various domains and tasks with simple instructions. The success of NLP motivates a parallel approach in computer vision.

Universal representation for diverse vision-related tasks presents unique challenges, notably the need for comprehensive perceptual abilities. Unlike NLP, which deals

通用表示在多样化的视觉相关任务中提出了独特的挑战，尤其是需要全面的感知能力。与主要处理文本的自然语言处理(NLP)不同，计算机视觉需要处理复杂的视觉数据，如物体位置、掩码轮廓和属性。在计算机视觉中实现通用表示需要熟练管理一系列复杂的任务，这些任务如图1所示，以二维方式组织：

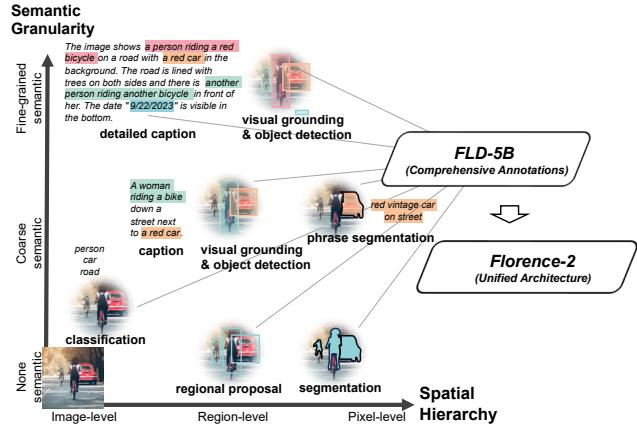


Figure 1. We aim to build a vision foundation model to enable extensive perception capabilities including spatial hierarchy and semantic granularity. To achieve this, a single unified model **Florence-2** is pre-trained on our **FLD-5B** dataset encompassing a total of 5.4B comprehensive annotations across 126M images, which are collected by our Florence data engine.

mainly with text, computer vision requires handling intricate visual data like object location, masked contours, and attributes. Attaining universal representation in computer vision demands adept management of a spectrum of complex tasks, organized two-dimensionally as illustrated in Figure 1:

- Spatial Hierarchy:** The model must discern spatial details across varying scales, understanding image-level concepts and fine-grained pixel specifics. Accommodating the intricate spatial hierarchy within vision demands the model's proficiency in handling diverse levels of granularity.
- Semantic Granularity:** Universal representation in computer vision should span a spectrum of semantic granularity. The model transitions from high-level captions to nuanced descriptions, enabling versatile understanding for diverse applications.

图1. 我们致力于构建一个视觉基础模型，以实现包括空间层次和语义粒度在内的广泛感知能力。为此，我们在FLD-5B数据集上预训练了一个统一的模型

Florence-2，该数据集包含由我们的Florence数据引擎收集的1.26亿张图像，共计54亿条全面注释

空间层次结构：模型必须能够辨别不同尺度上的空间细节，理解图像层面的概念以及细粒度的像素特性。适应视觉中复杂的空间层次结构要求模型具备处理多种粒度层次的能力

语义粒度：计算机视觉中的通用表示应涵盖一系列语义粒度。模型从高层次描述过渡到细致入微的刻画，从而实现对多样化应用场景的灵活理解

这一追求以其独特性和重大挑战为特征。一个关键障碍在于全面视觉标注的稀缺性，这阻碍了能够捕捉空间层次和语义粒度微妙差异的基础模型的开发。现有的数据集，如ImageNet [18]、COCO [48]和Flickr30k Entities [61]，专为特定应用定制，经过大量的人工标注。为了克服这一限制，必须在更大规模上为每张图像生成广泛的标注。

This pursuit is characterized by distinctiveness and substantial challenges. A key hurdle is the scarcity of *comprehensive visual annotations*, hindering the development of a foundational model capable of capturing the intricate nuances of spatial hierarchy and semantic granularity. Existing datasets, such as ImageNet [18], COCO [48], and Flickr30k Entities [61], tailored for specialized applications, are extensively labeled by humans. To overcome this constraint, it is imperative to generate extensive annotations for each image on a larger scale.

Another challenge is the absence of a *unified pre-training framework with a singular network architecture* that seamlessly integrates spatial hierarchy and semantic granularity in computer vision. Traditional models excel in tasks like object detection [26, 97], semantic segmentation [16, 82], and image captioning [45, 78] with task-specific design. However, it is essential to develop a comprehensive, unified model that is capable of adapting across various vision tasks in a task-agnostic manner, even accommodating new tasks with minimal or no task-specific fine-tuning.

The model *Florence* [95] pioneers the integration of spatial, temporal, and multi-modal aspects in computer vision through unified pre-training and network architecture. The first evolutionary version [95] excels in transfer learning via pre-training with noisy text-image pairs and task-specific fine-tuning using specialized adapters. However, it relies on large task-specific datasets and adapters, leaving gaps in addressing the above dual key challenges.

In this paper, we introduce *Florence-2*, a universal backbone achieved through multitask learning with extensive visual annotations. This results in a unified, prompt-based representation for diverse vision tasks, effectively addressing the challenges of limited comprehensive data and the absence of a unified architecture.

Multitask learning necessitates large-scale, high-quality annotated data. Our data engine, instead of relying on labor-intensive manual annotation, autonomously generates a comprehensive visual dataset called *FLD-5B*, encompassing a total of 5.4B annotations for 126M images. This engine consists of two efficient processing modules. The first module uses specialized models to collaboratively and autonomously annotate images, moving away from the traditional single and manual annotation approach. Multiple models work together to reach a consensus, reminiscent of the wisdom of crowds concept [33, 80, 89], ensuring a more reliable and unbiased image understanding. The second module iteratively refines and filters these automated annotations using well-trained foundational models.

By utilizing this extensive dataset, our model employs a sequence-to-sequence (seq2seq) architecture [17, 19, 66, 76], which integrates an image encoder and a multi-modality encoder-decoder. This design accommodates a spectrum of

通过利用这一广泛的数据集，我们的模型采用了序列到序列（seq2seq）架构 [17, 19, 66, 76]，该架构集成了图像编码器和多模态编码器-解码器。这一设计适应了多种视觉任务，无需针对特定任务进行架构修改，与自然语言处理（NLP）社区的理念一致，即在一致的底层结构下开发多功能模型。数据集FLD-5B中的所有注释均统一标准化为文本输出，便于采用统一的多任务学习方法，并使相同的损失函数进行一致的优化。其结果是产生了一个多功能的视觉基础模型——Florence-2，该模型能够在单一模型内执行多种任务，如目标检测、图像描述和定位，所有这些任务均由一组统一的参数控制。任务激活通过文本提示实现，反映了大型语言模型（LLMs）[65]所采用的方法

vision tasks without the need for task-specific architectural modifications, aligning with the ethos of the NLP community for versatile model development with a consistent underlying structure. All annotations in the dataset *FLD-5B*, are uniformly standardized into textual outputs, facilitating a unified multi-task learning approach with consistent optimization with the same loss function as the objective. The outcome is a versatile vision foundation model, *Florence-2*, capable of performing a variety of tasks, such as object detection, captioning, and grounding, all within a single model governed by a uniform set of parameters. Task activation is achieved through textual prompts, reflecting the approach used by Large Language Models (LLMs) [65].

Our approach attains a universal representation, demonstrating broad applicability across various visual tasks. Key results include:

- As a versatile vision foundation model, *Florence-2* achieves new state-of-the-art zero-shot performance in tasks such as captioning on COCO [48], visual grounding on Flickr30k [61], and referring expression comprehension on RefCOCO/+g [31, 56, 93].
- After fine-tuning with public human-annotated data, *Florence-2*, despite its compact size, competes with larger specialist models. Notably, the fine-tuned *Florence-2* establishes new state-of-the-art results on the benchmarks on RefCOCO/+g.
- The pre-trained *Florence-2* backbone enhances performance on downstream tasks, e.g. COCO object detection and instance segmentation, and ADE20K semantic segmentation, surpassing both supervised and self-supervised models. Compared to pre-trained models on ImageNet, ours improves training efficiency by 4× and achieves substantial improvements of 6.9, 5.5, and 5.9 points on COCO [48] and ADE20K [98] datasets, using Mask-RCNN [26], DINO [97], and UperNet [82] frameworks respectively.

2. Rethinking Vision Model Pre-training

In pursuit of a versatile vision foundation model, we revisit three predominant pre-training paradigms: supervised (e.g., ImageNet classification [18]), self-supervised (e.g., SimCLR [9], MoCo [25], BEiT [4], MAE [24]), and weakly supervised (e.g., CLIP [64], Florence [95], SAM [32]). Each paradigm captures unique aspects of visual data but is inherently limited by the constraints of single-task learning frameworks. Supervised pre-training excels in object recognition but lacks adaptability [38]; self-supervised algorithms reveal intricate features but may overemphasize certain attributes [8]; weakly supervised methods leverage unstructured textual annotations but yield only image-level understanding [64]. To build a unified vision foundation model suitable for various applications, we must explore

在追求一个多功能视觉基础模型的过程中，我们重新审视了三种主要的预训练范式：监督式（例如，ImageNet分类[18]）、自监督式（例如，SimCLR[9]、MoCo[25]、BEiT[4]、MAE[24]）以及弱监督式（例如，CLIP[64]、Florence[95]、SAM[32]）。每种范式都捕捉了视觉数据的独特方面，但本质上受到单任务学习框架的限制。监督式预训练在物体识别方面表现出色，但缺乏适应性[38]；自监督算法揭示了复杂的特征，但可能过度强调某些属性[8]；弱监督方法利用非结构化文本注释，但仅产生图像级别的理解[64]。为了构建一个适用于各种应用的统一视觉基础模型，我们必须探索能够克服单任务限制并整合文本和视觉语义的创新预训练策略

另一个挑战在于缺乏一种统一的预训练框架，该框架应具备单一网络架构，能够无缝整合计算机视觉中的空间层与语义粒度。传统模型在特定任务设计下，如目标检测[26, 97]、语义分割[16, 82]和图像描述[45, 78]等方面表现出色。然而，开发一个全面、统一的模型至关重要，该模型能够以任务无关的方式适应各种视觉任务，甚至能在极少或无需特定任务微调的情况下适应新任务。

*Florence*模型[95]通过统一的预训练和网络架构，率先在计算机视觉领域整合了空间、时间和多模态方面。其首个进化版本[95]凭借使用噪声文本-图像对进行预训练，并通过专用适配器进行任务特定的微调，在迁移学习方面表现出色。然而，该模型依赖于大规模任务特定数据集和适配器，在应对此类挑战方面仍存在不足。

本文介绍了*Florence-2*，一种通过多任务学习与广泛视觉注释实现的通用骨干。该网络为多种视觉任务提供了统一的、基于提示的表示，有效解决了综合数据有限及缺乏统一架构的挑战。

多任务学习需要大规模、高质量的标注数据。我们的数据引擎不依赖于劳动密集型的手动标注，而是自主生成了一个名为FLD-5B的全面视觉数据集，该数据集包含126M图像的54亿条标注。该引擎由两个高效的处理模块组成。第一个模块使用专门设计的模型协作自主地对图像进行标注，摒弃了传统的单且手动标注的方式。多个模型协同工作，达成共识，类似于群体智慧的概念[33, 80, 89]，从而确保了对图像更可靠且无偏见的理解。第二个模块则利用训练有素的基础模型对这些自动化标注进行迭代优化和筛选。

我们的方法实现了通用表征，展示了其在多种视觉任务中的广泛适用性。关键成果包括

作为一种多功能视觉基础模型，*Florence-2*在多项任务中实现了新的零样本性能突破，包括在COCO数据集上的图像描述[48]、Flickr30k数据集上的视觉定位[61]，以及RefCOCO/+g数据集上的指代表达理解。

经过使用公开的人工标注数据进行微调后，尽管*Florence-2*模型规模紧凑，却能与更大型的专业模型相媲美。尤为显著的是，微调后的*Florence-2*在RefCOCO/+g基准测试中创下了新的最先进成果。

预训练的*Florence-2*骨干网络在下游任务中显著提升了性能，例如在COCO目标检测与实例分割以及ADE20K语义分割任务上，均超越了有监督与自监督模型。与基于ImageNet预训练的模型相比，我们的模型在训练效率上提高了4倍，并在COCO [48]和ADE20K [98]数据集上，分别采用Mask-RCNN [26]、DINO [97]和UperNet [82]框架时，实现了6.9、5.5和5.9个百分点的显著提升。

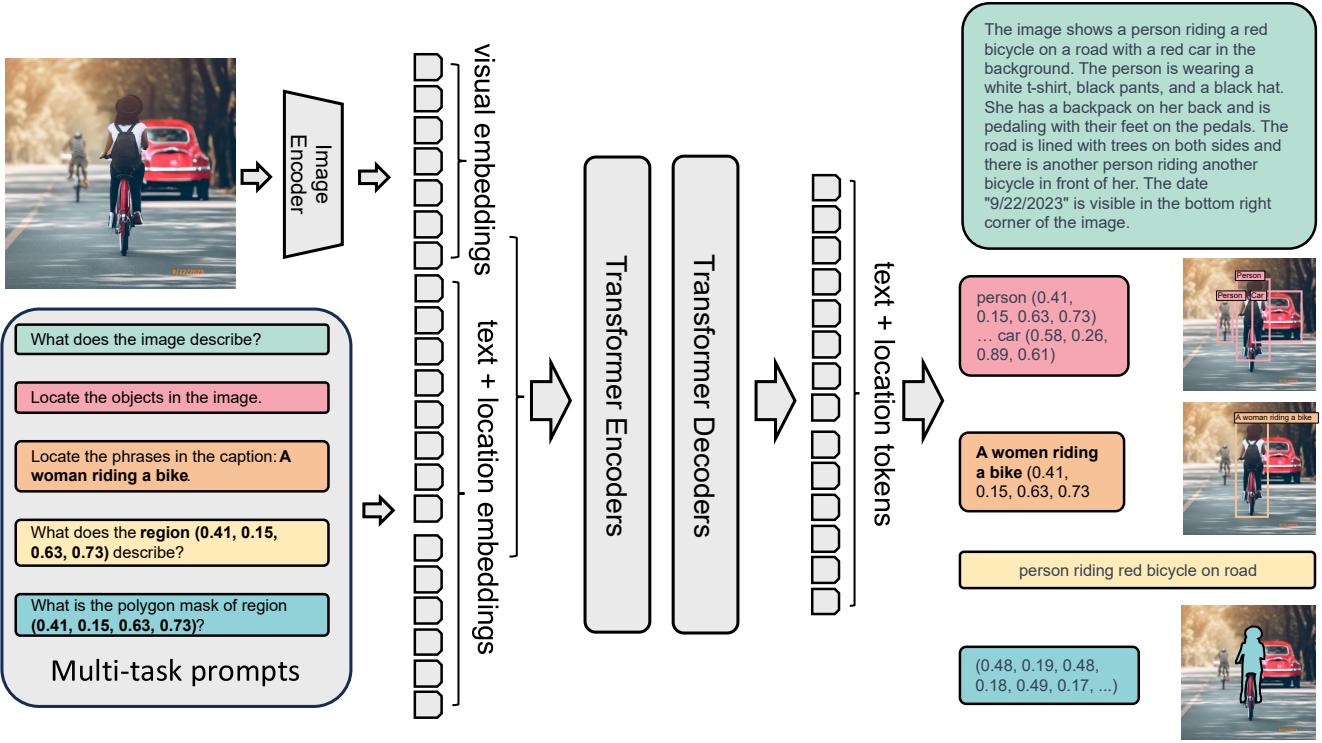


Figure 2. **Florence-2** consists of an image encoder and standard multi-modality encoder-decoder. We train **Florence-2** on our **FLD-5B** data in a unified multitask learning paradigm, resulting in a generalist vision foundation model, which can perform various vision tasks.

图2. Florence-2由图像编码器和标准的多模态编码器-解码器组成。我们在FLD-5B数据集上以统一的多任务学习范式对Florence-2进行训练，从而构建了一个通用的视觉基础模型，该模型能够执行多种视觉任务。

innovative pre-training strategies that overcome single-task limitations and integrate both textual and visual semantics.

Image understanding necessitates capturing multiple levels of granularity, from global semantics to local details, and comprehending spatial relationships between objects and entities in their semantic context. To address these core aspects of image understanding, our approach incorporates a diverse set of annotations, effectively capturing visual understanding nuances and bridging the gap between vision and language understanding.

2.1. Comprehensive Multitask Learning

To develop a versatile vision foundation model, we formulate a range of multitask learning objectives, each tailored to address specific aspects of visual comprehension. These objectives align with our predefined criteria: spatial hierarchy and semantic granularity, inspired by recent research on multitask learning [2, 12, 14, 15, 55, 79]. Our multitask learning approach incorporates three distinct learning objectives, each addressing a different level of granularity and semantic understanding:

- **Image-level understanding** tasks capture high-level semantics and foster a comprehensive understanding of images through linguistic descriptions [13, 18, 34, 91]. They enable the model to comprehend the overall

图像理解需要捕
捉从全局语义到
局部细节的多层次
粒度，并在语义背景
下理解物体与实体之
间的空间关系。为了
应对图像理解中的
这些核心问题，我们
的方法整合了多样化的
标注，有效地捕
捉了视觉理解的
细微差别，并弥
合了视觉与语言
理解之间的鸿沟。

为了开发一个
多功能的视觉基础
模型，我们制定了一
系列多任务学
习目标，每个目
标都针对视觉
理解的特定方面
进行定制。这些
目标与我们预先
设定的标准相一
致：空间层次和
语义粒度，这些
标准受到了近期
多任务学习研究
的启发
[2, 12, 14, 15, 55,
79]。我们的多
任务学习方法包
含了三个不同的
学习目标，每个
目标都针对不同
粒度和语义理解
层次：

context of an image and grasp semantic relationships and contextual nuances in the language domain. Exemplar tasks include image classification, captioning, and visual question answering.

