

Supplementary Material:

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

A. Data Analysis on *FLD-5B*

A.1. Annotation Statistics

The statistics for each annotation type within our dataset are presented in Table 1.

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 [3]. These lengthy 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 [17] and Object 365 [34]. 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 text based on the Florence-1 score.

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.

A.2. Semantic Coverage

Our text annotations include various types, analyzed using SpaCy [11] for semantic coverage. We categorize tokens based on part-of-speech tags into types, *e.g.*, objects, attributes, actions, and proper nouns, and introduce *token complexity* based on their connections in the dependency parsing tree. This study mainly focuses on the complexity

of objects and actions. Table 2 presents the statistics on the average number of semantic elements and their corresponding complexity. The results indicate that more detailed text annotations lead to an increase in semantic elements and their complexity, especially for actions, which significantly outnumber those in brief texts. This suggests traditional brief texts are less effective in describing image actions. Objects and actions in detailed texts have more semantic connections, with actions showing a significant increase in complexity.

A.3. 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 1a, 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 image concepts. In Figure 1b, 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. 1c and 1d, indicate a center bias, with region-text pairs displaying a more uniform distribution than text-phrase-region triplets.

B. Experiments

B.1. Setup for Pre-training

We initialize the weights of the image encoder and multi-modality encoder-decoder from UniCL [42] and BART [18], respectively. We adopt AdamW [25] with cosine learning rate decay [24] for training our models. We

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

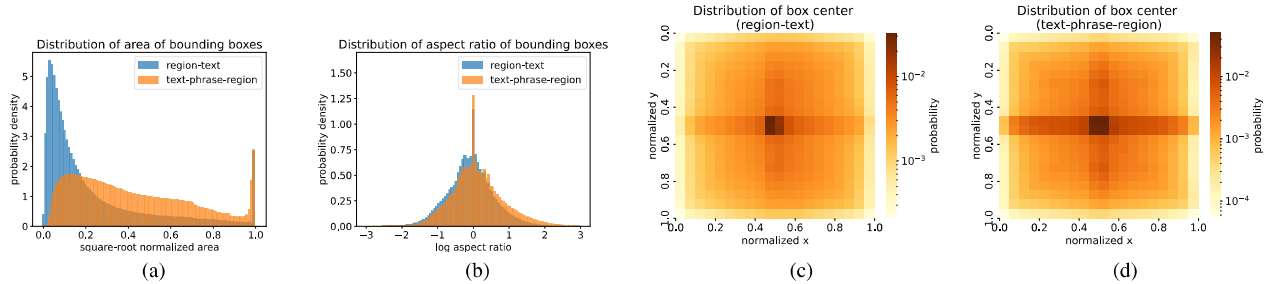


Figure 1. Distributions of bounding boxes in *FLD-5B* dataset.

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

leverage Deepspeed [32] 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 [4, 12, 31, 43, 45], 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.

B.2. 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 [21] experiments with Mask R-CNN [9], and COCO object detection [21] experiments with DINO [46] to further demonstrate the effectiveness of *Florence-2* pre-training. We train on the *train2017* split and evaluate on the *val2017* split.

For Mask R-CNN [9] experiments, we follow the common setup used in [22, 46], 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 [22, 37, 40, 41], or 100 epochs in [19], to achieve better results.

For DINO [46] experiments, we train DINO-4scale [46] detector for 12 epochs ($1 \times$) using the same data augmentation strategy as employed by [2].

First, our base model achieves a strong performance improvement compared to other approaches. As shown in Table 3, our DaViT-B model pre-trained by *Florence-2* sur-

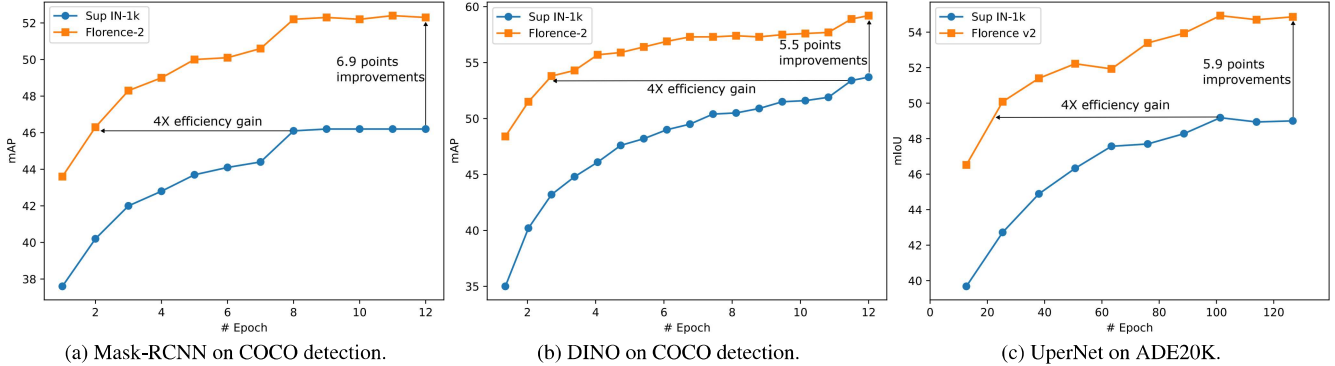


Figure 2. Training efficiency on COCO object detection and segmentation, and ADE20K semantic segmentation tasks.

