

Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection

Authors: Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang

ICLR 2024, 420 citations

Presented by Group 7: Daniel Cisneros, Suranadi Dodampagamage, Andrew El-Kommos, Bradley Racey, Salem Long



Outline

1. Background
2. Grounding DINO
3. Architecture
4. Experiments
5. Applications



Background

- Traditional Object Detection Models trained on fixed set of classes
- Adding new classes is expensive and time consuming
- Not flexible
- Open-set object detection allows the model to identify objects of unseen classes.



Closed-set Detection

Open-set Detection

Detect objects of pre-defined classes

COCO pre-defined categories



bench

person

Standard Object Detection

Detect objects that were not present in training data

Human-input novel categories



ear, lion, bench

worldcup

Zero-Shot Transfer to Novel Categories

Grounding DINO

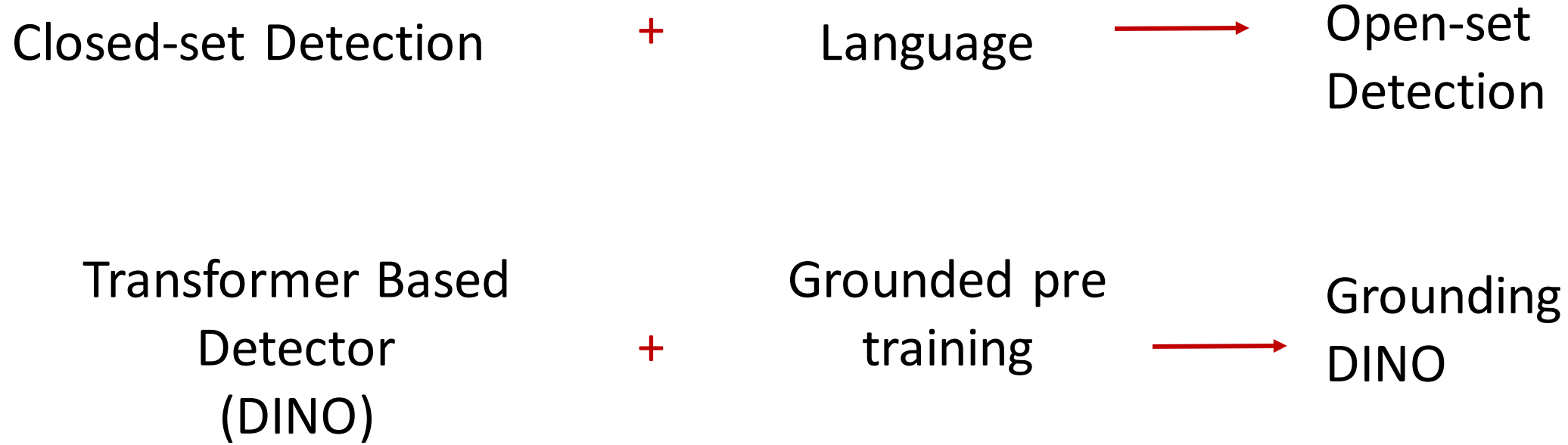


Introduction

- Strong Open-set object detector which combines transformer-based detector DINO with grounded pre-training
- Introduces language for unseen object generalization.
- Detects arbitrary objects with human inputs (Ex: category or referring expressions)



Grounding DINO



➔ DINO - DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection

✘ DINO - Emerging Properties in Self-Supervised Vision Transformers

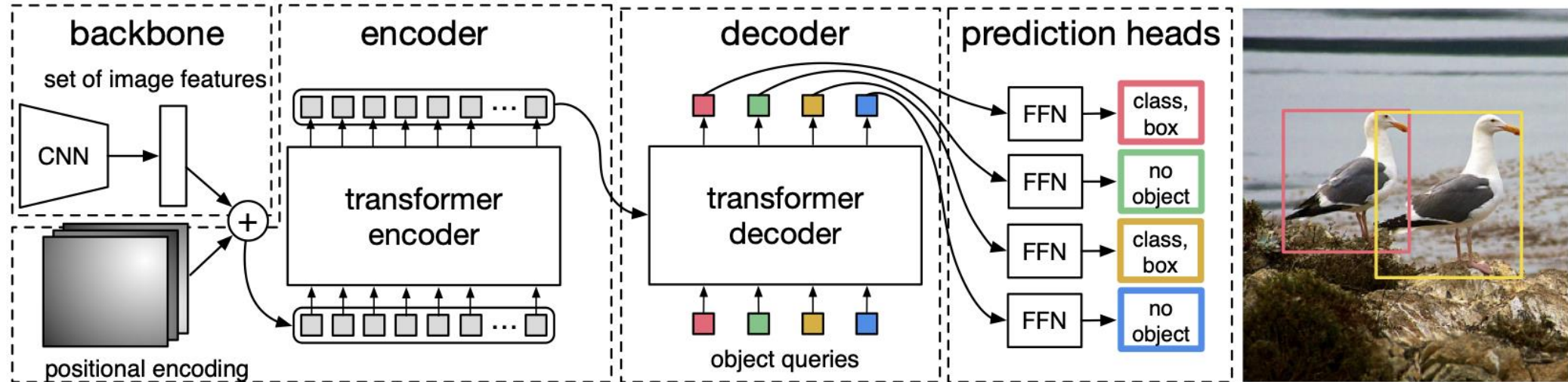


Related Work – Detection Transformers

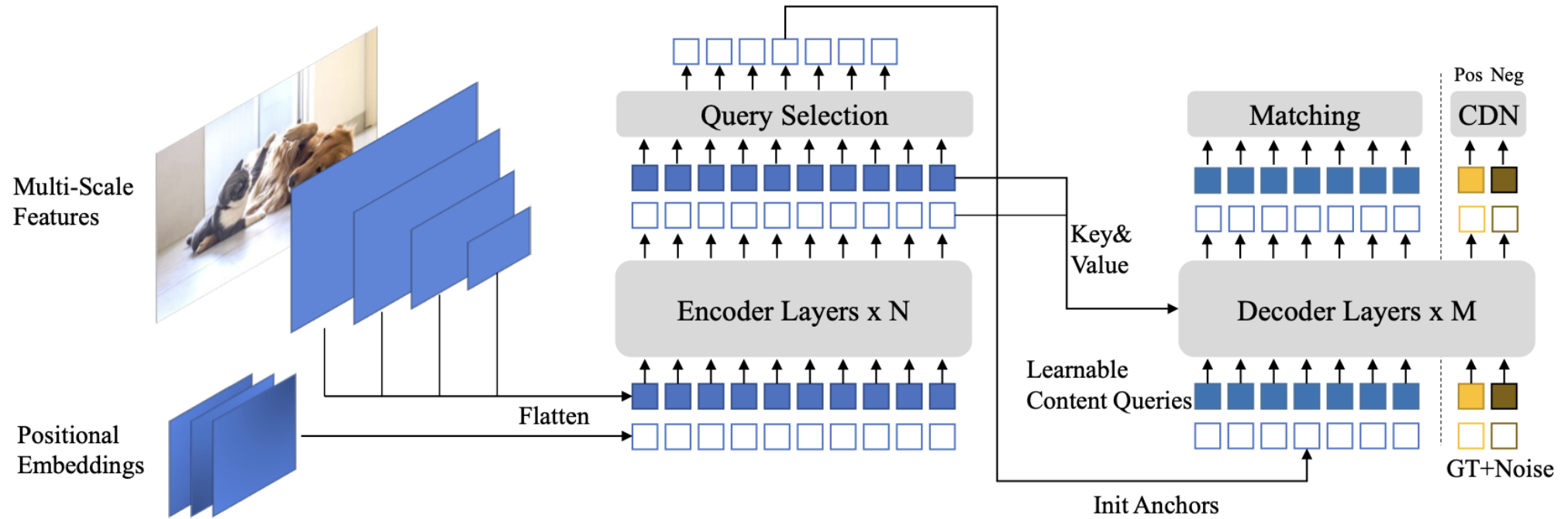
- Grounding DINO builds upon the DETR-like model DINO, transformer-based detector that leverages the strengths of the original DETR model.
- There are many other models built upon DETR
 - DAB-DETR
 - DN-DETR
- Mostly focused on closed-set object detection



