Aligning and Prompting Everything All at Once for Universal Visual Perception

Yunhang Shen[1], Chaoyou Fu[1], Peixian Chen[1], Mengdan Zhang[1]
Ke Li[1], Xing Sun[1], Yunsheng Wu[1], Shaohui Lin[2]*, Rongrong Ji[3]
[1]Tencent Youtu Lab [2]School of Computer Science and Technology, East China Normal University, Shanghai, China [3]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
{shenyunhang01, bradyfu24}@gmail.com, shlin@cs.ecnu.edu.cn, rrji@xmu.edu.cn
{peixianchen, davinazhang, tristanli, winfredsun, simonwu}@tencent.com

Abstract

*Vision foundation models have been explored recently to build general-purpose vision systems. However, predominant paradigms, driven by casting instance-level tasks as an object-word alignment, bring heavy cross-modality interaction, which is not effective in prompting object detection and visual grounding. Another line of work that focuses on pixel-level tasks often encounters a large annotation gap of things and stuff, and suffers from mutual interference between foreground-object and background-class segmentation. In stark contrast to the prevailing methods, we present APE, a universal visual perception model for **a**ligning and **p**rompting **e**verything all at once in an image to perform diverse tasks,* i.e.*, detection, segmentation, and grounding, as an instance-level sentence-object matching paradigm. Specifically, APE advances the convergence of detection and grounding by reformulating language-guided grounding as open-vocabulary detection, which efficiently scales up model prompting to thousands of category vocabularies and region descriptions while maintaining the effectiveness of cross-modality fusion. To bridge the granularity gap of different pixel-level tasks, APE equalizes semantic and panoptic segmentation to proxy instance learning by considering any isolated regions as individual instances. APE aligns vision and language representation on broad data with natural and challenging characteristics all at once without task-specific fine-tuning. The extensive experiments on over* 160 *datasets demonstrate that, with only one-suit of weights, APE outperforms (or is on par with) the state-of-the-art models, proving that an effective yet universal perception for anything aligning and prompting is indeed feasible. Codes and trained models are released at* https://github.com/shenyunhang/APE.

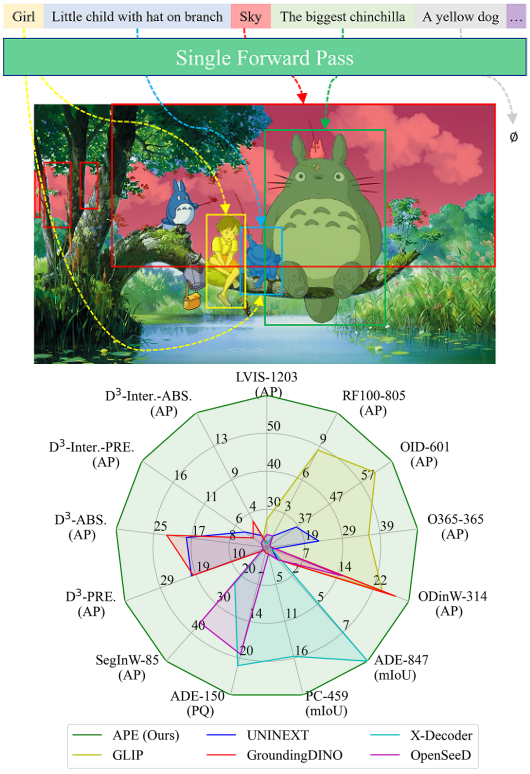

Figure 1. APE supports prompting thousands of things, stuff, and sentences in a single forward pass and performing various segmentation without granularity discrepancy.

1. Introduction

Developing vision systems that recognize and localize a wide range of basic concepts and can be transferable to novel concepts or domains, has emerged as an important research topic in the community. In light of the strong transferability demonstrated by LLMs, many researchers have attempted to build advanced vision foundation mod-

*Corresponding Author

els (VFMs) to serve general-purpose vision tasks.

Generally, the existent VFMs are roughly categorized into three groups. The first one is to learn all-purposed visual features via self-supervised learning, *e.g.*, DINO [2] and iBOT [53], and align text-image corpora with weakly-supervised learning, such as CLIP [35] and ALIGN [15]. Despite the promising feature transferability, the aforementioned methods often require individual adapters for downstream tasks. The second group aligns region and text representation for instance perception tasks, such as GLIP [21], UNINEXT [48], and GroundingDINO [27]. As they formulate this problem as a visual grounding problem with deep region-word fusion, it is incapable of operating on a large number of categories and grounding phrases at once [49]. The other group focuses on generic segmentation tasks, such as SAM [17], X-Decoder [56], OpenSeeD [52], and SEEM [57]. However, they usually suffer from the granularity discrepancy between foreground objects and background stuff, as foreground objects often perform object-level instance segmentation while background stuff corresponds to class-level semantic segmentation. Previous methods [20, 23, 52] decouple foreground and background learning with private queries and workflows, which involves manual prior knowledge to route each concept.

To develop VFMs that address the above problems, this work explores efficient promptable perception models and handles diverse semantic concepts for detection, foreground and background segmentation, and grounding. To address the heavy computational cost of vision-language fusion, we aggregate compact sentence representation with gated cross-modality interaction and efficiently adapt the concept of vocabularies and sentences into a common embedding space. To address the granularity discrepancy of foreground things and background stuff, we equalize their granularity by decomposing the category-level segmentation learning into the instance-level proxy objective, forming a single instance segmentation task. Then instance-level patterns are ready to project back to category-level segments during inference. By eliminating the discrepancy between foreground and background, it is granularity-friendly to learn from category-aware and category-agnostic segmentation data without distinguishing things and stuff manually.

To this end, we propose a APE, a universal visual perception model for **a**ligning and **p**rompting **e**verything all at once in an image, which performs foundational vision tasks with an instance-level region-sentence interaction and matching paradigm. We characterize several important capabilities that maximize APE's practicality in real-world scenarios from three perspectives: (1) Task generalization: APE is built based on DETR [1] framework to perform a wide array of semantic understanding tasks, which is capable of predicting labels, boxes, and masks for any object, region, and parts. Specifically, we unify object detection of common and long-tailed vocabularies, image segmentation for various granularity, and visual grounding, into an instance-level detection transformer framework. (2) Data diversity: APE is trained on broad data sources all at once, ranging from long-tailed categories, federated annotations, anything segmentation, and hybrid vocabulary- and sentence-described concepts. (3) Effective description prompting: It is feasible to query APE with thousands of text prompts for object vocabularies and sentence descriptions, which aggregates the word-level prompt embedding for effective gated cross-modality fusion and alignment.