passes previous best base model (ConvNext v2-B), which is pre-trained by FCMAE [37], by 0.7 AP_b using Mask RCNN. Importantly, while ConvNeXt v2-B leverages a $3\times$ schedule (36 epochs), our model efficiently employs a $1\times$ 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 4 and Figure 2, 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 4 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 [38] framework on ADE20k [47] dataset. We mostly follow the training and evaluation protocols from Swin [22]. Specifically, we use input size 512×512 and train the model for 40k iterations with a batch size of 64. We adopt the AdamW [25] 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 5, our base model outperforms the previous SoTA model, which is BEiT pre-trained ViT-B [1], by 1.3 and 1.4 points in single-scale and multi-scale testing protocol, respectively. With the same backbone architecture of DaViT-B [5], *Florence-2* pre-trained model achieves a remarkable improvement of 4.9 points and $4\times$ efficiency compared to the ImageNet-1k

Backbone	Pretrain	Mask R-CNN		DINO
		AP_b	AP_m	AP
ViT-B [19]	MAE, IN-1k	51.6	45.9	55.0
Swin-B [22]	Sup IN-1k	50.2	-	53.4
Swin-B [22]	SimMIM [39]	52.3	-	-
FocalAtt-B [41]	Sup IN-1k	49.0	43.7	-
FocalNet-B [40]	Sup IN-1k	49.8	44.1	54.4
ConvNeXt v1-B [23]	Sup IN-1k	50.3	44.9	52.6
ConvNeXt v2-B [37]	Sup IN-1k	51.0	45.6	-
ConvNeXt v2-B [37]	FCMAE	52.9	46.6	-
DaViT-B [5]	<i>Florence-2</i>	53.6	46.4	59.2

Table 3. **COCO object detection and instance segmentation results** using Mask-RCNN framework, and **COCO object detection results** using DINO-scale 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 [42]	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 4. 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.

pre-trained counterpart as demonstrated in Table 4 and Figure 2.

Backbone	Pretrain	mIoU	ms-mIoU
ViT-B [8]	Sup IN-1k	47.4	-
ViT-B [8]	MAE IN-1k	48.1	-
ViT-B [1]	BEiT	53.6	54.1
ViT-B [28]	BEiTv2 IN-1k	53.1	-
ViT-B [28]	BEiTv2 IN-22k	53.5	-
Swin-B [22]	Sup IN-1k	48.1	49.7
Swin-B [22]	Sup IN-22k	-	51.8
Swin-B [22]	SimMIM [39]	-	52.8
FocalAtt-B [41]	Sup IN-1k	49.0	50.5
FocalNet-B [40]	Sup IN-1k	50.5	51.4
ConvNeXt v1-B [23]	Sup IN-1k	-	49.9
ConvNeXt v2-B [37]	Sup IN-1k	-	50.5
ConvNeXt v2-B [37]	FCMAE	-	52.1
DaViT-B [5]	<i>Florence-2</i>	54.9	55.5

Table 5. **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 .

B.3. Ablation Studies

Multitask transfer. In this study, we aimed to identify the most effective pre-trained model for transfer learning across various downstream tasks in computer vision. We compared three different models, each pre-trained on a different combination of tasks:

- Image-level Model: pre-trained on image-level tasks only
- Image-Region Model: pre-trained on image-level and region-level tasks
- Image-Region-Pixel Model: pre-trained on image-level, region-level, and pixel-level tasks

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 3. 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).

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

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 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 11. We show the zero-shot performance on four downstream tasks in Table 6. 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 7 presents the zero-shot performance results on COCO caption, COCO object detection, Flickr30k grounding, and RefCOCO referring segmentation (RES) tasks. We

