The All-Seeing Project: Towards Panoptic Visual Recognition and Understanding of the Open World

Weiyun Wang[*1,2], **Min Shi**[*1,3], **Qingyun Li**[*1,4], **Wenhai Wang**[*1], **Zhenhang Huang**[*1], **Linjie Xing**[*1], **Zhe Chen**[1,5], **Hao Li**[1,6], **Xizhou Zhu**[1,7], **Zhiguo Cao**[3], **Yushi Chen**[4], **Tong Lu**[5], **Jifeng Dai**[†1,8], **Yu Qiao**[1]

[1]OpenGVLab, Shanghai AI Laboratory [2]Fudan University
[3]Huazhong University of Science and Technology [4]Harbin Institute of Technology
[5]Nanjing University [6]The Chinese University of Hong Kong
[7]SenseTime Research [8]Tsinghua University

Code: `https://github.com/OpenGVLab/all-seeing`
Demo: `https://huggingface.co/spaces/OpenGVLab/all-seeing`

Abstract

We present the All-Seeing (AS)[1] project: a large-scale data and model for recognizing and understanding everything in the open world. Using a scalable data engine that incorporates human feedback and efficient models in the loop, we create a new dataset (AS-1B) with over 1 billion regions annotated with semantic tags, question-answering pairs, and detailed captions. It covers a wide range of 3.5 million common and rare concepts in the real world, and has 132.2 billion tokens that describe the concepts and their attributes. Leveraging this new dataset, we develop the All-Seeing model (ASM), a unified framework for panoptic visual recognition and understanding. The model is trained with open-ended language prompts and locations, which allows it to generalize to various vision and language tasks with remarkable zero-shot performance, including region-text retrieval, region recognition, captioning, and question-answering. We hope that this project can serve as a foundation for vision-language artificial general intelligence research.

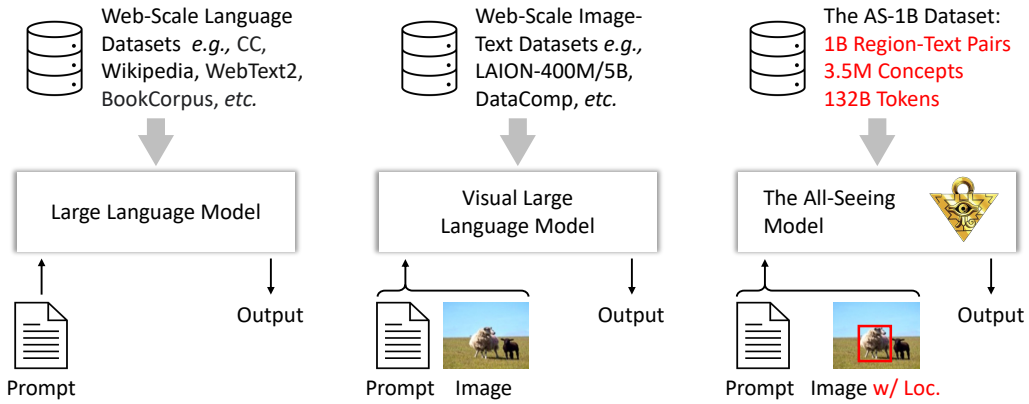
1 Introduction

Creating artificial general intelligence (AGI) systems that can match human intelligence and excel in any task across domains is the ultimate goal of artificial intelligence. Recent advancements in Large Language Models (LLMs) have demonstrated impressive zero-shot capabilities in user-tailored natural language processing (NLP) tasks, suggesting new avenues for achieving AGI. However, as shown in Fig. 1a, most popular LLMs [62, 64, 77, 78, 80, 19, 17] are limited to processing language information and lack the ability to perceive or understand the visual world.

Although there have been some recent developments [61, 111, 54, 48, 47, 22, 96, 55] in open-world visual understanding, they are primarily focused on understanding images as a whole, rather than recognizing and comprehending individual instances within the scene (see Fig. 1b). *This goes against the nature of the human visual system*, as described by the feature integration theory [81], which

[*]Equal contribution. This work is done when Weiyun Wang, Min Shi, and Qingyun Li are interns at Shanghai AI Laboratory. [†]Corresponding to Jifeng Dai <daijifeng@tsinghua.edu.cn>.

[1]"All-Seeing" is derived from "The All-Seeing Eye", which means having complete knowledge, awareness, or insight into all aspects of existence.



(a) Large Language Models (LLMs) possess extensive world knowledge and demonstrate impressive reasoning capabilities, but lack the ability to receive and comprehend visual information.

(b) Visual Large Language Models (VLLMs) can process both text and images, but they can only capture the holistic visual information of the whole image and understand it based on LLMs.

(c) Our All-Seeing Model (ASM) can comprehensively recognize and understand the objects or concepts in regions of interest, while maintaining the capabilities of VLLMs and LLMs.

Figure 1: **Overview and comparison of our All-Seeing project with other popular large foundation models.** To address the limitations of LLMs in understanding visual inputs and VLLMs in effectively leveraging region-aware information, we propose (1) a large-scale dataset AS-1B which consists of 2 billion region-text pairs, 3.5 million open-world concepts, and over 1 billion tokens of region-related question-answering and caption; and (2) the All-Seeing model (ASM), which is capable of recognizing and understanding context in arbitrary regions.

suggests that we attentively gather visual features and contexts in certain regions to achieve high-level understanding and recognition, rather than analyzing all information simultaneously.

To achieve instance-level visual understanding like humans, there are two major challenges as follows: (1) *The scarcity of open-world instance-text pair data.* As listed in Table 1, existing datasets, such as Visual Genome [43], have limitations in terms of data scale and open-world annotations. Laion-5B [72] only contains web-crawled image-text pairs without location information, and SA-1B [42] lacks semantic information. (2) *The lack of spatial information modeling in most existing models*. These models mainly focus on whole-image understanding as mentioned above.

In this work, we propose the All-Seeing (AS) project for open-world panoptic visual recognition and understanding, driven by the goal of creating a vision system that mimics human cognition. The term "panoptic" refers to including everything visible in one view [41]. The AS project addresses the challenges from both the data and model perspectives.

From the data aspect, we propose the All-Seeing 1B (AS-1B) dataset, consisting of over 1 billion region annotations in various formats, such as semantic tags, locations, question-answering pairs, and captions (refer to Fig. 2). AS-1B dataset is made possible by a scalable semi-automatic data engine, which significantly lowers the previously unaffordable expense of manually annotating a massive amount of open-world semantics. The data engine operates in a "data-human-model" loop, iteratively refining data quality. Initially, diverse models, including large language models (LLMs) [17], detection[88, 25, 51], captioning [48], and visual question answering models (VQA) [54, 111, 55], are employed as "annotators", which add semantic annotations to dense region proposals generated by state-of-the-art object detectors [42, 25, 51, 88]. Subsequently, human annotators verify the generated pseudo labels and provide feedback with high-quality data, which is then used to fine-tune the models to improve their performance. The enhanced models are then used to re-annotate the data, starting another iteration of the loop. As shown in Fig. 2, AS-1B contains a wide range of open-world concepts, including over 3.5 million different semantic tags ranging from common categories (*e.g.*, human, backpack) to fine-grained or rare categories with attributes (*e.g.*, old metal latches). AS-1B also encompasses the 3.3 billion visual question-answering pairs and 1.2 billion region captions for 1.2 billion regions.

In terms of the model perspective, we propose the All-Seeing model (ASM), a unified location-aware image-text foundation model. The model consists of two key components: a location-aware image tokenizer and an LLM-based decoder. The location-aware image tokenizer uses location information such as box, mask, and point set as conditions (see Fig. 1c) to extract image features, which contribute to the location capability of ASM. The LLM-based decoder inherits the world knowledge and reasoning capability from LLMs such as LLaMA [80], providing a strong foundation for visual recognition and understanding. In addition, to unify image-text aligning and generation tasks, we introduce a new decoding approach, where the aligning tasks are reformulated as a "special" generation task, enabling our model to generalize to various vision-language tasks with shared weights.

Compared to previous methods [67, 2, 48, 54, 111], our work offers several advantages as follows: (1) Our model not only excels in image-level understanding but also demonstrates exceptional capability in recognizing and comprehending objects at the instance level, closely aligning with human cognitive processes. (2) Our model is a unified framework that supports a wide range of image-text tasks, including discriminative tasks like image-text retrieval, as well as generative tasks such as visual captioning and question-answering. (3) Our model comes with AS-1B the largest dataset with open-world panoptic semantics. Data and models feed each other in the data engine, iteratively improving the model performance, data scale and diversity.

In summary, our contributions are three folds:

(1) We propose a new large-scale dataset (AS-1B) for open-world panoptic visual recognition and understanding, using an economical semi-automatic data engine that combines the power of off-the-shelf vision/language models and human feedback. As reported in Table 1, we have 159 times more semantic tags and 33 times more regions compared with its counterparts.

(2) Based on the dataset, we develop a unified vision-language foundation model (ASM) for open-world panoptic visual recognition and understanding. Aligning with LLMs, our ASM supports versatile image-text retrieval and generation tasks, demonstrating impressive zero-shot capability.

(3) We evaluate our model on a representative vision and vision-language tasks. Our ASM outperforms CLIP [67] by 10.4 and 14.3 (mAP) on COCO [53] and LVIS [31] in zero-shot region recognition tasks. When trained with AS-1B (region-level data) and LaionCOCO [71] (image-level data), our model achieves superior zero-shot and fine-tuned performance compared to recent image-level [47, 22, 87, 98, 35] and region-level [99, 92, 65] VLLMs.