Related Work – Detection Transformer (DeTr)



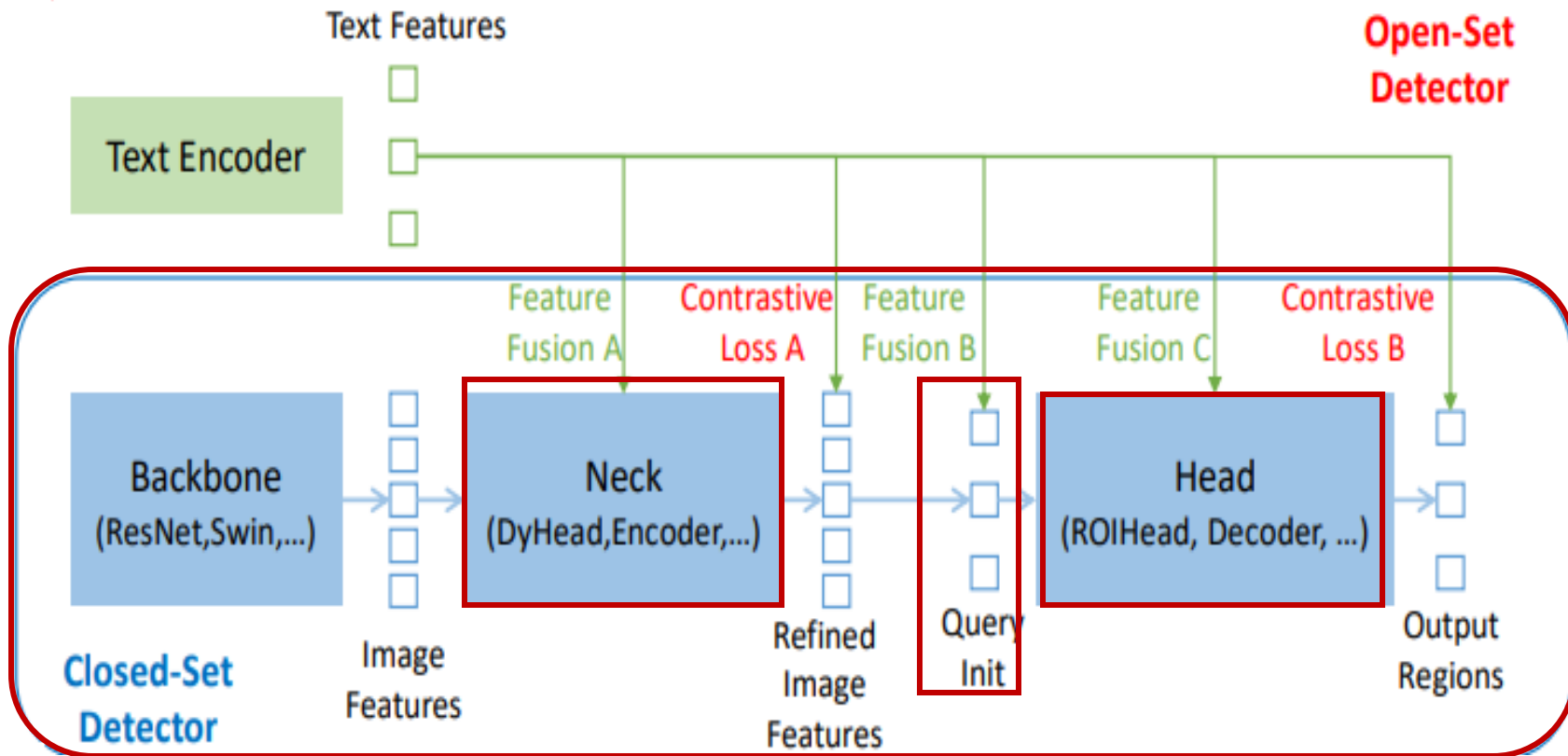
Related Work – DINO architecture



Related Work – Open-set Object Detection

- Most existing open-set detectors created by adapting closed-set detectors for open-set situations using language data.
- Generally, a closed set detector has 3 main modules
 - Backbone – Feature extraction
 - Neck – Feature enhancement
 - Head – Region refinement / box prediction





- Language features can fuse at different phases

- Phase A : Neck
 - Ex - GLIP
- Phase B : Query initialization
 - OV-DETR
- Phase C : Head

In existing open-set detectors,

- Multi-model feature fusion done in partial phases.
- REC is overlooked during evaluation

