OpenSeeD - Open-vocab segmentation & detection

ICCV 2023

Open vocabulary

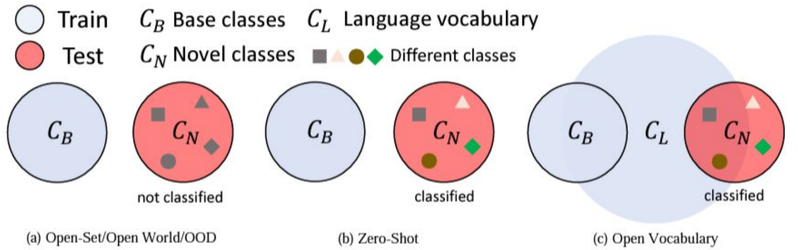
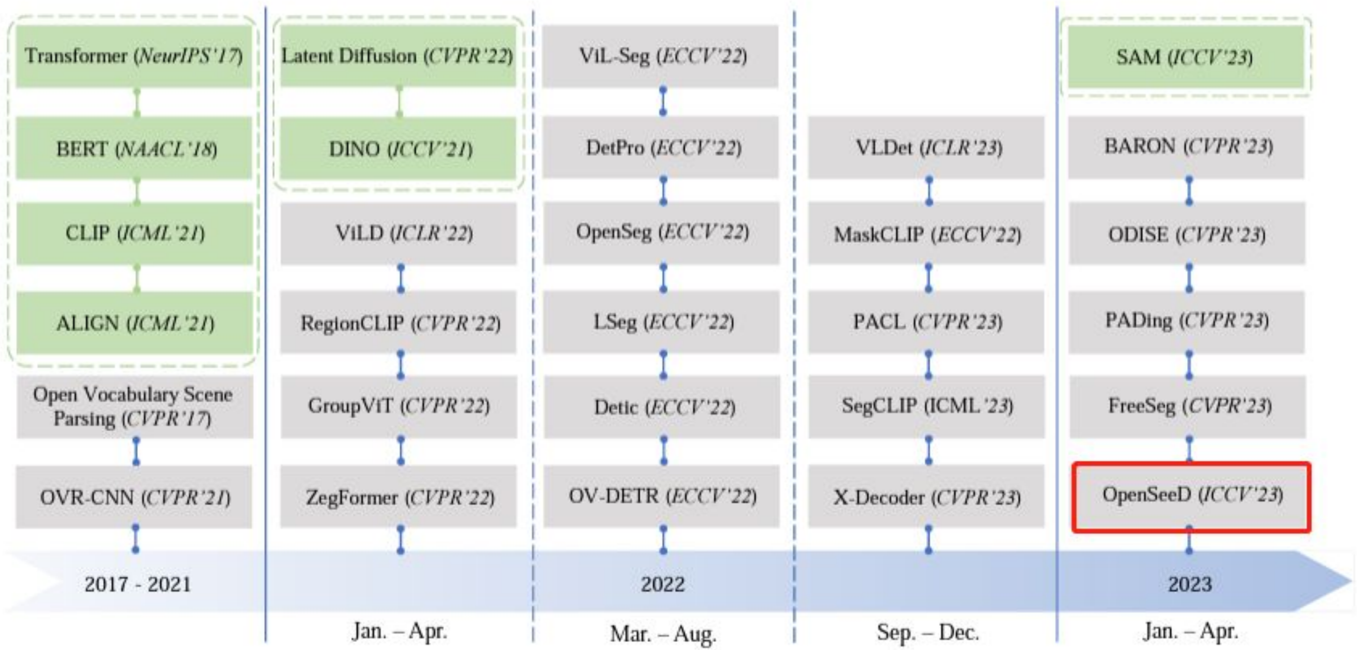


Fig. 1: Concepts comparison between open-set/open world/out-of-distribution detection (OOD), zero-shot and open vocabulary. Different shapes represent different novel categories. Colors represent the predictions of the novel objects. (a), in the open-set/Open World/OOD settings, the model only needs to identify novel classes and mark them as 'unknown.' (b) in the zero-shot setting, a model must classify novel classes into specific categories. (c) in the open vocabulary settings, the model can classify novel classes with the help of large language vocabulary knowledge C_L.

- **Open set:** all unseen classes are unknown.
- **Open vocabulary:** with the help of language vocabulary, the model could classify the unseen data into classes.