Benchmarked on over 160 datasets, APE achieves the state-of-the-art (SotA) or competitive performance with one-suit of weights at various visual perception tasks, demonstrating the generalization and practicality of APE as VFMs. We hope to facilitate the community on wide real-life applications. Extensive ablation studies also verify the efficiency and effectiveness of each proposed component.

Conclusively, our contributions are the following:

- We present a APE, a universal visual perception model for **a**ligning and **p**rompting **e**verything all at once in an image, which is trained on broad data at scale and provides SotA performance without task-specific fine-tuning.
- We reformulate the visual grounding as open-vocabulary detection with region-sentence vision-language interaction and matching, which significantly improves the efficiency of model querying for large-scale text prompts.
- We bridge the granularity gap of various segmentation patterns by transforming learning into the object-level proxy objective, which models thing and stuff categories equally without category-specific design.

2. Related Work

Unified vision-language models have recently drawn a lot of attention because of their great flexibility in generalizing to various tasks. Mainstream approaches can be roughly categorized into two main types of goals, namely, ground anything and segment anything. The former unified instance-level tasks, such as object detection and visual grounding, as region-word grounding learning, while the latter focuses on promptable and interactive learning for pixel-level segmentation with dense outputs.

2.1. Unified Detection and Grounding

Pix2Seq [3] and SeqTR [55] design a pixel-to-sequence interface for various tasks, such as object detection, instance segmentation, and visual grounding. MDETR [16] links the output tokens of DETR to specific words with region-word alignment. GLIP [21] formulates detection as grounding and learns instance-level visual representation with language-aware deep fusion. UNINEXT [48] further supports prompts in both text and image for instance-level, *i.e.*, foreground objects, perception tasks. Ground-

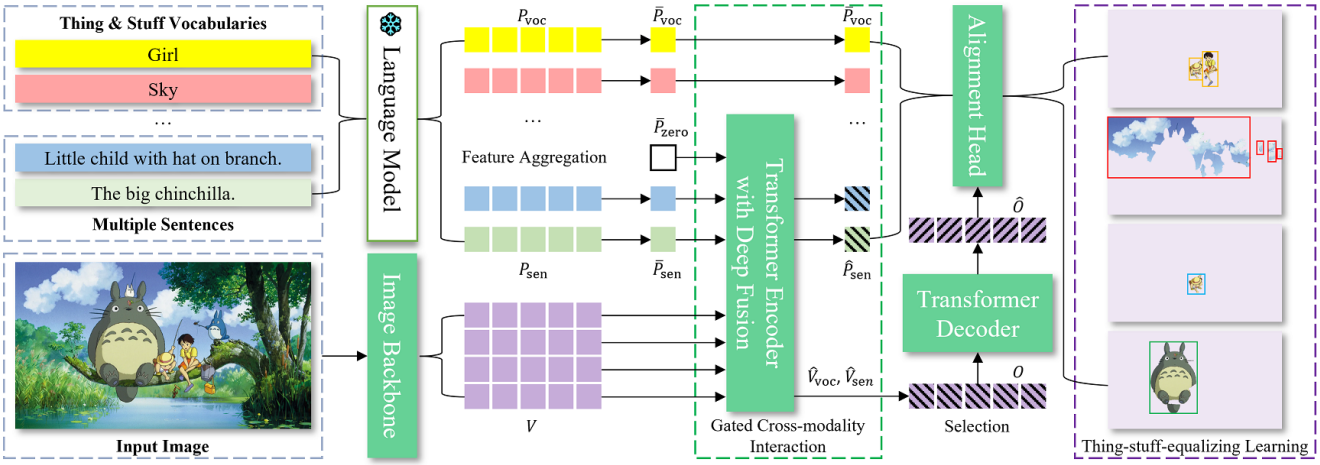

Figure 2. The overall framework of the proposed APE. First, the image backbone and language model extract discrete visual embeddings V and text embeddings P for a given image and corresponding text prompts, respectively. Second, the word-level text embedding P is further aggregated into sentence-level embeddings $\bar{P}$. Then, the cross-modality encoder fuses information from two modalities to condition the object queries O on text queries and update text embeddings $\hat{P}$. A transformer decoder generated final object embeddings $\hat{O}$ from object queries O. Finally, a visual-language alignment module to predict the correct pairings of regions and prompts.

ingDINO [27] introduces additional cross-modality to the encoder, query selection, and decoder of detection transformers.

While promising generalization performance is presented, we argue that casting detection as a grounding problem via object-word fusion and alignment leads to inefficient interaction between vision and language. Specifically, they can not prompt a large number of vocabularies or expressions in one forward due to the limit of GPU memory footprint and token length of text models. For example, LVIS [12] and D^3 [46] have 1203 vocabularies and 422 descriptions, on which the previous models [21, 27, 48] require about 30 and 422 forwards to infer on a single image. To address this drawback, we reformulate detection and grounding tasks as instance-level region-sentence matching, which is feasible to query models with thousands of concepts at scale. Meanwhile, APE also avoids the additional processing of extracting the root object in given sentences.

2.2. Unified Image Segmentation

Mask2Former [5] and MaskDINO [20] present a universal architecture capable of handling semantic, instance, and panoptic segmentation for close-set categories. ODISE [47] leverages the frozen internal representation of text-to-image diffusion models for open-vocabulary panoptic segmentation. X-Decoder [56] and SEEM [57] introduce a query-based segmentation architecture to support generic, referring, and interactive segmentation, and image-level vision-language understanding tasks. However, they suffer from the mutual interference between things and stuff within each query. To alleviate granularity discrepancy, OpenSeeD [52] and HIPIE [43] decouple foreground things and background stuff with separate decoders instead of one unified one.

However, decoupling learning [43, 52] requires manually defining categories into things and stuff for both training and inference, which does not apply to segmentation data without semantic labels, such as SA-1B [17]. In this paper, we formulate foreground and background equally with the proxy instance-level objective. And the instance-level outputs are ready to convert to segmentation predictions of different formats to satisfy the desired granularity of tasks, *i.e.*, semantic and panoptic segmentation.

3. Method

As shown in Fig. 2, APE consists of a vision backbone for image feature extraction, a language model for text feature extraction, a transformer encoder with cross-modality deep fusion [21], and a transformer decoder. APE is expected to output a set of scores, boxes, and masks for a given image and a set of prompts, which could contain a large number of (thing and stuff) vocabularies and sentences.

