Visual In-Context Prompting

Feng Li♠, Qing Jiang♯, Hao Zhang♠, Tianhe Ren†, Shilong Liu¶, Xueyan Zou§, Huaizhe Xu♠, Hongyang Li♯, Jianwei Yang‡1, Chunyuan Li‡, Lei Zhang†2, Jianfeng Gao‡2

♠ HKUST ‡ Microsoft Research, Redmond † IDEA ♯ SCUT ¶ Tsinghua § UW-Madison

1. Project Lead 2. Equal Advisory Contribution

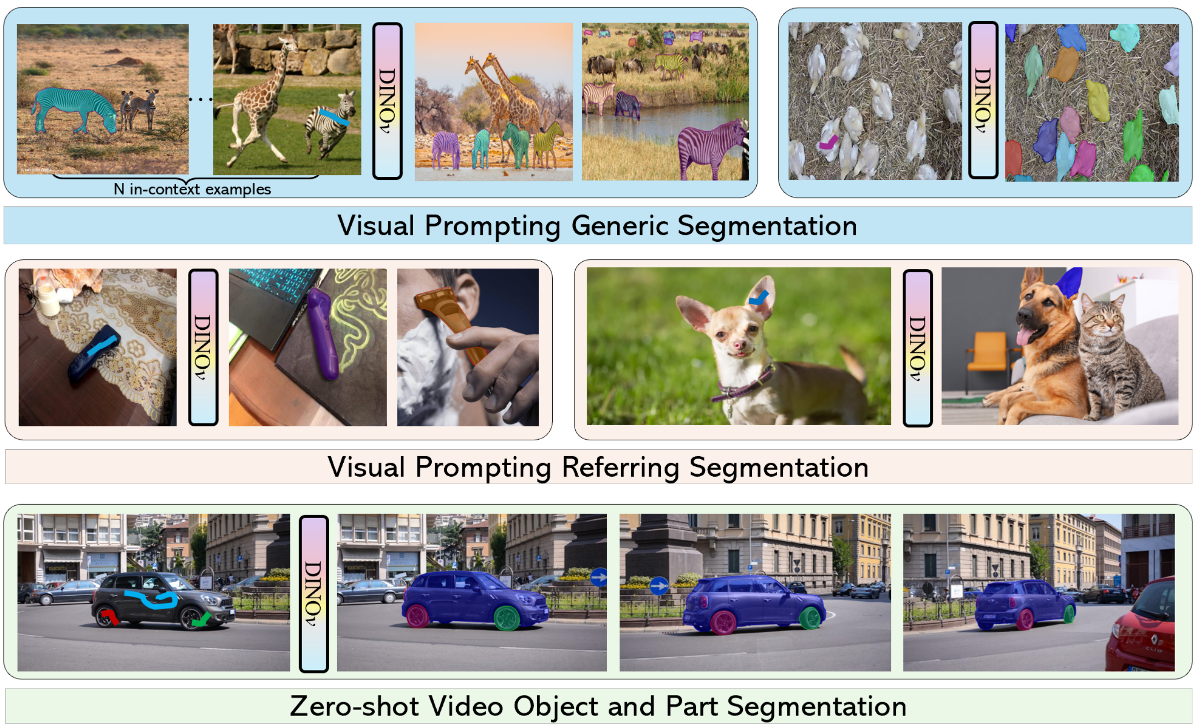


Figure 1. Our model *DINOv* supports generic and referring segmentation to associate multiple or single objects with the user input visual prompts. A user can input one or more in-context visual prompts (scribbles, masks, boxes, etc.) to improve the segmentation performance.

Abstract

In-context prompting in large language models (LLMs) has become a prevalent approach to improve zero-shot capabilities, but this idea is less explored in the vision domain. Existing visual prompting methods focus on referring segmentation to segment the most relevant object, falling short of addressing many generic vision tasks like open-set segmentation and detection. In this paper, we introduce a universal visual in-context prompting framework for both tasks, as shown in Fig. 1. In particular, we build on top of an encoder-decoder architecture, and develop a versatile prompt encoder to support a variety of prompts like strokes, boxes, and points. We further enhance it to take an arbitrary number of reference image segments as the context. Our extensive explorations show that the proposed visual in-context prompting elicits extraordinary referring and generic segmentation capabilities to refer and detect, yielding competitive performance to close-set in-domain datasets and showing promising results on many open-set segmentation datasets. By joint training on COCO and SA-1B, DINOv achieves 57.7 *PQ on COCO and* 23.2 *PQ on ADE20K. Code will be available at* `https://github.com/UX-Decoder/DINOv`

1. Introduction

The recent progress in large language models (LLMs) like GPT [1, 27] has shown promising results towards artificial general intelligence (AGI) by training unified models on large amounts of text data. These giant LLMs manifest themselves with intriguing emerging capabilities such as in-context learning. Nevertheless, similar paradigms have not yet succeeded in solving all vision tasks due to the diversity of scenarios in computer vision [15]. Some works [23, 49] have combined LLMs and vision models to tackle complex image understanding tasks with text outputs such as visual question answering, but challenges remain in fine-grained tasks that require pixel-level outputs, like instance masks, rather than just text.

To this end, the community has observed a growing interest in the development of language-enhanced vision foundation models. These models demonstrate profound competencies in open-world visual understanding tasks using text prompts, encompassing areas like open-set detection [19, 24, 44] and segmentation [38, 42, 44, 50]. Visual prompt, a different prompting mechanism has been explored in some recent segmentation models [13, 17, 51]. In these works, different visual prompting formats (*e.g.*, points, boxes and strokes, *etc*) have been explored to facilitate the segmentation of visual contents specified by users.