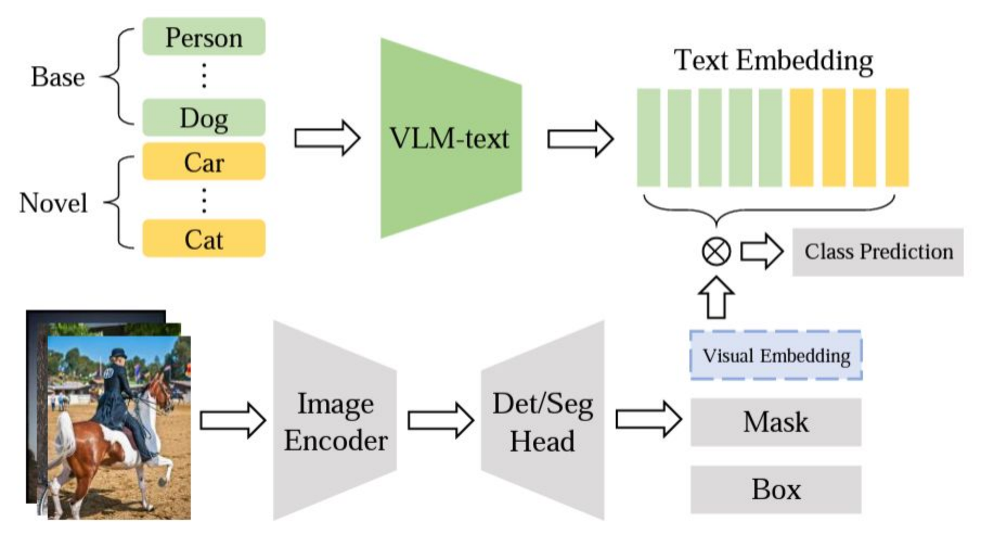
Open-vocab



Transformer (NeurIPS'17)
BERT (NAACL'18)
CLIP (ICML'21)
ALIGN (ICML'21)
Open Vocabulary Scene Parsing (CVPR'17)
OVR-CNN (CVPR'21)
Latent Diffusion (CVPR'22)
DINO (ICCV'21)
ViLD (ICLR'22)
RegionCLIP (CVPR'22)
GroupViT (CVPR'22)
ZegFormer (CVPR'22)
ViL-Seg (ECCV'22)
DetPro (ECCV'22)
OpenSeg (ECCV'22)
LSeg (ECCV'22)
Detic (ECCV'22)
OV-DETR (ECCV'22)
VLDet (ICLR'23)
MaskCLIP (ECCV'22)
PACL (CVPR'23)
SegCLIP (ICML'23)
X-Decoder (CVPR'23)
SAM (ICCV'23)
BARON (CVPR'23)
ODISE (CVPR'23)
PADing (CVPR'23)
FreeSeg (CVPR'23)
OpenSeeD (ICCV'23)
2017 - 2021
2022
2023
Jan. – Apr.
Mar. – Aug.
Sep. – Dec.
Jan. – Apr.