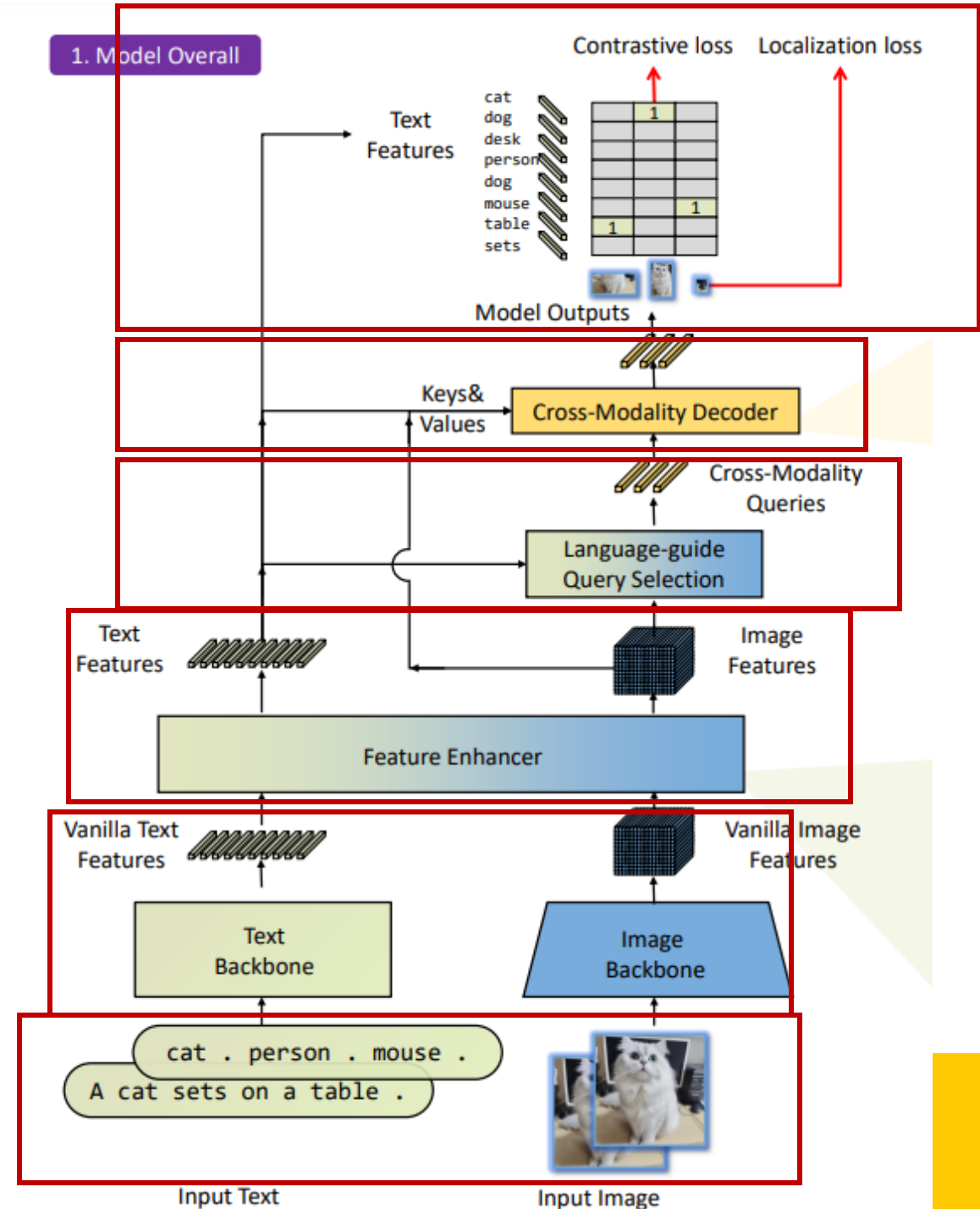
Model Architecture



- Dual-encoder-single-decoder architecture

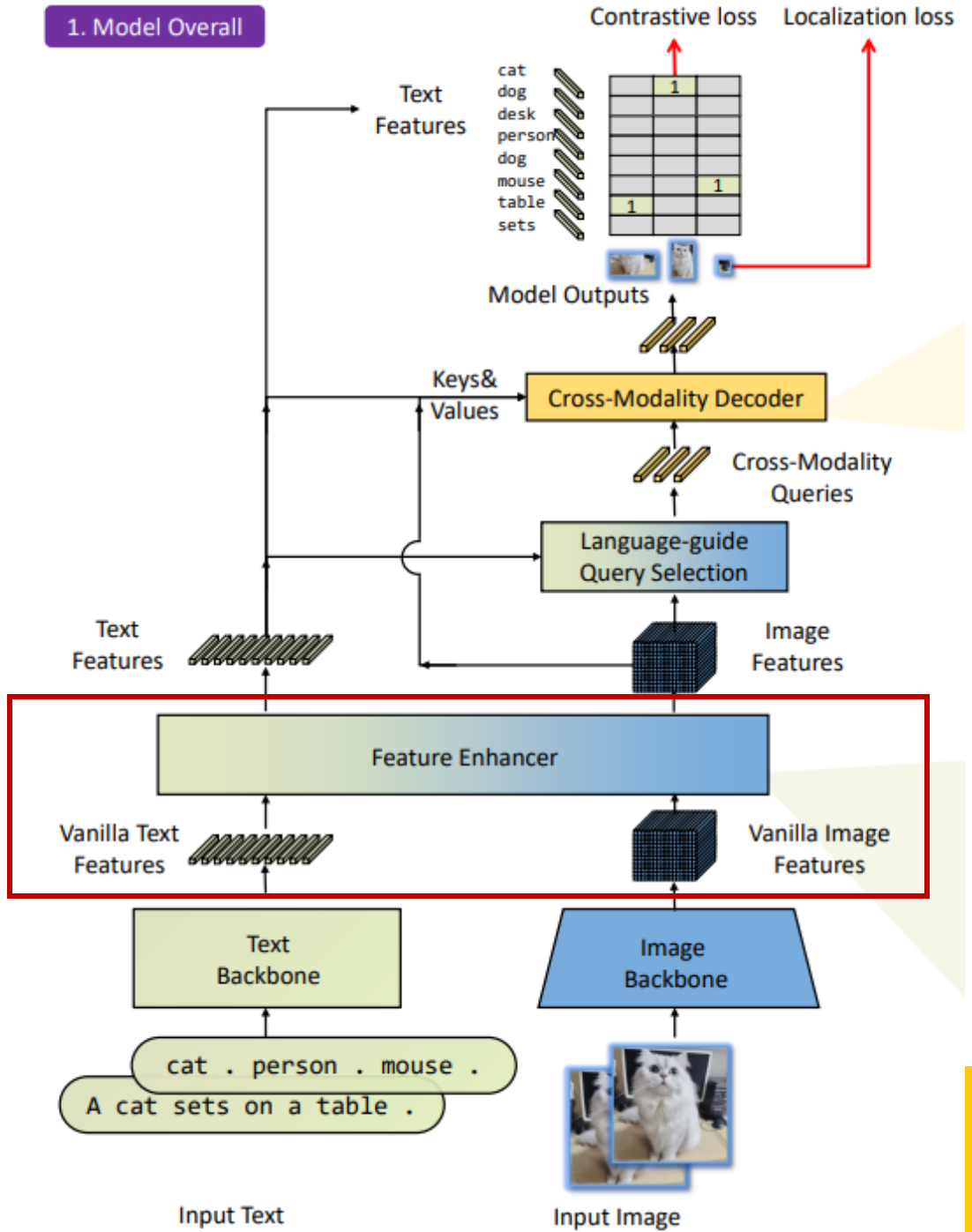
- Key Components :

- Image Backbone
- Text Backbone
- Feature Enhancer
- Language-Guided Query Selection
- Cross-Modality Decoder

