Grounding Dino

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica
Corso di Laurea Magistrale in AI & Robotics

SAPIENZA
UNIVERSITÀ DI ROMA

Yusupha Juwara

A.Y 2024

1/3 Grounding Dino

- https://www.youtube.com/watch?v=o1t8s5innZ8&ab_channel=WhyML

- For each (Image, Text) pair,
- we first extract vanilla image features and vanilla text features using an image backbone and a text backbone, respectively.
- The two vanilla features are fed into a feature enhancer module for cross-modality feature fusion.
- After obtaining cross-modality text and image features, we use a language-guided query selection module to select cross-modality queries from image features.
- Like the object queries in most DETR-like models, these cross-modality queries will be fed into a cross-modality decoder to probe desired features from the two modal features and update themselves.
- The output queries of the last decoder layer will be used to predict object boxes and extract corresponding phrases.

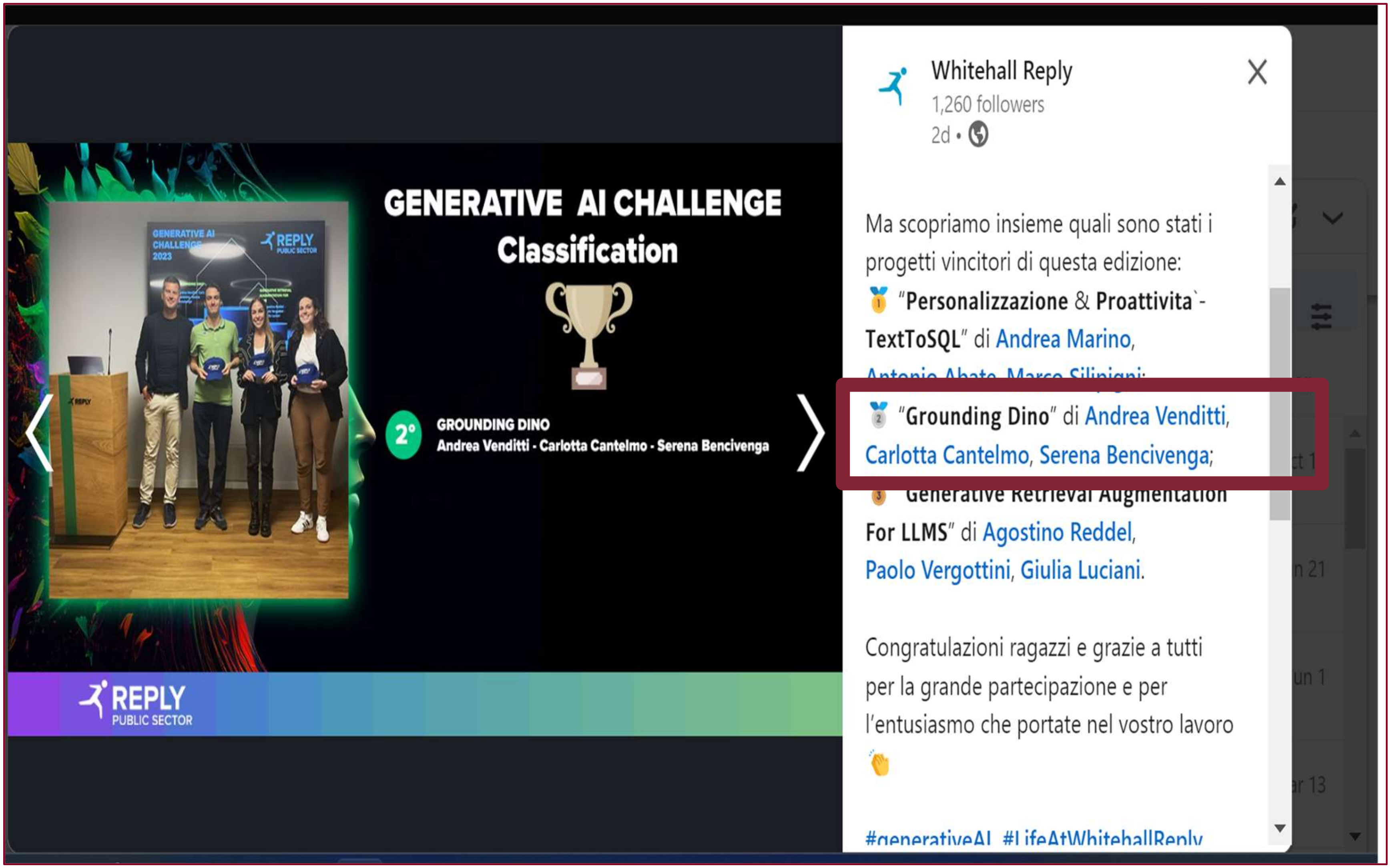