Method	Task	Text Training Data	Vision Training Annotations	Text Model	Vision Model	Highlight
Open Vocabulary Detection (Sec. 3.2)						
OVR-CNN [12]	OVOD	Captions	Bounding Boxes (Base)	BERT	Faster R-CNN	The first method that proposes OVOD and adopts grounded caption pre-training.
ViLD [13]	OVOD	None	Bounding Boxes (Base)	CLIP-text	CLIP-vision + Faster R-CNN	The first method that distills knowledge from the pre-trained CLIP model.
HierKD [153]	OVOD	None	Bounding Boxes (Base)	CLIP-text	CLIP-vision + ATSS	Distills Global-level knowledge from the pre-trained CLIP model into the one-stage detector.
VL-PLM [154]	OVOD	None	Bounding Boxes (Base)	CLIP-text	CLIP-vision + Faster R-CNN	Generate pseudo-labels for novel classes using pre-trained OVOD.
RegionCLIP [155]	OVOD	Captions	Bounding Boxes (Base + Pseudo Novel)	CLIP-text	CLIP-vision + Faster R-CNN	Creates region-text pairs as pseudo labels using CLIP and pre-train the detector in the first stage.
Detic [156]	OVOD	ImageNet	Bounding Boxes (Base + Pseudo Novel)	CLIP-text	Centernet2	Propose a weakly supervised approach that is training rare classes with image-level annotations.
DetPro [157]	OVOD	ImageNet	Bounding Boxes (Base + Pseudo Novel)	CLIP-text	Centernet2	learn continuous prompt representations for open vocabulary object detection based on the pre-trained vision-language model.
OV-DETR [158]	OVOD	None	Bounding Boxes (Base)	CLIP-text	CLIP-vision + DETR	Introduces a conditional binary matching mechanism to let DETR model generalize to queries from unseen classes.
CORA [159]	OVOD	None	Bounding Boxes (Base)	CLIP-text	CLIP-vision + DETR	Propose Anchor Pre-Matching strategy to reduce both training and inference time for conditional binary matching.
Prompt-OVD [160]	OVOD	None	Bounding Boxes (Base)	CLIP-text	CLIP-vision + DETR	Propose RoI-based masked attention and RoI pruning techniques by utilizing CLIP visual features to improve the novel object classification.
VLDet [35]	OVOD	Captions	Bounding Boxes (Base)	CLIP-text	Faster R-CNN	Aligns image regions with words in captions by a set matching method.
F-VLM [161]	OVOD	None	Bounding Boxes (Base)	CLIP-text	CLIP-vision + Mask R-CNN	Train the detector with frozen VLMs and combine scores of joint detection and VLMs.
BARON [162]	OVOD	None	Bounding Boxes (Base)	CLIP-text	Faster R-CNN	Aligning Bag of Regions for Open Vocabulary Object Detection.
OWLv2 [163]	OVOD	None	Bounding Boxes (Base + Pseudo Novel)	CLIP-text	ViT + detection head	Generate pseudo-labels from WebLI dataset and train the detector with the generated datasets.
MaMMUT [164]	OVOD	Captions	Bounding Boxes (Base + Pseudo Novel)	CLIP-text	ViT + detection head	Joint pre-train with multi-modal tasks to benefit the novel object detection.
Open Vocabulary Segmentation (Sec. 3.3)						
LSeg [15]	OVSS	None	Segmentation Masks (Base)	CLIP-text	ViT	Aligns text embeddings from the VLM model with pixel features.
ZegFormer [165]	OVSS	None	Segmentation Masks (Base)	CLIP-text	CLIP-vision + Query-based Transformer Decoder	Decouple segmentation and classification by generating class-agnostic segment masks then classify each mask.
OpenSeg [37]	OVSS	Captions	Segmentation Masks (Base)	ALIGN	EfficientNet-B7	Performs region-word grounding loss between mask features and word features.
MaskCLIP+ [166]	OVSS	None	Segmentation Masks (Pseudo All)	CLIP-text	CLIP-vision + DeepLabv2	Modifies CLIP so that it can output per-pixel feature maps.
OVSeg [167]	OVSS	Captions	Segmentation Masks (Base + Pseudo All)	CLIP-text	CLIP-vison + MaskFormer	Uses CLIP to match the proposed image regions with nouns in the captions to generate pseudo labels.
GroupVit [168]	OVSS	Captions	None	Transformer Encoder	ViT-S	Learns semantic segmentation only with caption data.
Vil-Seg [169]	OVSS	Captions	None	ViT-B	ViT-B	Learns semantic segmentation without pixel-level annotations using contrastive loss and clustering loss.
TCL [170]	OVSS	Captions	None	CLIP-text	CLIP-vision	Proposes a finer-grained contrastive loss for training without pixel-level annotations.
CGG [38]	OVIS	Captions	Segmentation Masks (Base)	BERT	Mask2Former	Fully exploits caption data using caption grounding and generation.
MaskCLIP [39]	OVPS	None	Segmentation Masks (Base)	CLIP-text	CLIP-vision + Mask2Former	Proposes Relative Mask Attention (RMA) modules to adapt cropped images to the pre-trained CLIP model.
ODISE [171]	OVPS	Captions	Segmentation Masks (Base)	CLIP-text	Stable Diffusion	Exploits the vision-language alignment learned by denoising diffusion models.
OVDiff [172]	OVSS	None	None	CLIP-text	Stable Diffusion	Proposes a prototype-based method. Use diffusion models to produce prototypes.
OpenSeed [173]	OVIS	None	Segmentation Masks & Boxes (Base)	UniCL	MaskDINO	Jointly learns from detection and segmentation data.
OVSegmentor [174]	OVSS	Captions	None	BERT	DINO	introduces masked entity completion and cross-image mask constituency objectives to improve training.
Cat-Seg [175]	OVSS	None	Segmentation Masks (Base)	CLIP-text	CLIP-image	Jointly aggregates the image and text embeddings from CLIP to form the segmentation predictions.
SAN [176]	OVSS	None	Segmentation Masks (Base)	CLIP-text	CLIP-vision	Proposes a lightweight side adaptor network to extract the CLIP model's knowledge.