Another critical LLM technique, in-context learning, has been less explored. In-context learning specifies the new task instruction using examples, and allows models to adapt to new tasks or domains without explicit retraining by providing relevant examples. One pioneering work in this area is SegGPT [34], which demonstrates the ability to output an image mask based on visual examples. However, these works focus on associating a user visual prompt with one most relevant object and have the limited ability to identify multiple objects of the same semantic concept. Therefore, these approaches fall short of addressing many generic vision tasks like open-set object detection and segmentation, which often require the segmentation of multiple objects of a given concept. On the other hand, while text-prompted vision models [24, 50] do not align with in-context learning methodologies, they exhibit notable flexibility in managing both referring and generic tasks in detection or segmentation, but with language prompts. In this paper, we strive to develop a model that supports *visual in-context prompting* for all types of image segmentation tasks. A comparison between our work and previous work is shown in Fig. 2. Besides supporting both single-image and cross-image visual prompting, our model distinguishes itself by effectively handling both referring and generic segmentation problems.

To achieve this goal, we build a model called ***DINOv*** to support versatile visual prompting capabilities, based on the unified detection and segmentation model MaskDINO [18]. *DINOv* follows the general encoder-decoder design with an extra prompt encoder to formulate and sample visual prompts. The decoder takes in segmentation queries and reference prompt queries to generate segmentation masks and target visual prompts, and we associate the output segmentation masks with the target prompt queries for the final output. We can define the visual in-context samples with a set of *reference image (Q) - visual prompt (A)* pairs. The visual prompt can be in various types, including mask, scribble, box, etc. With the in-context examples, our model takes in a target image and outputs the masks. The creation of target visual prompts involves a prompt encoder that extracts reference visual prompts from a Q-A pair. This is followed by a decoder to get the target visual prompt by attending reference visual prompts to the target image. During training, to construct positive and negative samples for generic segmentation, we sample reference visual prompts in a batch across different images. To address task and data discrepancies, we formulate generic latent queries and point queries for generic and referring segmentation, respectively. By joint training on COCO [21] and SA-1B [13] for generic and referring segmentation, our model attains competitive performance on in-domain segmentation tasks compared with text-prompted models and shows promising generalization capability on a wide range of open-set segmentation benchmarks using purely visual prompts.

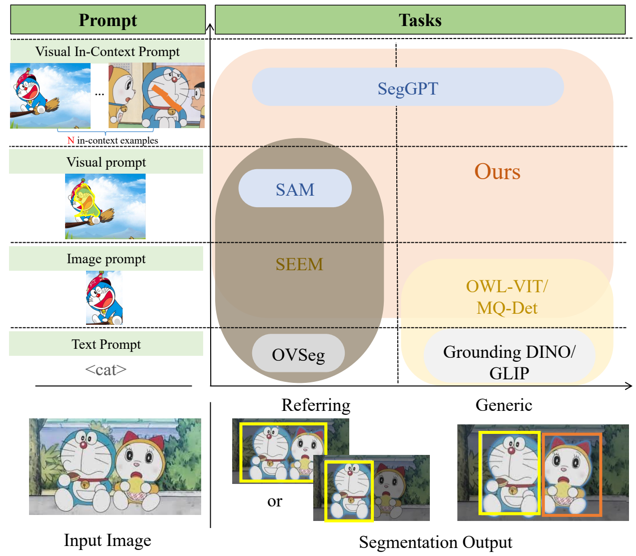
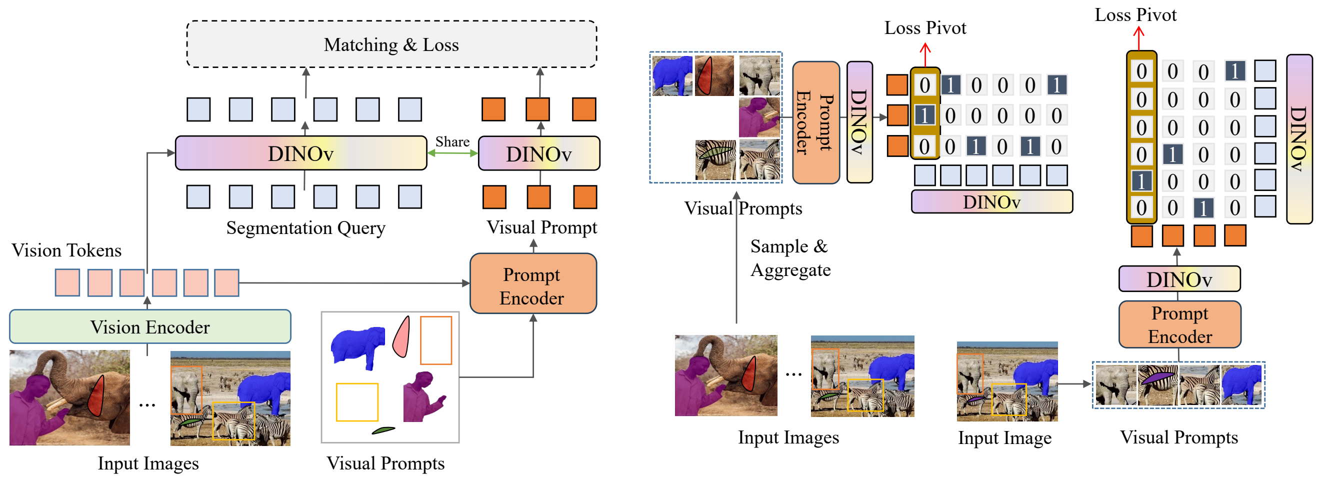


Figure 2. Comparison with related works. Generic: segment all objects of the same semantic concept that match the user prompt. Refer: segment a particular object with the user input visual prompts. Image prompt: crop the image regions as prompts. (single) Visual prompt: one image-prompt example to segment. In-context prompt: one or multiple image-prompt examples. We can do single-image and cross-image visual prompting tasks and support referring and generic segmentation.

