Towards Open-Ended Visual Recognition with Large Language Models

Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen

ByteDance

Abstract. Localizing and recognizing objects in the open-ended physical world poses a long-standing challenge within the domain of machine perception. Recent methods have endeavored to address the issue by employing a class-agnostic mask (or box) proposal model, complemented by an open-vocabulary classifier (*e.g.*, CLIP) using pre-extracted text embeddings. However, it is worth noting that these open-vocabulary recognition models still exhibit limitations in practical applications. On one hand, they rely on the provision of class names during testing, where the recognition performance heavily depends on this predefined set of semantic classes by users. On the other hand, when training with multiple datasets, human intervention is required to alleviate the label definition conflict between them. In this paper, we introduce the OmniScient Model (OSM), a novel Large Language Model (LLM) based mask classifier, as a straightforward and effective solution to the aforementioned challenges. Specifically, OSM predicts class labels in a generative manner, thus removing the supply of class names during both training and testing. It also enables cross-dataset training without any human interference, exhibiting robust generalization capabilities due to the world knowledge acquired from the LLM. By combining OSM with an off-the-shelf mask proposal model, we present promising results on various benchmarks, and demonstrate its effectiveness in handling novel concepts. Code and models are available at `https://github.com/bytedance/OmniScient-Model`.

1 Introduction

A persistent challenge in the realm of machine perception involves the accurate localization [5, 6, 24, 28, 49] and recognition [17, 29, 37] of objects in real-world settings. While considerable progress has been made on various standard benchmarks [32, 55, 67, 77], existing methods continue to grapple with the complexities of real-life scenarios where novel out-of-training dataset concepts frequently arise. To address this issue and enhance the practical utility of models, a common strategy is to decompose the problem into two components: class-agnostic mask/box proposal and mask/box classification, as highlighted in previous works [3, 11, 65, 87]. It has been observed that mask/box proposal models, when trained on a dataset such as COCO [44], can still effectively generalize to previously unseen concepts [33, 84]. Additionally, recent advancements, exemplified by Segment Anything Model (SAM) [36], have expanded the training dataset

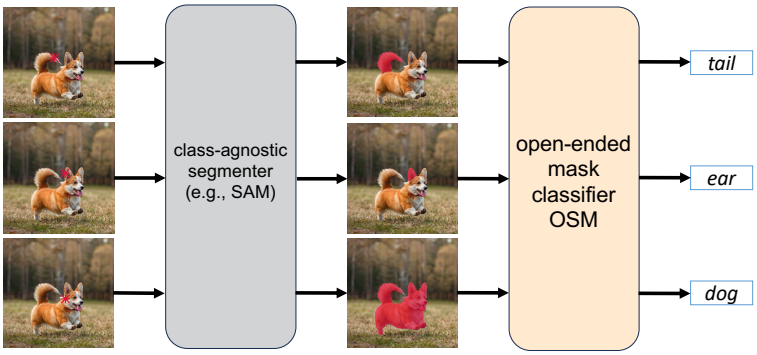

Fig. 1: Illustration of open-ended recognition. The open-ended recognition task is decomposed into two sub-tasks: class-agnostic mask proposal and open-ended mask classification. To tackle the task, we propose OSM (OmniScient Model), an open-ended recognition model that works hand in hand with an off-the-shelf class-agnostic mask proposal model (*e.g.*, SAM). Unlike existing open-vocabulary recognition models, OSM does not require any user-predefined vocabulary and instead directly predicts the class of each proposal with unconstrained vocabulary in a generative manner. As a result, OSM shows a great generalization ability. For example, we observe **emergent part predictions** such as tail and ear, while OSM has never seen such masks or labels during training (*i.e.*, we do not use any part segmentation datasets during training). Moreover, by obtaining masks from a class-agnostic segmenter, we can take advantage of it and take a wide range of prompt types including point, box, and mask.

to an extensive scale, encompassing 1.1 billion class-agnostic masks from 11 million images. This has yielded a mask proposal model characterized by robust zero-shot segmentation capabilities, generalizing to novel images and concepts. These developments present a promising avenue toward a solution to the first issue regarding class-agnostic object proposals.

Despite the remarkable achievements in the development of general proposal models, addressing the challenge of classifying novel concepts in real-world scenarios remains an unsolved issue. Many of the existing approaches leverage vision-language models (VLMs), such as CLIP [55] and ALIGN [32], which have been pretrained on extensive Internet datasets and have demonstrated outstanding performance in aligning images and text within a shared embedding space. Specifically, these techniques [19, 20, 23, 43, 69, 72–74, 78, 84] aim to train open-vocabulary classifiers that rely on the precomputed text embeddings derived from VLMs, as opposed to learning label embeddings directly from the training dataset. The dependency on VLM text embeddings highlights the inherent power and generalization capabilities of VLMs, which, to a certain extent, ensure the classifier's ability to generalize to novel concepts.

Nevertheless, it is important to acknowledge that while these methods have shown promise, they are still confronted with several challenges that impede their practical application. Firstly, these models typically operate under the assumption that class names are predefined during testing, a condition seldom met in

real-life scenarios. Furthermore, when utilizing multiple diverse datasets, complications arise when different label definitions or label space conflicts exist among them. Consequently, many current multi-dataset frameworks address this issue by training on each dataset with an individual decoder or classifier [26, 83, 85], or merge the label space manually [38], adding complexity to the process.