3.1. Description Prompting at Scale

Recently, many unified learning paradigms simultaneously train detection and grounding data, showing strong transferability to various object-level recognition tasks. In detail, the previous works, such as GLIP [21], GroundingDINO [27] and UNINEXT [48], reformulate object detection as phrase grounding by replacing region classifier with word-region alignment. However, such formulation is difficult to scale up for large-scale detection and grounding with thousand tokens in text prompts. The main reason is two-fold: First, the above methods require bidirectional language models, *i.e.*, the pre-trained BERT [7], to encode

text prompts, which can only encode sentences containing at most 512 tokens. Second, the above formulation heavily relies on vision-language fusion to perform cross-modality multi-head attention between words and regions at high dimension, *i.e.*, $2,048$, and brings heavy computational costs and GPU memory consumption. Although the length of text prompt is further limited to 256 in [21, 27, 48, 51], such fusion module also brings about $0.6 \sim 1.4\times$ additional memory footprint overall during inference in GLIP [21]. A practical solution can split the long text prompts into multiple prompts and query the model multiple times for both training and inference. However, such a workaround does not address the inherent problem.

To efficiently prompt a large number of vocabularies and sentences all at once, we reverse the widely-used paradigm that cast the classical object detection task into a grounding problem. Rather, we reformulate visual grounding as object detection to equally unify both localization tasks. Based on this reformulation, we re-design the text prompt, cross-modality fusion, and vision-language alignment strategy.

Independent Prompt. Given object classes, such as *Girl* and *Sky*, the previous methods [21, 27, 48, 51] concatenate all vocabularies into a single prompt: "Girl. Sky. . . . ", in which the corresponding concept embedding is modeled based on its relationship to the other words in the sentence. The overall prompt embedding $P_{\text{voc}} \in \Re^{1\times l\times d}$ has a sequence length of l and an embedding dimension of d. Similarly, the prompts for sentence descriptions are formed as: ["Little child with hat on branch", "The big chinchilla", . . .], thus obtaining embedding $P_{\text{sen}} \in \Re^{n\times l\times d}$, where n is the sentence number. We find that the correlation among vocabularies is not necessary, and modeling individual concepts alone is sufficient to identify different instances.

To this end, we blend the individual concepts of vocabularies or sentences as independent text prompts to compute their text embeddings. Thus, we construct a set of text prompts: ["Girl", "Sky", "Little child with hat on branch", "The big chinchilla", . . .], in which the number of concepts is not limited by the input sequence length of text models. Those diverse prompts are directly input into directional language models, such as CLIP [35] and Llama [41], getting the prompt embedding $\{P_{\text{voc}}, P_{\text{set}}\} \in \Re^{n\times l\times d}$ for n independent text prompts.

Sentence-level Embeddings. To reduce computational complexity and memory usage, we further compress word-level concept representation to sentence-level prompt embeddings. Specifically, we aggregate word-level embeddings $\{P_{\text{voc}}, P_{\text{sen}}\}$ to sentence-level ones $\{\bar{P}_{\text{voc}}, \bar{P}_{\text{sen}}\} \in \Re^{n\times d}$ via: $\bar{P}_{n,d} = \frac{1}{l}\sum_{j=0}^{l} P_{n,j,d}$, which performs average operator along the length axis. Albeit word-level prompt embeddings may have more fine-grained information, we find that sentence-level prompt embeddings provide comparable performance, as demonstrated in Sec. 4.2.

Gated Cross-modality Interaction. The original deep fusion [21] involves multi-head attention over a large number of input elements for open-vocabulary detection, which usually has thousands of vocabularies to learn. Thus, we further propose gated cross-modality interaction to restrict different types of prompts from vision-language fusion. Firstly, the interaction between image features and large-scale vocabularies is prohibitively expensive. Instead, an all-zero token $\bar{P}_{\text{zero}} \in \Re^{1\times 1\times d}$ serves as a special text embedding and inputs to the fusion module for all given vocabularies. In this situation, the fusion process is "static", as no language information is injected into vision features. The $\bar{P}_{\text{zero}}$ could provide explicit instructions to recognize primitive concepts and slightly tune vision feature $\hat{V}_{\text{voc}}$ and retain original language feature $\bar{P}_{\text{voc}}$. For sentence prompts, the corresponding sentence-level embedding $\bar{P}_{\text{set}}$ are injected into the vision feature V, which dynamically updates new vision feature $\hat{V}_{\text{set}}$ and language feature $\hat{P}_{\text{set}}$.

The proposed gated fusion enjoys two advantages: 1) It is feasible to model thousands of detection categories and fuse hundreds of grounding sentences with only a single forward during training and inference. 2) The previous work has shown that training detection data with a deep fusion module could hurt the zero-shot generalization to novel categories [21]. The gated interaction prevents such degeneration by explicitly prohibiting fusion for detection task.

Region-sentence Alignment. MDETR [16] first proposes to predict the span of tokens from a text prompt that refers to each matched object, which is so-called word-region alignment. From a practical perspective, it may not be necessary to detect each word in prompts. Instead, we predict objects corresponding to the whole prompt, which are a category or sentence. Concretely, we compute the alignment scores S between object embeddings $\hat{O}$ and prompt embeddings $\{\bar{P}_{\text{voc}}, \hat{P}_{\text{sen}}\}$ as: $S = \hat{O} \cdot (\bar{P}_{\text{voc}}, \hat{P}_{\text{set}})^{\top}$, where $S \in \Re^{n\times m}$. In such a way, detection categories are fixed anchors in the vision-language common embedding space, as prompt embeddings of vocabularies are not updated in the proposed gated cross-modality interaction.

To compensate for the loss of fine-grained information in sentence-level embedding, we adapt additional irrelevant prompts as negative queries, which imposes the model to have a close "look" at target prompts and reject the negative ones. In detail, we maintain a history embedding bank and select several embeddings as negative, which are concatenated with positive embeddings for fusion and alignment.

3.2. Thing-stuff-equalizing Alignment

Recently, MaskFormer [4] and Mask2Former [5] unify thing and stuff segmentation tasks by query-based Transformer architectures and perform mask classification. However, they are designed for segmentation tasks and usually have unsatisfactory detection performance. On the

other hand, MaskDINO [20] adopts the idea of mask classification from Mask2Former [5] to construct a segmentation branch on DINO [2]. However, MaskDINO [20] uses different strategies to learn thing and stuff categories, which requires manually identifying vision concepts into two groups ahead. Thus, it is difficult to incorporate generic semantic-aware segmentation data with large-scale class-agnostic data, such as SA-1B [17].