To summarize, our contributions are threefold:

- We are the first to extend visual in-context prompting to support generic vision tasks like open-set generic segmentation and detection, and achieve comparable performance with text prompt-based open-set models.



(a) General Framework (b) Visual Generic Segmentation (c) Visual Referring Segmentation

Figure 3. *DINOv* is a universal segmentation framework that can do generic segmentation and referring image segmentation. The vision encoder is used to extract image features.(b) An illustration of losses for visual generic segmentation. In the example, there are 6 visual prompts sampled from 6 masks from 3 categories. The visual prompts from the instances of the same class are averaged as the class embedding. Each colume of the matching matrix is a 3-dimension one-hot vector which is a one-hot class label of the instance; (c) An illustration of losses for visual referring segmentation. Each visual prompt is classified to one of the 6 instances.

- We build *DINOv*, a unified framework for referring segmentation and generic segmentation based on visual in-context prompting. This unification simplifies model design and allows our model to consume both semantically-labelled and unlabelled data for better performance.
- We conduct extensive experiments and visualizations to show that our model can handle generic, referring, and video object segmentation tasks. Our early attempts exhibit promising results on open-set segmentation and detection with visual prompting.

2. Method

2.1. Unified Formulation for Segmentation Tasks

In this paper, we concentrate on visual prompting tasks involving images, encompassing both generic segmentation and referring segmentation tasks. Given N reference images $\mathcal{I} = \{\mathbf{I}_1, ..., \mathbf{I}_N\} \in \mathcal{R}^{N\times H\times W\times 3}$ with the corresponding visual prompts $\mathcal{P} = \{p_1, ..., p_N\}$, *DINOv* aims to segment objects of interest on a new target image $\mathbf{I_t}$. The visual prompts include masks, boxes, scribbles, points, etc. The interested objects can be a particular object for referring segmentation or all objects of the same semantic concept for generic segmentation. Note that the reference image can be identical to the target image, in which the task reduces to single-image visual prompting segmentation.

To address these tasks, *DINOv* utilizes a comprehensive query-based encoder-decoder architecture. This architecture comprises a vision encoder, denoted as $\mathbf{Enc}$, responsible for extracting image features, a prompt encoder referred to as $\mathbf{PromptEncoder}$, designed to extract visual prompt features by combining image features and user-provided visual prompts, and a general decoder represented as $\mathbf{Decoder}$, which generates masks and visual concepts based on the segmentation query and visual prompt features. Upon receiving the input image and user-provided visual prompts, our initial step involves extracting image features denoted as $\mathbf{Z}$ using the vision encoder. Subsequently, we feed both the image features and visual prompts into the prompt encoder to extract the *reference visual prompt* $\mathcal{F}$ and subsequently sample the *query visual prompt* features $\mathbf{Q_p}$.Formally, we have:

$$\begin{aligned} \mathcal{Z} &= \mathbf{Enc}(\mathcal{I}), \mathbf{Z} = \mathbf{Enc}(\mathbf{I_t}) \\ \mathcal{F} &= \mathbf{PromptEncoder}(\mathcal{P}, \mathcal{Z}) \\ \mathbf{Q_p} &= \mathbf{PromptSample}(\mathcal{F}) \end{aligned} \tag{1}$$

In addition to the visual prompt features $\mathbf{Q_p}$, *DINOv* incorporates segmentation queries $\mathbf{Q_s}$ for proposal extraction. A shared decoder is employed to decode outputs for both $\mathbf{Q_s}$ and $\mathbf{Q_p}$ while performing cross-attention with respect to the target image feature $\mathbf{Z}$.

$$\begin{aligned} \mathbf{O_s} &= \mathbf{Decoder}\left(\mathbf{Q_s}; \mathbf{Z}\right) \\ \mathbf{O_p} &= \mathbf{Decoder}\left(\mathbf{Q_p}; \mathbf{Z}\right) \\ \langle \mathbf{M}, \mathbf{B}\rangle &= \mathbf{MaskHead}(\mathbf{O_s}) \\ \mathbf{C_g}, \mathbf{C_r} &= \mathbf{PromptClassifier}(\mathbf{O_s}, \mathbf{O_p}) \end{aligned} \tag{2}$$

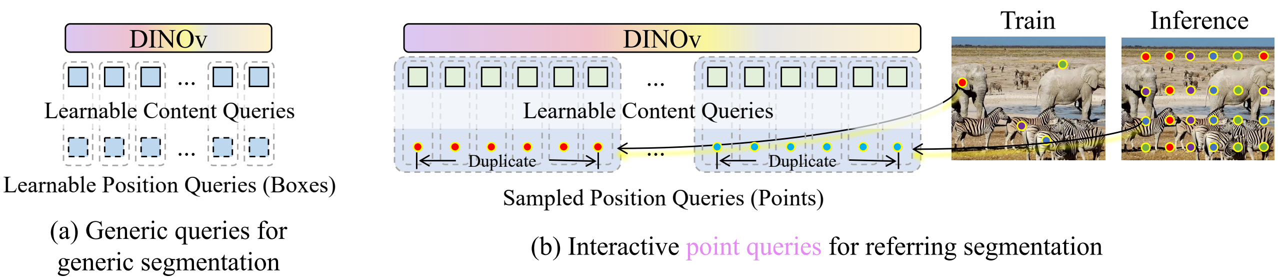


Figure 4. *DINOv* query formulation of generic and referring segmentation tasks.

