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 [‡]directional lead

Microsoft

Abstract

We introduce Florence-2*, a novel vision foundation model with a unified, prompt-based representation for various computer vision and vision-language tasks. While existing large vision models excel in transfer learning, they struggle to perform diverse 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.*

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 accross diverse applications. This trend is evident in natural language processing (NLP), where advanced models [6, 7, 19, 38, 52, 53] 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 tasks presents unique challenges, notably the need for comprehensive perceptual abilities. Unlike NLP, which deals mainly with text, computer vision requires handling intricate visual data like

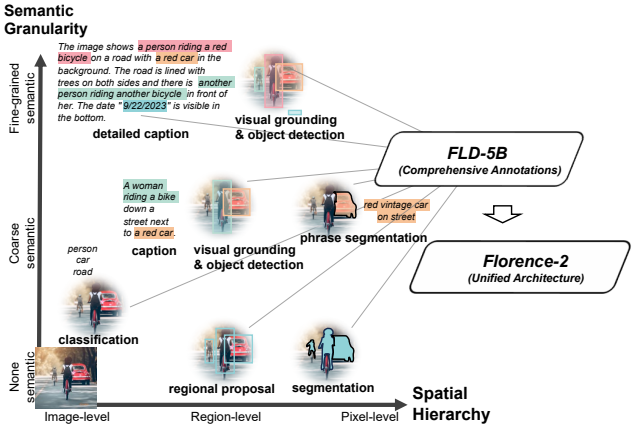

Figure 1. We aim to build a vision foundation model to enable extensive perception capabilities including spatial hierarchy and semantic granularity. To achieve this, a single unified model ***Florence-2*** is pre-trained on our ***FLD-5B*** dataset encompassing a total of 5.4B comprehensive annotations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

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.

This pursuit is characterized by distinctiveness and substantial challenges. A key hurdle is the scarcity of *com-*

prehensive 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 [41], and Flickr30k Entities [49], 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 [24, 75], semantic segmentation [16, 64], and image captioning [40, 61] with task-specific design. However, it is essential to develop a unified model capable of adapting across various vision tasks in a task-agnostic manner, even accommodating new tasks with minimal or no task-specific fine-tuning.

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 [30, 63, 67], 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 [17,19,53,59], which integrates an image encoder and a multi-modality encoder-decoder. This design accommodates a spectrum of 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) [52].

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 [41], visual grounding on Flick30k [49], and referring expression comprehension on RefCOCO/+/g [28, 45, 71].
- 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.

2. Rethinking Vision Model Pre-training

In pursuit of a vision foundation model, we revisit three predominant pre-training paradigms: supervised (*e.g.*, ImageNet classification [18]), self-supervised (*e.g.*, SimCLR [9], MoCo [23], BEiT [5], MAE [22]), and weakly supervised (*e.g.*, CLIP [51], Florence [73], SAM [29]). 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 [34]; 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 [51]. To build a unified vision model suitable for various applications, we must explore 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 [3, 12, 14, 15, 44, 62]. Our multitask learning approach incorporates three distinct learning objectives, each addressing a different level of granularity and semantic understanding:

