

General Object Foundation Model for Images and Videos at Scale

Junfeng Wu^{1*}, Yi Jiang^{2*}, Qihao Liu³, Zehuan Yuan², Xiang Bai^{1†}, Song Bai^{2†}

¹Huazhong University of Science and Technology, ²ByteDance Inc., ³Johns Hopkins University

Abstract

We present GLEE in this work, an object-level foundation model for locating and identifying objects in images and videos. Through a unified framework, GLEE accomplishes detection, segmentation, tracking, grounding, and identification of arbitrary objects in the open world scenario for various object perception tasks. Adopting a cohesive learning strategy, GLEE acquires knowledge from diverse data sources with varying supervision levels to formulate general object representations, excelling in zero-shot transfer to new data and tasks. Specifically, we employ an image encoder, text encoder, and visual prompter to handle multi-modal inputs, enabling to simultaneously solve various object-centric downstream tasks while maintaining state-of-the-art performance. Demonstrated through extensive training on over five million images from diverse benchmarks, GLEE exhibits remarkable versatility and improved generalization performance, efficiently tackling downstream tasks without the need for task-specific adaptation. By integrating large volumes of automatically labeled data, we further enhance its zero-shot generalization capabilities. Additionally, GLEE is capable of being integrated into Large Language Models, serving as a foundational model to provide universal object-level information for multi-modal tasks. We hope that the versatility and universality of our method will mark a significant step in the development of efficient visual foundation models for AGI systems. The model and code will be released at <https://glee-vision.github.io/>.

1. Introduction

Foundation model [7] is an emerging paradigm for building artificial general intelligence (AGI) systems, signifying a model trained on broad data that is capable of being adapted to a wide range of downstream tasks in an general paradigm. Recently, NLP foundation models such as BERT [22], GPT-3 [9], T5 [78] developed with unified

*Equal Technical Contribution. †Correspondence to Xiang Bai <xbai@hust.edu.cn> and Song Bai <songbai.site@gmail.com>.

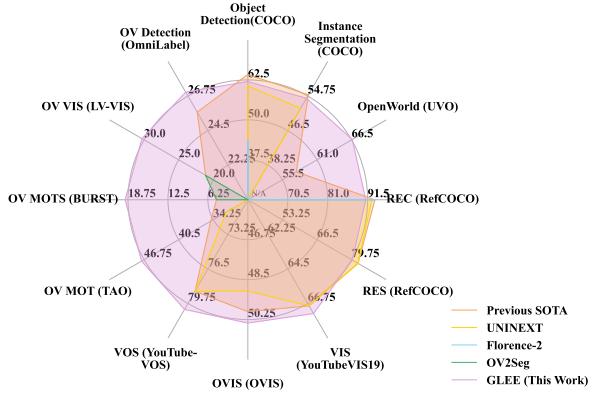


Figure 1. The performance of GLEE on a broad range of object-level tasks compared with existing models.

input-output paradigms and large-scale pre-training, have achieved remarkable generalization capabilities to address nearly all NLP tasks.

In computer vision, the diversity of task types and the lack of a unified from makes visual foundation models only serve specific subdomains, such as CLIP [77] for multi-modal visual model, MAE [35] for visual representations model, SAM [43] for segmentation model. Despite being widely studied, current visual foundation models are still focusing on establishing correlations between global image features and language descriptions or learning image-level feature representations. However, locating and identifying objects constitute foundational capabilities in computer vision systems, serves as a basis for solving complex or high level vision tasks such as segmentation, scene understanding, object tracking, event detection, and activity recognition and support a wide range of applications.

In this work, we advance the development of object-level foundation models within the visual domain. To address the aforementioned limitation, providing general and accurate object-level information, we introduce a general object visual foundation model, coined as GLEE, which simultaneously solve a wide range of object-centric tasks while ensuring SOTA performance, including object detection, instance segmentation, grounding, object tracking, interactive segmentation and tracking, etc., as shown in Figure 1. Through

a unified input and output paradigm definition, our model is capable of learning from a wide range of diverse data and predicting general object representations, which masks it to generalize well to new data and tasks in a zero-shot manner and achieve amazing performance. In addition, thanks to the unified paradigm, the training data can be scaled up at low cost by introducing a large amount of automatically labeled data, and further improve the zero-shot generalization ability of the model.

A general object foundation model framework. Our objective is to build an object visual foundation model capable of simultaneously addressing a wide range of object-centric tasks. Specifically, we employ an image encoder, a text encoder, and a visual prompter to encode multi-modal inputs. They are integrated into a detector to extract objects from images according to textual and visual input. This unified approach to handle multiple modalities enables us to concurrently solve various object-centric tasks, including detection [11, 58, 90, 132], instance segmentation [16, 34], referring expression comprehension [38, 62, 104, 131], interactive segmentation [1, 13, 135], multi-object tracking [21, 68, 111, 126, 129], video object segmentation [17, 18, 73, 110], video instance segmentation [37, 98, 101, 103, 113], and video referring segmentation [86, 102, 104], all while maintaining state-of-the-art performance.

A multi-granularity joint supervision and scaleable training paradigm. The design of the unified framework capable of addressing multiple tasks enables joint training on over five million images from diverse benchmarks and varying levels of supervision. Existing datasets differ in annotation granularity: detection datasets like Objects365 [88] and OpenImages [46] offer bounding boxes and category names; COCO [58] and LVIS [32] provide finer-grained mask annotations; RefCOCO [72, 120] and Visual Genome [44] provide detailed object descriptions. Additionally, video data enhance the temporal consistency of model, while open-world data contribute class-agnostic object annotations. An intuitive display of the supervision types and data scales of the datasets employed is presented in Figure 2. The unified support for multi-source data in our approach greatly facilitates the incorporation of additional manually or automatically annotated data, enabling easy scaling up of the dataset. Furthermore, the alignment of model optimization across tasks means that joint training serves not only as a unifying strategy but also as a mechanism to boost performance across individual tasks.

Strong zero-shot transferability to a wide range of object level image and video tasks. After joint training on data from diverse sources, GLEE demonstrates remarkable versatility and zero-shot generalization abilities. Extensive experiments demonstrate that GLEE achieves state-of-the-art performance compared to existing specialist and

generalist models in object-level image tasks such as detection, referring expression comprehension, and open-world detection, all without requiring any task-specific designs or fine-tuning. Furthermore, we showcase the extraordinary generalization and zero-shot capabilities of GLEE in large-vocabulary open-world video tracking tasks, achieving significantly superior performance over existing models even in a zero-shot transfer manner. Additionally, by incorporating automatically annotated data like SA1B [43] and GRIT [75], we are able to scale up our training dataset to an impressive size of 10 million images at a low cost, which is typically challenging to achieve for object-level tasks and further enhances the generalization performance. Moreover, we replace the SAM [43] component with GLEE in a multimodal Large Language Model (mLLM) [47] and observe that it achieves comparable results. This demonstrates that GLEE is capable of supplying the visual object-level information that modern LLMs currently lack, thus laying a solid foundation for an object-centric mLLMs.

2. Related Work

2.1. Visual Foundation Model

As foundation models [9, 19, 22, 78, 91] in the NLP field have achieved remarkable success, the construction of visual foundation models attracts increasing attention. Unlike NLP tasks that are predominantly unified under a text-to-text paradigm, tasks in Computer Vision still exhibit significant differences in form and definition. This disparity leads to visual foundation models typically being trained in a single-task learning frameworks, limiting their applicability to tasks within certain sub-domains. For instance, multi-modal visual foundation models like CLIP [77], ALIGN [41], Florence [121], BEIT3 [97], Flamingo[2] make significant advancements in efficient transfer learning and demonstrate impressive zero-shot capabilities on vision-language tasks by employing contrastive learning and masked data modeling on large-scale image-text pairs. DALL-E [79, 80] and Stable Diffusion [83] are trained on massive pairs of images and captions, enabling them to generate detailed image content conditioned on textual instruction. DINO [12], MAE [35], EVA [27], ImageGPT [14] obtain strong visual representations through self-supervised training on large-scale image data, which are then employed to transfer to downstream tasks. These foundation models learn image-level features, which are not directly applicable to object-level tasks. The recently proposed SAM [43], capable of segmenting any object of a given image based on visual prompt such as points and boxes, provides rich object-level information and demonstrates strong generalization capabilities. However, the object information lacks semantic context, limiting its application in object-level tasks. Unlike existing visual foundation models, we aim