A common architecture

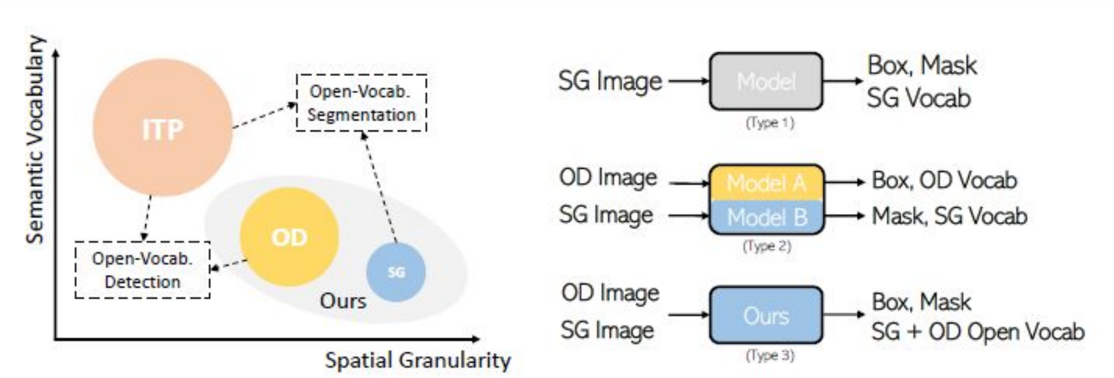


A common open-vocab method always focusing on a single task, e.g., object detection.

Motivation

- Most of common methods focused on how to improve the performance for either detection or segmentation.
- Transferring weak image-level supervision to fine-grained tasks usually requires sophisticated designs to mitigate the huge granularity gap and is vulnerable to noises in image-text pairs.

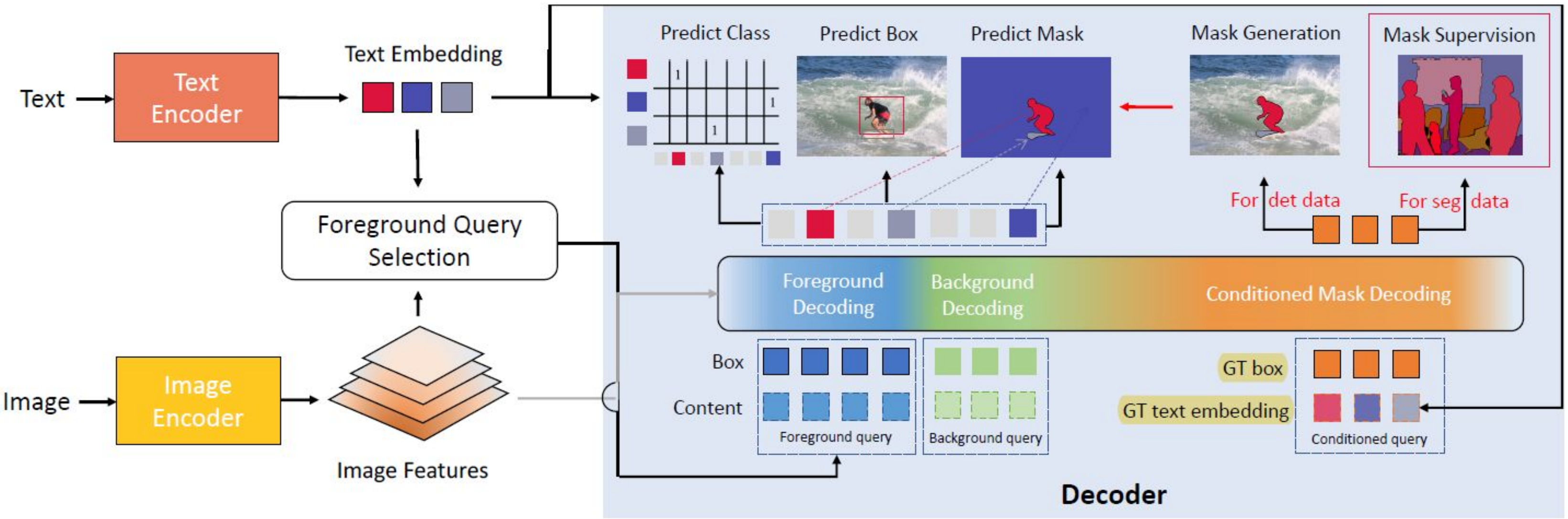
➔ *Can we bridge detection and segmentation that are cleaner and have a closer gap to attain a good open-vocabulary model for both?*



Challenges:

- The vocabulary shares commons but also bear substantial differences between the two tasks. We need to accommodate the two vocabularies and further go beyond towards open vocabulary.
- Semantic and panoptic segmentation tasks require segmenting not only foreground objects (things like **“dog” and “cat”**.) but also background concepts (stuff like **“sky” and “building”**), while detection task solely cares about foreground objects.
- Box supervision by nature is coarser than mask supervision.

