Introduction to
Human Interaction Recognition

Sungguk Cha

4 April 2022

Contents

- Introduction to Human Interaction Recognitions
- Benchmarks
- General Approaches

Introduction to
Human Interaction Recognition (HIR)

Introduction to

Human Interaction Recognition (HIR)

Fundamental visual recognition techniques are insufficient for HIR.

Introduction to

Human Interaction Recognition (HIR)

Fundamental visual recognition techniques are insufficient for HIR.

E.g., we want to recognize if a man is calling (talking on the phone)

Introduction to

Human Interaction Recognition (HIR)

Fundamental visual recognition techniques are insufficient for HIR.

E.g., we want to recognize if a man is calling (talking on the phone)

Figure. Man talking on the phone

Figure. Man holding the phone

Image sources:
https://www.123rf.com/photo_77373715_serious-man-calling-on-phone-and-working-on-laptop.html
https://www.dreamstime.com/young-man-holding-up-his-cell-phone-vest-handsome-standing-gray-blue-shirt-bowtie-isolated-white-background-image132759432

Introduction to

Human Interaction Recognition (

Fundamental visual recognition techniques

E.g., we want to recognize if a man is callin

First, we may detect human and phone.

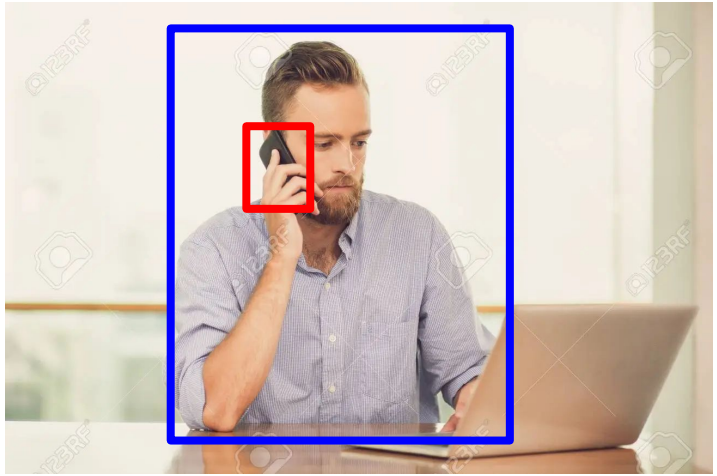


Figure. Man talking on the phone

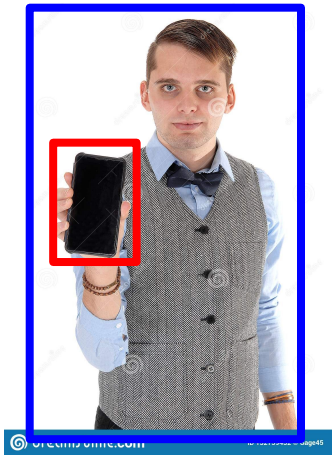


Figure. Man holding the phone

Image sources:
https://www.123rf.com/photo_77373715_serious-man-calling-on-phone-and-working-on-laptop.html
https://www.dreamstime.com/young-man-holding-up-his-cell-phone-vest-handsome-standing-gray-blue-shirt-bowtie-isolated-white-background-image132759432

Introduction to

Human Interaction Recognition (

Fundamental visual recognition techniques

E.g., we want to recognize if a man is callin

First, we may detect human and phone.
Next, we need to determine if the man is calling.

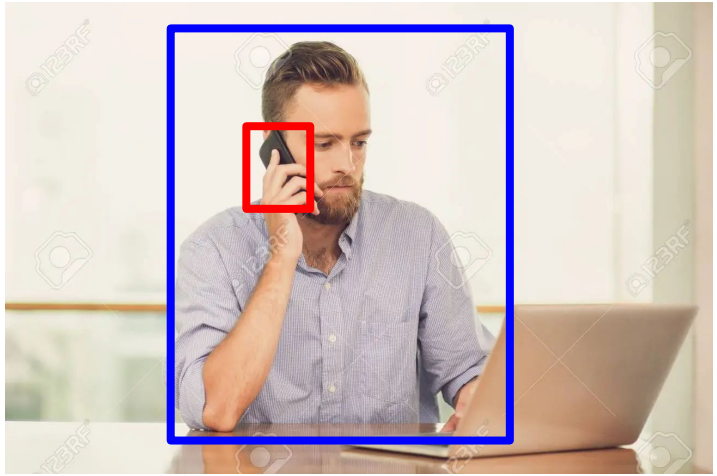


Figure. Man talking on the phone

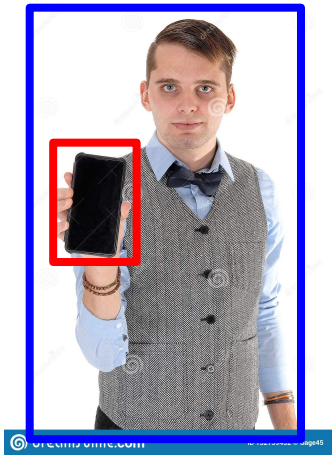


Figure. Man holding the phone

Image sources:
https://www.123rf.com/photo_77373715_serious-man-calling-on-phone-and-working-on-laptop.html
https://www.dreamstime.com/young-man-holding-up-his-cell-phone-vest-handsome-standing-gray-blue-shirt-bowtie-isolated-white-background-image132759432

Introduction to

Human Interaction Recognition (

Fundamental visual recognition techniques

E.g., we want to recognize if a man is callin

First, we may detect human and phone.
Next, we need to determine if the man is calling.
Technically, we can handle it naively.