To this end, we design a straightforward albeit surprisingly effective solution, where the granularity of the background is equalized to the foreground one, *i.e.*, the model is not aware of the difference between things and stuff. As stuff categories may mislead the instance-level prediction, stuff regions are composited into multiple disconnected instances, which are treated as standalone samples and aligned with proxy object-level objectives. During training, we apply connected-component labeling on stuff mask annotations, where subsets of connected components are proxy-ground-truth instances. For inference, we join all predictions of the same stuff categories as the final results. Given predicted scores $S \in \Re^{q\times c}$ and masks $M \in \Re^{q\times h\times w}$, the final semantic masks $\hat{M} \in \Re^{c\times h\times w}$ are accumulated as: $\hat{M}_{c,h,w} = \sum_{i=1}^{q} S_{i,c} M_{i,h,w}$, where q and c are the number of queries and categories, respectively.

3.3. Single-Stage Training with Diversity Data

Training Objective. To build foundation models that solve a variety of tasks simultaneously, APE is trained in a single stage without task-specific fine-tuning. The training objective is a linear combination of classification loss, localization loss, and segmentation loss for the encoder and decoder, respectively:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{class}} + \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{giou}}}_{\text{encoder and decoder}} + \underbrace{\mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{dice}}}_{\text{last layer of decoder}},$$

where $\mathcal{L}_{\text{class}}$ is Focal loss [26] to classify foreground and background regions for the encoder and align language and vision embeddings for the decoder. $\mathcal{L}_{\text{bbox}}$ and $\mathcal{L}_{\text{giou}}$ are L1 loss [37] and GIoU loss [38] for box regression, which is applied to both encoder and decoder. $\mathcal{L}_{\text{mask}}$ and $\mathcal{L}_{\text{dice}}$ are cross-entropy loss and dice loss [30] for mask segmentation, which supervises the last output of decoder only.

Training Data. With the above loss function, 10 datasets with different annotation types are employed to train APE. For object detection, APE simultaneously learns common vocabularies from MS COCO [25], Objects365 [39] and OpenImages [19], and long-tailed LVIS [12]. OpenImages and LVIS are also federated datasets with sparse annotations. For image segmentation, apart from mask annotations in MS COCO and LVIS, APE also learns class-agnostic segmentation data from SA-1B [17], which contains both things and stuff without semantic labels. For visual grounding, we joint Visual Genome [18], RefCOCO/+/g [29, 50], GQA [14], Flickr30K [34], and PhraseCut [44].

To handle the diverse data and meet the requirement of single-stage training, we propose three principles for multi-dataset and multi-task learning: First, the well-annotated detection and segmentation data supervise all classification losses, localization losses, and even segmentation losses when there exist pixel-level annotations. For federated datasets, such as LVIS and OpenImages, a federated loss is integrated into classification losses $\mathcal{L}_{\text{class}}$ in the decoder. For class-agnostic data from SA-1B, the classification losses $\mathcal{L}_{\text{class}}$ in the decoder is not trained. Second, grounding data is only used to learn classification losses $\mathcal{L}_{\text{class}}$ in the decoder, as most grounding data does not exhaustively annotate all images with all objects and the bounding boxes are often not as accurate as detection data. For grounding data with segmentation annotations, such as RefCOCO/+/g, all loss functions in the decoder are trained. Third, we set the dataset sampling ratio to 1.0 if the dataset has more than 100K images and 0.1 otherwise. We list the configures of sampling ratios and loss weights for all datasets in Tab. 11 of the supplement material.

Image-centri Grounding Samples. The previous methods construct grounding samples of the form $\{I, T, B\}$, where I is an image, T is a phrase that describes an instance in I, and B is the corresponding bounding box. However, compared to detection training where all annotations in an image are trained all at once, the above region-centri format makes grounding training inefficient, as the model is supervised by a single instance for each sample. APE is feasible to handle multiple phrase prompts in a single forward pass during the training and inference. Thus, we gather the grounding samples in the image-centri form of $\{I, (T_i, B_i), \ldots, (T_n, B_n)\}$, which groups the grounding annotations. The new image-centri format significantly reduces the number of training iterations while the model still receives the same amount of supervision. For example, there are on average 92 box-level annotations (region descriptions and object instances) per image in Visual Genome. The proposed image-centri format leads to 92× speedup over the traditional region-centri format. During training, to prevent multiple phrases referring to the same object, we apply NMS to all boxes with random scores.

4. Experiments

We show that APE serves as a strong general-purpose vision system after training, *i.e.*, one model weight for all. Specifically, APE is directly evaluated on over 160 datasets to detect, segment, and ground without fine-tuning. Implementation details are in Sec. 6.3 of the supplement material.

APE-L (A) is built on DETA [32] with our designs replacing the corresponding modules. It is based on the ViT-L [8] and only trained on detection and segmentation data, including COCO, LVIS, Objects365, OpenImages, and Vi-

Table 1. One suit of weights for open-vocabulary detection on multiple datasets.

Method	Backbone	LVIS-1203				RF100-805	OID-601	Objects365-365		ODinW-314				COCO-80	
		val		minval		100 val	val	val	minival	35 val		13 val		val	
		AP^b	AP^m	AP^b	AP^m	AP^b_{avg}	AP^b	AP^b	AP^b	AP^b_{avg}	AP^b_{med}	AP^b_{avg}	AP^b_{med}	AP^b	AP^m
MDETR [16]	ENB5	–	∅	–	∅	–	–	–	–	10.7	3.0	–	–	–	∅
OWL [31]	ViT-L	34.6	∅	–	∅	–	–	–	–	18.8	9.8	–	–	43.5	∅
GLIP [21]	Swin-L	26.9	∅	37.3	∅	8.6	61.4	36.2	39.0	23.4	11.0	52.1	57.6	49.8	∅
GLIPv2 [51]	Swin-H	–	–	50.1	–	–	–	–	–	–	–	55.5	–	**64.1**	47.4
UNINEXT [48]	ViT-H	14.0	12.2	18.3	16.0	–	36.1	23.0	25.5	–	–	–	–	60.6	**51.8**
G-DINO [27]	Swin-L	–	∅	33.9	∅	–	–	–	–	26.1	18.4	–	–	60.7	∅
OpenSeeD [52]	Swin-L	23.0	21.0	–	–	–	–	–	–	15.2	5.0	–	–	–	–
APE-Ti	ViT-Ti	36.6	33.0	42.1	38.1	8.2	54.0	24.5	26.4	20.0	10.4	42.5	47.4	44.5	37.8
APE-L (A)	ViT-L	55.1	48.7	60.1	53.0	8.0	66.5	46.0	47.5	25.6	10.0	55.2	64.2	56.1	47.0
APE-L (B)	ViT-L	57.0	50.5	62.5	55.4	9.6	**68.2**	47.2	48.9	**29.4**	**16.7**	**59.8**	**66.9**	57.7	48.6
APE-L (C)	ViT-L	56.7	50.7	62.5	55.6	10.4	66.6	46.4	47.9	29.3	15.4	59.7	66.7	57.4	48.6
APE-L (D)	ViT-L	**59.6**	**53.0**	**64.7**	**57.5**	**11.9**	66.7	**49.2**	**51.1**	28.8	19.9	57.9	64.9	58.3	49.3

Table 2. One suit of weights for open-vocabulary segmentation on multiple datasets.