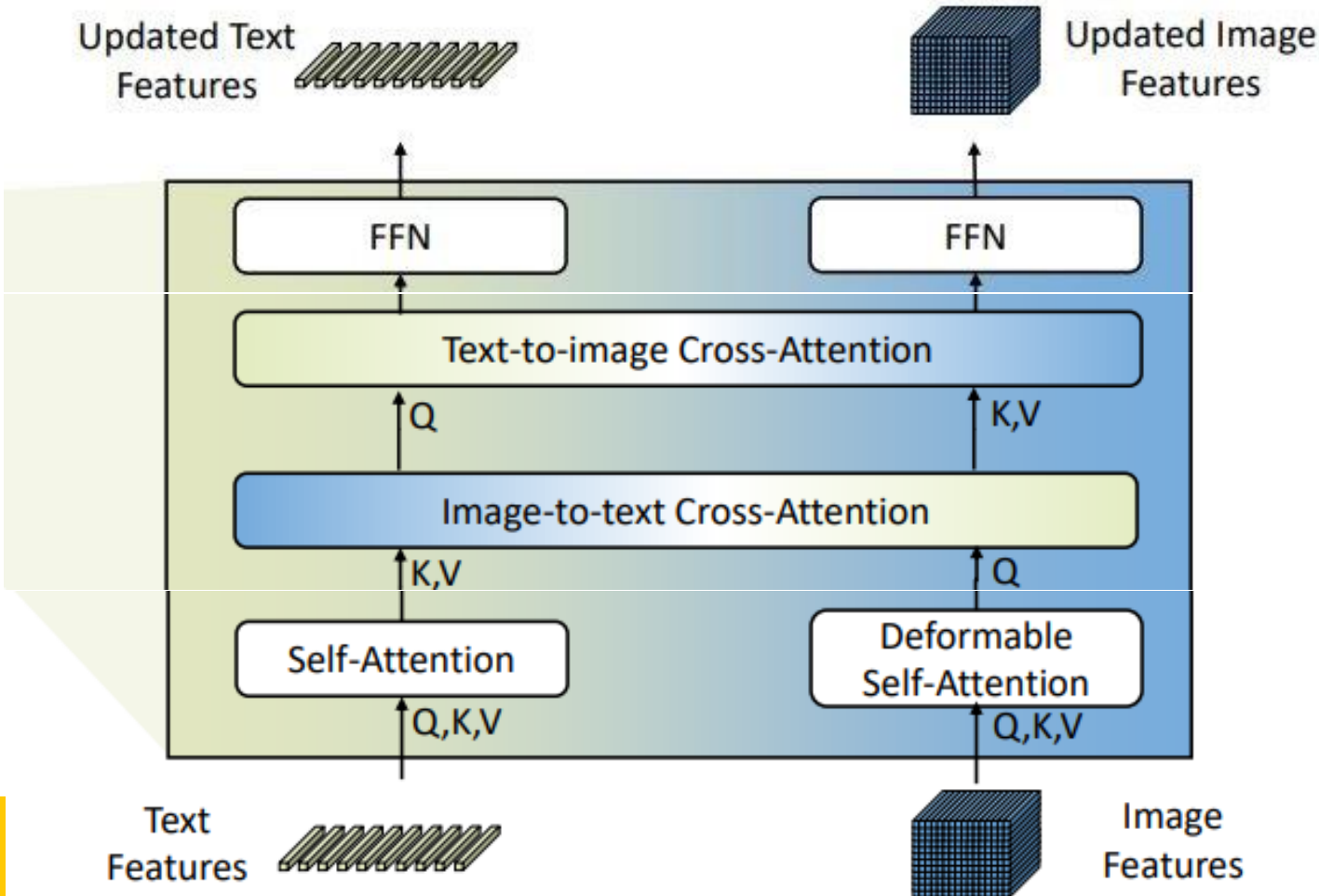


Feature Enhancer

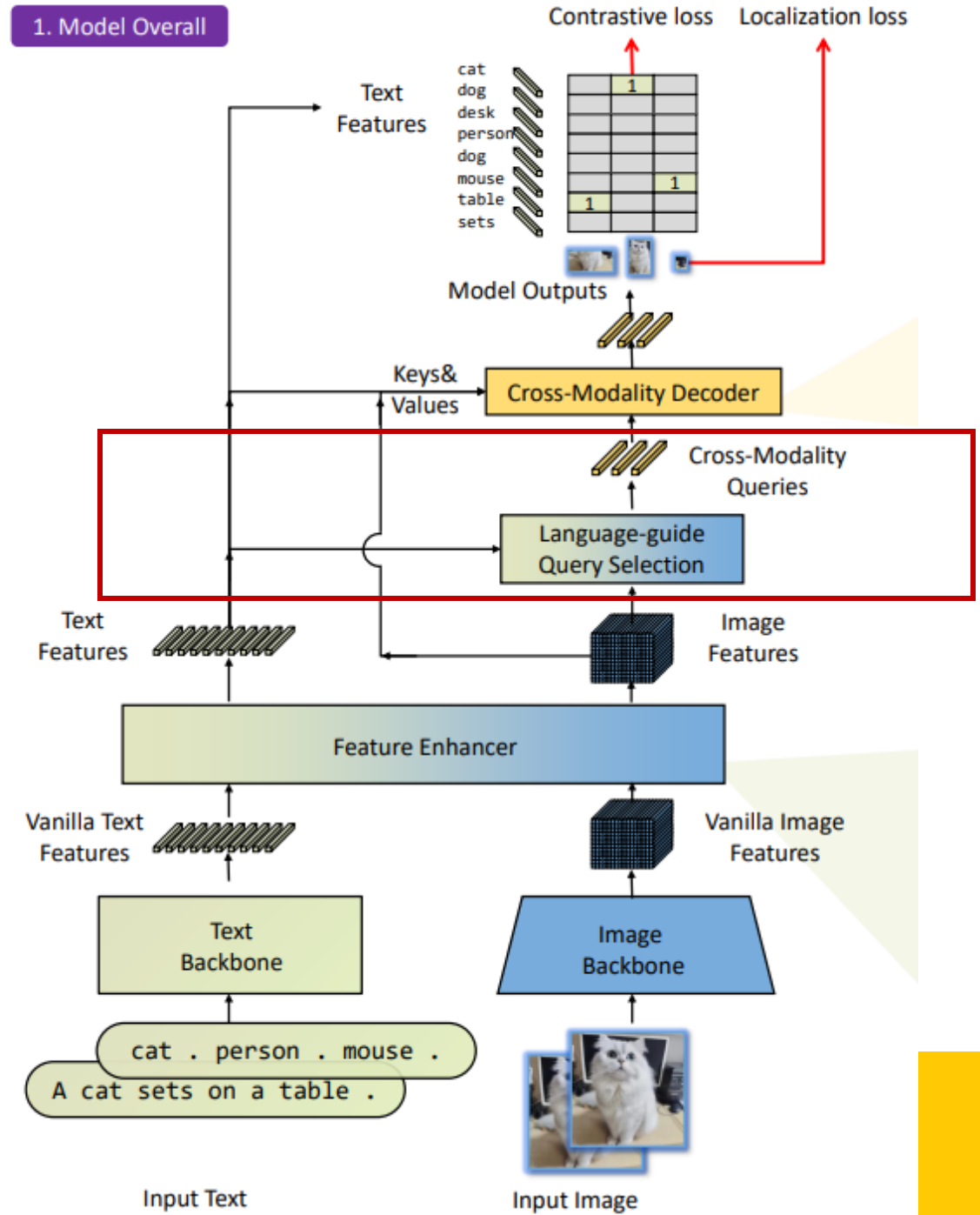
1. Model Overall



Feature Enhancer



Language-Guided Query Selection

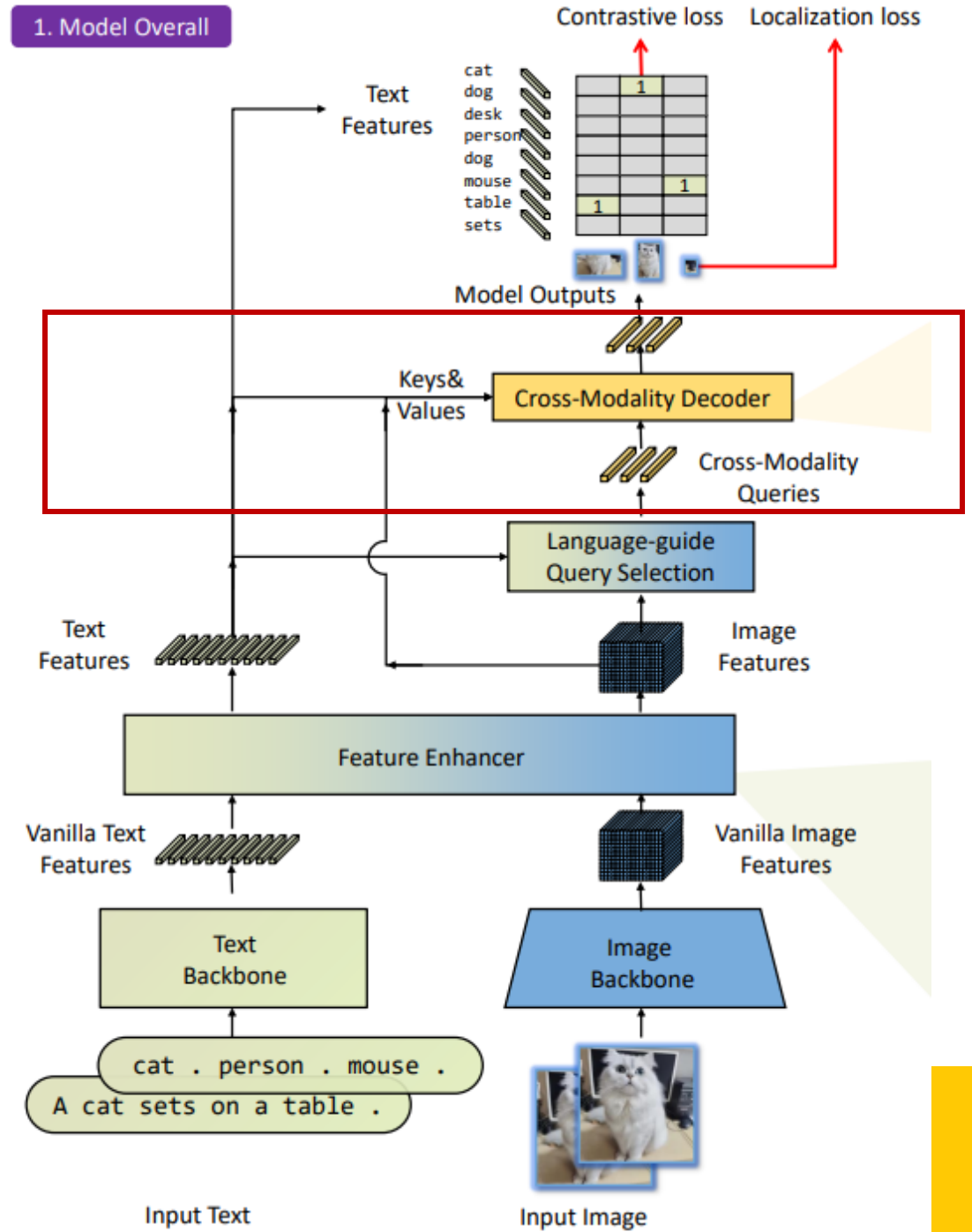


Language-Guided Query Selection

