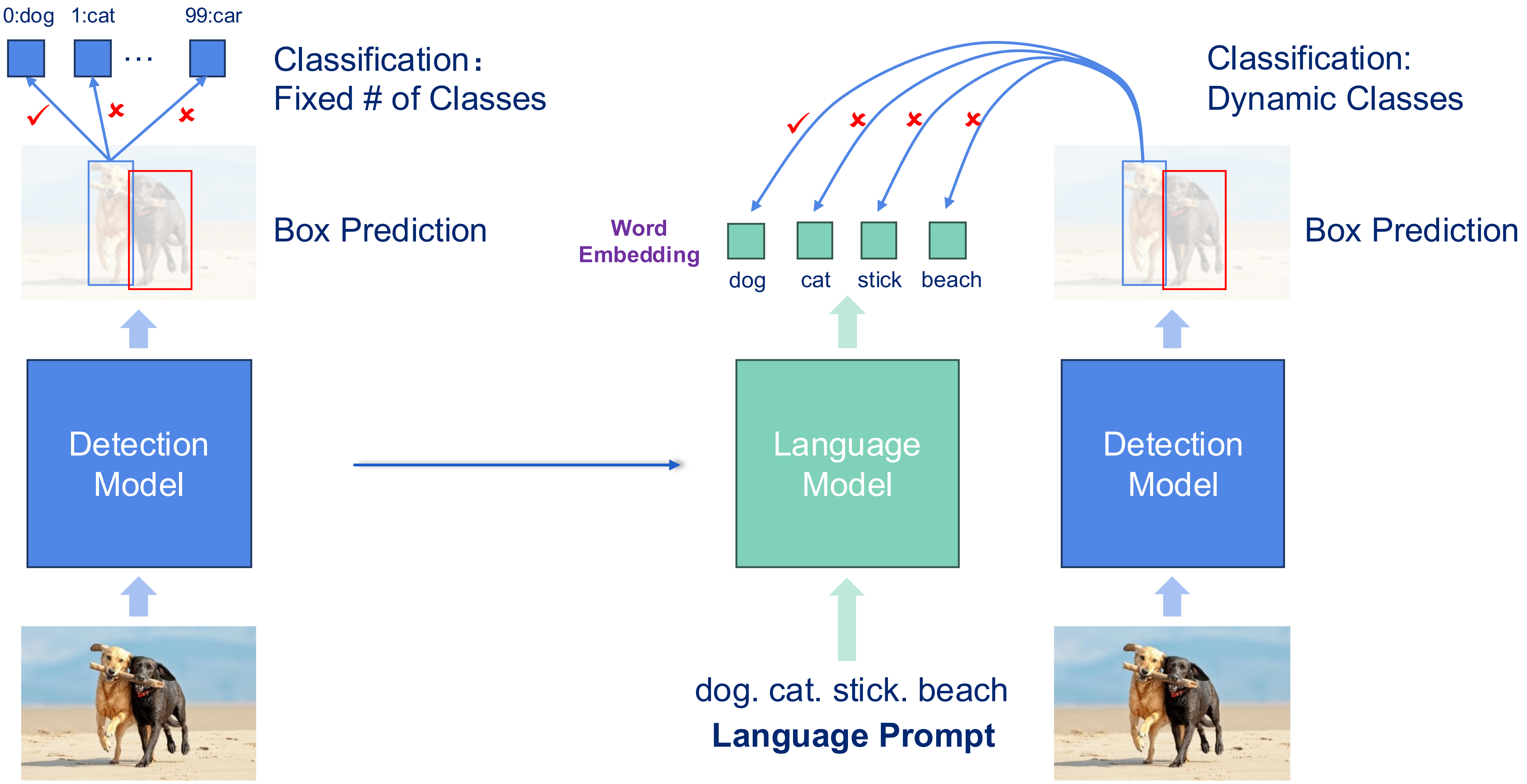
# T-Rex2: Towards Generic Object Detection via Text-Visual Prompt Synergy