• **Region/pixel-level recognition** tasks facilitate detailed object and entity localization within images, capturing relationships between objects and their spatial context. Tasks include object detection, segmentation, and referring expression comprehension.

区域/像素级识别任务有助于在图像中详细定位对象和实体，捕捉对象之间的关系及其空间上下文。这些任务包括目标检测、分割以及指代表达理解。

• **Fine-grained visual-semantic alignment** tasks require fine-grained understanding of both text and image. It involves locating the image regions that correspond to the text phrases, such as objects, attributes, or relations. These tasks challenge the ability to capture the local details of visual entities and their semantic contexts, as well as the interactions between textual and visual elements.

By combining these three learning objectives in a multitask learning framework, our foundation model learns to handle different levels of detail and semantic understanding. This strategic alignment enables our model to deal with various spatial details, distinguish levels of detail in understanding, and go beyond surface-level recognition—ultimately learning a universal representation for vision understanding.

通过将这三个学习目标整合到一个多任务学习框架中，我们的基础模型学会了处理不同层次的细节和语义理解。这种策略性的对齐使我们的模型能够应对各种空间细节，区分理解中的细节层次，并超越表面层次的识别——最终学习到一种通用的视觉理解表示。

我们介绍了基础模型Florence-2，该模型专为通用表示学习而设计，能够通过单一权重集和统一架构处理多种视觉任务。如图2所示，Florence-2采用序列到序列的学习范式[77]，将所有任务（在第2节中描述）整合到一个共同的语言建模目标下。该模型接收图像及任务提示作为任务指令，并以文本形式生成所需结果。它使用视觉编码器将图像转换为视觉标记嵌入，随后与文本嵌入拼接，并通过基于Transformer的多模态编码器-解码器进行处理以生成响应。在接下来的章节中，我们将详细解释每个模型组件

3. Model

We present the foundation model *Florence-2*, designed for universal representation learning, capable of handling various vision tasks with a single set of weights and a unified architecture. As depicted in Figure 2, *Florence-2* employs a sequence-to-sequence learning paradigm [77], integrating all tasks, described in Section 2, under a common language modeling objective. The model takes images coupled with task-prompt as task instructions, and generates the desirable results in text forms. It uses a vision encoder to convert images into visual token embeddings, which are then concatenated with text embeddings and processed by a transformer-based multi-modal encoder-decoder to generate the response. In the following sections, we will provide a detailed explanation of each model component.

任务表述。我们采用序列到序列框架
[10, 15, 55, 77]，以统一的方式处理各种视觉任务。如表13所示，我们将每个任务表述为一个翻译问题：给定输入图像和任务特定的提示，它们生成相应的输出响应。根据任务的不同，提示和响应可以是文本或区域：

文本：当提示或答案为纯文本且无特殊格式时，我们将其保留在最终的序列到序列格式中

区域：针对特定区域的任务，我们在分词器的词汇表中添加位置标记，用以表示量化后的坐标。我们创建了1000个区间，类似于[10, 11, 55, 79]，并根据任务需求采用定制化的格式来表示区域：

框表示法(x₀, y₀, x₁, y₁)：应用于目标检测和密集区域描述等任务中，其中位置标记对应于框的坐标。这些位置标记是框的左上角和右下角的坐标

四边形框表示法(x₀, y₀, ..., x₃, y₃)：在文本检测与识别任务中，采用位置标记来标识包围文本的四边形各顶点坐标。这些位置标记依次为四边形框四个角的坐标，从左上角开始，按顺时针方向排列

多边形表示法(x₀, y₀, ..., x_n, y_n)：针对分割任务，采用位置标记来代表多边形的顶点。这些位置标记按顺时针顺序排列，表示多边形各顶点的坐标

Task formulation. We adopt a sequence-to-sequence framework [10, 15, 55, 77] to address various vision tasks in a unified manner. As shown in Table 13, we formulate each task as a translation problem: Given an input image and a task-specific prompt, we generate the corresponding output response. Depending on the task, the prompt and response can be either text or region:

- **Text:** When the prompt or answer is plain text without special formatting, we maintain it in our final sequence-to-sequence format.
- **Region:** For region-specific tasks, we add location tokens to the tokenizer’s vocabulary list, representing quantized coordinates. We create 1,000 bins, similar to [10, 11, 55, 79], and represent regions using formats tailored to task requirements:
 - **Box representation** (x_0, y_0, x_1, y_1): Utilized in tasks such as object detection and dense region captioning, with location tokens corresponding to the box coordinates. The location tokens are the coordinates of the top-left and bottom-right corners of the box.
 - **Quad box representation** ($x_0, y_0, \dots, x_3, y_3$): For text detection and recognition tasks, using location tokens for each coordinate of the quadrilateral enclosing the text. The location tokens are the coordinates of each corner of the quad box, starting from the top-left and going clockwise.
 - **Polygon Representation** ($x_0, y_0, \dots, x_n, y_n$): For referring segmentation tasks, with location tokens representing the vertices of the polygon. The location tokens are the coordinates of the vertices of the polygon, in clockwise order.

By extending the tokenizer’s vocabulary to include location tokens, we enable the model to process region-specific tasks. 我们通过扩展分词器的词汇表以包含位置标记，使模型能够以统一的学习格式处理区域特定信息。这消除了为不同任务设计特定任务头部的需求，并允许采用更加以数据为中心的方法

information in a unified learning format. This eliminates the need to design task-specific heads for different tasks and allows for a more data-centric approach.

Vision encoder. We employ DaViT [20] as the vision encoder. It processes an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ (with H and W denoting height and width, respectively) into flattened visual token embeddings $\mathbf{V} \in \mathbb{R}^{N_v \times D_v}$, where N_v and D_v represent the number and dimensionality of vision tokens, respectively.

Multi-modality encoder decoder. We use a standard encoder-decoder transformer architecture to process visual and language token embeddings. We first obtain prompt text embeddings $\mathbf{T}_{prompt} \in \mathbb{R}^{N_t \times D}$ using our extended language tokenizer and word embedding layer [43]. Then, we concatenate vision token embeddings with prompt embeddings to form the multi-modality encoder module input, $\mathbf{X} = [\mathbf{V}', \mathbf{T}_{prompt}]$, where $\mathbf{V}' \in \mathbb{R}^{N_v \times D}$ is obtained by applying a linear projection and LayerNorm layer [3] to \mathbf{V} for dimensionality alignment.

Optimization objective. Given the input x combined from the image and the prompt, and the target y , we use the standard language modeling with cross-entropy loss for all the tasks.

$$\mathcal{L} = - \sum_{i=1}^{|y|} \log P_\theta(y_i | y_{<i}, x), \quad (1)$$

where θ are the network parameters, $|y|$ is the number of target tokens.

视觉编码器。我们采用DaViT [20]作为视觉编码器。它将输入图像 $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ （其中 H 和 W 分别表示高度和宽度）处理为扁平化的视觉标记嵌入 $\mathbf{V} \in \mathbb{R}^{N_v \times D_v}$ ，其中 N_v 和 D_v 分别表示视觉标记的数量和维度

多模态编码器-解码器。我们采用标准的编码器-解码器。Transformer架构来处理视觉与语言标记嵌入。首先，利用我们扩展的语言标记化器及词嵌入层[43]，获取提示文本嵌入 $\mathbf{T}_{prompt} \in \mathbb{R}^{N_t \times D}$ 。随后，将视觉标记嵌入与提示嵌入拼接，形成多模态编码器模块的输入 $\mathbf{X} = [\mathbf{V}', \mathbf{T}_{prompt}]$ ，其中 $\mathbf{V}' \in \mathbb{R}^{N_v \times D}$ 是通过对 \mathbf{V} 应用线性投影及LayerNorm层[3]进行维度对齐后得到的

优化目标。给定由图像和提示组合而成的输入 x 以及目标 y ，我们对所有任务采用标准的语言建模方法，并利用交叉熵损失函数进行优化

为了训练我们的Florence-2模型，我们需要一个全面、大规模、高质量的多任务数据集，涵盖图像数据的各个方面。鉴于此类数据的稀缺性，我们开发了一个新的多任务图像数据集。该数据集FLD-5B包含1.26亿张图像、5亿条文本注释、13亿条文本区域注释以及36亿条文本短语-区域注释，覆盖了不同的任务。我们详细阐述了数据收集和注释过程，包括对各种注释类型的适应性调整。图3所示的数据引擎管道将在后续章节中讨论

4. Data Engine

To train our *Florence-2* model, we require a comprehensive, large-scale, high-quality multitask dataset encompassing various image data aspects. Given the scarcity of such data, we have developed a new multitask image dataset. This dataset FLD-5B includes **126M** images, **500M** text annotations, and **1.3B** text-region annotations, and **3.6B** text-phrase-region annotations across different tasks. We extensively explain our data collection and annotation procedures, encompassing adaptations for various annotation types. The data engine pipeline, shown in Figure 3, will be discussed in subsequent sections.

4.1. Image Collection

We construct our data by gathering a diverse collection of images from various sources. We begin with the identification of three key tasks that act as primary sources for our image corpus: image classification, object detection, and image captioning. Consequently, we curate and combine five distinct datasets originating from the aforementioned tasks: ImageNet-22k [18], Object 365 [70], Open Images [40], Conceptual Captions [71], and LAION [68]

我们通过从各种来源收集各种图像来构建我们的数据。首先，我们确定了三个作为我们图像语料库主要来源的关键任务：图像分类、目标检测和图像描述。因此，我们策划并整合了五个来自上述任务的独特数据集：ImageNet-22k [18]、Object 365 [70]、Open Images [40]、Conceptual Captions [71]和通过[45]筛选的LAION [68]。这种组合共得到一个包含1.26亿张图像的数据集。

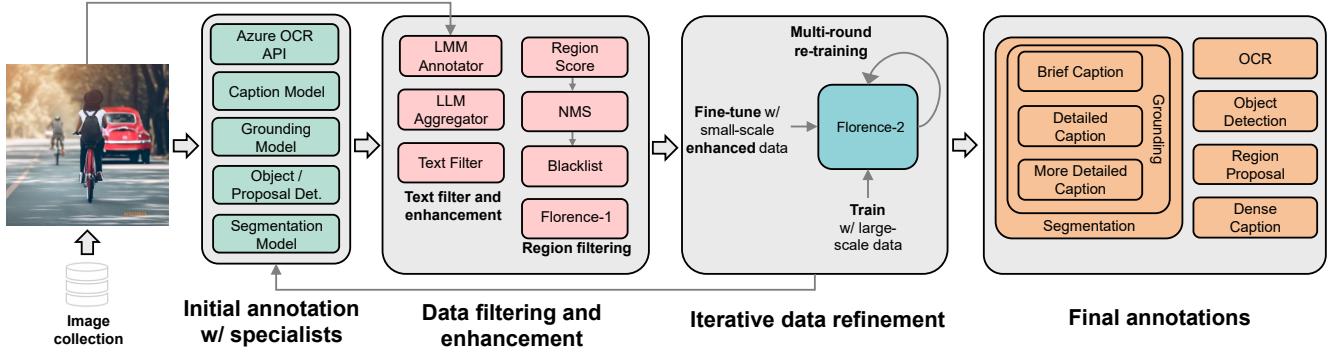


Figure 3. **Florence-2** data engine consists of three essential phases: (1) initial annotation employing specialist models, (2) data filtering to correct errors and remove irrelevant annotations, and (3) an iterative process for data refinement. Our final dataset (**FLD-5B**) of over **5B** annotations contains **126M** images, **500M** text annotations, **1.3B** region-text annotations, and **3.6B** text-phrase-region annotations.

图3. Florence-2数据引擎由三个关键阶段构成：(1) 利用专业模型进行初始标注，(2) 数据过滤以纠正错误并移除无关标注，(3) 数据精炼的迭代过程。我们的最终数据集（FLD-5B）包含超过50亿条标注，涵盖1.26亿张图像、5亿条文本标注、13亿条区域-文本标注及36亿条文本-短语-区域标注

我们的主要目标是生成能够有效支持多任务学习的全面注释。因此，我们的注释工作覆盖了一系列广泛的任务，这些任务被归纳为三个独立的注释类别：文本、区域-文本对以及文本-短语-区域三元组，如图4所示。数据注释工作流程包括三个关键阶段，每个阶段都确保了注释的准确性和质量：(1) 使用专业模型进行初步注释，(2) 数据过滤以纠正错误并去除非相关的注释，以及(3) 数据精炼的迭代过程。

filtered by [45]. This combination results in a dataset of 126 million images in total.

4.2. Data Annotation

Our primary objective is to generate comprehensive annotations that can support multitask learning effectively. Accordingly, our annotation endeavors span a comprehensive range of tasks, encapsulated within three discrete annotation categories: *text*, *region-text* pairs, and *text-phrase-region* triplets, which is illustrated in Figure 4. The data annotation workflow consists of three essential phases, each of which ensures the accuracy and quality of the annotations: (1) initial annotation employing specialist models, (2) data filtering to correct errors and remove irrelevant annotations, and (3) an iterative process for data refinement.

Initial annotation with specialist models. To initiate the annotation process for each annotation type, we employ synthetic labels obtained from specialist models. These specialist models are a combination of offline models trained on a diverse range of publicly available datasets and online services hosted on cloud platforms. They are specifically tailored to excel in annotating their respective annotation types.

It is worth noting that certain image datasets may already contain partial annotations for some annotation types. For instance, the Object 365 [70] dataset already includes human-annotated bounding boxes and corresponding categories as region-text annotations. In such cases, we merge the pre-existing annotations with the synthetic labels generated by the specialist models. This approach enhances the coverage and diversity of the annotations.

Moreover, specific annotations, such as detailed descriptions in the text annotation type, are represented by datasets of a considerably small size. This inherently poses challenges in obtaining high-performance specialist models. Consequently, we opt to omit these tasks during the initial annotation phase. Annotations for these tasks are generated

此外，特定类型的注释，如文本注释类型中的详细描述，其数据集规模相对较小。这本质上对高性能的专用模型构成了挑战。因此，在初始注释阶段，我们选择省略这些任务。这些任务的注释将在后续的数据迭代精炼过程中生成

later during the iterative data refinement process.

In summation, through the rigorous initial annotation procedures, we ensure that the aggregated dataset of 126 million images is comprehensively labeled across the majority of annotation types.

Data filtering and enhancement. The initial annotations obtained from the specialist models, while comprehensive, are susceptible to noise and imprecision. In response to this challenge, we have implemented a multifaceted filtering process to refine and eliminate undesired annotations. Our general filtering protocol mainly focuses on two data types in the annotations: text and region data.

首先，关于文本注释方面，我们受到DiHT [63]和开发基于SpaCy [28]的解析工具的启发。我们过滤掉包含过多对象的文本，因为这些文本往往会产生噪音，且可能无法准确反映对应图像中的实际内容。此外，我们通过测量依赖解析树中节点的程度来评估动作和对象的复杂性。我们保留具有一定最低动作和对象复杂性的文本，以确保图像中视觉概念的丰富性。

其次，关于区域标注，特别是边界框，我们在置信度分数阈值下移除了噪音框。作为补充，我们还采用了非极大值抑制技术以减少冗余或重叠的边界框。

Iterative data refinement. Using our filtered initial annotations, we trained a multitask model that processes sequences of data. Upon evaluating this model against our training images, we discerned a marked enhancement in its predictions, particularly in instances where original labels were marred by inaccuracies or extraneous noise, such as in alt-texts. Motivated by these findings, we integrated these updated annotations with our original ones and subjected the model to another training iteration. This cyclical re-training process利用我们筛选后的初始标注，我们训练了一个处理数据序列的多任务模型。在针对训练图像评估该模型时，我们发现其预测结果有了显著提升，尤其是在原始标签因不准确或额外噪声（如替代文本中的噪声）而受损的情况下。受这些发现启发，我们将这些更新后的标注与原始标注整合，并让模型进行了新一轮的训练。这一循环精炼过程逐步提升了我们训练数据集的质量

对于因数据不足而未能直接训练出稳健专家模型的任务，我们采
用了迭代训练模
型进行预训练。
随后，利用稀疏
数据集对这一预
训练模型进行微
调，结果显示其
性能优于在相
同数据上从头开始
训练的模型。因此，
我们利用这
一微调后的模型
作为专家，对包
含1.26亿张图像
的庞大数据集进
行标注，确保了
标注的全面覆盖

inement process incrementally improves the quality of our training dataset.

In the case of tasks we initially bypassed due to insufficient data for the training of a robust specialist model, we leveraged the iteratively trained model for pre-training purposes. Subsequent fine-tuning of this pre-trained model with the sparse dataset showcased superior performance compared to a model trained from scratch on the same data. Thus, we harness the fine-tuned model as a specialist for annotating our expansive dataset comprising 126 million images, ensuring comprehensive annotation coverage.

4.3. Annotation-specific Variations

In Section 4.2, we introduce our general annotation workflow. This section delves into each annotation type and the corresponding variations of the annotation procedure.

Text. Text annotations categorize images using three types of granularities: brief, detailed, and more detailed. The brief text includes only one sentence that demonstrates the most salient objects and activities, which is similar to COCO caption [13]. In contrast, the detailed text and more detailed text contain multiple sentences that describe the image with richer objects, attributes, and actions.