- Select image features that are more relevant to the text queries.
- Outputs cross modality queries for object detection

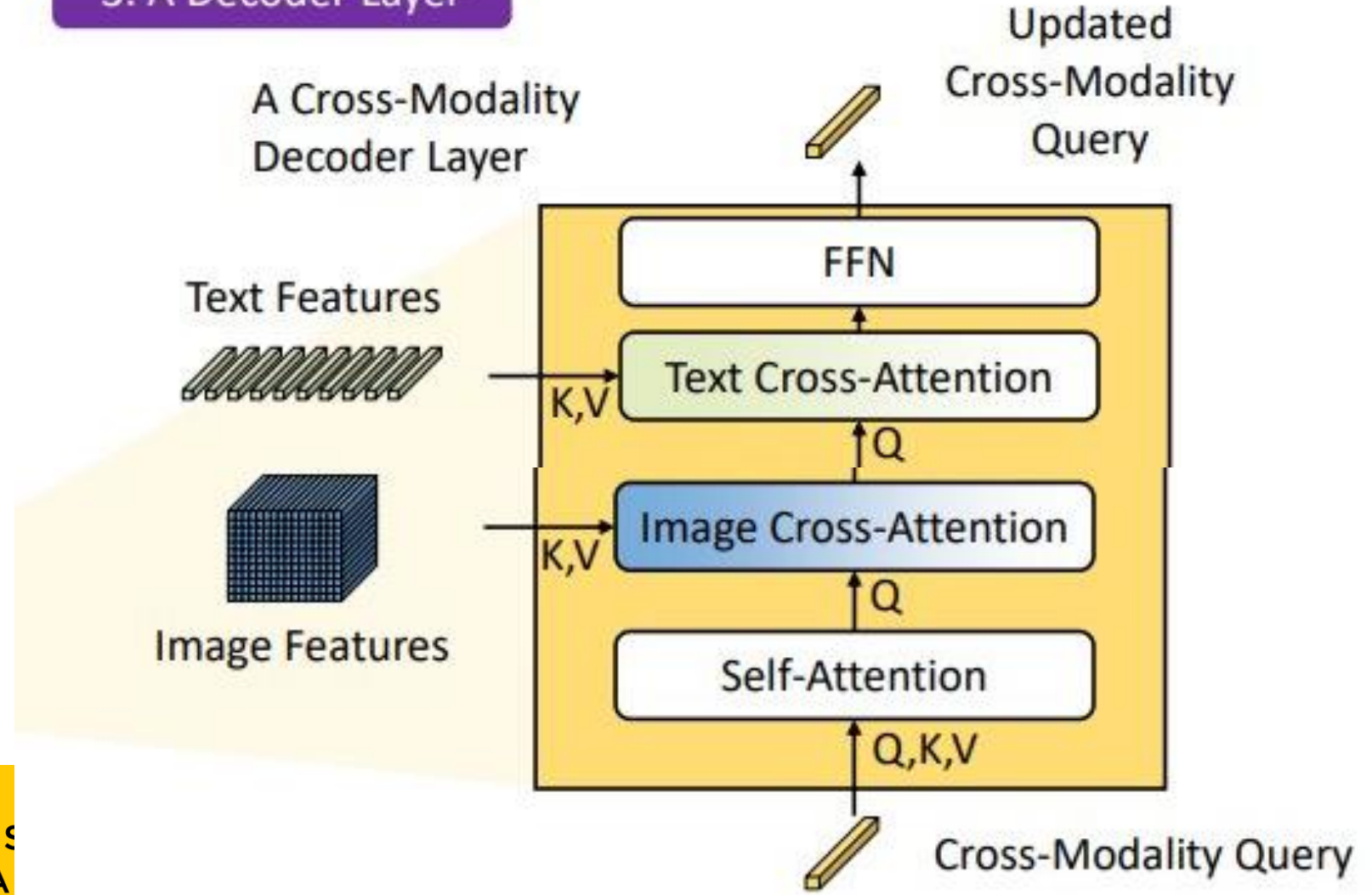


Cross-Modality Decoder



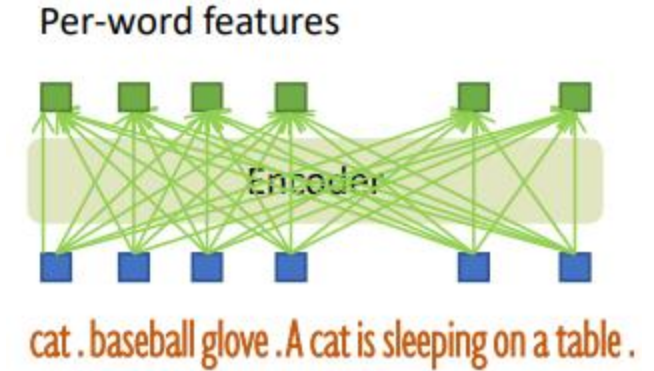
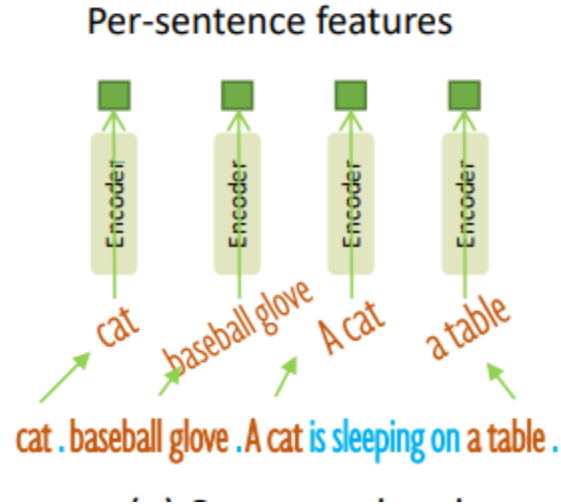
Cross-Modality Decoder