OpenSeeD



Text
Text Encoder
Text Embedding
Foreground Query Selection
Image
Image Encoder
Image Features
Predict Class
Predict Box
Predict Mask
Mask Generation
Mask Supervision
For det data
For seg data
Foreground Decoding
Background Decoding
Conditioned Mask Decoding
Box
Content
Foreground query
Background query
GT box
GT text embedding
Conditioned query
Decoder

Segmentation data: $D_m = \{I_i, (c_i, m_i)\}_{i=1}^{M}$

Detection data: $D_b = \{I_j, (c_j, b_j)\}_{j=1}^{N}$

Vocabulary V has k visual concepts, query Q

Image feature: $O = Enc_I(I)$

Text feature: $T = Enc_T(V)$

Decoding images will have masks, boxes, and decoded semantics

$$< P^m, P^b, P^s > = Dec(Q; O)$$

Classification score: $P^c = Sim(P^s, T)$

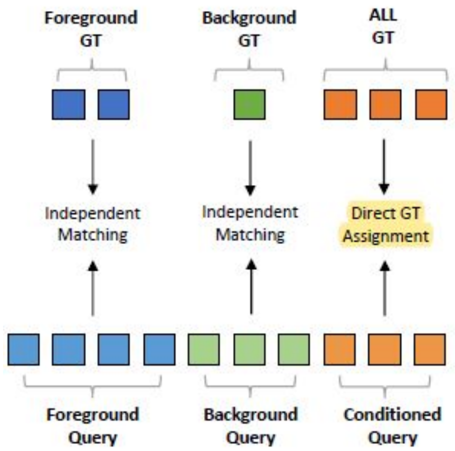
$$\mathcal{L}_{all} = \sum_{I,(\mathbf{c},\mathbf{m})\in\mathcal{D}_m} \overbrace{\Big(\mathcal{L}_m(\mathbf{P^m},\mathbf{m}) + \mathcal{L}_b(\mathbf{P^b},\hat{\mathbf{b}}) + \mathcal{L}_c(\mathbf{P^c},\mathbf{c})\Big)}^{\text{Segmentation loss}} + \sum_{I,(\mathbf{c},\mathbf{b})\in\mathcal{D}_b} \underbrace{\Big(\mathcal{L}_b(\mathbf{P^b},\mathbf{b}) + \mathcal{L}_c(\mathbf{P^c},\mathbf{c})\Big)}_{\text{Detection loss}}$$

- ➔ *Using the same queries for both tasks creates conflicts that can significantly degrade performance.*
- ➔ *Good box predictions are typically indicative of good masks, and vice versa.*

Bridge Task Gap

Semantic and panoptic segmentation require the recognition of both foreground and background, while detection focuses solely on localizing foreground objects.

Without loss of generality, we have defined the visual concepts that appear in instance segmentation and detection as foreground, while the stuff categories in panoptic segmentation are considered background. To mitigate the task discrepancy, we perform foreground and background decoding with foreground queries $\mathbf{Q_f}$ and background queries $\mathbf{Q_b}$, respectively. Specifically, for these two query types, our decoder predicts two sets of outputs: $\langle \mathbf{P}_f^m, \mathbf{P}_f^b, \mathbf{P}_f^c \rangle$ and $\langle \mathbf{P}_b^m, \mathbf{P}_b^b, \mathbf{P}_b^c \rangle$. We also divide the ground truths in segmentation dataset into two groups: $(\mathbf{c}_f, \mathbf{m}_f)$ and $(\mathbf{c}_b, \mathbf{m}_b)$, and then perform two independent Hungarian Matching processes for these two sets correspondingly, as shown in



(a) Label Assignment

Bridge Task Gap

Semantic and panoptic segmentation require the recognition of <u>both foreground and background</u>, while detection focuses solely on localizing <u>foreground objects</u>.