To address these challenges, we introduce the OmniScient Model (OSM), a novel generative framework towards open-ended recognition tasks. Instead of training the model to "select" correct classes from a predefined vocabulary, our approach focuses on training it to generate the desired class names. This paradigm shift means that the model no longer requires the prior knowledge of all possible class names provided by users, eliminating the necessity for a well-defined vocabulary during both training and testing phases. Consequently, this approach naturally accommodates the training and testing on datasets with varying label spaces, obviating the need for human intervention to harmonize the differences. Additionally, by building upon a pre-trained Large Language Model (LLM) [12, 63], OSM leverages the implicitly learned world knowledge [30, 59] encoded within the LLM, enhancing its ability to effectively generalize to novel concepts, further bolstering its utility and reliability.

We conduct meticulous experiments to validate the appropriateness of employing a generative model for discriminative tasks. Our investigation includes assessing the generative model's ability to effectively capture and adapt to the characteristics of a given training dataset and its associated vocabulary. We compare its performance to that of a discriminative model, primarily focusing on classification accuracy. Additionally, we introduce a Mode Query mechanism, which empowers the model to make predictions within a predefined vocabulary (referred to as vocabulary-specific predictions), or to provide open-ended predictions without vocabulary constraints (referred to as vocabulary-agnostic predictions). Finally, we integrate OSM with various off-the-shelf segmentors (*i.e.*, mask proposal models), such as kMaX-DeepLab [79] and SAM [36], and validate its effectiveness across several benchmarks.

2 Related Work

Open-Vocabulary Recognition Recently, exemplified by CLIP [55] and ALIGN [32], open-vocabulary recognition methods have demonstrated promising outcomes. These methods involve the pre-training of dual-encoder models (for image and text) using contrastive objectives on extensive collections of noisy image-text pairs. This pre-training process yields feature representations that possess cross-model capabilities, showcasing robust performance in zero-shot downstream tasks. Drawing inspiration from these advances, the field of open-vocabulary detection and segmentation [23, 51, 68, 78] has also witnessed remarkable breakthroughs, where class names provided during testing may not have been encountered during the training phase. A majority of these state-of-the-art techniques [20, 23, 43, 69, 78, 84] approach the problem by disentangling it into class-agnostic proposals, along with open-vocabulary proposal classification

by leveraging a pre-trained CLIP model. However, despite the impressive accomplishments of these open-vocabulary methods in recognizing unseen classes beyond the training dataset, they hinge on a strong yet brittle assumption that the semantic classes (*i.e.*, vocabulary) are known in advance and remain static, an assumption that can easily be disrupted in practical applications. In parallel with our research efforts, vocabulary-free image classification [14] and zero-guidance semantic segmentation [58] seek to address this challenge by dynamically generating vocabularies through processes such as parsing captions [39, 40] or retrieving them from external databases [60], and subsequently conducting predictions within this generated vocabulary in a discriminative manner. By contrast, our approach offers a straightforward solution by reformulating the open-ended classification problem as text generation [2, 18, 56, 61], naturally eliminating the need for a user-predefined vocabulary. Furthermore, while our method primarily focuses on object-level recognition, the works presented in [14] and [58] concentrates on image-level classification and semantic segmentation, respectively.

Large Language Models In recent years, research community has witnessed a remarkable surge in the development of Large Language Models (LLMs) [2, 52, 57, 62, 63]. These models have demonstrated impressive emergent capabilities, including in-context learning [2], instruction following [13, 70], and chain-of-thought reasoning [71]. However, a significant limitation of these LLMs is their inherent "blindness" to other modalities, such as visual inputs. More recently, the excitement surrounding multi-modal LLMs has surged, particularly with the introduction of GPT-4V [52]. Pioneering research [1, 4, 39, 45, 46, 66, 86] has illustrated a promising avenue for bridging the gap between language and vision modalities. This approach involves constructing modular models that typically consist of a frozen CLIP vision encoder, a trainable bridging module (*e.g.*, Perceiver Resampler in [1], Q-Former in [39], or a simple linear/MLP layer in [45, 46]), and a frozen LLM. Furthermore, [4, 53, 66, 75, 80, 81] add referring or grounding ability to the multi-modal LLM through taking bounding-box as inputs or outputs. The proposed OSM can be categorized as a modular multi-modal LLMs with referring capability. However, previous endeavors primarily aim to enhance multi-modal LLMs with bounding-box (as bounding-box can be naturally represented in text by referring to its coordinates) for conversation applications, which also require providing vocabulary in the input prompt [66, 80]. Our perspective underscores the value of enabling multi-modal LLMs to recognize *segmentation masks* and serve as standalone tools.

3 Method

In this section, we commence by detailing how we transform the conventional classification task into a text generation task, aligning with the principles outlined in [2, 56] (Sec. 3.1). Subsequently, we elucidate the construction of OSM (OmniScient Model), which follows the footsteps of previous modular vision-language models [16, 39, 46, 86] (Sec. 3.2). We also provide a comprehensive overview of our training and evaluation protocols (Sec. 3.3).

