# Recent Advances in Visual Grounding

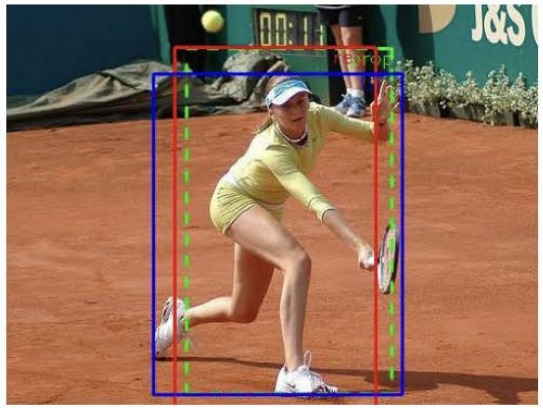
**报告人：叶加博（故研）**

# Outline

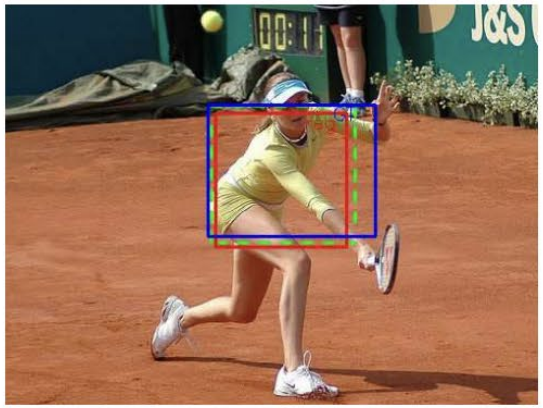
- Background
- Fully Supervised Methods
- Weakly Supervised Methods
- Semi-weakly Supervised Method
- Novel Datasets and Tasks
- Further Work

# Phrase Grounding

A young tennis player wearing a yellow shirt and shorts hits the tennis ball



Query 1: A young tennis player

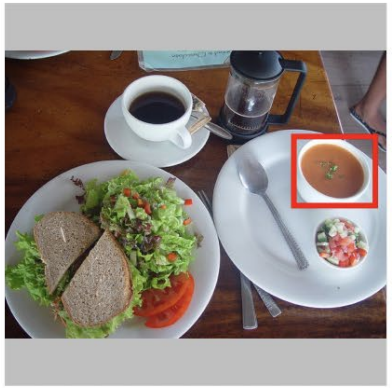


Query 2 : yellow shirt

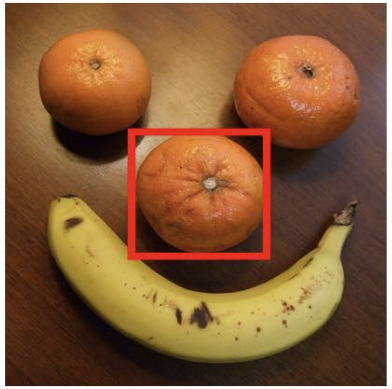
Query 3 : tennis ball

# Referring Expression Grounding

**(a) This is the giraffe on the right who is looking towards the camera**

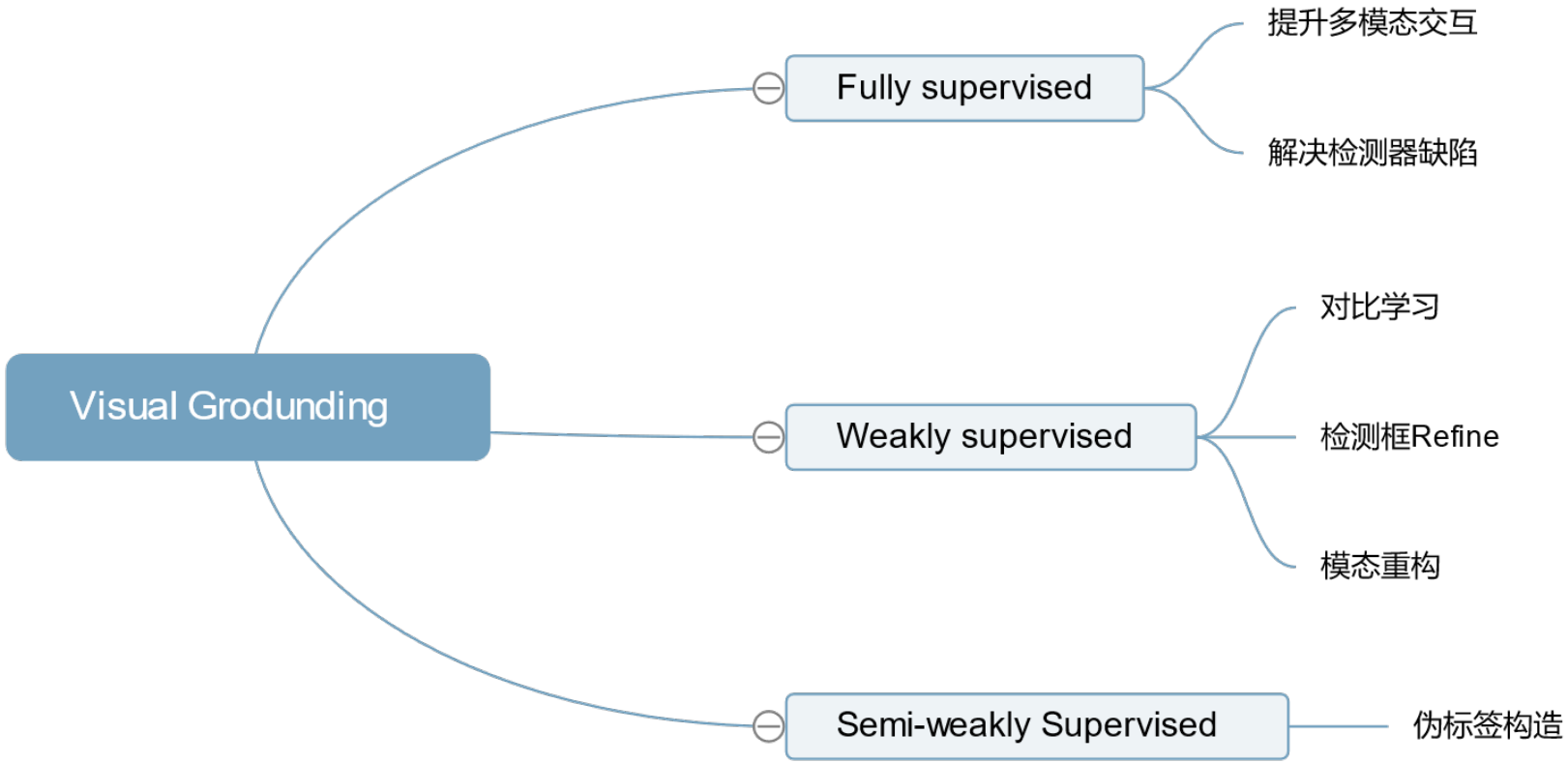


**(b) Bowl of tomato soup**



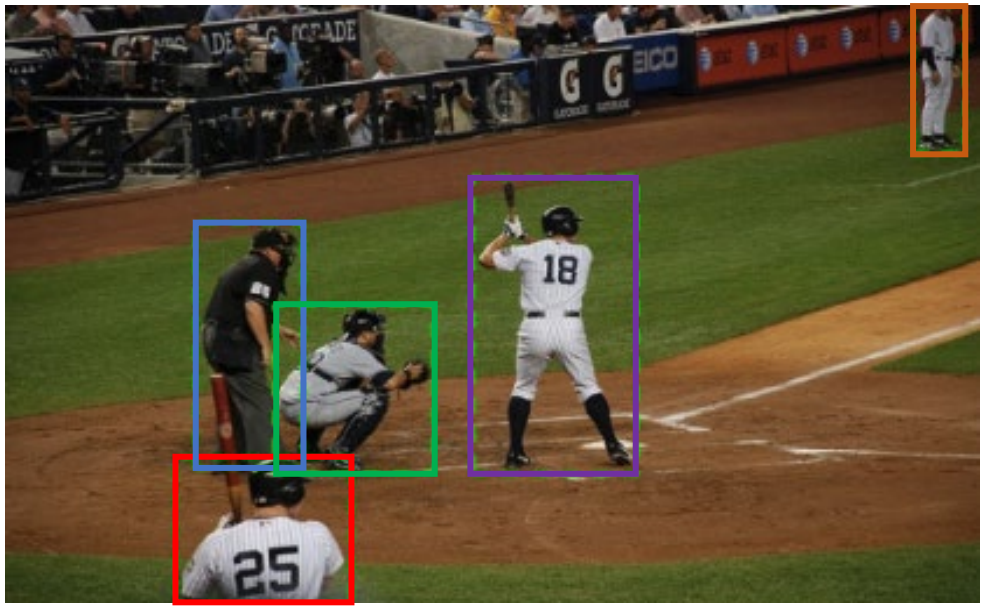
**(c) Orange between other oranges and a banana**

**(d) A man is eating sandwich by sitting along with other members**


Visual Grodunding
Fully supervised
提升多模态交互
解决检测器缺陷
Weakly supervised
对比学习
检测框Refine
模态重构
Semi-weakly Supervised
伪标签构造