蒋擎

3-29

# Paradigm Shift in Object Detection



0:dog 1:cat … 99:car

✓ ✗ ✗

Classification：
Fixed # of Classes

Box Prediction

Detection Model

**Word Embedding**

dog cat stick beach

✓ ✗ ✗ ✗

Classification:
Dynamic Classes

Box Prediction

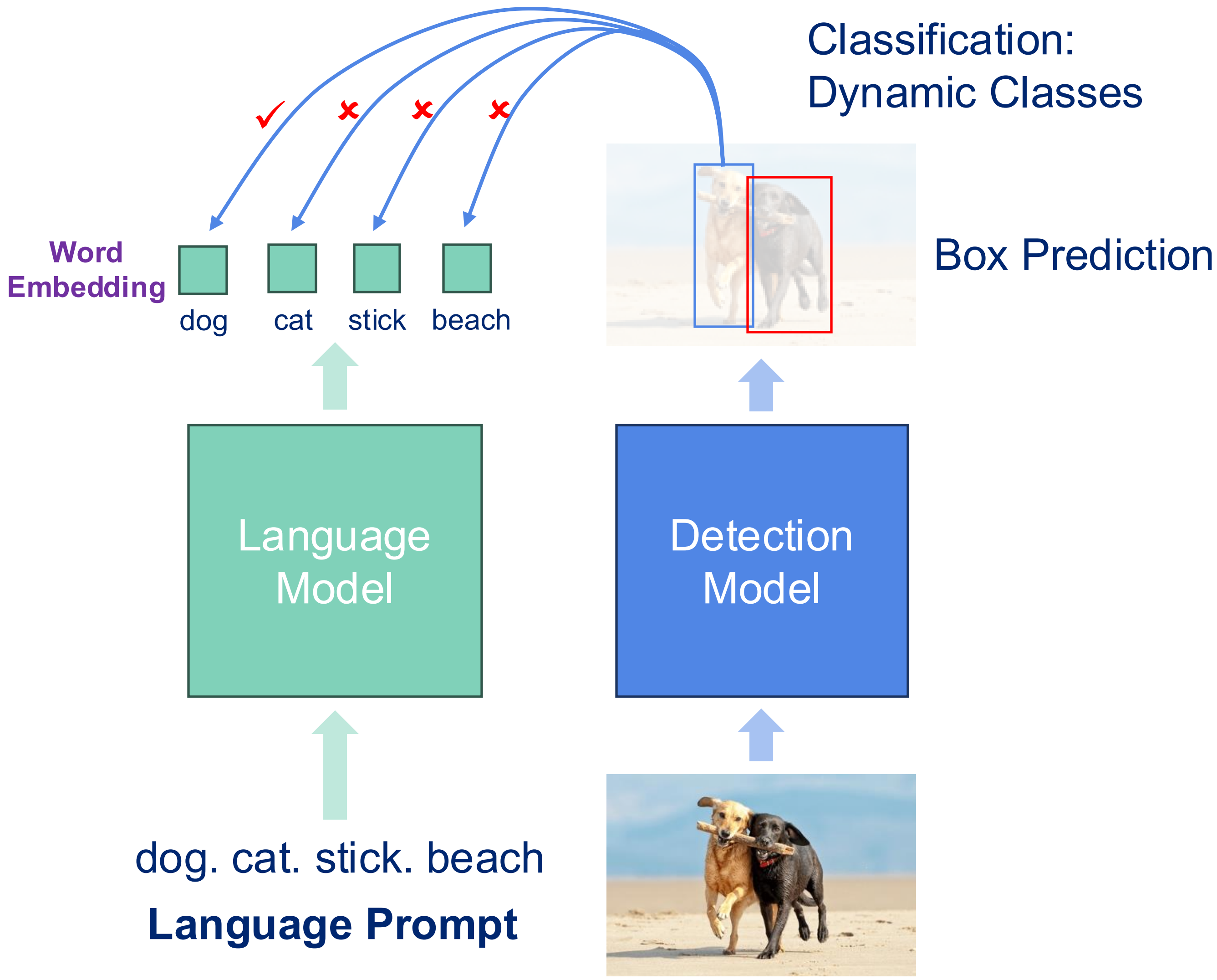
Language Model

Detection Model

dog. cat. stick. beach

**Language Prompt**

# Text Prompt v.s. Visual Prompt



Classification:
Dynamic Classes

✓ ✗ ✗ ✗

**Word Embedding**

dog cat stick beach

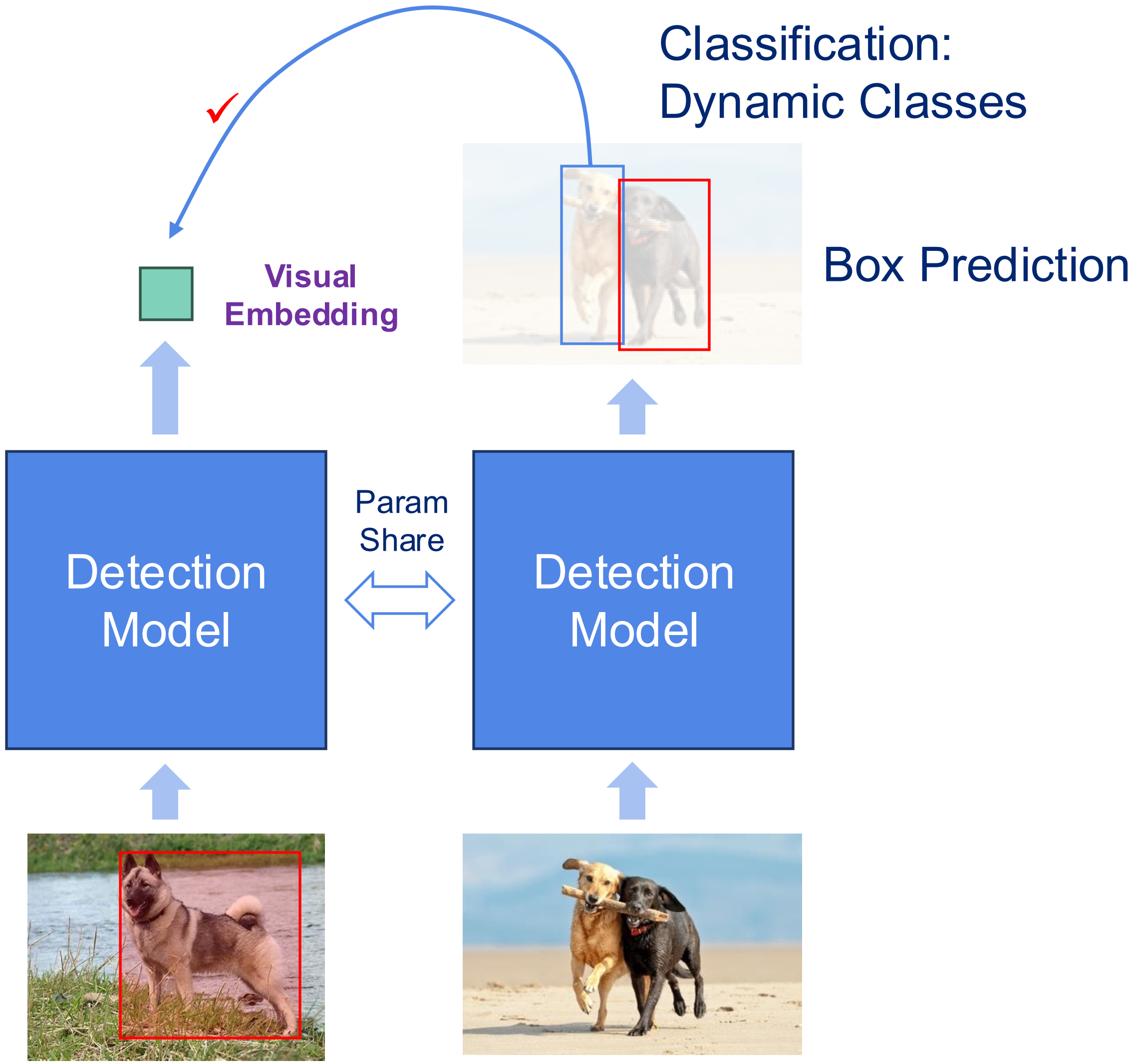
Box Prediction

Language Model

Detection Model

dog. cat. stick. beach

**Language Prompt**



Classification:
Dynamic Classes

✓

**Visual Embedding**

Box Prediction

Detection Model

Param Share

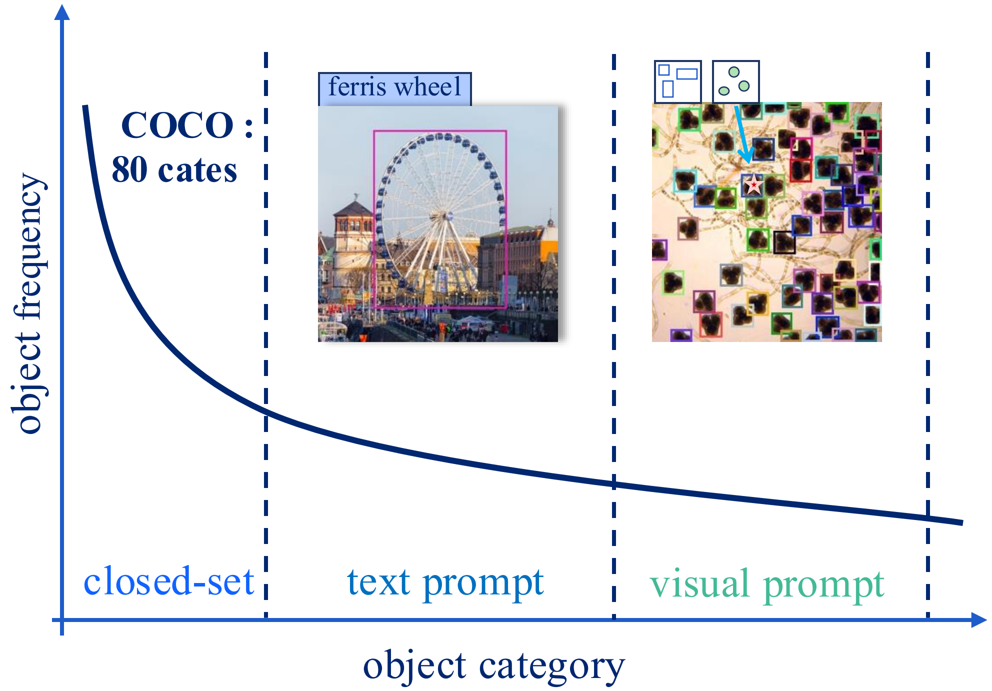
Detection Model

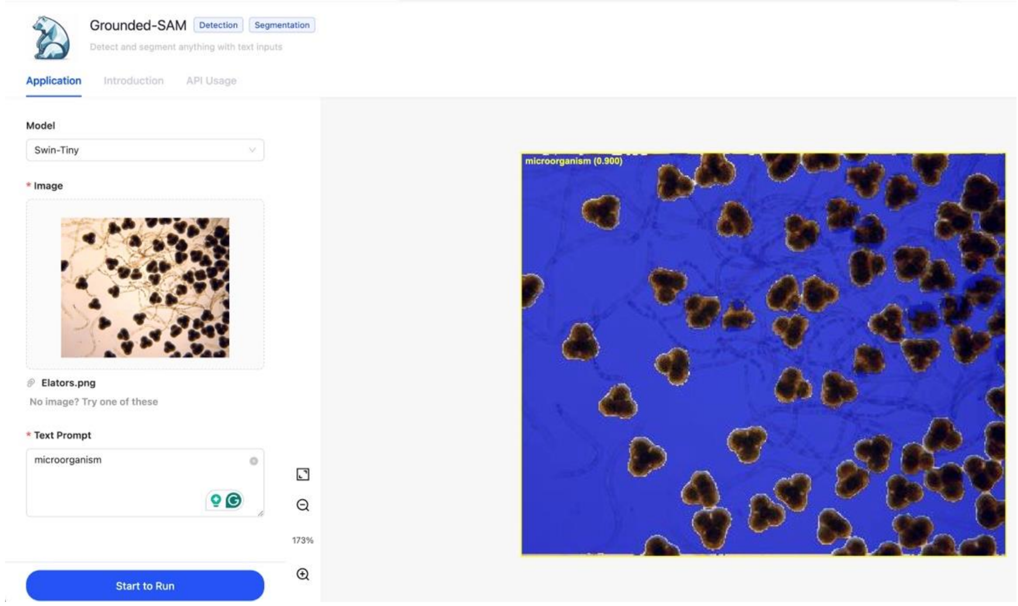
**Visual Prompt**

# Text Prompt v.s. Visual Prompt

## Text Prompt

- describe objects in natural language
- require modality alignment, suffers from long-tailed data shortage
- fall short in describe object that are hard to describe in language