3. A Decoder Layer



Sub-Sentence Level Text Feature

Existing approaches



Sentence Level representation

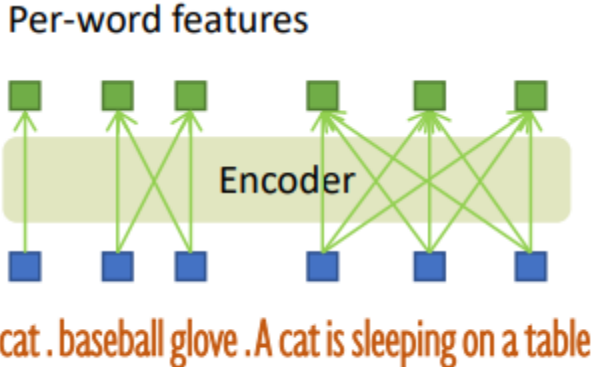
- Treats entire sentence as single unit.
- Removes influence between words and fine-grained information in sentence.

Word level Representation

- Treat each word separately
- Introduce unnecessary dependencies among categories

Sentence Level Text Feature

Proposed Method



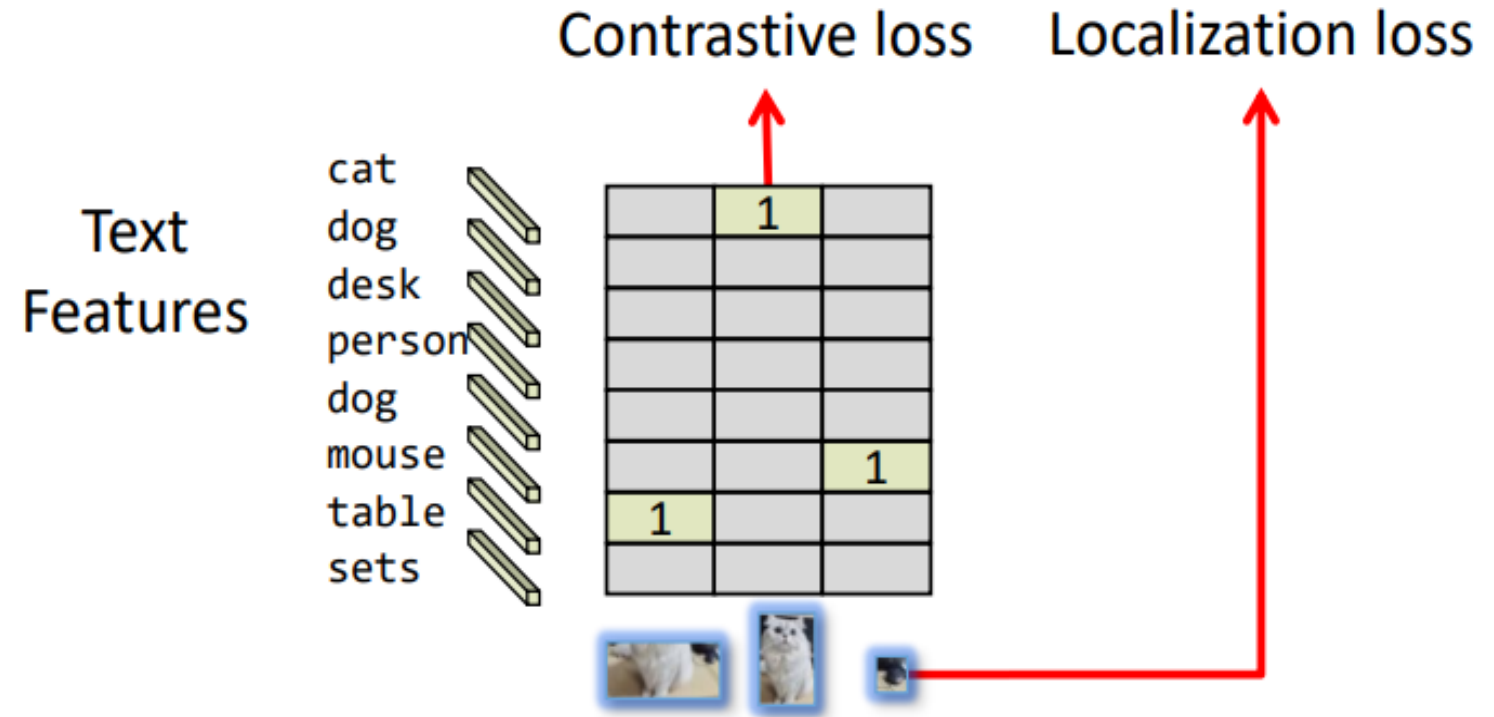
Sub-sentence level Representation

- Maintains per-word features for fine-grained understanding
- Introduces attention masks to block attention among unrelated words