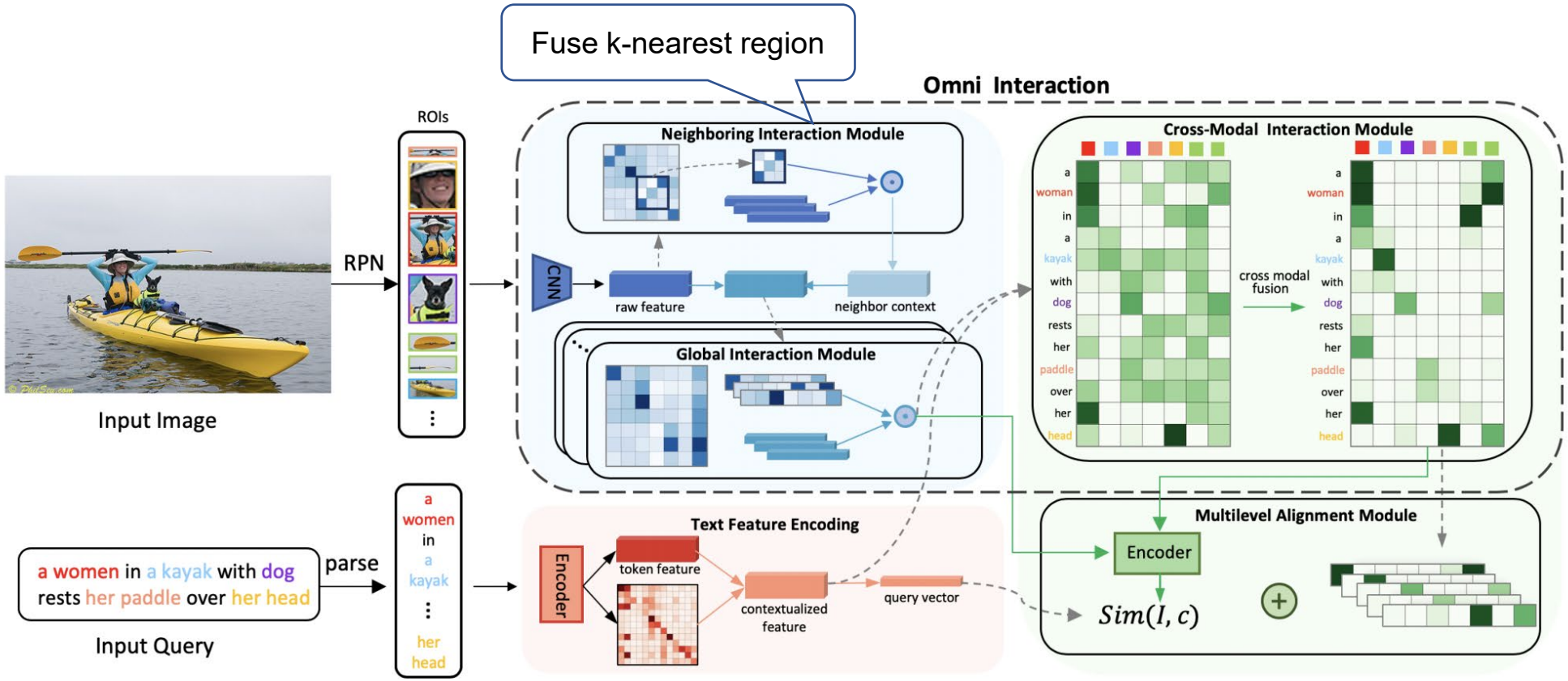
# Fully Supervised Methods

Fine-grained Alignment




man in white on the left holding a bat.
man in black on the left.
man in the middle.
man in the middle holding a bat.
man in the top right corner.