For the brief text, a *Florence-2* model is trained as the specialist on publicly available image caption and image-text datasets, creating an image-to-text model for initial annotations. Iterative refinement is used to minimize noise in these texts. For the detailed text, prompts including existing image annotations like the brief text and region-text annotations, are fed to large language models (LLMs) or large multimodal models (LMMs) to generate comprehensive descriptions. Due to the high cost of the large models, only a small set of detailed text and more detailed text are generated. These are used to fine-tune the caption specialist, developing a detailed description specialist for further annotations.

Region-text pairs. The region-text pairs provide descriptive textual annotation for semantic regions in the image. Semantic regions include regions of visual objects as well as text regions. The region is represented by a tight bounding box surrounds the region. Moreover, each region can be annotated with varying degrees of granularity, including phrases and sentences, that contribute to a richer understanding of the region.

Region-text pairs are annotated differently for text regions and visual object regions. Text regions are labeled using Azure AI Services' OCR API [1], while visual objects are initially annotated with a DINO object detector [97] trained on public datasets. Data filtering, including confidence thresholding and non-maximum suppression, removes noisy boxes. Textual annotations for the visual object regions are further enriched by brief text generated from an image-to-text model with cropped image regions. Each

区域-文本对的标注方式因文本区域和视觉对象区域而异。文本区域采用 Azure AI服务的OCR API[1]进行标注，而视觉对象则首先通过基于公开数据集训练的DINO目标检测器[97]进行初步标注。通过数据过滤，包括置信度阈值设定和非极大值抑制，去除了噪声框。视觉对象区域的文本注释通过图像到文本模型生成的简短文本进一步丰富，这些文本基于裁剪后的图像区域生成。每个区域随后获得三种文本注释：来自对象类别的短语、简短文本以及从简短文本中提取的名词词组块。*Florence-1*模型[95]负责确定与每个图像区域最相似的文本注释。

region then receives three textual annotations: phrase from object category, brief text, and noun phrase chunks from the brief text. The *Florence-1* [95] model determines the most similar textual annotation to each image region.

Text-phrase-region triplets. Text-phrase-region triplets consist of a descriptive text of the image, noun phrases in this text related to image objects, and region annotations for these objects. The text includes brief, detailed, and more detailed text generated earlier. For each text, the Grounding DINO model [50] identifies noun phrases and creates bounding boxes for them. Additionally, the SAM model [32] generates segmentation masks for each box, offering more precise object localization. During data filtering, a confidence score threshold is applied to both noun phrases and bounding boxes to ensure relevance. A black-list is also used to exclude irrelevant noun phrases like pronouns and abstract concepts.

5. Dataset

This section introduces the statistics and analysis of *FLD-5B* that we built using the data engine in Section 4. We begin with an overview of the dataset and compare it with the recent works. We then show further analyses of detailed annotation statistics, semantic coverage and spatial coverage in the established dataset.

5.1. Overview

Following the data engine, we build a large-scale training set (*FLD-5B*) of 126M images, more than **500M** text annotations, **1.3B** region-text annotations, and **3.6B** text-phrase-region annotations. Each image is annotated with text, region-text pairs, and text-phrase-region triplets and each annotation type has multiple instances varying in diverse granularity. An illustrative example of an image and its corresponding annotations can be found in Figure 4.

We provide a comparison between our data set and the existing data sets that are commonly used for training foundation models in Table 1. Our data set has several advantages over the previous ones, such as having more annotations in total and per image. Moreover, the annotations in our data set span multiple levels of spatial and semantic granularity, which allows for more diverse and comprehensive visual understanding tasks.

5.2. Data Analysis

Annotation statistics. The statistics for each annotation type within our dataset are presented in Table 2.

Firstly, we have around **500M** text annotations, including brief, detailed, and more detailed texts with different lengths. It is noteworthy that our detailed and more detailed text has 4x and 9x number of tokens compared with the brief text that is similar to COCO captions [13]. These lengthy

首先，我们拥有大约5亿条文本注释，包括简短、详细以及更详细的文本，其长度各不相同。值得注意的是，与类似COCO描述[13]的简短文本相比，我们的详细文本和更详细的文本的标记（词元）数量分别多4倍和9倍。这些较长的注释为全面的视觉理解提供了更为丰富的信息。

文本-短语-区域三元组。文本-短语-区域三元组由图像的描述性文本、与该文本中与图像对象相关的名词短语以及这些对象的区域标注组成。文本包括先前生成的简短、详细和更详细的描述，对于每个文本，Grounding DINO模型[50]识别名词短语并为其创建边界框。此外，SAM模型[32]为每个边界框生成分割掩码，提供更精确的对象定位。在数据过滤过程中，对名词短语和边界框应使用置信度分数阈值以确保相关性。还使用黑名单来排除不相关的名词短语，如代词和抽象概念

本节介绍了我们利用第4节中的数据引擎构建的*FLD-5B*的统计与分析。首先，我们对数据集进行了概述，并将其与近期研究成果进行了比较。随后，我们对已建立数据集中的详细注释统计、语义覆盖范围及空间覆盖范围进行了进一步的分析展示

依据数据引擎，我们构建了一个大规模训练集(*FLD-5B*)，包含1.26亿张图片，超过5亿条文本标注、13亿个区域-文本标注以及36亿个文本短语-区域标注三元组。每张图片均配以文本、区域-文本对及文本短语-区域三元组标注，且每种标注类型均包含多种粒度不同的实例。图4展示了一张图片及其对应标注的示例

我们在表1中提供了我们的数据集与常用于训练基础模型的现有数据集之间的比较。我们的数据集相较于之前的数据集具有多项优势，例如总注释数量及每张图像的注释数量更多。此外，我们数据集中的注释涵盖了空间和语义粒度上的多个层次，这使得其能够支持更为多样化和全面的视觉理解任务

注释统计。表2展示了我们数据集中每种注释类型的统计信息。

图4. FLD-5B数据集中图像及其对应标注的示例。FLD-5B中的每张图像均由Florence数据引擎标注，包含文本-区域-文本对及文本-短语-区域三元组，这些标注覆盖了多层次的空间结构、从简略到详细的渐进粒度以及广泛的语义范围，从而能够从多角度实现更为全面的视觉理解



Figure 4. An illustrative example of an image and its corresponding annotations in *FLD-5B* dataset. Each image in *FLD-5B* is annotated with text, region-text pairs, and text-phrase-region triplets by Florence data engine, which covers multiple spatial hierarchies, brief-to-detailed progressive granularity, and a wide semantics spectrum, enabling more comprehensive visual understanding from diverse perspectives.

Dataset	Rep. Model	#Images	#Annotations	Spatial hierarchy	Semantics granularity
JFT300M [21]	ViT	300M	300M	Image-level	Coarse
WIT [64]	CLIP	400M	400M	Image-level	Coarse
SA-1B [32]	SAM	11M	1B	Region-level	Non-semantic
GrIT [60]	Kosmos-2	91M	137M	Image & Region-level	Fine-grained
M3W [2]	Flamingo	185M	43.3M*	Multi-image-level	Fine-grained
FLD-5B (ours)	Florence-2 (ours)	126M	5B	Image & Region-level	Coarse to fine-grained

Table 1. Comparison with datasets in vision foundation model training. *Flamingo's annotations are counted in the number of documents, where each document may have multiple images.

表1. 与视觉基础模型训练数据集的比较。*Flamingo的标注以文档数量计，其中每个文档可能包含多张图像

annotations provide much richer information for comprehensive visual understanding.

In addition, our dataset has around **1.3B** region-text annotations, which is more than 30x larger than the academic object detection datasets such as OpenImages [40] and Object 365 [70]. On average, each image has around 5 regions, and each region is annotated with either a phrase or a relatively longer brief text. Note that the regional brief text (2.55 avg tokens) is shorter than typical brief text annotation (7.95 avg tokens), as the regional brief text annotation actually includes a mixture of phrase, noun chunks, and brief

此外，我们的数据集包含约13亿个区域-文本注释，这一数量是诸如OpenImages [40]和Object 365 [70]等学术目标检测数据集的30倍以上。平均而言，每张图像包含约5个区域，每个区域均标注有短语或相对较长的简短文本。值得注意的是，区域简短文本（平均2.55个词元）比典型的简短文本注释（平均7.95个词元）要短，因为区域简短文本注释实际上包含了基于Florence-1评分的短语、名词块和简短文本的混合。更多细节可参见第4.3节——区域-文本对

text based on the Florence-1 score. More details can be found from Section 4.3 - region-text pairs.

Moreover, we collect text-phrase-region annotations that include more than **3.6B** phrase-region pairs for the **500M** text annotations. Specifically, the brief text annotation has 4.27 average phrase-region pairs, while detailed and more detailed text annotation has more than 10 pairs, indicating that the richer text annotation covers more objects and their corresponding phrases in the text.

此外，我们收集了文本-短语-区域标注，其中包含超过36亿个短语-区域对，对应于5亿条文本标注。具体而言，简短的文本标注平均包含4.27个短语-区域对，而详细和更为详细的文本标注则拥有超过10对，这表明更丰富的文本标注涵盖了文本中更多的对象及其对应的短语。

Semantic coverage. Our text annotations comprise various text types, addressing different levels of detail. To assess

语义覆盖范围，我们的文本注释包含各种文本类型，涉及不同的细节层次。为了评估语义覆盖范围，我们采用SpaCy [28]进行标记化和解析，这一做法受到DiHT [63]的启发。该过程产生词性(POS)标签以及标记之间的依存关系解析树。我们基于词性标签建立启发式规则，将标记分类为语义元素类型，例如对象、属性、动作和专有名词。此外，我们引入标记复杂性的概念，通过将该标记在依存关系解析树（视为无向图时）中的总度数进行测量。这种复杂性反映语义联系的丰富程度。在我们的研究中，我们重点关注对象和动作的复杂性测量。

Annotation Type	Text Type	#Image Annotations	#Avg Tokens	#Regions	#Avg Regions	#Avg Regional Tokens
Text	Brief	235M	7.95	-	-	-
	Detailed	126M	31.65	-	-	-
	More detailed	126M	70.53	-	-	-
Region-Text	Phrase	126M	-	681M	5.42	1.19
	Brief	126M	-	681M	5.42	2.55
Text-Phrase-Region	Brief	235M	7.95	1007M	4.27	1.93
	Detailed	126M	31.65	1289M	10.25	1.49
	More detailed	126M	70.53	1278M	10.17	1.35

Table 2. Annotation statistics of *FLD-5B* dataset.
表2. FLD-5B数据集的注释统计

semantic coverage, we employ SpaCy [28] for tokenization and parsing, inspired by DiHT [63]. This process yields part-of-speech (POS) tags and the dependency parsing tree among tokens. We establish heuristic rules based on POS tags, categorizing tokens into semantic element types, *e.g.*, objects, attributes, actions, and proper nouns. Additionally, we introduce the concept of *token complexity*, measured by the total degrees of the token in the dependency parsing tree when treated as an undirected graph. This complexity reflects the richness of semantic connections. In our study, we focus on measuring the complexity of objects and actions.

表3展示了关于语义要素平均数量及其对应复杂度的统计数据。结果表明，随着文本注释中细节的增加，所有测量值均呈现上升趋势。值得注意的是，平均动作数量增幅最为显著，详细与更详细的文本相较于简略文本分别增长了7倍和15倍。这一现象凸显了传统简略文本注释在描述图像动作方面的局限性。相对而言，专有名词的增长较为有限，这可能是由于专家在描述对象时更倾向于使用概括性词汇而非具体的专有名词。在复杂度测量方面，无论是对对象还是动作，在详细文本注释中都显示出更多的语义联系。动作复杂度的提升尤为明显，这与我们观察到的动作数量增加的趋势相吻合。

Table 3 presents the statistics on the average number of semantic elements and their corresponding complexity. The results indicate that all measurements increase with the inclusion of more details in text annotations. Notably, average actions experience the most significant boost, with detailed and more detailed text exhibiting $7\times$ and $15\times$ increases, respectively, compared to brief text. This highlights the limitations of traditional brief text annotations in describing image actions. Conversely, the increment in proper nouns is relatively low, potentially because specialists often describe objects more generally than using specific proper nouns. In terms of complexity measurements, both objects and actions show more semantic connections in detailed text annotations. The complexity of actions exhibits a higher improvement, aligning with our observation of the increasing number of actions.

Spatial coverage. Our region-text and text-phrase-region annotations, represented by bounding boxes and masks, capture the location of visual concepts within images. The distribution of box areas, as shown in Figure 5a, reveals more small boxes in region-text pairs and a uniform box size distribution in text-phrase-region triplets. This difference stems from the divergent origins of these boxes: object detectors for region-text pairs and a grounding model for text-phrase-region triplets, which aligns boxes to textual phrases representing both localized and overarching

空间覆盖范围。我们的区域-文本和文本-短语-区域注释，由边界框和掩码表示，捕捉图像中视觉概念的位置。如图5a所示，框面积的分布显示，在区域-文本对中有更多的小框，而在文本-短语-区域三元组中框大小分布均匀。这种差异源于这些框的不同起源：区域-文本对的物体检测器和文本-短语-区域三元组的基础模型，该模型将框与表示局部和整体图像概念的文本-短语对齐。在图5b中，展示了宽高比的对数网格分布。区域-文本对和文本-短语-区域三元组呈现出类似的对称分布，涵盖了广泛的宽高比范围。如图5c和5d所示，每种注释类型的框中心热图表明存在中心偏差，区域-文本对比文本-短语-区域三元组显示出更均匀的分布。

Text Type	Brief	Detailed	More detailed
#Image Annotations	235M	126M	126M
#Avg Tokens	7.95	31.65	70.53
#Avg Objects	3.23	13.31	28.06
#Avg Attributes	2.80	7.27	16.25
#Avg Actions	0.58	4.21	8.76
#Proper Nouns	1.10	2.40	2.41
Avg Object Complexity	2.80	4.00	4.02
Avg Action Complexity	1.14	3.63	4.38

Table 3. Statistics of the average number of semantic elements and corresponding complexity in *FLD-5B* dataset.

表3. FLD-5B数据集中语义元素平均数量及其对应复杂度的统计信息

image concepts. In Figure 5b, the log-format distribution of aspect ratios is illustrated. Region-text pairs and text-phrase-region triplets exhibit similar symmetric distributions, covering a wide range of aspect ratios. Heatmaps of the box center for each annotation type, shown in Figures. 5c and 5d, indicate a center bias, with region-text pairs displaying a more uniform distribution than text-phrase-region triplets.

6. Experiments

Our *Florence-2* models are trained on *FLD-5B* to learn a universal image representation. We conduct our experiments in three main parts: (1) We evaluate the *zero-shot* performance of our method on various tasks to show its inherent ability to handle multiple tasks without any extra fine-tuning on task-specific data using *one single generalist* model. (2) We show the adaptability of our method by further training *one single generalist* model with additional supervised data on a wide range of tasks, achieving competitive state-of-the-art performance. (3) We examine the performance of the learned visual representation on the downstream tasks as the backbone to show the superiority of our pre-training method over previous approaches.

我们的Florence-2模型在FLD-5B数据集上进行了训练，以学习一种通用的图像表示方法。我们的实验主要分为三个部分：(1) 我们评估了该方法在各种任务上的零样本性能，以展示其内在能力，即无需针对特定任务数据进行额外微调，仅使用一个通用模型即可处理多种任务。(2) 我们通过在一个通用模型上进一步使用大量监督数据进行训练，展示了该方法的适应性，并在广泛的任务上实现了具有竞争力的最先进性能。(3) 我们检验了所学习视觉表示在下游任务中作为骨干网络的性能，以展示我们的预训练方法相较于先前方法的优越性。

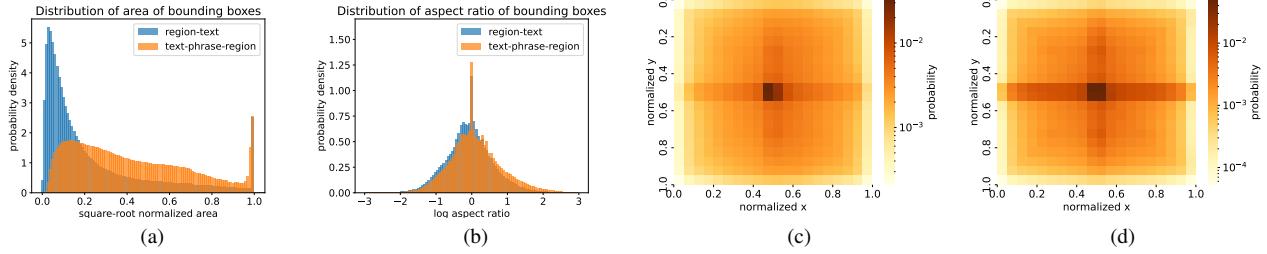


Figure 5. Distributions of bounding boxes in FLD-5B dataset.
图5. FLD-5B数据集中边界框的分布情况

6.1. Setup 我们研究了两种不同规模的模型变体：Florence-2-B模型，拥有2.32亿参数，以及Florence-2-L模型，拥有7.71亿参数。各模型的详细架构见表15。我们分别从UniCL [87]和BART [43]初始化图像编码器和多模态编码器-解码器的权重。

We investigate two model variants with different sizes: Florence-2-B model with 232 million parameters and Florence-2-L model with 771 million parameters. The detailed architectures of each model are given in Table 15. We initialize the weights of the image encoder and multi-modality encoder-decoder from UniCL [87] and BART [43], respectively.

We adopt AdamW [54] with cosine learning rate decay [53] for training our models. We leverage DeepSpeed [67] and mixed precision to improve the training efficiency. The maximum learning rate is set at $1e-4$ for the base model and $1e-5$ for the large model. A linear warm-up to the maximum learning rate is applied during the first 5,000 optimization steps.