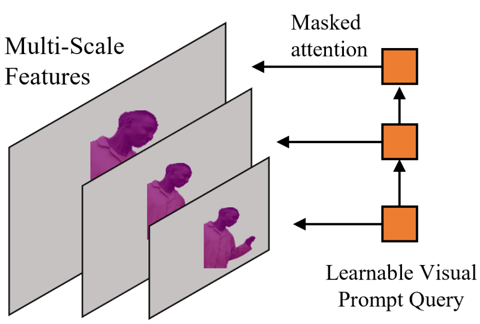


Figure 5. Prompt encoder to encode visual prompt from reference images. We use three masked cross-attention from the vision encoder small feature map to large feature map.

Here, $\mathbf{O_s}$ represents the decoded segmentation query features, $\mathbf{O_p}$ corresponds to the decoded *target visual prompt* features, while $\mathbf{M}$ and $\mathbf{B}$ denote the predicted masks and boxes, respectively. Furthermore, we have $\mathbf{C_g}$ and $\mathbf{C_r}$ as the predicted matching scores for generic segmentation and referring segmentation tasks. These scores are derived through the use of a PromptClassifier, which computes the similarity between $\mathbf{O_s}$ and $\mathbf{O_p}$.

PromptClassifier. We clarify the definition of the prompt classifier, denoted as **PromptClassifier**$(\cdot, \cdot)$, for both generic segmentation and referring segmentation tasks here. In the case of generic segmentation tasks like instance and panoptic segmentation, the typical objective is to classify object features $\mathbf{O_s}$ into respective categories. When employing visual prompting for generic segmentation tasks, the distinction lies in the utilization of visual prompt features $\mathbf{O_p}$ as class embeddings. This is illustrated in the following equation:

$$\mathbf{C_g} = \mathbf{g}(\mathbf{O_s})\mathbf{g}(\mathbf{O_p}^T), \mathbf{C_g} \in N_p \times N_s \tag{3}$$

where N_p and N_s are the number of visual prompts and object features. g is the linear projection for generic segmentation task. Each of N_s objects is classified into one of N_p classes. For visual referring segmentation, the objective differs. Here, each visual prompt is employed to identify the most closely matched instance within the target image. This task can be framed as a classification problem, where each visual prompt is assigned to a specific instance within the target image. It's important to note that during our training phase, the target image and the reference image are identical. The matching score matrix for referring segmentation is structured as follows:

$$\mathbf{C_r} = \mathbf{h}(\mathbf{O_p})\mathbf{h}(\mathbf{O_s}^T), \mathbf{C_r} \in N_s \times N_q \tag{4}$$

h is the linear projection for referring segmentation task. Fig. 3(b) and (c) provide an illustrative representation of the two tasks. In our implementation, the generic segmentation task involves finding the most suitable visual prompt for each mask proposal, effectively pivoting the loss from a query to all prompts. Conversely, the referring segmentation task focuses on matching a given visual prompt to a specific mask proposal, with the loss pivot transitioning from a prompt to all proposals. As indicated in Equations 3 and 4, the **PromptClassifier** for both generic and referring segmentation tasks share a similar formulation. Consequently, they can share the entire framework, except for the two distinct linear layers denoted as $\mathbf{g}$ and $\mathbf{h}$.

2.2. Visual Prompt Formulation

The heart part of our *DINOv* is the proposed visual prompting mechanism. As shown in Eq. 1 and Eq. 2, we employ two modules to get the final visual prompt:

- A **PromptEncoder** to encode the *reference visual prompt* $\mathcal{F}$ from the reference image features (followed by a sampling process to get *query visual prompt* $\mathbf{Q_p}$).
- A **Decoder** (shared with the segmentation decoder) to decode outputs for the *target visual prompt* $\mathbf{O_p}$ by interacting with the target image features.

This design allows our model to first encode the *reference visual prompt* and then adapt the prompt to the target image in a flexible way. As we attempt to express visual concepts through visual prompts, a straightforward way is to employ a pre-trained vision encoder (*e.g.*, CLIP [29]) to process the reference images guided by user prompts [26]. However, it may encounter several challenges: (i) the vision encoder takes cropped images as inputs, which causes substantial domain shift, especially for small objects [47]; (ii) The visual features extracted from CLIP tend to be more semantic and may not meet the demands in VOS tasks. As we will show in our ablation study, employing a CLIP vision encoder to extract visual prompts has a clear inferior generalization ability.

To address these issues, we reuse the vision encoder in our model and develop a simple yet effective prompt encoder. It extracts visual features corresponding to the locations indicated by various forms of visual prompts. To capture visual details of different granularities, we have incorporated multiple layers (default to 3) of the Mask Cross Attention Layer, as shown in Fig. 5. Each layer takes the image features extracted at different levels (output multi-scale features from the vision encoder, ranging from lower to higher resolutions) as inputs, utilizes the regions defined by the visual inputs as masks, and employs learnable queries to process the features at the corresponding positions to get the visual prompt features.

2.3. Prompt Sampling

We introduce two prompt sampling strategies tailored for referring segmentation and generic segmentation.

Referring segmentation. In the case of referring segmentation, we employ a "self-referring" approach during training, wherein the reference image is identical to the target image. Here, we sample a prompt from an instance and train the model to refer to the same instance. This approach allows us to leverage extensive segmentation data, such as SA-1B, for training our model effectively. Despite being trained on the same instances, our model demonstrates the capability to perform cross-image referring during inference. As illustrated in Fig. 3(c), we can change the target images to various different images, enabling the model to effectively engage in cross-image referring tasks.