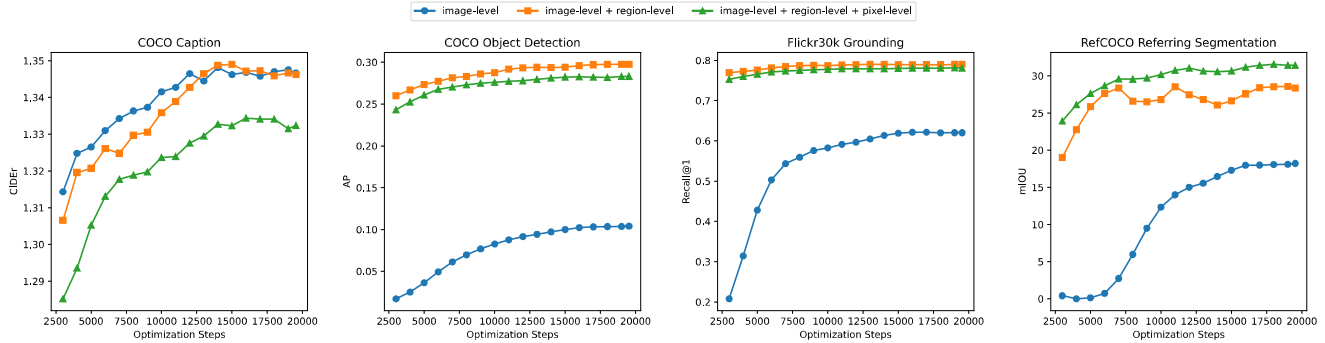


Figure 3. **Multitask transfer.** We conduct experiments with three different versions of *Florence-2* models, each trained on a different level of image annotation: image level, image and region level, and image, region, and pixel level. We then evaluate the transfer learning performance of these models on four downstream tasks: COCO caption, COCO object detection, Flickr30k grounding, and Refcoco referring segmentation.

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

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. 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 allows for rapid and efficient scaling of pre-training datasets, enabling exploration of the impact of larger data sizes on model performance in various computer vision tasks. This approach ensures a sustainable and scalable annotation process as demand for high-quality labeled data increases. In summary, our data scaling approach offers a more efficient

V Pre	L Pre	Caption	Detection	Grounding	RES	
		CIDEr	AP	Recall@1	mIOU	oIOU
<i>Freeze Vision Encoder</i>						
✓	✓	120.0	6.9	66.3	9.9	13.6
<i>Unfreeze Vision Encoder</i>						
	✓	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 8. **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.

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 8

We observe that freezing the vision encoders does not affect the performance of image-level understanding tasks, but significantly reduces region-level or pixel-level task effectiveness (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 [10, 15], image-text contrastive learning [31, 45]), 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 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 how training configurations affect a foundation model’s performance in region-level and pixel-level tasks. Unfreezing the vision backbone significantly enhances its ability to learn from regions and pixels, aiding in various downstream tasks. Additionally, language pre-training weights aid tasks requiring text understanding but are less impactful for purely vision-based tasks.

C. 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 segmentation	Region-Text	Image, text, region	Region
Region to text	Region-Text	Image, text, region	Text
Text detection and recognition	Region-Text	Image, text	Text, region

Table 9. Supported Tasks and annotations used for *Florence-2* pretraining.

D. Supervised Data Collection for Generalist Model Fine-tuning

Task	Dataset
Caption	COCO [3]
Text Caption	TextCaps [35]
Paragraph caption	Stanford Paragraph Caption [14]
Detailed caption	Localized Narratives [30]
Detection	COCO [20], Object365* [34], Open Images* [16]
Phrase Grounding	Flickr30k, Object365* [34], Open Images* [16]
Referring expression	RefCOCO-mix (RefCOCO, RefCOCO+, RefCOCOg) [13, 26, 44]
Referring expression segmentation	RefCOCO-mix (RefCOCO, RefCOCO+, RefCOCOg) [13, 26, 44]
Region to category	COCO [20], Object365* [34], Open Images* [16]
Region to polygon	COCO [20] (after deduplicating RefCOCO-mix val)
VQA	VQAv2 [6], OKVQA [27], AOKVQA [33], TextVQA [36], ViZWiz VQA [7]
OCR	Subset from <i>FLD-5B</i> OCR (2 million samples)

Table 10. 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.

E. Model Configuration

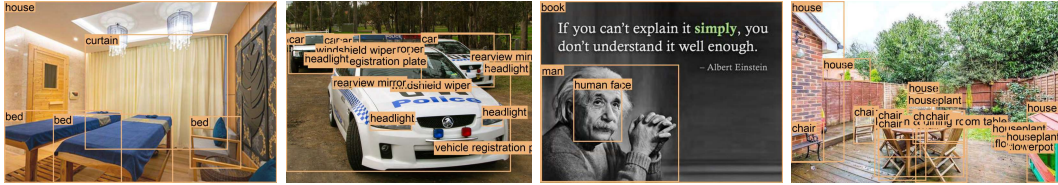
Model	Image Encoder (DaViT)				Encoder-Decoder (Transformer)			
	dimensions	blocks	heads/groups	#params	encoder layers	decoder layers	dimensions	#params
<i>Florence-2-B</i>	[128, 256, 512, 1024]	[1, 1, 9, 1]	[4, 8, 16, 32]	90M	6	6	768	140M
<i>Florence-2-L</i>	[256, 512, 1024, 2048]	[1, 1, 9, 1]	[8, 16,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

Table 11. Model configuration of different size.