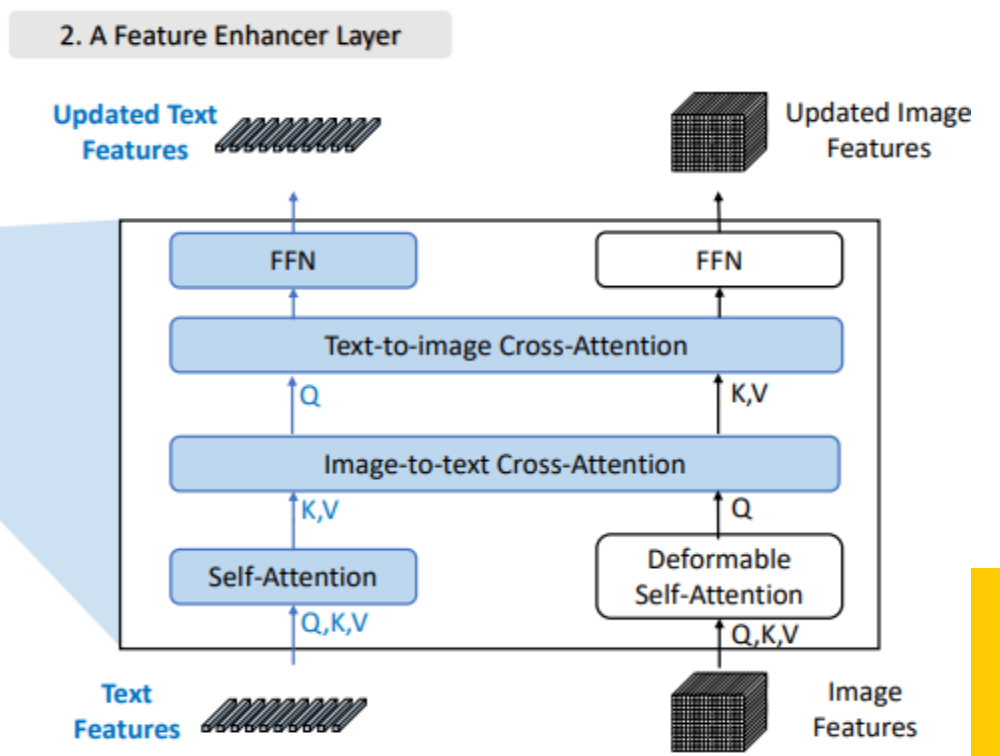
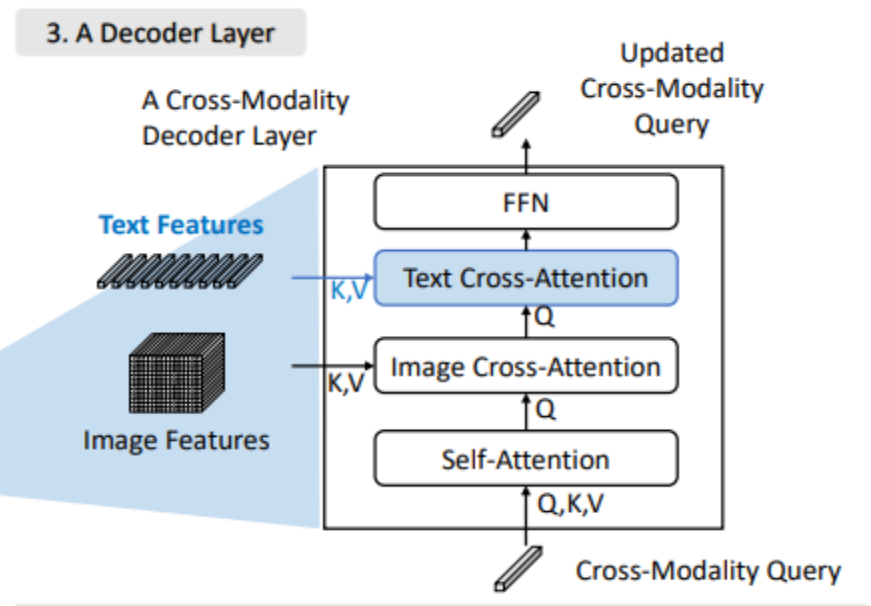
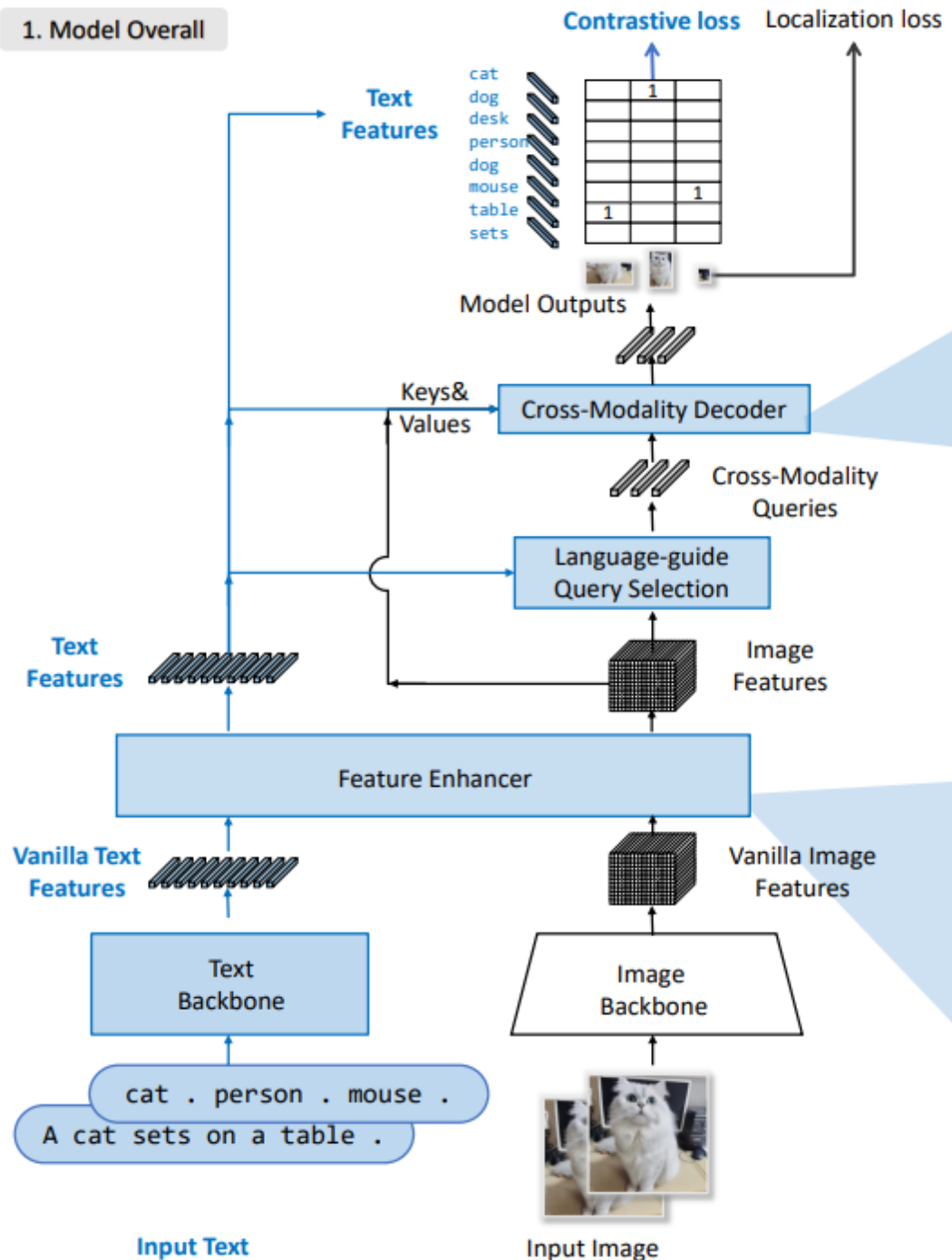
Loss Function

- Contrastive loss for classification
- L1 loss and GIOU for bounding box regression



DINO vs Grounding DINO





Experiments



Zero-Shot Transfer of Grounding DINO

COCO Benchmark

Model	Backbone	Pre-Training Data	Zero-Shot 2017val	Fine-Tuning 2017val/test-dev
Faster R-CNN	RN50-FPN	-	-	40.2 / -
Faster R-CNN	RN101-FPN	-	-	42.0 / -
DyHead-T [5]	Swin-T	-	-	49.7 / -
DyHead-L [5]	Swin-L	-	-	58.4 / 58.7
DyHead-L [5]	Swin-L	O365,ImageNet21K	-	60.3 / 60.6
SoftTeacher [52]	Swin-L	O365,SS-COCO	-	60.7 / 61.3
DINO(Swin-L) [58]	Swin-L	O365	-	62.5 / -
DyHead-T† [5]	Swin-T	O365	43.6	53.3 / -
GLIP-T (B) [26]	Swin-T	O365	44.9	53.8 / -
GLIP-T (C) [26]	Swin-T	O365,GoldG	46.7	55.1 / -
GLIP-L [26]	Swin-L	FourODs,GoldG,Cap24M	49.8	60.8 / 61.0
DINO(Swin-T)† [58]	Swin-T	O365	46.2	56.9 / -
Grounding-DINO-T (Ours)	Swin-T	O365	46.7	56.9 / -
Grounding-DINO-T (Ours)	Swin-T	O365,GoldG	48.1	57.1 / -
Grounding-DINO-T (Ours)	Swin-T	O365,GoldG,Cap4M	48.4	57.2 / -
Grounding-DINO-L (Ours)	Swin-L	O365,OI [19],GoldG	52.5	62.6 / 62.7 (63.0 / 63.0)*
Grounding-DINO-L (Ours)	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	60.7	62.6 / -