2 Related Work

The Emergence of Large Language Models. In recent years, based on the large-scale text corpora [28, 83, 69, 116, 82, 100], the field of Large Language Models (LLMs) has witnessed remarkable progress [69, 8, 56, 70, 106]. Prominent models such as ChatGPT [62] and GPT-4 [61] have demonstrated excellent performance across various tasks, showcasing their potential for semantic understanding, dialogue generation, programming, mathematical problem-solving, and more. However, there is a growing concern that these leading institutes are becoming increasingly conservative in sharing the technical details of their models and roadmaps. To catch up with the performance of ChatGPT, the open-source community has devoted substantial efforts [80, 90, 77, 17, 102, 29, 104]. For instance, Self-Instruct [90] introduced an iterative bootstrapping algorithm that leverages off-the-shelf LLMs and a seed set of manually-written instructions to expand the instruction collection. Alpaca [77] utilized the Self-Instruct technique to generate high-quality instruction-following data, which was then used to fine-tune the LLaMA [80] model. Vicuna [17] demonstrated that fine-tuning on user-shared ChatGPT conversations can spark dialog and improve instruction-following capabilities. Furthermore, there has been a focus on improving multilingual capabilities, particularly in Chinese, with LLMs like Chinese-Alpaca [21], GLM-130B [102], InternLM [78], MOSS [19], and others. These LLMs have shown excellent proficiency in learning world knowledge, which lays the groundwork for open-world understanding.

Datasets for Visual Recognition and Understanding. The dataset plays a critical role in the advancement of deep learning models, especially in the field of visual recognition and comprehension. Prior to the era of large-scale models, datasets are primarily closed-world or have limited data scale, including CIFAR-10/100 [44], ImageNet [23], and iNaturalist [84] for image classification,

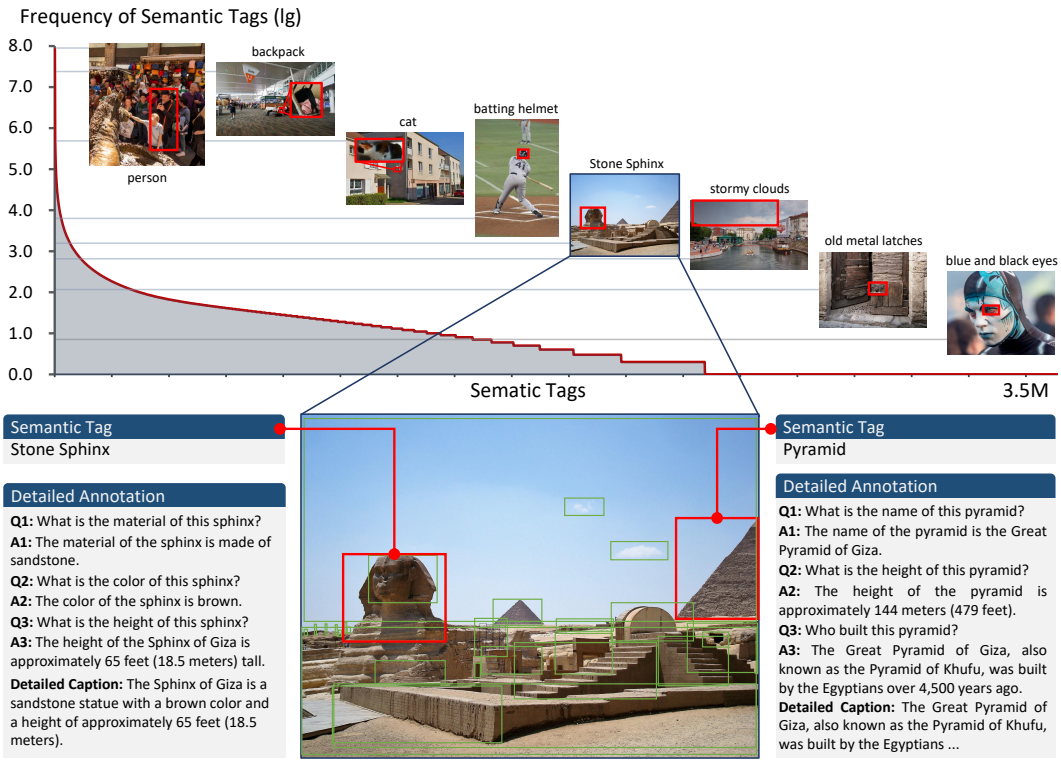


Figure 2: **Semantic concepts and annotations in the AS-1B dataset**. The semantic tags in AS-1B dataset encompass a wide range of concepts, from common objects to rare and fine-grained categories with attributes. Beyond brief semantic tags, detailed annotations, including visual-question-answering pairs and region captions are also provided.

Dataset	#Images	#Regions	#Concepts	#Tokens	Location	Semantic
Image-Level						
ImageNet-22K [23]	15M	–	22,000	–	–	Closed-Set
COCO Caption [15]	0.1M	–	–	8.4M	–	Closed-Set
SBU [63]	0.8M	–	–	14.6M	–	Open-World
CC12M [11]	12.4M	–	–	250.9M	–	Open-World
YFCC15M [38]	15M	–	–	1.0B	–	Open-World
COYO700M [9]	700M	–	–	15.0B	–	Open-World
Laion-5B [72]	5B	–	–	135.0B	–	Open-World
Class-Agnostic						
SA-1B [42]	11M	1.1B	–	–	Open-World	–
Region-Level						
COCO [53]	0.1M	0.9M	80	–	Closed-Set	Closed-Set
LVIS [31]	0.1M	1.5M	1,203	–	Closed-Set	Closed-Set
Objects365 [74]	0.6M	10.1M	365	–	Closed-Set	Closed-Set
Open Images [45]	1.5M	14.8M	600	–	Closed-Set	Closed-Set
BigDetection [10]	3.5M	36.0M	600	–	Closed-Set	Closed-Set
V3Det [86]	0.2M	1.5M	13,029	–	Closed-Set	Closed-Set
Visual Genome [43]	0.1M	0.3M	18,136	51.2M	Open-World	Open-World
AS-1B (ours)	**11M**	**1.2B**	**3.5M**	**132.2B**	Open-World	Open-World

Table 1: **Comparison with popular vision and vision-language datasets**. "#" denotes the number of something. We see that the proposed AS-1B dataset has a significantly larger data scale and diversity than prior region-level datasets.

Pascal VOC [24], COCO [53], LVIS [31], OpenImages [45], ADE20K [109], and Cityscape [20]

for visual location, as well as SBU [63], CC3M [75], CC12M [11], YFCC15M [79], and VQA [4], VQA 2.0 [30], ICDAR 2015 [40], SCUT-CTW1500 [101] for visual understanding. Additionally, datasets like Visual Genome [43] and Visual7W [115] integrate visual location and understanding, offering more comprehensive tasks to describe the visual world. However, these datasets have limited semantics and fail to encompass diverse scenarios in the open world, which hinders the generalization ability of models. To achieve open-world capability, CLIP [67] and ALIGN [37] propose training models using web-scale image-text pairs collected from the internet. Subsequent works, such as Laion-400M [73], Laion-5B [72], COYO-700M [9] and DataComp [27], have also been introduced for open-source research. However, these approaches only include descriptions or question-answering pairs corresponding to the entire image, resulting in models struggling to accurately recognize and understand specific objects at the instance level. Recently, Kirillov et al. introduced SA-1B [42], which provides open-world location information such as boxes and masks but still lacks semantic details. So existing datasets cannot meet the requirements of data scale, open-world location and semantics necessary for achieving visual AGI models, thus posing challenges in supporting human-like panoptic visual recognition and understanding.

Models for Visual Recognition and Understanding. Significant advancements have been made in the field of visual recognition and understanding in recent years. Previous methods [33, 39, 16, 113, 14, 41, 93, 46] mainly concentrate on the close-set recognition while recent works begin to focus on the open world understanding. Models trained with contrastive learning-based methods, including CLIP [67], ALIGN [37], EVA [26] and FLIP [52], are able to recognize and understand the open world semantics under an image-text matching framework while the lack of generation ability limits their applicability. To address this limitation, subsequent works, such as SimVLM [91], UniPerceiver [114], VL-BERT [7], VLMo [6], BEiT-3 [89], ALBEF [49], CoCa [98], as well as Flamingo [2], have incorporated generative training tasks. However, these models are trained from scratch and do not capitalize on the powerful perception capabilities of existing powerful vision foundation models for image, and Large Language Models for text, increasing the cost of developing new models. The recent progress of LLMs [61, 62, 68, 69, 8] initiates a new era, leading to the emergency of many LLM-based multimodal models [48, 47, 111, 54, 96, 104, 60, 87, 12] and interactive systems [94, 55, 76, 110, 50, 112, 95]. However, these works are only capable of recognizing the entire image, lacking the ability to comprehend specific regions within the image. Some concurrent methods, such as ChatSpot [107], Shikra [13], KOSMOS-2 [65], and GPT4RoI [105] begin to focus on location-aware understanding. However, without the support of large-scale instance-level visual understanding data, the generalization ability of these models is still limited. Besides, these models only support generative tasks, limiting their application to discriminative tasks, such as image-text retrieval and zero-shot object recognition. In this work, we propose a unified location-aware image-text foundation model, based on ViT-g [26] and Husky [55]. Our model supports both image-text matching and generation tasks, expanding its range of applications and contributing to the advancement of AGI models.