Model: Swin-Tiny

* Image: Elators.png

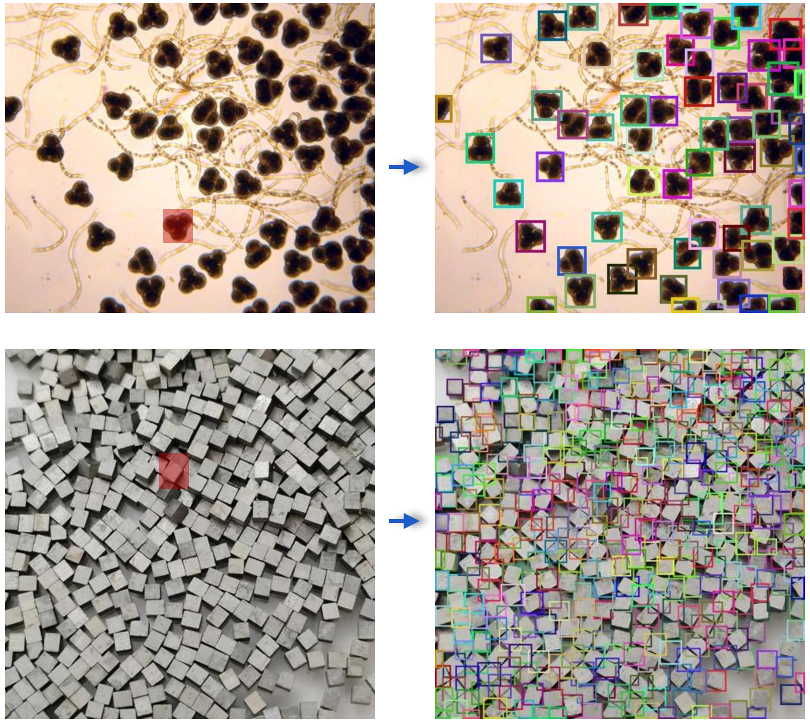
No image? Try one of these

* Text Prompt: microorganism

# Text Prompt v.s. Visual Prompt

## Visual Prompt

- describe objects through visual examples
- less effective at capturing the general concept

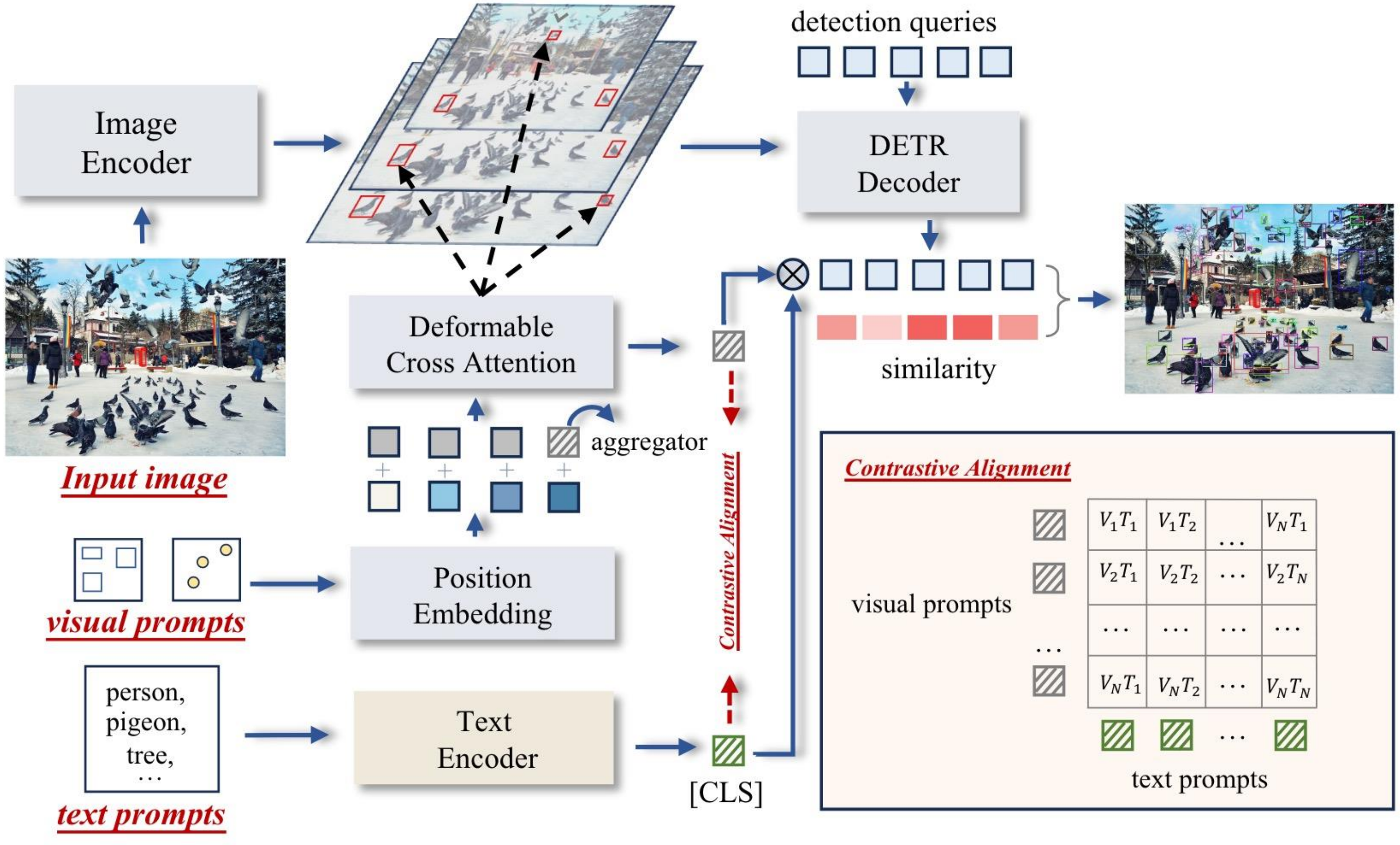


“ Dog ”

require many examples to convey a general concept

# T-Rex2: Visual-Text Prompt Synergy

# T-Rex2: Method



| | | | |
|---|---|---|---|
| $V_1T_1$ | $V_1T_2$ | … | $V_NT_1$ |
| $V_2T_1$ | $V_2T_2$ | … | $V_2T_N$ |
| … | … | … | … |
| $V_NT_1$ | $V_NT_2$ | … | $V_NT_N$ |

DINO-based End-to-End model

**Visual Prompt Encoder**: Deformable Cross Attention

$$B = \text{Linear}(\text{PE}(b_1, ...b_K); \theta_B) : \mathbb{R}^{K \times 4D} \rightarrow \mathbb{R}^{K \times D}$$

$$P = \text{Linear}(\text{PE}(p_1, ...p_K); \theta_P) : \mathbb{R}^{K \times 2D} \rightarrow \mathbb{R}^{K \times D}$$

$$Q = \begin{cases} \text{Linear}\left(\text{CAT}\left([C; C'], [B; B']\right); \varphi_B\right), \text{box} \\ \text{Linear}\left(\text{CAT}\left([C; C'], [P; P']\right); \varphi_P\right), \text{point} \end{cases}$$

$$Q'_j = \begin{cases} \text{MSDeformAttn}(Q_j, b_j, \{\boldsymbol{f}_i\}_{i=1}^{L}), \text{box} \\ \text{MSDeformAttn}(Q_j, p_j, \{\boldsymbol{f}_i\}_{i=1}^{L}), \text{point} \end{cases}$$

$$V = \text{FFN}(\text{SelfAttn}(Q'))[-1]$$

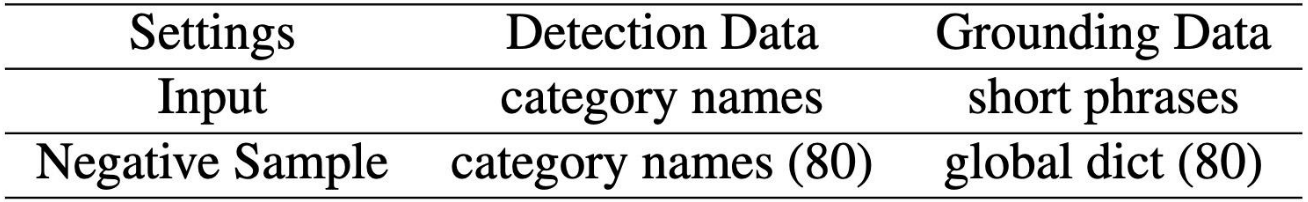
**Text Prompt Encoder**: CLIP

**Modality Alignment**: Contrastive Learning

$$\mathcal{L}_{align} = -\frac{1}{K} \sum_{i=1}^{K} \log \frac{\exp(v_i \cdot t_i)}{\sum_{j=1}^{K} \exp(v_i \cdot t_j)}$$