2/3 Grounding Dino

- Open-set object detector can detect arbitrary objects with human inputs such as category names or referring expressions. -> language + closed-set object detector
- REC typically involves understanding and localizing objects in an image based on natural language descriptions or references, such as "the red ball" or "the object on the left." -> Describing objects with attributes
- 3 important modules, a backbone for feature extraction, a neck for feature enhancement, and a head for box prediction (cross modality).
- Dual-encoder-single-decoder architecture
  - Image backbone -> image feature extraction
  - Text backbone -> text feature extraction
- contrastive loss between object regions/queries and text features
  - encourages the model to learn to associate the object queries with relevant text information
- Bounding box regression loss
  - measures how well the predicted bounding box coordinates align with the ground truth bounding boxes.
  - Losses used: L1 loss and Generalized Intersection over Union (GIOU) loss

3/3 Grounding Dino

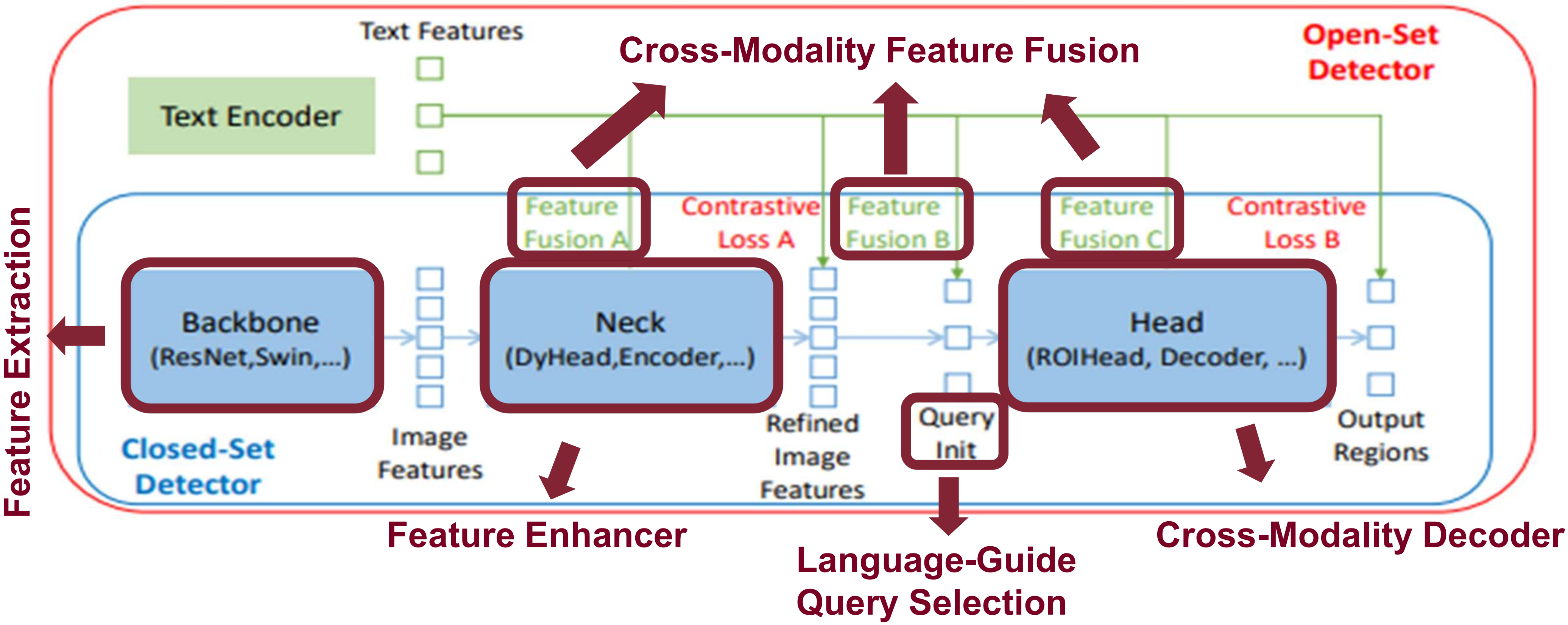
- To make the closed-set detector capable of detecting “novel objects”, it needs to learn "language-aware region embeddings." This implies that the model should understand the relationship between language (human-provided descriptions or labels) and regions of interest within an image.
- Each region of interest to “novel categories”
- Outputs multiple pairs of object boxes and noun phrases for a given (Image, Text) pair
- The classification of regions into novel categories is done in a "language-aware semantic space." This means that language descriptions and image regions are connected in a way that enables the model to understand and use language to categorize or identify objects.
- Language-Guided Query Selection
  - Aim -> Detect objects from an image specified by an input text
  - select features that are more relevant to the input text as decoder queries.
  - outputs num-query (900) indices to extract features to initialize queries.