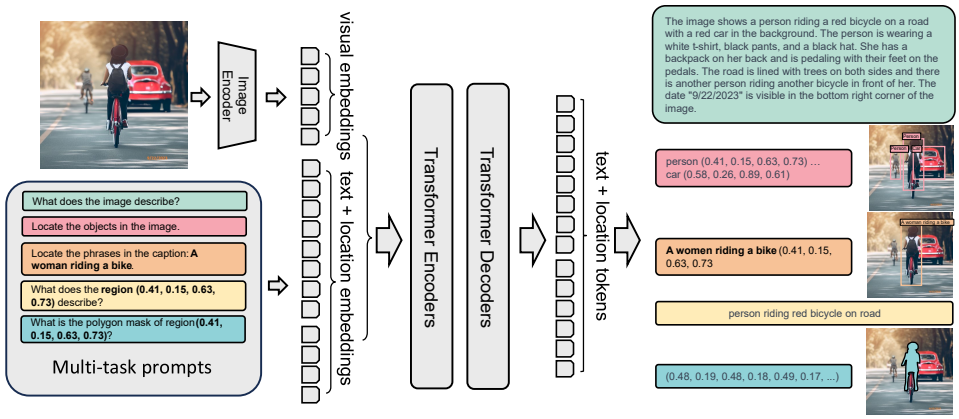

Figure 2. ***Florence-2*** consists of an image encoder and standard multi-modality encoder-decoder. We train ***Florence-2*** on our ***FLD-5B*** data in a unified multitask learning paradigm, resulting in a generalist vision foundation model, which can perform various vision tasks.

- **Image-level understanding** tasks capture high-level semantics and foster a comprehensive understanding of images through linguistic descriptions [13, 18, 31, 69]. They enable the model to comprehend the overall 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.

- **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.

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 [60], 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.

Task formulation. We adopt a sequence-to-sequence framework [10, 15, 44, 60] to address various vision tasks in a unified manner. As shown in **??**, 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, 44, 62], 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** $(x_0, y_0, ..., x_3, y_3)$: 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** $(x_0, y_0, ..., x_n, y_n)$: 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 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 $\mathbf{T}_{prompt} \in \mathbf{R}^{N_t \times D}$ using our extended language tokenizer and word embedding layer [38]. 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 [4] 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.

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

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

4. Data Engine

To train our *Florence-2* model, we require a comprehensive, large-scale, high-quality multitask dataset encompassing various image data aspects. 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 [55], Open Images [35], Conceptual Captions [56], and LAION [54] filtered by [40]. 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. **??** demonstrates examples of our annotations in our data. 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 and enhancement 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, which 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. If image datasets already contain partial annotations, like the Object 365 dataset [55] with human-annotated bounding boxes and categories, we combine these with our synthetic labels to improve annotation coverage and diversity. However, we exclude certain tasks, such as detailed text, from initial labeling due to the difficulty of achieving high-performance specialist models, adding them later during data refinement. Ultimately, this ensures our 126 million image dataset is thoroughly labeled across most annotation types.

Data filtering and enhancement. We employ a multi-faceted filtering process to address the noise and imprecesion in the initial annotations from specialist models. Inspired by DiHT [50], we develop a parsing tool using SpaCy [25] to filter text annotations, discarding texts with too many objects and selecting texts with a minimum complexity of actions and objects to maintain rich visual concepts. For region data, particularly bounding boxes, we eliminate those below a confidence threshold and apply non-maximum suppression to remove redundancies, enhancing the quality of our annotations. In addition, we