Method	Backbone	ADE-847	PC-459	ADE-150				SegInW-85		PC-59	BDD-40		VOC-20	Cityscapes-19		
		val	val	val				25 val		val	val		val	val		
		mIoU	mIoU	PQ	AP^m	AP^b	mIoU	AP^m_{avg}	AP^m_{med}	mIoU	PQ	mIoU	mIoU	PQ	AP^m	mIoU
X-Decoder [56]	DaViT-L	9.2	16.1	21.8	13.1	–	**29.6**	22.3	32.3	**64.0**	17.8	47.2	**97.7**	38.1	24.9	**52.0**
OpenSeeD [52]	Swin-L	–	–	19.7	15.0	17.7	23.4	36.1	38.7	–	**19.4**	**47.4**	–	**41.4**	**33.2**	47.8
OpenSeg [11]	ENB7	8.8	12.2	–	–	–	28.6	–	–	48.2	–	–	72.2	–	–	–
OVSeg [24]	Swin-B	9.0	12.4	-	–	–	29.6	–	–	55.7	–	–	94.5	–	–	–
APE-Ti	ViT-Ti	6.4	14.2	20.9	19.4	24.3	23.7	43.1	37.8	46.8	14.7	41.1	85.9	29.5	22.3	40.9
APE-L (A)	ViT-L	4.1	15.9	26.6	23.8	28.7	28.9	47.4	48.0	51.2	15.2	41.8	90.4	28.6	27.4	37.9
APE-L (B)	ViT-L	9.2	21.0	26.4	23.5	28.5	29.0	46.4	53.7	58.3	13.4	35.3	95.8	26.9	26.6	37.2
APE-L (C)	ViT-L	**9.4**	20.1	26.1	23.8	28.8	28.5	47.8	49.9	58.6	16.2	43.9	95.5	32.8	30.7	42.6
APE-L (D)	ViT-L	9.2	**21.8**	**27.2**	**24.4**	**29.6**	**30.0**	**49.6**	**52.2**	58.5	17.4	45.7	96.5	33.3	30.3	44.2

sual Genome. **APE-L (B)** is enhanced with Visual Genome region descriptions and RefCOCO/+/g. It is designed to verify the effectiveness of grounding data. **APE-L (C)** adds class-agnostic data SA-1B for training. **APE-L (D)** further include incorporates GQA, PhraseCut, and Flickr30k. **APE-Ti** reduces the size of image backbone to ViT-Ti [8]. Details on datasets are in Sec. 6.2 of the supplement material. All APE models are jointly trained on the corresponding datasets in a single stage without any fine-tuning. We list data usages in Tab. 10 of supplementary. Compared to other methods, our framework feeds the least number of images to models during training.

4.1. One Model Weight for All

To investigate the generalization ability of APE, we evaluate our models on various domain- and task-specific datasets.

Object Detection. In Tab. 1, we evaluate on the well-established benchmarks, including large-vocabulary and long-tailed dataset, *e.g.*, LVIS [12], and common object detection datasets, *e.g.*, Objects365 [39], OpenImages [19] and MSCOCO [25]. "∅" indicates that the task is beyond the model capability. "–" indicates that the work does not have a reported number. APE achieves the state-of-the-art or competitive performance across all benchmarks simultaneously. We find that the existing methods, such as GLIP, OWL, and UNINEXT, while including Objects365 during training, fall short in delivering strong performance on Objects365. The proposed APE is notably superior to them, proving that APE remembers all seen concepts well.

In addition, we further introduce Roboflow [6] and ODinW [21], which consist of 100 and 35 datasets, respectively, with different imagery domains, to evaluate generalizability under real-world scenarios. APE achieves a new SotA on both Roboflow and ODinW, validating APE can handle a large-scale of diverse concepts in the wild.

Image Segmentation. We then compare APE with the previous works on various segmentation tasks. We report PQ, AP^m, and mIoU for panoptic, instance, and semantic segmentation, respectively. Overall, APE achieves significantly better performance on PC-459, ADE20K, and SegInW with 459, 150, and 85 categories, respectively, and comparable performance for BDD, VOC, and Cityscapes with only 40, 20, and 19 categories, respectively. SegInW consists of 25 diverse segmentation datasets. The results demonstrate that APE has superior generalization ability to detect and segment a wide range of object categories in real-world scenarios. Note that we only use instance-level annotations for training, which puts APE at a disadvantage in the evaluation of panoptic-level results in terms of PQ. This is because the panoptic task requires non-overlap instance predictions, while APE produces overlapping segments.

Visual Grounding. We evaluate the model's ability to ground objects in natural language on description detection dataset (D^3) [46]. Following work in [46], we evaluate each image with only the descriptions that existed in the image as intra-scenario. For inter-scenario, all references in the dataset are used to query the models. It is noted that, for both intra-scenario and inter-scenario settings, other methods only process a single description for one forward, while APE only needs a single forward to query all references.

As demonstrated in Tab. 3, APE outperforms all existing methods for all metrics in D^3. Specifically, APE achieves significant improvement in the inter-scenario evaluation. APE is naturally capable of rejecting irrelevant

Table 3. One suit of weights for visual grounding on D^3.