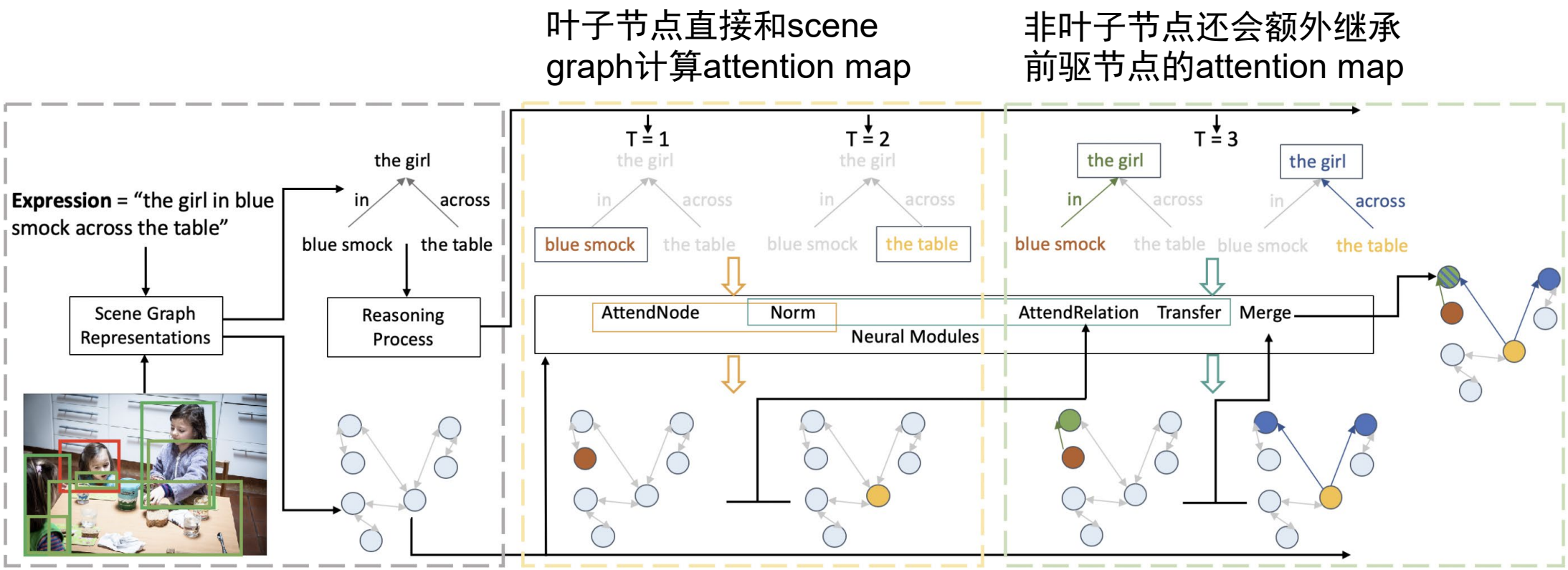
# Improve Multi-modal Interaction



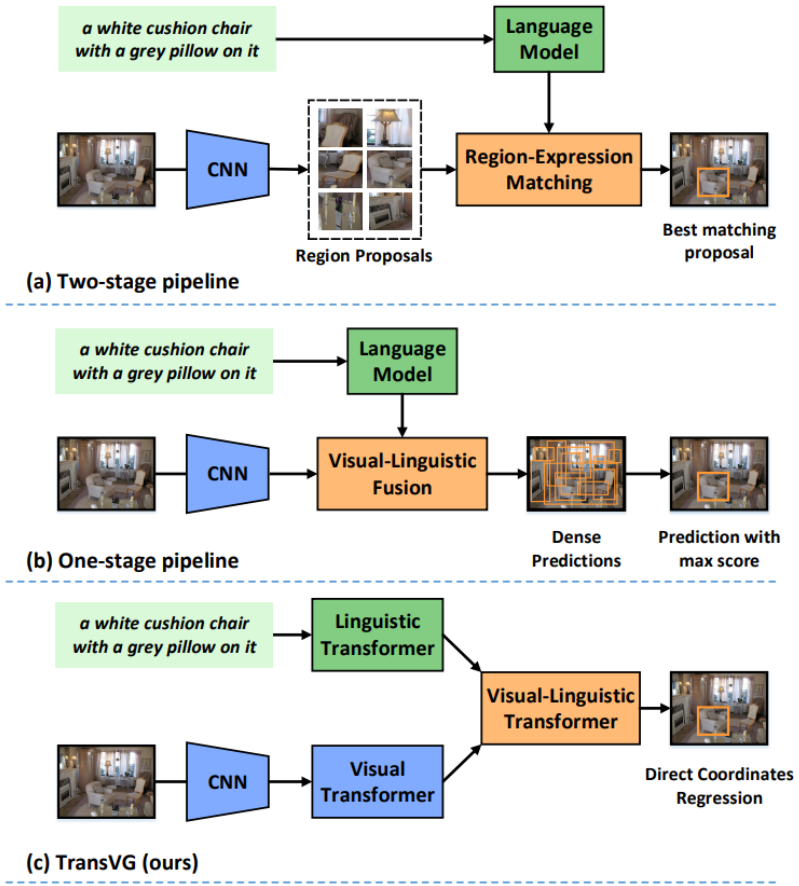


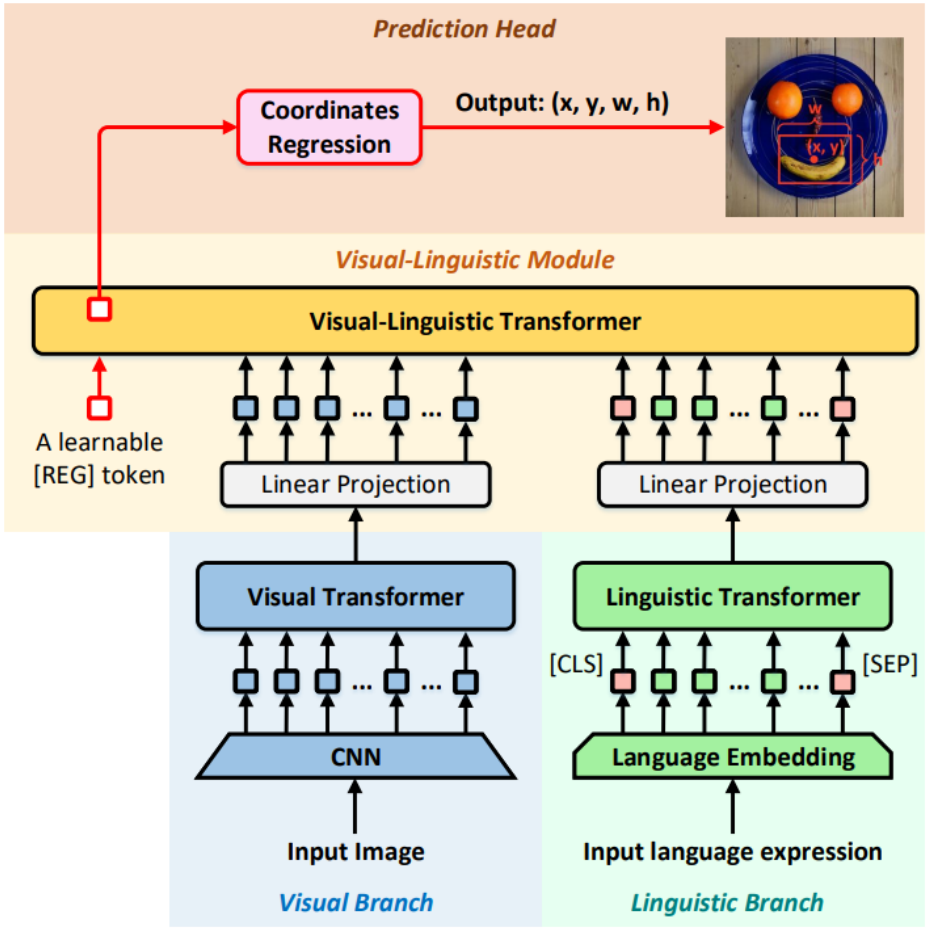

*Multimedia 2020, Cross-Modal Omni Interaction Modeling for Phrase Grounding*

# Improve Multi-modal Interaction



# Improve Multi-modal Interaction



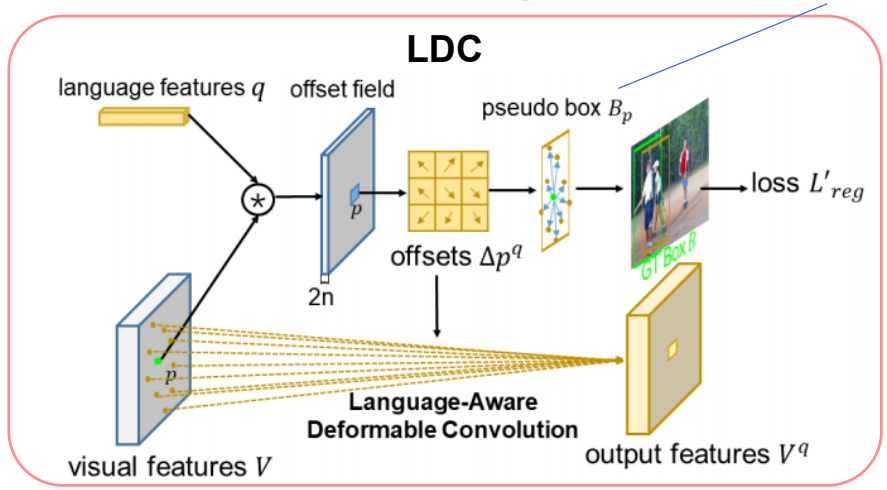



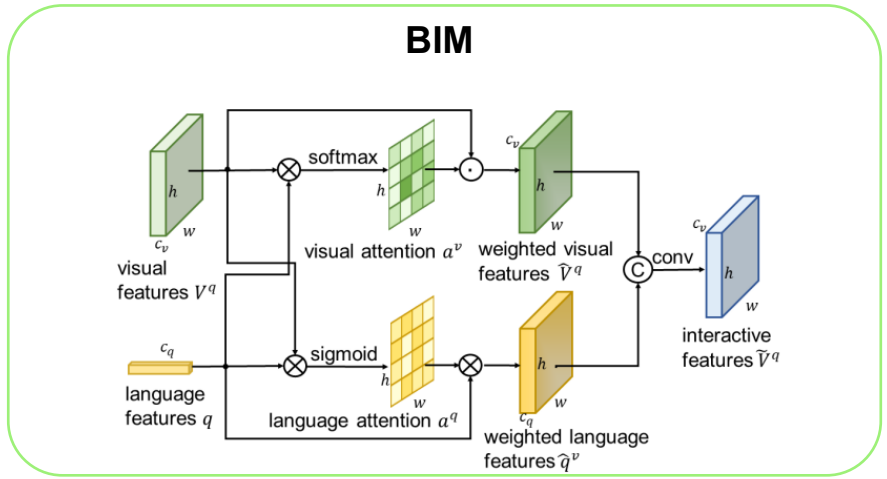


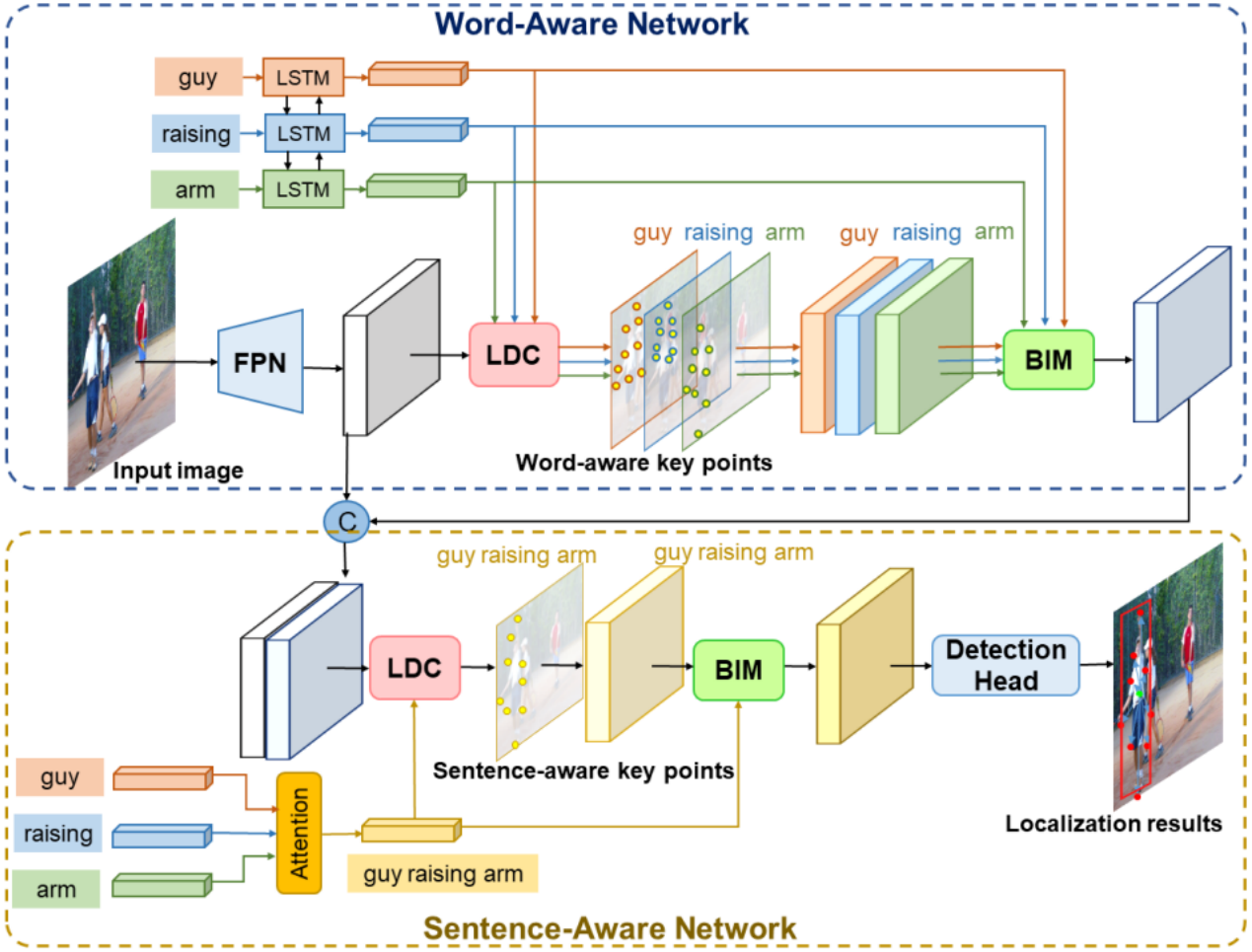

*arxiv 2021, TransVG: End-to-End Visual Grounding with Transformers*

# Improve Detection Structure

$$B_p = \{mean(x_n), mean(y_n), \lambda_x std(x_n), \lambda_y std(y_n)\}$$



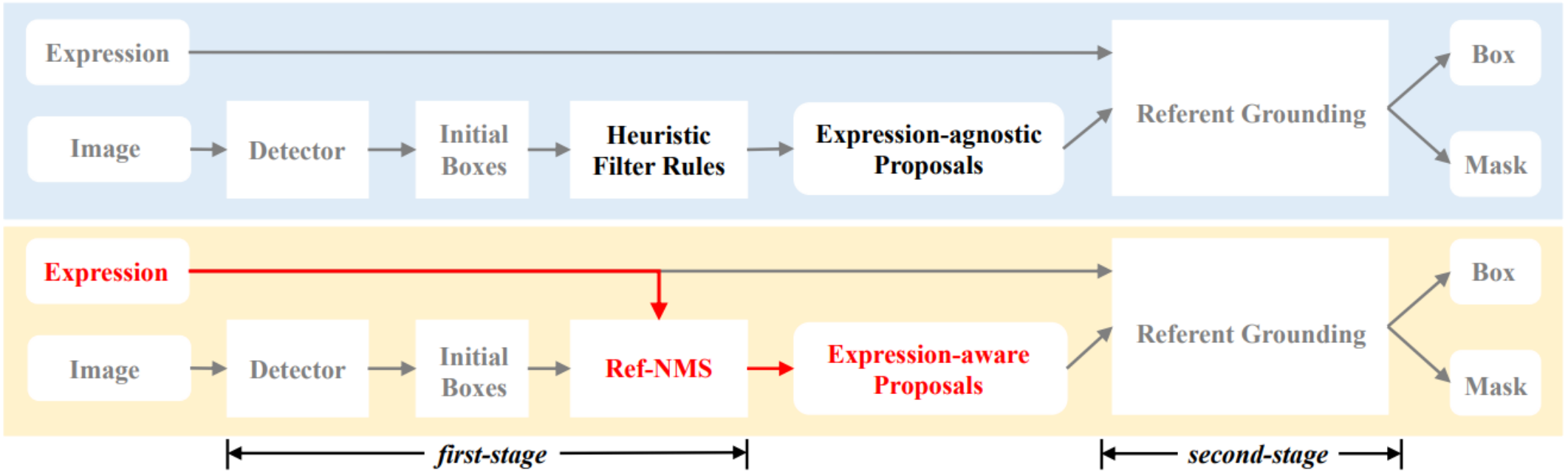






*Multimedia 2020, Language-Aware Fine-Grained Object Representation for Referring Expression Comprehension*