Generic segmentation. The sampling strategies are slightly different during training and inference:

- Training. In the training process, it is crucial to create both positive and negative visual prompt samples. To achieve this, we generate visual prompts by utilizing a large image training batch. As depicted in Algorithm 1, our approach begins by grouping together **reference visual prompt** $\mathcal{F}$ of the same semantic category across all images within a training batch. For each semantic category, we then randomly select a variable number of in-context examples, ranging from 1 to N, and perform an aggregation process to generate *reference visual prompt* tokens $\mathbf{Q_p}$, where each *reference visual prompt* token corresponds to a specific semantic category. $\mathbf{Q_p}$ is subsequently fed into the decoder, where it interacts with the target image to produce the final *target visual prompt* $\mathbf{O_p}$. Consequently, the number of semantic categories corresponds to the number of target visual prompts. It's important to note that a given batch of images may not encompass all semantic categories present in the dataset, resulting in a variable number of semantic categories during the training process.

- Inference. During the inference stage, using the COCO Dataset as an example, we pre-extract the respective visual prompt features based on mask prompts for all semantic categories established during the training phase. For evaluation purposes, we adopt a random selection approach, where we choose $\mathbf{N}$ features for each semantic category. By default, $\mathbf{N}$ is set to 16. These selected features act as representative visual prompt features for each category. This practice ensures that our inference stage maintains the same number of categories as in traditional open-set evaluation, effectively preventing any potential information leakage.

We also provide the pseudo-code of prompting sampling in Appedix E.

2.4. Decoder Query Formulation

In *DINOv*, we designed two types of segmentation queries to address two different tasks as depicted in Fig. 4. For generic segmentation, the query is a number of learnable ones similar to MaskDINO [16]. For the visual referring task, we adopt the interactive point query following Semantic-SAM [17], so that we can exploit the rich granularities in SA-1B [13]. Similar to Semantic-SAM, the visual prompts (points or boxes) are both converted into anchor box format, and then the position of each visual prompt will be encoded into position queries. Each position query is duplicated and subsequently combined with content queries of different granularities as the final segmentation queries. For the training on SA-1B, in order to avoid excessive computational overhead on the model, we selectively sample a subset of points contained within this visual concept as positive point queries. Concurrently, we randomly sample a subset of points from the remaining areas to serve as negative points. During the inference stage, we sample the initial point position queries on 20×20 uniformly distributed grid as the initial point position for a single frame.

3. Experiments

We jointly train on SA-1B [13] and COCO [21]. We evaluate on open-set segmentation and video object segmentation. We provide the detailed experiment setting, implementation details, and metrics in Appendix. A, B, and C.

3.1. Generic Segmentation and Detection

We evaluate our visual prompt based generic segmentation performance in Table 1.

In-domain Segmentation on COCO. Compared to other models trained for visual prompts, we achieve significantly better results. For example, we surpass SegGPT [34] and Painter [33] by 14.3 PQ and 25.5 PQ. In addition, With just a few visual in-context prompts for each category, our

Table 1. **One suit of weights** for generic visual in-context segmentation on multiple datasets. Our model is trained on COCO and SA-1B data. Note: "–" denotes the model does not have number reported or does not have the ability for the specific task. $\star$ means it is the test set results. † FC-CLIP adopts a frozen CLIP for open-set (text), we prompt the FC-CLIP with CLIP visual features to simulate visual promoting. # FC-CLIP and ODISE rely on frozen CLIP and Stable Diffusion knowledge. Mask DINO [18] is our baseline for comparison.

Method	Semantic Data	Type	COCO (in-domain)				ADE (out-domain)				SegInW (out-domain)	
			PQ	mask AP	box AP	mIoU	PQ	mask AP	box AP	mIoU	AP-Average	AP-Median
Mask2Former-T [2]	COCO		53.2	43.3	46.1	63.2	–	–	–	–	–	
Mask2Former-B [2]	COCO		56.4	46.3	49.5	67.1	–	–	–	–	–	–
Mask2Former-L [2]	COCO		57.8	48.6	52.1	67.4	–	–	–	–	–	–
OneFormer-L [9]	COCO	Closed-set	57.9	48.9	–	67.4	–	–	–	–	–	–
MaskDINO-L [16]	COCO		58.3	50.6	56.2	67.5	–	–	–	–	–	–
Pano/SegFormer-B [36]	COCO		55.4	–	–	–	–	–	–	–	–	–
GLIPv2-H [45]	COCO+O365+GOLDG+...		–	48.9*	–	–	–	–	–	–	–	–
MaskCLIP (L) [5]	YFCC100M		–	–	–	–	–	15.1	6.0	–	23.7	–
#ODISE-H [37]	COCO (Stable diffusion))		45.6	38.4	–	52.4	23.4	13.9	–	28.7	–	–
#FC-CLIP-L [42]	COCO (CLIP)	Text Open-set	54.4	44.6	–	63.7	26.8	16.8	–	34.1	–	–
OpenSeed-T [44]	COCO+O365		55.4	47.6	52.0	64.0	19.8	14.1	17.0	22.9	33.9	21.5
X-Decoder-T [50]	COCO+CC3M+..		51.4	40.5	43.6	62.8	18.8	9.8	–	25.0	22.7	15.2
X-Decoder-L [50]	COCO+CC3M+..		56.9	46.7	–	67.5	21.8	13.1	–	29.6	36.1	38.7
OpenSeed-L [44]	COCO+O365		59.5	53.2	58.2	68.6	19.7	15.0	17.7	23.4	36.1	38.7
FC-CLIP†-L [42]	COCO		–	–	–	–	2.3	4.1	–	7.8	–	–
SegGPT-L [34]	COCO+ADE+VOC+..		43.4	–	–	–	–	–	–	–	–	–
Painter-L [33]	COCO+ADE+NYUv2	Visual Prompt	34.4	–	–	–	–	–	–	–	–	–
DINOv-T (Ours)	COCO		49.0	41.5	45.2	57.0	19.4	11.4	12.8	21.9	39.5	41.6
DINOv-L (Ours)	COCO		57.7	50.4	54.2	66.7	23.2	15.1	14.3	25.3	40.6	44.6