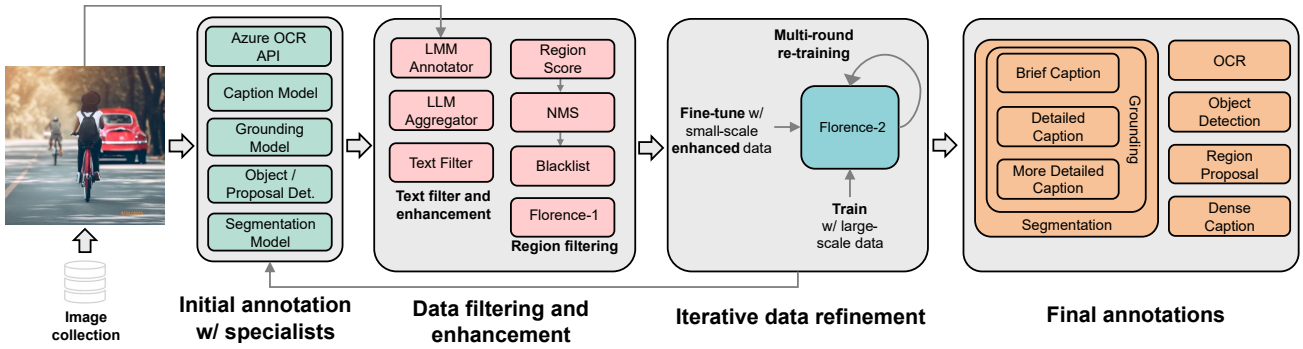

Figure 3. ***Florence-2* data engine** consists of three essential phrases: (1) initial annotation employing specialist models, (2) data filtering and enhancement to correct errors, and (3) an iterative process for data refinement. Our final dataset (***FLD-5B***) of over **5B** annotations contains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

also adopt large language model (i.e., GPT-4 [47]) and large multimodal model (i.e., GPT-4V [2]) to enhance text annotations with more details.

Iterative data refinement. We enhance our training dataset's quality through an iterative refinement process, using a multitask model trained on initially filtered annotations. The model shows improved accuracy, particularly where the initial labels are noisy or inaccurate. By integrating the model's refined predictions back into our dataset and retraining, we progressively improve our annotations. For tasks initially set aside due to data scarcity, we use the multitask model for pre-training, then fine-tune it with the limited data, achieving better results than training from scratch. This fine-tuned model then serves as a specialist for annotating our 126 million image dataset, ensuring extensive and accurate 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 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, *Florence-2* model is trained as the specialist on image-text datasets for initial annotations. Iterative refinement minimizes 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 provide descriptive textual annotations for semantic regions in images, including visual object and text regions, each enclosed within tight bounding boxes. These annotations offer varying levels of granularity, from phrases to sentences, enhancing region comprehension.

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 [75] 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 a specialist caption model with cropped image regions. Each region then receives three textual annotations: phrase from object category, brief text, and noun phrase chunks from the brief text. The Florence-1 [73] model determines the most similar textual annotation to each image region.

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 [43] identifies noun phrases and creates bounding boxes for them. Additionally, the SAM model [29] 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 blacklist is also used to exclude irrelevant noun phrases like pronouns and abstract concepts.

5. *FLD-5B* Dataset

Following the data engine, we build a large-scale training set (*FLD-5B*) of 126M images, more than **500M** text

Annotation Type	Text Type	#Image Anno.	#Avg Tokens	#Regions
Text	Brief	235M	7.95	-
	Detailed	126M	31.65	-
	More detailed	126M	70.53	-
Region-Text	Phrase	126M	-	681M
	Brief	126M	-	681M
Text-Phrase-Region	Brief	235M	7.95	1007M
	Detailed	126M	31.65	1289M
	More detailed	126M	70.53	1278M

Table 1. Annotation statistics of *FLD-5B* dataset.

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. The statistics for each annotation type within our dataset are presented in Table 1. We present the detailed analysis on *FLD-5B* dataset in **??**.

We provide a comparison between our dataset and the existing datasets that are commonly used for training foundation models in Table 2. Our dataset has several advantages over the previous ones, such as having more annotations in total and per image. Moreover, the annotations in our dataset span multiple levels of spatial and semantic granularity, which allows for more diverse and comprehensive visual understanding tasks.

6. Experiments

Our *Florence-2* models are trained on *FLD-5B* to learn a universal image representation. 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 and training setup are given in **????**. 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.

6.1. Zero-shot Evaluation Across Tasks

We present a powerful vision foundation model that does not require task-specific supervised annotations for fine-tuning. The **zero-shot** performance of our model is shown in Table 3. For image-level tasks, *Florence-2-L* achieves a 135.6 CIDEr score on the COCO caption benchmark [41], utilizing less than 1% of the parameters compared to the 80B Flamingo [3] 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 [49] Recall@1, and approximately 4%, 8%, and 8% absolute improvements on Refcoco, Refcoco+, and Refcocog [72], respectively, compared to the Kosmos-2 [48] model, which has 1.6B parameters. Additionally, our pre-trained model attains a 35.8% mIOU in the Refcoco referring expression segmentation (RES) [72] task, a capability not supported by prior foundation models.