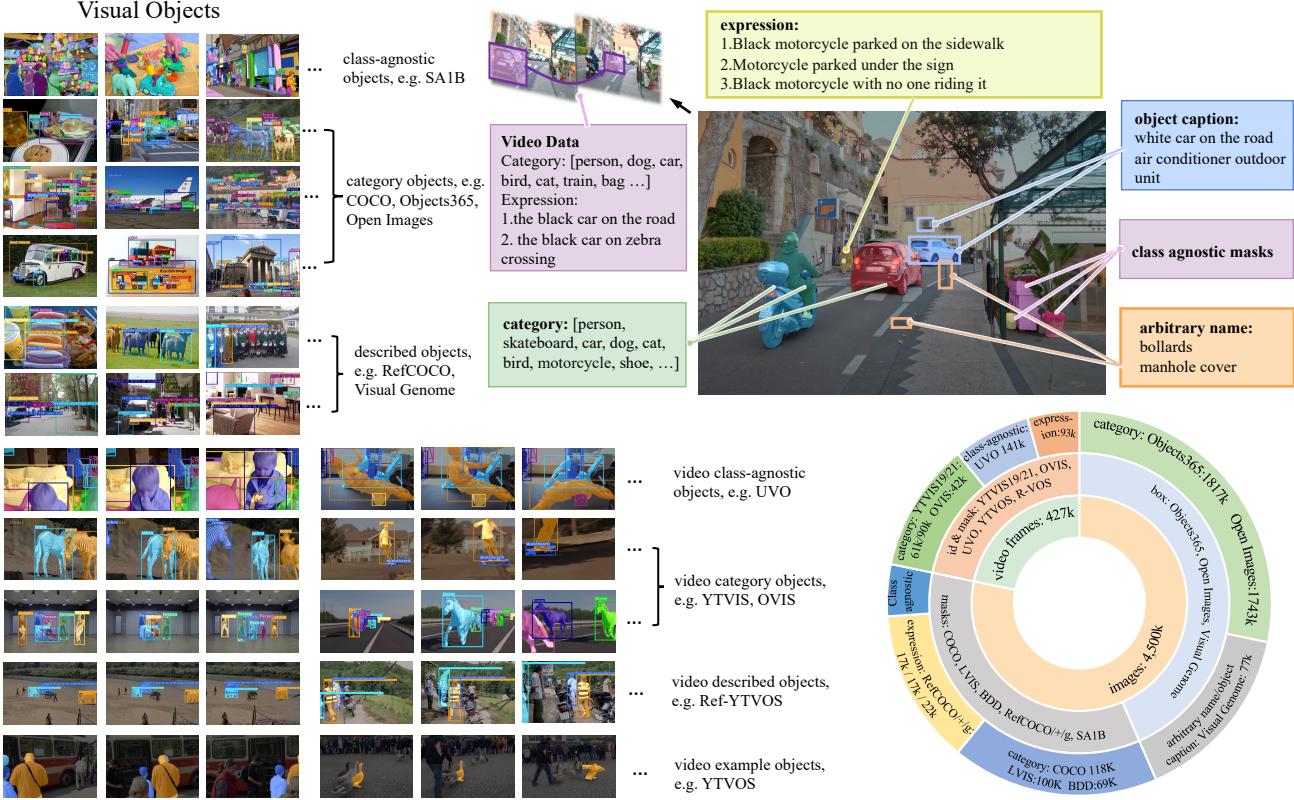


Figure 2. An illustrative example showcasing annotations of varying granularities from different datasets, along with the scale of data we utilized. Training on datasets from multiple sources endows the model with more universal representations.

to develop an object foundation model that directly solve downstream tasks without the need for additional parameters or fine-tuning.

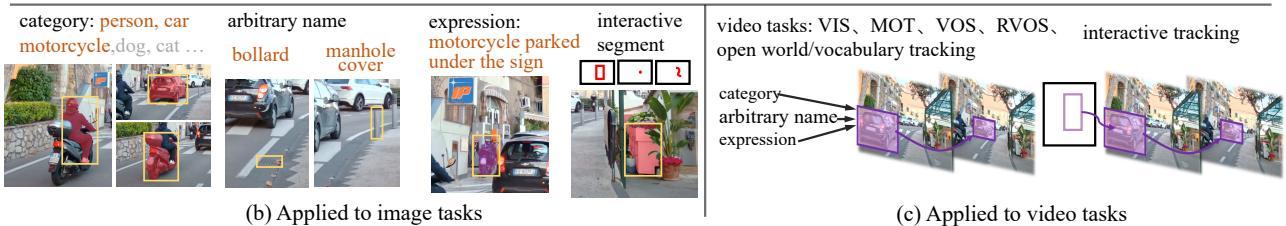
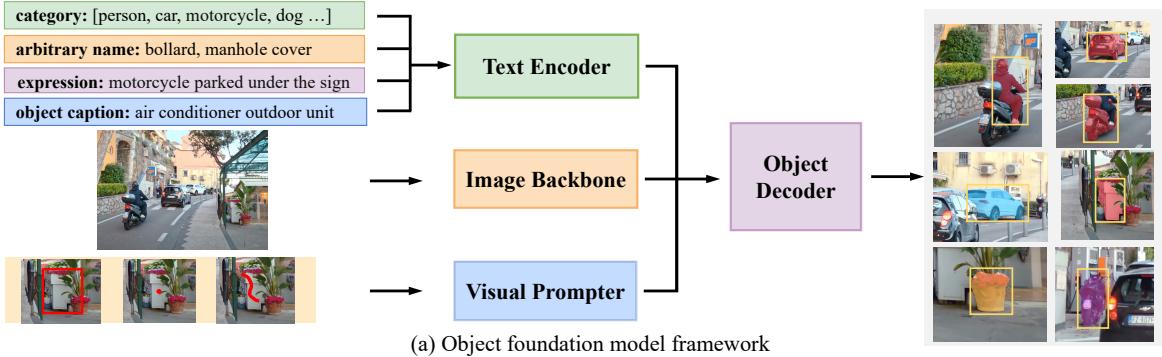
2.2. Unified and General Model

Unified models share similarities with foundation models in the aspect of multi-task unification for their ability to handle multiple vision or multi-modal tasks within a single model. MuST [30] and Intern [87] propose to train across multiple vision tasks and solving them simultaneously. Inspired by the success of sequence-to-sequence NLP models [9, 78], models such as Uni-Perceiver [133], OFA [94], Unified-IO [66], Pix2Seq v2 [15], and UniTAB [114] propose modeling various tasks as sequence generation tasks within a unified paradigm. While these approaches have demonstrated promising cross-task generalization capabilities, they focus mainly on image-level understanding tasks. In addition, their auto-regressive generation of boxes and masks results in significantly slower inference speeds and the performance still falls short of state-of-the-art task-specific models. Building upon on detectors [50, 132], Uni-Perceiver v2 [51] and UNINEXT [112] utilize unified maximum likelihood estimation and object retrieval to support various tasks, effectively resolves the challenges of local-

ization. Nonetheless, they are trained on closed-set data, thereby not exhibiting zero-shot generalization capabilities. X-decoder [134] and SEEM [135] construct a generalized decoding model capable of predicting pixel-level segmentation and language tokens. Diverging from unified models, the proposed GLEE not only directly addresses object-level tasks in a unified manner but also provides universal object representations, which generalize well to new data and tasks, serving as a cornerstone for a broader range of tasks that require detailed object information.

2.3. Vision-Language Understanding

Open-vocabulary detection (OVD) and Grounding models both necessitate the localization and recognition of as many objects as possible. With the recent advancements in vision-language pre-training [41, 77, 119, 121], a commonly employed strategy for OVD involves transferring the knowledge from pre-trained vision-language models (VLMs) to object detectors [31, 45, 71]. Another group of studies leverages extensive image-text pair datasets to broaden the detection vocabulary [28, 52, 57, 116, 122, 128]. However, these language-based detectors are inherently constrained by the capabilities and biases of language models, making it challenging to excel simultaneously in both localization



video tasks: VIS, MOT, VOS, RVOS, open world/vocabulary tracking



Figure 3. **Framework of GLEE.** The text encoder accepts textual descriptions in various forms from diverse data sources, including object categories, names, captions, and referring expressions. The visual prompter encodes points, bounding boxes, or scribbles into corresponding visual representations. The object decoder takes them and image features to predict objects in images. (b) illustrates the application of GLEE to image tasks tailored for different language descriptions and visual prompts. (c) demonstrates the application across various object-level video tasks.

and recognition. Our objective is to optimally utilize existing datasets to construct a general object-level foundation model, aims to not only detect and identify objects effectively but also to offer universal object representations for a wide range of downstream tasks

3. Method

3.1. Formulation

The proposed GLEE consists of an image encoder, a text encoder, a visual prompter, and an object decoder, as illustrated in Figure 3. The text encoder processes arbitrary descriptions related to the task, including object categories, names in any form, captions about objects, and referring expressions. The visual prompter encodes user inputs such as points, bounding boxes, or scribbles during interactive segmentation into corresponding visual representations of target objects. Then they are integrated into a detector for extracting objects from images according to textual and visual input. We build the object decoder upon MaskDINO [50] with a dynamic class head by compute similarity between object embedding from detector and text embedding from the text encoder. Given an input image $I \in \mathcal{R}^{3 \times H \times W}$, we first extract multi-scale features Z with backbones such as ResNet [33]. Then we feed them into the object decoder and adopt three prediction heads (classification, detection, and segmentation) on the output embedding $q_d \in \mathcal{R}^{N \times C}$ from decoder. Following other object segmentation mod-

els [16, 50, 55], we construct a $1/4$ resolution pixel embedding map $M_p \in \mathcal{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$ which is obtained by upsampling and fusing multi-scale feature maps from the backbone and Transformer encoder. Finally, we obtain each binary mask prediction $m \in \mathcal{R}^{N \times \frac{H}{4} \times \frac{W}{4}}$ via a dot product between the N mask embeddings and pixel embedding map:

$$m = FFN(q_d) \otimes M_p, \quad (1)$$

where FFN is a 3-layer feed forward head with ReLU activation function and a linear projection layer.

To support arbitrary vocabularies and object descriptions, we replace the FFN classifier with text embeddings following DetCLIP [115]. Specifically, we feed K category names as separate sentences into the text encoder Enc_L and use the average of each sentence tokens as the output text embedding $e_t \in \mathcal{R}^{K \times D}$ for each category or description. Then we compute the alignment scores $S_{align} \in \mathcal{R}^{N \times K}$ between object embedding and text embedding:

$$S_{align} = q_d \cdot W_{i2t} \otimes e_t, \quad (2)$$

where $W_{i2t} \in \mathcal{R}^{C \times D}$ is image-to-text projection weights. We use logits S_{align} to replace traditional classification logits to compute Hungarian matching cost during training and assign categories to objects during inference. To make the original visual features prompt-aware, an early fusion module is adopted before Transformer encoder following

UNINEXT [112], which tasks image feature from backbone and prompt embedding as input and perform bi-directional cross-attention between them.

3.2. Task Unification

Based on the above designs, GLEE can be used to seamlessly unify a wide range of object perception tasks in images and videos, including object detection, instance segmentation, grounding, multi-target tracking (MOT), video instance segmentation (VIS), video object segmentation (VOS), interactive segmentation and tracking, and supports open-world/large-vocabulary image and video detection and segmentation tasks.