Table 2. **One suit of weights** on ODinW benchmark. Average and median AP across 35 datasets are reported for simplicity.

Model	Pretrain Data	Average	Median
MDETR [12]	GOLDG, REFC	10.7	3.0
GLIP-T [19]	Object365	11.4	1.6
OpenSeed (T) (Ours)	Object365, COCO	14.2	3.1
OpenSeed (L) (Ours)	Object365, COCO	15.2	5.0
DINOv (T) (Ours)	COCO, SAM	14.9	5.4
DINOv (L) (Ours)	COCO, SAM	15.7	4.78

model achieves comparable results with previous close-set or open-set models on COCO. For example, the panoptic segmentation performance gap between *DINOv* and our baseline Mask DINO is only 0.6 PQ (57.7 PQ vs 58.3 PQ).

Out-domain open-set segmentation on ADE20K. After training with visual prompt on COCO and SAM, we do zero-shot evaluation on ADE20K to validate its open-set segmentation capability when seeing novel visual concepts. To our best knowledge, it is the first time to use visual prompt for open-set segmentation. Compared with previous text-prompted open-set models, we achieve comparable or better performance with only COCO semantic data and no semantic knowledge from large pre-trained models. Especially, compared with our baseline OpenSeed, we achieve better performance with much fewer data. Note that FC-CLIP [42] employs a frozen CLIP to do text-based open-set segmentation. As the text and visual features are aligned in CLIP, we also attempt to prompt a pre-trained FC-CLIP with visual features from CLIP to test its open-set ability with visual prompts. However, its visual prompting performance largely lags behind its text-prompted results. Therefore, it is non-trivial to transfer a multi-modal text-based open-set model to do visual-prompted recognition well. The results indicate that visual prompts can generalize well to new concepts.

Segmentation and detection in the wild. We also validate the generalization capability of visual prompting on some diversified and domain-specific datasets including SegInW and ODinW, which in total encompass more than 60 datasets. These datasets contain many real-scenario or rare categories. As these datasets all focus on instance-level segmentation, we report the average and median AP (AP-Average and AP-Median) over all datasets. We first evaluate the Segmentation in the Wild (SegInW) benchmark, which consists of 25 datasets. With visual prompting, *DINOv* achieves a significant performance improvement over our baseline OpenSeed. For example, Our best AP-Average exceeds OpenSeed by 4.5 AP. We further evaluate Object Detection in the Wild (ODinW), which is composed of 35 datasets with bounding box annotations. As shown in Table 2, though we only employ much fewer semantically labeled data, we achieve better performance compared with previous models under similar settings.

3.2. Video Object Segmentation

Video object segmentation (VOS) aims to segment an interested object in a video by giving text or visual clues. Our model focuses on the semi-supervised setting, which segments a particular object throughout a video by giving visual clues in the first frame. In *DINOv*, the visual prompt originates from one single image (generic/referring segmentation) or other images in one batch (generic segmentation). Therefore, our model has learned to prompt with visual features from other images. Therefore, *DINOv* is able to do video object segmentation (VOS) by replacing current frame visual prompt features with previous frames. For more accurate tracking, we also store the visual features of the predicted mask in previous frames. These fea-

Table 3. **Zero-shot** video object segmentation. Without training with video or pairwise image data, our approach is able to do video object segmentation in a zero-shot manner. (#Concurrent work.)

Method	Segmentation Data	Type	Refer-Type	Zero-Shot	DAVIS17			DAVIS16-Interactive			YouTube-VOS 2018				
					JF	J	F	JF	J	F	G	Js	Fs	Ju	Fu
With Video Data															
AGSS [20]	VOS+DAVIS		Mask		67.4	64.9	69.9	–	–	–	71.3	71.3	65.5	75.2	73.1
AGAME [11]	(Synth)VOS+DAVIS		Mask		70.0	67.2	72.7	–	–	–	66.0	66.9	*	61.2	*
SWEM [22]	Image+VOS+DAVIS		Mask		84.3	81.2	87.4	–	–	–	82.8	82.4	86.9	77.1	85.0
XMem [3]	Image+VOS+DAVIS	Video	Mask		–	–	–	–	–	–	86.1	85.1	89.8	80.3	89.2
SiamMask [32]	COCO+VOS		Box		*	54.3	58.5	69.8	71.7	67.8	*	60.2	58.2	45.1	47.7
MiVOS [4]	BL30K+VOS+DAVIS		Mask		84.5	81.7	87.4	91.0	89.6	92.4	82.6	81.1	85.6	77.7	86.2
ReferFormer-B [35]	RefCOCO(+/g)+VOS+DAVIS		Text		61.1	58.1	64.1	–	–	–	*	*	*	*	*
UNINEXT-T [41]	Image+Video		Mask		74.5	71.3	77.6	–	–	–	77.0	76.8	81.0	70.8	79.4
UNINEXT-L [41]	Image+Video	Generalist	Mask		77.2	73.2	81.2	–	–	–	78.1	79.1	83.5	71.0	78.9
UNINEXT-L [41]	Image+Video		Text		66.7	62.3	71.1	–	–	–	*	*	*	*	*
Without Video Data															
Painter-L [33]	COCO+ADE+NYUv2		Mask	✓	34.6	28.5	40.8	–	–	–	24.1	27.6	35.8	14.3	18.7
SegGPT-L [34]	COCO+ADE+VOC+...		Mask	✓	75.6	72.5	78.6	–	–	–	74.7	75.1	80.2	67.4	75.9
PerSAM-L [46]	SAM+DAVIS		Mask	✗	60.3	56.6	63.9	–	–	–	*	*	*	*	*
SEEM-T [51]	COCO+LVIS	Generalist	Mask	✓	60.4	57.6	63.3	62.7	58.9	66.4	51.4	55.6	44.1	59.2	46.9
SEEM-L [51]				✓	58.9	55.0	62.8	62.2	58.3	66.0	50.0	57.2	38.2	61.3	43.3
DINOv-T (Ours)	COCO+SAM		Mask	✓	73.3	71.0	75.7	77.0	72.9	81.2	60.9	65.3	70.0	52.3	57.9
DINOv-L (Ours)				✓	72.3	69.8	74.8	75.4	71.3	79.4	59.6	61.7	65.7	52.3	58.8