# Improve Detection Structure






*AAAI 2021, Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding*

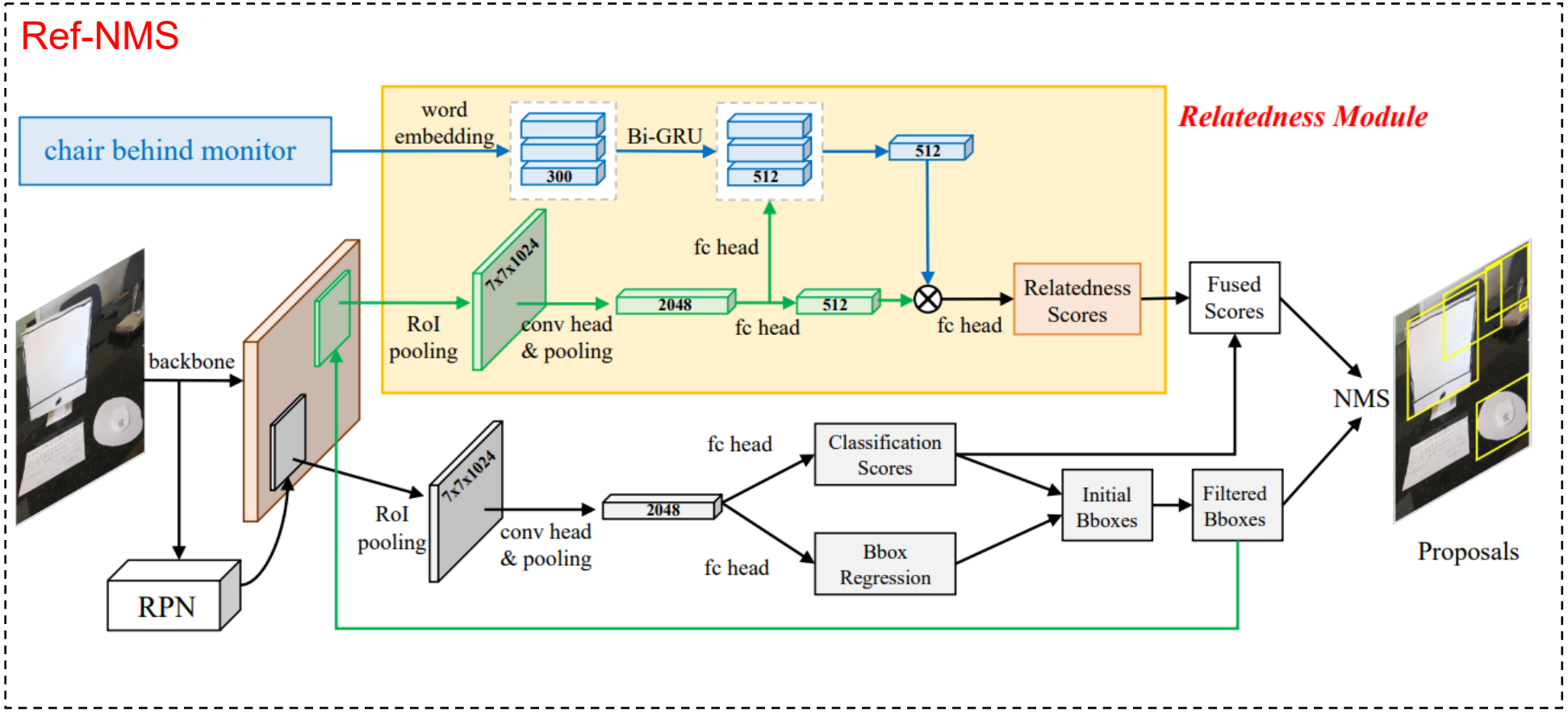
# Improve Detection Structure

## Traditional NMS

同分类的IoU大于阈值的boxes保留classification score分数最高的box。

## Ref-NMS

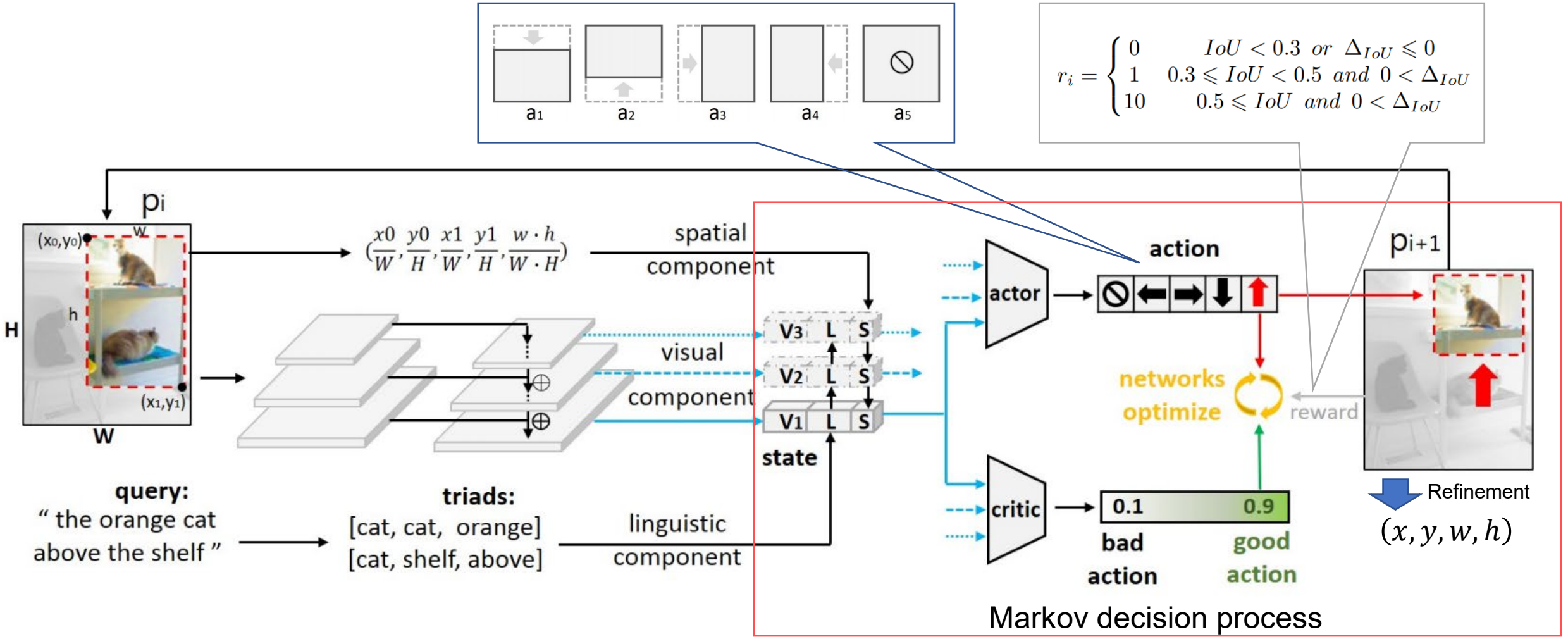
计算relatedness score和classification score共同指导NMS过程






*AAAI 2021, Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding*

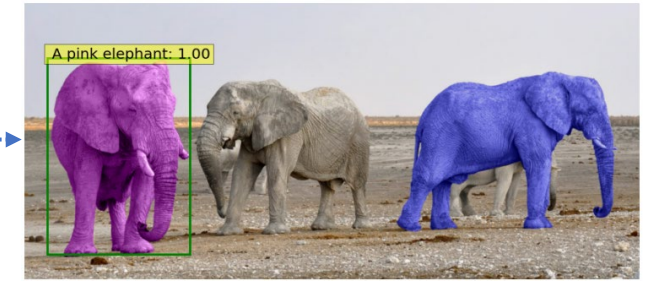
# Improve Detection Structure



# Improve Detection Structure

- Detector is text-agnostic
- Detector is trained on a fixed vocabulary of objects and attributes

Unable to detect unseen object




Object Detector

ROIs

A women in a kayak with dog rests her paddle over her head

Text Encoder

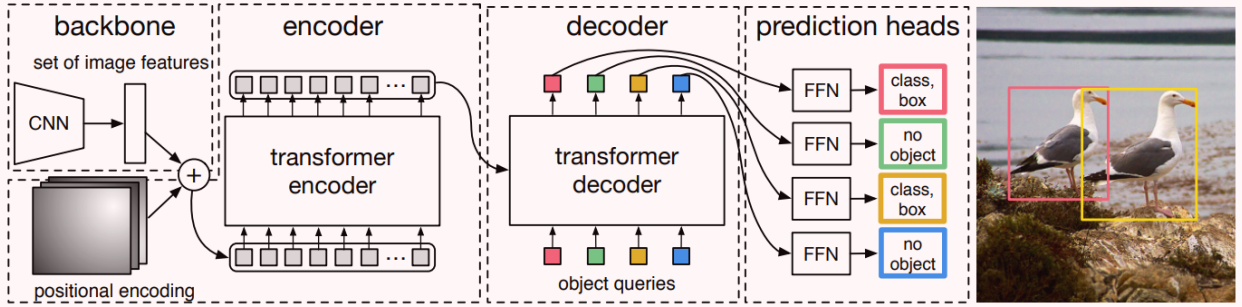
Co-TRM → TRM → $(h_{v0}, h_{v1}, \cdots, h_{vT})$

Co-TRM → TRM → $(h_{w0}, h_{w1}, \cdots, h_{wT})$

$\mathbf{k}\times$

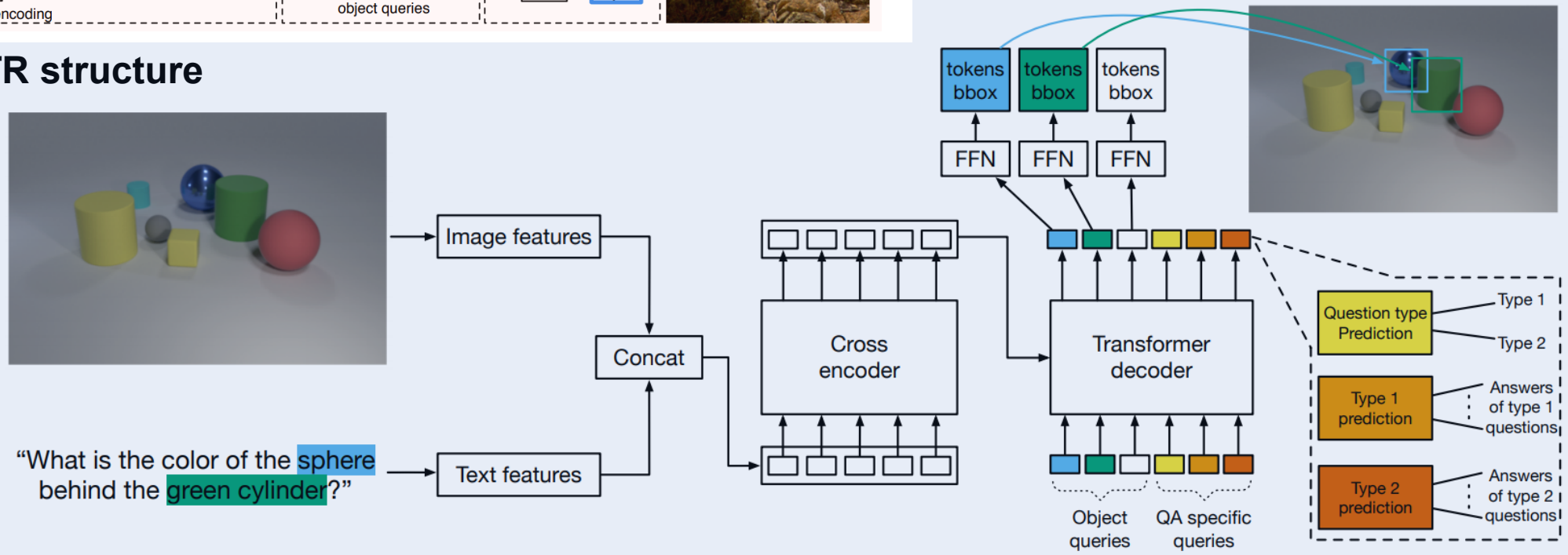
# Improve Detection Structure

**DETR structure**




- Bounding box regression
- Phrase span prediction
- Contrastive learning

**MDETR structure**






*CVPR 2021, MDETR - Modulated Detection for End-to-End Multi-Modal Understanding*