$$
\begin{aligned}
\mathcal{L}_{all} = \sum_{I,(\mathbf{c},\mathbf{m})\in\mathcal{D}_m} & \overbrace{\Big(\mathcal{L}_m(\mathbf{P}^{\mathbf{m}}_{\mathbf{f}}, \mathbf{m}_f) + \mathcal{L}_b(\mathbf{P}^{\mathbf{b}}_{\mathbf{f}}, \hat{\mathbf{b}}_f) + \mathcal{L}_c(\mathbf{P}^{\mathbf{c}}_{\mathbf{f}}, \mathbf{c}_f)\Big)}^{\text{Segmentation loss for foreground}} \\
& + \overbrace{\Big(\mathcal{L}_m(\mathbf{P}^{\mathbf{m}}_{\mathbf{b}}, \mathbf{m}_b) + L_b(\mathbf{P}^{\mathbf{b}}_{\mathbf{b}}, \hat{\mathbf{b}}_b) + \mathcal{L}_c(\mathbf{P}^{\mathbf{c}}_{\mathbf{b}}, \mathbf{c}_b)\Big)}^{\text{Segmentation loss for background}} \\
& + \sum_{I,(\mathbf{c},\mathbf{b})\in\mathcal{D}_b} \underbrace{\Big(\mathcal{L}_b(\mathbf{P}^{\mathbf{b}}_{\mathbf{f}}, \mathbf{b}) + \mathcal{L}_c(\mathbf{P}^{\mathbf{c}}_{\mathbf{f}}, \mathbf{c})\Big)}_{\text{Detection loss for foreground}}
\end{aligned}
\tag{4}
$$

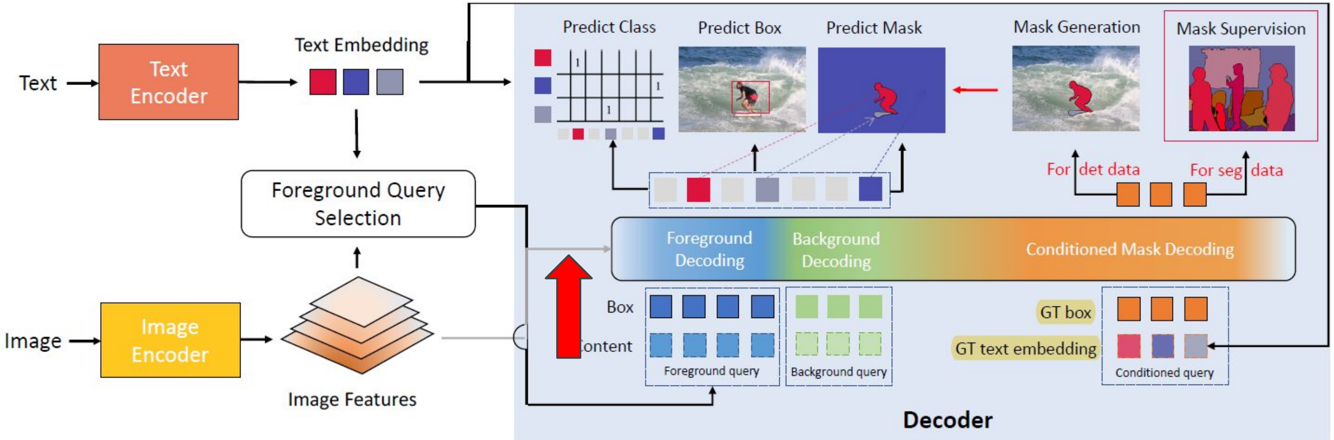
- ➔ *How to determine foreground and background queries?*

Bridge Task Gap

Language-guided foreground query selection

$$\mathbf{E}^b = \text{Head}(\mathbf{O}), \mathbf{E}^c = \text{Sim}(\mathbf{O}, \mathbf{T}) \quad (5)$$

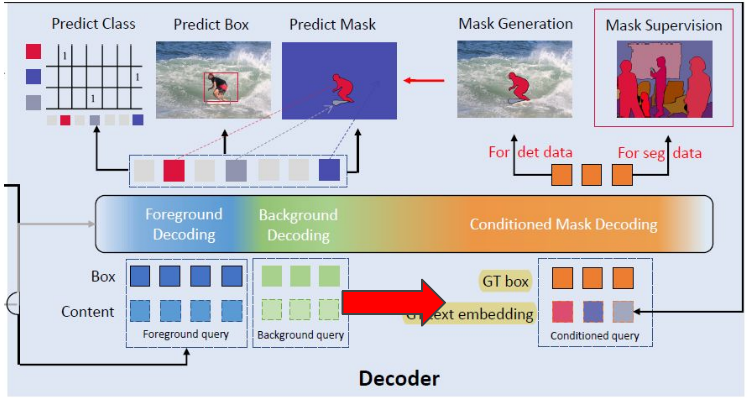
where Head is the box head. Then we select L_f top-ranked entries from $\mathbf{E}^b$ and $\mathbf{O}$ according to the scores in $\mathbf{E}^c$. These



Bridge Data Gap

Ultimate goal: $$\mathcal{L}_{all} = \sum_{I,(\mathbf{c},\mathbf{m},\mathbf{b})\in\mathcal{D}} \mathcal{L}_m(\mathbf{P}^{\mathbf{m}}, \mathbf{m}) + \mathcal{L}_b(\mathbf{P}^{\mathbf{b}}, \mathbf{b}) + \mathcal{L}_c(\mathbf{P}^{\mathbf{c}}, \mathbf{c})$$