Method		Backbone	Intra-scenario						Inter-scenario											
			Full		*Presence*		*Absence*		*Full*		*Presence*		*Absence*		*Full*		*Presence*		*Absence*	
			AP^b	AP^m	AP^b	AP^m	AP^b	AP^m	AP^b	AP^m	AP^b	AP^m	AP^b	AP^m	AR^b	AR^m	AR^b	AR^m	AR^b	AR^m
OFA	[42]	R152	4.2	∅	4.1	∅	4.6	∅	0.1	∅	0.1	∅	0.1	∅	17.1	∅	16.7	∅	18.4	∅
CORA	[45]	R50	6.2	∅	6.7	∅	5.0	∅	2.0	∅	2.2	∅	1.3	∅	10.0	∅	10.5	∅	8.7	∅
OWL-ViT	[31]	ViT-L	9.6	∅	10.7	∅	6.4	∅	2.5	∅	2.9	∅	2.1	∅	17.5	∅	19.4	∅	11.8	∅
UNINEXT	[48]	ViT-H	20.0	–	20.6	–	18.1	–	3.3	–	3.9	–	1.6	–	45.3	–	46.7	–	41.4	–
G-DINO	[27]	Swin-B	20.7	∅	20.1	∅	22.5	∅	2.7	∅	2.4	∅	3.5	∅	51.1	∅	51.8	∅	48.9	∅
OFA-DOD	[46]	R101	21.6	∅	23.7	∅	15.4	∅	5.7	∅	6.9	∅	2.3	∅	47.4	∅	49.5	∅	41.2	∅
APE-Ti		ViT-Ti	29.1	26.7	29.8	27.3	26.8	24.8	11.8	10.9	12.8	11.7	8.9	8.5	69.8	64.5	70.2	64.5	68.4	64.3
APE-L (A)		ViT-L	25.1	22.6	24.5	22.0	26.9	24.4	16.4	14.8	15.9	14.3	17.9	16.3	63.1	57.4	63.5	57.5	62.1	57.1
APE-L (B)		ViT-L	30.0	26.8	29.9	26.6	30.3	27.3	20.0	18.0	20.5	18.4	**18.6**	**16.8**	79.3	70.8	79.0	70.2	79.9	72.4
APE-L (C)		ViT-L	27.8	25.6	27.9	25.5	27.3	25.8	20.4	**18.7**	21.2	**19.3**	18.1	16.9	79.9	72.8	80.1	72.6	79.0	73.3
APE-L (D)		ViT-L	**37.5**	**34.4**	**38.8**	**35.4**	**33.9**	**31.7**	**21.0**	18.3	**22.0**	**19.3**	17.9	15.4	**82.7**	**73.7**	**82.8**	**73.5**	**82.4**	**74.3**

Table 4. Ablation study of unified detection and grounding. O-W: object-word fusion. R-S: region-sentence fusion.

Training	Fusion	Bank	Text	COCO		LVIS		RefCOCO						RefCOCO+						RefCOCOg					
				val		val		val		testA		testB		val		testA		testB		umd-val		umd-test		google-val	
				AP^b	AP^m	AP^b	AP^m	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU	P@1	oIoU
RefC	×	×	CLIP	–	–	–	–	80.5	65.1	84.6	69.0	72.5	58.5	70.4	53.7	77.6	59.4	59.1	44.2	79.0	62.6	72.8	55.2	72.2	55.7
	×	×	T5	–	–	–	–	74.2	57.5	78.5	61.7	69.3	53.2	57.7	41.6	64.0	46.4	49.8	35.1	67.3	49.6	61.0	42.6	61.8	44.2
	O-W	×	CLIP	–	–	–	–	85.5	71.7	89.1	74.1	81.3	67.1	73.4	57.7	80.7	63.4	64.4	48.9	83.0	67.9	78.0	61.7	78.0	61.9
	O-W	×	BERT	–	–	–	–	84.6	70.7	88.6	73.6	79.6	66.4	72.4	57.1	79.0	61.7	62.2	47.2	82.9	67.4	77.7	60.9	78.0	62.1
RefC	R-S	×	CLIP	–	–	–	–	80.7	65.8	85.0	69.7	74.8	60.3	68.9	52.3	76.2	58.7	59.2	43.5	74.1	56.7	72.5	55.2	70.7	53.8
	R-S	√	CLIP	–	–	–	–	82.7	68.2	87.1	72.2	77.4	62.9	72.3	56.7	79.7	62.6	61.3	46.1	79.5	63.5	74.8	57.6	73.8	57.5
	R-S	√	Llama2	–	–	–	–	85.3	71.5	89.1	75.3	80.0	66.4	73.9	58.9	81.8	66.4	64.0	49.0	85.1	68.8	74.5	57.5	75.2	59.9
+COCO	×	×	CLIP	50.9	43.7	–	–	81.6	67.0	85.5	71.2	77.9	63.4	69.8	53.7	76.3	58.8	61.4	45.3	78.9	62.0	74.9	57.1	74.7	57.1
	×	×	T5	50.9	43.7	–	–	78.3	62.6	81.7	65.3	75.8	60.2	60.6	45.0	66.3	49.6	53.6	38.6	71.9	54.5	67.7	48.6	68.9	50.2
	O-W	×	CLIP	49.2	42.2	–	–	85.1	71.5	87.9	73.4	83.4	69.8	71.8	56.0	77.5	60.5	64.5	49.2	80.3	64.7	76.8	60.3	77.5	61.1
	O-W	×	BERT	50.2	43.2	–	–	85.6	71.5	87.8	73.3	83.2	68.8	71.2	55.4	77.0	60.2	63.8	48.8	81.1	66.2	77.7	60.9	79.1	62.9
	R-S	×	CLIP	50.9	43.7	–	–	84.0	70.7	88.0	74.4	80.7	67.7	73.2	57.8	79.6	63.5	64.5	49.5	80.7	64.7	76.2	59.0	77.0	60.3
+LVIS	×	×	CLIP	–	–	36.8	32.9	78.6	63.7	83.2	68.8	73.4	59.0	69.5	53.2	75.3	57.7	61.0	45.2	77.6	61.1	73.6	55.1	73.9	56.7
	O-W	×	CLIP	–	–	33.0	29.6	86.6	73.5	88.6	74.9	83.5	70.9	74.7	59.5	79.8	63.3	66.3	50.9	82.9	68.2	79.4	62.9	79.9	63.7
	O-W	×	BERT	–	–	25.6	23.0	84.8	70.9	87.0	72.9	82.3	68.2	71.7	56.1	76.6	59.9	64.7	49.5	79.9	64.3	77.6	60.8	77.2	61.4
	R-S	×	CLIP	–	–	37.8	34.1	83.3	69.7	87.2	73.8	78.2	64.8	73.3	57.9	80.2	63.7	64.5	48.7	79.4	63.6	75.6	58.0	75.6	58.9
	R-S	√	CLIP	–	–	37.8	33.9	85.2	71.8	88.5	74.8	81.6	68.2	75.0	59.9	81.1	64.4	66.1	50.3	81.2	66.0	77.6	59.7	78.0	62.1

prompts, while the previous methods almost completely fail in this setting. This demonstrated that formulating visual grounding as open-vocabulary detection with sentence-object vision-language fusion and alignment not only significantly improves the efficiency of prompting a lot of text queries, but also obtains a strong ability to reject negative references. And *Full*, *Presence*, and *Absence* denote evaluation on all descriptions, presence descriptions only, and absence descriptions only. We find that APE is less biased toward the presence descriptions and well handles the absence descriptions. We further conduct experiments on RefCOCO/+/g in Tab. 12 of supplement material, APE surpasses all other methods without further fine-tuning.