Implementation



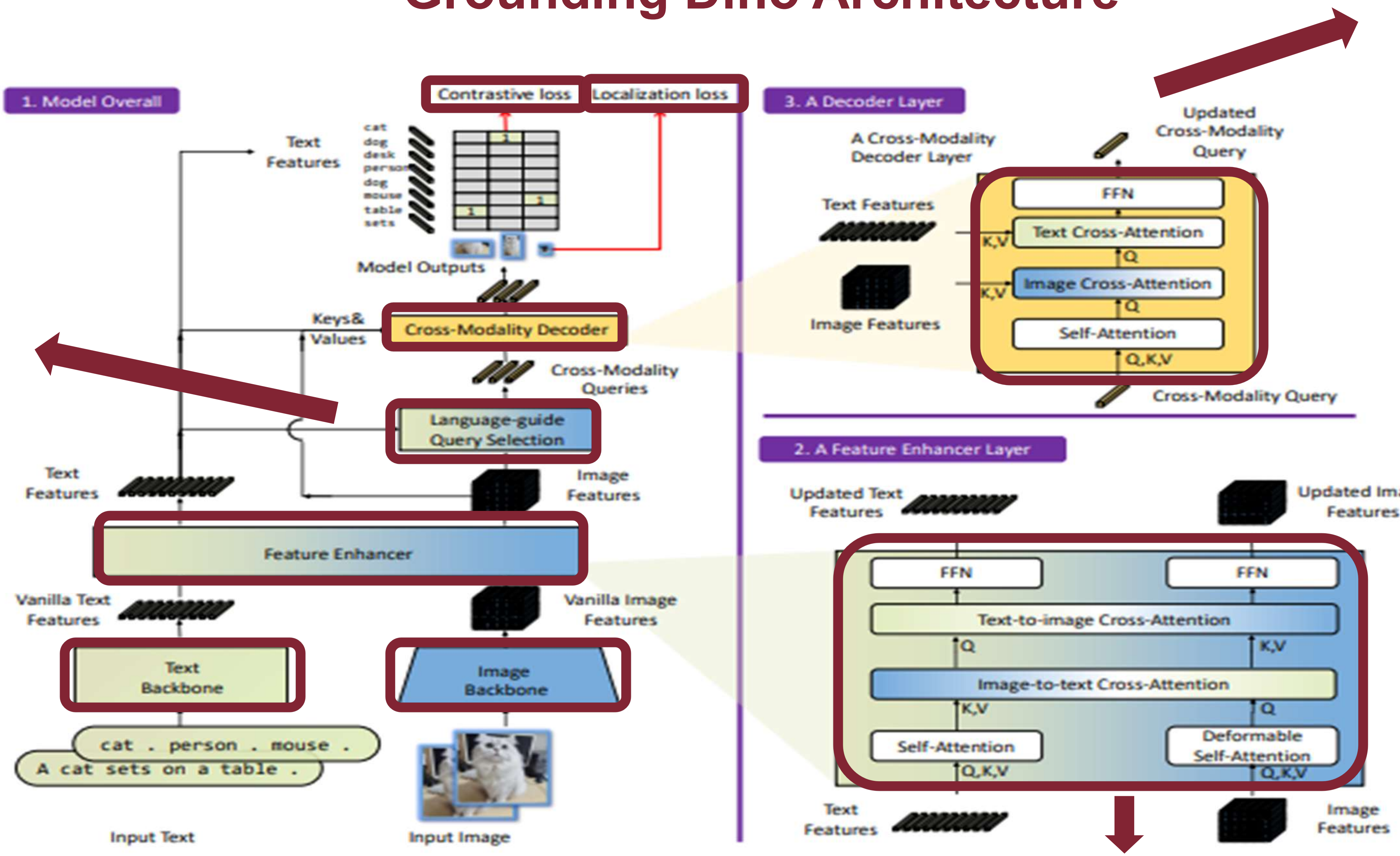
Grounding Dino

- Open-set object detector
- Referring expression comprehension (REC)
  - Describing objects with attributes (noun phrases)
- Zero-Shot Transfer for Model Generalization
- Dual-encoder-single-decoder architecture
  - Next page