We train our models with a mini-batch size of 2048/3072 (base/large) and an image size of 384×384 until reaching 3 billion effective training samples. Similar to [15, 29, 64, 92, 95], we further conduct high-resolution tuning with an image size of 768×768 for 0.5 billion samples for the base model and 0.1 billion samples for the large model.

6.2. Zero-shot Evaluation Across Tasks

We present a powerful vision foundation model that does not require task-specific supervised annotations for finetuning. The **zero-shot** performance of our model is shown in Table 4. For image-level tasks, Florence-2-L achieves a 135.6 CIDEr score on the COCO caption benchmark [48], utilizing less than 1% of the parameters compared to the 80B Flamingo [2] model (which has an 84.3 CIDEr score). For region-level grounding and referring expression comprehension tasks, Florence-2-L establishes a new record in zero-shot performance achieving a 5.7 improvement in Flickr30k [61] Recall@1, and approximately 4%, 8%, and 8% absolute improvements on Refcoco, Refcoco+, and Refcocog [94], respectively, compared to the Kosmos-2 [60] model, which has 1.6B parameters. Additionally, our pre-trained model attains a 35.8% mIOU in the Refcoco referring expression segmentation (RES) [94] task, a capability not supported by prior foundation models.

我们提出了一种强大的视觉基础模型，该模型无需针对特定任务的监督注释进行微调。我们模型的零样本性能如表4所示。对于图像级任务，Florence-2-L在COCO字幕基准测试[48]中获得了135.6的CIDEr分数，与拥有800亿参数的Flamingo模型[2]（其CIDEr分数为84.3）相比，使用的参数不到1%。对于区域级定位和指代表达理解任务，Florence-2-L在零样本性能上创下了新纪录，在Flickr30k[61]的Recall@1上提升了5.7%，与拥有16亿参数的Kosmos-2模型[60]相比，在Refcoco、Refcoco+和Refcocog[94]上分别实现了约4%、8%和8%的绝对提升。此外，我们的预训练模型在Refcoco指代表达分割（RES）[94]任务中获得了35.8%的mIOU，这是先前的基础模型所不具备的能力。

6.3. Generalist Model with Public Supervised Data

We demonstrate the versatility and effectiveness of our model as a vision foundation that can be transferred to various downstream tasks. We fine-tune Florence-2 models by adding a collection of public datasets that cover image-level, region-level, pixel-level tasks, yielding one generalist model for various vision tasks. The details of the dataset collection are provided in Appendix B. Tables 5 and 6 compare our model with other state-of-the-art models. Our key findings are:

Simple design for strong performance. Florence-2 demonstrates *strong* performance with *standard* multi-modality Transformer encoder-decoder without special designs, particularly for region-level and pixel-level tasks. For example, Florence-2-L outperforms PolyFormer [49] on both RefCOCO REC task and RES task by 3.0 Accuracy@0.5 and 3.54 mIOU respectively, where PolyFormer [49] adapts specifically designed regression-based prediction head for coordinates. Florence-2-L also outperforms previous SOTA method UNINEXT [84] on RefCOCO by 0.8 Accuracy@0.5, where UNINEXT [84] is based on advanced object detector Deformable DETR [100] and DINO [97].

Competitive performance with fewer parameters. Florence-2-L achieves competitive performance without the need for LLMs, showcasing efficiency in handling diverse tasks while maintaining a compact size. For instance, Florence-2-L attains a CIDEr score of 140.0 on the COCO Caption karpathy test split [30], outperforming models with significantly more parameters, such as Flamingo (80B parameters, 138.1 CIDEr score).

Adaptable generalization across task levels. Florence-2 demonstrates competitive performance across image-level, pixel-level, and region-level tasks, emphasizing its adaptability and effectiveness in addressing various challenges in computer vision and natural language processing. For example, in the TextVQA task, Florence-2-L sets a new state-of-the-art performance with an accuracy of 81.5 without any external OCR token input, surpassing previous SOTA methods.

在任务层次上具有适应性泛化能力。Florence-2在图像级、像素级和区域级任务中展现出有竞争力的性能，这凸显出其在应对计算机视觉和自然语言处理中的各种挑战时的适应性和有效性。例如，在文本视觉问答（TextVQA）任务中，Florence-2-L在不使用任何外部光学字符识别（OCR）标记输入的情况下，以81.5的准确率达到了新的最优性能，超越了之前的最优方法[12, 15]

我们展示了模型作为视觉基础的多功能性和有效性，其可迁移至多种下游任务。通过整合一系列涵盖图像级、区域级、像素级任务的公共数据集，我们对Florence-2模型进行了微调，从而生成了一个适用于多种视觉任务的通用模型。数据集收集的详细信息见附录B。表5和表6将我们的模型与其他最先进的模型进行了比较。我们的主要发现如下：

简洁设计，卓越性能。Florence-2采用标准的多模态Transformer编码器-解码器架构，无需特殊设计，尤其在区域级和像素级任务上展现出强劲性能。例如，在RefCOCO的REC任务和RES任务中，Florence-2分别以3.0的Accuracy@0.5和3.54的mIOU超越了专门为坐标预测设计的回归预测头的PolyFormer [49]。此外，Florence-2-L在RefCOCO上以0.8的Accuracy@0.5超越了基于先进目标检测器Deformable DETR [100]和DINO [97]的先前SOTA方法UNINEXT [84]

在参数较少的情况下实现竞争性表现。Florence-2-L无需依赖大型语言模型（LLMs）即可实现竞争性性能，展示了其在处理多样化任务时的效率，同时保持了紧凑的模型规模。例如，Florence-2-L在COCO Caption karpathy测试集[30]上获得了140.0的CIDEr分数，优于参数规模显著更大的模型，如Flamingo（800亿参数，CIDEr分为138.1）。

表4. 通用视觉基础模型的零样本性能表现。这些模型在训练过程中未接触评估任务的训练数据。Florence-2模型基于FLD-5B数据集进行预训练。COCO图像描述评估采用Karpathy测试集划分

Method	#params	COCO Cap.		TextCaps		COCO Det.		Flickr30k		Refcoco			Refcoco+			Refcog		Refcoco RES	
		test CIDEr	val CIDEr	val CIDEr	val2017 mAP	test R@1	val	test-A Accuracy	test-B	val	test-A Accuracy	test-B	val	test Accuracy	val	test mIoU			
Flamingo [2]	80B	84.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
Kosmos-2 [60]	1.6B	-	-	-	-	78.7	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7	-				
Florence-2-B	0.23B	133.0	118.7	70.1	34.7	83.6	53.9	58.4	49.7	51.5	56.4	47.9	66.3	65.1	34.6				
Florence-2-L	0.77B	135.6	120.8	72.8	37.5	84.4	56.3	61.6	51.4	53.6	57.9	49.9	68.0	67.0	35.8				

Table 4. **Zero-shot** performance of generalist vision foundation models. The models do not see the training data of the evaluation tasks during training. Florence-2 models are pre-trained on *FLD-5B* dataset. Karpathy test split is used for COCO caption evaluation.

Method	#params	COCO Caption		NoCaps		TextCaps		VQAv2		TextVQA		VizWiz VQA	
		Karpathy test CIDEr	val CIDEr	val CIDEr	val CIDEr	val CIDEr	val CIDEr	test-dev Acc	test-dev Acc	test-dev Acc	test-dev Acc	test-dev Acc	test-dev Acc
Specialist Models													
CoCa [92]	2.1B	143.6	122.4	-	-	82.3	-	-	-	-	-	-	-
BLIP-2 [44]	7.8B	144.5	121.6	-	-	82.2	-	-	-	-	-	-	-
GIT2 [78]	5.1B	145	126.9	148.6	81.7	67.3	71.0	-	-	-	-	-	-
Flamingo [2]	80B	138.1	-	-	82.0	54.1	65.7	-	-	-	-	-	-
PaLI [15]	17B	149.1	127.0	160.0 Δ	84.3	58.8 / 73.1 Δ	71.6 / 74.4 Δ	-	-	-	-	-	-
PaLI-X [12]	55B	149.2	126.3	147 / 163.7 Δ	86.0	71.4 / 80.8 Δ	70.9 / 74.6 Δ	-	-	-	-	-	-
Generalist Models													
Unified-IO [55]	2.9B	-	100	-	77.9	-	57.4	-	-	-	-	-	-
Florence-2-B	0.23B	140.0	116.7	143.9	79.7	63.6	63.6	-	-	-	-	-	-
Florence-2-L	0.77B	143.3	124.9	151.1	81.7	73.5	72.6	-	-	-	-	-	-

Table 5. Performance of specialist and generalist models on captioning and VQA tasks. **Specialist Models** refer to those that are fine-tuned specifically for each task, while **Generalist Models** denote a single model fine-tuned in a task-agnostic manner, applicable across all tasks.

Δ indicates usage of external OCR as input.

表5. 专家模型与通用模型在图像描述生成和视觉问答任务上的表现。专家模型指针对每项任务进行专门微调的模型，而通用模型则表示以任务无关方式微调的单一模型，适用于所有任务。 Δ 表示使用了外部OCR作为输入

这些成就凸显了 Florence-2 在处理多样化任务时的高效性，同时保持了紧凑的规模，使其在人工智能研究和应用不断发展的领域中成为一项独特且宝贵的资产。

ods [12, 15].

These achievements emphasize Florence-2's efficiency in handling diverse tasks while maintaining a compact size, making it a unique and valuable asset in the ever-evolving landscape of AI research and applications.

6.4. Downstream Tasks Fine-tuning

In this section, we investigate the performance of our single model fine-tuning on downstream tasks. This experiment highlights the superiority of Florence-2 pre-training over previous approaches, as it demonstrates the effectiveness of the learned universal image representation. We use the base size model with about 80M parameters in our experiments to ensure fair comparison with other methods.

Object detection and segmentation. We conduct COCO object detection and instance segmentation [48] experiments with Mask R-CNN [26], and COCO object detection [48] experiments with DINO [97] to further demon-

目标检测与分割。我们使用Mask R - CNN[26]进行COCO目标检测和实例分割[48]实验，并且使用DINO[97]进行COCO目标检测[48]实验，以进一步证明Florence - 2预训练的有效性。我们在train2017数据集上进行训练，并在val2017数据集上进行评估

strate the effectiveness of Florence-2 pre-training. We train on the *train2017* split and evaluate on the *val2017* split.

For Mask R-CNN [26] experiments, we follow the common setup used in [51, 97], we use the standard 1 \times (12 epochs) schedule with multi-scale training for all experiments. The learning rate is stepped down by a factor of 0.1 at the 67% and 89% of training epochs. We do not use any additional augmentation (such as random crop, mosaic, etc) or optimization techniques (such as EMA, weight normalization) during training to ensure a fair comparison. We do not use any test time augmentation (TTA) either. Thanks to the strong universal representation learned by Florence-2 pre-training, we do not require longer training epochs, such as 36 epochs in [51, 81, 85, 86], or 100 epochs in [46], to achieve better results.

For DINO [97] experiments, we train DINO-4scale [97] detector for 12 epochs (1 \times) using the same data augmentation strategy as employed by [7].

针对DINO [97]的实验，我们采用与[7]相同的数据增强策略，对DINO-4scale [97]检测器进行了12个周期(1 \times)的训练

针对Mask R-CNN [26]的实验，我们遵循[51, 97]中采用的通用设置，所有实验均采用标准1 \times (12周期)的训练计划，并辅以多尺度训练。学习率在训练周期的67%和89%时分别降低0.1倍。为确保公平比较，训练过程中未使用任何额外的数据增强(如随机裁剪、马赛克等)或优化技术(如指移动平均、权重重归一化)。同样，我们也没有采用任何测试时数据增强(TTA)。得益于 Florence-2 预训练所学习的强大通用表示能力，我们无需像[51, 81, 85, 86]中那样延长训练周期至36个周期，或如[46]中那样达到100个周期，即可获得更优的结果。

表6. 专家模型与通用模型在区域级任务上的表现。专家模型指针对每项任务专门进行微调的模型，而通用模型则指以任务无关方式微调的单一模型，适用于所有任务

Method	#params	COCO Det.	Flickr30k	Refcoco			Refcoco+			Refcocog	Refcoco RES
		val2017 mAP	test R@1	val Accuracy	test-A	test-B	val Accuracy	test-A	test-B	val Accuracy	val mIoU
Specialist Models											
SeqTR [99]	-	-	-	83.7	86.5	81.2	71.5	76.3	64.9	74.9	74.2
PolyFormer [49]	-	-	-	90.4	92.9	87.2	85.0	89.8	78.0	85.8	85.9
UNINEXT [84]	0.74B	60.6	-	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
Ferret [90]	13B	-	-	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3
Generalist Models											
UniTAB [88]	0.23B	-	-	88.6	91.1	83.8	81.0	85.4	71.6	84.6	84.7
Florence-2-B	0.23B	41.4	84.0	92.6	94.8	91.5	86.8	91.7	82.2	89.8	82.2
Florence-2-L	0.77B	43.4	85.2	93.4	95.3	92.0	88.3	92.9	83.6	91.2	91.7
											80.5

Table 6. Performance of specialist and generalist models on region-level tasks. **Specialist Models** refer to those that are fine-tuned specifically for each task, while **Generalist Models** denote a single model fine-tuned in a task-agnostic manner, applicable across all tasks.

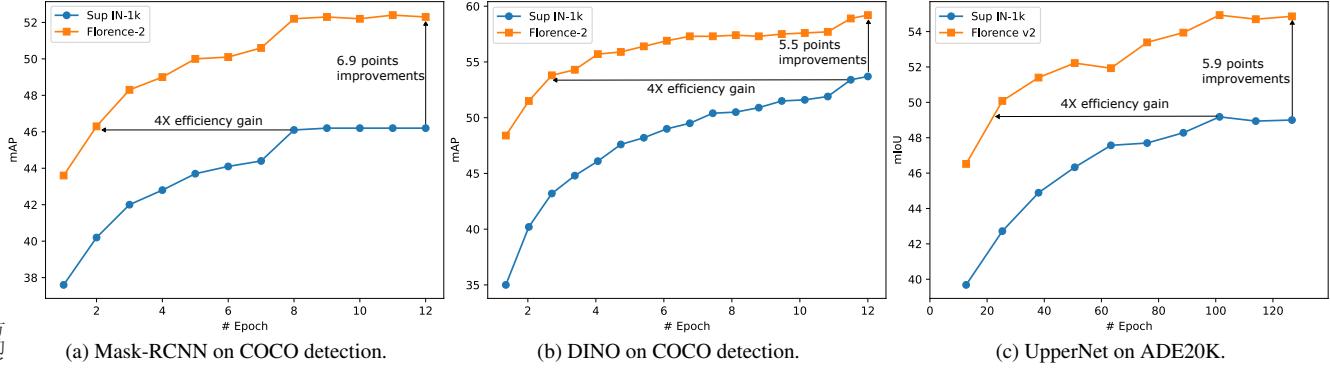


Figure 6. Training efficiency on COCO object detection and segmentation, and ADE20K semantic segmentation tasks.

图6. 在COCO目标检测与分割以及ADE20K语义分割任务上的训练效率

首先，与其他方法相比，我们的基础模型实现了显著的性能提升。如表7所示，由Florence-2预训练的DaViT-B模型在使用Mask RCNN时，超越了之前最佳的基础模型（由FCMAE [81]预训练的ConvNext v2-B），AP提升了0.7。重要的是，尽管ConvNeXt v2-B采用了3倍训练计划（36个周期），我们的模型得益于强大的预训练通用表示，仅需1倍训练计划（12个周期）即可高效完成。在DINO框架下，我们的模型显著优于ViT-B，AP提升了4.2。

其次，我们的预训练展现了更高的训练效率。如表8和图6所示，与采用监督式ImageNet-1k预训练的模型相比，采用Florence-2预训练的我们的模型在Mask-RCNN和DINO框架下分别实现了4倍的效率提升，以及显著的6.9 AP和5.5 AP的性能提升。

Third, our pre-training provides a good generic representation without extensive fine-tuning. Table 8 indicates that the models with *Florence-2* pre-training maintains competitive performances when the first two stages

First, our base model achieves a strong performance improvement compared to other approaches. As shown in Table 7, our DaViT-B model pre-trained by *Florence-2* surpasses previous best base model (ConvNext v2-B), which is pre-trained by FCMAE [81], by 0.7 AP using Mask RCNN. Importantly, while ConvNeXt v2-B leverages a 3× schedule (36 epochs), our model efficiently employs a 1× schedule (12 epochs) thanks to our powerful pre-trained universal representation. For DINO framework, our model significantly outperforms the ViT-B, achieving a notable improvement of 4.2 AP.

Second, our pre-training demonstrates higher training efficiency. As shown in Table 8 and Figure 6, compared to the model with supervised ImageNet-1k pre-training, our model with *Florence-2* pre-training achieves 4x efficiency and a significant improvement of 6.9 AP and 5.5 AP with Mask-RCNN and DINO framework, respectively.

Third, our pre-training provides a good generic representation without extensive fine-tuning. Table 8 indicates that the models with *Florence-2* pre-training maintains competitive performances when the first two stages

are frozen with only 0.3 and 0.2 drops for Mask-RCNN and DINO, respectively. Moreover, our approach with completely frozen backbone can outperform the model with supervised ImageNet-1k pre-training by 1.6 and 2.4 for Mask-RCNN and DINO.