➔ *We can generate boxes by masks in segmentation task.*
➔ *We cannot have masks given only coarse location (box) and categories.*



Given the ground-truth concepts and boxes, $(\mathbf{c}, \mathbf{b})$, we employ the decoder to decode the mask:

$$\mathbf{P}^{\mathbf{m}} = \mathrm{Dec}\left((\mathbf{t}, \mathbf{b}); \mathbf{O}\right) \tag{7}$$

where t is the text features extracted for the concepts. Based

➔ *Can we learn from segmentation data a good mapping which generalizes well to detection data <u>with different categories</u>?*

Bridge Data Gap

➔ *Can we learn from segmentation data a good mapping which generalizes well to detection data <u>with different categories</u>?*

Verification: train a model which learns to decode masks conditioned on GT concepts and boxes on COCO, and then evaluate the conditioned decoding performance on ADE20K.

Table 1: Results for models trained on COCO without and with conditioned mask decoding. Models are evaluated on COCO and ADE20K validation set. "final" and "early" means the model at the final and early training stage, respectively. "Convert box into mask" means we directly convert the GT boxes into rectangle masks for evaluation.

Method	COCO Mask AP	ADE Mask AP
OpenSeeD (T)	45.1	8.6
OpenSeeD (T) w conditioned train & eval (final)	53.2	**46.4**
OpenSeeD (T) w conditioned train & eval (early)	42.2	14.8
Convert box into mask	16.2	25.4

Figure 6: Background and foreground masks generated with boxes and text as the condition for sample images in ADE20K [57].

<u>Online Assistance:</u> train one model and generate the masks on the fly. Instead of directly using the generated masks as mask supervision, we use the masks to assist in matching predictions and GT instances because the mask quality is not strong enough for supervision. especially in the early stage.

<u>Offline Assistance:</u> train our model with conditioned mask decoding until convergence and generate mask annotations for detection data.

Table 2: One suit of weights for open-vocabulary segmentation on multiple datasets in a zero-shot manner. Our model is pre-trained on COCO and Objects365 data. 'SEG' indicates segmentation data (COCO), 'DET' indicates detection data (Objects365), and ITP indicates image-text pairs/referring/-captioning data. The values in gray are supervised results. ⋆ X-Decoder (L) is not open-source, so we cannot evaluate its performance on LVIS.

Method	Training Data			ADE				Cityscapes			LVIS		BDD		SCAN-20		SCAN-41	SUN
	SEG	DET	ITP	PQ	mask AP	box AP	mIoU	PQ	mask AP	mIoU	mask AP	box AP	PQ	mIoU	PQ	mIoU	mIoU	mIoU
MSeg (B) [23]	✓	✗	✗	33.7	32.6	–	19.1	46.9	24.8	51.1	–	–	–	44.9	–	33.4	–	29.6
MDETR [20]	✗	✓	✓	–	–	–	–	–	–	–	–	24.2	–	–	–	–	–	–
LSeg+ (B) [26]	✓	✗	✗	–	–	–	18.0	–	–	–	–	–	–	–	–	–	–	–
ZegFormer (B) [16]	✓	✗	✗	–	–	–	16.4	–	–	–	–	–	–	–	–	–	–	–
OpenSeg (B) [12]	✓	✗	✗	–	–	–	21.1	–	–	–	–	–	–	–	–	–	–	–
OpenSeg (B) [12]	✓	✗	✓	–	–	–	26.4	–	–	–	–	–	–	–	–	–	–	–
MaskCLIP (L) [7]	✓	✗	✗	15.1	6.0	–	23.7	–	–	–	–	–	–	–	–	–	–	–
ODISE (H) [46]	✓	✗	✓	23.5	13.9	–	28.7	–	–	–	–	–	–	–	–	–	–	–
GLIP (T) [29]	✗	✓	✗	–	–	–	–	–	–	–	–	18.5	–	–	–	–	–	–
X-Decoder (T) [60]	✓	✗	✓	18.8	9.8	–	25.0	37.2	16.0	47.3	9.6	–	16.4	42.4	30.7	37.8	21.7	34.5
OpenSeeD (T)	✓	✓	✗	**19.8**	**14.1**	**17.0**	22.9	**37.3**	**26.2**	46.1	**19.4**	**21.8**	**17.2**	**44.8**	**39.7**	**45.1**	**25.2**	**39.0**
X-Decoder (L) [60]	✓	✗	✓	21.8	13.1	–	29.6	38.1	24.9	52.0	⋆	–	17.8	47.2	39.5	49.5	29.7	43.0
OpenSeeD (L)	✓	✓	✗	19.7	**15.0**	**17.7**	23.4	**41.4**	**33.2**	47.8	**21.0**	**23.0**	**19.4**	**47.4**	**42.2**	48.7	27.4	41.9