Detection and Instance Segmentation. For detection task, a fixed-length category list is given and all objects in the category list are required to be detected. For a dataset with category list length K , the text input can be formulated as $\{p_k\}_{k=1}^K$ where p_k represents for the k -th category name, e.g., $P = \{\text{"person"}, \text{"bicycle"}, \text{"car"}, \dots, \text{"toothbrush"}\}$ for COCO [58]. For datasets with large vocabulary, calculating the text embedding of all categories is very time-consuming and redundant. Therefore, for datasets with a category number greater than 100, such as objects365 [88] and LVIS [32], suppose there are \hat{K} positive categories in an image, we take the \hat{K} positive categories and then pad the category number to 100 by randomly sampling from the negative categories. For instance segmentation, we enable the mask branch and add mask matching cost with mask loss.

Grounding and Referring Segmentation. These tasks provide reference textual expressions, where objects are described with attributes, for example, Referring Expression Comprehension (REC) [120, 131], Referring Expression Segmentation (RES) [62, 120], and Referring Video Object Segmentation (R-VOS) [86, 102] aim at finding objects matched with the given language expressions like “The fourth person from the left”. For each image, we take the all the object expressions as text prompt and feed the them into the text encoder. For each expressions, we apply global average pooling along the sequence dimension to get text embedding e_t . The text embeddings are feed into early fusion module and additionally interact with object queries through self-attention module in the decoder.

MOT and VIS. Both Multi-object Tracking (MOT)[4, 21, 68, 126, 129] and Video Instance Segmentation (VIS)[37, 76, 103, 113] need to detect and track all the objects in the predefined category list, and VIS requires additional mask for the objects. These two tasks can be considered as extended tasks of detection and instance segmentation on videos. We found that with sufficient image exposure, object embeddings from the decoder effectively distinguish objects in a video, showing strong discriminability and temporal consistency. As a result, they can be directly employed for tracking without the need for an additional

tracking head. Training on image-level data can address straightforward tracking scenarios, but in cases of severe occlusion scenes, such as OVIS [76], image-level training cannot guarantee that the model exhibits strong temporal consistency under occlusion conditions. Therefore, for occlusion scenarios, it is essential to utilize video data for training. Following IDOL [103], we sample two frames from a video and introduce contrastive learning between frames to make the embedding of the same object instance closer in the embedding space, and the embedding of different object instances farther away. During Inference, the detected objects are tracked by simple bipartite matching of the corresponding object queries following MinVIS [39].

Visual Prompted Segmentation. Interactive segmentation [8, 13, 63, 84, 89, 100, 109] takes various forms of visual prompt, such as points, boxes, or scribbles, to segment the specified objects within an image. On the other hand, VOS aims to segment the entire object throughout the entire video based on a mask provided in the first frame of the video. We extract visual prompt embeddings twice in the model. First, we crop the prompt square area from RGB image and send it to the backbone to obtain the visual prompt feature of the corresponding area, and send it to the early fusion module before the Transformer encoder. Second, we sample fine-grained visual embeddings from the pixel embedding map M_p according to visual prompt and make them interacted with object queries through self-attention module in the Transformer decoder layer, as the same with text embeddings.

3.3. Training Unification

Tasks with Dynamic Loss. We jointly train GLEE in an end-to-end manner on over 5 million images from diverse benchmarks with various levels of supervision. Different loss functions are selected for training on various datasets. There are six types of losses for our GLEE: semantic loss, box loss, mask loss, confidence loss, contrastive tracking loss, and distillation loss. For all tasks with category list or object expressions, we apply focal loss [59] as semantic loss on logits S_{align} to align the text concepts with object features. For box prediction, we use a combination of L1 loss and generalized IoU loss [81]. The mask loss is defined as a combination of the Dice loss [70] and Focal loss. For the Visual Prompt Segmentation tasks, we employ an additional FFN to predict the confidence score for each object queries supervised by focal loss. Following IDOL [103], for video tasks, we sample two frames and apply contrastive tracking loss on the object query from the last layer of decoder:

$$\mathcal{L}_{embed} = \log[1 + \sum_{\mathbf{k}^+} \sum_{\mathbf{k}^-} \exp(\mathbf{v} \cdot \mathbf{k}^- - \mathbf{v} \cdot \mathbf{k}^+)], \quad (3)$$

Method	Type	Generic Detection & Segmentation								Referring Detection & Segmentation				OpenWorld			UV0	
		COCO-val		COCO-test-dev		LVIS				RefCOCO		RefCOCO+		RefCOCOg				
		AP_box	AP_mask	AP_box	AP_mask	AP_box	AP _r -box	AP _r -mask	AP _r -mask	P@0.5	oIoU	P@0.5	oIoU	P@0.5	oIoU	AR _{mask}		
MDETR [42]	Specialist Models	-	-	-	-	-	-	-	-	87.5	-	81.1	-	83.4	-	-	-	
SeqTR [131]		-	-	-	-	-	-	-	-	87.0	71.7	78.7	63.0	82.7	64.7	-	-	
PolyFormer (L) [62]		-	-	-	-	-	-	-	-	90.4	76.9	85.0	72.2	85.8	71.2	-	-	
ViTDet-L [55]		57.6	49.8	-	-	51.2	-	46.0	34.3	-	-	-	-	-	-	-	-	
ViTDet-H [55]		58.7	50.9	-	-	53.4	-	48.1	36.9	-	-	-	-	-	-	-	-	
EVA-02-L [26]		64.2	55.0	64.5	55.8	65.2	-	57.3	-	-	-	-	-	-	-	-	57.7	
ODISE [107]		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Mask2Former (L) [16]		-	50.1	-	50.5	-	-	-	-	-	-	-	-	-	-	-	-	
MaskDINO (L) [50]		-	54.5	-	54.7	-	-	-	-	-	-	-	-	-	-	-	-	
UniTAB (B) [114]		-	-	-	-	-	-	-	-	88.6	-	81.0	-	84.6	-	-	-	
OFA (L) [94]	Generalist Models	-	-	-	-	-	-	-	-	90.1	-	85.8	-	85.9	-	-	-	
Pix2Seq v2 [15]		46.5	38.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Uni-Perceiver-v2 (B) [51]		58.6	50.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Uni-Perceiver-v2 (L) [51]		61.9	53.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
UNINEXT (R50) [112]		51.3	44.9	-	-	36.4	-	-	-	89.7	77.9	79.8	66.2	84.0	70.0	-	-	
UNINEXT (L) [112]		58.1	49.6	-	-	-	-	-	-	91.4	80.3	83.1	70.0	86.9	73.4	-	-	
UNINEXT (H) [112]		60.6	51.8	-	-	-	-	-	-	92.6	82.2	85.2	72.5	88.7	74.7	-	-	
GLIPv2 (B) [123]		-	-	58.8	45.8	-	-	-	-	-	-	-	-	-	-	-	-	
GLIPv2 (H) [123]		-	-	60.6	48.9	-	-	-	-	-	-	-	-	-	-	-	-	
X-Decoder (B) [134]		-	45.8	-	45.8	-	-	-	-	-	-	-	-	-	-	-	-	
X-Decoder (L) [134]		-	46.7	-	47.1	-	-	-	-	-	-	-	-	-	-	-	-	
Florence-2 (L) [106]		43.4	-	-	-	-	-	-	-	93.4	-	88.3	-	91.2	-	-	-	
GLEE-Lite	Foundation Models	55.0	48.4	54.7	48.3	44.2	36.7	40.2	33.7	88.5	77.4	78.3	64.8	82.9	68.8	66.6	-	
GLEE-Plus		60.4	53.0	60.6	53.3	52.7	44.5	47.4	40.4	90.6	79.5	81.6	68.3	85.0	70.6	70.6	-	
GLEE-Pro		62.0	54.2	62.3	54.5	55.7	49.2	49.9	44.3	91.0	80.0	82.6	69.6	86.4	72.9	72.6	-	

Table 1. Comparison of GLEE to recent specialist and generalist models on object-level image tasks. For REC and RES tasks, we report Precision@0.5 and overall IoU (oIoU). For open-world instance segmentation task, we reported the average recall of 100 mask proposals (AR@100) on the UV0 [96].

where \mathbf{k}^+ and \mathbf{k}^- are the object queries belong to the same object and other objects from the reference frame, respectively. For the text encoder, we distill the knowledge from the teacher CLIP text encoder to ensure the text embedding in pre-trained vison-language embedding space. We apply an L1 loss between our text encoder and CLIP text encoder to minimize their distance:

$$\mathcal{L}_{text} = \frac{1}{K} \sum_{i=0}^K \|Enc_L(p_i) - Enc_{CLIP}(p_i)\|. \quad (4)$$

Data Scale Up. A visual foundation model should be able to easily scale up the training data and achieve better generalization performance. Thanks to the unified training paradigm, the training data can be scaled up at low cost by introducing a large amount of automatically labeled data from SA1B [43] and GRIT [75]. SA1B provides large and detailed mask annotations, which enhance the object perception capabilities of model, while GRIT offers a more extensive collection of referring-expression-bounding-box pairs, improving the object identification abilities and the understanding capability of descriptions. Ultimately, we introduced 2 million SA1B data points and 5 million GRIT data points into the training process, bringing the total training data to 10 million.

4. Experiments

4.1. Experimental Setup

Datasets and Training Strategy. We conducted training in three stages. Initially, we performed pretraining

for object detection task on Objects365 [88] and Open-Images [46], initializing the text encoder with pretrained CLIP [77] weights and keeping the parameters frozen. In the second training step, we introduced additional instance segmentation datasets, including COCO [58], LVIS [32], and BDD [118]. Furthermore, we treat three VIS datasets: YTVIS19 [113], YTVIS21 [108], and OVIS [76], as independent image data to enrich the scenes. For datasets that provide descriptions of objects, we included RefCOCO [120], RefCOCO+ [120], RefCOCOg [72], Visual Genome [44], and RVOS [86]. Since Visual Genome contains multiple objects in a single image, we treat it as detection task and used both object descriptions and object noun phrases as categories, with a total of 200 dynamic category lists per batch. Additionally, we introduced two open-world instance segmentation datasets, UV0 [96] and a subset of SA1B [43]. For these two datasets, we set the category name for each object to be 'object' and train in instance segmentation paradigm. During the second step, text encoder is unfrozen but supervised by distillation loss to ensure the predicted text embedding in CLIP embedding space. After the second step, GLEE demonstrated state-of-the-art performance on a range of downstream image and video tasks and exhibited strong zero-shot generalization capabilities, unless otherwise specified, all the experimental results presented below were obtained by the model at this stage.