LVIS Benchmark

Model	Backbone	Pre-Training Data	MiniVal [18]	
			AP	APr/APc/APf
MDETR [18]*	RN101	GoldG,RefC	24.2	20.9/24.9/24.3
Mask R-CNN [18]*	RN101	-	33.3	26.3/34.0/33.9
GLIP-T (C)	Swin-T	O365,GoldG	24.9	17.7/19.5/31.0
GLIP-T	Swin-T	O365,GoldG,Cap4M	26.0	20.8/21.4/31.0
Grounding-DINO-T	Swin-T	O365,GoldG	25.6	14.4/19.6/32.2
Grounding-DINO-T	Swin-T	O365,GoldG,Cap4M	27.4	18.1/23.3/32.7
Grounding-DINO-L	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	33.9	22.2/30.7/38.8

ODinW Benchmark

Model	Language Input	Backbone	Model Size	Pre-Training Data	Test	
					AP _{average}	AP _{median}
<i>Zero-Shot Setting</i>						
MDETR [18]	✓	ENB5 [48]	169M	GoldG,RefC	10.7	3.0
OWL-ViT [35]	✓	ViT L/14(CLIP)	>1243M	O365, VG	18.8	9.8
GLIP-T [26]	✓	Swin-T	232M	O365,GoldG,Cap4M	19.6	5.1
OmDet [61]	✓	ConvNeXt-B	230M	COCO,O365,LVIS,PhraseCut	19.7	10.8
GLIPv2-T [59]	✓	Swin-T	232M	O365,GoldG,Cap4M	22.3	8.9
DetCLIP [53]	✓	Swin-L	267M	O365,GoldG,YFCC1M	24.9	18.3
Florence [55]	✓	CoSwinH	≈841M	FLD900M,O365,GoldG	25.8	14.3
Grounding-DINO-T(Ours)	✓	Swin-T	172M	O365,GoldG	20.0	9.5
Grounding-DINO-T(Ours)	✓	Swin-T	172M	O365,GoldG,Cap4M	22.3	11.9
Grounding DINO L(Ours)	✓	Swin-L	341M	O365,OI,GoldG,Cap4M,COCO,RefC	26.1	18.4
<i>Few-Shot Setting</i>						
DyHead-T [5]	✗	Swin-T	≈100M	O365	37.5	36.7
GLIP-T [26]	✓	Swin-T	232M	O365,GoldG,Cap4M	38.9	33.7
DINO-Swin-T [58]	✗	Swin-T	49M	O365	41.2	41.1
OmDet [61]	✓	ConvNeXt-B	230M	COCO,O365,LVIS,PhraseCut	42.4	41.7
Grounding-DINO-T(Ours)	✓	Swin-T	172M	O365,GoldG	46.4	51.1
<i>Full-Shot Setting</i>						
GLIP-T [26]	✓	Swin-T	232M	O365,GoldG,Cap4M	62.6	62.1
DyHead-T [5]	✗	Swin-T	≈100M	O365	63.2	64.9
DINO-Swin-T [58]	✗	Swin-T	49M	O365	66.7	68.5
OmDet [61]	✓	ConvNeXt-B	230M	COCO,O365,LVIS,PhraseCut	67.1	71.2
DINO-Swin-L [58]	✗	Swin-L	218M	O365	68.8	70.7
Grounding-DINO-T(Ours)	✓	Swin-T	172M	O365,GoldG	70.7	76.2

Referring Object Detection

Method	Backbone	Pre-Training Data	Fine-tuning	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
MAttNet [54]	R101	None	✓	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
VGTR [9]	R101	None	✓	79.20	82.32	73.78	63.91	70.09	56.51	65.73	67.23
TransVG [7]	R101	None	✓	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
VILLA.L* [10]	R101	CC, SBU, COCO, VG	✓	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
RefTR [27]	R101	VG	✓	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
MDETR [18]	R101	GoldG,RefC	✓	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
DQ-DETR [46]	R101	GoldG,RefC	✓	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44
GLIP-T(B)	Swin-T	O365,GoldG		49.96	54.69	43.06	49.01	53.44	43.42	65.58	66.08
GLIP-T	Swin-T	O365,GoldG,Cap4M		50.42	54.30	43.83	49.50	52.78	44.59	66.09	66.89
Grounding-DINO-T (Ours)	Swin-T	O365,GoldG		50.41	57.24	43.21	51.40	57.59	45.81	67.46	67.13
Grounding-DINO-T (Ours)	Swin-T	O365,GoldG,RefC		73.98	74.88	59.29	66.81	69.91	56.09	71.06	72.07
Grounding-DINO-T (Ours)	Swin-T	O365,GoldG,RefC	✓	89.19	91.86	85.99	81.09	87.40	74.71	84.15	84.94
Grounding-DINO-L (Ours)*	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	✓	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02

.



Ablations

#ID	Model	COCO minival		LVIS minival
		Zero-Shot	Fine-Tune	Zero-Shot
0	Grounding DINO (Full Model)	46.7	56.9	16.1
1	w/o encoder fusion	45.8	56.1	13.1
2	static query selection	46.3	56.6	13.6
3	w/o text cross-attention	46.1	56.3	14.3
4	word-level text prompt	46.4	56.6	15.6

Transfer from DINO to Grounding DINO

Model	Pre-Train Data		COCO minival Zero-Shot	LVIS minival Zero-Shot	ODinW Zero-Shot
	DINO Pre-Train	Grounded Fine-Tune			
Grounding-DINO-T (from scratch)	-	O365	46.7	16.2	14.5
	-	O365,GoldG	48.1	25.6	20.0
Grounding-DINO-T (from pre-trained DINO)	O365	O365	46.5	17.9	13.6
	O365	O365,GoldG	46.4	26.1	18.5

Applications : Image Editing

- Combine Grounding DINO with Stable Diffusion
- Combine the Grounding DINO with GLIGEN for fine-grained image editing



Combine Grounding DINO with Stable Diffusion

Input

Detection Result

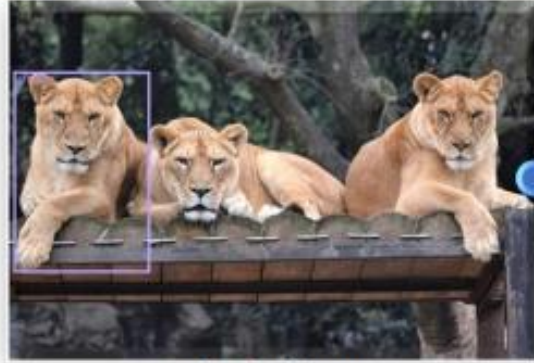
Edited Images



Detection prompt : green mountain
Generation Prompt :red mountain



Detection prompt : the running girl
Generation Prompt (modify background): The Wandering Earth



The left lion



Prompt (modify detected objects): Dog



The bottom man with his head up
Referring Object Detection
(Referring Expression Comprehension)



Prompt (modify background): All people
around the world cheer with a worldcup.

Combine the Grounding DINO with GLIGEN

Input



Detection Result



Edited Images



Detection prompt :dog, cat

Generation Prompt :a cake and a phone

Phrase Prompt: a cake; a phone.



Detection prompt : a sketch person

Generation Prompt :a woman and a man are talking

Phrase Prompt: a woman; a man

Limitations

- Grounding DINO lacks the capability for segmentation task
- Limited training data restricts its performance potential



Future Work

- Conduct larger-scale training to further enhance the scalability
- Improve training of Grounding DINO leveraging pre-trained DINO
- Enhance zero-shot performance in REC tasks



Conclusion

- Extends DINO to open-set object detection
- Introduces better fusion approach for cross-modality information
- Propose sub-Sentence Level Representation for effectively leveraging detection data
- Extends open-set detection to REC tasks



Thank you!