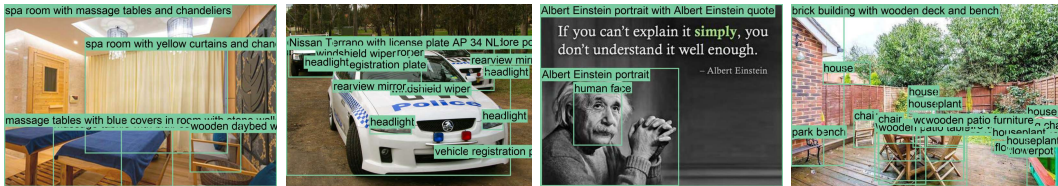
F. More Examples of Annotations in *FLD-5B*



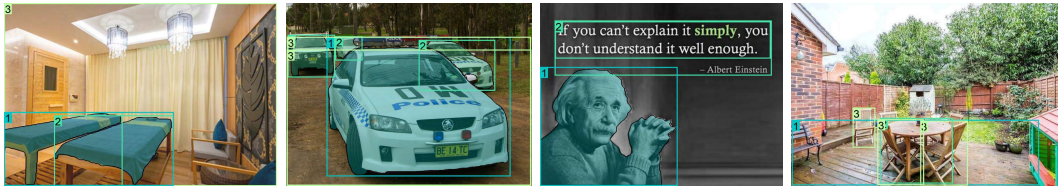
(a) Region only



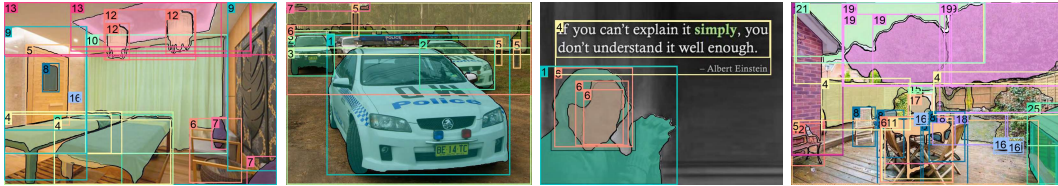
(b) Region w/ phrases



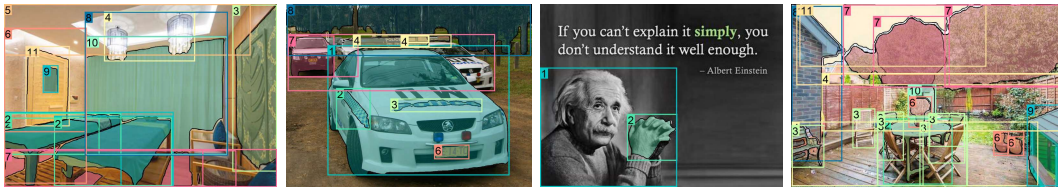
(c) Region w/ brief text



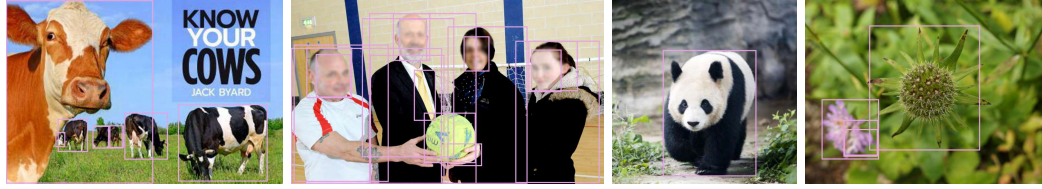
(d) Text-phrase-region w/ brief text



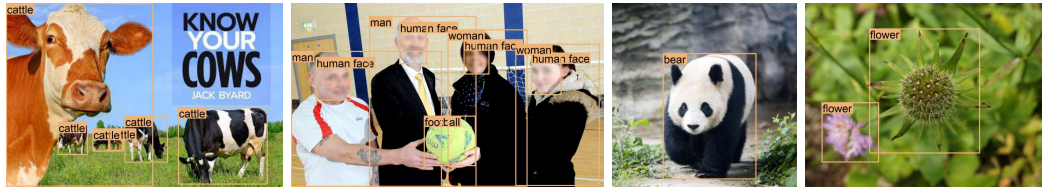
(e) Text-phrase-region w/ detailed text



(f) Text-phrase-region w/ more detailed text
Figure 4. Examples of annotations in *FLD-5B*.