6.2. 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 **??**. Tables 4 and 5 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 [42] on both RefCOCO REC task and RES task by 3.0 Accuracy@0.5 and 3.54 mIOU respectively, where PolyFormer [42] adapts specifically designed regression-based prediction head for coordinates. *Florence-2-L* also outperforms previous SOTA method UNINEXT [65] on RefCOCO by 0.8 Accuracy@0.5, where UNINEXT [65] is based on advanced object detector Deformable DETR [77] and DINO [75].

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 [27], 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 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 methods [12, 15].

Dataset	Rep. Model	#Images	#Annotations	Spatial hierarchy	Semantics granularity
JFT300M [58]	ViT [21]	300M	300M	Image-level	Coarse
WIT [51]	CLIP [51]	400M	400M	Image-level	Coarse
SA-1B [29]	SAM [29]	11M	1B	Region-level	Non-semantic
GrIT [48]	Kosmos-2 [48]	91M	137M	Image & Region-level	Fine-grained
M3W [3]	Flamingo [3]	185M	43.3M*	Multi-image-level	Fine-grained
FLD-5B (ours)	*Florence-2* (ours)	126M	5B	Image & Region-level	Coarse to fine-grained

Table 2. 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.

Method	#params	COCO Cap.	NoCaps	TextCaps	COCO Det.	Flickr30k	Refcoco			Refcoco+			Refcocog		Refcoco RES
		test	val	val	val2017	test	val	test-A	test-B	val	test-A	test-B	val	test	val
		CIDEr	CIDEr	CIDEr	mAP	R@1	Accuracy			Accuracy			Accuracy		mIoU
Flamingo [3]	80B	84.3	-	-	-	-	-	-	-	-	-	-	-	-	-
Kosmos-2 [48]	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 3. **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.

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.

7. Related Works

7.1. Vision-Language Foundation Models

Recent vision-language pre-training models [26, 51, 73] 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 [46, 57]. These models, trained on weakly large-scale image-text data, have been extended to more downstream tasks such as object detection, achieving state-of-the-art performance with task-specific adaptation heads. Other studies [3,40,61,70] use a multi-modality decoder for autoregressive text prediction, employing language modeling pre-training objectives. The methods for combining vision and language embeddings vary: GIT [61] concatenates vision and text tokens for the decoder input with a causal attention mask, and CoCa [70] utilizes attentional poolers with learnable queries to select task-specific vision tokens.

Beyond image captioning pre-training task, some research [15,44,62] 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, 44, 62]. 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 image-level captions. It shares the same encoder-decoder design as other multi-modality encoder-decoder models [15, 44] 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 evolution in computer vision has shifted from focusing on single-perspective datasets, like image classification [18], to multi-perspective, comprehensive annotations for each visual data point [32, 35, 41]. Datasets such as MS-COCO [13, 41] and Visual Genome [32] offer rich spatial and semantic annotations, enhancing model interactions across annotations. However, their size is limited due to the cost of human verification. Our large-scale datasets maintain comprehensive annotations, including text, region-text pairs, and text-phrase-region triplets, with less human involvement.

Scalable annotations. Over the past decade, vision datasets

Method	#params	COCO Caption Karpathy test CIDEr	NoCaps val CIDEr	TextCaps val CIDEr	VQAv2 test-dev Acc	TextVQA test-dev Acc	VizWiz VQA test-dev Acc
Specialist Models							
CoCa [70]	2.1B	143.6	122.4	-	82.3	-	-
BLIP-2 [39]	7.8B	144.5	121.6	-	82.2	-	-
GIT2 [61]	5.1B	145	126.9	148.6	81.7	67.3	71.0
Flamingo [3]	80B	138.1	-	-	82.0	54.1	65.7
PaLI [15]	17B	149.1	127.0	160.0$^{\triangle}$	84.3	58.8 / 73.1$^{\triangle}$	71.6 / 74.4$^{\triangle}$
PaLI-X [12]	55B	149.2	126.3	147 / 163.7$^{\triangle}$	86.0	71.4 / 80.8$^{\triangle}$	70.9 / 74.6$^{\triangle}$
Generalist Models							
Unified-IO [44]	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 4. 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. $^{\triangle}$ indicates usage of external OCR as input.