Semantic segmentation. We conduct semantic segmentation experiments with UperNet [82] framework on ADE20k [98] dataset. We mostly follow the training and evaluation protocols from Swin [51]. Specifically, we use input size 512×512 and train the model for 40k iterations with a batch size of 64. We adopt the AdamW [54] optimizer with the optimal learning rate searched from {8e-4, 4e-4, 2e-4, 1e-4}.

Our results show a similar trend to the object detection experiments. As illustrated in Table 9, our base model outperforms the previous SoTA model, which is BEiT pre-trained ViT-B [4], by 1.3 and 1.4 points in single-scale and multi-scale testing protocol, respectively. With the same backbone architecture of DaViT-B [20], *Florence-2* pre-trained model achieves a remarkable improvement of 4.9 points and 4× efficiency compared to the ImageNet-1k pre-

我们的结果呈现出与目标检测实验相似的趋势。如表9所示，在单尺度测试协议和多尺度测试协议下，我们的基础模型分别比之前的最先进的(SoTA)模型（即经过BEiT预训练的ViT-B[4]）高出1.3个点和1.4个点。在与DaViT-B[20]相同的主干架构下，*Florence-2*预训练模型相较于在ImageNet-1k上预训练的对应模型取得了显著的提升，提升了4.9个点并且效率提高了4倍，如表8和图6所示。

语义分割。我们在ADE20k数据集上采用UperNet框架进行语义分割实验。我们主要遵循Swin的训练与评估协议。具体而言，我们使用512×512的输入尺寸，以64的批量大小训练模型40,000次迭代。我们采用AdamW优化器，并从{8e-4, 4e-4, 2e-4, 1e-4}中搜索最优学习率。

表7. 使用Mask-RCNN框架的COCO目标检测与实例分割结果，以及采用DINO-4scale框架的COCO目标检测结果。所有条目均采用基础尺寸模型以确保公平比较。在Mask-RCNN实验中，我们的方法采用 $1\times$ 训练计划（12个周期），ViT-B使用100个周期，其余均采用 $3\times$ 计划（36个周期）。对于DINO实验，除ViT-B使用50个周期外，所有条目均采用 $1\times$ 训练计划

Backbone	Pretrain	Mask R-CNN		DINO
		AP _b	AP _m	AP
ViT-B [46]	MAE, IN-1k	51.6	45.9	55.0
Swin-B [51]	Sup IN-1k	50.2	-	53.4
Swin-B [51]	SimMIM [83]	52.3	-	-
FocalAtt-B [86]	Sup IN-1k	49.0	43.7	-
FocalNet-B [85]	Sup IN-1k	49.8	44.1	54.4
ConvNeXt v1-B [52]	Sup IN-1k	50.3	44.9	52.6
ConvNeXt v2-B [81]	Sup IN-1k	51.0	45.6	-
ConvNeXt v2-B [81]	FCMAE	52.9	46.6	-
DaViT-B [20]	<i>Florence-2</i>	53.6	46.4	59.2

Table 7. **COCO object detection and instance segmentation results** using Mask-RCNN framework, and **COCO object detection results** using DINO-4scale framework. All the entries use a base size model to ensure a fair comparison. For Mask-RCNN experiments, our method utilizes $1\times$ schedule (12 epochs), ViT-B use 100 epochs, all others use $3\times$ (36 epochs). For DINO experiments, all the entries use $1\times$ schedule except for ViT-B which uses 50 epochs.

Pretrain	Frozen stages	Mask R-CNN		DINO	UperNet
		AP _b	AP _m	AP	mIoU
Sup IN1k	n/a	46.7	42.0	53.7	49
UniCL [87]	n/a	50.4	45.0	57.3	53.6
<i>Florence-2</i>	n/a	53.6	46.4	59.2	54.9
<i>Florence-2</i>	[1]	53.6	46.3	59.2	54.1
<i>Florence-2</i>	[1, 2]	53.3	46.1	59.0	54.4
<i>Florence-2</i>	[1, 2, 3]	49.5	42.9	56.7	49.6
<i>Florence-2</i>	[1, 2, 3, 4]	48.3	44.5	56.1	45.9

Table 8. Downstream task fine-tuning on COCO and ADE20K dataset. **COCO object detection** using Mask R-CNN and DINO. **ADE20K semantic segmentation** using UperNet. All entries use DaViT-B with 80M parameters as the backbone and standard $1\times$ schedule.

表8. 在COCO和ADE20K数据集上的下游任务微调。使用Mask R-CNN和DINO进行COCO目标检测。使用UperNet进行ADE20K语义分割。所有条目均采用80M参数的DaViT-B作为骨干网络，并遵循标准的 $1\times$ 训练计划

trained counterpart as demonstrated in Table 8 and Figure 6.

6.5. Ablation Studies

多任务迁移
在本研究中，我们旨在识别出在计算机视觉领域中，针对各种下游任务进行迁移学习时最有效的预训练模型。我们比较了三种不同的模型，每种模型均基于不同任务组合进行了预训练：

- **Image-level Model:** pre-trained on image-level tasks only
图像级别模型：仅在图像级别任务上进行预训练
- **Image-Region Model:** pre-trained on image-level and region-level tasks
图像区域模型：在图像级别和区域级别任务上进行预训练
- **Image-Region-Pixel Model:** pre-trained on image-level, region-level, and pixel-level tasks
图像-区域-像素模型：在图像级别、区域级别和像素级别任务上进行预训练

图像-区域-像素模型：在图像级别、区域级别和像素级别任务上进行预训练

Backbone	Pretrain	mIoU	ms-mIoU
ViT-B [24]	Sup IN-1k	47.4	-
ViT-B [24]	MAE IN-1k	48.1	-
ViT-B [4]	BEiT	53.6	54.1
ViT-B [59]	BEiT v2 IN-1k	53.1	-
ViT-B [59]	BEiT v2 IN-22k	53.5	-
Swin-B [51]	Sup IN-1k	48.1	49.7
Swin-B [51]	Sup IN-22k	-	51.8
Swin-B [51]	SimMIM [83]	-	52.8
FocalAtt-B [86]	Sup IN-1k	49.0	50.5
FocalNet-B [85]	Sup IN-1k	50.5	51.4
ConvNeXt v1-B [52]	Sup IN-1k	-	49.9
ConvNeXt v2-B [81]	Sup IN-1k	-	50.5
ConvNeXt v2-B [81]	FCMAE	-	52.1
DaViT-B [20]	<i>Florence-2</i>	54.9	55.5

Table 9. **ADE20K semantic segmentation results** using UperNet. The input size is 512×512 for all the entries, except for models with BEiT pre-trained, which use the input size of 640×640 . 表9. 使用UperNet进行ADE20K语义分割的结果。所有条目的输入尺寸均为 512×512 ，除了采用BEiT预训练的模型，其输入尺寸为 640×640

在预训练阶段，我们针对FLD5B数据集的子集，对所有模型进行了优化，确保每个模型都处理了相同数量的有效样本（7200万）

随后，这些模型被迁移到一个包含四项下游任务的综合数据集上，每项任务代表了不同层次的任务粒度：COCO图像描述（图像级任务）、COCO目标检测（区域级任务）、Flickr30k定位（区域级任务）以及RefCOCO指称分割（像素级任务）

结果如图7所示。结果表明，在所有三个层次任务上预训练的Image-Region-Pixel模型，在四个下游任务中始终展现出具有竞争力的性能

在COCO字幕任务中，图像-区域-像素模型最初的表现不及图像级模型和图像-区域模型，但最终达到了仅略逊于其他模型的性能（133.4 CIDEr），而其他模型的得分为134.6 CIDEr

在COCO目标检测任务中，图像-区域像素模型相较于图像级模型表现出显著优势（28.3对比0.1），且仅略逊于图像-区域模型（27.9）

在Flickr30k的定位任务中，图像-区域-像素模型展现出强劲的性能（召回率@1为78.1），与图像-区域模型（召回率@1为79.1）相当，并显著优于图像级模型（召回率@1为62.0）

在RefCOCO指称分割任务中，图像-区域-像素模型明显优于图像级模型和图像-区域模型，相较于其他模型（28.4和18.2 mIoU），其性能达到了最高水平（31.6 mIoU）

For pre-training, we optimize all models for the same number of effective samples (72M) on a subset of our *FLD-5B* dataset. These models are then transferred to a combined dataset with four downstream tasks, each representing a different level of task granularity: COCO caption (image-level task), COCO object detection (region-level task), Flickr30k grounding (region-level task), RefCOCO referring segmentation (pixel-level task).

The results are shown in Figure 7. The results demonstrate that Image-Region-Pixel Model, pre-trained on all three levels of tasks, consistently demonstrated competitive performance across the four downstream tasks.

For the COCO caption task, Image-Region-Pixel Model initially performs worse than Image-level Model and Image-Region Model but eventually achieve a final performance (133.4 CIDEr) that is only slightly worse than the other models (134.6 CIDEr).

For the COCO object detection task, Image-Region-Pixel Model outperforms Image-level Model by a significant margin (28.3 vs. 0.1) and was only slightly worse than Image-Region Model (29.7).

For the Flickr30k grounding task, Image-Region-Pixel Model shows strong performance (78.1 recall@1), comparable to Image-Region Model (79.1 recall@1) and significantly better than Image-level Model (62.0 recall@1).

For the RefCOCO referring segmentation task, Image-Region-Pixel Model clearly outperforms both Image-level Model and Image-Region Model, achieving the highest performance (31.6 mIoU) compared to the other models (28.4 and 18.2 mIoU).

Our findings suggest that the Image-Region-Pixel Model, which is pre-trained on tasks at the image, region, and pixel levels, is able to learn general features that can be transferred to other tasks. This is particularly evident in the COCO object detection task, where the Image-Region-Pixel Model achieves a performance (28.3 mIoU) that is significantly higher than the Image-level Model (0.1 mIoU) and the Image-Region Model (29.7 mIoU). This demonstrates the effectiveness of multi-task learning and the proposed pre-training strategy for computer vision tasks.

图7. 多任务迁移。我们使用三个不同版本的Florence-2模型进行实验，每个模型在不同层次的图像标注上进行训练：图像级别、图像和区域级别，以及图像、区域和像素级别。随后，我们评估这些模型在四个下游任务上的迁移学习性能：COCO图像描述、COCO目标检测、Flickr30k定位以及Refcoco参考分割

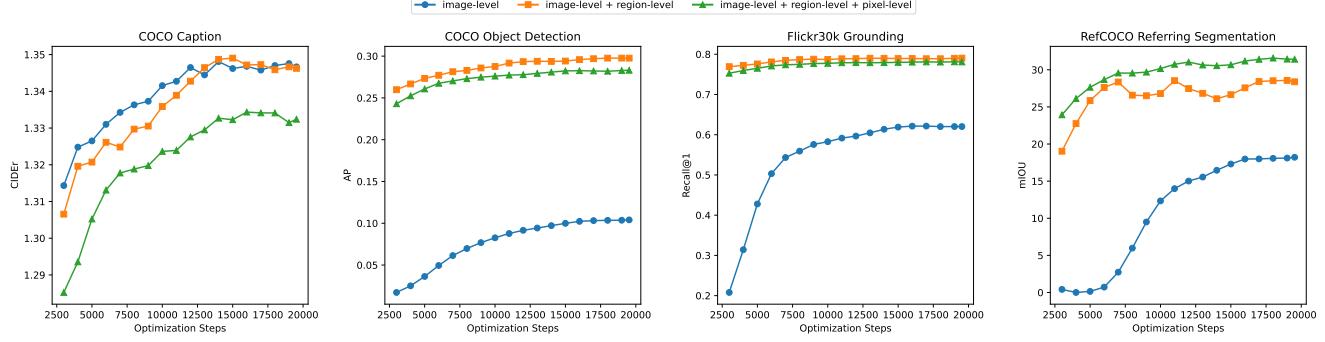


Figure 7. **Multitask transfer.** We conduct experiments with three different versions of *Florence-2* models, each trained on a different level of image annotation: image level, image and region level, and image, region, and pixel level. We then evaluate the transfer learning performance of these models on four downstream tasks: COCO caption, COCO object detection, Flickr30k grounding, and Refcoco referring segmentation.

Model	Caption	Detection	Grounding	RES	
	CIDEr	AP	Recall@1	mIOU	oIOU
Base	118.7	19.7	76.3	18.6	17.8
Large	124.4	22.6	78.2	21.5	19.1

Table 10. **Model scaling.** Zero-shot performance on COCO caption and COCO object detection, Flickr30k grounding, RefCOCO referring expression segmentation(RES).

表10. 模型扩展。在COCO字幕生成和COCO目标检测、Flickr30k定位、RefCOCO指代表分割（RES）上的零样本性能

and pixel levels, is the most effective base model for transfer learning across various computer vision tasks. This model shows strong performance on all four downstream tasks we evaluated, and consistently outperforms the Image-level Model and matches or exceeds the Image-Region Model in performance. By pre-training a model on tasks at different levels of granularity, we can ensure that the base model is better prepared to handle a diverse range of downstream tasks, offering a versatile and robust solution for transfer learning in computer vision.

Model scaling. We aimed to investigate the impact of increasing model capacity on zero-shot performance on various downstream tasks in computer vision. We compared two models: *Florence-2-B* and *Florence-2-L*, which have 232M and 771M parameters, respectively. The model architectures are described in Table 15. We show the zero-shot performance on four downstream tasks in Table 10. The large model clearly outperforms the base model across various downstream tasks.

Data scaling. We conducted experiments to study how zero-shot performance on various computer vision tasks is affected by the scale of pre-training data. We used four different data sizes for pre-training: 0.12M, 0.36M, 1.2M, and 12M images. All models were trained with the same effective sample size (72M) on a subset of *FLD-5B* data.

Table 11 presents the zero-shot performance results on

表11展示了在COCO字幕（COCO caption）、COCO目标检测（COCO object detection）、Flickr30k实体定位（Flickr30k grounding）和RefCoco指代表分割（RefCoco referring segmentation, RES）任务上的零样本性能结果。我们可以观察到随着预训练数据量的增加，在下游任务上的零样本性能呈现提升的趋势（除了RES任务，在该任务上120万数据量比1200万数据量的性能略好）

Data size	Caption	Detection	Grounding	RES	
	CIDEr	AP	Recall@1	mIOU	oIOU
0.12M	102.8	16.1	74.0	15.9	16.6
0.36M	114.3	18.7	75.8	16.6	16.4
1.2M	118.1	18.9	76.3	19.3	18.4
12M	118.7	19.7	76.3	18.6	17.8

Table 11. **Data scaling.** Zero-shot performance on COCO caption, COCO object detection, Flickr30k grounding, COCORef referring segmentation.

表11. 数据缩放。在COCO字幕、COCO目标检测、Flickr30k定位、COCORef指代分割任务上的零样本性能表现

COCO caption, COCO object detection, Flickr30k grounding, and RefCoco referring segmentation (RES) tasks. We can observe a trend of improved zero-shot performance on the downstream tasks as the pre-training data size increases (except for RES, 1.2M data has slightly better performance compared to 12M).

Our experiments on data scaling demonstrate that larger pre-training data sizes generally lead to improved zero-shot performance across a variety of downstream tasks in computer vision. This finding suggests that investing in larger pre-training datasets can provide a more effective and versatile foundation for handling a wide range of downstream tasks.

Our approach to scaling data is significantly more efficient than relying solely on human annotations, as most of the annotation generation is performed using model inference. By leveraging specialist models to generate annotations, we can substantially reduce the time and cost associated with manual annotation efforts, which often involve labor-intensive processes and may be subject to human errors or inconsistencies.

Furthermore, utilizing model-generated annotations enables us to scale the pre-training datasets more rapidly and efficiently, allowing us to explore the impact of larger data

此外，利用模型生成的注释能够让我们更快速、高效地扩展预训练数据集，从而使我们能够探究更大的数据量对计算机视觉中各类下游任务的模型性能的影响。这不仅有助于开发更有效、更多功能的基本模型，而且随着对高质量标注数据需求的持续增长，还能确保注释过程保持可持续性和可扩展性

我们对数据规模进行的实验表明，在计算机视觉领域的多种下游任务中，更大的预训练数据规模通常能够带来零样本性能的提升。这一发现表明，投资于更大的预训练数据集可以为处理广泛的下游任务提供更有效且多功能的基础。

我们的数据扩展方法显著优于仅依赖人工标注的方式，因为大部分标注生成工作是通过模型推理完成的。通过利用专业模型来生成标注，我们能够大幅减少与手动标注相关的时间和成本，这些手动标注通常涉及劳动密集型过程，并且可能受到人为错误或不一致性的影响。

表12. 基础组件。在COCO字幕、COCO目标检测、Flickr30k定位以及COCOREf指代分割任务上的零样本性能表现。V Pre与L Pre分别表示采用视觉与语言预训练初始化

V Pre L Pre	Caption	Detection	Grounding	RES	
	CIDEr	AP	Recall@1	mIOU	oIOU
Freeze Vision Encoder					
✓ ✓	120.0	6.9	66.3	9.9	13.6
Unfreeze Vision Encoder					
✓	81.3	4.9	69.0	15.3	15.6
✓	117.4	19.6	75.2	21.5	19.3
✓ ✓	118.7	19.7	76.3	18.6	17.8

Table 12. **Basic components.** Zero-shot performance on COCO caption, COCO object detection, Flickr30k grounding, and COCORef referring segmentation. V Pre and L Pre indicate that using vision and language pre-training initialization, respectively.

sizes on model performance across various downstream tasks in computer vision. This not only facilitates the development of more effective and versatile foundation models but also ensures that the annotation process remains sustainable and scalable as the need for high-quality labeled data continues to grow.

