# All You Need to Know About Florence-2!

Sunidhi Ashtekar

8 min read · Jun 25, 2024

Listen Share More

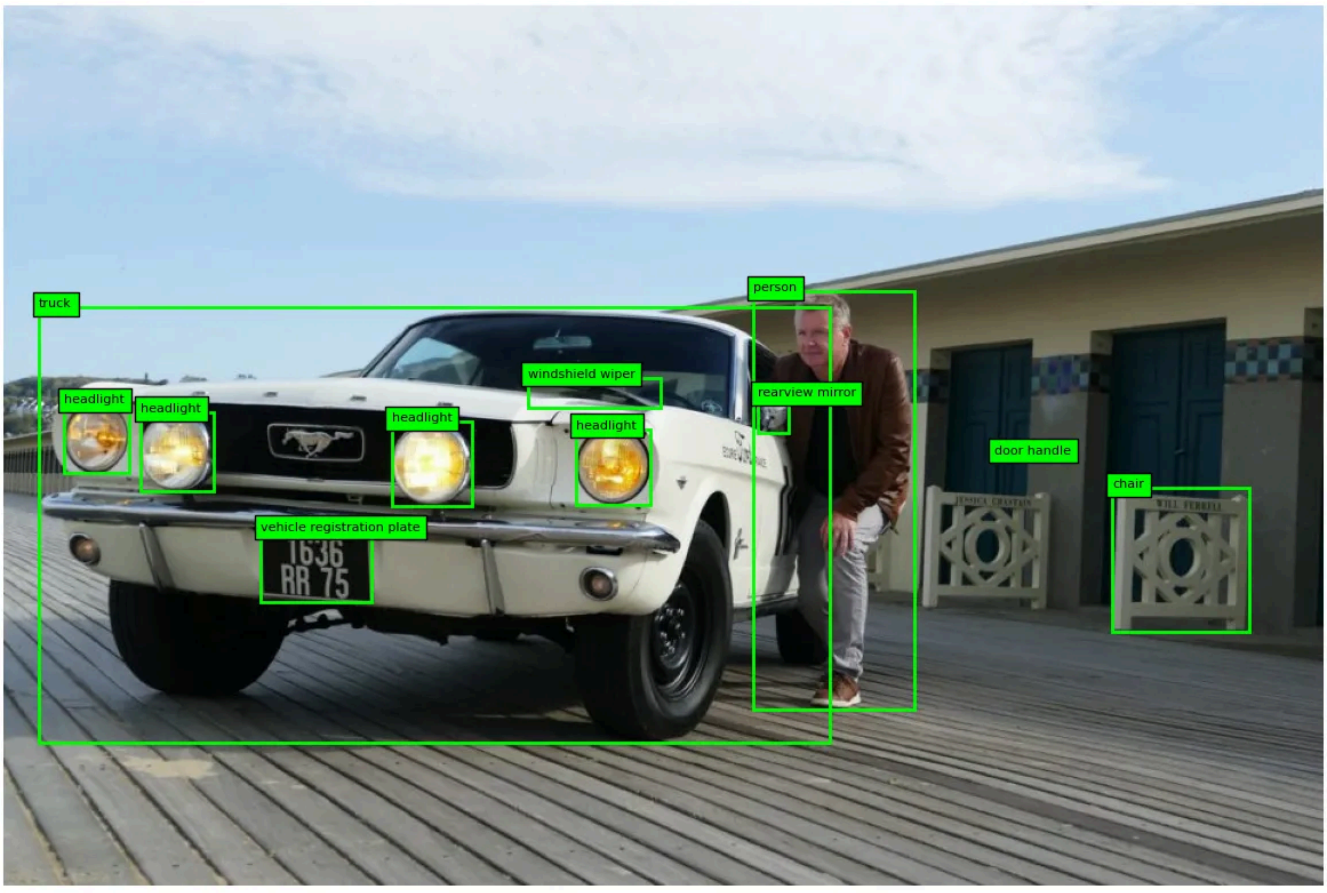



Image from Florence-2 Official Repository

## Introduction

Florence-2 is a lightweight vision-language foundation model developed by Microsoft Azure AI and open-sourced under the MIT license. It aims to achieve a unified, prompt-based representation for diverse vision and vision-language tasks, including captioning, object detection, grounding, and segmentation. Despite its

compact size, Florence-2 rivals much larger models like Kosmos-2 in performance. Its strength lies in the extensive FLD-5B dataset, which consists of 126 million images and 5.4 billion annotations, enabling robust zero-shot and fine-tuning capabilities.