Building upon this, we introduce the SA1B and GRIT datasets to scale up the training set, resulting in a model named **GLEE-scale**, which exhibited even stronger zero-shot performance on various downstream tasks. Since im-

Method	Tracking Any Object (TAO [20])				BURST [3]						LV-VIS [93]		
	TETA	LocA	AssocA	ClsA	ALL		Common		Uncommon		AP	AP _b	AP _n
					HOTA	mAP	HOTA	mAP	HOTA	mAP			
Tracktor [5]	24.2	47.4	13.0	12.1	-	-	-	-	-	-	-	-	-
DeepSORT [99]	26.0	48.4	17.5	12.1	-	-	-	-	-	-	-	-	-
Tracktor++ [20]	28.0	49.0	22.8	12.1	-	-	-	-	-	-	-	-	-
QDTrack [74]	30.0	50.5	27.4	12.1	-	-	-	-	-	-	-	-	-
TETer [53]	33.3	51.6	35.0	13.2	-	-	-	-	-	-	-	-	-
OVTrack† [54]	34.7	49.3	36.7	18.1	-	-	-	-	-	-	-	-	-
STCN Tracker† [3]	-	-	-	-	5.5	0.9	17.5	0.7	2.5	0.6	-	-	-
Box Tracker† [3]	-	-	-	-	8.2	1.4	27.0	3.0	3.6	0.9	-	-	-
Detic [130]-SORT† [6]	-	-	-	-	-	-	-	-	-	-	12.8	21.1	6.6
Detic [130]-XMem †[17]	-	-	-	-	-	-	-	-	-	-	16.3	24.1	10.6
OV2Seg-R50† [93]	-	-	-	-	-	3.7	-	-	-	-	14.2	17.2	11.9
OV2Seg-B† [93]	-	-	-	-	-	4.9	-	-	-	-	21.1	27.5	16.3
UNINEXT (R50) [112]	31.9	43.3	35.5	17.1	-	-	-	-	-	-	-	-	-
GLEE-Lite†	40.1	56.3	39.9	24.1	22.6	12.6	36.4	18.9	19.1	11.0	19.6	22.1	17.7
GLEE-Plus†	41.5	52.9	40.9	30.8	26.9	17.2	38.8	23.7	23.9	15.5	30.3	31.6	29.3
GLEE-Pro†	47.2	66.2	46.2	29.1	31.2	19.2	48.7	24.8	26.9	17.7	23.9	24.6	23.3

Table 2. Comparison of GLEE to recent specialist and generalist models on object-level video tasks in a zero-shot manner. Evaluation metrics of BURST are reported separately for ‘common’, ‘uncommon’ and ‘all’ classes. The mAP computes mask IoU at the track level, HOTA is a balance of per-frame detection accuracy (DetA) and temporal association accuracy (AssA), and TETA that deconstructs detection into localization and classification components. The AP, AP_b, and AP_n in LV-VIS mean the average precision of overall categories, base categories, and novel categories. † does not use videos for training. The under-performance of Pro relative to Plus on LV-VIS is due to Pro employing larger training and inference resolutions, which prove to be sub-optimal for this specific dataset

age data alone is insufficient for the model to learn temporal consistency features, we incorporated sequential video data from YTVIS, OVIS, RVOS, UVOS, and VOS to improve its performance if specifically note.

Implementation Details. In our experiments, we developed GLEE-Lite, GLEE-Plus, and GLEE-Pro using ResNet-50 [33], Swin-Large [64], and EVA-02 Large [26] as the vision encoder respectively. Following MaskDINO [50], we adopt deformable transformer in object decoder, and use 300 object queries. Query denoising and Hybrid matching are kept to accelerate convergence and improve performance. During pretraining, we set a minibatch to 128 on 64 A100 GPUs, for 500,000 iterations. For joint-training, we train GLEE on 64 A100 GPUs for 500,000 iterations, further training details, data pre-processing methods, and data sampling strategies can be found in the supplementary materials. More detailed information on data usage and model training is available in the supplementary materials.

4.2. Comparison with Generalist Models

We demonstrate the universality and effectiveness of our model as an object-level visual foundation model, directly applicable to various object-centric tasks while ensuring state-of-the-art performance without needing fine-tuning. We compare our approach with existing specialist and generalist models in image-level tasks, including detection, referring expression comprehension, and open-world instance

segmentation. We report detection and instance segmentation results on both the COCO validation [58] set and LVIS val v1.0 [32]. While sharing almost identical image sets, LVIS is distinguished by its annotations of over 1,200 object categories, showcasing a long-tail distribution. This distinction makes LVIS more representative of challenging real-world scenarios due to its broader category coverage. As indicated in Table 1, our model outperforms all generalist models on both COCO and LVIS benchmarks. Even when compared to other state-of-the-art specialist approaches, which are tailored with specific design, our model remains highly competitive. This demonstrates that GLEE, while mastering universal and general object representations, concurrently maintains advanced capabilities in object detection and segmentation. This characteristic is vitally important for adapting to a broad spectrum of downstream tasks requiring precise object localization. For the REC and RES tasks, we evaluated our model on RefCOCO [120], RefCOCO+ [120], and RefCOCOg [72], as shown in Table 1, GLEE achieved comparable results with SOTA specialist methods PolyFormer [62], demonstrating strong capability to comprehend textual descriptions and showcasing potential to adapt to a broader range of multimodal downstream tasks. In open-world instance segmentation tasks, we treated “object” as the category name, instructing the model to identify all plausible instance in an image in a class-agnostic manner. GLEE outperforms previous arts ODISE [107] by 8.9 points, demonstrating the

Model	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	Avg
GLIP-T	56.2	12.5	18.4	70.2	50.0	73.8	72.3	57.8	26.3	56.0	49.6	17.7	44.1	46.5
GLIP-L	61.7	7.1	26.9	75.0	45.5	49.0	62.8	63.3	68.9	57.3	68.6	25.7	66.0	52.1
GLEE-Lite	61.7	7.9	23.2	72.6	41.9	51.6	32.9	51.1	35.0	59.4	45.6	21.8	56.9	43.2
GLEE-Lite-Scale	61.2	5.0	23.9	71.9	46.2	57.8	25.6	56.8	33.1	60.6	57.1	25.3	52.5	44.4
GLEE-Plus	67.8	10.8	38.3	76.1	47.4	19.2	29.4	63.8	66.7	63.8	62.6	15.3	66.5	48.3
GLEE-Plus-Scale	67.5	12.1	39.7	75.8	50.3	41.1	42.4	66.4	64.0	62.8	61.8	17.5	63.8	51.2
GLEE-Pro	68.9	16.5	37.6	77.2	23.3	40.1	44.7	68.2	66.2	66.1	63.2	18.1	65.8	50.5
GLEE-Pro-Scale	69.1	13.7	34.7	75.6	38.9	57.8	50.6	65.6	62.7	67.3	69.0	30.7	59.1	53.4

Table 3. Zero-shot performance on 13 ODinW datasets.

Method	Backbone	YTVIS 2019 val [113]			OVIS val [76]		
		AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅
IFC [40]		42.8	65.8	46.8	13.1	27.8	11.6
SeqFormer [101]		47.4	69.8	51.8	15.1	31.9	13.8
IDOL [103]		49.5	74.0	52.9	30.2	51.3	30.0
VITA [36]		49.8	72.6	54.5	19.6	41.2	17.4
GenVIS [37]	ResNet-50	51.3	72.0	57.8	34.5	59.4	35.0
DVIS [124]		52.6	76.5	58.2	34.1	59.8	32.3
NOVIS [69]		52.8	75.7	56.9	32.7	56.2	32.6
UNINEXT		53.0	75.2	59.1	34.0	55.5	35.6
GLEE-Lite		53.1	74.0	59.3	27.1/32.3	45.4/52.2	26.3/33.7
SeqFormer [101]		59.3	82.1	66.4	-	-	-
VITA [36]		63.0	86.9	67.9	27.7	51.9	24.9
IDOL [103]		64.3	87.5	71.0	42.6	65.7	45.2
GenVIS [37]	Swin-L	63.8	85.7	68.5	45.4	69.2	47.8
DVIS [124]		64.9	88.0	72.7	49.9	75.9	53.0
NOVIS [69]		65.7	87.8	72.2	43.5	68.3	43.8
GLEE-Plus	EVA02-L	63.6	85.2	70.5	29.6/40.3	50.3/63.8	28.9/39.8
UNINEXT	ConvNeXt-L	64.3	87.2	71.7	41.1	65.8	42.0
UNINEXT	ViT-H	66.9	87.5	75.1	49.0	72.5	52.2
GLEE-Pro	EVA02-L	67.4	87.1	74.1	38.7/ 50.4	59.4/71.4	39.7/ 55.5

Table 4. Performance comparison of our GLEE on video instance segmentation tasks.

capability of identifying all plausible instance that might be present in an open-world scenario.