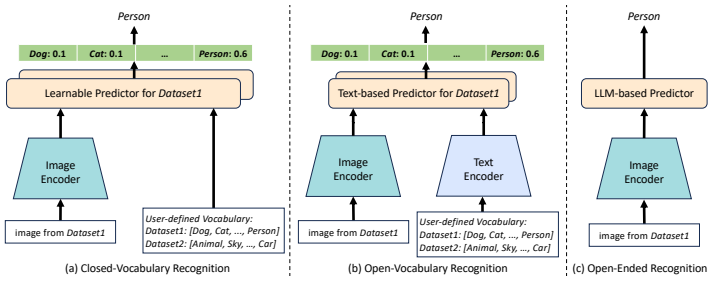

Fig. 2: Recognition scheme comparisons. (a) In the closed-vocabulary recognition setting, the sets of semantic classes are fixed during both training and testing. A learnable predictor (*e.g.*, 1×1 convolution layer) is used for each training dataset. (b) In the open-vocabulary recognition setting, the sets of semantic classes can be different during training and testing, allowing detection of novel concepts during testing by leveraging a pretrained CLIP backbone. The text-based predictor (*i.e.*, the text embeddings of the predefined set of semantic classes) is different for each dataset. (c) In the open-ended recognition setting, the model directly predicts the class names in a generative manner, removing the need to predefine the semantic classes during training and testing. Additionally, it enables the cross-dataset training in an easier way (*e.g.*, no need to involve humans to resolve the label definition conflicts between datasets).

3.1 Problem Formulation of Classification

Without loss of generality, we focus our discussion on mask classification. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a collection of M segmentation masks $\mathbf{M} \in \mathbb{R}^{H \times W \times M}$ (from an off-the-shelf segmenter, *e.g.*, SAM [36]), our objective is to predict a semantic class for each of these masks:

$$\{y_i\}_{i=1}^{M} = \{(m_i, c_i)\}_{i=1}^{M}, \tag{1}$$

where m_i is the i-th mask from $\mathbf{M}$ and c_i is its predicted class, belonging to the set of predefined semantic classes C, which is assumed to be known during both training and testing phases. In a closed-vocabulary setting, models focus solely on the target classes, implying that the set of predefined semantic classes are identical during both training and testing (*i.e.*, $C_{train} = C_{test}$, where the subscript denotes the training or testing phase). By contrast, in an open-vocabulary setting, this assumption is relaxed by allowing for the possibility that C_{test} may include novel categories that were not seen during training (*i.e.*, $C_{test} \neq C_{train}$). Nevertheless, in both cases, it is essential to have access to the category names of C_{train} and C_{test} during both the training and testing stages. As a result, the recognition performance heavily hinges on the careful design of C_{train} and C_{test} (called prompt engineering in the literature [23, 78]).

The aforementioned assumption (*i.e.*, the access to C_{train} and C_{test}) plays a pivotal role in contemporary recognition frameworks, whether operating in a closed-vocabulary or open-vocabulary context. These frameworks typically rely on computing similarity logits across semantic class candidates and selecting the

candidate with the highest probability as the final prediction. While these methods have demonstrated effectiveness and success across various tasks and benchmarks over the past decades, they are not without critical limitations. Firstly, it is practically impossible to predefine and encompass all potential semantic classes present in the real world. This limitation poses a significant challenge in open-vocabulary recognition since it necessitates the prior definition of novel concepts within the vocabulary. Furthermore, many of these methods are constructed around a handcrafted and meticulously designed label space, with the expectation of covering common concepts that should ideally have unambiguous definitions. However, the manual curation of label spaces may not be scalable, particularly when researchers aim to expand their models to encompass all available datasets from various sources. This process may require labor-intensive tasks such as meticulous manual merging [38] or conducting separate training [26, 83].

To address those challenges, we depart from the conventional approach in visual recognition and propose a novel paradigm named open-ended visual recognition. In this paradigm, we make the bold assumption that the vocabulary C remains **unknown** during both training and testing. We note that during training we only need to access the target class for each mask, without the need to know the existence of all the other possible classes in C, which is required for existing methods relying on softmax-based prediction. This shift in perspective is illustrated in Fig. 2 for a holistic comparison of the different paradigms. Rather than selecting a prediction class from a predefined vocabulary, our approach involves directly predicting the class name of the target object. Essentially, this reformulates the recognition task as a text generation problem. Mathematically, we frame open-ended recognition as an endeavor to maximize the conditional likelihood of the class name under a forward autoregressive factorization:

$$p(c_i) = \prod_{j=0}^{N} p(c_{i,j}|c_{i,0}, \cdots, c_{i,j-1}), \tag{2}$$

where $c_{i,j}$ corresponds to the j-th text token within the class names for c_i.

3.2 Model Architecture

The architectural overview of OSM is presented in Fig. 3. In alignment with the established modular vision-language models [39, 46, 86], OSM comprises three principal components: a frozen CLIP-ViT, a trainable MaskQ-Former, and a frozen Large Language Model (LLM). Our approach incorporates specialized design enhancements aimed at optimizing the model for object-level recognition, which we detail in the following paragraphs:

High-Resolution Feature Extraction with Frozen CLIP-ViT A frozen vision transformer backbone, pre-trained in the CLIP style, has become the standard choice in existing multi-modal LLM designs. The appeal of CLIP-ViT lies in its dual advantages: it provides a robust and adaptable feature representation for input images, and its feature space is well-suited for seamless conversion into language tokens, which the LLM can comprehend as inputs.