### Model Architecture

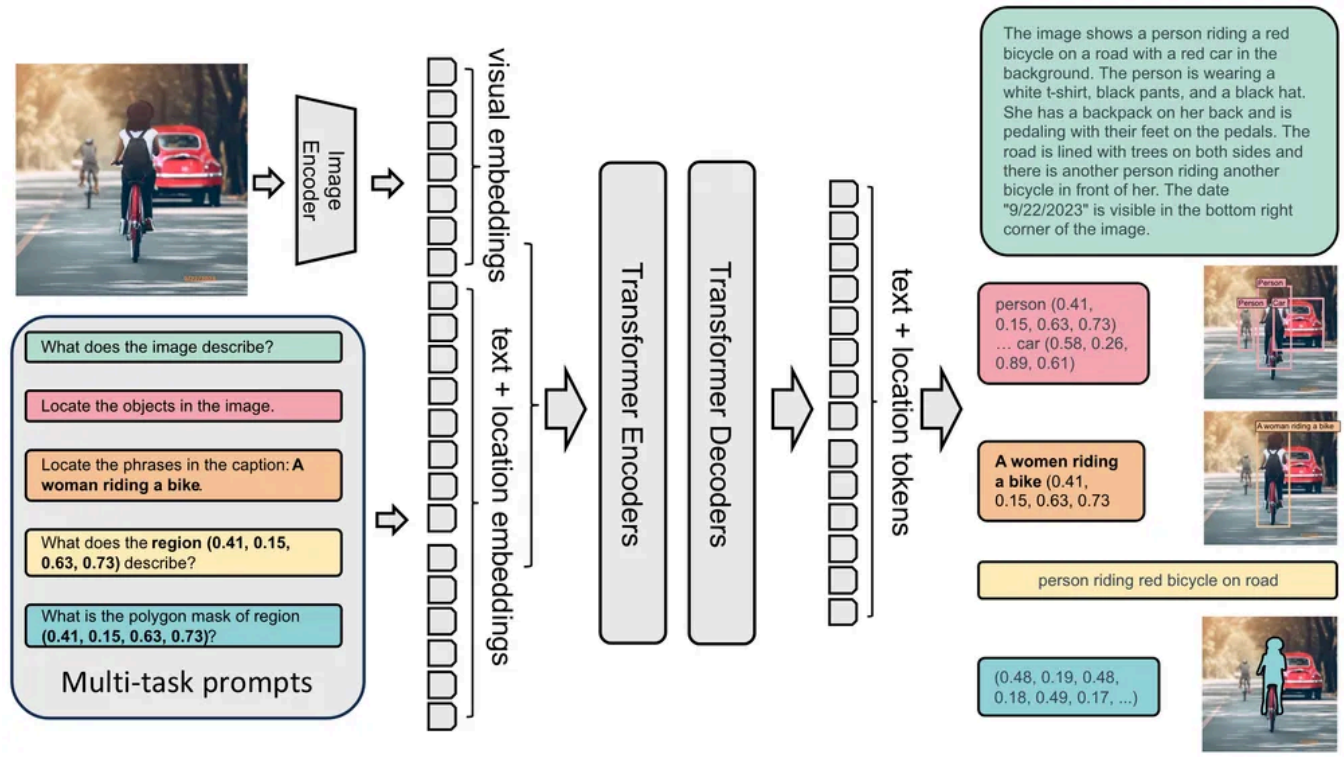



Image from Florence-2 Official Paper

1. **Image Encoder DaViT (Diverse Vision Transformer):**

Florence-2 utilizes the DaViT architecture as its image encoder, transforming input images into flattened visual token embeddings.

2. **Multi-Modality Encoder-Decoder:**



Tasks are handled as translation problems, where an image and a task-specific prompt are provided as input, resulting in a textual response.

For purely textual tasks, the input prompt and output are in text form.

3. **Region-Specific Tasks**:

For tasks involving specific regions, location. tokens representing coordinates are added to the tokenizer's vocabulary.

- **Box Representation (x0, y0, x1, y1)**: Location tokens correspond to the top-left and bottom-right corners of a box.
- **Polygon Representation (x0, y0, ..., xn, yn)**: Location tokens represent the vertices of a polygon in clockwise order.