In summary, our data scaling approach offers a more efficient alternative to traditional human annotation methods by harnessing the power of specialist models for annotation generation. This strategy enables us to accelerate the pre-training process, optimize model performance, and effectively manage the ever-increasing demand for labeled data in the field of computer vision.

Training settings. We analyze the basic model training settings for the two primary components of our model, namely the vision encoder and the multi-modality encoder-decoder. The experiment results are presented in Table 12

We observe that freezing the vision encoders does not affect the performance on tasks that require image-level understanding, but it significantly degrades the performance on tasks that require region-level or pixel-level understanding (e.g., AP on COCO object detection drops from 19.7 to 6.9). Previous methods for pre-training vision foundation models mainly focus on image-level tasks (e.g., image classification [27, 38], image-text contrastive learning [64, 95]), which may not provide them with sufficient region-level and pixel-level skills for downstream tasks. Therefore, it is important to unfreeze the vision backbone, enabling it to learn region-level and pixel-level features for various downstream tasks.

The effect of language pre-training weights on multi-modal encoder-decoder tasks varies depending on the task. Tasks that require more text understanding, such as captioning and grounding, benefit slightly from using language pre-training weights (e.g., COCO caption, Flickr30k grounding). Tasks that are mostly vision-focused, such as object detection and region segmentation, do not gain much from

语言预训练权重对多模态编码器-解码器任务的影响因任务而异。需要更多文本理解的任务，如图像描述和定位，使用语言预训练权重（例如，COCO图像描述、Flickr30k定位）略有裨益。主要侧重于视觉的任务，如目标检测和区域分割，使用语言预训练权重获益不大（对于COCO目标检测，增益仅为0.1；对于仅使用定位标记的RES任务，下降为2.91 mIOU）。

using language pre-training weights (for COCO object detection, the gain is only 0.1; for RES tasks, which use only localization tokens, the drop is 2.91 mIOU).

We investigate the effects of different training configurations on the performance of a foundation model in region-level and pixel-level tasks. We find that unfreezing the vision backbone is crucial for enhancing the model's ability to learn from regions and pixels, which is beneficial for transferring to various downstream tasks. Moreover, we observe that using language pre-training weights can help the model in tasks that require text understanding, but have less impact on tasks that are purely vision-based. These results offer useful guidance for choosing the best training settings for different computer vision tasks.

我们研究了不同训练配置对基础模型在区域级别和像素级别任务中表现的影响。研究发现，解冻视觉骨干网络对于增强模型从区域和像素中学习的能力至关重要，这有利于向各种下游任务的迁移。此外，我们观察到，使用语言预训练权重能够帮助模型在需要文本帮助理解的任务中表现更佳，但对于纯视觉任务的影响则较小。这些结果为不同计算机视觉任务选择最佳训练设置提供了有价值的指导。

7. Related Works

7.1. Vision-Language Foundation Models

Recent vision-language pre-training models [29, 64, 95] have demonstrated impressive zero-shot transfer abilities to vision-language alignment and image classification tasks, thanks to the alignment of vision and text embeddings extracted from respective encoders through contrastive learning objectives [58, 74]. These models (e.g., [95]), trained on weakly large-scale image-text data, have been further extended to more downstream tasks such as object detection, achieving state-of-the-art performance with task-specific adaptation heads.

近期的视觉-语言预训练模型[29, 64, 95]展现了在视觉-语言对齐和图像分类任务上显著的零样本迁移能力，这得益于通过对比学习目标[58, 74]从各自编码器中提取的视觉与文本嵌入的对齐。这些模型（例如[95]）在弱监督的大规模图像-文本数据上训练，已进一步扩展到更多下游任务，如物体检测，通过任务特定的适应头实现了最先进的性能。

In contrast, other studies [2, 45, 78, 92] propose using a multi-modality decoder to predict text in an autoregressive manner with language modeling pre-training objectives. Techniques for fusing vision and language embeddings vary: GIT [78] concatenates vision and text tokens as decoder input and designs a causal attention mask, CoCa [92] uses attentional poolers with learnable queries to select task-specific vision representations which are then cross-attended via the decoder, and Flamingo [2] pools a fixed number of vision tokens with a Perceiver Resampler and adds new learnable cross-attention layers to the decoder while freezing the pre-trained vision encoder and text decoder.

相比之下，其他研究[2, 45, 78, 92]提出使用多模态解码器以自回归方式预测文本，并采用语言建模预训练目标。融合视觉和语言嵌入的技术各不相同：GIT[78]将视觉和文本标记连接起来作为解码器输入，并设计了一个因果注意力掩码；CoCa[92]使用带有可学习查询的注意力池化器来选择任务特定的视觉表示，然后通过解码器进行交叉注意；Flamingo[2]使用感知器重采样器对固定数量的视觉标记进行池化，并在解码器中添加新的可学习交叉注意层，同时冻结预训练的视觉编码器和文本解码器。

Beyond image captioning pre-training task, some research [15, 55, 79] attempts to formulate more vision tasks in a unified sequence-to-sequence learning paradigm, including object detection and image segmentation. Customized special tokens accommodate representations beyond pure text, such as bounding boxes [10, 55, 79]. This approach uses the same architecture for pre-training and downstream tasks, potentially using the same set of weights for all tasks. Our method, which falls into this category, aims to obtain foundation models that understand dense information beyond simple image-level captions. It shares the same encoder-decoder design as other multi-modality encoder-

除了图像描述预训练任务外，一些研究[15, 55, 79]尝试将更多视觉任务统一为序列到序列的学习范式，包括目标检测和图像分割。定制的特殊标记适应了纯文本之外的表示，例如边界框[10, 55, 79]。这种方法在预训练和下游任务中使用相同的架构，可能对所有任务使用相同的权重集。我们的方法属于这一类别，旨在获得能够理解超越简单图像级别的密集信息的基础模型。它与其他多模态编码器-解码器模型[15, 55]共享相同的编码器-解码器设计，这些模型适用于序列到序列学习，但使用我们构建的大规模综合注释数据，而不是结合现有的稀疏注释数据。

综上所述，我们分析了模型两个主要组件的基本训练设置，即视觉编码器和多模态解码器。实验结果显示，通过利用专业模型进行标注生成，为传统的手工标注方法提供了一种更为高效的替代方案。这一策略使我们能够加速预训练过程，优化模型性能，并有效应对计算机视觉领域对标注数据日益增长的需求。

训练设置。我们分析了模型两个主要组件的基本训练设置，即视觉编码器和多模态解码器。实验结果显示

我们观察到，冻结视觉编码器不会影响需要图像级别理解的任务性能，但会显著降低需要区域级别或像素级别的任务性能（例如，COCO目标检测的平均精度从19.7降至6.9）。以往预训练视觉基础模型的方法主要集中在图像级别任务（如图像分类[27, 38]、图像-文本对比学习[64, 95]），这可能无法为下游任务提供足够的区域级别和像素级别的技能。因此，解冻视觉骨干网至关重要，使其能够学习适用于各种下游任务的区域级别和像素级别的特征。

语言预训练权重对多模态编码器-解码器任务的影响因任务而异。需要更多文本理解的任务，如图像描述和定位，使用语言预训练权重（例如，COCO图像描述、Flickr30k定位）略有裨益。主要侧重于视觉的任务，如目标检测和区域分割，使用语言预训练权重获益不大（对于COCO目标检测，增益仅为0.1；对于仅使用定位标记的RES任务，下降为2.91 mIOU）。

decoder models [15, 55] adapted for sequence-to-sequence learning, but uses our built large-scale comprehensive annotation data instead of combining existing sparse annotated data.

7.2. Vision Datasets

Comprehensive annotations. The quest for comprehensive understanding of visual scenes, the holy grail of computer vision [36], has evolved from focusing on individual datasets each targeting a single perspective, *e.g.*, image classification [18], to providing multi-perspective [36, 40, 48], comprehensive annotations for every visual data point. Notable datasets like MS-COCO [13, 48] and Visual Genome [36] integrate various types of annotations, enabling richer understanding in spatial and semantic granularities and better model interactions across annotations. However, due to the high cost of human verification, these annotations are limited in size. Our datasets, while large-scale, maintain comprehensive annotations covering text, region-text pairs, and text-phrase-region triplets, with reduced human involvement.

Scalable annotations. Over the past decade, vision datasets have rapidly scaled up from thousands [37, 42] to billion examples [29, 96] to encompass more visual concepts for better generalization. This shift is evident in recent foundation models that employ massive quantities of data [5]. These large datasets typically collect images from the web and parse noisy annotations from the corresponding metadata, such as category label from query [75, 96], short description from alt-text [29, 64], as well as detailed description from interleaved text [2, 41]. Despite their diversity, these annotations suffer from randomness and limited types (*i.e.*, texts only). Some works [32, 45] attempt to scale up annotations using pseudo-label generation with iteratively trained models, which offer higher quality without significant diversity loss. Our data pipeline extends these large-scale, web-crawled noisy annotations with higher-quality, autonomous annotations generated from multiple specialist models. The pipeline iteratively refines labels and completes missing pieces, resulting in a scalable and comprehensive dataset for learning a unified visual representation.

8. Conclusion

The Florence Project endeavors to develop a foundational vision model endowed with a diverse array of perceptual capabilities, encompassing spatial hierarchy and semantic granularity. To this end, we construct *FLD-5B* dataset containing an extensive collection of 126M images paired with 5B comprehensive annotations, which are collected by the Florence data engine. Subsequently, we pre-train *Florence-2* on this rich dataset through comprehensive multitask learning in a unified manner. *Florence-2* has ex-

hibited remarkable zero-shot capabilities that extend across a wide spectrum of visual tasks, such as captioning, object detection, visual grounding, and referring segmentation, among others. The experimental findings underscore the potency of the universal representation pre-trained by *Florence-2*, revealing its substantial contributions to the enhancement of a multitude of downstream tasks.

Acknowledgment. We would like to express our heartfelt gratitude to all the contributors from the Azure AI team who worked on the Florence project. We sincerely appreciate Misha Bilenko for the invaluable guidance and support. Our thanks are extended to Yi-Ling Chen, Mengchen Liu, Yen-Chun Chen and Dongdong Chen for engaging in helpful discussions and to Yunsheng Li for their assistance with segmentation annotations. Deep appreciation is also expressed to Qingfen Lin, Ryan Menezes, Kuan Lu, Gabe Blanco, Shohei Ono, Ping Jin, Jiahe Zhou, Xiong Qiao, Tong Bai, Xingchao Peng, Pei Guo, Lihang Li for providing valuable feedback in downstream applications discussions. Special thanks to Cha Zhang, Jinyu Li, Min Gao, Christina Sun, Oliver Ernst, Kevin Pan, Mei Gao for their work on data annotation support and insightful discussions in data pipeline. Furthermore, we would like to thank Thomas Soemo, Nguyen Bach for their constructive feedback.

References

- [1] Azure ai services. <https://azure.microsoft.com/en-us/products/ai-services/?activetab=pivot:azureopenaiservicetab>. Accessed: 2023-10-13. 6
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3, 7, 9, 10, 14, 15
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 4
- [4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 2, 11, 12
- [5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 15
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

- Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 10
- [8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 2
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2
- [10] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection, 2022. 4, 14
- [11] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 4
- [12] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 3, 10
- [13] 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. 3, 6, 15, 20
- [14] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger, 2023. 3
- [15] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2022. 3, 4, 9, 10, 14, 15
- [16] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 2
- [17] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3, 4, 15
- [19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1, 2
- [20] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 74–92. Springer, 2022. 4, 11, 12
- [21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 7
- [22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 20
- [23] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 20
- [24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 12
- [25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2
- [26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 10
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 14
- [28] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020. 5, 8

- [29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 9, 14, 15
- [30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2014. 9
- [31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 20
- [32] 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. *arXiv preprint arXiv:2304.02643*, 2023. 2, 6, 7, 15
- [33] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19, 2007. 2
- [34] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017. 3
- [35] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 20
- [36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 15
- [37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 15
- [38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2, 14
- [39] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, mar 2020. 20
- [40] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 4, 7, 15
- [41] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekerman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023. 15
- [42] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 15
- [43] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. 1, 4, 9
- [44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 10
- [45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 5, 14, 15
- [46] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 10, 12
- [47] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 20
- [48] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 9, 10, 15
- [49] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 9, 11
- [50] 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. 6
- [51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 10, 11, 12
- [52] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 12

- [53] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 9
- [54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 9, 11
- [55] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks, 2022. 3, 4, 10, 14, 15
- [56] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2, 20
- [57] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. 20
- [58] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 14
- [59] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. 2022. 12
- [60] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 7, 9, 10, 34, 35
- [61] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2, 9
- [62] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 20
- [63] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. *arXiv preprint arXiv:2301.02280*, 2023. 5, 8
- [64] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 7, 9, 14, 15
- [65] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1, 2
- [66] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1, 2
- [67] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 9
- [68] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4
- [69] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. 20
- [70] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 4, 5, 7, 20
- [71] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 4
- [72] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. 20
- [73] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 20
- [74] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 14
- [75] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 15
- [76] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 2
- [77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4
- [78] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022. 2, 10, 14
- [79] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022. 3, 4, 14
- [80] Nic M Weststrate, Susan Bluck, and Judith Glück. Wisdom of the crowd. *The Cambridge handbook of wisdom*, pages 97–121, 2019. 2

- [81] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. [10](#), [11](#), [12](#)
- [82] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. [2](#), [11](#)
- [83] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhiliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. [12](#)
- [84] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. [9](#), [11](#)
- [85] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. [10](#), [12](#)
- [86] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. [10](#), [12](#)
- [87] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space, 2022. [9](#), [12](#)
- [88] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. [11](#)
- [89] Sheng Kung Michael Yi, Mark Steyvers, Michael D Lee, and Matthew J Dry. The wisdom of the crowd in combinatorial problems. *Cognitive science*, 36(3):452–470, 2012. [2](#)
- [90] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. [11](#)
- [91] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. [3](#)
- [92] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. [9](#), [10](#), [14](#)
- [93] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. [2](#), [20](#)
- [94] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–85, Cham, 2016. Springer International Publishing.
- [95] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Lu-wei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. [2](#), [6](#), [9](#), [14](#)
- [96] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. [15](#)
- [97] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. [2](#), [6](#), [9](#), [10](#)
- [98] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. [2](#), [11](#)
- [99] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. [11](#)
- [100] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. [9](#)

A. Supported Tasks and Annotations in Florence-2

Task	Annotation Type	Prompt Input	Output
Caption	Text	Image, text	Text
Detailed caption	Text	Image, text	Text
More detailed caption	Text	Image, text	Text
Region proposal	Region	Image, text	Region
Object detection	Region-Text	Image, text	Text, region
Dense region caption	Region-Text	Image, text	Text, region
Phrase grounding	Text-Phrase-Region	Image, text	Text, region
Referring expression comprehension	Region-Text	Image, text	Text, region
Open vocabulary detection	Region-Text	Image, text	Text, region
Referring segmentation	Region-Text	Image, text	Text, region
Region to text	Region-Text	Image, text, region	Text
Text detection and recognition	Region-Text	Image, text	Text, region

Table 13. Supported Tasks and annotations used for *Florence-2* pretraining.

B. Supervised Data Collection for Generalist Model Fine-tuning

Task	Dataset
Caption	COCO [13]
Text Caption	TextCaps [72]
Paragraph caption	Standford Paragraph Caption [35]
Detailed caption	Localized Narratives [62]
Detection	COCO [47], Object365* [70], Open Images* [39]
Phrase Grounding	Flickr30k, Object365* [70], Open Images* [39]
Referring expression	RefCOCO-mix (RefCOCO, RefCOCO+, RefCOCOg) [31, 56, 93]
Referring expression segmentation	RefCOCO-mix (RefCOCO, RefCOCO+, RefCOCOg) [31, 56, 93]
Region to category	COCO [47], Object365* [70], Open Images* [39]
Region to polygon	COCO [47] (after deduplicating RefCOCO-mix val)
VQA	VQAv2 [22], OKVQA [57], AOKVQA [69], TextVQA [73], ViZWiz VQA [23]
OCR	Subset from <i>FLD-5B</i> OCR (2 million samples)

Table 14. Collection of dataset for finetuning one single generalist model for downstream tasks evaluation. * indicates using the annotations from *FLD-5B*, which merges original annotations with ours.

C. Model Configuration

Model	Image Encoder (DaViT)				Encoder-Decoder (Transformer)			
	dimensions	blocks	heads/groups	#params	encoder layers	decoder layers	dimensions	#params
<i>Florence-2-B</i>	[128, 256, 512, 1024]	[1, 1, 9, 1]	[4, 8, 16, 32]	90M	6	6	768	140M
<i>Florence-2-L</i>	[256, 512, 1024, 2048]	[1, 1, 9, 1]	[8, 16, 32, 64]	360M	12	12	1024	410M

Table 15. Model configuration of different size.

D. More Examples of Annotations in FLD-5B



Figure 8. Examples of annotations in FLD-5B.