# T-Rex2: Training Strategy

## Text Prompt



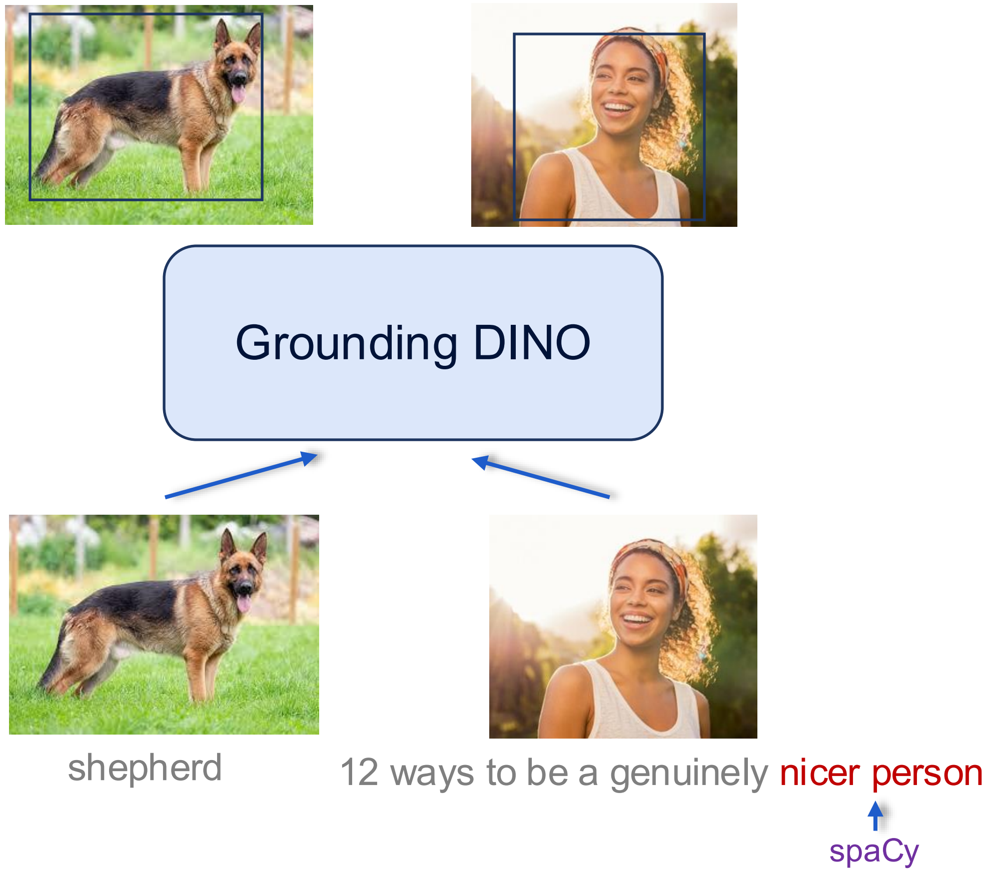
| Settings | Detection Data | Grounding Data |
|---|---|---|
| Input | category names | short phrases |
| Negative Sample | category names (80) | global dict (80) |

## Visual Prompt

- “Current image prompt, current image detect”: for each category in a training set image, we randomly choose between one to all available GT boxes to use as visual prompts. We convert these GT boxes into their center point with a 50% chance for point prompt training.