Table 3: Task-specific transfer of *OpenSeeD* to different segmentation and VL tasks. We directly evaluate the COCO performance without finetuning. Note: "–" denotes the model does not have number reported or does not have the ability for the specific task. ⋆ means it is the test set results. The results in the bracket are trained with 1280×1280 image size. Note that the results of GLIPv2 and X-Decoder on COCO are fine-tuned while those of *OpenSeeD* are reported without fine-tuning.

Method	Type	ADE				Cityscapes			COCO			
		PQ	mask AP	box AP	mIoU	PQ	mask AP	mIoU	PQ	mask AP	box AP	mIoU
Mask2Former (T) [5]	Closed-set	39.7	26.4	28.8	47.7	63.9	39.1	80.5	53.2	43.3	46.1	63.2
Mask2Former (B) [5]		–	–	–	53.9	66.1	42.8	82.7	56.4	46.3	49.5	67.1
Mask2Former (L) [5]		48.1	34.2	36.4	56.1	66.6	43.6	82.9	57.8	48.6	52.1	67.4
OneFormer (L) [18]		48.6	35.9	–	57.0	67.2	45.6	83.0	57.9	48.9	–	67.4
MaskDINO (L) [28]		–	–	–	–	–	–	–	58.3	50.6	56.2	67.5
Pano/SegFormer (B) [44]		–	–	–	51.0	–	–	–	55.4	–	–	–
kMaX-DeepLab (L) [52]		48.7	–	–	54.8	–	–	–	58.1	–	–	–
GLIPv2 (T) [55]	Open-vocabulary	–	–	–	–	–	–	–	–	42.0⋆	–	–
GLIPv2 (B) [55]		–	–	–	–	–	–	–	–	45.8⋆	–	–
GLIPv2 (H) [55]		–	–	–	–	–	–		–	48.9⋆	–	–
X-Decoder (T) [60]		41.6	27.7	28.8	51.0	61.3	36.2	78.7	52.6	41.3	43.6	62.4
OpenSeeD (T)		**47.2**	**35.1**	**39.4**	**52.2**	**63.9**	**38.2**	**80.3**	**55.4**	**47.6**	**52.0**	**64.0**
X-Decoder (L) [60]		49.6	35.8	–	58.1	65.6	42.2	81.7	56.9	46.7	–	67.5
OpenSeeD (L)		**53.1 (53.7)**	**42.0 (42.6)**	**46.4 (46.9)**	**58.6 (58.4)**	**69.2**	**49.3**	**84.5**	**59.5**	**53.2**	**58.2**	**68.6**

Table 4: One suit of weights on the SeginW benchmark in a zero-shot manner.

Model	Med.	Avg	Air-Par.	Bottles	Br. Tum.	Chicken	Cows	Ele.-Sha.	Eleph.	Fruits	Gar.	Gin.-Gar.	Hand	Hand-Metal	House-Parts	HH.-Items	Nut.-Squi.	Phones	Poles	Puppies	Rail	Sal.-Fil.	Stra.	Tablets	Toolkits	Trash	W.M
X-Decoder (T) [60]	15.2	22.7	10.5	19.0	1.1	12.0	12.0	1.2	65.6	66.5	28.7	7.9	0.6	22.4	5.5	50.6	62.1	29.9	3.6	48.9	0.7	15.0	41.6	15.2	9.5	19.3	16.2
OpenSeeD (T)	21.5	**33.9**	12.2	27.4	5.0	68.7	21.5	0.3	73.3	72.9	7.3	6.2	92.4	62.3	0.5	55.0	63.6	2.4	4.6	63.8	5.4	15.6	85.3	32.0	4.8	14.5	51.0
X-Decoder (L) [60]	22.3	32.3	13.1	42.1	2.2	8.6	44.9	7.5	66.0	79.2	33.0	11.6	75.9	42.1	7.0	53.0	68.4	15.6	20.1	59.0	2.3	19.0	67.1	22.5	9.9	22.3	13.8
OpenSeeD (L)	38.7	**36.1**	13.0	39.7	2.1	82.9	40.9	4.7	72.9	76.4	16.9	13.6	92.7	38.7	1.8	50.0	40.0	7.6	4.6	74.6	1.8	15.0	82.8	47.4	15.4	15.3	52.3

Thanks!