The model takes images and task prompts as input, generating the desired results in text format. It uses a DaViT vision encoder to convert images into visual token embeddings. These embeddings are then concatenated with BERT-generated text embeddings and processed by a transformer-based multi-modal encoder-decoder to generate the response.

### Dataset FLD-5B

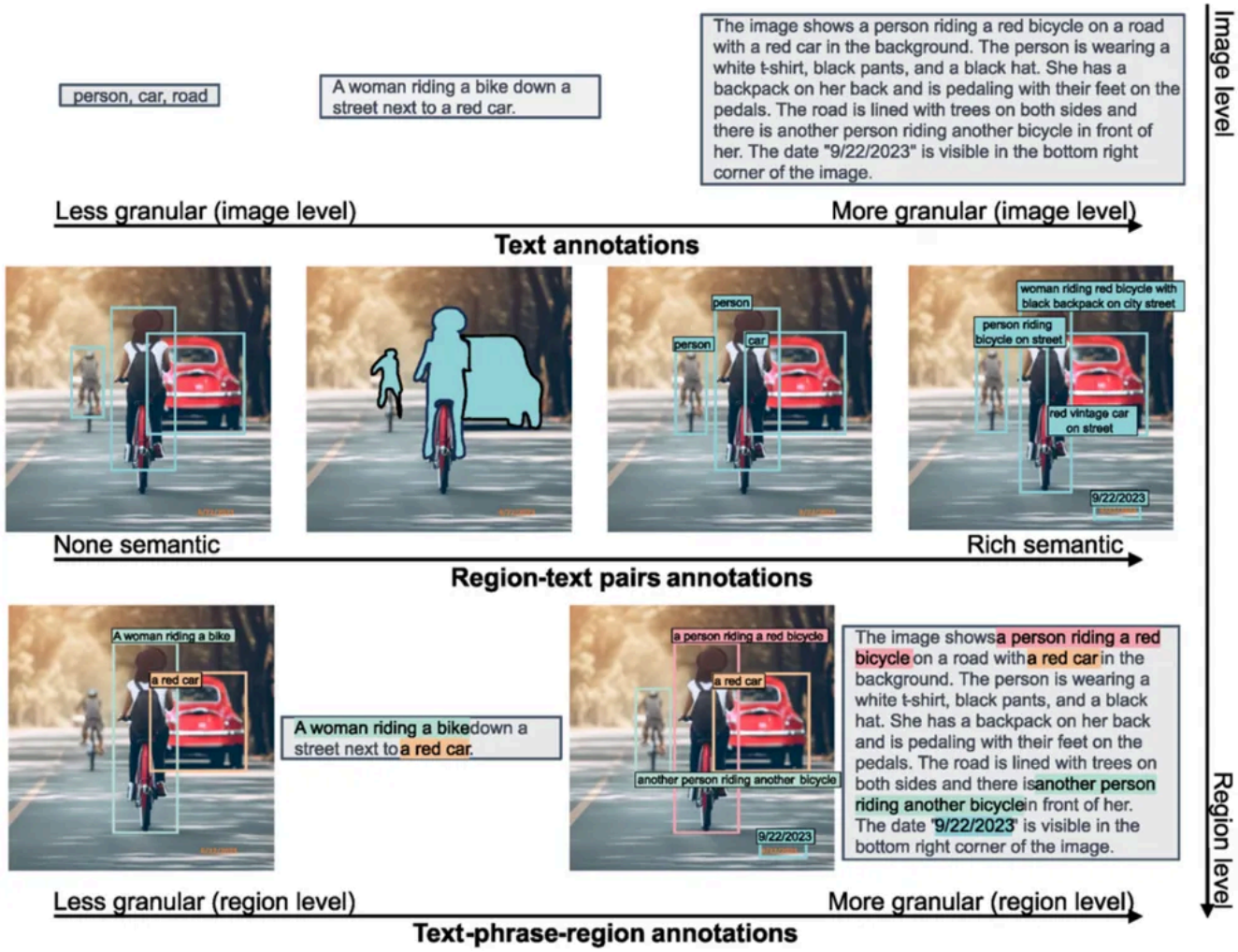



Image from Florence-2 Official Paper

Florence-2's exceptional performance is rooted in its training on the FLD-5B dataset. This dataset was meticulously created using an iterative strategy involving automated image annotation and model refinement, ensuring both scale and quality. The lack of large, unified datasets necessitated the creation of FLD-5B, as existing datasets like SA-1B and COCO are either limited in scope or size. Manual labelling being prohibitively expensive, the authors opted for automation using specialized models.