# T-Rex2: Data Engine

## Text Prompt Data Engine



Grounding DINO

shepherd

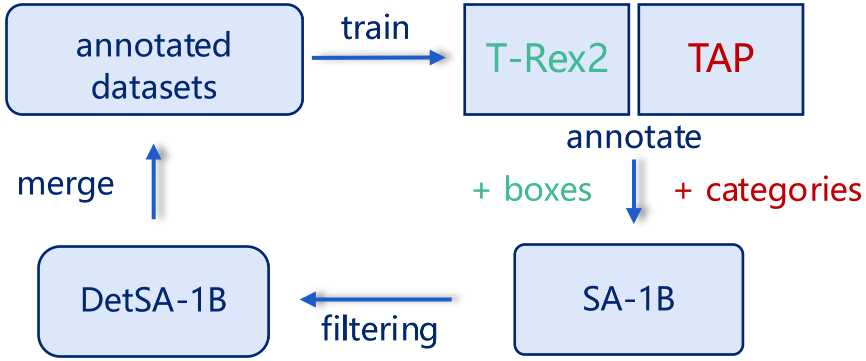
12 ways to be a genuinely nicer person

spaCy

Image classification data

Image caption data

## Visual Prompt Data Engine



annotated datasets

train

T-Rex2

TAP

annotate

+ boxes

+ categories

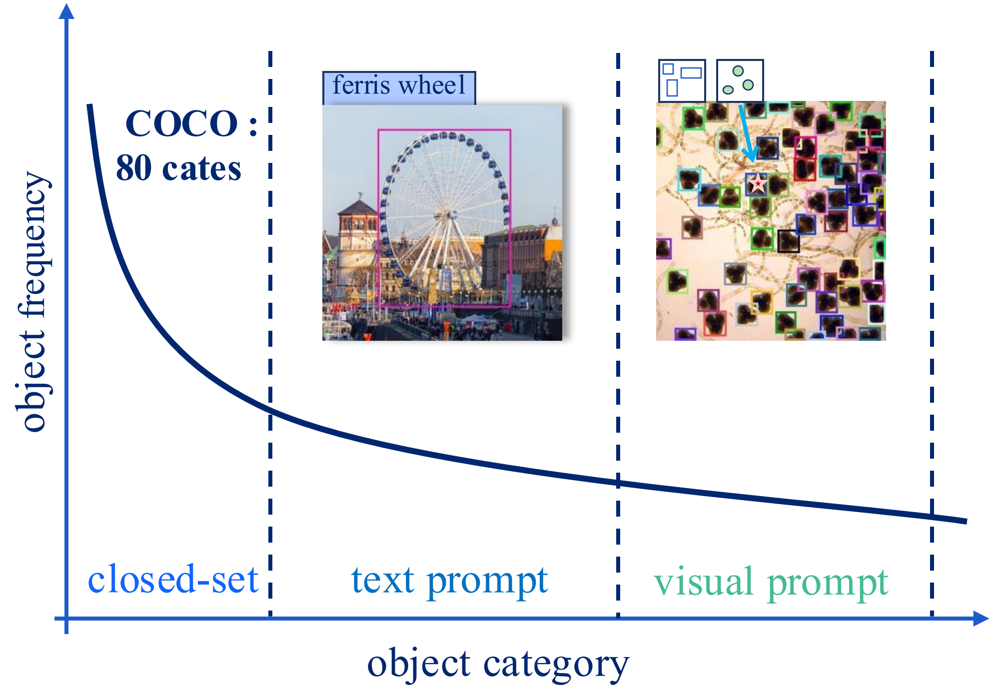
merge

DetSA-1B

filtering

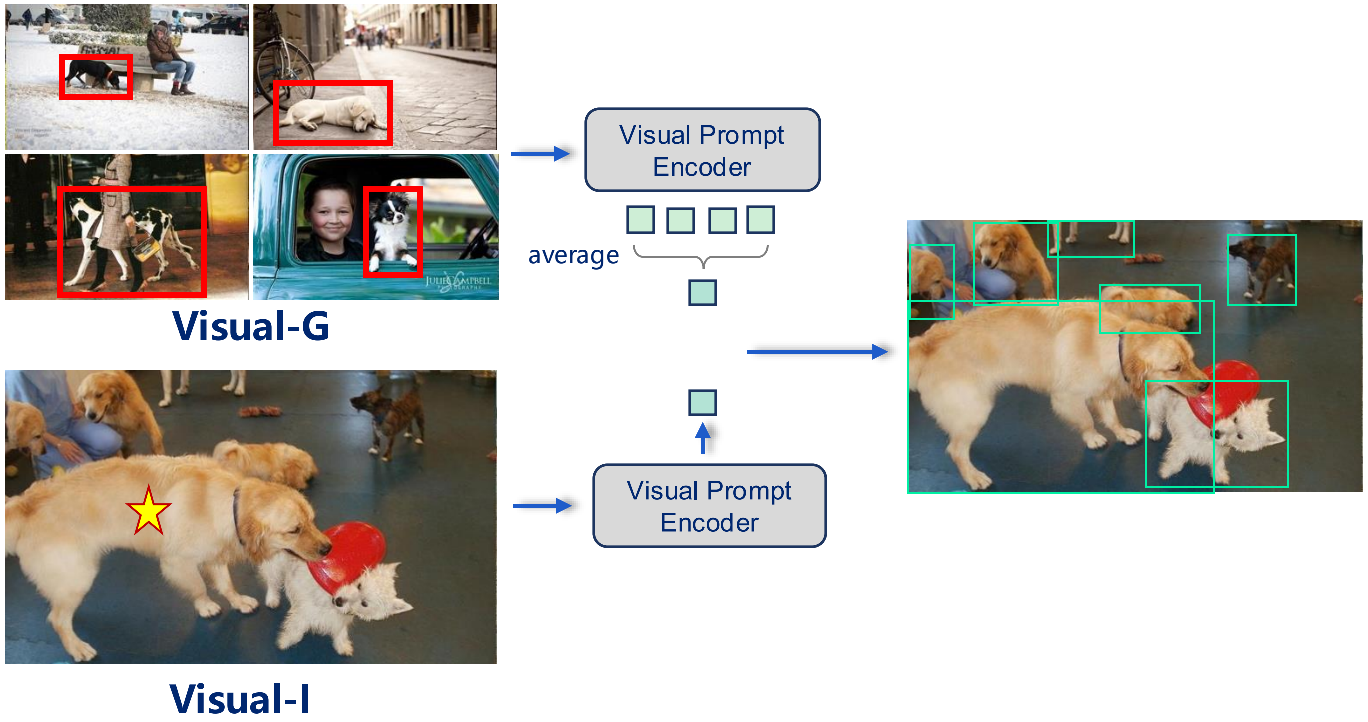
SA-1B

# T-Rex2: Experiments



- **Q1**: Whether the category coverage of Text Prompt and Visual Prompt follows the distribution in the figure?
- **Q2**: Whether Text Prompt and Visual Prompt can benefit with each other?

# T-Rex2: Metrics and Benchmarks



## Text Prompt Metric (AP)

- We use all the category names of the benchmark as text prompt inputs.

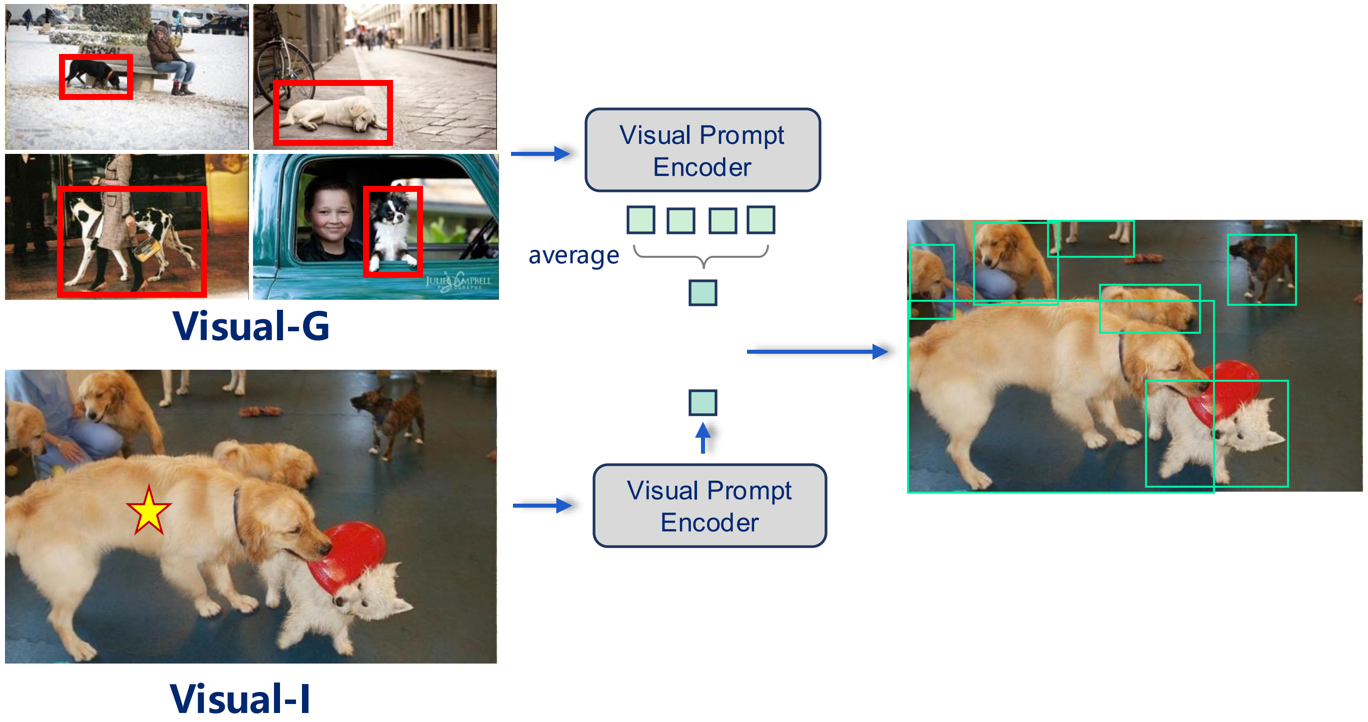
## Visual Prompt Metric (AP)

- **Visual-G**: Generic visual prompt
- **Visual-I**: Interactive visual prompt

Evaluation benchmarks (Zero-Shot):

- COCO (80 cates)
- LVIS (1203 cates): frequent: common: rare = 405:461:337 for val, 389:345:70 for minival
- ODinW (35 datasets)
- Roboflow100 (100 datasets)

# T-Rex2: Metrics and Benchmarks



## Visual Prompt Metric (AP)

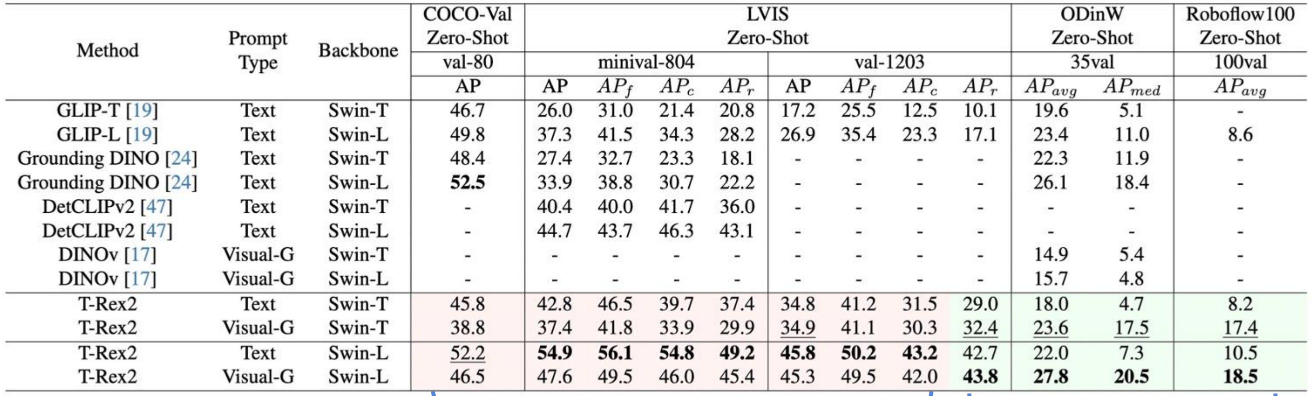
- **Visual-G**: Generic visual prompt
- **Visual-I**: Interactive visual prompt

Evaluation benchmarks (Zero-Shot):

- COCO (80 cates)
- LVIS (1203 cates): frequent: common: rare = 405:461:337 for val, 389:345:70 for minival
- ODinW (35 datasets)
- Roboflow100 (100 datasets)