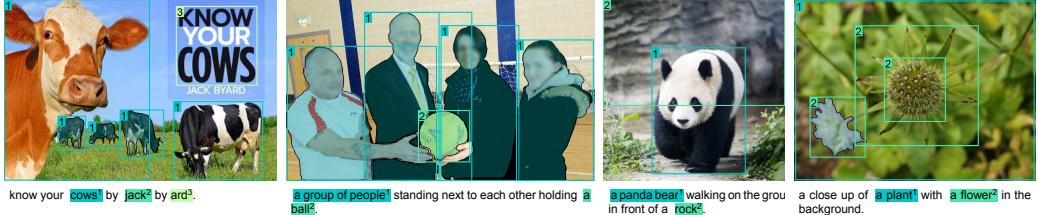
(a) Region only



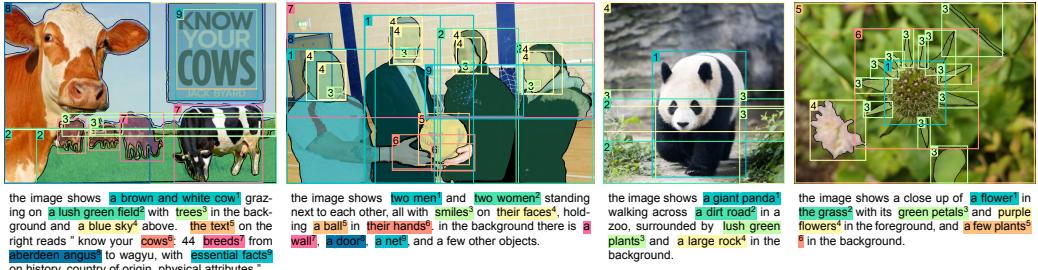
(b) Region w/ phrases



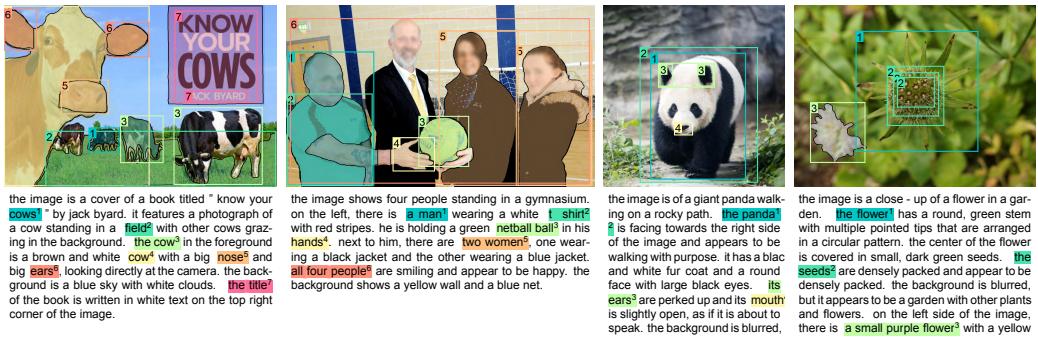
(c) Region w/ brief text



(d) Text-phrase-region w/ brief text



(e) Text-phrase-region w/ detailed text



(f) Text-phrase-region w/ more detailed text

Figure 9. Examples of annotations in FLD-5B (continued).

E. Qualitative Evaluation and Visualization Results

E.1. Detailed Image Caption

Detailed Image Caption

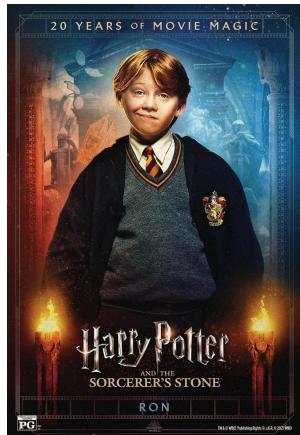
Prompt: Describe with a paragraph what is shown in the image.



The image is a hand-drawn illustration of a glass of gin and tonic. The glass is filled with ice cubes, lime slices, and a slice of lemon. There is a straw sticking out of the top of the glass. The background is a light green color with a grungy texture. On the right side of the image, there is a list of ingredients written in black ink. The text reads "Gin & Tonic" and "2oz gin, 5oz tonic water, 2-4 ice cubes lime slice for garnish".



The image shows the interior of a grocery store with a large variety of fruits and vegetables on display. The store has a high ceiling with green and white striped awnings, and the floor is made of wood. There are rows of wooden crates filled with different types of fruits, including apples, oranges, lemons, limes, and limes. The crates are arranged in neat rows, and there are price tags attached to each crate. In the background, there are shelves stocked with various items such as fruits, vegetables, and other grocery items. The overall atmosphere of the store is bright and colorful, with a sense of freshness and abundance.



The image is a movie poster for the film "Harry Potter and the Sorcerer's Stone" starring Ron Weasley. The poster features a young boy, Ron Weasley, standing in front of a dark background with a castle-like structure in the background. He is wearing a school uniform with a Gryffindor crest on his jacket and a red tie. He has blonde hair and is looking directly at the camera with a serious expression on his face. The title of the film is written in white text at the top of the poster, with the tagline "20 years of movie magic" written in smaller text below.



The image is a digital illustration of a girl hugging a white cat. The girl is wearing a pink sweater and has long brown hair. She is sitting on a green surface with several potted plants and flowers around her. The plants have green leaves and pink and white flowers. There are also two butterflies fluttering around the scene. The background is white. The overall style of the illustration is cartoon-like and playful.

Figure 10. Detailed captioning prediction results.

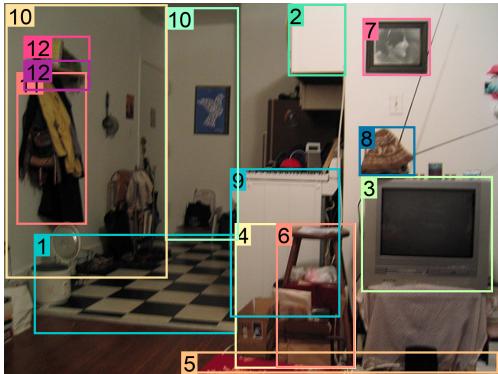
E.2. Visual Grounding

Visual Grounding

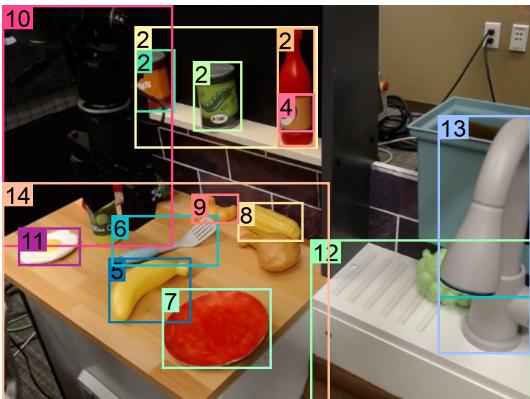
Prompt: Locate the phrases in the caption: {caption}



The image shows a group of five cartoon monsters. On the left side, there is a brown monster¹ with horns and a big smile on its face. Next to it, there are two smaller monsters², one black and one green. The black monster³ has two large horns on its head and is standing in the center of the group. The green monster⁴ on the right side is a green monster with big eyes and a long antennae. It is standing on its hind legs with its arms stretched out to the sides. In the middle of the image, there appears to be a small blue monster⁵ with a round head and two antennae on its back. The background is light beige with small green circles scattered around.



The image shows a cluttered room with a black and white checkered floor¹. On the right side of the image, there is a small white cabinet² with a television³ on top of it. Next to the cabinet, there are several items⁴ scattered on the floor, including a red blanket⁵, a wooden stool⁶, and a pile of trash. On top of the cabinet is a picture frame⁷ and a hat⁸. In the center of the room is a white refrigerator⁹ with a few items on top. The walls¹⁰ are painted white and there are a few clothes¹¹ hanging on a rack¹² on the left wall. The room appears to be in disarray, with some items strewn about and others scattered around.

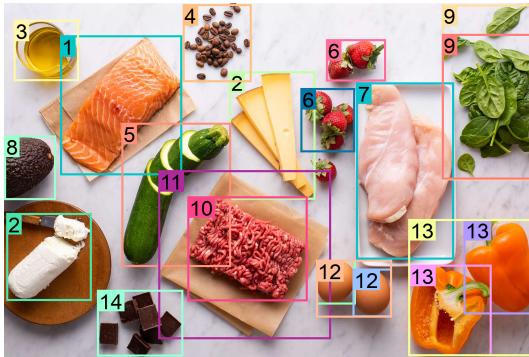


The image shows a kitchen countertop with various kitchen items on it. On the left side of the countertop, there is a microscope with a black body and a white lens¹. Next to the microscope, there are two bottles of condiments² - one with a red label³ and the other with green. On top of the microscope is a yellow banana⁴, a blue spatula⁵, a red plate⁶, and a yellow corn⁷ on the cob. In the center of the image, there appears to be a frying pan⁸ with a fried egg⁹ on it, and on the right side is a white sink¹⁰ with a white faucet¹¹. The countertop¹² is made of wood and has a gray tile backsplash.

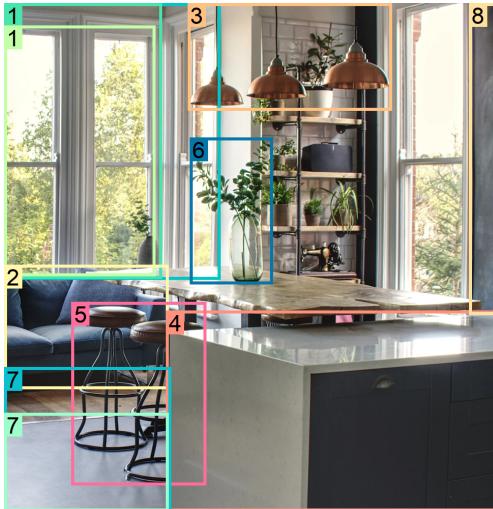
Figure 11. Visual grounding prediction results.

Visual Grounding

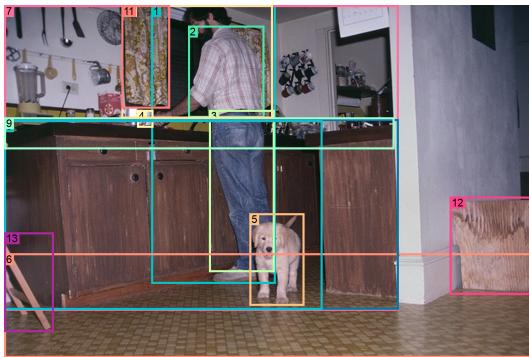
Prompt: Locate the phrases in the caption: {caption}



The image is a flat lay of various food items arranged on a white marble countertop. On the left side of the image, there is a piece of salmon¹. Next to it, there are slices of cheese², a glass of oil³, coffee beans⁴, a zucchini⁵, a bunch of strawberries⁶, two chicken breasts⁷, an avocado⁸ and a few whole spinach leaves⁹. In the center of the table, there appears to be a pile of ground beef¹⁰ on paper¹¹, two eggs¹², two orange bell peppers¹³, and some dark chocolate bars¹⁴. The items are arranged in a way that suggests they are being prepared for a meal.



The image shows a modern kitchen with a large window on the left side. The window¹ has a view of trees and greenery outside. On the left side of the image, there is a blue sofa² with a wooden coffee table in front of it. Above the table, there are three copper pendant lights³ hanging from the ceiling. There is a large island⁴ with a white countertop. There are two bar stools⁵ next to the table. In the center of the kitchen, there is a bottle green plants⁶ on the table. The floor⁷ is made of light-colored wood and the walls⁸ are painted in a dark blue color.



The image shows a man¹ standing in a kitchen with a small dog. The man¹ is wearing a plaid shirt² and jeans³ and is holding a red cup⁴ in his hand. The dog⁵ is a light brown color and is standing on a tiled floor⁶. The kitchen⁷ has wooden cabinets⁸ and a countertop⁹ with various kitchen utensils hanging on the wall. There is a window¹⁰ with yellow curtains¹¹ in the background. On the right side of the image, there is a wooden cutting board¹² and a wooden stool¹³.

Figure 12. Visual grounding prediction results. (continued)

E.3. Dense Region Caption

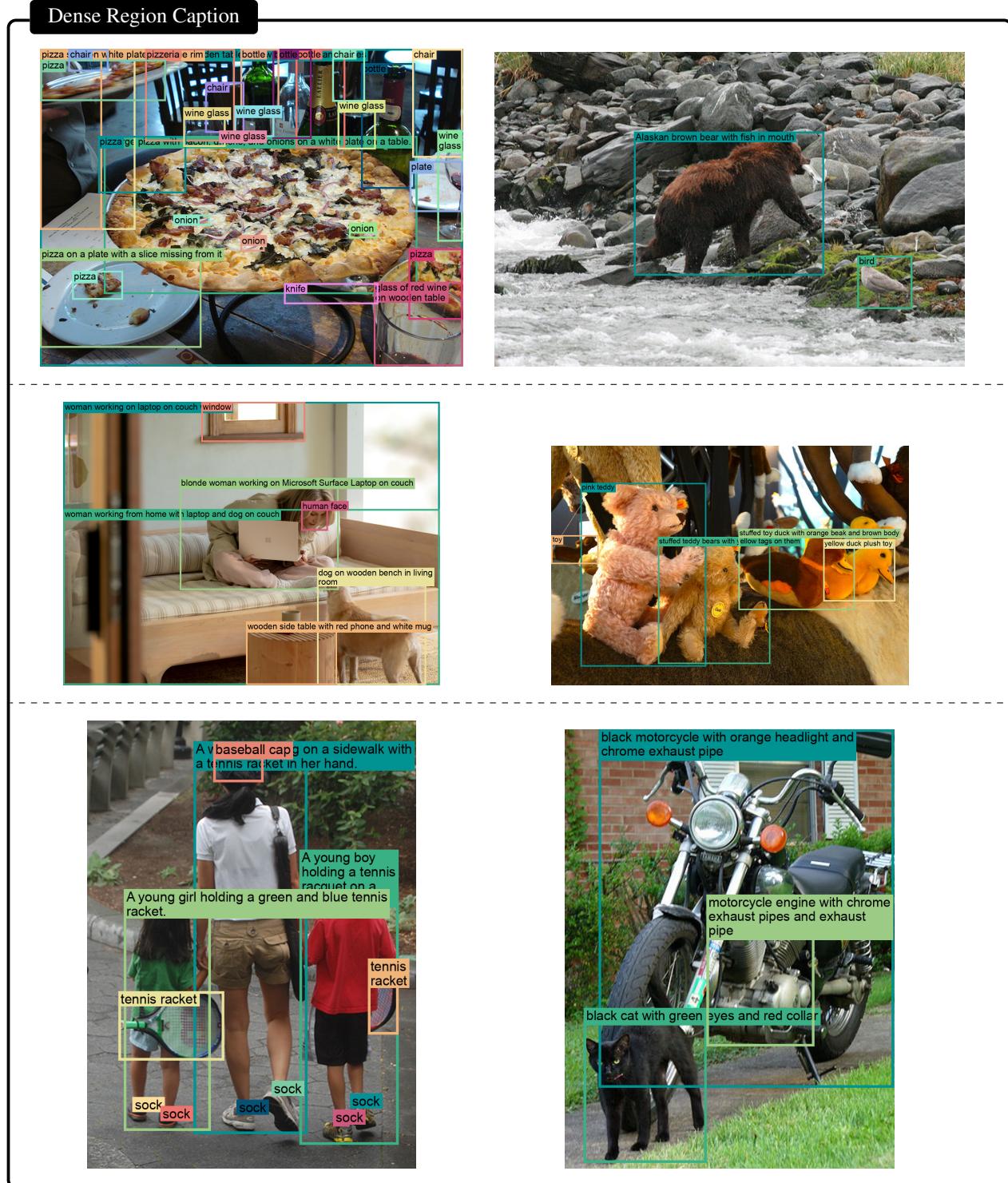


Figure 13. Dense region caption prediction results.

E.4. Open Vocabulary Detection

Open Vocabulary Object Detection

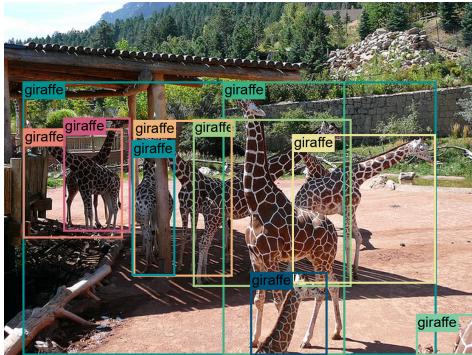
Prompt: Locate Five Alive juice box (and) Colgate toothpaste in the image.



Prompt: Locate Chewbacca in the image.



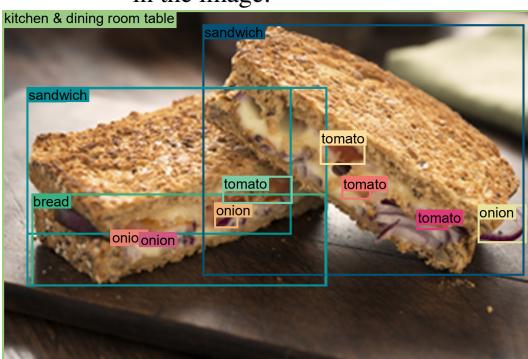
Prompt: Locate giraffe in the image.



Prompt: Locate Mercedes-Benz (and) M2 (and) Audi in the image.



Prompt: Locate the objects with category name in the image.



Prompt: Locate the objects with category name in the image.