3 The All-Seeing Dataset (AS-1B)

In this section, we introduce the All-Seeing-1B (AS-1B) dataset for open-world panoptic visual recognition and understanding. The dataset consists of 1.2 billion regions in 11 million images[2]. Each region is annotated with comprehensive information, including categories, locations, attributes, captions, and question-answer pairs. Compared with the previous visual recognition datasets like ImageNet [23] and COCO [53], visual understanding datasets like Visual Genome [43] and Laion-5B [72], *the proposed AS-1B dataset stands out due to its rich and diverse instance-level location annotation and corresponding detailed object concepts and descriptions.*

3.1 Data Annotation Engine

We develop a semi-automatic data engine that efficiently uses a wide range of state-of-the-art foundation models as annotators, reducing the enormous labeling cost to an acceptable level. As depicted in Fig. 3, the process of the data engine begins by generating noisy pseudo data using well-trained off-the-shelf foundation models from diverse fields. Subsequently, these pseudo data are iteratively refined through multiple loops with the aid of models fine-tuned on human feedback

[2]Images source from SA-1B [42]

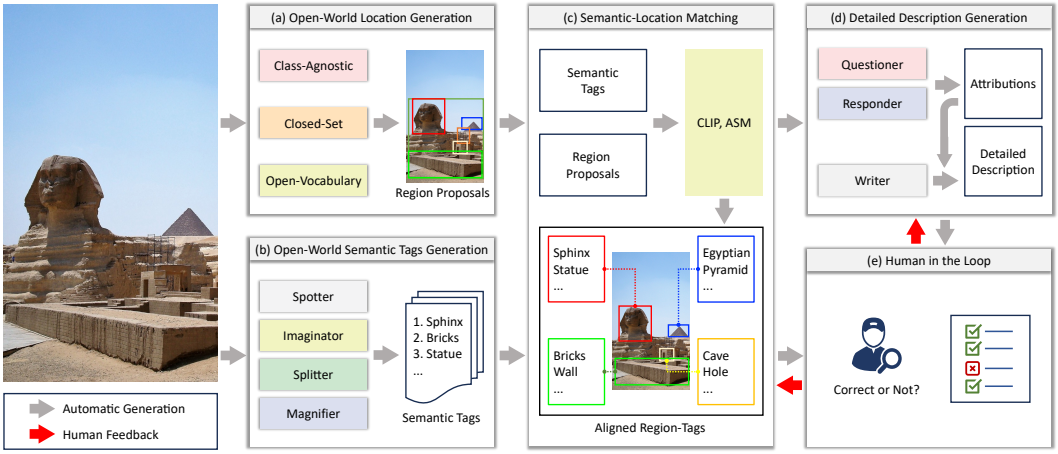


Figure 3: **Data engine for AS-1B dataset**. Our data engine consists of an automatic annotation pipeline (*i.e.*, (a), (b), (c), (d)) and human verification stage (*i.e.*, (e)). We combine strong object detectors, LLMs, and VLLMs to produce open-world locations and annotations for different regions. The automatic annotations are sampled and verified by human experts. Automated annotations are used together with human validation results to train region-aware alignment and generation models, which are then used in the automated annotation pipeline to improve data quality.

data. By employing this "data-human-model" cycle, we can generate a large number of region-level annotations with exceptional quality.

As the core component of the data engine, the pseudo data generation pipeline consists of five steps as follows: (1) Creating open-world location (*e.g.*, bounding box, mask, point set) with an ensemble of state-of-the-art class-agnostic, visual grounding, and closed-set perception models [42, 51, 88, 25]; (2) Generating open-world semantic tags using the combination of image captioning models [48, 111] and LLMs [17]; (3) Matching the semantic tags to proper regions with image-text aligning models such as CLIP [67]; (4) Using LLM [17] and VQA models [55] to generate the attributions of each region based on the matched semantic tags; (5) Generating detailed captions based on the semantics and attributions of each region.

3.2 Open-World Localization

To obtain comprehensive locations of all instances in an image, we combine the results of state-of-the-art perception models from different fields, including (1) **class-agnostic model**: we adopt the SAM [42] to provide initial proposals of most objects in view. (2) **closed-set detection model**: we use InternImage-H [88] and EVA-02 [25] trained on BigDetection [10] and LVIS [31], respectively, to detect the common-seen objects. (3) **grounding model**: we use GLIP [51] to ground open-world semantics generated by LLMs [111] (see Sec. 3.3). All the bounding boxes are gathered together to ensure that all possible objects in view are covered.

Due to the incomparable score ranges of different models, directly using non-maximum suppression (NMS) to eliminate duplicated proposals from multiple resources is infeasible. Therefore, we develop an effective strategy that keeps all the semantics while removing highly-overlapped regions. As shown in Alg. 1, the merging strategy works as follows: (1) We start by initializing the result region proposal set $\mathcal{R}$ with the class-agnostic bounding boxes generated by SAM. (2) When a set of region proposals $\mathcal{R}'$ from a new source (*e.g.*, closed-set/grounding detector) comes in, we calculate the Intersection over Union (IoU) between the regions in $\mathcal{R}'$ and $\mathcal{R}$. (3) If the IoU between a new region $r' \in \mathcal{R}'$ and an existing region $r \in \mathcal{R}$ is greater than a threshold T_{IoU}, the region r' is removed, and its closed-set/grounding tags are appended to the tag list of the matched region r. (3) Finally, the remaining low-IoU regions in $\mathcal{R}'$ along with their tags are added to $\mathcal{R}$. By employing this strategy, we sequentially combine the results of SAM, InternImage, EVA-02 and GLIP to obtain comprehensive location information for an image.

Algorithm 1 Region Proposal Merging

Input:

Existing region proposals $\mathcal{R}$

New region proposals $\mathcal{R}'$

IoU threshold T_{IoU}

Output:

Merged region proposals $\mathcal{R}$

for region $r' \in \mathcal{R}'$ **do**
  Calculate IoU between r' and proposals in $\mathcal{R}$
  if maximum IoU > T_{IoU} **then**
    Merge semantic tags from r' into the semantic tag of corresponding regions in $\mathcal{R}$
    Delete r'
  else
    Add r' into $\mathcal{R}$
  end if
end for

3.3 Open-World Semantic

Manually labeling billions of regions for an open-world semantic description is impractical due to the enormous cost and time required. On the other hand, generating annotations with off-the-shelf multi-modal models is also non-trivial, as it demands sufficient world knowledge and context-related reasoning capabilities to accurately label diverse objects in the wild. To remedy these challenges, we draw inspiration from the recent advancements in Large Language Models (LLMs) [8, 80, 19, 77, 17, 78, 102] and Visual Large Language Models (VLLMs) [61, 54, 48, 55, 87, 111, 96], we leverage a series of LLMs and VLLMs as "semantic generators" and tap into their vast world knowledge and reasoning abilities for open-world semantic generation. These "semantic generators" can be specialized for producing short semantic tags (such as category names and brief attributes) or detailed annotations (including attributes, question-answering pairs, and captions) based on specially-designed prompts.

3.3.1 Semantic Tags

To generate as many semantic tags as possible for a view, different instructions are employed to harness the diverse capabilities of LLMs and VLLMs, turning them into annotators with different focuses and skills. Specifically, we have (1) a **spotter**, which identifies major instances and provides an overview of the scenes, (2) a **imaginator** that leverages world knowledge to imagine plausible objects, (3) a **splitter** that divides complicated objects into parts, as well as (4) which zooms on each region to produce region-specific candidates. These models complement each other to create a powerful system that can generate comprehensive open-world semantic tags for each region and the entire image. Here are the details of each model:

Spotter. This module aims to list the prominent and major objects present in the given image. To achieve this, we use MiniGPT4 [111] to provide an overall caption of the input image. From the generated captions, we extract noun phrases to serve as the semantic tags shared by all the regions in the input image. In addition, we also add an OCR detector [18] to detect the texts as semantic tags in the scenes. Note that the generated caption will also be passed to other annotators, which gives visual signal for the LLMs, serving as their eyes.

Imaginator. Although the "spotter" can find out the major objects in the scenes, it fails to identify many insignificant objects. To address this limitation, we develop an "imaginator" to further expand the semantic tag list with plausible imagination. The "imaginator" emulates human-like thinking. When provided with descriptions of a particular scene, humans can effortlessly imagine the potential objects present. For instance, if informed that an image depicts a group of children standing in a classroom, one may envision objects like "teacher", "blackboard", and "stationery". In our data engine, we utilize Vicuna [17] to imagine possible objects in scenes based on the captions generated

by the “spotter”, and then extend the set using web search engines [66]. The “imaginator” excels at supplementing scene-specific object candidates, such as suggesting “airport stuff” instead of simply “person”. This significantly enhances the concept diversity within this project.

Splitter. This model is proposed to divide the generated concepts into more fine-grained parts. We find that some region proposals only cover a part of the objects, such as the wing of a plane or the windshield of a car. However, most of the existing perception or caption models are not capable of detecting parts. To this end, we further instruct the Vicuna [17] to divide the semantic tag into parts. For example, “building" will be decomposed into “roof”, “door”, “windows” and “walls”. We tailor the prompt for LLM so that the model only divides the semantic tag that represents a concrete object into parts. LLM is instructed to ignore the semantic candidate that is non-physical or cannot be further divided, such as “water”, “sky”, etc.

Magnifier. Although hundreds of open-world semantic tags can be generated by the aforementioned annotators for each image, there still exists some regions whose semantics are absent from the generated tag lists. So we introduce a “magnifier” to zoom in on each region and add semantic tags for them. We simply crop the region and use a caption model to describe the cropped image, and then extract the noun phrases, which are used as the semantic candidates exclusive for the corresponding regions. In this model, we use BLIP [48] for efficiency.

3.3.2 Detailed Descriptions

To provide detailed descriptions that include attributes and statuses of each region, we develop a pipeline that expands the region description using the open-world location and its matched semantic tags (see Sec. 3.4 for location-semantic matching). Similar to how we generate semantic tags, we utilize a series of skilled LLMs, including (1) a **questioner** that asks specific questions about the attributes or status of a given semantic tag; (2) a **responder** that provides the accurate answers for these questions based on the region’s content; and (3) a **writer** responsible for composing a detailed caption for each region, according to the generated semantic tags, attributes, and status.

Questioner. Given semantic tag, to determine its commonly-used attributes, we use Vicuna [17] as a questioner to generate three questions about the attributes or statuses. The prompt is shown below. In this way, we leverage the world knowledge and reasoning capabilities of LLMs to identify the most relevant attribute of an object.