# T-Rex2: **Answer for Q1** (category coverage)

## Zero-Shot Generic Object Detection

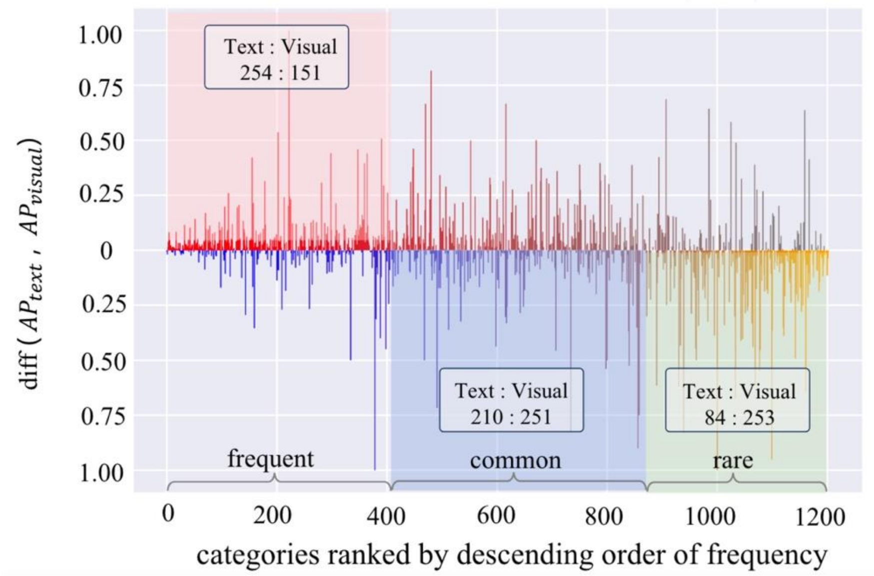


| Method | Prompt Type | Backbone | COCO-Val Zero-Shot | LVIS Zero-Shot | | | | | | | | ODinW Zero-Shot | | Roboflow100 Zero-Shot |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | val-80 | minival-804 | | | | val-1203 | | | | 35val | | 100val |
| | | | AP | AP | $AP_f$ | $AP_c$ | $AP_r$ | AP | $AP_f$ | $AP_c$ | $AP_r$ | $AP_{avg}$ | $AP_{med}$ | $AP_{avg}$ |
| GLIP-T [19] | Text | Swin-T | 46.7 | 26.0 | 31.0 | 21.4 | 20.8 | 17.2 | 25.5 | 12.5 | 10.1 | 19.6 | 5.1 | - |
| GLIP-L [19] | Text | Swin-L | 49.8 | 37.3 | 41.5 | 34.3 | 28.2 | 26.9 | 35.4 | 23.3 | 17.1 | 23.4 | 11.0 | 8.6 |
| Grounding DINO [24] | Text | Swin-T | 48.4 | 27.4 | 32.7 | 23.3 | 18.1 | - | - | - | - | 22.3 | 11.9 | - |
| Grounding DINO [24] | Text | Swin-L | **52.5** | 33.9 | 38.8 | 30.7 | 22.2 | - | - | - | - | 26.1 | 18.4 | - |
| DetCLIPv2 [47] | Text | Swin-T | - | 40.4 | 40.0 | 41.7 | 36.0 | - | - | - | - | - | - | - |
| DetCLIPv2 [47] | Text | Swin-L | - | 44.7 | 43.7 | 46.3 | 43.1 | - | - | - | - | - | - | - |
| DINOv [17] | Visual-G | Swin-T | - | - | - | - | - | - | - | - | - | 14.9 | 5.4 | - |
| DINOv [17] | Visual-G | Swin-L | - | - | - | - | - | - | - | - | - | 15.7 | 4.8 | - |
| T-Rex2 | Text | Swin-T | 45.8 | 42.8 | 46.5 | 39.7 | 37.4 | 34.8 | 41.2 | 31.5 | 29.0 | 18.0 | 4.7 | 8.2 |
| T-Rex2 | Visual-G | Swin-T | 38.8 | 37.4 | 41.8 | 33.9 | 29.9 | 34.9 | 41.1 | 30.3 | 32.4 | 23.6 | 17.5 | 17.4 |
| T-Rex2 | Text | Swin-L | 52.2 | **54.9** | **56.1** | **54.8** | **49.2** | **45.8** | **50.2** | **43.2** | 42.7 | 22.0 | 7.3 | 10.5 |
| T-Rex2 | Visual-G | Swin-L | 46.5 | 47.6 | 49.5 | 46.0 | 45.4 | 45.3 | 49.5 | 42.0 | **43.8** | **27.8** | **20.5** | **18.5** |

common and frequent case
**Text prompt better**

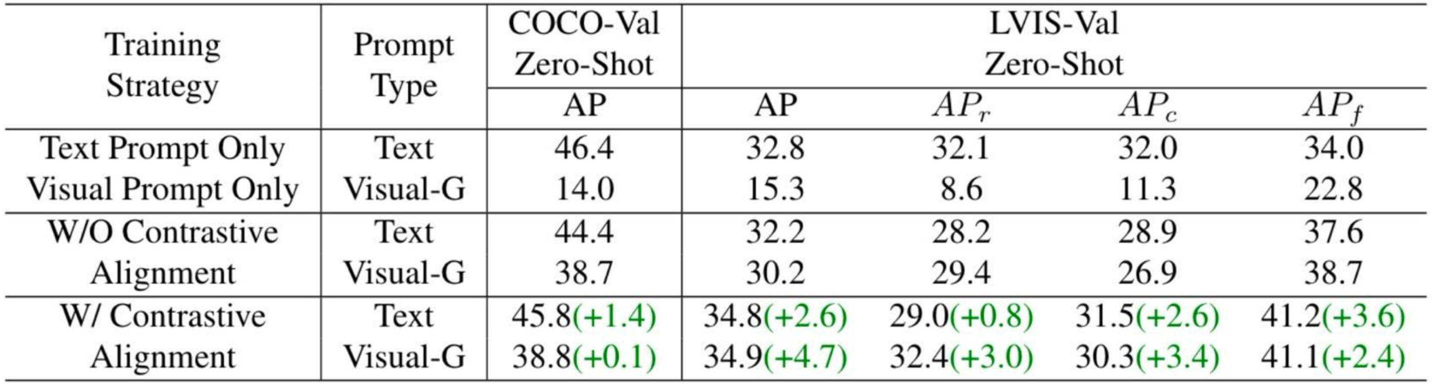
rare and novel case
**Visual prompt better**

## Text prompt v.s. Visual prompt on LVIS



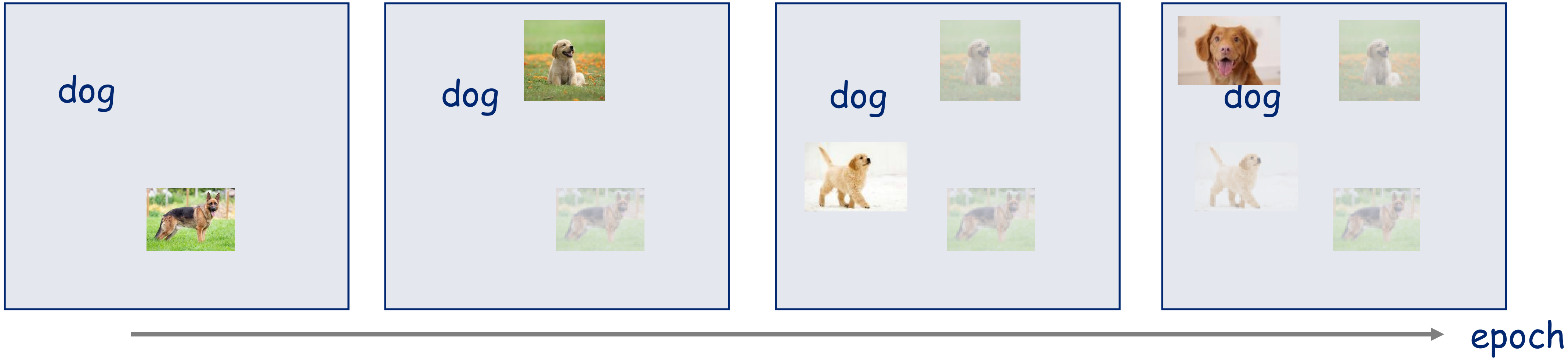
- Text prompt is good at common and frequent object, while visual prompt succeed in rare and novel scenarios.

# T-Rex2: **Answer for Q2** (benefits of synergy)



| Training Strategy | Prompt Type | COCO-Val Zero-Shot | LVIS-Val Zero-Shot | | | |
|---|---|---|---|---|---|---|
| | | AP | AP | $AP_r$ | $AP_c$ | $AP_f$ |
| Text Prompt Only | Text | 46.4 | 32.8 | 32.1 | 32.0 | 34.0 |
| Visual Prompt Only | Visual-G | 14.0 | 15.3 | 8.6 | 11.3 | 22.8 |
| W/O Contrastive Alignment | Text | 44.4 | 32.2 | 28.2 | 28.9 | 37.6 |
| | Visual-G | 38.7 | 30.2 | 29.4 | 26.9 | 38.7 |
| W/ Contrastive Alignment | Text | 45.8(+1.4) | 34.8(+2.6) | 29.0(+0.8) | 31.5(+2.6) | 41.2(+3.6) |
| | Visual-G | 38.8(+0.1) | 34.9(+4.7) | 32.4(+3.0) | 30.3(+3.4) | 41.1(+2.4) |

- W/O training with text prompt, visual prompts only have limited generic detection capability.