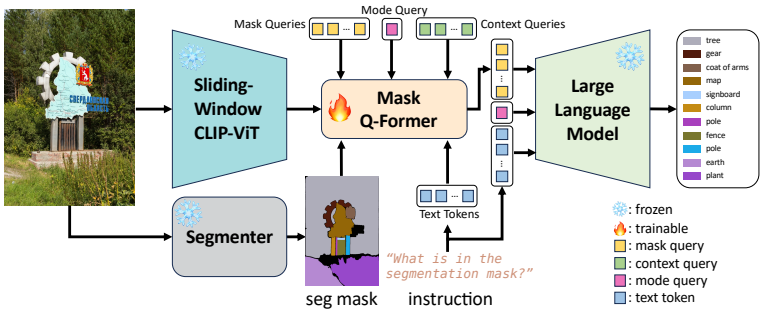

Fig. 3: An overview of the proposed OSM, consisting of a frozen CLIP-ViT that extracts high-resolution features in a sliding-window manner, a trainable MaskQ-Former that resamples pixel features in a mask-aware manner, and a frozen LLM, which predicts a semantic class for each corresponding mask in a generative manner without a predefined vocabulary. OSM can be combined with any off-she-shelf segmenter, *e.g.*, SAM [36] and kMaX-DeepLab [79]. The proposed MaskQ-Former takes as input (1) Mask Queries, (2) Context Queries, and (3) Mode Query. The Mask Queries focus on the mask regions proposed by the off-the-shelf segmenter, while the Context Queries attend to a broader region derived from the mask. The Mode Query consists of two modes: vocabulary-specific and vocabulary-agnostic, allowing the model to perform with dataset-specific and dataset-agnostic vocabularies, respectively. Note that we have two separate Model Query for MaskQ-Former and LLM respectively, and only the Mask Queries from MaskQ-Former are fed into LLM.

Nonetheless, the usage of CLIP-ViT, while successful in many multi-modal LLM applications such as image captioning [9, 54] and visual question answering [25,31], has its limitations. It was originally pre-trained on lower resolutions, typically at resolution 224×224. This lower resolution can hinder its performance, especially when tasked with object-level recognition. Moreover, previous research [78] has observed that a frozen ViT exhibits weak generalization capabilities across varying input resolutions.

Despite the widespread use of frozen ViT backbones in multi-modal LLM models, it is evident that a 224×224 input resolution falls short, particularly for object-level recognition. Typical adaptations, such as windowed attention [47] as seen in ViTDet [42], may not be applicable to a completely frozen ViT backbone. To address this limitation, we propose a straightforward strategy to extract more effective features using a frozen ViT at a higher resolution, for example, 896×896. Specifically, we employ a sliding-window feature extraction approach at the input level, where each window size matches that of the ViT's pre-trained image size. Afterwards, a global positional embedding is added to compensate the missing location information across windows. In our experiments, we empirically prove that this seemingly simple strategy is surprisingly effective, yielding significantly improved performance in feature extraction from high-resolution inputs.

MaskQ-Former We employ a visual resampler, such as Q-Former [39] or Perceiver Resampler [1], to bridge the gap between the encoded image features and inputs suitable for the LLM. This visual resampler typically consists of a

stack of transformer decoders [64] that transform image tokens into a reduced set of query tokens, which are usually far fewer in number compared to image tokens. However, existing visual resamplers like those used in [1, 39], employ a set of queries that globally attend to image features without considering the segmentation mask priors.

In response to this limitation, we introduce a novel variant called MaskQ-Former. The MaskQ-Former takes a segmentation mask as input and performs masked cross-attention [10], as depicted in Fig. 4. It consists of two sets of learnable queries: mask queries and context queries. The mask queries execute masked cross-attention, restricting their focus to the mask region, while the context queries attend to a broader region derived from the mask, such as the bounding box region, to provide complementary contextual information. This contextual information is essential for precise and unbiased recognition [7, 8, 76].

The MaskQ-Former effectively summarizes the mask region while retaining access to essential contextual content. Information exchange between the mask queries and context queries is facilitated through the self-attention layer. Notably, all parameters are shared between the mask and context queries, except for the learnable query initialization, resulting in negligible additional costs. In the end, we retain only the mask queries as inputs to the LLM, ensuring computational efficiency.

Mode Query While our primary objective is to enable OSM to perform effectively in an open-ended setting, where it can make predictions without prior knowledge of any vocabulary, we acknowledge the importance of versatility. OSM has the capability to perform accurately when required to align with a specific vocabulary. To achieve this, we introduce Mode Query, consisting of vocabulary-specific and vocabulary-agnostic queries, drawing inspiration from prefix tuning techniques [41]. These queries leverage the strong instruction-following capabilities of the LLM, enhancing the model's adaptability across diverse scenarios. Concretely, we propose appending a dedicated learnable query for each vocabulary to both the MaskQ-Former and LLM inputs. During training, when utilizing datasets from various sources, the corresponding vocabulary-specific query for each dataset is activated, allowing the model to effectively "memorize" the associated vocabulary of each dataset, thereby improving alignment during prediction. Additionally, we include a general vocabulary-agnostic query that is activated during training on any dataset to keep the open-ended recognition ability.

This approach provides flexibility during testing. We can activate a vocabulary-specific query to ensure that the model's predictions align better with the desired vocabulary, or we can activate the vocabulary-agnostic query to facilitate open-ended predictions. This adaptability enhances OSM's utility across a spectrum of real-world scenarios, making it a versatile tool for a wide range of applications.