4.3. Zero-shot Evaluation Across Tasks

Zero-shot Transfer to Video Tasks. The proposed GLEE is capable of adapting to new data and even new tasks in a zero-shot manner, without the need for additional fine-tuning. We evaluate its zero-shot capability on three large-scale, large-vocabulary open-world video tracking datasets: TAO [20], BURST [3], and LV-VIS [93]. TAO comprises 2,907 high-resolution videos across 833 categories. BURST builds upon TAO, encompassing 425 base categories and 57 novel categories. LV-VIS offers 4,828 videos within 1,196 well-defined object categories. These three benchmarks require the model to detect, classify, and track all objects in videos, while BURST and LV-VIS additionally require segmentation results from the model. In Table 2, we compare the performance of our proposed model with existing specialist models. **Notably, the GLEE here is from the second training stage, which has not been exposed to images from these three datasets nor trained on video-level data.** Despite these constraints, GLEE achieves state-of-the-art performance that significantly exceeds existing

methodologies. Specifically, GLEE surpasses the previous best method OVTrack by 36.0% in TAO, nearly triples the performance of the best baseline in BURST, and outperforms OV2Seg [93] by 43.6% in LV-VIS. This outstanding performance strongly validates the exceptional generalization and zero-shot capabilities of GLEE in handling object-level tasks across a range of benchmarks and tasks.

We additionally provide performance comparison on classic video segmentation tasks, including VIS, VOS, and RVOS. As shown in Table 4, on the YTVIS2019 [113] benchmark, our model achieves SOTA results across various model sizes, surpassing all specialist models with complex designs to enhance temporal capabilities and the video unified model UNINEXT [112]. On the OVIS [76] benchmark, which features lengthy videos with extensive object occlusions where temporal capabilities of object features are particularly crucial, our model does not directly reach SOTA. However, after a few hours of simple fine-tuning, it still achieves SOTA performance. This further validates the versatility and generalization capabilities of our model. More details on zero-shot evaluations for video tasks and demonstrations of interactive segmentation and tracking can be found in the Sec 7 of supplementary materials.

Zero-shot Transfer to Real-world Downstream Tasks.

To measure generalization on diverse real-world tasks, we evaluate zero-shot performance on OmniLabel [85], which is a benchmark for evaluating language-based object detectors and encourages complex and diverse free-form text descriptions of objects. As show in Table 5, compared to language-based detectors trained on large-scale caption data, GLEE significantly outperforms previous models in P-categ. However, due to the limited captions in our training dataset, it scores lower in AP-descr. By incorporating a more diverse set of box-caption data from the GRIT [75] to sclae up our training set, the AP-descr can be elevated to a level comparable with existing models. We conduct additional experiments on the “Object Detection in the Wild” (ODinW) benchmark [48], which is a suite of datasets covering a wide range of domains. We report the average mAP on the subset of 13 ODinW detection datasets introduced in [52], and report the per-dataset performance in a zero-shot manner, as shown in Table 3. GLEE performs better than GLIP [52] on the average of 13 public datasets, show-

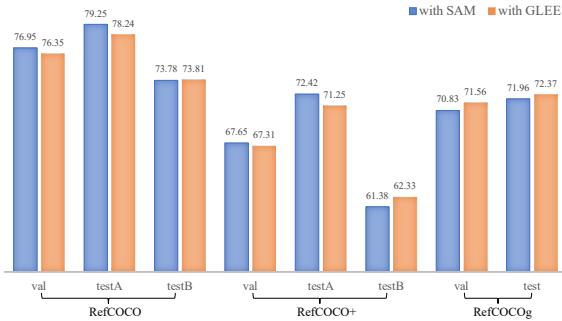


Figure 4. The performance comparison of replacing SAM with GLEE in LISA, GLEE achieves the same effectiveness as SAM in extracting objects.

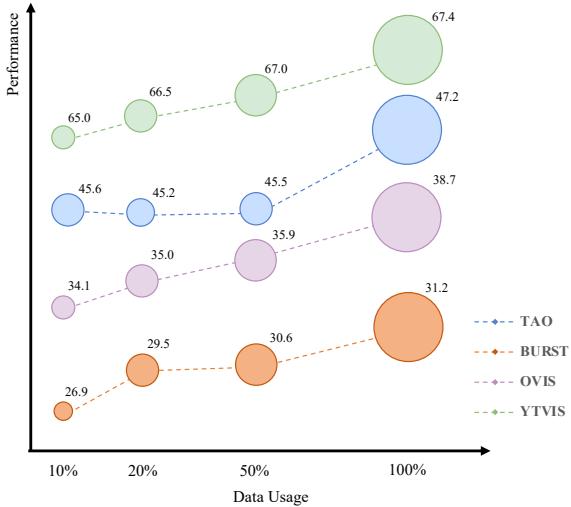


Figure 5. **Data scaling.** The performance of GLEE-Pro after training on 10%, 20%, 50%, 100% of the total data on TAO, BURST, OVIS, YTVIS19. Increased scale of training data result in enhanced zero-shot performance across diverse downstream tasks.

casing its robust generalization capability. Furthermore, it is evident that by introducing automatically labeled data at a low cost for scaling up the training data, the zero-shot capabilities can be further enhanced, this reveals that GLEE has greater potential through scale-up. A more comprehensive report on the per-dataset few-shot performance on ODinW is available in the supplementary materials to assess the adaptability of GLEE to other datasets.

4.4. Serve as Foundation Model

To explore whether GLEE can serve as a foundation model for other architectures, we selected LISA [47] for analysis, a mVLLM that combines LLAVA [61] with SAM [43] for reasoning segmentation. We substituted its vision backbone with a frozen, pretrained GLEE-Plus and fed the object queries from GLEE into LLAVA and remove decoder of LISA. We directly dot product the output SEG tokens with

Images Method	AP	AP-categ		AP-descr		AP-descrpos		AP-descrS		AP-descrM		AP-descr_L	
		AP	AP-categ	AP-descr	AP-descrpos	AP-descrS	AP-descrM	AP-descr_L					
RegionCLIP [127]	2.7	2.7	2.6	3.2	3.6	2.7	2.3						
Detic [130]	8.0	15.6	5.4	8.0	5.7	5.4	6.2						
MDETR [42]	-	-	4.7	9.1	6.4	4.6	4.0						
GLIP-T [52]	19.3	23.6	16.4	25.8	29.4	14.8	8.2						
All GLIP-L [52]	25.8	32.9	21.2	33.2	37.7	18.9	10.8						
FIBER-B [25]	25.7	30.3	22.3	34.8	38.6	19.5	12.4						
GLEE-Lite	20.3	37.5	14.0	19.1	23.0	12.7	10.0						
GLEE-Lite-Scale	22.7	35.5	16.7	22.3	33.7	14.3	10.2						
GLEE-Plus	25.4	46.7	17.5	23.9	28.4	16.3	12.5						
GLEE-Plus-Scale	27.0	44.5	19.4	25.9	36.0	17.2	12.4						

Table 5. Evaluation on the OmniLabel benchmark. The final AP value is the geometric mean of categories (AP-categ) and free-form descriptions (AP-descr).

GLEE feature map to generate masks. As shown in Figure 4, after training for the same number of steps, our modified LISA-GLEE achieved comparable results to the original version, demonstrating the versatility of representations from GLEE and its effectiveness in serving other models.

4.5. Ablation

We conducted experiments to investigate the impact of training data scale on zero-shot performance across various tasks. To this end, we trained GLEE-Pro with 10%, 20%, 50%, 100% of the training data to evaluate the performance on zero-shot transfer tasks, including TAO, BURST, OVIS, and YTVIS as illustrated in the Figure 5. Our data scaling experiments reveal that increased sizes of training datasets result in enhanced zero-shot performance across diverse downstream tasks. This outcome implies that larger pre-training datasets are a valuable investment, offering a more effective and adaptable basis for a broad spectrum of downstream tasks. Thanks to the unified training approach of GLEE, we can efficiently incorporate any manually or automatically annotated data into our training process to achieve enhanced generalization capabilities.

5. Conclusion

We introduce GLEE, a cutting-edge object-level foundation model designed to be directly applicable to a wide range of object-level image and video tasks. Crafted with a unified learning paradigm, GLEE learns from diverse data sources with varying levels of supervisions. GLEE achieves state-of-the-art performance on numerous object-level tasks and excels in zero-shot transfer to new data and tasks, showing its exceptional versatility and generalization abilities. Additionally, GLEE provides general visual object-level information, which is currently missing in modern LLMs, establishing a robust foundation for object-centric mLLMs.

References

- [1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. 2018. 2
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736, 2022. 2
- [3] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1674–1683, 2023. 7, 8
- [4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019. 5
- [5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019. 7
- [6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 7
- [7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1
- [8] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 105–112. IEEE, 2001. 5
- [9] 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. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 3
- [10] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 1
- [11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2
- [12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2
- [13] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5230–5238, 2017. 2, 5
- [14] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Hee woo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2
- [15] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 3, 6
- [16] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 4, 6
- [17] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 2, 7
- [18] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2
- [19] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2
- [20] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 7, 8
- [21] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129(4):845–881, 2021. 2, 5
- [22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2
- [23] Henghui Ding, Chang Liu, Shuteng He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2302.01872*, 2023. 1, 2
- [24] Zihan Ding, Tianrui Hui, Shaofei Huang, Si Liu, Xuan Luo, Junshi Huang, and Xiaoming Wei. Progressive multimodal interaction network for referring video object segmentation. *The 3rd Large-scale Video Object Segmentation Challenge*, 2021. 2
- [25] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *NeurIPS*, 35:32942–32956, 2022. 9
- [26] 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. 6, 7

- [27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 2
- [28] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Prompt-det: Towards open-vocabulary detection using uncurated images. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, page 701–717, 2022. 3
- [29] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 1
- [30] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8856–8865, 2021. 3
- [31] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 3
- [32] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2, 5, 6, 7, 1
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 7
- [34] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2
- [35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2
- [36] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. In *Advances in Neural Information Processing Systems*, 2022. 8
- [37] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14623–14632, 2023. 2, 5, 8
- [38] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14, pages 108–124. Springer, 2016. 2
- [39] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 5
- [40] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. In *NeurIPS*, 2021. 8
- [41] Chao Jia, Yinfen 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*, pages 4904–4916. PMLR, 2021. 2, 3
- [42] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 6, 9
- [43] 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. In *ICCV*, 2023. 1, 2, 6, 9
- [44] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 6, 1
- [45] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2023. 3
- [46] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2, 6, 1
- [47] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 9
- [48] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevator: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS*, 2022. 8, 1, 2
- [49] Dezhuang Li, Ruqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You only infer once: Cross-modal meta-transfer for referring video object segmentation. In *AAAI*, 2022. 2
- [50] 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 *CVPR*, 2023. 3, 4, 6, 7, 1
- [51] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhui Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Pro-*

- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2700, 2023. 3, 6
- [52] 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 *CVPR*, pages 10965–10975, 2022. 3, 8, 9, 2
- [53] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *ECCV*, 2022. 7
- [54] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5567–5577, 2023. 7
- [55] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 4, 6
- [56] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021. 2
- [57] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 3
- [58] 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 *ECCV*, 2014. 2, 5, 6, 7, 1
- [59] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5
- [60] Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, and Wei Liu. Swem: Towards real-time video object segmentation with sequential weighted expectation-maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1362–1372, 2022. 2
- [61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 9
- [62] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 2, 5, 6, 7
- [63] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 5
- [64] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 7
- [65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1
- [66] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3
- [67] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. In *NeurIPS*, 2023. 2
- [68] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 2, 5
- [69] Tim Meinhardt, Matt Feiszli, Yuchen Fan, Laura Leal-Taixe, and Rakesh Ranjan. Novis: A case for end-to-end near-online video instance segmentation. *arXiv preprint arXiv:2308.15266*, 2023. 8
- [70] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 2016. 5
- [71] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, page 728–755, 2022. 3
- [72] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2, 6, 7, 1
- [73] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2
- [74] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, 2021. 7
- [75] 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. 2, 6, 8, 1
- [76] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, pages 1–18, 2022. 5, 6, 8, 1, 2
- [77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-

- ing transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6
- [78] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1, 2, 3
- [79] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2
- [80] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2
- [81] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5
- [82] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *CVPR*, 2020. 2
- [83] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [84] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “grabcut” interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 5
- [85] Samuel Schulter, Vijay Kumar B G, Yumin Suh, Konstantinos M. Dafnis, Zhixing Zhang, Shiyu Zhao, and Dimitris Metaxas. Omnilabel: A challenging benchmark for language-based object detection. In *ICCV*, 2023. 8
- [86] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 2, 5, 6, 1
- [87] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jiacao Liu, et al. Intern: A new learning paradigm towards general vision. *arXiv preprint arXiv:2111.08687*, 2021. 3
- [88] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2, 5, 6, 1
- [89] Konstantin Sofiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 5
- [90] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021. 2
- [91] Hugo Touvron, Thibaut Lavige, 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. 2
- [92] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam R-CNN: Visual tracking by re-detection. In *CVPR*, 2020. 2
- [93] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4057–4066, 2023. 7, 8
- [94] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3, 6
- [95] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 2
- [96] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 6, 1, 2
- [97] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhajit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2
- [98] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2
- [99] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 7
- [100] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 256–263, 2014. 5
- [101] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xi-ang Bai. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2022. 2, 8
- [102] Jannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 2, 5

- [103] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, pages 588–605. Springer, 2022. 2, 5, 8
- [104] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Segment every reference object in spatial and temporal spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2538–2550, 2023. 2
- [105] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 1
- [106] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*, 2023. 6
- [107] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 6, 7
- [108] Ning Xu, Linjie Yang, Jianchao Yang, Dingcheng Yue, Yuchen Fan, Yuchen Liang, and Thomas S. Huang. Youtubevos dataset 2021 version. <https://youtubevos.org/dataset/vis/>. 6, 1, 2
- [109] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381, 2016. 5
- [110] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtubevos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2, 1
- [111] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022. 2
- [112] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 3, 5, 6, 7, 8, 2
- [113] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 5, 6, 8, 1
- [114] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 3, 6
- [115] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *NeurIPS*, 2022. 4
- [116] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023. 3
- [117] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019. 2
- [118] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6, 1
- [119] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 3
- [120] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2, 5, 6, 7, 1
- [121] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2, 3
- [122] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 3
- [123] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, pages 36067–36080, 2022. 6
- [124] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1282–1291, 2023. 8
- [125] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *CVPR*, 2020. 2
- [126] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Byte-track: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 2, 5
- [127] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 9
- [128] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 3

- [129] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. [2](#), [5](#)
- [130] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368. Springer, 2022. [7](#), [9](#)
- [131] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. [2](#), [5](#), [6](#)
- [132] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. [2](#), [3](#)
- [133] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022. [3](#)
- [134] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. [3](#), [6](#)
- [135] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023. [2](#), [3](#)

General Object Foundation Model for Images and Videos at Scale

Supplementary Material

In this supplementary material, we first provide more detailed information on data usage and model training in Sec 6. Subsequently, in Sec 7, we supplement additional zero-shot and fine-tuning results on classic object-level video tasks, such as VOS and RVOS. In Sec 8, detailed few-shot experimental results on the ODinW [48] benchmark are provided to validate the transferability of GLEE to various real-world tasks. Finally, in Sec 9, we showcase the results in interactive segmentation and tracking for images and videos.

6. Datasets and Implementation Details

Data Preparation. To ensure the generalization of GLEE as an object-level foundation model, we conduct joint training using a substantial amount of data with region-level annotations from both images and videos. Existing datasets exhibit variations in annotation granularity: detection datasets such as Objects365 [88] and OpenImages [46] provide bounding boxes and category names; COCO [58] and LVIS [32] offer more detailed mask annotations; RefCOCO [72, 120] and Visual Genome [44] include comprehensive object descriptions. Furthermore, video datasets [76, 86, 96, 108, 110, 113] contribute to the temporal consistency of models, and open-world data [43, 96] enrich the annotations with class-agnostic object information. A comprehensive list of the datasets we utilized, along with their respective sizes and annotation granularities, is presented in Table 6. We extracted subsets of 500,000 and 2,000,000 images from the SA1B [43] dataset for joint training in stage 2 and scale-up training respectively. To ensure that objects from SA1B are at the object-level rather than the part-level, we apply mask IoU based NMS and use area as NMS score to eliminate part-level object annotations. For GRIT [75] data, we extract 5,000,000 samples for scale-up training to enhance the richness of object descriptions.

Model and Training Details. Following the image backbone, text encoder, and visual prompter, we incorporate a 6-layer deformable transformer encoder and a 9-layer decoder to serve as our Object Decoder following MaskDINO [50]. We adopt 300 object queries, query denoising, and hybrid matching to accelerate convergence and improve performance. During the pretraining phase of stage 1, we sample data from Objects365 and OpenImages in a 1:1 ratio, with the batch size of 128 for 500,000 training iterations. Moving to stage 2, we train GLEE for 500,000 iterations on all image-level data jointly according to the ratios outlined in Table 7. For the scale-up training, we set the

dataset	Sizes		Annotations			
	images	objects	semantic	box	mask	track id
Detection Data						
Objects365 [88]	1817287	26563198	category	✓	-	-
OpenImages [46]	1743042	14610091	category	✓	-	-
LVIS [32]	100170	1270141	category	✓	✓	-
COCO [58]	118287	860001	category	✓	✓	-
BDD [118]	69863	1274792	category	✓	✓	-
Grounding Data						
RefCOCO [120]	16994	42404	description	✓	✓	-
RefCOCOg [72]	21899	42226	description	✓	✓	-
RefCOCO+ [120]	16992	42278	description	✓	✓	-
VisualGenome [44]	77396	3596689	description	✓	-	-
GRIT [75]	5117307	9090607	description	✓	-	-
OpenWorld Data						
UVQ [96]	16923	157624	-	✓	✓	-
SA1B [43]	2147712	99427126	-	✓	✓	-
Video Data						
YTVIS19 [113]	61845	97110	category	✓	✓	✓
YTVIS21 [108]	90160	175384	category	✓	✓	✓
OVIS [76]	42149	206092	category	✓	✓	✓
UVQ-dense [96]	45270	657990	-	✓	✓	✓
VOS [110]	94588	156310	-	✓	✓	✓
RefVOS [86]	93857	159961	description	✓	✓	✓

Table 6. The tasks GLEE learns to complete and the datasets used in training.

sampling ratios for SA1B and GRIT to 5.0 in Table 7, and train for an extra 500,000 iterations. We used AdamW [65] optimizer with base learning rate of 1×10^{-4} , and weight decay of 0.05, learning rate is decayed at the 400,000 iterations by a factor of 0.1. Learning rates of the image backbone and text encoder are multiplied by a factor of 0.1. For the ResNet-50 backbone and Swin backbone, we use scale augmentation [105], resizing the input images such that the shortest side is at least 480 and at most 800 pixels while the longest at most 1333. For EVA02-L backbone, we use the large-scale jittering (LSJ) [29] augmentation with a random scale sampled from range 0.1 to 2.0 followed by a fixed size crop to 1536×1536.

7. Transfer to Video Tasks

To substantiate the effectiveness of GLEE across diverse object-level video tasks, we present the performance on VOS and RVOS tasks in Table 8 and Table 9 respectively.

VOS. Video object segmentation (VOS) aims at segmenting a particular object throughout the entire video clip sequence. We evaluate GLEE on semi-supervised VOS [10] that gives the first-frame mask of the target object on YouTube-VOS 2018 [110] and MOSE [23]. Given the first-frame mask of the target object, we first crop the prompt square area from RGB image and send it to the image backbone to obtain the visual prompt feature of the corresponding area, and send it to the early fusion module be-

Datasets	OpenImages	Objects365	LVIS	VisualGenome	COCO	RefCOCO-mixed	SA1B	UV0-frame	BDD	YTVIS19	YTVIS21	OVIS	Ref-YTBVOS
Ratio	1.5	1.5	1.5	2	1.5	2.5	2.5	0.2	0.15	0.3	0.3	0.3	0.3

Table 7. The data sampling ratios during the joint-training of stage 2. RefCOCO-mixed refers to the mixed dataset of RefCOCO [120], RefCOCO+ [120], RefCOCOG [72], and the last four video datasets are treated as independent image data for training.