1. **Automated Annotation**: The dataset was built using a strategy that leverages automated image annotation and iterative model refinement. This approach ensures that the dataset is comprehensive and consistently updated with high-quality annotations.

2. **Annotation Types**: FLD-5B includes a variety of annotations listed as follows,

- **Image-Level Annotations:** Providing overall image descriptions and categorizations.

- **Region-Level Annotations**: Highlighting specific regions within images, such as objects or areas of interest.
- **Pixel-Level Annotations**: Offering detailed information at the pixel level, essential for tasks like segmentation.
- **Granularity**: The dataset contains boxes, masks, and various levels of captions, providing rich and detailed annotations.

### 3. Data Engine:

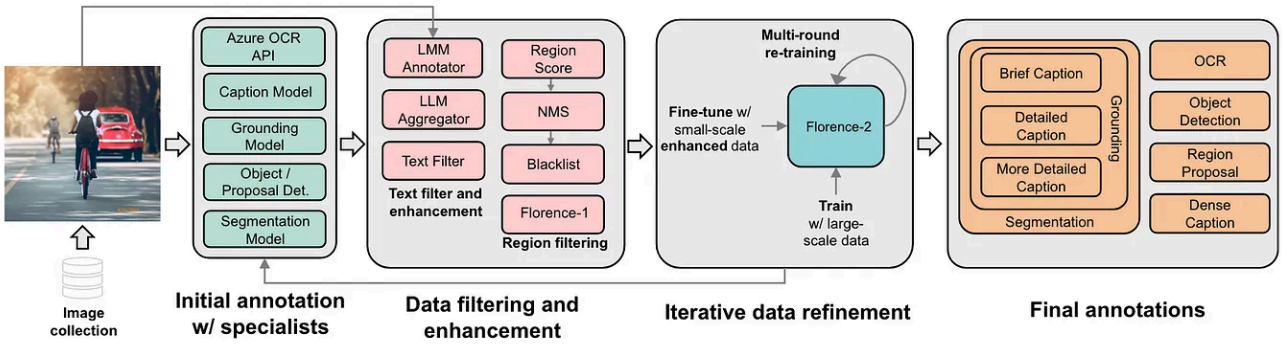



Figure 3. ***Florence-2* data engine** consists of three essential phrases: (1) initial annotation employing specialist models, (2) data filtering to correct errors and remove irrelevant annotations, and (3) an iterative process for data refinement. Our final dataset (***FLD-5B***) of over **5B** annotations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

Image from Florence-2 Official Paper

The creation of FLD-5B involved several stages listed below,

- **Initial Annotation**: Images are collected from various sources and annotated using specialist models such as the Azure OCR API for text recognition, a Caption Model for generating image captions, a Grounding Model for linking phrases to image regions, an Object/Proposal Detector for identifying regions of interest, and a Segmentation Model for segmenting images into different regions.
- **Data Filtering**: The initial annotations undergo a rigorous filtering process to correct errors and remove irrelevant data. Key components include LMM Annotator and LLM Aggregator for text data, a Text Filter for enhancing textual annotations, and Region Filtering using techniques like Region Score, Non-Maximum Suppression (NMS), and Blacklist filtering. Florence-1 is also used for additional filtering to ensure data quality.

- **Iterative Refinement:** The annotations are refined through multi-round re-training and fine-tuning of the Florence-2 model. The process begins with small-scale enhanced data and progresses to large-scale data, iteratively improving annotation accuracy and quality.

- **Final Annotations:** The resulting dataset, FLD-5B, includes 126 million images, 500 million text annotations, 1.3 billion region-text annotations, and 3.6 billion text-phrase-region annotations. Final annotations include brief, detailed, and more detailed captions, grounding, segmentation, OCR, object detection, region proposals, and dense captions.

Notably, FLD-5B doesn't introduce new images, all images are sourced from existing computer vision datasets. By combining these elements, FLD-5B provides a rich and diverse set of visual data that allows Florence-2 to learn effectively across various tasks, ensuring robust performance and versatility.