3.3 Training and Evaluation Protocols

Datasets To create a robust training and evaluation framework, we ensemble six publicly available segmentation datasets, encompassing diverse image distributions, domains, and segmentation tasks. These datasets include COCO

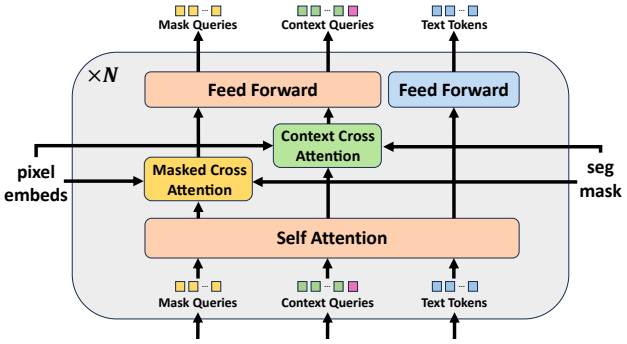

Fig. 4: An overview of MaskQ-Former. The parameters of Masked Cross Attention layer and Context Cross Attention layer are shared. We append the Mode Query to Context Queries. Moreover, Mask Queries only attend to the mask region in cross-attention layer, while Context Queries may attend to a larger region derived from the mask. All queries/tokens will communicate with each other in the self-attention layer.

panoptic segmentation [44], ADE20K panoptic segmentation [82], Cityscapes panoptic segmentation [15], LVIS instance segmentation [27], ADE-847 semantic segmentation [82], and PC-459 semantic segmentation [21].

Training Protocols During training, we adopt an instruction tuning approach [13,46,70] to seamlessly integrate training with the LLM. For each training iteration, we randomly select an image and its corresponding ground-truth mask from a dataset. We randomly choose an instruction template and insert the actual class name. This approach enables training the model using a straightforward next-token prediction loss without the need for intricate designs. We default to the template *What is in the segmentation mask?* and greedy search decoding for testing.

OSM is jointly trained on various datasets, with each batch comprising 32, 64, 16, 8, 16, and 8 samples from COCO, LVIS, ADE-847, PC-459, ADE-20K, and Cityscapes, respectively. In each training batch, half of the data activate vocabulary-specific queries corresponding to their respective datasets, while the other half activate vocabulary-agnostic queries. We use AdamW optimizer [34,50] with learning rate 4×10^{-5} and weight decay 0.05. The learning rate follows a cosine decay schedule. Training continues until the model has processed a total of 6 million masks.

Evaluation Protocols Our model is evaluated on the validation set of each dataset, using two types of masks: ground-truth masks or masks produced by an off-the-shelf segmenter. When using ground-truth masks as inputs, we purely assess mask classification accuracy. Specifically, a prediction is considered correct only when the predicted class name **exactly** matches the class name in the ground-truth annotation. To enhance the reliability of this metric, we augment the ground-truth class names with synonyms sourced from [23, 78]. Additionally, we consider plural and singular formats of class names. It is important to note that these synonyms are not used during model training, as they may not always be semantically aligned (*e.g.*, "person", "man", and "woman" are

synonyms in COCO and LVIS). As a result, we report two metrics: Accuracy (Acc) and Not-in-Vocabulary (NIV), which represent the percentage of predictions correctly match ground-truth classes, or the predictions do not fall into the dataset's vocabulary, respectively. The metric Acc directly evaluates the model's classification capacity, while NIV reflects the model's generalizability or degrees of overfitting to the trained datasets.

Additionally, we consider a more practical application where OSM is connected to an off-the-shelf mask proposal model, such as kMaX-DeepLab [79] or SAM [36]. We directly evaluate the model's performance on the established academic benchmarks, including panoptic segmentation and semantic segmentation, using panoptic quality (PQ) [35] and mIoU [21], respectively.

4 Experimental Results

In this section, we first provide the settings used for the ablation studies and our final model. We then evaluate OSM with ground-truth masks along with ablation studies in Sec. 4.1, followed by the setting using an off-the-shelf mask proposal model in Sec. 4.2.

Default Settings for Ablations Unless otherwise specified, we use the default setting below for ablation studies: We resize both the image and mask during training until the longer side reaches a length of 896 pixels, and then pad the shorter side to match this length. We apply minimal data augmentation, limited to random flipping. The context queries in MaskQ-Former attend to the whole image. We initialize OSM with InstructBLIP [16] pre-trained weight, which uses EVA-ViT-g/224 [22] as vision encoder, and Vicuna-7B [12] as LLM. We use 32 mask queries, 32 context queries, and 1 mode query which is randomly selected between vocab-agnostic query (shared across datasets) and vocab-specific query (one per dataset).

Settings for Final Models Based on the findings in the ablation studies (we detail the results later), for our final model, we increase the image resolution to 1120 and the context queries attend to the bounding box region that is $0.5\times$ larger than the box-constrained mask region. We also use random scale jittering in the range of $[0.5, 1.5]$.

4.1 Mask Classification with Ground-Truth Masks

Generative Model for Discriminative Tasks In Tab. 1, we demonstrate that a generative model can effectively capture the training dataset, yielding predictions well-aligned with the training vocabulary. Specifically, as shown in the top few rows of the table ("*single dataset*"), we first train OSM separately on each of the six segmentation datasets, and evaluate its mask classification accuracy using ground-truth masks. Remarkably, the model, although tasked with unrestricted generation of class names, consistently delivers predictions well within the vocabulary of its respective dataset. This is evident by the high percentage of accurate predictions (*i.e.*, high Acc scores) and the very low percentage of

Table 1: Mask classification accuracy across the six segmentation datasets, using ground-truth masks. Note that OSM (vocab-agnostic) and OSM (vocab-specific) are obtained from the same model and weights, but activate vocabulary-agnostic or vocabulary-specific queries during inference, respectively. NIV: Not-in-Vocabulary. †: Our final model setting.