(a) Region only



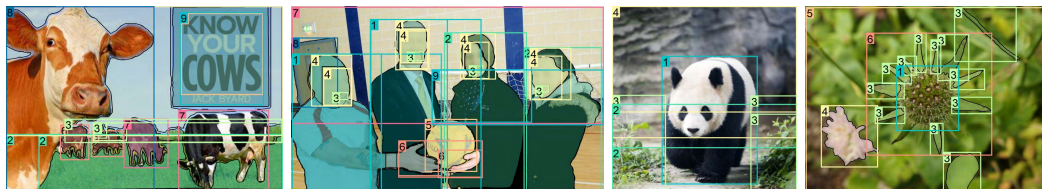
(b) Region w/ phrases



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

G. Qualitative Evaluation and Visualization Results

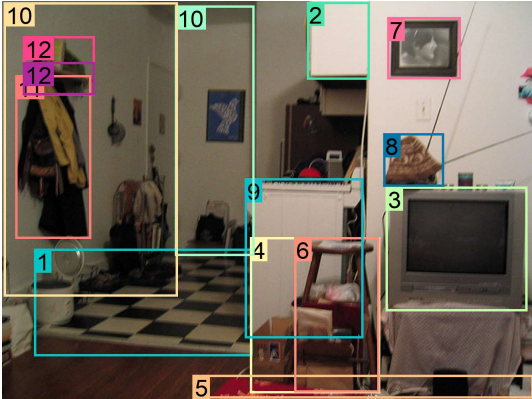
G.1. Visual Grounding

Visual Grounding

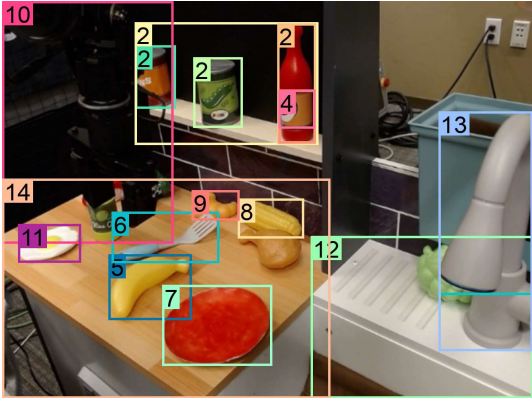
Prompt: Locate the phrases in the caption: {caption}



The image shows a group of five cartoon monsters. On the left side, there is a brown monster with horns and a big smile on its face. Next to it, there are two smaller monsters, one black and one green. The black monster has two large horns on its head and is standing in the center of the group. The green monster 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 monster 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 floor. On the right side of the image, there is a small white cabinet with a television on top of it. Next to the cabinet, there are several items scattered on the floor, including a red blanket, a wooden stool, and a pile of trash. On top of the cabinet is a picture frame and a hat. In the center of the room is a white refrigerator with a few items on top. The walls are painted white and there are a few clothes hanging on a rack on the left wall. The room appears to be in disarray, with some items strewn about and others scattered around.

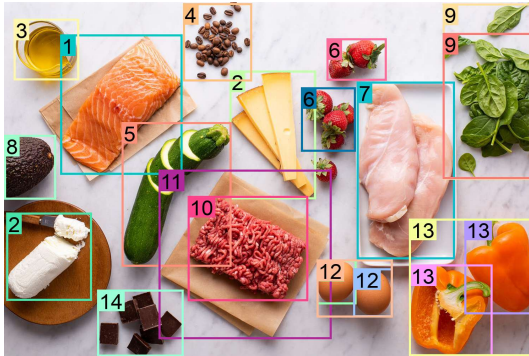


The image shows a kitchen countertop with various kitchen items on it. On the left side of the countertop, there is a microscope with a black body and a white lens. Next to the microscope, there are two bottles of condiments - one with a red label and the other with green. On top of the microscope is a yellow banana, a blue spatula, a red plate, and a yellow corn on the cob. In the center of the image, there appears to be a frying pan with a fried egg on it, and on the right side is a white sink with a white faucet. The countertop is made of wood and has a gray tile backsplash.

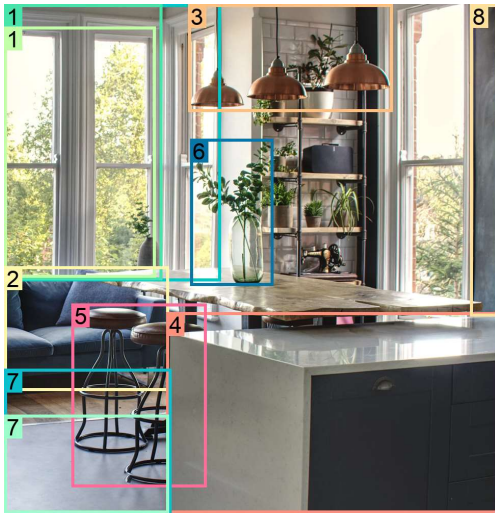
Figure 6. Visual grounding prediction results.

Visual Grounding

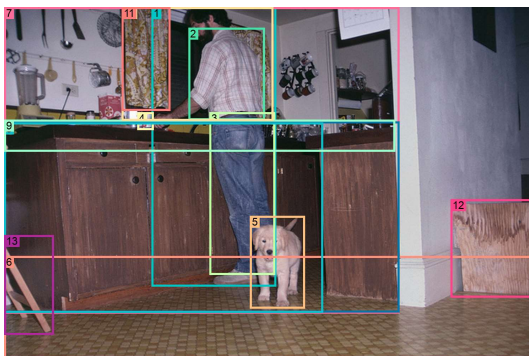
Prompt: Locate the phrases in the caption: {caption}



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, a 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 of 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 man² 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 7. Visual grounding prediction results. (continued)

G.2. Detailed Image Caption

Detailed Image Caption

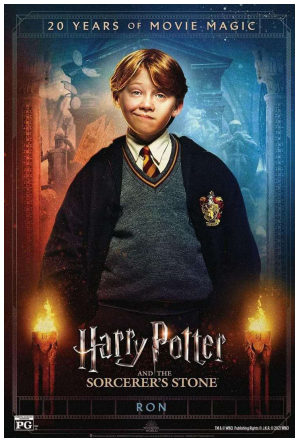
Prompt: Describe with a paragraph what is shown in the image.