> **Prompt:** I will give you some objects. Please list 3 questions about the given objects. These questions must be answerable based on a photograph of the object and cannot rely on any outside knowledge. Some examples are listed as follows:
>
> **Human**: Person
> **Assistant**: Q1: What is the sex of this person? Q2: What is the hairstyle of this person? Q3: What is this person doing?
>
> **Human**: {Semantic Tag}
> **Assistant**:

Responder. After obtaining the questions related to a semantic tag, we employ Husky [55], an LLM-based VQA model, to generate the responses to each question. The responses are generated in several sentences, taking into account the content of the region. An example prompt is shown below. This approach enables us to gather additional information about a region while preventing the inclusion of irrelevant content.

> **Human**: What is the material of this sphinx? **Assistant**:

Writer. Based on the question-answering pairs, we proceeded to rephrase them into a single sentence, resulting in a detailed description of the region. The prompt used during annotation is “Please paraphrase the following sentences into one sentence. {answer for question 1} {answer for question 2} {answer for question 3}”. It is notable that both the question-answering pairs from previous steps and the region captions from this step are valuable for visual recognition and understanding models.

3.4 Matching Location and Semantic

Given the generated open-world location and semantic labels, we devise a matching pipeline to select and appropriate tags for each region. Semantic tags that are most related to the region will be picked.

In the matching process, we employ a region-text aligning model to measure the similarity between a certain region and its semantic tag list. For each region, the semantic tag list is constructed by LLMs (*i.e.*, "spotter", "imaginator", and "divider") and closed-set/grounding object detectors. Initially, in the first iteration of the data engine, we use a CLIP model [67] for the region-text alignment, where the input is the cropped region. Subsequently, we upgrade the model to our All-Seeing Model.

In addition, in the first round of data engine, we find that only using CLIP led to erroneous results as it cannot tell which candidate is the major object in the bounding boxes. For example, a bounding box that perfectly frames a person can be classified as a "backpack" if the person is carrying a backpack. To remedy this, we use CLIPSeg [58] to generate the mask for each candidate, and the original CLIP confidence is modulated with the corresponding mask area. In this way, the candidate belonging to the main object in the region can be selected.

3.5 Human Verification

Albeit efficient, annotations from the automated pipeline still contains some noise due to the cropping process, which might discard essential context information. For instance, a lampshade hanging on the ceiling could be mistakenly described as a "cup" due to its similar shape and color. Therefore, to enhance the data quality, we find it crucial to include human verification.

Semantic tags. We design a data sampling strategy and simplify the task for annotators by focusing on picking the incorrect ones from the top-5 candidates in each region. In the real world, concepts exhibit long-tail distribution as shown in Fig. 2. Therefore, many rare concepts will be missed if the region is randomly sampled for validation. To address this issue, we implement a concept-wise sampling strategy. Specifically, we collect a list of concepts in the first 1M images in the AS-1B dataset. From this list, we select most concepts for verification. We randomly sample 6 regions from the least frequent concepts and 90 regions from the concepts with the highest number of regions. During the human verification process, the semantic tag list for the sampled regions is provided to the annotators, who are then tasked with filtering out any incorrect tags.

Visual Question-Answering Pairs. Although using LLMs/VLLMs greatly reduces the annotation cost of generating visual question-answer pairs, there are still some issues that may introduce noise into the data. (1) The answer to the question is wrong since the VLLM is not perfect. (2) The generated question for the semantic tag may be unanswerable according to the given image content. (3) The semantic tag assigned to a region may be incorrect, leading to meaningless generated questions. For example, if a region containing a dog is wrongly labeled as a cat, asking about the color of the cat would be nonsensical.

To address these issues, we perform a two-stage verification procedure. In the first stage, human annotators are provided with the image, location (bounding box), and corresponding question-answer pairs. They are then asked to annotate the visual question-answer pair with one of four choices: correct answer, wrong answer, unanswerable question, or wrong semantic tag. Samples annotated as "correct answer" are retained, while those annotated as "wrong answer" are re-annotated with a correct answer generated by human annotators in the second stage. Samples annotated as "unanswerable question" or "wrong semantic tag" are annotated with a rejection answer, such as "This question is unanswerable according to the image" or "The object in this region is incorrectly labeled", respectively.

Verification Review. We engaged 50 human annotators to perform verification on the annotations generated by our model. To guarantee the quality of this verification process, we additionally request 10 experts to review the verified annotations. These experts are selected based on their domain knowledge and experience in annotation tasks. To streamline the process, we organize the regions requiring review into groups of 100. Each group is assigned to one expert, who checks the accuracy and consistency of the annotations within the group. Any package with an accuracy rate below 95% will be sent back for re-verification by another annotator. This review process double-checks the annotations, further ensuring their reliability and validity for our models.

3.6 Data Engine Iteration

To continuously improve the data quality, we implement a "data-human-model" loop that maximizes the utilization of both human-verified data and models. As depicted Alg. 2, the data engine iteration comprises three steps as follows: (1) The images are processed with the annotation pipeline which produces automatic annotations. (2) The ASM model is then trained using these coarse annotations, enabling it to perform both discriminative and generative tasks such as region-text matching and region captioning. (3) The automatic annotations are sampled and reviewed and corrected by human annotators, yielding high-quality human annotations. This verified data is then used to fine-tune the ASM model, thereby enhancing its performance. (4) The fine-tuned model is utilized to re-rank the semantic tags and generate more accurate region captions and answers. Repeat the third and fourth steps until the data quality meets the requirements. By following this data iteration process, we ensure continuous optimization of data quality, ultimately leading to superior results.

Algorithm 2 Data Engine

Input:

Iteration Number n
Images $\mathcal{I}$
Models $\mathcal{M}$
Annotation Pipeline $P(\mathcal{M}, \mathcal{I})$

Output:

Annotations: $\mathcal{A}$
Improved Models $\mathcal{M}$

Generate initial annotation $\mathcal{A}_0$ by off-the-shelf models;
Train ASM with $\mathcal{A}_0$, yield $\mathcal{M}_0$;
$i \leftarrow 0$
while $i < n$ **do**
    Perform Human verification on $\mathcal{A}_i$, yield $\mathcal{A}'_i$;
    Fine-tune $\mathcal{M}_i$ with $\mathcal{A}'_i$, obtain $\mathcal{M}_{i+1}$;
    Obtain Annotation $\mathcal{A}_{i+1}$ by $P(\mathcal{M}_{i+1}, \mathcal{I})$;
    $i \leftarrow i + 1$
end while

4 The All-Seeing Model (ASM)

4.1 Overal Architecture

Our objective is to create a unified framework that supports contrastive and generative image-text tasks at both the image level and region levels. By leveraging pre-trained LLMs and powerful vision foundation models (VFMs), this model demonstrates promising performance in discriminative tasks like image-text retrieval and zero classification, as well as generative tasks such as visual question answering (VQA), visual reasoning, image captioning, region captioning/VQA, etc. Additionally, our model shows potential in grounding tasks like phrase grounding and referring expression comprehension, with the assistance of a class-agnostic detector.

As illustrated in Fig. 4, our All-Seeing Model (ASM) comprises three key designs: (1) a **location-aware image tokenizer** extracting features from both the image and region levels based on the input image and bounding box, respectively. (2) a **trainable task prompt** that is incorporated at the beginning of the vision and text tokens to guide the model in distinguishing between discriminative and generative tasks. In the case of the discriminative task, a trainable align token is appended to the input sequence to gather the overall representation, and its embedding is then used in the matching process. (3) an **LLM-based decoder** that is utilized to extract vision and text features for discriminative tasks, as well as to auto-regressively generate response tokens in generative tasks.

The training objective of ASM contains two objectives: next token prediction and region-text aligning, as formulated in Eqn. 1. The primary objective focuses on enhancing the model's generation capability,

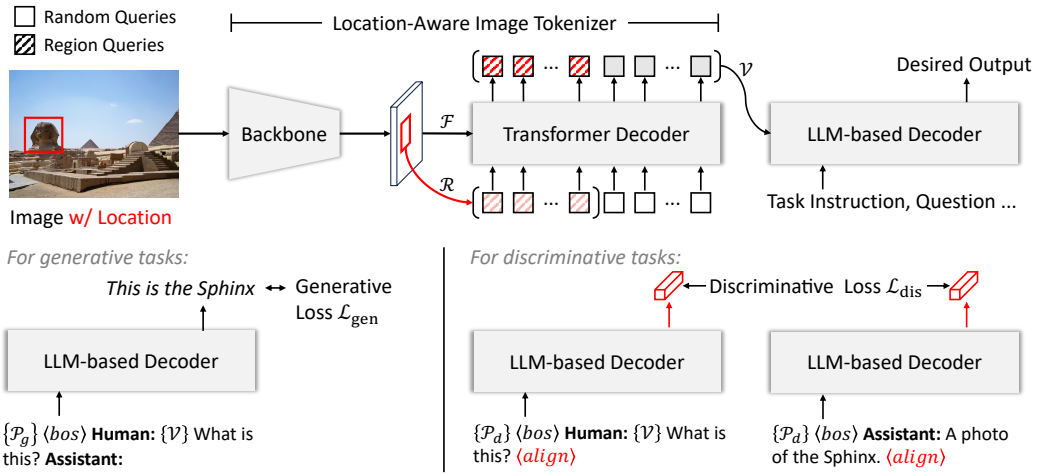


Figure 4: **Architecture and task modeling of the All-Seeing Model (ASM)**. ASM incorporates a location-aware image tokenizer to perform region-text alignment tasks. Image-level and region-level features are encoded as visual tokens $\mathcal{V}$, and fed into the LLM-based decoder along with the users' text input. ASM employs a specific prompt design that allows the LLM decoder to handle both generative tasks and discriminative tasks using a unified architecture with shared parameters. We add soft prompt tokens (*i.e.*, $\mathcal{P}_g$ and $\mathcal{P}_d$) to indicate the desired tasks and use an "$\langle align \rangle$" token to perform image-text alignment at the LLM's output. $\langle bos \rangle$ denotes the beginning token of a sentence.

whereas the secondary objective aims to improve its discriminative and retrieval capabilities.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{dis}}, \tag{1}$$

where the generation loss $\mathcal{L}_{\text{gen}}$ is for the next token prediction, and is the same as the loss of GPT series [60, 69, 8, 61]. The discriminative loss $\mathcal{L}_{\text{dis}}$ is for tasks like region-text aligning/retrieval. The discriminative loss follows the contrastive loss of CLIP [67], where each region is treated as an image when calculating the loss.