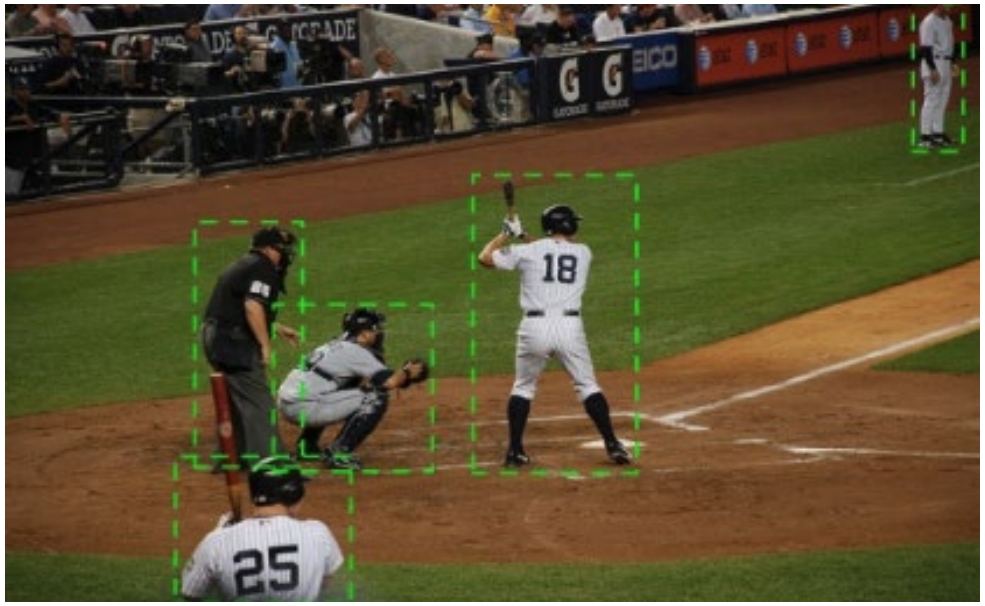
# Fully Supervised Performance

Accuracy@0.5 comparison

| | **RefCOCO** | | **RefCOCO+** | | **RefCOCOg** | **Flickr30k** | **ReferIt** |
|---|---|---|---|---|---|---|---|
| | testA | testB | testA | testB | test | test | test |
| COI Net | - | - | - | - | - | 77.51 | 66.16 |
| SGMN | 86.67 | **85.36** | 78.66 | 69.77 | 81.42 | - | - |
| DIGN | - | - | - | - | - | 78.73 | 65.15 |
| TransVG | 82.72 | 78.35 | 70.70 | 56.94 | 67.73 | **79.10** | **70.73** |
| HFRN | 83.12 | 75.51 | 72.53 | 57.09 | 69.08 | - | - |
| Ref-NMS (Mattnet) | 82.71 | 73.94 | 71.29 | 58.40 | 68.67 | - | - |
| IS | 74.27 | 68.10 | 71.05 | 58.25 | 70.05 | - | - |
| MDETR | **90.40** | 82.67 | **85.52** | **72.96** | **83.31** | - | - |

# Weakly Supervised Methods

Coarse-grained Alignment

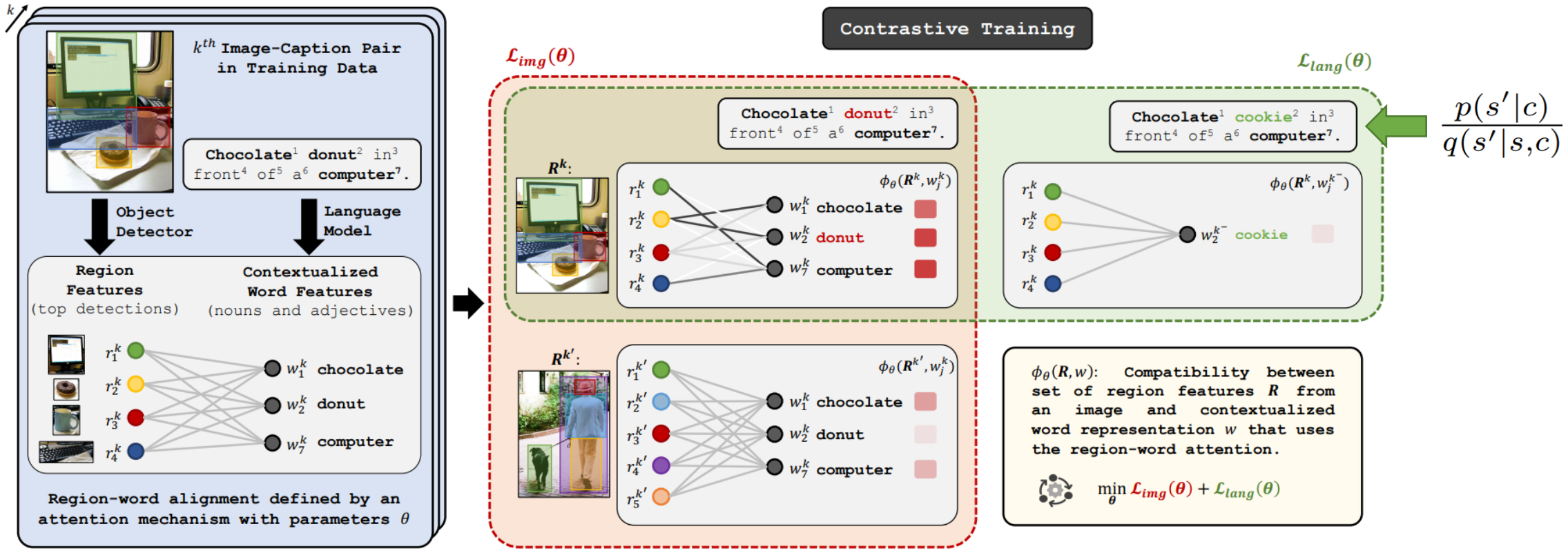



man in white on the left holding a bat.
man in black on the left.
man in the middle.
man in the middle holding a bat.
man in the top right corner.

# Contrastive Learning

$w_l'$: constructed negative phrase representations

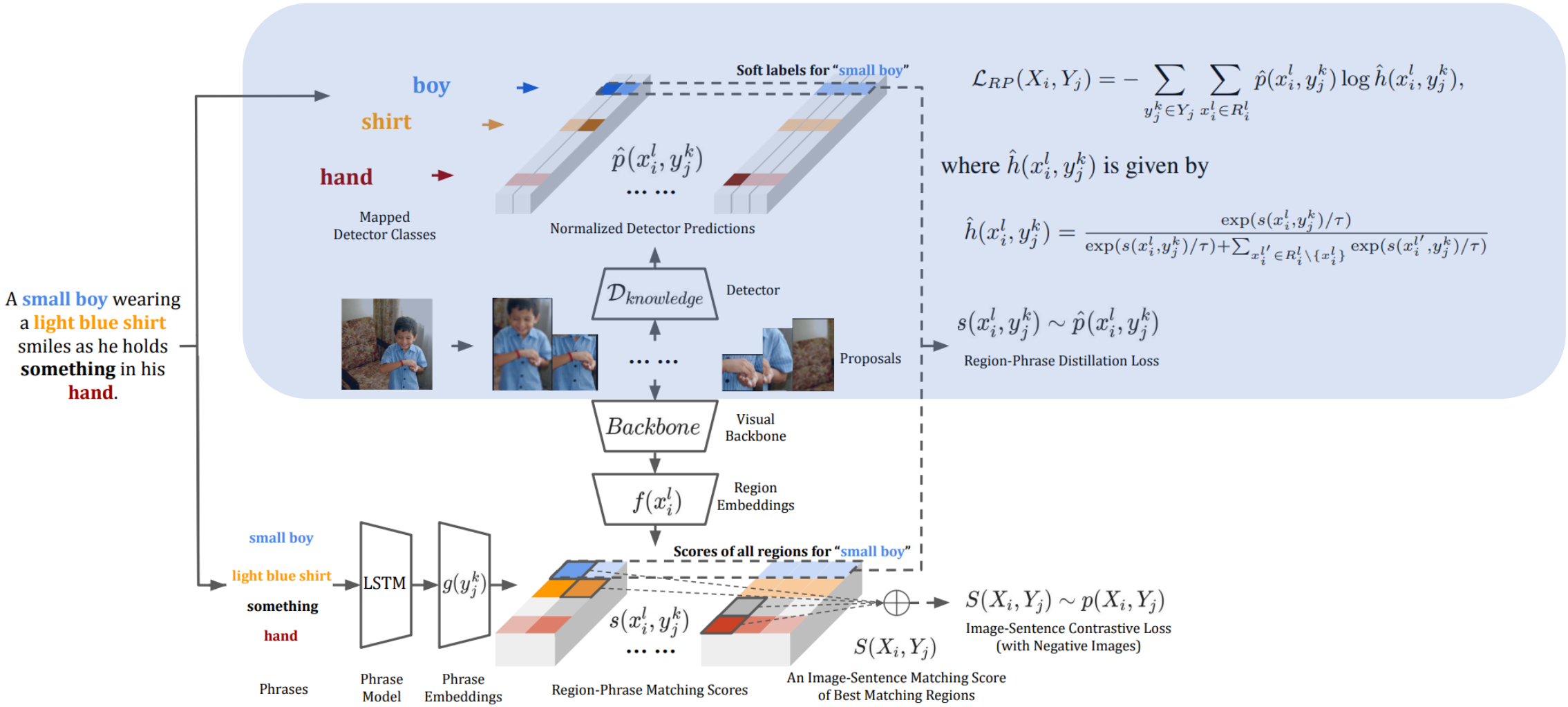
$$\mathcal{L}_{\text{lang}}(\theta) = \mathbb{E}_{\mathcal{B}}\left[-\log\left(\frac{e^{\phi_\theta(\mathbf{R},w)}}{e^{\phi_\theta(\mathbf{R},w)} + \sum_{l=1}^{25} e^{\phi_\theta(\mathbf{R},w_l')}}\right)\right]$$




$$\mathcal{L}_{\text{img}}(\theta) = \mathbb{E}_{\mathcal{B}}\left[-\sum_{j=1}^{n}\log\left(\frac{e^{\phi_\theta(\mathbf{R},w_j)}}{e^{\phi_\theta(\mathbf{R},w_j)} + \sum_{i=1}^{k-1} e^{\phi_\theta(\mathbf{R}_i',w_j)}}\right)\right]$$