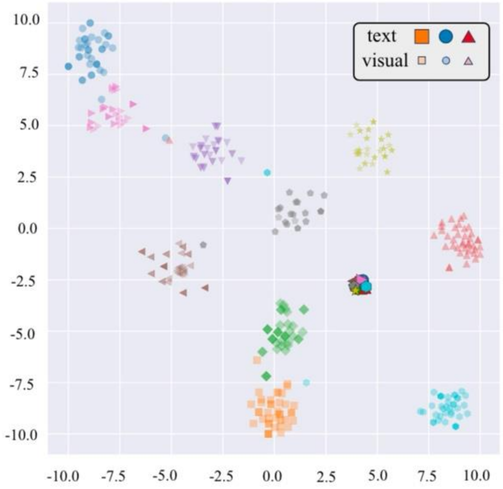


During the training process, the visual samples vary significantly from iter to iter, making it difficult to learn a common representation, whereas text prompt only needs to optimize the same embedding.

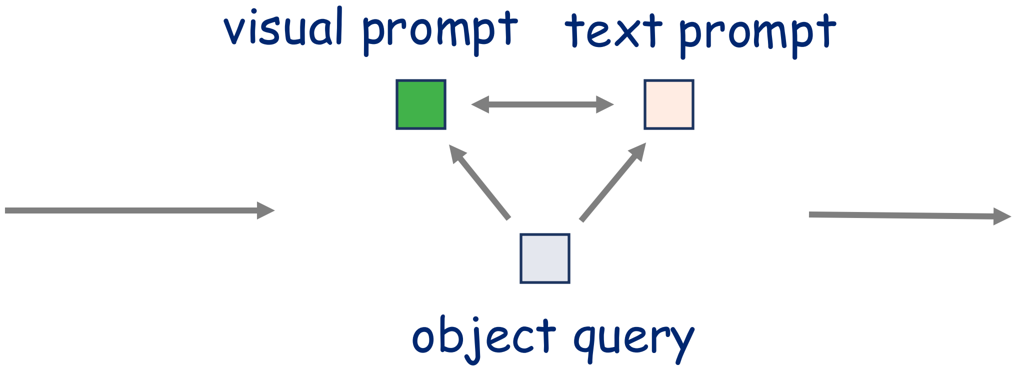
# T-Rex2: **Answer for Q2** (benefits of synergy)

| Training Strategy | Prompt Type | COCO-Val Zero-Shot | LVIS-Val Zero-Shot | | | |
|---|---|---|---|---|---|---|
| | | AP | AP | $AP_r$ | $AP_c$ | $AP_f$ |
| Text Prompt Only | Text | 46.4 | 32.8 | 32.1 | 32.0 | 34.0 |
| Visual Prompt Only | Visual-G | 14.0 | 15.3 | 8.6 | 11.3 | 22.8 |
| W/O Contrastive Alignment | Text | 44.4 | 32.2 | 28.2 | 28.9 | 37.6 |
| | Visual-G | 38.7 | 30.2 | 29.4 | 26.9 | 38.7 |
| W/ Contrastive Alignment | Text | 45.8(+1.4) | 34.8(+2.6) | 29.0(+0.8) | 31.5(+2.6) | 41.2(+3.6) |
| | Visual-G | 38.8(+0.1) | 34.9(+4.7) | 32.4(+3.0) | 30.3(+3.4) | 41.1(+2.4) |

- W/O training with text prompt, visual prompts only have limited generic detection capability.
- Naïve joint training without explicit alignment can improve visual prompt but harm text prompt.



(a) w/o contrastive align



visual prompt

text prompt

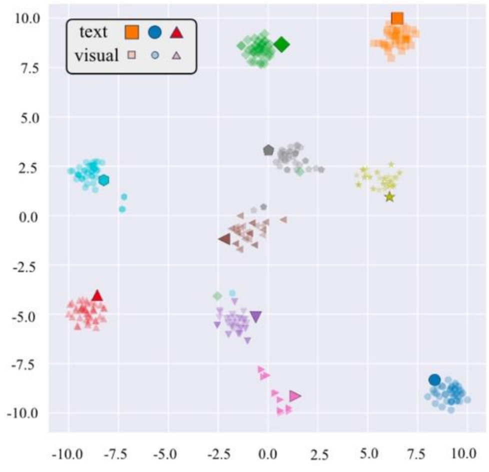
object query

object query needs to bridge the distance between itself and both the visual prompt and the text prompt. But the visual prompt is not aligned with the text prompt, which makes optimization difficult.

# T-Rex2: **Answer for Q2** (benefits of synergy)

| Training Strategy | Prompt Type | COCO-Val Zero-Shot | LVIS-Val Zero-Shot | | | |
|---|---|---|---|---|---|---|
| | | AP | AP | $AP_r$ | $AP_c$ | $AP_f$ |
| Text Prompt Only | Text | 46.4 | 32.8 | 32.1 | 32.0 | 34.0 |
| Visual Prompt Only | Visual-G | 14.0 | 15.3 | 8.6 | 11.3 | 22.8 |
| W/O Contrastive Alignment | Text | 44.4 | 32.2 | 28.2 | 28.9 | 37.6 |
| | Visual-G | 38.7 | 30.2 | 29.4 | 26.9 | 38.7 |
| W/ Contrastive Alignment | Text | 45.8(+1.4) | 34.8(+2.6) | 29.0(+0.8) | 31.5(+2.6) | 41.2(+3.6) |
| | Visual-G | 38.8(+0.1) | 34.9(+4.7) | 32.4(+3.0) | 30.3(+3.4) | 41.1(+2.4) |

- W/O training with text prompt, visual prompts only have limited generic detection capability.
- Naïve joint training without explicit alignment can improve visual prompt but harm text prompt.
- With the proposed contrastive alignment, both prompt modalities can gain improvement.

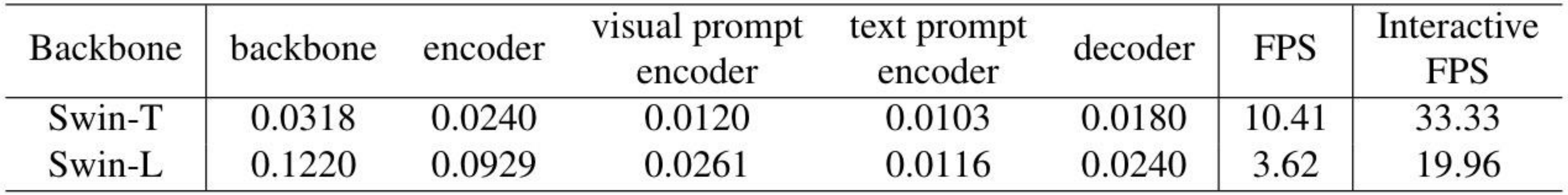


(b) w/ contrastive align

- Text prompt serve as anchor, which aggregates visual prompt
- Visual prompts act as a continuous source of refinement for text prompts.

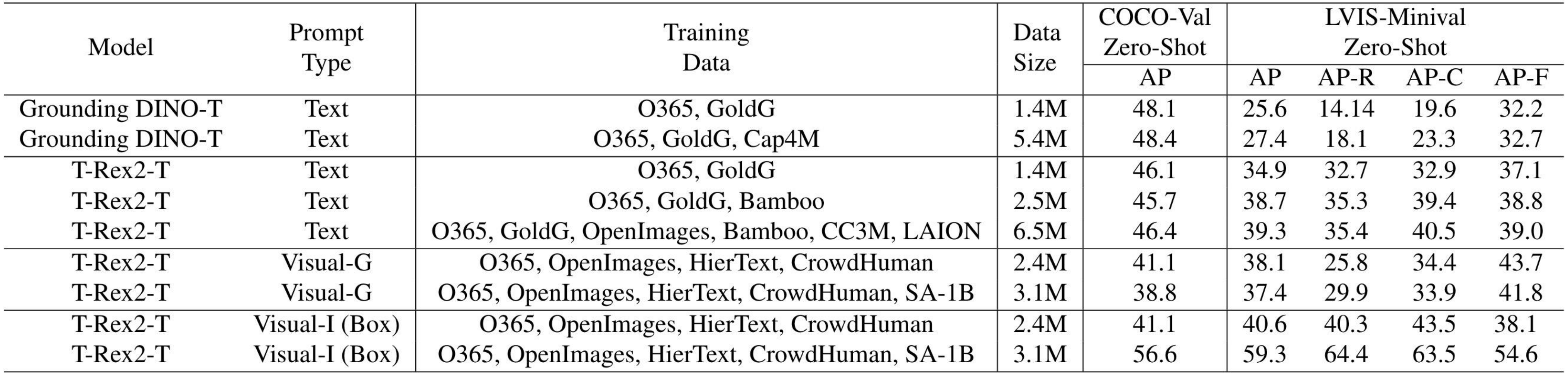
# T-Rex2: **More Analysis**

## Inference Speed



| Backbone | backbone | encoder | visual prompt encoder | text prompt encoder | decoder | FPS | Interactive FPS |
|---|---|---|---|---|---|---|---|
| Swin-T | 0.0318 | 0.0240 | 0.0120 | 0.0103 | 0.0180 | 10.41 | 33.33 |
| Swin-L | 0.1220 | 0.0929 | 0.0261 | 0.0116 | 0.0240 | 3.62 | 19.96 |