### Key Contributions

1. **Unified Representation**

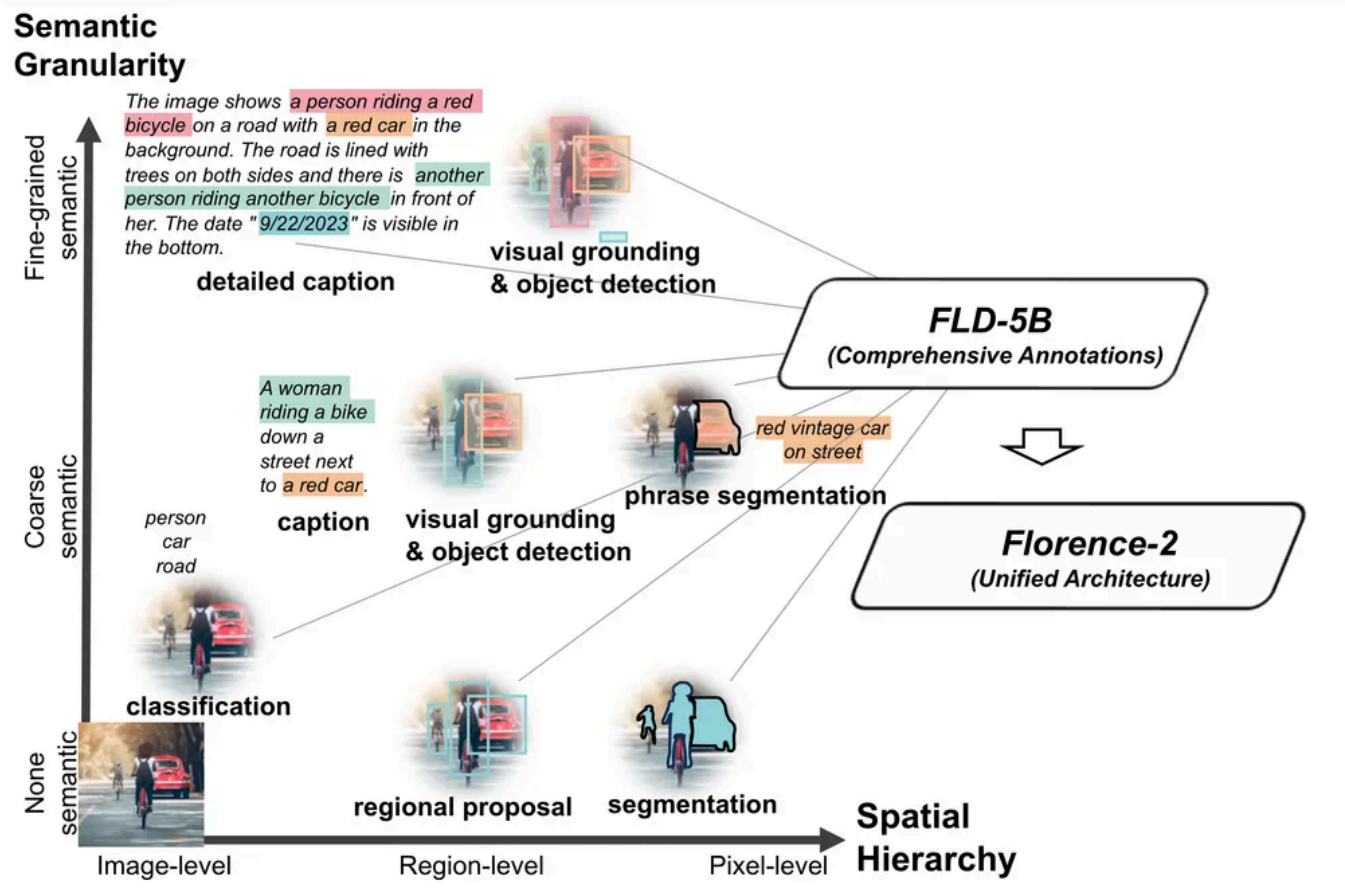


Image from Florence-2 Official Paper

Florence-2 employs a seq2seq framework to handle various vision tasks using a single model, simplifying the architecture and eliminating the need for task-specific modifications. Vision tasks differ in spatial hierarchy and semantic granularity; instance segmentation offers detailed object locations without semantic context, while image captioning captures relationships but not exact locations. By unifying these representations, Florence-2 effectively manages over 10 tasks with a single model, streamlining the process and enhancing versatility.