$R_i'$: region set in other images of the mini-batch

# Contrastive Learning





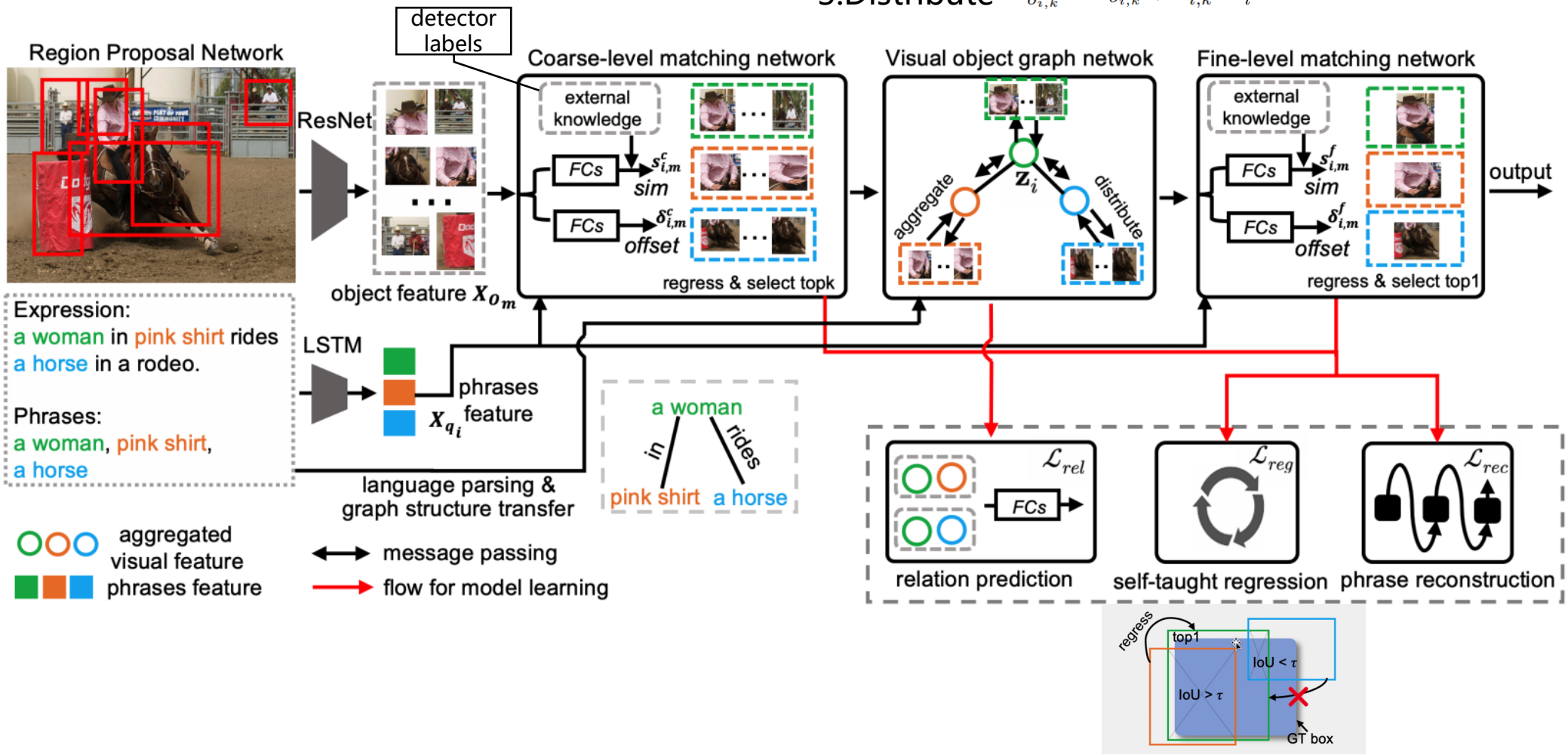

*CVPR 2021, Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation*

# Coarse to Fine

1.Aggregate $\mathbf{z}_i = \sum_{k=1}^{K} \alpha_{i,k}^{c} \cdot \mathbf{x}_{o_{i,k}}$

2.Message passing

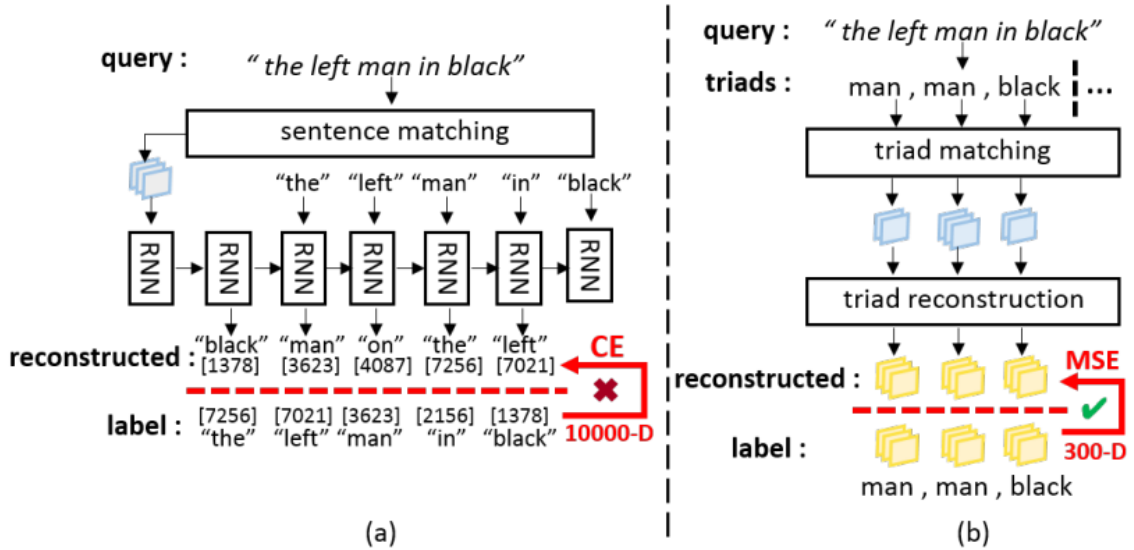
3.Distribute $\mathbf{x}'_{o_{i,k}} = \mathbf{x}_{o_{i,k}} + \alpha_{i,k}^{c} \cdot \mathbf{z}'_i$



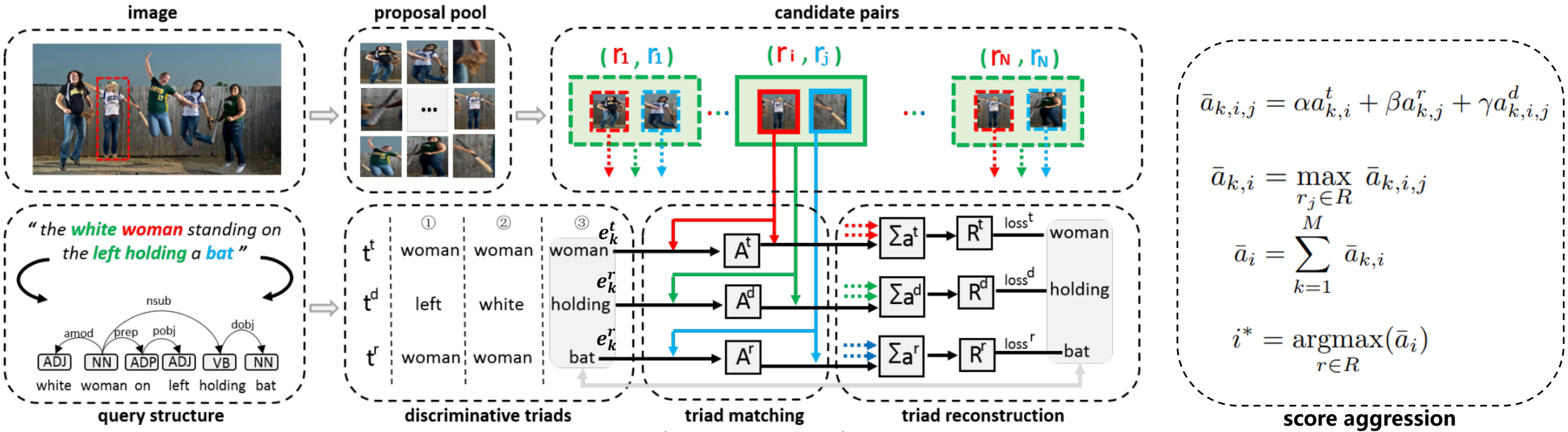
# Triad-Level Reconstruction

**Weakness of traditional reconstruction**

- Too large vocabulary
- Lack of discriminative



# Triad-Level Reconstruction




$$\bar{a}_{k,i,j} = \alpha a^t_{k,i} + \beta a^r_{k,j} + \gamma a^d_{k,i,j}$$

$$\bar{a}_{k,i} = \max_{r_j \in R} \bar{a}_{k,i,j}$$

$$\bar{a}_i = \sum_{k=1}^{M} \bar{a}_{k,i}$$

$$i^* = \operatorname{argmax}_{r \in R}(\bar{a}_i)$$

$$a^t_{k,i} = match(f^v_i, e^t_k)$$

$$a^d_{k,i,j} = match([f^v_i, f^v_j], e^r_k)$$

$$a^r_{k,j} = match(f^v_j, e^r_k)$$



*TPAMI 2021, Discriminative Triad Matching and Reconstruction for Weakly Referring Expression Grounding*