tures, denoted as memory visual prompts, will be averaged together with the first frame's given prompt to construct the visual prompt of the current frame. Details of the memory visual prompt and ablations are in Appendix. D. By default, the memory length is set to 8. In Table 3, we conduct (interactive) video object segmentation evaluation on DAVIS17, DAVIS2016-Interactive, and Youtube-VOS 2018. The results of DAVIS2017 and Youtube-VOS 2018 indicate our model achieves better performance than SEEM and PerSAM. In addition, *DINOv* can also do interactive VOS, and our performance on DAVIS16-Interactive achieves significantly better performance compared with models not using video data for training.

3.3. Ablation

Effectiveness of Query Formulation. In Table 4, we ablate the effectiveness of using different query formulations for different tasks. The results indicate our double query formulation outperforms using only one type of query.

Effectiveness of Visual Prompt Formulation. In Table 5, we attempt to use a pre-trained CLIP vision encoder to encode the features of the visual prompt by cropping the prompted region into images for CLIP to process. As CLIP features contain rich semantics with few appearance features, which could not apply to referring segmentation tasks. Therefore, we ablate on generic segmentation tasks and find that the final model could not generalize well on open-set datasets like ADE. This result verifies our hypothesis that CLIP vision features could not generalize well on in-context visual prompting.

Effectiveness of Unifying Tasks and Data. We unify visual generic segmentation and visual referring segmentation to use both semantically labeled data (COCO) and data with

Table 4. **Ablation** of using difference queries to do both in-context reference and generic segmentation. By default, we use both generic query and interactive query. We remove one type of query at a time to ablate their effectiveness.

Method	COCO				DAVIS17		
	PQ	mask AP	box AP	mIoU	JF	J	F
DINOv-SwinT	49.6	42.7	47.0	58.0	73.3	71.0	75.7
only point query	45.2	**31.0**(11.7)	**34.7**(-12.3)	52.7	71.4	68.8	74.0
only generic query	46.2	**38.3**(-4.4)	**41.5**(-6.0)	53.3	68.9	66.5	71.3

Table 5. **Ablation** of using different ways to encode the visual prompt on our Swin-T model. Under the same setting, we change our prompt encoding method and use a pre-trained CLIP to crop and encode the prompted objects in the image.

Prompt Encoding	COCO (in-domain)				ADE (out-domain)			
	PQ	mask AP	box AP	mIoU	PQ	mask AP	box AP	mIoU
Ours	49.6	42.7	47.0	58.0	19.4	11.4	12.8	21.9
CLIP	48.5	40.7	43.5	54.9	12.6	1.4	1.3	13.3

only segmentation annotations (SA-1B). In Table 6, the results indicate that employing both datasets improves each individual task.

Training batch size for generic segmentation. In Table 7, the results show that increasing training batch size consistently improves the generic segmentation performance. The reason for this phenomenon is that a larger batch size helps to sample more positive and negative visual in-context examples across different images, which better matches the inference setting with random visual examples.

Inference In-Context Examples. In Fig. 6, we ablate the impact of using different in-context lengths. Increasing the in-context example exhibits diminishing returns, especially when the number of examples is more than eight.

Table 6. **Ablation** of the effectiveness of unifying tasks and data.

Method	Data	COCO				DAVIS17		
		PQ	mask AP	box AP	mIoU	JF	J	F
DINOv-SwinT	COCO, SAM	49.6	42.7	47.0	58.0	73.3	71.0	75.7
DINOv-SwinT	COCO	48.9	41.7	45.9	57.1	63.3	60.8	65.7
DINOv-SwinT	SAM	N/A	–	–	–	68.4	66.0	70.8

Table 7. **Ablation** of image batchsize sampling in training.