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 8. Detailed captioning prediction results.

G.3. Dense Region Caption

Dense Region Caption

Prompt: Locate the objects in the image, with their descriptions.

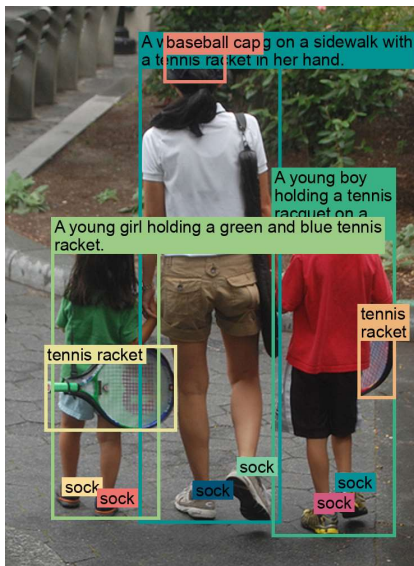
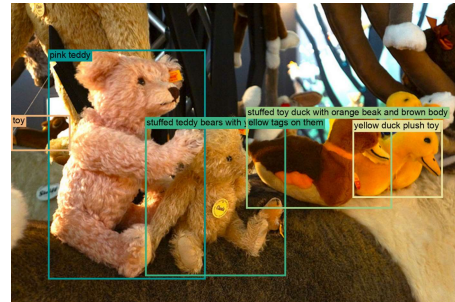
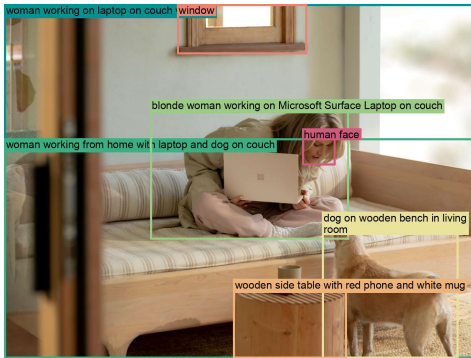
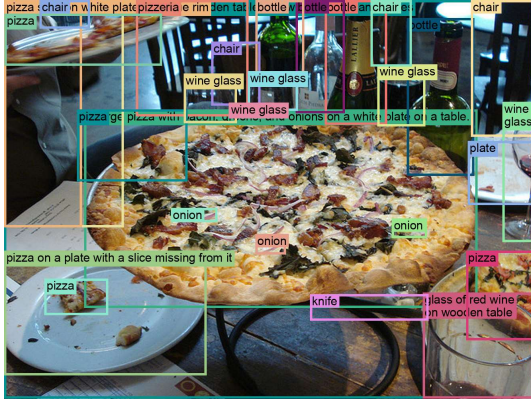


Figure 9. Dense region caption prediction results.

G.4. Open Vocabulary Detection

Open Vocabulary Object Detection

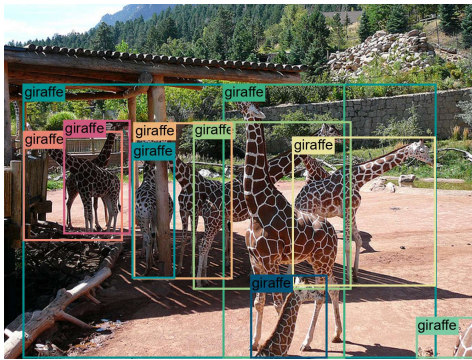
Prompt: Locate Five Alive juice box (and) Colgate toothpaste in the image.



Prompt: Locate Chewbacca in the image.



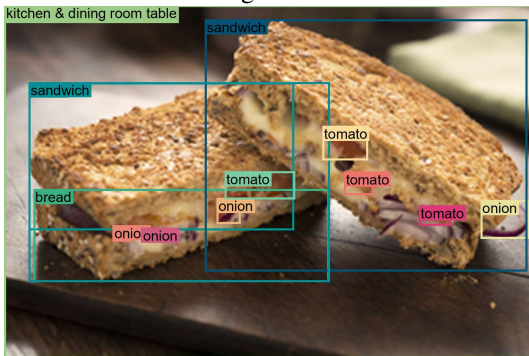
Prompt: Locate giraffe in the image.



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.

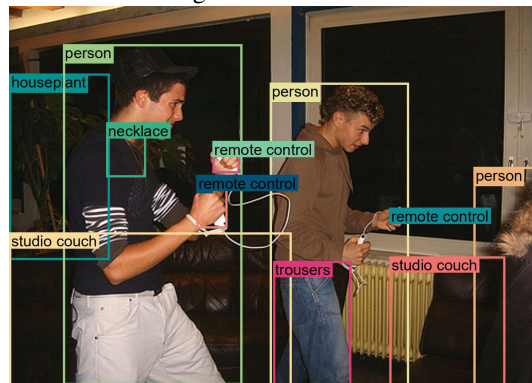
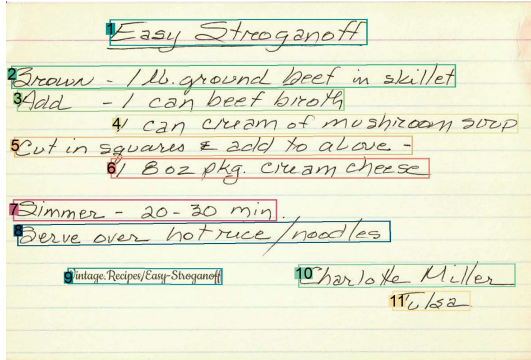


Figure 10. Open vocabulary object detection prediction results.