# Weakly Supervised Performance

Recall@K comparison of phrase grounding

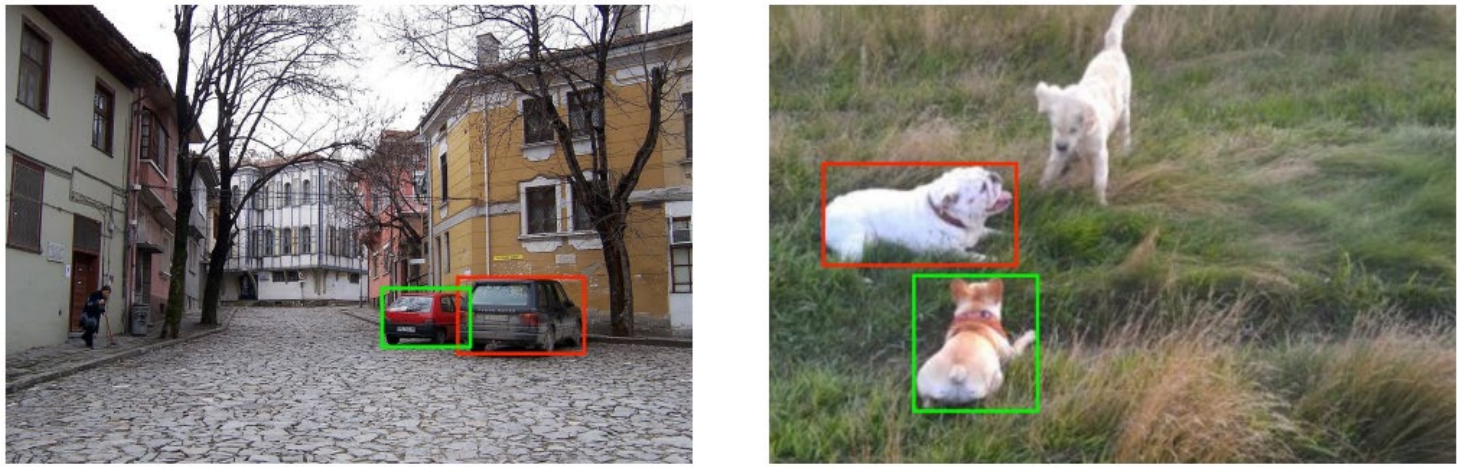
| | **Flickr30k** | | | **ReferIt** |
|---|---|---|---|---|
| | R@1/Acc | R@5 | R@10 | R@1/Acc |
| info-ground | 51.67 | **77.69** | **83.25** | - |
| Distill | 50.96 | - | - | 27.59 |
| ReIR | **59.27** | - | - | **37.68** |

- *Recall@1 is equivalent to the accuracy*

Acc@0.5 comparison of referring expression grounding

| | **RefCOCO** | | **RefCOCO+** | | **RefCOCOg** |
|---|---|---|---|---|---|
| | testA | testB | testA | testB | val |
| TransVG(fully baseline) | 82.72 | 78.35 | 70.70 | 56.94 | 67.02 |
| KPRN(weakly baseline) | 34.74 | 36.53 | 32.75 | 36.76 | 35.44 |
| DTWREG | **41.14** | **37.72** | **40.01** | **38.08** | **43.24** |

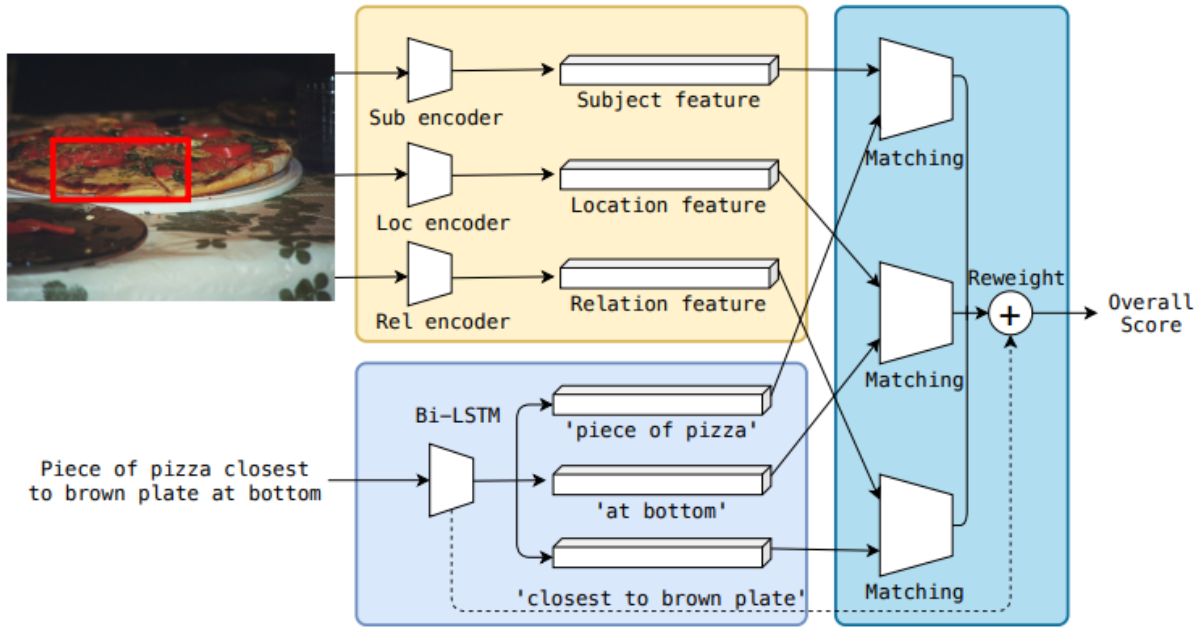
# Semi-Supervised Visual Grounding

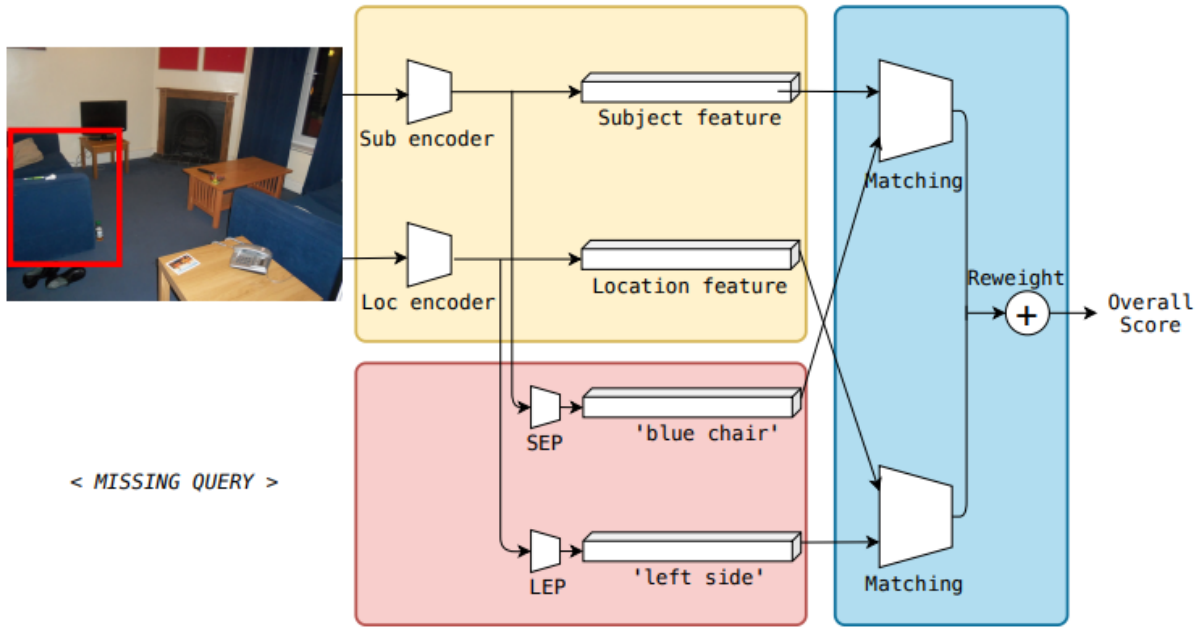


(a) (b)

Figure 1. Images in the training set where only some objects (those shown in red boxes) are labeled. Red boxes in the two images are annotated as '*black car*' and '*white dog lying on the grass on the left*' respectively, while objects in green boxes only have bounding boxes and category names, '*car*' and '*dog*', associated with them.

# Semi-Supervised Visual Grounding

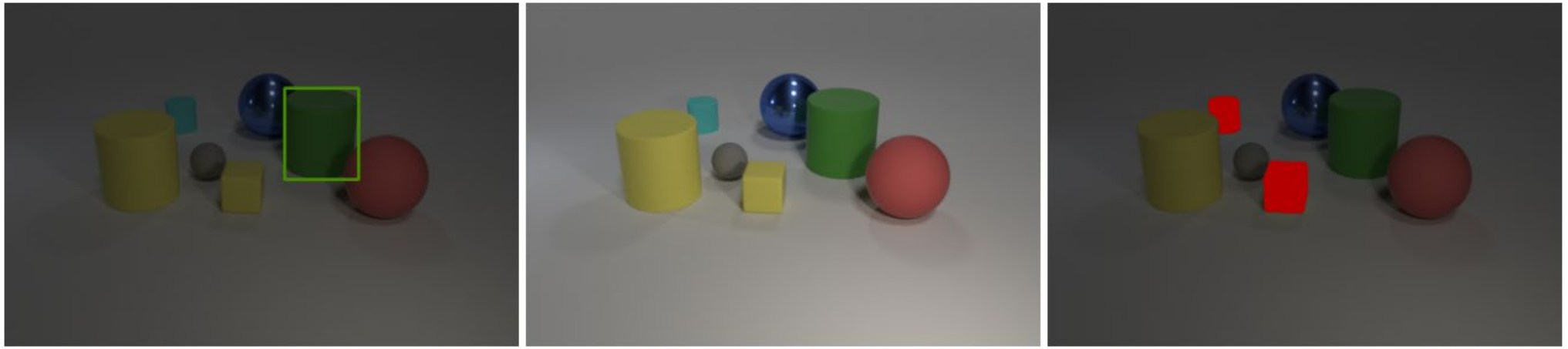


# Novel Datasets and Tasks

- CLEVR-Ref+
- Ref-Reasoning
- Cops-Ref
- KB-Ref

# CLEVR-Ref+



*The big thing(s) that are behind the second one of the big thing(s) from front and to the right of the first one of the large sphere(s) from left*

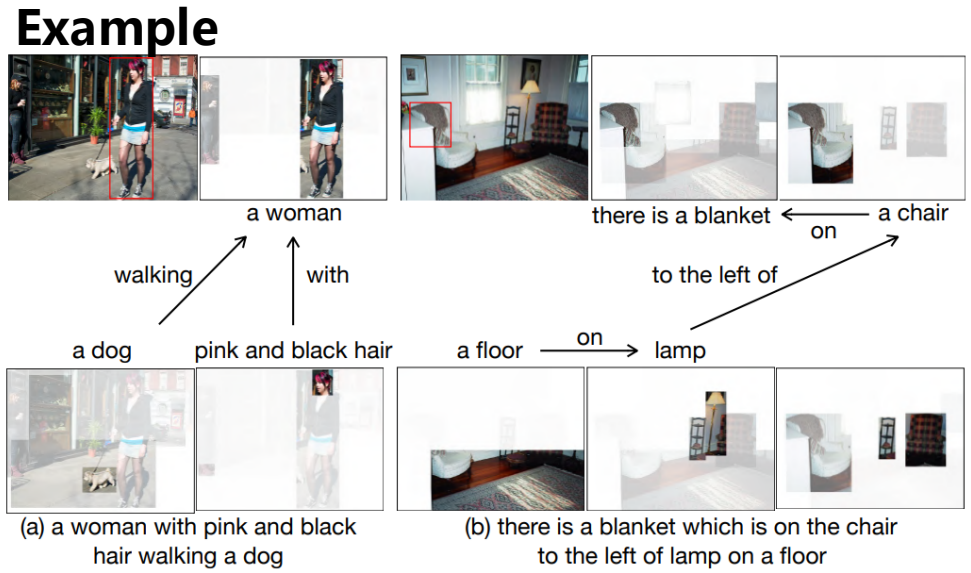
*Any other things that are the same size as the fifth one of the thing(s) from right*