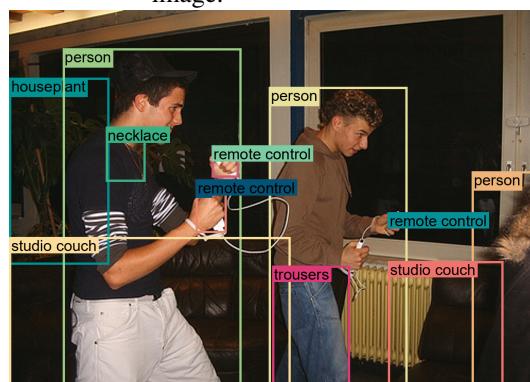
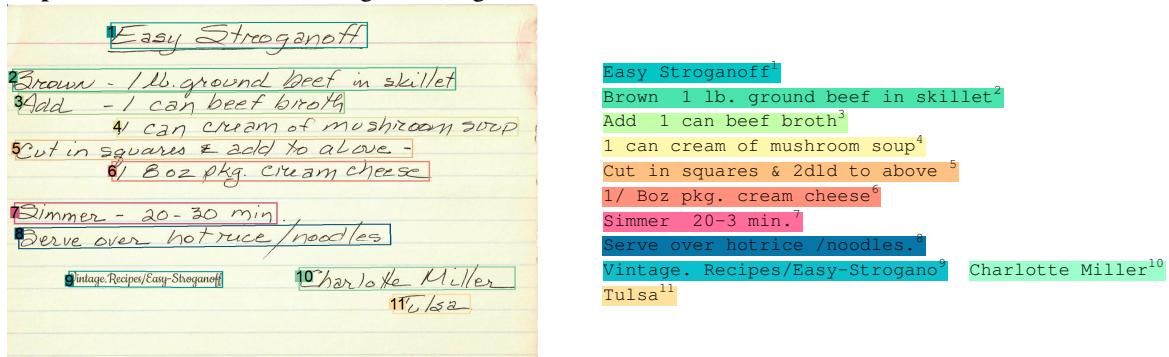


Figure 14. Open vocabulary object detection prediction results.

E.5. OCR

Ocr with region

Prompt: What is the text in the image, with regions?



COFFEE+TEA¹ BLENDED²
 \$1.69/\$1.89/\$2.09³ \$3.49/\$3.99⁴
 Hot Coffee/Tea⁵ Taro⁶
 Iced Coffee/ Tea⁷ Mango⁸
 Hot Chocolate⁹ Honeydew¹⁰
 \$3.49/\$ 3.99¹¹ Strawberry¹² Mocha¹³
 Thai Iced Tea / Coffee¹⁴ Caramel¹⁵
 \$1.99/\$2.29/\$2.59¹⁶ SPECIALTY Brew !!¹⁷
 Jasmine GreenTea¹⁸



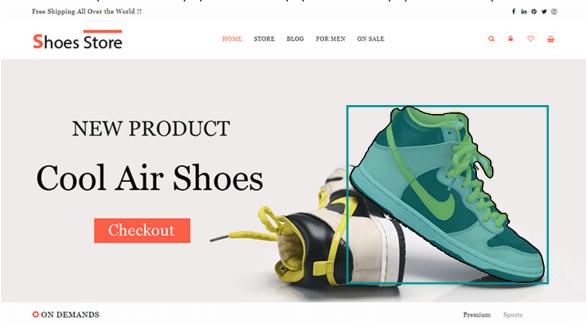
LEONARDO¹
 DiCAPRIO²
 ROBERT³
 DE NIRO⁴
 LILY⁵
 GLADSTONE⁶
 A MARTIN SCORSESE PICTURE⁷
 KILLERS!⁸
 OF⁹ FLOWER¹⁰
 MOON¹¹
 SCREENLY ERIC ROTH AND MARTIN SCORSESE DIRECTED¹²
 BYMARTIN SCORSESE¹³
 ONLY IN THEATRES OCTOBER 20¹⁴

Figure 15. Ocr with region prediction results.

E.6. Region to segmentation

Region to Segmentation

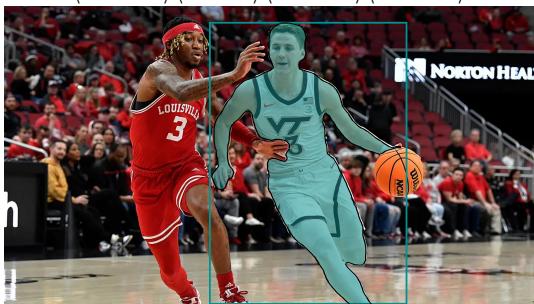
Prompt: What is the polygon mask of region
`<loc_586><loc_294><loc_929><loc_814>`



Prompt: What is the polygon mask of region
`<loc_541><loc_266><loc_692><loc_627>`



Prompt: What is the polygon mask of region
`<loc_386><loc_53><loc_759><loc_998>`



Prompt: What is the polygon mask of region
`<loc_317><loc_314><loc_893><loc_904>`



Prompt: What is the polygon mask of region
`<loc_583><loc_66><loc_794><loc_331>`



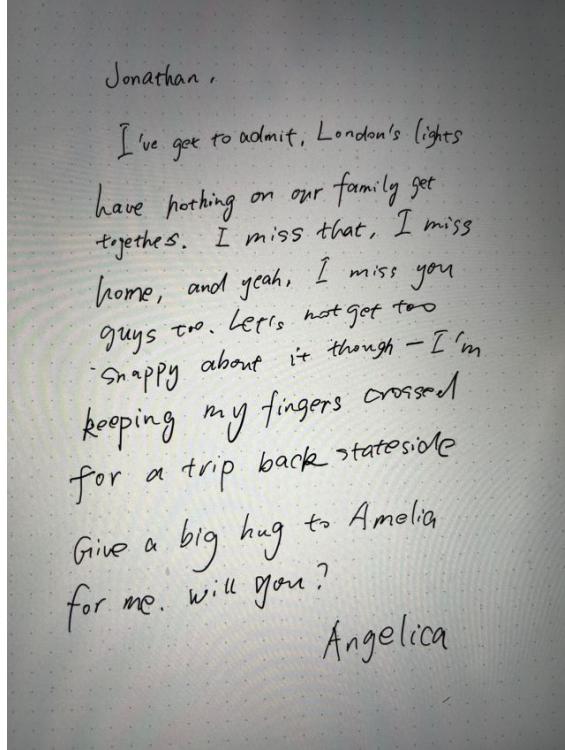
Prompt: What is the polygon mask of region
`<loc_102><loc_7><loc_375><loc_648>`



Figure 16. Region to segmentation prediction results.

F. Comparision with LMMs on Detailed Caption Task

Comparison with LMMs on Detailed Image Caption Task



Florence-2: The image is a photograph of a handwritten note from a person named Angelica. The note is written in black ink on a white background. The handwriting is neat and legible, and it appears to be a handwritten letter. The person's name, Jonathan, is written at the top of the note in a cursive font. Below the name, there is a note that reads, "I've got to admit, London's lights have nothing on our family get together. I miss that, I miss you guys too. Let's not get too happy about it though I'm keeping my fingers crossed for a trip back stateside. Give a big hug to Amelia for me. Will you?"

GPT-4V^a: The image shows a handwritten note addressed to Jonathan. The note conveys a personal message from the sender, Angelica, who admits to missing family gatherings and the feeling of home, contrasting it with the lights of London. Angelica expresses a longing to return stateside and asks Jonathan to give Amelia a big hug on her behalf. The tone of the note is casual and warm, with a sense of fondness and a hint of homesickness. The note is written informally, suggesting a close relationship between the sender and the recipients.

LLaVA^b: The image is a handwritten letter on a piece of paper, written by a person named Jonathan. The letter is addressed to someone named Angela, and it appears to be a heartfelt message expressing love and affection. The letter is written in cursive, giving it a personal and intimate touch.

MiniGPT-4^c: The image is a handwritten letter from an unknown person to someone named Angel. The letter is written in black ink on a piece of white paper. The letter is addressed to Angel and begins with the words, "Dear Angel,". The letter goes on to describe the writer's feelings about a recent event or situation, but the details are unclear. The letter ends with the words, "Sincerely, [Writer's Name]."

^a<https://chat.openai.com>

^b<https://llava.hliu.cc>

^c<https://huggingface.co/spaces/Vision-CAIR/minigpt4>

Figure 17. Comparison with LMMs on detailed caption task.

Comparison with LMMs on Detailed Image Caption Task



Florence-2: The image is a close-up of a red apple with a green label on it. The label has the text "FUJI 94131 USA" and a barcode on it, indicating that the apple is from the United States. The apple appears to be fresh and has a shiny texture. The background is white, making the apple stand out.

GPT-4V^a: The image displays a Fuji apple with a sticker indicating it is organic, along with the identifying number 94131, and specifying that it is from the USA. The apple has a reddish-pink skin with some yellow and green coloration, which is characteristic of Fuji apples. The apple is placed on a white background, which appears to be a textured fabric.

LLaVA^b: The image features a large, fresh, and ripe red apple sitting on a white surface. The apple has a sticker on it, which is likely a barcode or a label providing information about the product. The apple is prominently displayed, taking up a significant portion of the image.

MiniGPT-4^c: The image shows a red apple with a label on it that reads "Organic" in white letters. The apple is sitting on a white surface.

^a<https://chat.openai.com>

^b<https://llava.hliu.cc>

^c<https://huggingface.co/spaces/Vision-CAIR/minigpt4>

Figure 18. Comparison with LMMs on detailed caption task (continued).

Comparison with LMMs on Detailed Image Caption Task



Florence-2: The image shows a highway with multiple lanes. There are several vehicles on the road, including a black van and a red truck. On the left side of the image, there is a white semi-truck with the word "Martini" written on it. Above the highway, there are several bridges with cars and trucks on them. The sky is blue and there are trees and grass on the right side. In the center of the highway is a green sign that reads "Exit 94, Chamblee Tucker Rd, Mercer Univ."

GPT-4V^a: The image shows a highway scene with several lanes of traffic. There are overpasses with cars and trucks traveling over them, and the highway itself appears to be busy with vehicles, including a prominent black van in the foreground. A road sign indicates that Exit 94 for Chamblee-Tucker Rd and Mercer University is 1 1/4 miles ahead. The presence of the sign suggests that drivers in the right lanes will soon have the opportunity to take the exit to reach these destinations. The infrastructure suggests a well-traveled interstate or highway system, possibly in an urban or suburban area given the overpass complexity and road signage for local roads and institutions. The vehicles on the road are diverse, indicating a mix of commercial and personal transport.

LLaVA^b: The image depicts a busy highway with multiple lanes of traffic, including cars and trucks. There is a large truck driving on the highway, and several other trucks are also visible in the scene. The highway is filled with cars, and there is a traffic sign above the lanes, indicating the direction to Tucker Road. The scene captures the hustle and bustle of a typical day on a highway filled with vehicles.

MiniGPT-4^c: The image shows a highway with multiple lanes of traffic moving in both directions. There are several large bridges spanning the highway, and in the distance, there are mountains and a blue sky. There are also several large trucks and cars on the highway, including a black van in the center of the image. The road surface is smooth and well-maintained.

^a<https://chat.openai.com>

^b<https://llava.hliu.cc>

^c<https://huggingface.co/spaces/Vision-CAIR/minigpt4>

Figure 19. Comparison with LMMs on detailed caption task (continued).

Comparison with LMMs on Detailed Image Caption Task



Florence-2: The image shows a kitchen countertop with a laptop on it. The laptop is open and the screen is displaying a blue abstract design. On the countertop, there are various kitchen utensils hanging on a metal rack. The rack is attached to the wall above the sink and has a variety of kitchen tools hanging on it, including a grater, a spatula, a ladle, a spoon, and a whisk. There are also a few other kitchen items scattered around the counter. The countertop is made of wood and there is a wire rack at the bottom of the image.

GPT-4V^a: The image shows a modern kitchen setting where culinary tools are organized neatly on a rack against the wall. These tools include a variety of utensils such as a grater, tongs, a whisk, ladles, and spatulas. On the counter, there is an open laptop with a vibrant blue wallpaper on the screen, which suggests that someone might be using it for a recipe or as entertainment while cooking. In front of the laptop, there appears to be a cutting board with some chopped ingredients and a knife, indicating that food preparation is underway or has recently taken place. There are also some dried chilies and a plastic container that might contain other ingredients or spices, further suggesting cooking activities. The counter also has an electrical outlet and some covered sockets, and below the main counter, there are storage bins that likely hold additional kitchen supplies. The environment looks lived-in and practical, with a mix of technology and traditional cooking tools.

LLaVA^b: The image features a wooden kitchen counter with a laptop computer sitting on top of it. The laptop is open and displaying a picture of a cake. The counter is also cluttered with various kitchen utensils, including multiple spoons and knives, as well as a bowl. The scene suggests that someone might be using the laptop for cooking or baking purposes, possibly following a recipe or watching a tutorial.

MiniGPT-4^c: The image shows a kitchen counter with a laptop computer on it, a set of cooking utensils hanging from a rack above it, and a knife block on the counter. The laptop appears to be open, with a blue screen displaying an image of a kitchen. The utensils hanging from the rack include a spatula, a ladle, a whisk, and a cutting board. The knife block contains several knives, including a chef's knife, a paring knife, and a bread knife.

^a<https://chat.openai.com>

^b<https://llava.hliu.cc>

^c<https://huggingface.co/spaces/Vision-CAIR/minigpt4>

Figure 20. Comparison with LMMs on detailed caption task (continued).

G. Comparison with Kosmos-2 on Detailed Caption and Grounding Tasks

Comparison on detailed caption and grounding tasks.



The image features a home office with **two white desks**, one on the left and the other on the right. The desks are connected by **a white metal frame**, creating a functional and stylish workspace. **A laptop** is placed on the desk on the far left, and **a keyboard** is visible on the other desk. **Two chairs** are placed in front of the desks, one closer to the left desk and the second one on its right side.

(a) Result from Kosmos-2.



The image shows a modern home office setup with two white **desks**¹ and **two black chairs**². The **desks**³ are arranged in a corner of the room with a large window on the left side. On the right side of the desk, there is **a computer monitor**⁴, **a keyboard**⁵, a mouse, **a printer**⁶, and a **laptop**⁷. Above the computer monitor and keyboard, there are two framed **pictures**⁸ hanging on the wall. **The walls**⁹ are painted in a light green color and **the floor**¹⁰ is made of light-colored wood. **The floor**¹¹ is covered with a beige area rug with a geometric pattern. The overall style of the space is minimal and contemporary.

(b) Result from Florence-2.

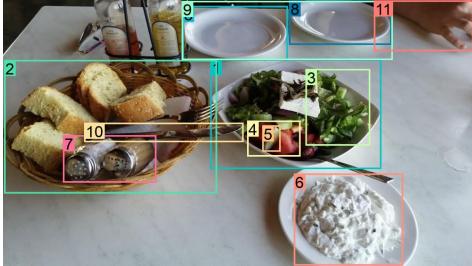
Figure 21. Systematic comparison with Kosmos-2 [60] on detailed caption and grounding tasks. The models generate both the detailed caption and grounding results. The results of Kosmos-2 are from <https://huggingface.co/spaces/ydshieh/Kosmos-2>.

Comparison on detailed caption and grounding tasks.



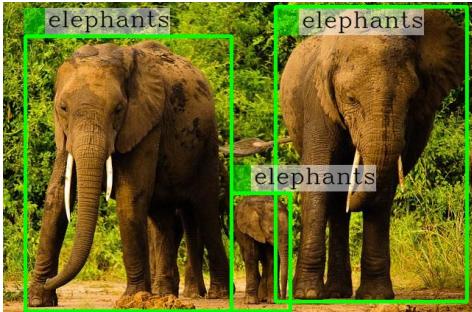
The image features a white dining table with a variety of food items on it. On the table, there is a bowl of bread, a bowl with a salad, and a plate with a side of yogurt. In addition to the food, there are **two cups** placed on the table. One cup is located near the left side of the table and the other cup is situated on the right side. The table is set with a fork and a knife, ready for a meal.

(c) Result from Kosmos-2.



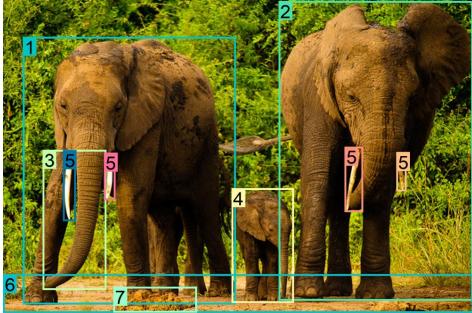
The image shows a table with a basket of bread and a plate of **salad**¹. **The basket**² is made of woven straw and has several slices of bread in it. Next to the basket, there is **a small bowl**³ of salad with a variety of vegetables, including **lettuce**⁴, **tomatoes**⁵, **cucumbers**⁶, and **feta cheese**⁷. There are also **two salt**⁸ and pepper **shakers**⁹ on the table. On the right side of the table, there are **two white plates**¹⁰ with a dollop of white sauce on them. The table is covered with a white tablecloth and there are **a few other dishes**¹¹ and **utensils**¹² scattered around. **A person's hand**¹³ can be seen in the top right corner of the image.

(d) Result from Florence-2.



The image features a family of **elephants** walking together in a lush green forest. The **elephants** are walking in a line, with one of them walking in front of the rest of the family. The family consists of a mother and her two calves, with the mother and the two calves walking side by side. The forest is filled with green trees, adding to the natural beauty of the scene.

(e) Result from Kosmos-2.



The image shows a group of three elephants standing in a dirt field with trees and bushes in the background. **The elephants**¹ are standing close together, with the largest elephant in the center and two smaller ones on either side. **The largest elephant**² on the left is standing with its **trunk**³ extended, while **the smaller one**⁴ is standing next to it. **All three elephants**⁵ have **tusks**⁶ and appear to be in their natural habitat. **The ground**⁷ is covered in dirt and there is **a small pile of dirt**⁸ in front of them. The overall mood of the image is peaceful and serene.

(f) Result from Florence-2.

Figure 22. Systematic comparison with Kosmos-2 [60] on detailed caption and grounding tasks. The models generate both the detailed caption and grounding results. The results of Kosmos-2 are from <https://huggingface.co/spaces/ydshieh/Kosmos-2>. (continued)