4.2 Location-Aware Image Tokenizer

To achieve location-aware image tokenizing, we introduce a query-based image tokenizer that conditions its queries on location information, such as bounding boxes, masks, or points. As depicted in Fig. 4, we first encode the input image using the ViT-g/14 [26] backbone, resulting in image features $\mathcal{F} \in \mathbb{R}^{H \times W \times D}$, where H and W denote the spatial size of the feature maps and D denotes the feature dimension. Next, we use the RoIAlign [33] to extract the region features $\mathcal{R} \in \mathbb{R}^{H_r \times W_r \times D}$ from the image features $\mathcal{F}$, according to the given bounding box (or mask, point set). Here, H_r and W_r denote the spatial size of the RoI features. We then flatten the region features $\mathcal{R}$, use two fully-connection (FC) layers to project them into $\mathcal{Q}_r \in \mathbb{R}^{G \times D_q}$, which has the same shape as randomly initialized query tokens $\mathcal{Q}' \in \mathbb{R}^{G \times D_q}$. Here, G denotes the number of tokens in a query group[3], and D_q denotes the dimension of a query token. Subsequently, the $\mathcal{Q}_r$ of N bounding boxes and $\mathcal{Q}'$ are concatenated to form location-aware query tokens $\mathcal{Q} \in \mathbb{R}^{(N+1) G \times D_q}$. These location-aware query tokens $\mathcal{Q}$ are then passed through a transformer decoder with 12 blocks to extract output features. Finally, the output features are projected to match the feature dimension D_t of the LLM and are used as the soft prompt $\mathcal{V} \in \mathbb{R}^{(N+1) G \times D_t}$ for subsequent decoding processes. Particularly, when no location information is provided, the bounding box is assumed to cover the entire image. This method guarantees a consistent approach for both local region and whole image tokenization.

4.3 LLM-Based Decoder

To develop a unified LLM-based framework that can effectively handle both generation tasks and discriminative tasks, we utilize Husky-7B [55] as our foundation language model to handle various

[3] A query group represents a randomly initialized query or a set of query tokens conditioned by a bounding box.

vision-language tasks under the guidance of user instructions and learnable soft prompts that contain image-level and region-level visual information.

For generative tasks, the input sequence comprises three types of tokens, including (1) learnable generative task prompt $\mathcal{P}_g \in \mathbb{R}^{M \times D_t}$, which informs the model that it should perform a generative task. (2) location-aware image tokens $\mathcal{V}$ that contain the extracted image-level and region-level information from the input image and (3) user prompt that expresses his/her requirements. Given such an input sequence, the LLM generates text tokens sequentially in an autoregressive manner until an end token $\langle eos \rangle$ is reached. An example prompt is provided below:

Prompt #1: "$\{\mathcal{P}_g\}$ $\langle bos \rangle$ **Human:** $\{\mathcal{V}\}$ What is this? **Assistant:**",

where the token number of task prompt M is set to 5. $\langle bos \rangle$ represents the beginning of the sentence.

For discriminative tasks, different from the CLIP-based framework that directly aligns the output feature from vision and language encoders, we introduce a trainable align token $\langle align \rangle$ to extract the holistic representation of the current input sequence. An example prompt for encoding input image is shown as follows:

Prompt #2: "$\{\mathcal{P}_d\}$ $\langle bos \rangle$ **Human:** $\{\mathcal{V}\}$ What is this? $\langle align \rangle$",

where $\mathcal{P}_d \in \mathbb{R}^{M \times D_t}$ represents the learnable task prompt used for discriminative tasks.

Similarly, the input sequence of input text consists of soft prompt tokens that indicate task information, as well as text tokens that represent the corresponding region caption or object class name. We omit the vision tokens to avoid information leakage. Here is an example prompt:

Prompt #3: "$\{\mathcal{P}_d\}$ $\langle bos \rangle$ **Assistant:** A photo of the Sphinx. $\langle align \rangle$".

During the process of region-text matching, we can achieve image-text retrieval by simply computing the similarity of the embedding of the $\langle align \rangle$ token. It is notable that the learnable task prompt and align tokens used in Prompt #2 and #3 are shared, while the task prompt differs between generative tasks (Prompt #1) and discriminative tasks (Prompt #2 and #3).

Compared to the CLIP-based framework, our LLM-based decoder offers two advantages: (1) Our approach builds upon LLMs, allowing us to leverage the powerful world knowledge and reasoning capability of LLMs. (2) Both the image and text aligning embedding of our method are generated by an LLM, which bridges the gap between the pre-training task for the language model and the language-image pre-training task.

5 Data Analysis

We conduct an in-depth analysis of our AS-1B dataset. We begin by showcasing the abundance of data in terms of quantity. Next, we explore the data diversity and open-world semantics captured in AS-1B. Finally, we thoroughly analyze the data quality of the initial automatic annotation pipeline and explain how we have improved it through data engineering and human feedback.

5.1 Data Scale

Statistics. The AS-1B dataset consists of a vast collection of 1.2 billion region-text pairs extracted from 11 million images, encompassing 3.5 million distinct semantic tags. Regions in the dataset are categorized into five different resolution scales: tiny, small, medium, large, xlarge, and huge. As indicated in Table 2, the distribution of region resolutions follows a roughly normal distribution. Over half of the regions are on the medium or large scale. In Sec. 3.2, we utilize several region proposal generators, including SAM [42], InternImage [88], EVA-02 [25], and GLIP [51], to generate region proposals for the AS-1B dataset. Table 3 presents the proportion of regions provided by each model in the 1.2 billion regions. SAM generates 36.4% of the regions, while the other three models contribute to 63.6% of the regions. Therefore, although our dataset shares images with SA-1B [42] and has a similar number of regions, the actual regions are different due to the use of diverse region proposal generators.

Region Type	Area Range	Proportion	(V)LLMs	BLIP	InternImage	EVA-02	GLIP
Tiny	$< 20^2$	4.2%	33.8%	16.5%	24.6%	25.1%	0.0%
Small	$20^2 \sim 40^2$	8.7%	34.5%	14.3%	24.6%	25.9%	0.7%
Medium	$40^2 \sim 100^2$	35.8%	55.6%	22.9%	8.3%	11.6%	1.7%
Large	$100^2 \sim 200^2$	23.7%	58.5%	26.2%	5.0%	7.9%	2.3%
Xlarge	$200^2 \sim 500^2$	18.3%	62.6%	27.1%	3.0%	4.3%	3.0%
Huge	$> 500^2$	9.5%	69.7%	24.9%	1.6%	1.2%	2.7%
All	–	100%	55.4%	24.0%	8.2%	10.4%	2.1%

Table 2: **Region statistics and semantic sources**. The percentage of semantic tags generated by different models at each resolution are reported. LLM/VLLMs [17, 111, 48] contribute significantly to the semantic diversity of our dataset.

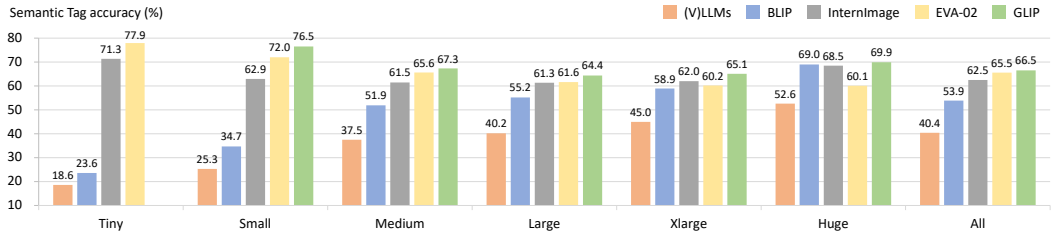


Figure 5: **The accuracy of semantic tags from different sources**. LLM/VLLMs [17, 111, 48] show lower accuracy than other models, especially on low resolution regions.

Each region is also annotated with detailed question-answer pairs and a caption, which yields a total of 3.3 billion visual question-answering pairs and 1.2 billion detailed region captions. As seen in Table 4, the average token number of the answers is 16.91, while the average token number of the composited caption is 34.84. The total number of tokens in our detailed region captions amounts to approximately 42.2 billion. This extensive collection of detailed captions provides valuable textual descriptions of regions within the images.

Comparisons. When comparing the AS-1B dataset with popular datasets containing region-level annotations, AS-1B stands out with a significantly larger number of regions. It has about 33 times more regions than the current largest detection dataset, BigDetection [10]. While AS-1B has fewer images compared to close-set classification datasets [23] or vision-language datasets [72], it compensates with valuable region annotations. Additionally, AS-1B offers an abundant collection of detailed region annotations. Compared to the largest region-level dataset, Visual Genome [43], AS-1B's detailed region annotation is about 1941 times larger than Visual Genome's 1.7 million pairs of VQA annotations and 222 times larger than its 5.4 million region captions.