G.5. OCR

Ocr with region

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



Easy Stroganoff¹
 Brown 1 lb. ground beef in skillet²
 Add 1 can beef broth³
 1 can cream of mushroom soup⁴
 Cut in squares & add to above⁵
 1/ Boz pkg. cream cheese⁶
 Simmer 20-3 min.⁷
 Serve over hot rice / noodles.⁸
 Vintage. Recipes/Easy-Strogano⁹ Charlotte Miller¹⁰
 Tulsa¹¹



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 Mocha¹⁴
 Thai Iced Tea / Coffee¹³ Caramel¹⁵
 \$1.99/\$2.29/\$2.59¹⁶ SPECIALTY Brew !!¹⁷
 Jasmine Green Tea¹⁸



LEONARDO¹
 DICAPRIO²
 ROBERT³
 DE NIRO⁴
 LILY⁵
 GLADSTONE⁶
 A MARTIN SCORSESE PICTURE⁷
 KILLERS⁸
 OF⁹ FLOWER¹⁰
 MOON¹¹
 SCREENPLAY BY ERIC ROTH AND MARTIN SCORSESE DIRECTED BY MARTIN SCORSESE¹²
 ONLY IN THEATRES OCTOBER 20¹³

Figure 11. OCR with region prediction results.

G.6. Region to segmentation

Region to Segmentation

Prompt: What is the polygon mask of region $\langle \text{loc}_{.586} \rangle \langle \text{loc}_{.294} \rangle \langle \text{loc}_{.929} \rangle \langle \text{loc}_{.814} \rangle$



Prompt: What is the polygon mask of region $\langle \text{loc}_{.317} \rangle \langle \text{loc}_{.314} \rangle \langle \text{loc}_{.893} \rangle \langle \text{loc}_{.904} \rangle$



Prompt: What is the polygon mask of region $\langle \text{loc}_{.541} \rangle \langle \text{loc}_{.266} \rangle \langle \text{loc}_{.692} \rangle \langle \text{loc}_{.627} \rangle$



Prompt: What is the polygon mask of region $\langle \text{loc}_{.583} \rangle \langle \text{loc}_{.66} \rangle \langle \text{loc}_{.794} \rangle \langle \text{loc}_{.331} \rangle$



Prompt: What is the polygon mask of region $\langle \text{loc}_{.386} \rangle \langle \text{loc}_{.53} \rangle \langle \text{loc}_{.759} \rangle \langle \text{loc}_{.998} \rangle$



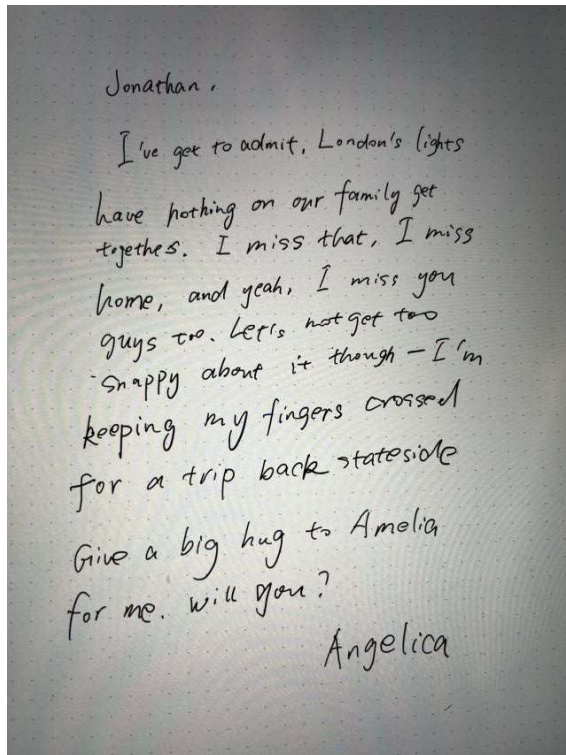
Prompt: What is the polygon mask of region $\langle \text{loc}_{.102} \rangle \langle \text{loc}_{.7} \rangle \langle \text{loc}_{.375} \rangle \langle \text{loc}_{.648} \rangle$



Figure 12. Region to segmentation prediction results.

H. Comparison 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 13. 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 14. 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 15. 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/minigt4>

Figure 16. Comparison with LMMs on detailed caption task (continued).