## Ablation of Data Engines

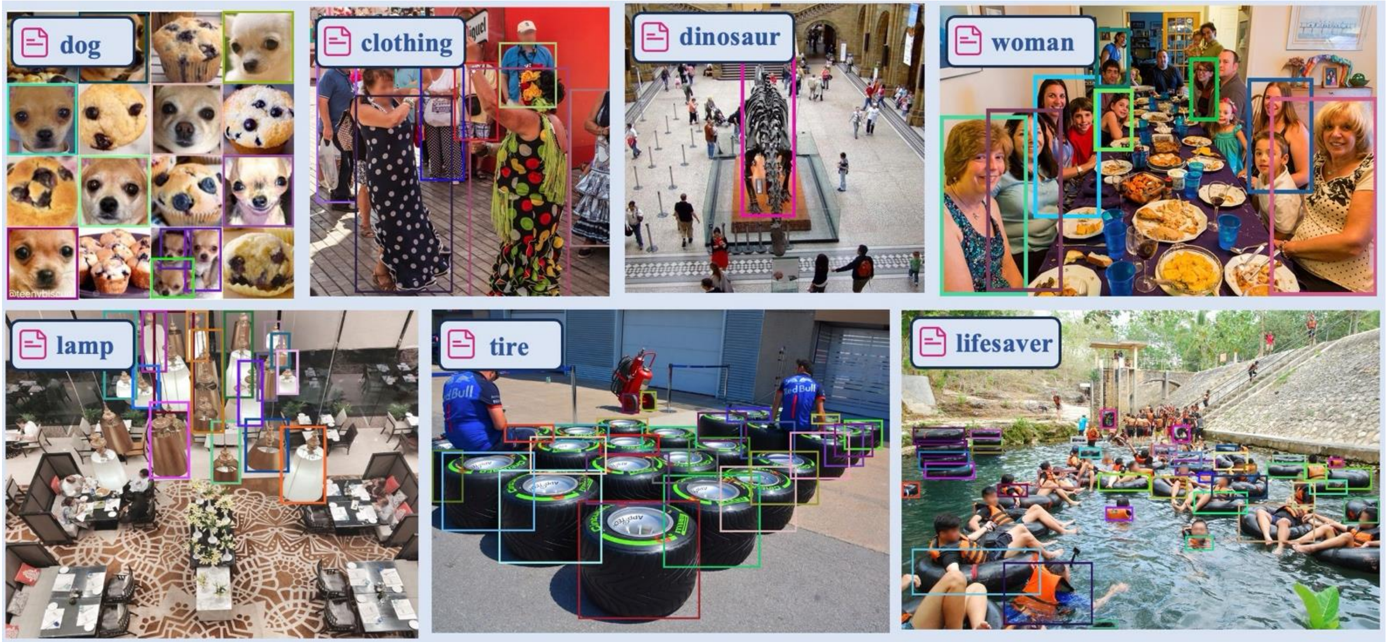


| Model | Prompt Type | Training Data | Data Size | COCO-Val Zero-Shot | LVIS-Minival Zero-Shot | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | AP | AP | AP-R | AP-C | AP-F |
| Grounding DINO-T | Text | O365, GoldG | 1.4M | 48.1 | 25.6 | 14.14 | 19.6 | 32.2 |
| Grounding DINO-T | Text | O365, GoldG, Cap4M | 5.4M | 48.4 | 27.4 | 18.1 | 23.3 | 32.7 |
| T-Rex2-T | Text | O365, GoldG | 1.4M | 46.1 | 34.9 | 32.7 | 32.9 | 37.1 |
| T-Rex2-T | Text | O365, GoldG, Bamboo | 2.5M | 45.7 | 38.7 | 35.3 | 39.4 | 38.8 |
| T-Rex2-T | Text | O365, GoldG, OpenImages, Bamboo, CC3M, LAION | 6.5M | 46.4 | 39.3 | 35.4 | 40.5 | 39.0 |
| T-Rex2-T | Visual-G | O365, OpenImages, HierText, CrowdHuman | 2.4M | 41.1 | 38.1 | 25.8 | 34.4 | 43.7 |
| T-Rex2-T | Visual-G | O365, OpenImages, HierText, CrowdHuman, SA-1B | 3.1M | 38.8 | 37.4 | 29.9 | 33.9 | 41.8 |
| T-Rex2-T | Visual-I (Box) | O365, OpenImages, HierText, CrowdHuman | 2.4M | 41.1 | 40.6 | 40.3 | 43.5 | 38.1 |
| T-Rex2-T | Visual-I (Box) | O365, OpenImages, HierText, CrowdHuman, SA-1B | 3.1M | 56.6 | 59.3 | 64.4 | 63.5 | 54.6 |

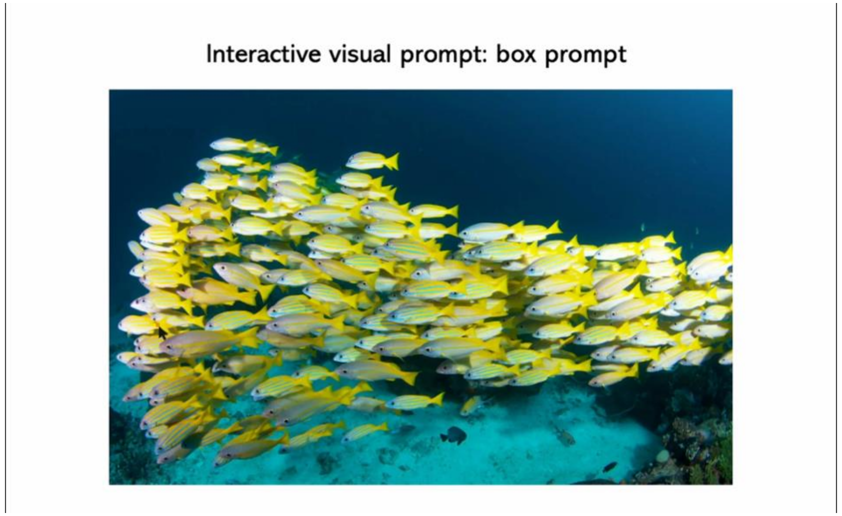
# T-Rex2: **Applications 1 Interactive Object Detection**

**Scenarios**: Counting, Annotation

# T-Rex2: **Applications 2 Open Vocabulary Object Detection**



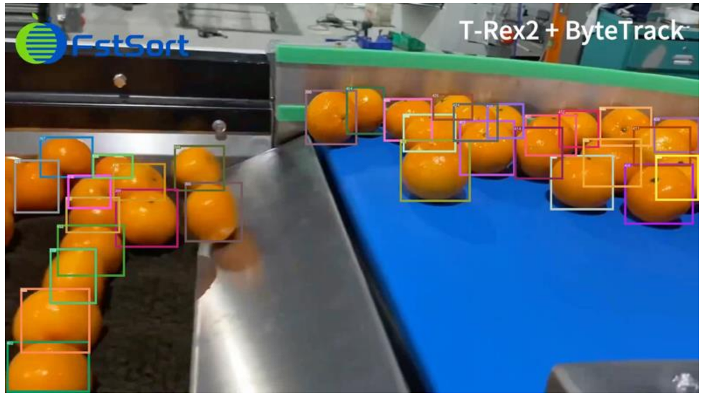
# T-Rex2: **Applications 3 Generic Object Detection**



**static image**

**video object detection**

**video + tracking**



# Conclusion

- **Visual prompt and text prompt together can lead to generic object detection**

**Paper:** https://arxiv.org/pdf/2403.14610.pdf

**Homepage:** https://deepdataspace.com/home

**Demo:** https://deepdataspace.com/playground/ivp

**Github:** https://github.com/IDEA-Research/T-Rex

**HuggingFace:** https://huggingface.co/spaces/Mountchicken/T-Rex2