**2. Comprehensive Dataset**

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

Image from Florence-2 CVPR 2024 poster

The extensive FLD-5B dataset, provides a rich foundation for the model. This dataset includes image-level, region-level, and pixel-level annotations, enabling Florence-2 to learn from a wide range of visual data and deliver robust performance across different tasks.

**3. Multitask Learning**

Florence-2's architecture and training strategy allow it to handle diverse vision tasks effectively. The model demonstrates strong performance in both zero-shot and fine-tuned settings, showcasing its ability to adapt to various vision challenges without needing task-specific adjustments.

**Vision Tasks Performed by Florence-2**

Florence-2 is designed to handle a variety of vision and vision-language tasks through its unified, prompt-based representation. The key vision tasks performed by Florence-2 include:

- **Caption**: Generating brief textual descriptions of images, capturing the essence of the scene.
- **Detailed Caption**: Producing more elaborate textual descriptions, providing richer information about the image.
- **More Detailed Caption**: Creating comprehensive textual descriptions that include extensive details about the image.
- **Region Proposal**: Identifying regions of interest within an image to focus on specific areas.
- **Object Detection**: Locating and identifying objects within an image, providing bounding boxes and labels for each detected object.
- **Dense Region Caption**: Generating textual descriptions for densely packed regions within an image.
- **Phrase Grounding**: Associating phrases in a text description with specific regions in an image, linking textual descriptions to visual elements.
- **Referring Expression Comprehension**: Identifying regions in an image that correspond to natural language expressions, making it adept at tasks that require fine-grained visual-textual alignment.
- **Open Vocabulary Detection**: Detecting objects in an image using a flexible and extensive vocabulary.
- **Referring Segmentation**: Segmenting regions in an image based on referring expressions, providing detailed object boundaries.
- **Region to Text**: Converting regions of an image into corresponding textual descriptions.
- **Text Detection and Recognition**: Detecting and recognizing text within an image, providing both text and region information.

Florence-2's ability to perform these tasks with a unified model architecture highlights its versatility and robustness, setting new standards in the field of computer vision and multimodal AI.

### Experiments and Results

1. **Zero-Shot Performance**

- **Image Captioning:** Florence-2 achieved a remarkable 135.6 CIDEr score on the COCO caption benchmark.
- **Visual Grounding and Referring Expression Comprehension:** The model set new state-of-the-art performance levels, outperforming existing models across various benchmarks in these tasks.