5.2 Data Diversity

Statistics. A distinctive feature of AS-1B is its vast inclusion of open-world concepts, demonstrated through two key aspects: 1) a large number of semantic tags and 2) long and informative detailed descriptions. Fig. 6 visually demonstrates the wide range of open-world concepts present in AS-1B. The dataset covers diverse categories, including fine-grained categories like "lynx", proper nouns such as "The Sphinxs", object parts like "charging cords", and attributes like "pink and white baby cribs". In Fig. 2, we display the frequency distribution of semantic tags, revealing a clear long-tail pattern. The most frequent semantic tags predominantly represent broad category names, while less frequent tags correspond to fine-grained category names or instances with specific attributes.

In Table 2, we analyze the sources of each semantic tag to understand how open-world concepts are enriched. We report the proportion of sources for the top-1 semantics in the semantic tags at different scales. The results reveal that 55% of the top-1 semantic candidates are from the LLM, while 24% originate from the BLIP (the "magnifier" in Sec. 3.3). Interestingly, only 19% of the top-1 candidates are generated from the closed-set detectors, InternImage, and EVA-02. This highlights

Model	SAM	InternImage	EVA-02	GLIP
Proportion	36.4%	20.5%	22.5%	20.6%

Table 3: **The proportion of region proposals generated by different models**. Only 40% regions are generated from SAM.

Type	Number	#Tokens	Average Tokens
Question	3.3B	34.6B	10.50
Answer	3.3B	55.4B	16.91
Caption	1.2B	42.2B	34.84

Table 4: **The statistics of detailed description in AS-1B dataset**. The overall number of tokens reaches 132.2 billion.

Type	Correct answer	Wrong answer	Invalid question	Wrong semantic
Proportion	47.1%	18.6%	19.0%	15.3%

Table 5: **The statistics of attribute question-answering**. The answers generated by the "responder" had an accuracy of 47.1%. Wrong semantic denotes that the semantic tags are incorrect.

that the majority of concepts in the AS-1B dataset are obtained from open-world sources, especially the LLMs and VLLMs.

As for the detailed region caption, the VQA-based generation approach in AS-1B has proven advantageous, resulting in longer and more informative region descriptions. A more straight-forward way is to directly ask the VLLM to generate region captions. However, without guidance from semantic tags and questions, the model tends to output inaccurate information or hallucinations.

Comparisons. Instead of using fixed labels from a pre-defined set, the AS-1B dataset employs flexible and open-world semantic tags to label each region. Table 1 highlights that AS-1B contains a significantly larger number of semantic tags and concepts compared to close-set classification datasets or object detection datasets. For example, the number of semantic tags in AS-1B is approximately 159 times greater than the widely-used classification dataset ImageNet-22k [23], and it is 268 times larger than the category number in V3Det [86].

5.3 Data Quality

The Accuracy of Automatic Annotations. We evaluated the data quality using two metrics: *top-1 accuracy* and *semantic tag accuracy*. Top-1 accuracy refers to the probability that the top-1 candidates are correct, as selected by the human annotators. On the other hand, semantic tag accuracy denotes the probability the generated semantic tags are selected by the annotators. In the verified annotations, we obtained a top-1 accuracy of 54.8% and a candidate accuracy of 47.0%.

As shown in Figure 5, we find that different models in the annotation pipeline exhibit complementary behavior. The LLM and BLIP models show lower accuracy for small regions as they are not robust for the cropped low-resolution images. In contrast, close-set detectors perform better on these small regions, providing more accurate semantic candidates. For larger regions, LLMs and VLLMs become more accurate. Hence, the inclusion of close-set detectors can provide a trade-off between data quality and open-world semantics. This interplay of models contributes to the overall improvement of data quality in AS-1B.

As discussed in Sec. 3.5, the detailed region descriptions are also verified by human experts using a similar procedure. The human annotators are tasked with classifying the VQA pairs into four situations: 1) the question is proper, and the answer is correct; 2) the answer is incorrect; 3) the generated question is unanswerable given the image (*e.g.*, the production date of a car); 4) the semantic tag is wrong. As shown in Table 5, the accuracy of question-answer pairs is 47.1%.

Consumption Analysis. Here we focus on the consumption and efficiency of human verification in the context of the semi-automatic data engine we constructed. This approach significantly reduces the human labor required for data refinement compared with annotating all the data by humans. For verifying semantic tags, it takes approximately 10 seconds for one annotator to complete one region. Verifying every 1 million regions would take about 2,750 working hours. Considering a group of 50 annotators in our case, the entire verification process takes approximately 15 days. If we were to

annotate all regions, the annotation consumption would become 1,000 times larger, approximately 42 years. Such a large-scale human annotation effort would be unaffordable.

Moreover, for detailed captions with longer texts, the verification process would take even longer, *e.g.*, 15 seconds for each VQA annotation. Therefore, for large-scale annotation involving billions of regions in our case, utilizing models to annotate data at scale and correcting the models' bias with limited human annotation proves to be both feasible and efficient.

6 Experiments

We analyze and compare the proposed ASM with a CLIP-based baseline model and leading Multi-modality Large Language models (VLLMs) on representative vision tasks including zero-shot region recognition, image-level caption and region-level caption. Additionally, since using conventional image captioning metrics to evaluate LLM-based models can be limiting [107], we also perform human subject evaluation to compare our model with existing powerful VLLMs [111, 54].

6.1 Implementation Details

Training Setting. The training of the All-Seeing Model (ASM) involves three types of labels obtained from the AS-1B dataset, including region-level semantic tags, question-answer pairs, and detailed captions. The semantic tags are used for aligning regions with corresponding text, while the other annotations are used to train the text generation task. In addition, we also include LaionCOCO [71] in our training process, since the image-level caption data from LaionCOCO is beneficial for ASM's ability to comprehend the whole images.

We adopt a multi-task training approach that combines text generation and region-text alignment tasks to train our ASM. The batch size for text generation is set to 256, while for region text alignment it is set to 32,768. We employ the AdamW optimizer [57] with the β_1 of 0.9, the β_2 of 0.999, and the weight decay of 0. During training, the learning rate is initialized as 5×10^{-4} and includes a linear warmup that lasts until the first 10% of training steps. The warmup is followed by a cosine decay strategy with a minimum learning rate of 0. Unless otherwise specified, the image resolution for all experiments is set to 224×224. We initialize the model parameters using Husky [55] and train the model for one epoch. In addition, we also provide a second-stage fine-tuning setting to further improve the effectiveness of ASM. Specifically, we utilize high-quality multi-modal data MiniGPT-4 [111], LLaVA-150k [54], and COCO caption dataset [15] as image-level text generation data, along with VG [43] and RefCOCOg [59] datasets as region-level text data. Human-verified region annotations are also included. During fine-tuning, we set the learning rate to 5×10^{-5} and apply a weight decay of 0. The other settings remain the same as during pre-training. The fine-tuned ASM is denoted as ASM-FT.

Baseline Model. To make comparison with recent popular multi-modality large language models (VLLMs) [111, 54, 47] that only focus on processing the entire image, we crop a region from the image and input it to these model for region-level visual recognition and understanding. However, this cropping may result in the loss of some contextual information from the entire image. For better comparison, we implement a simple region-text contrastive model based on CLIP [67] as a baseline. The baseline model, named Region-Aware CLIP (R-CLIP), is equipped with an RoIAlign layer [33] on the feature maps obtained from the vision encoder in the CLIP model. To initialize the model weights, we leverage OpenCLIP [36] (ViT-L/14) and then train the CLIP model on our AS-1B dataset. The model is trained for $10,000$ steps with a batch size of 32,768. Other training settings is the same as those of ASM.

6.2 Text Generation

Evaluation Setting. We evaluate the image-level caption ability of our model on Flickr30K [97] and NoCaps [1] dataset. We report the CIDEr [85] and SPICE [3] metric on these benchmarks. To assess the region-level caption ability, we also evaluate ASM on the Visual Genome [43] and RefCOCOg [59]. On the region caption task, we adopt both the Meteor [5] and CIDEr [85] metric as our evaluation metrics. The Meteor, CIDEr, and SPICE metrics are computed by COCOEvalCap[4].

[4]https://github.com/salaniz/pycocoevalcap

Figure 6: **Examples of the semantic tags**. Benefitting from the world knowledge of LLMs/VLLMs, the AS-1B dataset covers diversity semantic tags in the real world.

Results. For region-level captioning, as shown in Table 6, our ASM model surpasses the concurrent region-aware VLLMs, Kosmos-2 [65], by 1.4 points on the RefCOCOg dataset, under the zero-shot setting. After the second-stage fine-tuning, our ASM model has achieved a new record for referring expression generation on RefCOCOg. Besides, on the Visual Genome (VG) dataset, although the

Figure 7: **Examples of the detailed region annotations**. Visual question-answering pairs and captions are provided based on the semantic tags. Failure cases are marked in red.

Meteor score of zero-shot ASM is inferior to GRiT [92], ASM-FT achieves significantly better results than GRiT given relevant data.

In addition, our model also excels at image-level captioning, as presented in Table 7, our ASM model demonstrates promising zero-shot performance on Flickr30K [97] and NoCaps [1] dataset. Specifically, under the zero-shot setting, our model achieves a CIDEr score of 77.9 without the second-stage fine-tuning and 87.7 after the second-stage fine-tuning, which outperforms all the concurrent VLLMs, such as InstructBLIP [22], Shikra-13B [13] and Kosmos-2 [65]. Furthermore, on the NoCaps dataset, ASM also achieves comparable performance compared to the baselines under the zero-shot setting. These results indicate that our ASM model retains a strong image-level comprehension ability while also being region-aware.

In summary, these results highlight the strong region-level text generation capabilities of our model, while also showcasing its ability to comprehend the entire image. The promising zero-shot performance of ASM further demonstrates the effectiveness of our proposed AS-1B dataset. Moreover, the unified model structure of ASM enables it to effectively utilize diverse data sources during training, enhancing its overall performance.