Method	YT-VOS 2018 val [110]				MOSE val [23]			
	\mathcal{G}	\mathcal{J}_s	\mathcal{F}_s	\mathcal{J}_u	\mathcal{F}_u	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}
Memory	STM [73]	79.4	79.7	84.2	72.8	80.9	-	-
	SWEM [60]	82.8	82.4	86.9	77.1	85.0	50.9	46.8
	STCN [18]	83.0	81.9	86.5	77.9	85.7	50.8	46.6
	XMem [17]	86.1	85.1	89.8	80.3	89.2	57.6	53.3
Non-Memory	SiamMask [95]	52.8	60.2	58.2	45.1	47.7	-	-
	Siam R-CNN [92]	73.2	73.5	-	66.2	-	-	-
	TVOS [125]	67.8	67.1	69.4	63.0	71.6	-	-
	FRTM [82]	72.1	72.3	76.2	65.9	74.1	-	-
	UNINEXT-R50 [112]	77.0	76.8	81.0	70.8	79.4	-	-
	UNINEXT-L [112]	78.1	79.1	83.5	71.0	78.9	-	-
	UNINEXT-H [112]	78.6	79.9	84.9	70.6	79.2	-	-
	GLEE-Lite	80.4	80.2	85.5	74.3	81.4	56.1	51.8
							60.4	

Table 8. Performance comparison of our GLEE on video object segmentation tasks.

fore the Transformer encoder. Then we sample fine-grained visual embeddings from the pixel embedding map M_p inside the given mask area and make them interacted with object queries through self-attention module in the Transformer decoder layer. We conduct fine-tuning of GLEE-Lite jointly on YouTube-VOS [110], YTVIS2019 [113], YTVIS2021 [108], OVIS [76], and UV0-video [96] for 40,000 iterations. The evaluation is performed on YouTube-VOS and MOSE, as shown in the Table 8. It is noteworthy that semi-supervised VOS is almost dominated by space-time memory networks [17, 18, 60, 73] which construct a memory bank for each object in the video. GLEE achieves the best results among all non-memory-based methods on YouTube-VOS and even demonstrating competitive results compared to memory-based methods on the more challenging MOSE dataset.

RVOS. Referring Video Object Segmentation (R-VOS) aims at finding objects matched with the given language expressions in a given video and segment them. Ref-YouTube-VOS [86] is a popular R-VOS benchmarks, which are constructed by introducing language expressions for the objects in the original YouTube-VOS [110] dataset. As same as semi-supervised VOS, region similarity \mathcal{J} , contour accuracy \mathcal{F} , and the averaged score $\mathcal{J} \& \mathcal{F}$ are adopted as the metrics. Given an object expression and a video, we send the description into the text encoder, select the object query with the highest confidence score and compute its mask. Additionally, we introduce temporal consistency by adding

the similarity between the 300 object queries of the current frame and the object query selected in the previous frame to the current confidence score. We directly evaluate the GLEE trained from stage 2 on Ref-YouTube-VOS. As shown in Table 9, GLEE outperforms all previous R-VOS approaches and unified method.

Method	Backbone	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}
CMSA [117]	ResNet-50	36.4	34.8	38.1
YOFO [49]		48.6	47.5	49.7
ReferFormer [102]		58.7	57.4	60.1
UNINEXT [112]		61.2	59.3	63.0
PMINet + CFBI [24]	Ensemble	54.2	53.0	55.5
CITD [56]		61.4	60.0	62.7
ReferFormer [102]	Video-Swin-B	64.9	62.8	67.0
SOC [67]		67.3	65.3	69.3
UNINEXT [112]	ConvNext-L	66.2	64.0	68.4
UNINEXT [112]	ViT-H	70.1	67.6	72.7
GLEE-Plus	Swin-L	67.7	65.6	69.7
GLEE-Pro	EVA02-L	70.6	68.2	72.9

Table 9. Performance comparison of our GLEE on Ref-YouTube-VOS task.

8. Object Detection in the Wild

To further validate transferability of GLEE on diverse real-world detection tasks, we assess its few-shot transfer ability on the ODinW [48] dataset. We vary the amount of task-specific annotated data from X-shot, providing at least X examples per category, to using all the available data in the training set, following the procedure established by GLIP [52]. We fine-tune the models on the provided data using the same hyper-parameters across all models in a full-model tuning regime. For manually designed prompts, we revise the category names for the two datasets (“Cottontail-Rabbit” to “rabbit” and “Cow/Chanterelle” to “Cow/Chanterelle mushroom”) to provide language guidance. Models train with a batch size of 4 and a learning rate of 1×10^{-4} , undergoing 200, 300, 400, 600, and 2000 iterations for the 1, 3, 5, 10, and ALL shot splits, respectively. The optimal model is selected based on the validation split for each train/val split. For each few-shot setting, we train the models three times using different random seeds for train/val splits, and the average score and standard deviation on the test split are reported, as shown in the Table 10.