methods	COCO		LVIS		ADE20K		Cityscapes		ADE-847		PC-459		Avg	
	Acc	NIV	Acc	NIV	Acc	NIV	Acc	NIV	Acc	NIV	Acc	NIV	Acc	NIV
single dataset														
Train on Each Dataset	85.5	$5.3e^{-5}$	68.3	$2.4e^{-3}$	82.3	$4.3e^{-4}$	79.4	$4.9e^{-4}$	76.9	$3.3e^{-3}$	80.9	$6.9e^{-3}$	78.9	$2.3e^{-3}$
multiple datasets														
Learnable Embed	84.3	0.00	67.3	0.00	82.3	0.00	81.0	0.00	76.0	0.00	82.5	0.00	78.9	0.00
Text Embed	83.4	0.00	65.4	0.00	81.5	0.00	81.6	0.00	75.1	0.00	81.7	0.00	78.1	0.00
OSM vocab-agnostic	74.9	11.1	56.8	10.0	80.6	2.31	81.1	0.01	75.6	0.60	77.8	5.69	74.5	4.95
OSM vocab-specific	84.7	0.10	67.0	0.62	82.1	0.55	81.1	$7.0e^{-5}$	76.4	0.49	80.8	1.90	78.7	0.61
OSM † vocab-agnostic	79.5	8.75	64.6	8.22	83.8	2.11	88.7	0.01	76.6	0.81	80.6	3.74	79.0	3.94
OSM † vocab-specific	87.0	0.11	72.7	0.94	85.2	0.31	88.6	0.01	78.1	0.49	83.0	0.55	82.4	0.40

predictions falling outside the vocabulary (*i.e.*, low NIV scores), showcasing the generative model's capacity to operate for a discriminative task.

Next, we explore the more interesting setting, where all six datasets are used for training ("*multiple datasets*" in the table), where OSM still maintains a high accuracy for each individual dataset, even in the presence of potential label conflicts. Specifically, the proposed Mode Query scheme effectively alleviates the label conflicts between datasets, where the vocabulary-specific queries ("vocab-specific" in the table) better learns the associated vocabulary for each dataset, while the vocabulary-agnostic ("vocab-agnostic") maintains the open-ended recognition ability (indicated by higher NIV scores). Notably, this achievement is non-trivial and underscores the value of the proposed Mode Query.

Additionally, we establish two discriminative baselines for comparisons. The first one (denoted as "Learnable Embed") replaces the frozen LLM with six learnable linear layers, each tailored to a specific dataset. The second one (named "Text Embed") initializes the classification layer with pre-extracted text embeddings and applies it individually to each dataset, approximating the approach presented in [26,83]. As shown in the table, our generative model OSM performs comparably to the strong baseline "Learnable Embed" on the average (78.7% *vs.* 78.9% Acc) and outperforms the "Text Embed" baseline. Our findings highlight that the generative model can perform on par with the discriminative models, even in discriminative tasks, underscoring its versatility and effectiveness.

Finally, in the last two rows of the table (denoted as OSM †), using our final model setting (*e.g.*, larger input size) can further significantly improve the performance for both vocabulary-agnostic and vocabulary-specific settings.

Adaptation to Higher Input Resolution In contrast to many multi-modal Large Language Models (LLM) approaches that directly employ the frozen CLIP-ViT, we emphasize the critical importance of higher input resolution for achieving accurate object-level recognition. However, we recognize that frozen Vision Transformers (ViTs) often exhibit inferior performance when adapting to larger input resolutions compared to their pre-training resolutions, as doc-

Table 2: Ablation studies on OSM design choices. The ablated design choices include (a) image input resolution, (b) the sliding-window stride of the CLIP-ViT backbone to extract high-resolution image features, (c) employment of Mode Query, and (d) the box region size attended by the context queries in MaskQ-Former.

input res.	224	448	672	896	1120	1344	1568
Avg Acc	57.3	70.1	75.2	78.7	79.9	76.6	75.9
Avg NIV	0.89	0.84	1.01	0.61	0.74	3.5	3.5

(a) Input Resolution

sliding stride	Global	224	168	112
Avg Acc	71.5	78.7	79.5	79.6
Avg NIV	0.68	0.61	0.59	0.56

(b) Sliding-Window Stride

vocab queries	None	Vocab-Agnostic	Vocab-Specific
Avg Acc	74.8	74.5	78.7
Avg NIV	4.45	4.95	0.61

(c) Mode Query

ratio	Global	0.0×	0.1×	0.2×	0.3×	0.4×	0.5×	0.6×
Avg Acc	78.7	79.5	80.5	80.6	80.9	80.9	81.3	81.0
Avg NIV	0.61	0.51	0.43	0.50	0.62	0.56	0.47	0.50

(d) Context Enlargement Ratio

umented in [78]. To address this limitation, we introduce a simple yet highly effective sliding-window approach for obtaining enhanced features from a frozen ViT when processing higher-resolution inputs.

As illustrated in Tab. 2a, our experiments consistently demonstrate performance gains as input resolution increases, particularly from 224×224 to 448×448, reflecting an impressive improvement of (+12.8%). This underscores the pivotal role of a larger input resolution in achieving superior object-level recognition performance. The benefits persist until the input resolution reaches 1120×1120, while larger input resolution leads to a performance drop, potentially because each sliding-window fails to capture semantic meaningful feature. Notably, the "Avg NIV" metric remains relatively stable across all experiments, indicating that the performance boost primarily stems from improved mask classification rather than a better overfitting with the respective vocabulary.