Grounding Dino Architecture

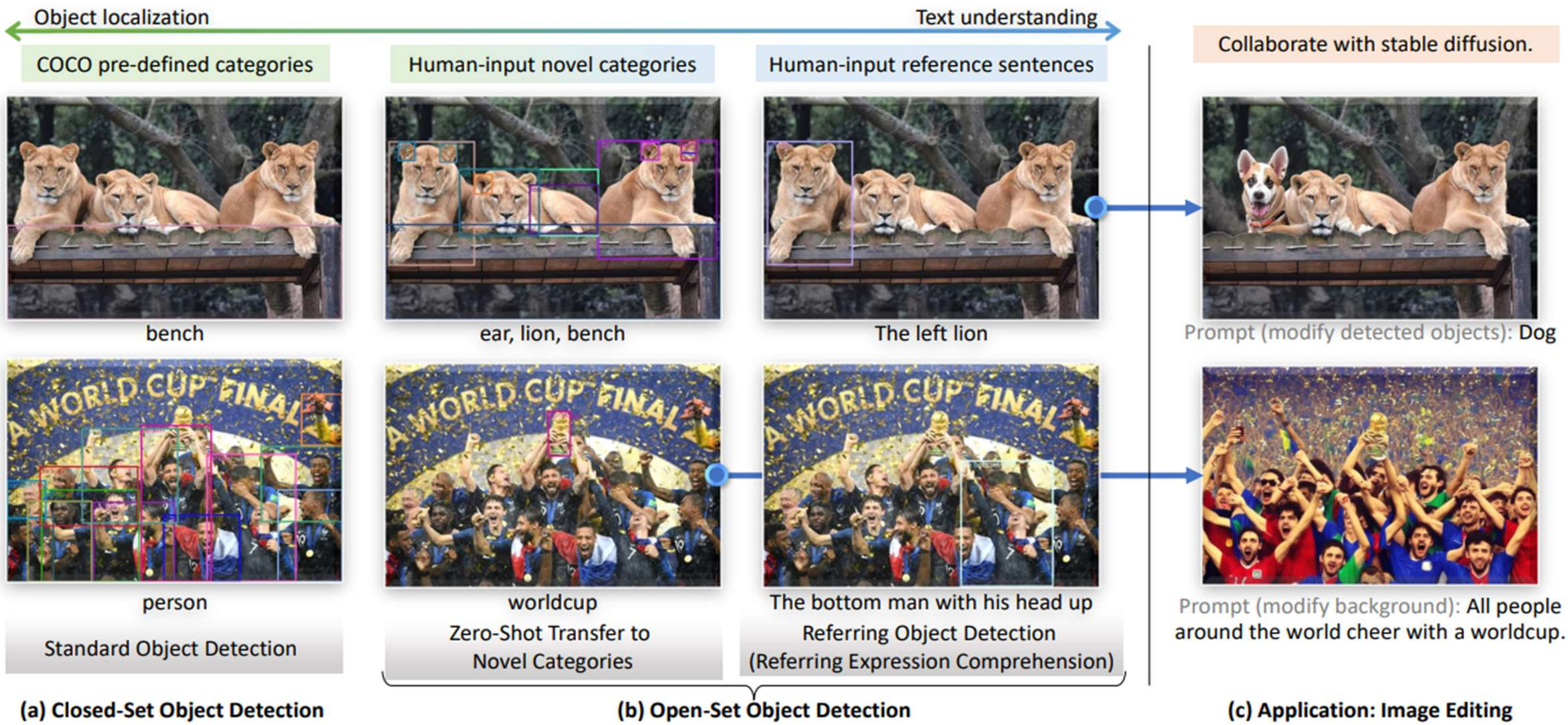
select cross-modality queries from image features



Outputs multiple pairs of object boxes and noun phrases

cross-modality feature fusion

Grounding Dino Example



Grounding Dino Summary

Model	Model Design: Base Detector	Model Design: Fusion Phases (Fig. 2)	Model Design: use CLIP	Text Prompt: Represent. Level (Sec. 3.4)	Closed-Set Settings: COCO	Zero-Shot Transfer: COCO	Zero-Shot Transfer: LVIS	Zero-Shot Transfer: ODinW	Referring Detection: RefCOCO/+/g
ViLD [13]	Mask R-CNN [15]	-	✓	sentence	✓	partial label	partial label		
RegionCLIP [62]	Faster RCNN [39]	-	✓	sentence	✓	partial label	partial label		
FindIt [21]	Faster RCNN [39]	A		sentence	✓	partial label			fine-tune
MDETR [18]	DETR [2]	A,C		word			fine-tune	zero-shot	fine-tune
DQ-DETR [46]	DETR [2]	A,C		word	✓		zero-shot		fine-tune
GLIP [26]	DyHead [5]	A		word	✓	zero-shot	zero-shot	zero-shot	
GLIPv2 [59]	DyHead [5]	A		word	✓	zero-shot	zero-shot	zero-shot	
OV-DETR [56]	Deformable DETR [64]	B	✓	sentence	✓	partial label	partial label		
OWL-ViT [35]	-	-	✓	sentence	✓	partial label	partial label	zero-shot	
DetCLIP [53]	ATSS [60]	-	✓	sentence			zero-shot	zero-shot	
OmDet [61]	Sparse R-CNN [47]	C	✓	sentence	✓			zero-shot	
Grounding DINO (Ours)	DINO [58]	A,B,C		sub-sentence	✓	zero-shot	zero-shot	zero-shot	zero-shot

A comparison of open-set object detectors.

References

- Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." *arXiv preprint arXiv:2303.05499* (2023).