Model	Zero-shot	Visual Genome		RefCOCOg	
		Meteor	CIDEr	Meteor	CIDEr
GRiT [92]	✗	17.1	142.0	15.2	71.6
SLR [99]	✗	-	-	15.4	59.2
SLR+Rerank [99]	✗	-	-	15.9	66.2
Kosmos-2 (Few-shot,k=2) [65]	✗	-	-	13.8	62.2
Kosmos-2 (Few-shot,k=4) [65]	✗	-	-	14.1	62.3
Kosmos-2 [65]	✓	-	-	12.2	60.3
ASM	✓	12.6	44.2	13.6	41.9
ASM-FT	✗	**18.0**	**145.1**	**20.8**	**103.0**

Table 6: **Performance on the region-level captioning task.** "-FT" denotes ASM with second-stage fine-tuning.

Model	Zero-shot	Flickr30k		NoCap	
		CIDEr	SPICE	CIDEr	SPICE
MetaVLM [32]	✓	43.4	11.7	-	-
VinVL [103]	✓	-	-	95.5	13.5
LEMON [34]	✓	-	-	106.8	14.1
Flamingo-3B [2]	✓	60.6	-	-	-
Flamingo-9B [2]	✓	61.5	-	-	-
SimVLM [91]	✓	-	-	110.3	14.5
CoCa [98]	✓	-	-	120.6	15.5
BLIP [48]	✓	-	-	113.2	14.7
BLIP-2 [47]	✓	-	-	121.6	**15.8**
InstructBLIP [22]	✓	82.8	-	**123.1**	-
Shikra-13B [13]	✓	73.9	-	-	-
Kosmos-1 [35]	✓	67.1	14.5	-	-
Kosmos-2 [65]	✓	66.7	-	-	-
ASM (ours)	✓	77.9	17.3	104.8	14.5
ASM-FT (ours)	✓	**87.7**	**18.7**	117.2	15.6

Table 7: **Zero-shot performance on the image-level captioning tasks.** Our ASM shows comparable or even better performance than models dedicated to image-level captioning.

6.3 Zero-shot Region Recognition

Evaluation Setting. We use zero-shot region recognition to evaluate the region-text alignment ability of our model. We use COCO [53] and LVIS [31] detection dataset for evaluation. Since our current focus is not on object localization, we use the ground-truth boxes and use the model to predict the categories given the corresponding texts following RegionCLIP [108]. We report the mean Average Precision (mAP) metrics for this evaluation.

Results. As shown in Table 8, both our baseline model R-CLIP and the proposed ASM achieve promising zero-shot region recognition performance. On the COCO dataset, R-CLIP outperforms the original CLIP by 9.7 mAP, and ASM further increases the performance by 10.4 mAP. On the more challenging LVIS dataset with 1,203 categories, R-CLIP outperforms CLIP by 7.7 mAP, and ASM achieves a more significant improvement of 14.3 mAP over CLIP. These results demonstrate the effectiveness of region-text data in AS-1B dataset and the proposed ASM in region-text alignment tasks. Notably, our ASM simultaneously performs caption and region recognition tasks with the same weight, showcasing its versatility and efficiency.

These results demonstrate that, despite the semantic tags in AS-1B contain some noise, we can still fine-tune a robust region-aware CLIP model with minor modifications. The result suggests

Model	COCO				LVIS			
	mAP	AP_S	AP_M	AP_L	mAP	AP_S	AP_M	AP_L
CLIP [67]	58.9	50.7	70.4	58.3	47.1	40.3	59.2	57.4
OpenCLIP [36]	63.3	47.8	75.6	60.9	49.1	37.4	62.8	66.5
R-CLIP (our baseline)	68.6	61.4	75.4	79.3	54.8	49.3	60.6	66.6
ASM (ours)	**69.3**	**64.3**	**78.0**	**71.0**	**61.4**	**56.7**	**67.9**	**69.2**

Table 8: **Zero-Shot object recognition performance.** We report the zero-shot recognition accuracy on COCO and LVIS dataset. The ground-truth boxes are used for inference.

Data Scale	COCO	LVIS
1M	67.8	54.0
2M	67.5	**55.0**
5M	**68.6**	54.8

Table 9: **Zero-shot object recognition performance (mAP)** with different training data scale.

Data Cleaning	COCO	LVIS
✗	61.8	46.5
✓	**67.8**	**54.0**

Table 10: **Zero-shot object recognition performance (mAP)** with and without data cleaning.

Human Data	Input Scale	COCO	LVIS
✗	224	67.8	54.8
✓	224	**70.2**	**55.0**
✗	896	76.7	65.7
✓	896	**80.0**	**68.4**

Table 11: **Zero-shot object recognition performance (mAP)** with and without fine-tuning on human-verified annotations.

that region-text data in AS-1B dataset can be beneficial in enabling the model to learn region-text alignment by considering both the region itself and its context.

6.4 Data Engineering

Here, we use quantitative results to show the impact of data quantity and data engineering. Considering the cost of the experiment, we use our baseline model R-CLIP. We use the Zero-shot object recognition metrics as in Sec. 6.3 to inspect the impact of data engineering, *i.e.*, we use the ground-truth boxes and use R-CLIP to determine the categories following RegionCLIP [108]. Unless otherwise specified, we train the model with semantic tags from 1M images in the AS-1B dataset.

Data Scale up. We find that scaling up the semantic tags can be helpful for zero-shot region recognition. To verify this, we train our baseline R-CLIP with different amounts of semantic tags. As shown in Table 9, with more training data (from 1M to 5M images), the R-CLIP model attains higher Zero-shot object recognition performance.

Data Cleaning. Data cleaning and post-processing are important. In practice, the original data annotation pipeline outputs a total of 2.14 billion regions. We devise a simple data cleaning strategy: (1) we sample the top 100 regions with the highest CLIP score at different scales from each image in the AS-1B dataset and (2) we further re-rank the semantic candidates with CLIPSeg [58], as discussed in Sec. 3.4. This data cleaning process will compress the original 2.14B regions into 1.2B regions. As shown in Table 10, adding data cleaning can significantly improve the mAP by 6.0% and 7.5% on COCO and LVIS datasets.

How human verification improves the model? An important part of our data engine is to improve the model with human feedback. In this way, the improved model can be used to refine the initial data which is automatically generated. In this section, we investigate the effectiveness of human verification process. We fine-tune the trained R-CLIP model with human-verified region annotations, and find that a small number of human labels can significantly boost the model performance.

Specifically, to make the most of human labels, we utilized both the positive and negative candidates marked by the human annotators. When calculating the contrastive loss, for each region, we randomly selected one positive candidate and use all the unselected candidates as negative samples. Compared with the image-to-text part in the original CLIP-style contrastive loss, each region will be compared with more negative text samples. The unselected candidates can be viewed as valuable hard samples, indicating when the model will make mistakes.

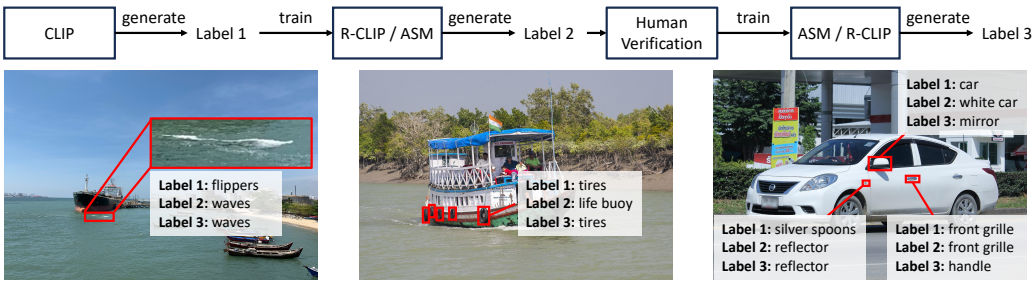


Figure 8: **Visualization of the data iteration process**. The iteration process improves the label accuracy. We visualize three types of models: (1) **Label 1**: labels produced the original CLIP; (2) **Label 2**: labels produced by R-CLIP or ASM, trained with **Label 1** as input data; (3) **Label 3**: labels produced by R-CLIP or ASM which is further tuned with human verification data.

In practice, we use a batch size of 1024 and a learning rate of 5e-4 to fine-tune the pre-trained model on the human data for four epochs with only 40k human verified semantic tags. Table 11 shows that fine-tuning the model with human data can yield significant performance gain: +2.4 and +3.3 COCO mAP on the resolution of 224 and 896. This demonstrates that a small amount of human data can correct the model's bias and hard cases thus improving performance. The effectiveness of human verification lays the foundation for data quality improvement in the data engine iterations. To intuitively show the data quality improvements, we show the coarse labeling results for CLIP as well as the output of R-CLIP output before and after the human data fine-tuning in Fig. 8. The original CLIP is unreliable at lower resolutions, *e.g.*, the reflectors and handles on the white cars are categorized into wrong classes. R-CLIP pre-trained on AS-1B data performs better in these low-resolution areas. However, it may fail to recognize some objects due to noisy labels, *e.g.*, labeling the tires hung by the boat as a "life buoy". The human data fine-tuning process can correct the pre-trained R-CLIP.

6.5 Human Evaluation

As discussed in ChatCaptioner [110], using conventional image caption metrics such as Meteor [5] and CIDEr [85] may not reliably evaluate relatively lengthy texts generated from LLM-based models. To better assess the text generation ability from a human perspective, we conducted a user study.

Evaluation Setting. In our user study, we involve a total of 5 participants to evaluate the performance of the All-Seeing Model (ASM) along with two other powerful VLLMs: MiniGPT4 [111], and LLaVA [54]. We evaluate image and region-level captioning. For the evaluation, we randomly select 20 samples from each of the Visual Genome, RefCOCOg, COCO, and Flickr30K datasets. Participants are asked to choose the most informative captions without any factual errors or hallucination. Aside from model outputs, we also add the ground truth captions as options, which can be viewed as human outputs.