Sliding-Window Stride We validate our sliding-window design in Tab. 2b, where direct application of the frozen ViT with high-resolution inputs ("Global") results in significantly inferior performance (−7.2%), consistent with the observations in [78]. Moreover, our findings reveal that employing the sliding-window approach with overlapping windows further enhances results, although the incremental benefit diminishes as the overlap increases. Considering the significant additional computational costs coming from overlapping window, we do not use it in our final setting.

Effect of Mode Query It is evident from our experiments that the inclusion of Mode Query plays a pivotal role in the effectiveness of OSM. As demonstrated in Tab. 2c, training OSM across multiple datasets without these queries may result in better generalization capabilities but compromised alignment to specific datasets. This is evident through a lower "Avg Acc" and a higher "Avg NIV". However, with the integration of the proposed Mode Query, OSM exhibits the ability to operate in both "closed-ended" mode (vocab-specific) and "open-ended" mode (vocab-agnostic). This allows the model to strike a balance between generalization and alignment, preserving both essential capabilities.

Context is Important for Recognition We investigate the critical role of context, as detailed in Tab. 2d. Here, "Global" signifies that the context attention may encompass the entire image, whereas "0.0×" refers to a tightly

Table 3: Comparisons with other discriminative models. OSM uses mask proposals from [79]. We mainly compare with the generalist models (LMSeg and DaTaSeg) and list Mask2Former as the specialist model for reference. Note that DaTaSeg [26] uses ADE20K-semantic training data instead of ADE20K-panoptic (thus mark in gray).

	proposal network	COCO	ADE20K		Cityscapes	
methods	backbone	PQ	PQ	mIoU	PQ	mIoU
specialist models (one model per dataset)						
Mask2Former [10]	ResNet50 [29]	51.9	39.7	46.1	62.1	77.5
Mask2Former [10]	Swin-L [47]	57.8	48.1	54.5	66.6	82.9
generalist models (one model for all datasets)						
LMSeg [83]	ResNet50 [29]	38.6	35.4	45.2	54.8	80.9
DaTaSeg [26]	ResNet50 [29]	49.0	29.8	48.1	-	-
DaTaSeg [26]	ViTDet-L [42]	53.5	33.4	54.0	-	-
OSM	ResNet50 [29]	53.3	43.8	50.0	59.5	77.0
OSM	ConvNeXt-L [48]	56.1	49.7	55.2	64.7	80.2

constrained bounding box that encircles the segmentation mask closely. The notation "$k\times$" indicates the expansion of the bounding box by a factor of "$k\times$" on each side. The results in the table underscore the significance of context. Even a tightly defined bounding box offers a noteworthy improvement over global context (+0.8%). Notably, the benefits become more pronounced as we progressively transition to a looser bounding box, with the most substantial gain occurring at "0.5×" (+2.6%) compared to the global context configuration. This underlines the importance of context for accurate recognition, with an optimal balance between tight and loose bounding boxes yielding superior results.

4.2 Mask Classification with Off-the-shelf Mask Proposal Model

Benchmarking with Other Generalists In addition to evaluating OSM with ground-truth masks, we also provide a practical assessment by integrating OSM with an off-the-shelf mask proposal model. We employ mask proposals generated by kMaX-DeepLab [79] and then apply OSM for classifying these mask proposals. We focus on the comparisons with other generalist segmentation models that are jointly trained with multiple segmentation datasets, similar to our setting. Specifically, we compare with text embedding-based methods like LMSeg [83] and DaTaSeg [26] across various datasets, including COCO panoptic, ADE20K panoptic and semantic, Cityscapes panoptic, and semantic segmentation. As outlined in Tab. 3, OSM consistently achieves higher Panoptic Quality (PQ) or mean Intersection-over-Union (mIoU) scores in comparison to other discriminative methods. Specifically, with ResNet50 proposal model backbone, OSM outperforms LMSeg [83] by +14.7, +8.4, and +4.7 PQ on COCO, ADE20K, Cityscapes respectively. Compared to DaTaSeg [26], OSM also improve the COCO PQ by +4.3, +2.6, the ADE20K mIoU by +1.9, +1.2 for ResNet50 and Large backbone variants, respectively. OSM also shows comparable performance to the specialist model Mask2Former [10].

Table 4: Comparisons in open-vocabulary settings on ADE20K (panoptic), ADE-847 (semantic), PC-459 (semantic) *val* sets. †: Methods that can only perform semantic segmentation. ∗: Methods using geometric ensemble from another frozen CLIP. We obtain ODISE and FC-CLIP's non-ensemble results by running their official code and setting $\alpha = 0$, $\beta = 0$ in the Equation (7) of the FC-CLIP paper [78].

zero-shot datasets	ADE20K			ADE-847	PC-459
methods	PQ	AP	mIoU	mIoU	mIoU
OpenSeg† [23]	-	-	21.1	6.3	9.0
MaskCLIP [20]	15.1	6.0	23.7	8.2	10.0
FC-CLIP [78]	17.8	11.1	20.8	4.1	11.8
ODISE [72]	19.5	10.8	23.8	6.0	9.4
OSM	21.4	12.4	26.9	8.8	19.0
OVSeg†* [43]	-	-	29.6	9.0	12.4
SAN†* [73]	-	-	33.3	13.7	15.7
ODISE* [72]	23.4	13.9	28.7	11.0	13.8
FC-CLIP* [78]	26.8	16.8	34.1	14.8	18.2
OSM *	26.9	16.2	33.6	13.3	19.8