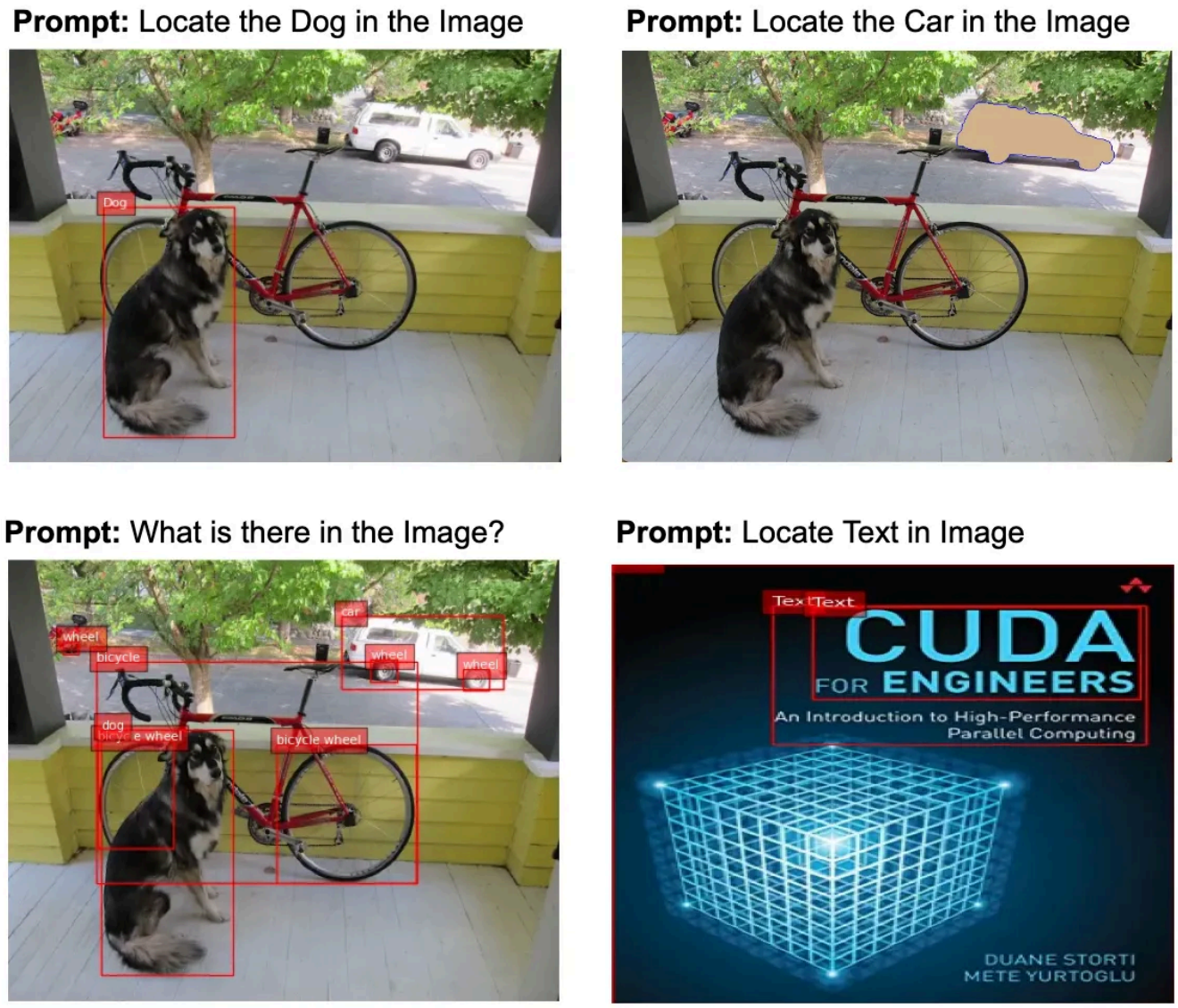
**2. Fine-Tuning**

- **State-of-the-Art Results:** Fine-tuning with public human-annotated data enabled Florence-2 to compete with larger specialist models, achieving new state-of-the-art results on benchmarks like RefCOCO, RefCOCO+, and RefCOCOg.
- **Object Detection and Segmentation:** The model showed improved performance on COCO object detection and ADE20K semantic segmentation tasks, surpassing both supervised and self-supervised models.

**3. Efficiency and Scalability**

- **Training Efficiency:** Pre-training on the FLD-5B dataset significantly enhanced training efficiency, allowing Florence-2 to achieve better results with fewer epochs compared to models pre-trained on ImageNet.
- **Model Scaling:** Florence-2 demonstrated that increasing the model's capacity leads to improved zero-shot performance across various tasks, showcasing its scalability and robustness.
- **Model Variants:** Florence-2 is available in two variants: Florence-2-base with 0.23 billion parameters and Florence-2-large with 0.77 billion parameters. These smaller sizes enable deployment even on mobile devices.

- **Comparison with Kosmos-2:** Despite its compact size, Florence-2 outperforms Kosmos-2 across all benchmarks, achieving better zero-shot results even though Kosmos-2 has 1.6 billion parameters.




Florence-2 Inference for Object Detection, Segmentation, Visual grounding tasks

## Conclusion

I hope this guide has provided a comprehensive overview of the newly launched Florence-2 vision foundation model, highlighting its innovative features and improvements over previous iterations. Florence-2 demonstrates remarkable capabilities in visual recognition, image generation, and multimodal understanding, making it a robust and versatile tool for various AI applications.

Stay tuned for my next blog where I will dive into the fine-tuning of Florence-2, providing insights and techniques to further enhance its performance for specific tasks.

If you have any questions, recommendations, or critiques, please don't hesitate to reach out on LinkedIn. I'm open to discussions and eager to hear your feedback or assist with any challenges you might encounter.

**References**

To further explore the concepts, I recommend visiting the following resources:

- Florence-2 Official Repository: Explore models and datasets related to Florence-2 on Hugging Face
- Florence-2 Official Paper
- Florence-2 Online Demo
- Images taken in this blog

This article illustrates the potential of Florence-2's cutting-edge architecture and innovations. To know more about implementation details and comparisons with other state-of-the-art models, go through the official paper. Whether for academic research, industry applications, or personal projects, Florence-2 offers a powerful and efficient solution for exploring and innovating within the vast domain of computer vision.

Multimodal | Computer Vision | Object Detection | Segmentation | Machine Learning