4.2. Ablations on Unified Detection and Grounding

Different cross-modality interactions could have a large impact on description references, and vocabulary concepts usually robust to the fusion and alignment strategies. Thus, we first perform an in-depth study of various module combinations on both object detection and visual grounding. Unless otherwise specified, we conduct ablation experiments with R-50 [13] with an input size of $800 \times 1,333$.

Region-sentence formulation vs. *Object-word formulation.* In the first part of Tab. 4, we find that the fusion module is helpful for visual grounding task and boosts the performance for all metrics. While sentence-level text embedding may lose fine-grained information, region-sentence fusion still achieves comparable performance with object-word fusion, as shown in the second part of Tab. 4.

Effectiveness of History Embedding Bank. To compensate for the loss of fine-grained information in sentence-level text embedding, we add the proposed history embedding bank to region-sentence fusion, and the performance of the visual grounding task is significantly improved. Because the text embedding bank introduces negative descriptions that do not exist in images and imposes the model to learn relevant information from language guidance.

Effectiveness of Joint Detection and Grounding Training. We further combine RefCoco/+/g RefC with MSCOCO or LVIS with a dataset ratio of $1:1$. In the third and fourth parts of Tab. 4, the proposed region-sentence fused models with text embedding bank catch up with, and even surpass, the counterparts of object-word fusion. We also find that, for large-vocabulary LVIS, object-word fusion could deteriorate the performance of object detection, while the proposed formulation also improves the detection results, demonstrating the effectiveness of APE.

Effectiveness of Text Models. We also ablate the influence of text models in Tab. 4. We find that contrastive pre-trained models, *i.e.*, CLIP [35], have better performance than BERT [7] and T5 [36] with masked image modeling for both detection and grounding tasks. Meanwhile, large language models [41] also improve grounding results.

Computational Cost. We test the additional computational cost of the cross-modality fusion by measuring decreased speed and increased memory in percentage. Following GLIP [21], for training, we use 2 and 1 images per batch for APE-R50 and APE-ViT-L, respectively. For inference, we use batch size 1 for all models. For a fair comparison, we disable the segmentation part and use 5 feature maps with strides ranging from 8 to 128. Tab. 5 shows that APE brings less proportion of additional com-

Table 5. Comparison of computational cost in terms of the proportions of decreased speed (FPS) and increased memory (GB) for cross-modality interaction. "∅" indicates that the task is beyond the model capability.

Model	Inference										Train	
	Detection				Grounding							
	COCO 80 classes		LVIS 1203 classes		1 sentences		128 sentences		1280 sentences			
	FPS	GB	FPS	GB	FPS	GB	FPS	GB	FPS	GB	FPS	GB
GLIP-T	↓ 47% 4.8 → 2.5	↑ 140% 1.0 → 2.4	↓ 98% 4.4 → 0.08	↑ 140% 1.0 → 2.4	↓ 47% 4.8 → 2.5	↑ 140% 1.0 → 2.4	∅	∅	∅	∅	↓ 41% 2.7 → 1.6	↑ 39% 11.5 → 16.0
APE-R50	↓ 32% 10.3 → 6.9	↑ 61% 1.3 → 2.2	↓ 34% 10.1 → 6.6	↑ 48% 1.7 → 2.5	↓ 29% 9.3 → 6.5	↑ 61% 1.3 → 2.1	↓ 39% 9.2 → 5.5	↑ 80% 1.3 → 2.4	↓ 76% 9.1 → 2.1	↑ 270% 1.3 → 5.1	↓ 35% 1.1 → 0.7	↑ 25% 7.7 → 9.7
GLIP-L	↓ 40% 0.5 → 0.3	↑ 60% 4.8 → 7.7	↓ 96% 0.5 → 0.017	↑ 60% 4.8 → 7.7	↓ 40% 0.5 → 0.3	↑ 60% 4.8 → 7.7	∅	∅	∅	∅	↓ 30% 1.1 → 0.8	↑ 18% 19.7 → 23.4
APE-L	↓ 11% 2.2 → 2.0	↑ 19% 4.2 → 5.0	↓ 8% 2.1 → 1.9	↑ 19% 4.4 → 5.2	↓ 12% 2.1 → 1.8	↑ 19% 4.2 → 5.0	↓ 14% 2.1 → 1.8	↑ 24% 4.2 → 5.3	↓ 46% 2.0 → 1.1	↑ 89% 4.3 → 8.1	↓ 15% 0.6 → 0.5	↑ 12% 10.9 → 12.3

Table 6. Ablations study of thing-stuff-equalizing learning with various training data.

Training Data	Equalize Thing&Stuff	Step	COCO-Stuff	COCO								
			mIoU	mIoU	PQ	SQ	RQ	PQ^{th}	SQ^{th}	RQ^{th}	PQ^{st}	SQ^{st}	RQ^{st}
COCO-Panoptic	×	90k	33.9	55.7	47.4	81.0	57.2	53.8	82.9	64.3	37.8	78.1	46.5
	√	90k	37.6	58.0	48.2	81.3	57.9	54.4	83.5	64.6	38.8	78.1	47.7
COCO-Instance	×	90k	43.0	52.2	44.8	80.9	54.1	52.3	83.0	62.5	33.5	77.6	41.4
	√	90k	45.5	54.5	45.5	81.3	54.7	52.3	83.4	62.2	35.3	78.0	43.3
LVIS, COCO-Stuff	×	375k	42.0	52.8	44.8	80.9	54.1	52.3	83.0	62.5	33.5	77.6	41.4
	√	375k	46.2	55.4	45.5	81.2	54.6	52.3	83.4	62.1	35.2	78.0	43.3
LVIS, COCO-Stuff, RefC	×	375k	42.8	53.8	44.8	81.7	53.9	49.3	82.9	58.8	38.0	80.0	46.5
	√	375k	46.1	55.2	45.3	81.9	54.4	49.2	83.0	58.5	39.3	80.3	48.1

Table 7. Ablation study of using SA-1B [17].

Training Data	Step	COCO val		LVIS val	
		AP^b	AP^m	AP^b	AP^m
COCO	90k	50.0	42.6	–	–
COCO, SA-1B	180k	50.1	43.4	10.3	9.2
LVIS	180k	–	–	37.3	33.4
LVIS, SA-1B	375k	–	–	39.1	35.3

Table 8. System comparisons of instance segmentation.