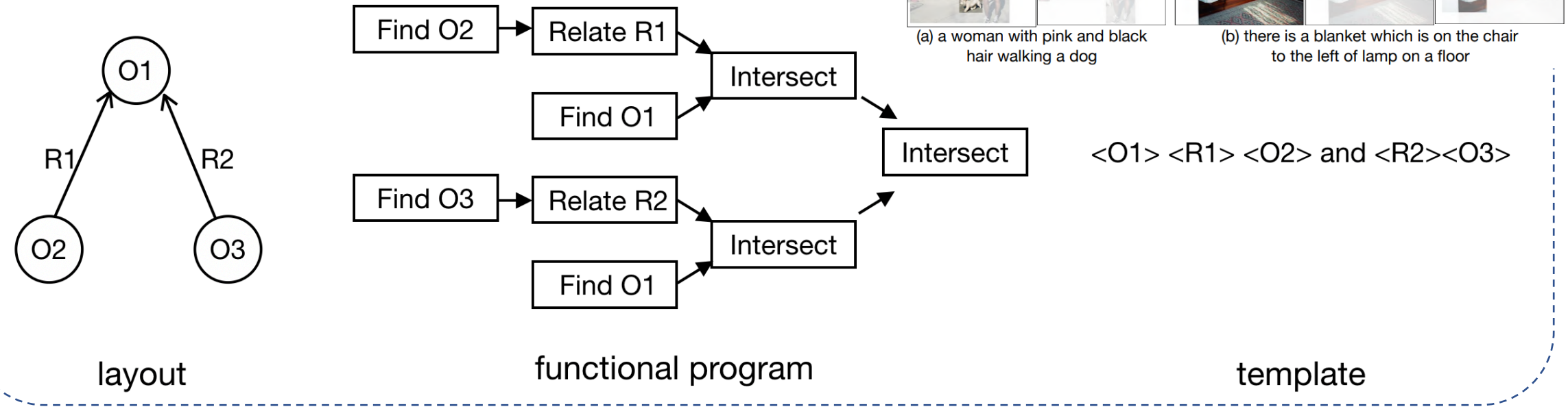
- 85,000 images (from CLEVR)
- 850,000 referring expressions



*CVPR 2019, CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions*

# Ref-Reasoning



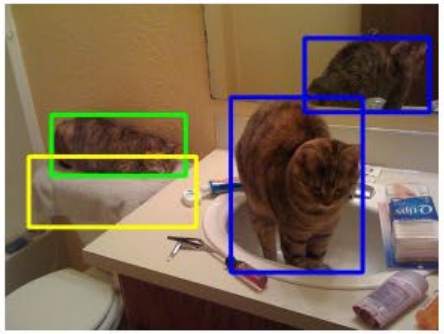




- 83,989 images (from GQA)
- 791,956 referring expressions

# Cops-Ref

**Reasoning tree**: cat (left, sleeping) $\xrightarrow{\text{resting}}$ towel (white)

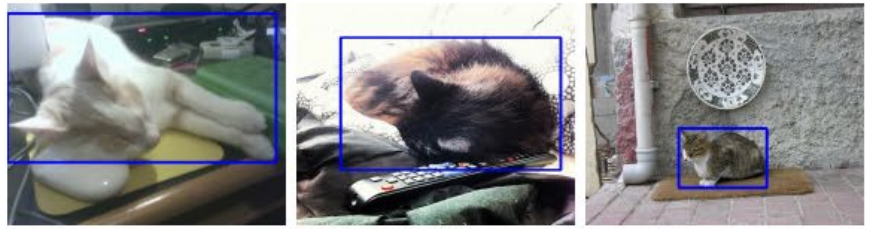
**Expression**: *The cat on the left that is sleeping and resting on the white towel.*



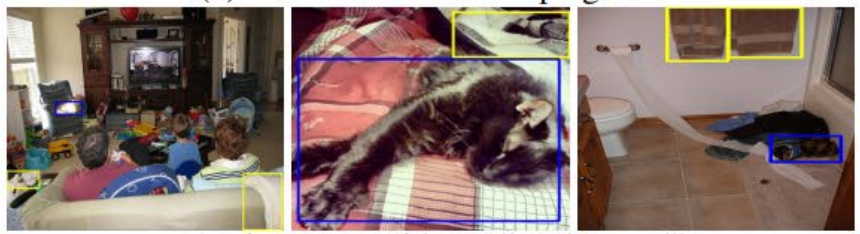
(a) The image with the target "cat"

(b) Distractors of different categories

(c) Distractors with "cat"



(d) Distractors with "sleeping cat"



(e) Distractors with "cat" and "towel"

- 75,299 images (from GQA)
- 148,712 referring expressions

**Task:**

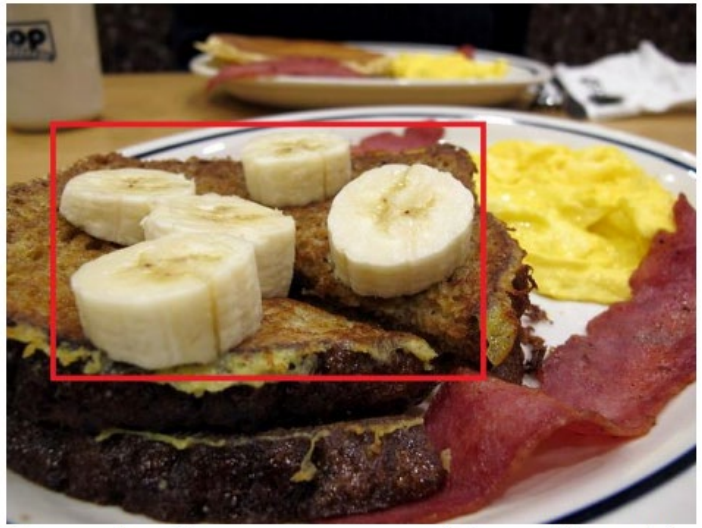
$$r_{i^*,j^*} = \underset{r_{i,j}, i\in[1,N], j\in[1,J_i]}{\arg\max} s(r_{i,j}|q)$$

$i$-th image and $j$-th region from $I_i$

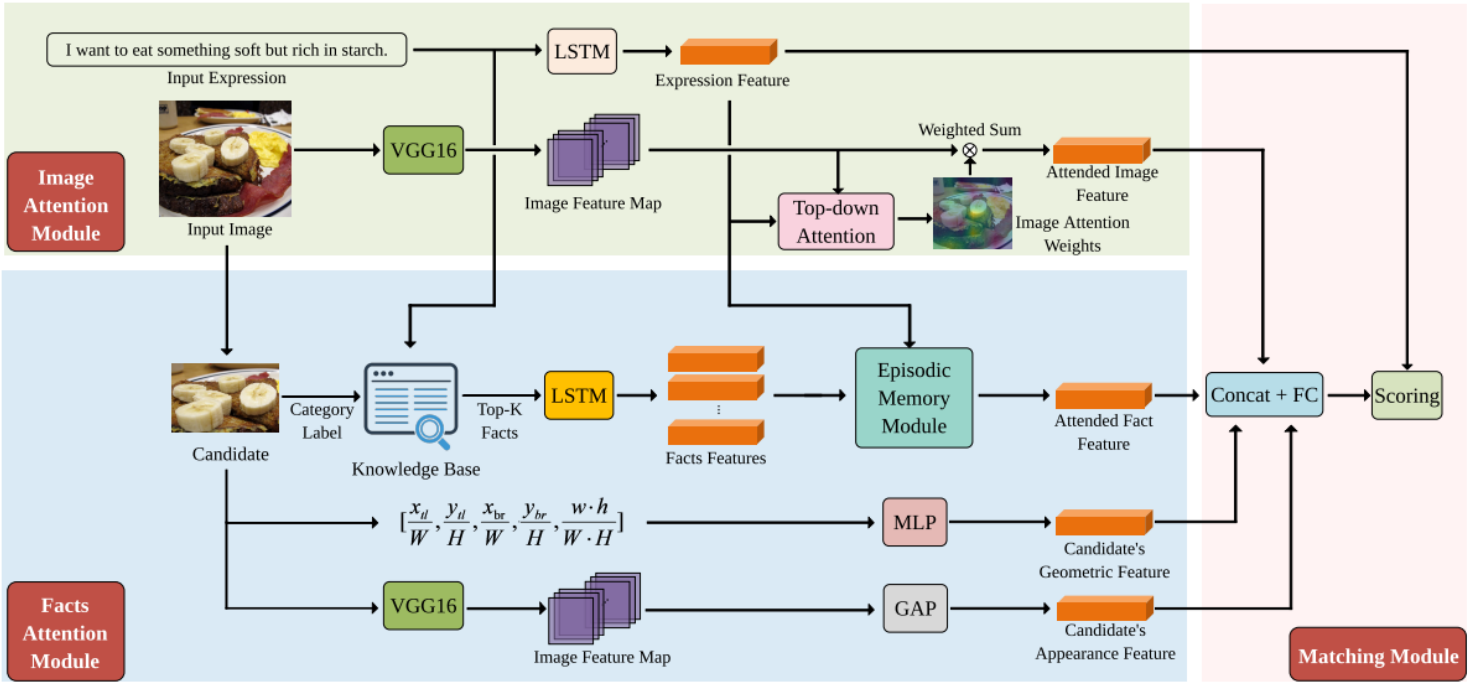


*CVPR 2020, Cops-Ref: A new Dataset and Task on Compositional Referring Expression Comprehension*

# KB-Ref



Expression: I want to eat something soft but rich in starch.
Target object: Banana
External knowledge: A banana is soft flesh rich in starch covered with a rind, which may be green, yellow, red, purple, or brown when ripe.




- 16,917 images (from Visual Genome)
- 43,284 referring expressions

# Further Work

- Fine-grained Cross-Modal Interaction
- Multi-Stage Detection
- Query-aware Detection
- Self-Supervised Learning