Method	#Batchsize for Prompt Sampling	COCO			
		PQ	mask AP	box AP	mIoU
DINOv-SwinT	1	28.9	23.2	25.3	33.7
DINOv-SwinT	4	45.1	37.0	40.4	50.6
DINOv-SwinT	8	47.3	39.2	43.1	53.1
DINOv-SwinT	32	47.8	40.3	44.1	56.2
DINOv-SwinT	64	49.0	45.2	41.5	57.0

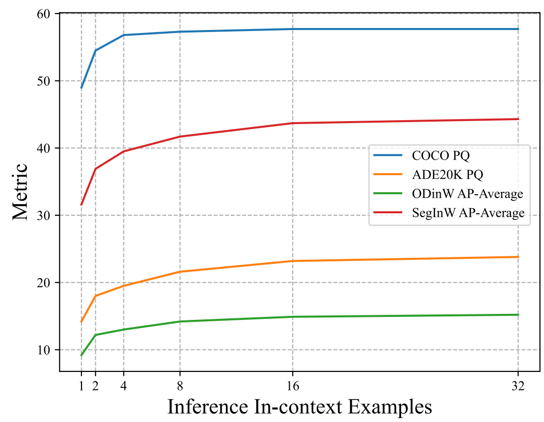


Figure 6. *DINOv* query formulation of different tasks.

4. Related Works

4.1. Visual Perception Through Text Prompt

Innovations in open-vocabulary object detection [7, 12, 19, 24, 26, 43, 44, 47] and open-vocabulary segmentation [6, 8, 14, 31, 37, 44], have shown great potential in generic visual perception, by leveraging large pre-trained vision-language models like CLIP [30] and ALIGN [10]. These approaches demonstrate significant strides in zero-shot and few-shot performance, adapting to a variety of visual contexts through text prompts. However, the reliance on text alone introduces limitations due to linguistic ambiguity and the potential mismatch between textual descriptions and complex visual scenes [40]. This highlights the ongoing need to refine the integration of visual inputs for more accurate and comprehensive image perception.

4.2. Visual Perception Through Image Example

Building upon the foundations set by text-based visual perception methodologies, the field has seen a notable shift towards incorporating image examples to enhance accuracy and context sensitivity. OV-DETR [43] extends its open-vocabulary object detection capability beyond text, by utilizing both the image encoder and text encoder from CLIP [30], allowing for object detection guided by visual examples. Similarly, OWL-ViT [26] leverages large-scale image text examples in its contrastive pre-training phase, and propose to adopt image example for one-shot image-conditioned object detection. MQ-Det [40] utilizes image examples to enhance text descriptions for better open-vocabulary object detection performance. These methods typically adopt the image encoder in CLIP to extract visual features from given image examples for a more accurate perception of objects and scenes, and demonstrate that visual examples can bridge the gap between textual ambiguity and the complex nature of visual perception.

4.3. Visual Perception Through Visual Prompt

Different from image example-based methods that take an image as input, which are then processed by multi-modal encoder like CLIP [30], visual prompt-based methods typically use visual instructions (e.g. point, mask, scribble, and refereed regions of another image) to guide a model for a specific visual task. SAM [13], for instance, introduces a promotable model for interactive segmentation, fostering research in computer vision foundation models. SEEM [51] stands out as an interactive and versatile model for segmenting objects, accommodating various types of prompts, and is semantic-aware compared to SAM. Semantic-SAM [17] excels in semantic awareness and recognizing granularity, and can handle panoptic and part segmentation. SegGPT [34] unifies various segmentation tasks by formulating segmentation as an in-context coloring problem.

5. Conclusion

We present *DINOv*, a unified framework for in-context visual prompting to accommodate both referring segmentation and generic segmentation tasks. To effectively formulate in-context visual prompts, we designed a simple prompt encoder to encoder reference visual prompts from the reference image and adopted a shared decoder to decode the final target visual prompts from the target image. We also formulate generic latent queries and point queries to align different tasks and data. The experimental results indicate that *DINOv* demonstrates impressive referring and generic segmentation capabilities to refer and detect with in-context visual prompting. Notably, *DINOv* delivers competitive performance compared to close-set segmentation on in-domain datasets and show promising results on many open-set segmentation benchmarks. We hope our early exploration of visual in-context prompting could inspire the community.

Limitations. We employ limited semantically labeled data (COCO), which can be scaled up for better performance and extended to text prompts for multi-modal understanding.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 6

[3] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 7

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, 2021. 7

[5] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 6

[6] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 8

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 8

[8] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 8

[9] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. *arXiv preprint arXiv:2211.06220*, 2022. 6

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 8

[11] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation, 2018. 7

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 6, 8

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2, 5, 8, 12

[14] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021. 8

[15] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1, 2023. 2

[16] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 5, 6, 12

[17] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2, 5, 8, 12

[18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 2, 6

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 6, 8, 12

[20] Huaijia Lin, Xiaojuan Qi, and Jiaya Jia. Agss-vos: Attention guided single-shot video object segmentation. In *ICCV*, 2019. 7

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5, 12

[22] 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. 7

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 8

[25] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 12

[26] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 4, 8

[27] OpenAI. Gpt-4 technical report, 2023. 1

[28] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 12

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 8

[31] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 8

[32] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019. 7

[33] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning, 2023. 5, 6, 7

[34] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2, 5, 6, 7, 8

[35] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. *arXiv preprint arXiv:2201.00487*, 2022. 7

[36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 6

[37] 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, 8

[38] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models, 2023. 2

[39] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 12

[40] Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. Multimodal queried object detection in the wild. *arXiv preprint arXiv:2305.18980*, 2023. 8

[41] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. *arXiv preprint arXiv:2303.06674*, 2023. 7

[42] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487*, 2023. 2, 6

[43] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 8

[44] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 2, 6, 8

[45] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 6

[46] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot, 2023. 7

[47] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei 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. 5, 8

[48] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 12

[49] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 2

[50] 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. *arXiv preprint arXiv:2212.11270*, 2022. 2, 6, 12

[51] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 2, 7, 8