Evaluation with Open-Vocabulary Benchmarks We also evaluate OSM aginst state-of-the-art open-vocabulary methods. To ensure an open-vocabulary setting (*i.e.*, the target datasets are never seen during training), we train OSM with COCO and LVIS data only and evaluate on ADE20K, ADE-847, and PC-459 in a zero-shot manner [20]. During testing, we use the same mask proposal model from [78], and replace the classification head with OSM. Furthermore, we map OSM's prediction to target vocabulary using text embedding similarity between predicted class name and target vocabulary class names. Following prior arts [72, 78], we also apply geometric ensemble to enhance the results with the frozen CLIP predictions. We report results with and without ensemble. As shown in Tab. 4, when not using the geometric ensemble method, OSM shows superior scores against state-of-the-art open-vocabulary methods, indicating its strong performance. It is noteworthy that when the geometric ensemble is applied — specifically, using an off-the-shelf CLIP for prediction ensemble, which is not inherently synergistic with OSM's approach of not predicting a class distribution due to the absence of a predefined set of classes — OSM continues to show advantages.

5 Conclusion

In this study, we introduced a novel challenge in the domain of visual recognition, referred to as open-ended visual recognition, and introduced OSM, a generative framework designed to address this challenge. OSM is a mask-aware multi-modal LLM capable of processing segmentation masks as inputs and generating semantic class predictions in a generative manner, without relying on a predefined vocabulary. Our empirical findings demonstrate that this generative model yields promising recognition accuracy and exhibits significant potential for real-world applications, particularly in handling novel concepts that extend beyond predefined vocabularies.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS (2022)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
4. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR (2015)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI (2017)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
9. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
10. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
11. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. NeurIPS (2021)
12. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/`
13. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
14. Conti, A., Fini, E., Mancini, M., Rota, P., Wang, Y., Ricci, E.: Vocabulary-free image classification. NeurIPS (2023)
15. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
16. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL (2018)

19. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: CVPR (2022)
20. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary panoptic segmentation with maskclip. In: ICML (2023)
21. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
22. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: CVPR (2023)
23. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: ECCV (2022)
24. Girshick, R.: Fast r-cnn. In: ICCV (2015)
25. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
26. Gu, X., Cui, Y., Huang, J., Rashwan, A., Yang, X., Zhou, X., Ghiasi, G., Kuo, W., Chen, H., Chen, L.C., Ross, D.: Dataseg: Taming a universal multi-dataset multi-task segmentation model. NeurIPS (2023)
27. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR (2019)
28. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
30. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: ICLR (2021)
31. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
32. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
33. Kim, D., Lin, T.Y., Angelova, A., Kweon, I.S., Kuo, W.: Learning open-world object proposals without learning to classify. In: ICRA (2022)
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
35. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: CVPR (2019)
36. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV (2023)
37. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS (2012)
38. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: Mseg: A composite dataset for multi-domain semantic segmentation. In: CVPR (2020)
39. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
40. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
41. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: ACL (2021)
42. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: ECCV (2022)

43. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: CVPR (2023)
44. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
45. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
46. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2023)
47. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
48. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR (2022)
49. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
50. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
51. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: ECCV. pp. 728–755. Springer (2022)
52. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
53. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
54. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
56. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog (2019)
57. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research (2020)
58. Rewatbowornwong, P., Chatthee, N., Chuangsuwanich, E., Suwajanakorn, S.: Zero-guidance segmentation using zero segment labels. In: ICCV (2023)
59. Roberts, A., Raffel, C., Shazeer, N.: How much knowledge can you pack into the parameters of a language model? In: EMNLP (2020)
60. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NeurIPS (2022)
61. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. NeurIPS (2014)
62. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., et al.: Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
63. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
64. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)

65. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: CVPR (2021)
66. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., Dai, J.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. NeurIPS (2023)
67. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., Qiao, Y.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: CVPR (2023)
68. Wang, X., Li, S., Kallidromitis, K., Kato, Y., Kozuka, K., Darrell, T.: Hierarchical open-vocabulary universal image segmentation. NeurIPS **36** (2024)
69. Wang, Z., Li, Y., Chen, X., Lim, S.N., Torralba, A., Zhao, H., Wang, S.: Detecting everything in the open world: Towards universal object detection. In: CVPR (2023)
70. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. In: ICLR (2022)
71. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS (2022)
72. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: CVPR (2023)
73. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: CVPR (2023)
74. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: ECCV (2022)
75. Yang, L., Wang, Y., Li, X., Wang, X., Yang, J.: Fine-grained visual prompting. NeurIPS **36** (2024)
76. Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N.: Context prior for scene segmentation. In: CVPR (2020)
77. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
78. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. NeurIPS (2023)
79. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: k-means Mask Transformer. In: ECCV (2022)
80. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023)
81. Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., Kang, B.: Bubogpt: Enabling visual grounding in multi-modal llms. arXiv preprint arXiv:2307.08581 (2023)
82. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)
83. Zhou, Q., Liu, Y., Yu, C., Li, J., Wang, Z., Wang, F.: Lmseg: Language-guided multi-dataset segmentation. In: ICLR (2023)
84. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022)
85. Zhou, X., Koltun, V., Krähenbühl, P.: Simple multi-dataset detection. In: CVPR (2022)
86. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

87. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2020)