Results. The human evaluation results in Table 12 indicate that captions generated by our ASM are preferred over those from MiniGPT4 and LLaVA. While LLaVA and MiniGPT4 may produce longer captions for region-level tasks (VG and RefCOCOg), they often introduce over-association, hallucinations, and factual errors. In contrast, ASM generates captions with moderate length and more accurate information. On RefCOCOg, Flickr30K, and NoCaps datasets, ASM even outperforms human annotations with longer and more detailed captions. This is because human annotators tend to write short captions, while users prefer longer, detailed captions generated by ASM, which also contain fewer factual errors. For image-level generation tasks, ASM produces captions with similar length to those from MiniGPT4 and LLaVA but is more frequently favored by users.

The results clearly demonstrate the effectiveness of ASM and the AS-2B dataset. The VQA-based annotation pipeline provides region-specific information with less irrelevant content, reducing the occurrence of hallucinations. Moreover, human verification further enhances the data quality, leading to significantly better performance on region-level tasks.

Model	Visual Genome		RefCOCOg		Flickr30K		NoCaps	
	Rate	Length	Rate	Length	Rate	Length	Rate	Length
Human	47.8	13.6	10.3	6.3	30.0	16.0	27.3	15.1
LLaVA [54]	4.3	110.8	15.4	100.6	17.5	114.0	9.1	108.4
MiniGPT4 [111]	8.7	110.9	15.4	113.5	14.2	114.6	13.6	101.0
ASM (ours)	39.2	37.5	46.1	33.6	38.3	112.4	50.0	102.1

Table 12: **Human evaluation results on caption tasks.** We ask the users to select the caption that contains the most information regarding the image/region while does not producing any factual errors.

7 Conclusion

In this paper, we present the All-Seeing (AS) Project, which develops a comprehensive system for panoptic visual recognition and understanding in the open world from both dataset and model perspectives. In terms of data, we elaborate a semi-automatic data engine consisting of an automatic annotation pipeline and a human verification step. Using this data engine, we annotated the AS-1B dataset comprising over 1 billion region-level comprehensive annotations, with controllable costs. From the model aspect, we propose a region-aware multi-modal large language model called the All-Seeing Model (ASM). The ASM utilizes a unified LLM decoder to model both region-text alignment and image-conditioned text generative tasks. Leveraging the AS-1B dataset and other high-quality data, ASM achieves state-of-the-art results on image and region-level tasks. We also fine-tune a region-aware CLIP model exclusively on the AS-1B dataset, surpassing the original CLIP by significant margins in region recognition. We believe that the data engine, AS-1B dataset, and the ASM model proposed in the All-Seeing Project will inspire further research and development towards empowering artificial intelligence systems with an "all-seeing eye," enabling them to achieve a deeper understanding of the world.

Credit Attribution of Equal Contribution Authors

Weiyun Wang is responsible for the implementation and experiments of ASM, constructing the detailed annotation data pipeline, optimizing the reasoning efficiency of LLM/VLLM-related annotation modules, refactoring the code of R-CLIP and improving its efficiency, implementing the code of the open-world semantic generation, and drafting the corresponding method and experiment sections.

Min Shi is responsible for managing the construction of the data engine, joint debugging the data engine, constructing the semantic tag annotation pipeline, designing data cleaning and conducting data-related analysis and experiments, implementing part of the R-CLIP's code and main experiments of R-CLIP, implementing the open-world semantic matching, participating in the human verification process, drafting partial manuscripts and revising the manuscript.

Qingyun Li is responsible for the main part of the open-world localization, optimizing partial localization models, implementing the main code of the R-CLIP, some refining experiments, setting up the human evaluation platform for ASM, and drafting partial manuscripts.

Wenhai Wang is the technical manager of the AS project, responsible for the task decomposition, prototyping, and optimization suggestions of each part of the project, and drafted and revised the entire manuscript.

Zhenghang Huang is responsible for the main part of the human verification process, setting up the human verification platform, implementing part of the location annotators, communicating and guiding the manual annotation team, and drafting partial manuscripts.

Linjie Xing is responsible for optimizing most of the location annotator and part of the semantic generation modules, implementing part of the location annotators, reviewing part of the human verification results, and drafting partial manuscripts.

Special acknowledgment to Xizhou Zhu and Hao Li for the preliminary idea and verification of the AS project.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Int. Conf. Comput. Vis.*, 2019. 15, 17

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inform. Process. Syst.*, 2022. 3, 5, 18

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Eur. Conf. Comput. Vis.*, 2016. 15

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Int. Conf. Comput. Vis.*, 2015. 5

[5] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics, 2005. 15, 20

[6] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inform. Process. Syst.*, 2022. 5

[7] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. Vl-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127*, 2022. 5

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.*, 2020. 3, 5, 7, 11

[9] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 4, 5

[10] Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. Bigdetection: A large-scale benchmark for improved object detector pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 4, 6, 13

[11] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 4, 5

[12] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 5

[13] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 5, 17, 18

[14] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 5

[15] 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. 4, 15

[16] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *Int. Conf. Learn. Represent.*, 2023. 5

[17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 1, 2, 3, 6, 7, 8, 13

[18] EasyOCR contributors. Easyocr. `https://github.com/JaidedAI/EasyOCR`, 2023. 7

[19] MOSS contributors. Moss. `https://github.com/OpenLMLab/MOSS`, 2023. 1, 3, 7

[20] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 4

[21] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023. 3

[22] Wenliang Dai, Junnan Li, Dongxu Li, AnthonyMeng Huat, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1, 3, 17, 18

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 3, 4, 5, 13, 14

[24] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88:303–338, 2010. 4

[25] 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. 2, 6, 12

[26] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 5, 11

[27] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 5

[28] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 3

[29] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022. 3

[30] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 5

[31] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3, 4, 6, 18

[32] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 18

[33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, 2017. 5, 11, 15

[34] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 18

[35] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 3, 18

[36] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. July 2021. 15, 19

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning.*, 2021. 5

[38] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. Real-time analysis and visualization of the yfcc100m dataset. In *Proceedings of the 2015 workshop on community-organized multimodal mining: opportunities for novel solutions*, pages 25–30, 2015. 4

[39] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Int. Conf. Comput. Vis.*, 2021. 5

[40] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 5

[41] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 5

[42] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4, 5, 6, 12

[43] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123:32–73, 2017. 2, 4, 5, 13, 15

[44] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[45] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.*, 128(7):1956–1981, 2020. 4

[46] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 5

[47] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 3, 5, 15, 18

[48] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning.*, 2022. 1, 2, 3, 5, 6, 7, 8, 13, 18

[49] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inform. Process. Syst.*, 2021. 5

[50] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 5

[51] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 6, 12

[52] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 5

[53] 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 *Eur. Conf. Comput. Vis.*, 2014. 3, 4, 5, 18

[54] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 3, 5, 7, 15, 20, 21

[55] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 1, 2, 5, 6, 7, 8, 11, 15

[56] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023. 3

[57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2019. 15

[58] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 9, 19

[59] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 15

[60] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023. 5, 11

[61] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3, 5, 7, 11

[62] TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022. 1, 3, 5

[63] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Adv. Neural Inform. Process. Syst.*, 2011. 4, 5

[64] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Adv. Neural Inform. Process. Syst.*, 2022. 1

[65] 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. 3, 5, 16, 17, 18

[66] Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Visual semantic complex network for web images. In *Int. Conf. Comput. Vis.*, 2013. 8

[67] 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.*, 2021. 3, 5, 6, 9, 11, 15, 19

[68] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018. 5

[69] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI*, 2019. 3, 5, 11

[70] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[71] Christoph Schuhman, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en. 2022. 3, 15

[72] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Adv. Neural Inform. Process. Syst.*, 2022. 2, 4, 5, 13

[73] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[74] 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 *Int. Conf. Comput. Vis.*, 2019. 4

[75] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Ann. Meeting of the Assoc. for Comput. Linguistics*, 2018. 5

[76] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 5

[77] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 2023. 1, 3, 7

[78] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. 2023. 1, 3, 7

[79] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5

[80] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 3, 7

[81] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 1

[82] Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018. 3

[83] Daniel A Updegrove, Sheldon B Smith, and Wendy Rickard Bollentin. Ccnews: An online forum for newsletter editors. In *Proceedings of the 16th annual ACM SIGUCCS conference on user services*, 1988. 3

[84] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[85] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 15, 20

[86] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. *arXiv preprint arXiv:2304.03752*, 2023. 4, 14

[87] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 3, 5, 7

[88] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Int. Conf. Comput. Vis.*, 2023. 2, 6, 12

[89] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 5

[90] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *Ann. Meeting of the Assoc. for Comput. Linguistics*, 2022. 3

[91] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *Int. Conf. Learn. Represent.*, 2022. 5, 18

[92] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 3, 17, 18

[93] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 5

[94] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 5

[95] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *Int. Conf. Learn. Represent.*, 2022. 5

[96] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 5, 7

[97] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 15, 17

[98] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022. 3, 5, 18

[99] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3, 18

[100] Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68, 2021. 3

[101] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017. 5

[102] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *Int. Conf. Learn. Represent.*, 2022. 3, 7

[103] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021. 18

[104] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 3, 5

[105] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 5

[106] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

[107] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*, 2023. 5, 15

[108] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 18, 19

[109] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *Int. J. Comput. Vis.*, 127:302–321, 2019. 4

[110] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023. 5, 20

[111] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 3, 5, 6, 7, 13, 15, 20, 21

[112] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023. 5

[113] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Int. Conf. Learn. Represent.*, 2021. 5

[114] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 5

[115] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 5

[116] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Int. Conf. Comput. Vis.*, 2015. 3