Method	Backbone	Size	COCO val		LVIS val	
			AP^b	AP^m	AP^b	AP^m
MaskDINO [20]	SwinL	1024	59.0	52.3	–	–
ViTDet [22]	ViT-H	1024	60.4	52.0	53.4	48.1
EVA-02 [9]	ViT-L	1536	62.3	53.8	60.1	53.5
APE	ViT-L	1536	**62.7**	**54.1**	**60.9**	**55.3**

putational costs than GLIP. It is feasible to query APE with a large number of prompts all at once, which validates the efficiency of the proposed gated cross-modality interaction.

4.3. Ablations on Thing-stuff-equalizing Alignment

We further conduct ablations on the proposed thing-stuff-equalizing alignment, which formulates both thing and stuff categories as instance-level learning. We train our models on COCO panoptic segmentation and evaluate them on COCO stuff and panoptic segmentation. In Tab. 6, the result indicates that our unification significantly enhances both segmentation performances in terms of PQ and mIoU. We also combine LVIS and COCO stuff as training data, and the results show that our improvement is also helpful for large-scale vocabularies.

To validate the compatibility of class-agnostic and semantic-aware annotations, we jointly train with SA-1B, LVIS, and MSCOCO. As shown in Tab. 7, the result indicates that SA-1B significantly helps the instance-level detection and segmentation for large-scale vocabularies.

4.4. Performance on Single Dataset

We further evaluate the performance of APE on a single benchmark **without** additional training data. For the long-tailed LVIS, we choose FedLoss [54] as the classification loss to remedy the impact of unbalanced data distribution. We compare the performances of instance segmentation in Tab. 8, and all methods are **not** pre-trained on Objects365. Our method suppresses all other models and achieves state-of-the-art results on MSCOCO and LVIS.

5. Conclusion

We present a APE, a universal visual perception model for **a**ligning and **p**rompting **e**verything all at once in an image to perform diverse tasks, *i.e.*, detection, segmentation, and grounding, as an instance-level sentence-object matching paradigm. APE is trained on broad data with natural and challenging characteristics, such as Zipfian distribution of categories, federated annotations, anything segmentation, and mixed vocabulary and sentence concepts. The extensive experiments show that APE outperforms (or is on par with) the existing SotA models with only one-suit of weights, proving that an effective yet universal perception for anything prompting and alignment at scale is indeed feasible.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NO. 62102151), Shanghai Sailing Program (21YF1411200), CCF-Tencent Rhino-Bird Open Research Fund, the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education (KLATASDS2305), Fundamental Research Funds for the Central Universities.

References

[1] 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 (ECCV)*, 2020. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 5

[3] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A Language Modeling Framework for Object Detection. In *The International Conference on Learning Representations (ICLR)*, 2021. 2

[4] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-Pixel Classification Is Not All You Need for Semantic Segmentation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 4

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-Attention Mask Transformer for Universal Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4, 5

[6] Floriana Ciaglia, Francesco Saverio Zuppichini, Paul Guerrie, Mark McQuade, and Jacob Solawetz. Roboflow 100: A Rich, Multi-Domain Object Detection Benchmark. 2022. 6

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 3, 7

[8] 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. In *The International Conference on Learning Representations (ICLR)*, 2021. 5, 6

[9] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A Visual Representation for Neon Genesis. *arXiv preprint arXiv:2303.11331*, 2023. 8, 12

[10] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 12

[11] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung Yi Lin. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In *European Conference on Computer Vision (ECCV)*, 2022. 6

[12] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5, 6, 12

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 12

[14] Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[15] 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. In *International Conference on Machine Learning (ICML)*, 2021. 2

[16] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR – Modulated Detection for End-To-End Multi-Modal Understanding. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 4, 6, 13, 14

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. 2, 3, 5, 8

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 2017. 5

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. the Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision (IJCV)*, 2018. 5, 6

[20] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: Towards a Unified Transformer-Based Framework for Object Detection and Segmentation. 2022. 2, 3, 5, 8

[21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded Language-Image Pre-Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 6, 7, 13, 14

[22] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring Plain Vision Transformer Backbones for Object Detection. In *European Conference on Computer Vision (ECCV)*, 2022. 8

[23] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[24] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-Vocabulary Semantic Segmentation with

Mask-Adapted CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[25] 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 *European Conference on Computer Vision (ECCV)*, 2014. 5, 6, 12

[26] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 5

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. 2023. 2, 3, 4, 6, 7, 13, 14

[28] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *The International Conference on Learning Representations (ICLR)*, 2019. 12

[29] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *International Conference on 3D Vision*, 2016. 5

[31] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple Open-Vocabulary Object Detection with Vision Transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 6, 7

[32] Jeffrey Ouyang, Zhang Jang, Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. NMS Strikes Back. *arXiv preprint arXiv:2212.06137*, 2022. 5, 12

[33] 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. 2023. 14

[34] Bryan A Plummer, Liwei Wang, · Chris, M Cervantes, Juan C Caicedo, Julia Hockenmaier, Svetlana Lazebnik, B A Plummer, L Wang, J C Caicedo, and J Hockenmaier. Flickr30k Entities: Collecting Region-To-Phrase Correspondences for Richer Image-To-Sentence Models. *International Journal of Computer Vision (IJCV)*, 2015. 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 4, 7

[36] 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. *Journal of Machine Learning Research*, 2020. 7

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 5

[38] Hamid Rezatofighi, Nathan Tsoi, Junyoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[39] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Yu † Gang, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 5, 6

[40] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale. 2023. 12

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael, Smith Ranjan, Subramanian Xiaoqing, Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. 4, 7

[42] 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. In *International Conference on Machine Learning (ICML)*, 2022. 7

[43] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical Open-Vocabulary Universal Image Segmentation. 2023. 3, 13

[44] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-Based Image Segmentation in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[45] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: Adapting CLIP for Open-Vocabulary Detection with Region Prompting and Anchor Pre-Matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7

[46] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described Object Detection: Liberating Object Detection with Flexible Expressions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3, 6, 7, 12, 15, 16, 17

[47] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-To-Image Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 13

[48] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal Instance Perception As Object Discovery and Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 6, 7, 12, 13

[49] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-Training for Open-World Detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[50] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[51] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying Localization and Vision-Language Understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 4, 6, 12, 13

[52] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. a Simple Framework for Open-Vocabulary Segmentation and Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 6, 13

[53] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. In *The International Conference on Learning Representations (ICLR)*, 2022. 2

[54] Xingyi Zhou, Vladlen Koltun, Philipp Krähenb, and Krähenb¨ Krähenbühl. Probabilistic Two-Stage Detection. *arXiv preprint arXiv:2103.07461*, 2021. 8

[55] 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 (ECCV)*, 2022. 2

[56] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized Decoding for Pixel, Image, and Language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 6, 12, 13, 18

[57] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment Everything Everywhere All at Once. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 13