Model	Shot	Tune	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	Avg
DyHead coco	1	Full	31.7 \pm 3.1	14.3 \pm 2.4	13.1 \pm 2.0	63.6 \pm 1.4	40.9 \pm 7.0	67.0 \pm 3.6	34.6 \pm 12.1	45.9 \pm 3.8	10.8 \pm 5.0	34.0 \pm 3.3	12.0 \pm 10.4	6.1 \pm 1.3	40.9 \pm 7.4	31.9 \pm 3.3
DyHead coco	3	Full	44.1 \pm 0.7	19.2 \pm 3.0	22.6 \pm 1.3	64.8 \pm 1.7	54.4 \pm 2.5	78.9 \pm 1.3	61.6 \pm 10.3	50.0 \pm 2.1	20.8 \pm 3.5	44.9 \pm 1.9	34.4 \pm 11.1	20.6 \pm 2.4	57.9 \pm 2.3	44.2 \pm 0.3
DyHead coco	5	Full	44.9 \pm 1.5	22.2 \pm 3.0	31.7 \pm 1.0	65.2 \pm 1.5	55.6 \pm 3.7	78.7 \pm 3.9	50.1 \pm 13.7	48.7 \pm 4.8	22.8 \pm 3.3	52.0 \pm 1.2	39.8 \pm 6.7	20.9 \pm 1.5	48.0 \pm 2.8	44.7 \pm 1.7
DyHead coco	10	Full	48.4 \pm 1.2	27.5 \pm 1.4	39.3 \pm 2.7	62.1 \pm 5.9	61.6 \pm 1.4	81.7 \pm 3.4	58.8 \pm 9.0	52.9 \pm 3.2	30.1 \pm 3.2	54.1 \pm 3.3	44.8 \pm 4.9	26.7 \pm 2.4	63.4 \pm 2.8	50.1 \pm 1.6
DyHead coco	All	Full	60.1	27.6	53.1	76.5	79.4	86.1	69.3	55.2	44.0	61.5	70.6	56.6	81.0	63.2
DyHead o365	1	Full	25.8 \pm 3.0	16.5 \pm 1.8	15.9 \pm 2.7	55.7 \pm 6.0	44.0 \pm 3.6	66.9 \pm 3.9	54.2 \pm 5.7	50.7 \pm 7.7	14.1 \pm 3.6	33.0 \pm 11.0	11.0 \pm 6.5	8.2 \pm 4.1	43.2 \pm 10.0	33.8 \pm 3.5
DyHead o365	3	Full	40.4 \pm 1.0	20.5 \pm 4.0	26.5 \pm 1.3	57.9 \pm 2.0	53.9 \pm 2.5	76.5 \pm 2.3	62.6 \pm 13.3	52.5 \pm 5.0	22.4 \pm 1.7	47.4 \pm 2.0	30.1 \pm 6.9	19.7 \pm 1.5	57.0 \pm 2.3	43.6 \pm 1.0
DyHead o365	5	Full	43.5 \pm 1.0	25.3 \pm 1.8	35.8 \pm 0.5	63.0 \pm 1.0	56.2 \pm 3.9	76.8 \pm 5.9	62.5 \pm 8.7	46.6 \pm 3.1	28.8 \pm 2.2	51.2 \pm 2.2	38.7 \pm 4.1	21.0 \pm 1.4	53.4 \pm 5.2	46.4 \pm 1.1
DyHead o365	10	Full	46.6 \pm 0.3	29.0 \pm 2.8	41.7 \pm 1.0	65.2 \pm 2.5	62.5 \pm 0.8	85.4 \pm 2.2	67.9 \pm 4.5	47.9 \pm 2.2	28.6 \pm 5.0	53.8 \pm 1.0	39.2 \pm 4.9	27.9 \pm 2.3	64.1 \pm 2.6	50.8 \pm 1.3
DyHead o365	All	Full	53.3	28.4	49.5	73.5	77.9	84.0	69.2	56.2	43.6	59.2	68.9	53.7	73.7	60.8
GLIP-T	1	Full	54.8 \pm 2.0	18.4 \pm 1.0	33.8 \pm 1.1	70.1 \pm 2.9	64.2 \pm 1.8	83.7 \pm 3.0	70.8 \pm 2.1	56.2 \pm 1.8	22.9 \pm 0.2	56.6 \pm 0.5	59.9 \pm 0.4	18.9 \pm 1.3	54.5 \pm 2.7	51.1 \pm 0.1
GLIP-T	3	Full	58.1 \pm 0.5	22.9 \pm 1.3	40.8 \pm 0.9	65.7 \pm 1.6	66.0 \pm 0.2	84.7 \pm 0.5	65.7 \pm 2.8	62.6 \pm 1.4	27.2 \pm 2.7	61.9 \pm 1.8	60.7 \pm 0.2	27.1 \pm 1.2	70.4 \pm 2.5	54.9 \pm 0.2
GLIP-T	5	Full	59.5 \pm 0.4	23.8 \pm 0.9	43.6 \pm 1.4	68.7 \pm 1.3	66.1 \pm 0.6	85.4 \pm 0.4	72.3 \pm 0.0	62.1 \pm 2.0	27.3 \pm 1.2	61.0 \pm 1.8	62.7 \pm 1.0	34.5 \pm 0.6	66.6 \pm 2.3	56.4 \pm 0.4
GLIP-T	10	Full	59.1 \pm 1.3	26.3 \pm 1.1	46.3 \pm 1.6	67.3 \pm 1.5	67.1 \pm 0.7	87.8 \pm 0.5	72.3 \pm 0.0	57.7 \pm 1.7	34.6 \pm 1.7	65.4 \pm 1.4	61.6 \pm 1.0	39.3 \pm 1.0	74.7 \pm 2.3	58.4 \pm 0.2
GLIP-T	All	Full	62.3	31.2	52.5	70.8	78.7	88.1	75.6	61.4	51.4	65.3	71.2	58.7	76.7	64.9
GLIP-L	1	Full	64.8 \pm 0.6	18.7 \pm 0.6	39.5 \pm 1.2	70.0 \pm 1.5	70.5 \pm 0.2	69.8 \pm 18.0	70.6 \pm 4.0	68.4 \pm 1.2	71.0 \pm 1.3	65.4 \pm 1.1	68.1 \pm 0.2	28.9 \pm 2.9	72.9 \pm 4.7	59.9 \pm 1.4
GLIP-L	3	Full	65.6 \pm 0.6	22.3 \pm 1.1	45.2 \pm 0.4	72.3 \pm 1.4	70.4 \pm 0.4	81.6 \pm 13.3	71.8 \pm 0.3	65.3 \pm 1.6	67.6 \pm 1.0	66.7 \pm 0.9	68.1 \pm 0.3	37.0 \pm 1.9	73.1 \pm 3.3	62.1 \pm 0.7
GLIP-L	5	Full	66.6 \pm 0.4	26.4 \pm 2.5	49.5 \pm 1.1	70.7 \pm 0.2	71.9 \pm 0.2	88.1 \pm 0.0	71.1 \pm 0.6	68.8 \pm 1.2	68.5 \pm 1.7	70.0 \pm 0.9	68.3 \pm 0.5	39.9 \pm 1.4	75.2 \pm 2.7	64.2 \pm 0.3
GLIP-L	10	Full	66.4 \pm 0.7	32.0 \pm 1.4	52.3 \pm 1.1	70.6 \pm 0.7	72.4 \pm 0.3	88.1 \pm 0.0	67.1 \pm 3.6	64.7 \pm 3.1	69.4 \pm 1.4	71.5 \pm 0.8	68.4 \pm 0.7	44.3 \pm 0.6	76.3 \pm 1.1	64.9 \pm 0.7
GLIP-L	All	Full	69.6	32.6	56.6	76.4	79.4	88.1	67.1	69.4	65.8	71.6	75.7	60.3	83.1	68.9
GLEE-Lite	1	Full	61.3 \pm 0.5	19.2 \pm 3.1	27.2 \pm 3.4	70.8 \pm 3.3	52.8 \pm 15.1	70.7 \pm 7.5	49.2 \pm 22.0	58.1 \pm 5.4	28.8 \pm 11.0	57.9 \pm 10.0	57.7 \pm 0.6	22.2 \pm 7.9	57.0 \pm 4.5	48.7 \pm 0.9
GLEE-Lite	3	Full	62.6 \pm 0.1	25.5 \pm 3.8	29.1 \pm 1.5	72.9 \pm 4.1	65.8 \pm 1.7	83.0 \pm 4.4	66.8 \pm 3.4	61.7 \pm 10.4	40.0 \pm 3.0	61.2 \pm 3.5	44.9 \pm 12.9	26.7 \pm 3.5	64.5 \pm 6.8	54.2 \pm 2.3
GLEE-Lite	5	Full	62.8 \pm 0.4	28.0 \pm 3.1	33.8 \pm 2.2	71.7 \pm 2.7	64.0 \pm 4.4	81.6 \pm 4.1	64.9 \pm 5.2	60.1 \pm 12.4	39.1 \pm 1.0	59.7 \pm 3.0	49.2 \pm 14.5	30.8 \pm 1.3	69.2 \pm 7.8	55.0 \pm 3.7
GLEE-Lite	10	Full	62.1 \pm 0.9	32.0 \pm 1.6	39.3 \pm 2.0	71.2 \pm 1.5	64.4 \pm 2.7	88.0 \pm 2.7	64.3 \pm 9.8	65.5 \pm 1.5	36.4 \pm 4.2	62.1 \pm 3.4	54.8 \pm 10.9	38.8 \pm 1.2	70.6 \pm 4.0	57.7 \pm 6.6
GLEE-Lite	All	Full	62.8	37.9	52.9	73.6	76.5	88.9	69.7	65.0	51.1	58.9	67.4	57.2	82.3	64.9
GLEE-Plus	1	Full	68.2 \pm 2.2	20.4 \pm 0.2	43.9 \pm 4.1	75.5 \pm 1.6	68.4 \pm 2.7	50.6 \pm 29.0	47.3 \pm 0.8	70.4 \pm 4.0	64.6 \pm 0.5	67.7 \pm 1.5	62.3 \pm 1.0	30.0 \pm 9.2	71.6 \pm 7.7	57.0 \pm 0.8
GLEE-Plus	3	Full	70.6 \pm 0.9	24.8 \pm 2.1	47.6 \pm 0.8	79.5 \pm 0.7	69.0 \pm 2.0	83.1 \pm 5.9	66.2 \pm 1.3	75.6 \pm 3.5	65.3 \pm 1.1	69.0 \pm 0.8	65.7 \pm 4.2	38.1 \pm 3.1	76.3 \pm 4.6	63.9 \pm 1.2
GLEE-Plus	5	Full	69.9 \pm 0.9	29.6 \pm 2.9	48.8 \pm 1.2	75.0 \pm 1.7	67.7 \pm 5.1	83.6 \pm 9.9	68.5 \pm 3.2	71.6 \pm 5.9	61.6 \pm 4.0	67.7 \pm 0.8	66.8 \pm 4.5	38.8 \pm 1.1	78.9 \pm 1.0	63.7 \pm 1.0
GLEE-Plus	10	Full	69.3 \pm 1.2	32.5 \pm 1.9	50.8 \pm 0.9	76.4 \pm 0.6	70.7 \pm 0.9	88.2 \pm 1.2	68.9 \pm 3.3	68.2 \pm 3.0	60.0 \pm 1.9	69.3 \pm 1.5	62.6 \pm 0.3	41.7 \pm 3.1	81.7 \pm 1.7	64.6 \pm 1.7
GLEE-Plus	All	Full	70.4	34.8	54.1	76.4	74.5	89.7	68.6	67.6	57.8	69.2	71.4	57.1	82.9	67.3
GLEE-Pro	1	Full	70.9 \pm 1.7	24.5 \pm 2.3	46.7 \pm 0.4	76.4 \pm 0.8	68.2 \pm 3.8	60.4 \pm 7.8	58.9 \pm 2.7	68.2 \pm 4.5	58.5 \pm 8.8	67.6 \pm 0.8	69.2 \pm 0.2	31.8 \pm 2.4	70.8 \pm 7.6	59.4 \pm 1.5
GLEE-Pro	3	Full	72.3 \pm 0.4	28.4 \pm 0.5	49.6 \pm 2.2	76.1 \pm 1.3	69.3 \pm 3.9	79.4 \pm 9.5	67.4 \pm 3.5	74.1 \pm 4.9	63.7 \pm 2.0	68.4 \pm 0.6	68.3 \pm 2.1	42.1 \pm 5.3	76.9 \pm 1.6	64.3 \pm 1.3
GLEE-Pro	5	Full	71.4 \pm 0.9	33.4 \pm 1.5	50.6 \pm 4.3	73.8 \pm 3.9	71.9 \pm 0.3	83.6 \pm 6.8	66.6 \pm 1.8	72.5 \pm 4.3	59.1 \pm 4.8	68.7 \pm 1.4	69.7 \pm 1.3	39.5 \pm 4.8	77.4 \pm 3.2	64.5 \pm 0.9
GLEE-Pro	10	Full	71.1 \pm 1.9	37.8 \pm 2.1	54.2 \pm 1.2	73.9 \pm 7.2	70.7 \pm 1.3	90.9 \pm 1.4	66.0 \pm 9.4	73.9 \pm 6.8	57.8 \pm 3.9	69.4 \pm 0.9	62.9 \pm 6.3	44.3 \pm 3.8	79.8 \pm 0.6	65.6 \pm 0.4
GLEE-Pro	All	Full	72.6	36.5	58.1	80.5	74.1	92.0	67.0	76.5	66.4	70.5	66.4	55.7	80.6	69.0

Table 10. Per-dataset performance compared with DyHead, GLIP-T, and GLIP-L. For PascalVOC, we report the mAP (IoU=0.50:0.95) using the COCO evaluation script, to be consistent with other 12 datasets. “Full” denotes full-model tuning.

9. Interactive Segmentation and Tracking

As described in Sec 7, GLEE achieves interactive segmentation and tracking by introducing a visual prompt. Sending points, boxes, or scribbles along with the image to the model enables the segmentation of specified objects. Moreover, by feeding the mask from the previous frame and its corresponding prompt feature into early fusion and self-attention, GLEE performs segmentation in the current frame based on the segmentation results from the previous frame. The features of objects in the previous frame serve as referring features at this point. As illustrated in the Figure 6, we showcase the interactive segmentation results of different prompts on images and videos. Please visit our project homepage to experience more custom interactive image and video segmentation effects through our online demo.



Figure 6. The visualization results of interactive segmentation and tracking. For image-level interactive segmentation, GLEE supports sending points, boxes, or scribbles as a visual prompts to the model, enabling direct segmentation of the specified object. In the case of video object segmentation, using the masked feature from the first frame as a prompt referring features allows segmentation of the corresponding object in subsequent frames of the video.