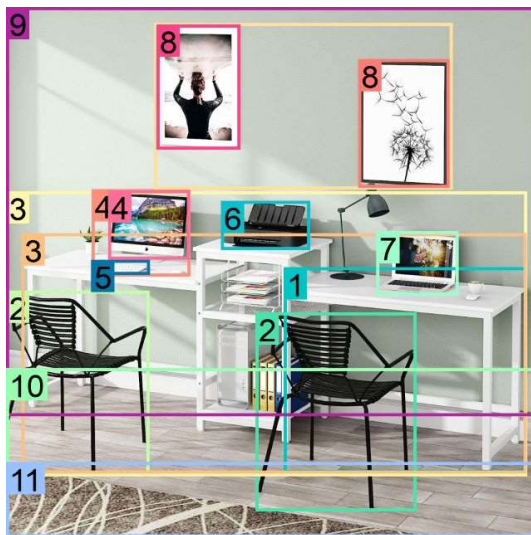
I. Comparison with Kosmos-2 on Detailed Caption and Grounding Tasks

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**³ 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*.

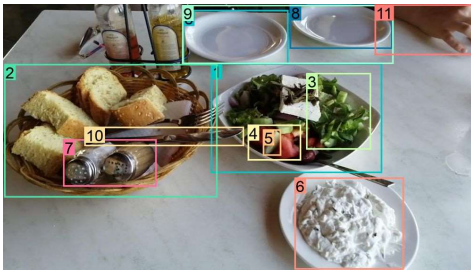
Figure 17. Systematic comparison with Kosmos-2 [29] on detailed caption and grounding tasks. The models generate both the detailed caption and grounding results. The results of Kosmos-2 are from <https://huggingface.co/spaces/ydshieh/Kosmos-2>.

Comparison on detailed caption and grounding tasks.



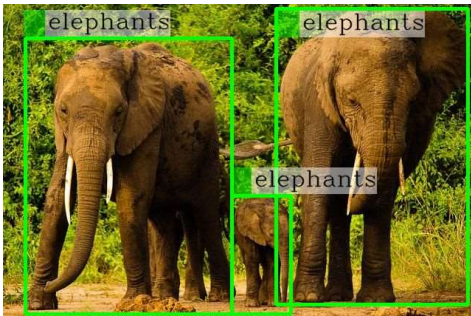
The image features a white dining table with a variety of food items on it. On the table, there is a bowl of bread, a bowl with a salad, and a plate with a side of yogurt. In addition to the food, there are two cups placed on the table. One cup is located near the left side of the table and the other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



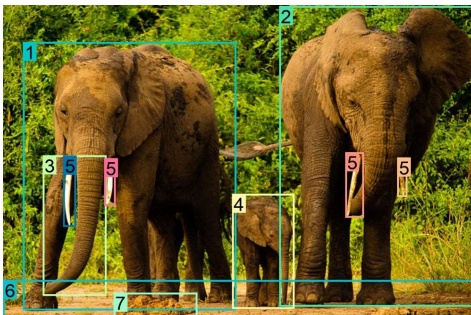
The image shows a table with a basket of bread and a plate of salad. The basket is made of woven straw and has several slices of bread in it. Next to the basket, there is a small bowl of salad with a variety of vegetables, including lettuce, tomatoes, cucumbers, and feta cheese. There are also two salt and pepper shakers on the table. On the right side of the table, there are two white plates with a dollop of white sauce on them. The table is covered with a white tablecloth and there are a few other dishes and utensils scattered around. A person's hand 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 are standing close together, with the largest elephant in the center and two smaller ones on either side. The largest elephant on the left is standing with its trunk extended, while the smaller one is standing next to it. All three elephants have tusks and appear to be in their natural habitat. The ground is covered in dirt and there is a small pile of dirt in front of them. The overall mood of the image is peaceful and serene.

(f) Result from Florence-2.

Figure 18. Systematic comparison with Kosmos-2 [29] 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)

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 3, 4
- [2] 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. 2
- [3] 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. 1, 7
- [4] 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
- [5] 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. 3, 4
- [6] 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. 7
- [7] 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. 7
- [8] 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. 4
- [9] 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] 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. 5
- [11] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020. 1
- [12] 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. 2
- [13] 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. 7
- [14] 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. 7
- [15] 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. 5
- [16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, mar 2020. 7
- [17] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, 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. 1
- [18] 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
- [19] 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. 2, 3
- [20] 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. 7
- [21] 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
- [22] 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. 2, 3, 4
- [23] 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. 3, 4
- [24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 1
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1, 3

- [26] 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. 7
- [27] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. 7
- [28] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. 2022. 4
- [29] 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. 21, 22
- [30] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 7
- [31] 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, 5
- [32] 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. 2
- [33] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. 7
- [34] 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. 1, 7
- [35] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. 7
- [36] 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. 7
- [37] 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. 2, 3, 4
- [38] 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. 3
- [39] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang 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. 3, 4
- [40] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 2, 3, 4
- [41] 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. 2, 3, 4
- [42] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space, 2022. 1, 3
- [43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 2
- [44] 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. 7
- [45] 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
- [46] 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
- [47] 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. 3