Method	#params	COCO Det.	Flickr30k	RefCOCO			RefCOCO+			RefCOCOg		RefCOCO RES
		val2017	test	val	test-A	test-B	val	test-A	test-B	val	test	val
		mAP	R@1	Accuracy			Accuracy			Accuracy		mIoU
Specialist Models												
SeqTR [76]	-	-	-	83.7	86.5	81.2	71.5	76.3	64.9	74.9	74.2	-
PolyFormer [42]	-	-	-	90.4	92.9	87.2	85.0	89.8	78.0	85.8	85.9	76.9
UNINEXT [65]	0.74B	60.6	-	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4	-
Ferret [68]	13B	-	-	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3	-
Generalist Models												
UniTAB [66]	-	-	-	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	78.0
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 5. 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.

have rapidly scaled up from thousands [33, 37] to billion examples [26, 74] to encompass more visual concepts for better generalization. This shift is evident in recent foundation models that employ massive quantities of data [6]. These large datasets typically collect images from the web and parse noisy annotations from the corresponding metadata, such as category label from query [58, 74], short description from alt-text [26, 51], as well as detailed description from interleaved text [3, 36]. Despite their diversity, these annotations suffer from randomness and limited types (*i.e.*, texts only). Some works [29, 40] 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 diverse 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 exhibited 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. Experimental results highlight its potent universal representation, significantly enhancing a wide range of downstream tasks.

References

[1] Azure ai services. https://azure.microsoft.com/en-us/products/ai-services?activetab=pivot:azureopenaiservicetab. Accessed: 2023-10-13. 5

[2] Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Accessed: 2023-11-17. 5

[3] 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. 2, 6, 7, 8

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 4

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 2

[6] 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, 8

[7] 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

[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. 3, 7

[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. 3

[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. 2, 6, 8

[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, 5, 7

[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. 2

[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. 2, 3, 6, 7, 8

[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, 7

[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

[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] 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

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF con-*

ference on computer vision and pattern recognition, pages 9729–9738, 2020. 2

[24] 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

[25] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020. 4

[26] 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. 7, 8

[27] 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. 6

[28] 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

[29] 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, 5, 7, 8

[30] 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

[31] 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

[32] 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. 7

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 8

[34] 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

[35] 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

[36] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, 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. 8

[37] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 8

[38] 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

[39] 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. 8

[40] 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, 4, 7, 8

[41] 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, 6, 7

[42] 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. 6, 8

[43] 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. 5

[44] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks, 2022. 2, 3, 7, 8

[45] 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

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 7

[47] OpenAI. Gpt-4 technical report, 2023. 5

[48] 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. 6, 7

[49] 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 *Pro-*

ceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015. 2, 6

[50] 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. 4

[51] 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, 8

[52] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1, 2

[53] 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

[54] 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

[55] 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

[56] 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

[57] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 7

[58] 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. 7, 8

[59] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[60] 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. 3

[61] 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, 7, 8

[62] 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. 2, 3, 7

[63] Nic M Weststrate, Susan Bluck, and Judith Glück. Wisdom of the crowd. *The Cambridge handbook of wisdom*, pages 97–121, 2019. 2

[64] 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

[65] 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. 6, 8

[66] 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. 8

[67] 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

[68] 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. 8

[69] 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

[70] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 7, 8

[71] 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

[72] 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. 6

[73] 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, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2, 5, 7

[74] 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. 8

[75] 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, 5, 6

[76] 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. 8

[77] 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. 6