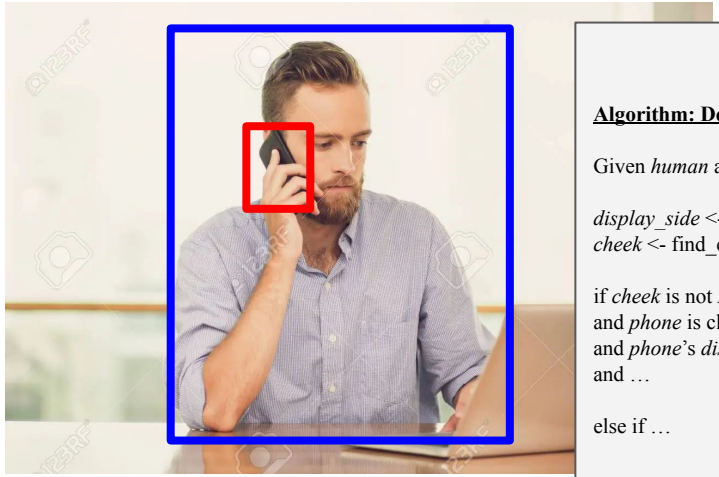


Figure. Man talking on the phone

Algorithm: Determine If Calling

Given *human* and *phone*

display_side <- find_phone_display_side(phone)
cheek <- find_cheek()

if *cheek* is not *None*
and *phone* is close to *cheek*
and *phone*'s *display_side* is toward *cheek*
and …

else if …

gure. Man holding the phone

Image sources:
https://www.123rf.com/photo_77373715_serious-man-calling-on-phone-and-working-on-laptop.html
https://www.dreamstime.com/young-man-holding-up-his-cell-phone-vest-handsome-standing-gray-blue-shirt-bowtie-isolated-white-background-image132759432

Introduction to

Human Interaction Recognition (

Fundamental visual recognition techniques

E.g., we want to recognize if a man is callin

First, we may detect human and phone.
Next, we need to determine if the man is calling.
Technically, we can handle it naively.

It is impractical and expensive engineering.

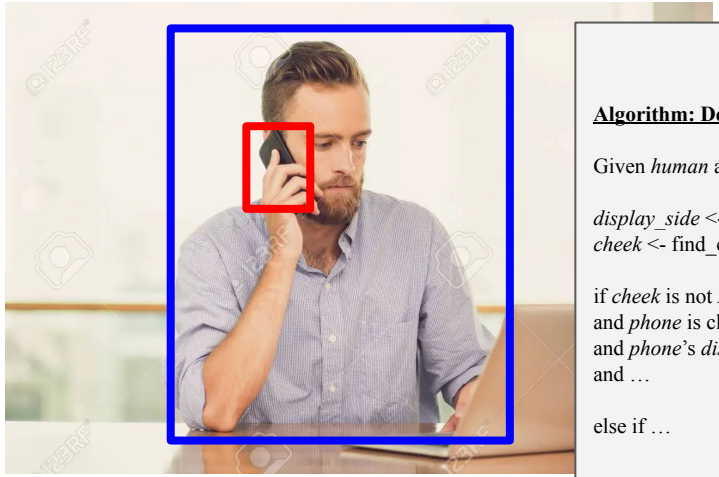


Figure. Man talking on the phone

Algorithm: Determine If Calling

Given *human* and *phone*

display_side <- find_phone_display_side(phone)
cheek <- find_cheek()

if *cheek* is not *None*
and *phone* is close to *cheek*
and *phone*'s *display_side* is toward *cheek*
and …

else if …

gure. Man holding the phone

Image sources:
https://www.123rf.com/photo_77373715_serious-man-calling-on-phone-and-working-on-laptop.html
https://www.dreamstime.com/young-man-holding-up-his-cell-phone-vest-handsome-standing-gray-blue-shirt-bowtie-isolated-white-background-image132759432

Introduction to

Human Interaction Recognition (

Fundamental visual recognition techniques

E.g., we want to recognize if a man is callin

First, we may detect human and phone.
Next, we need to determine if the man is calling.
Technically, we can handle it naively.

It is impractical and expensive engineering.
Solution to the demand, **HIR** arises.

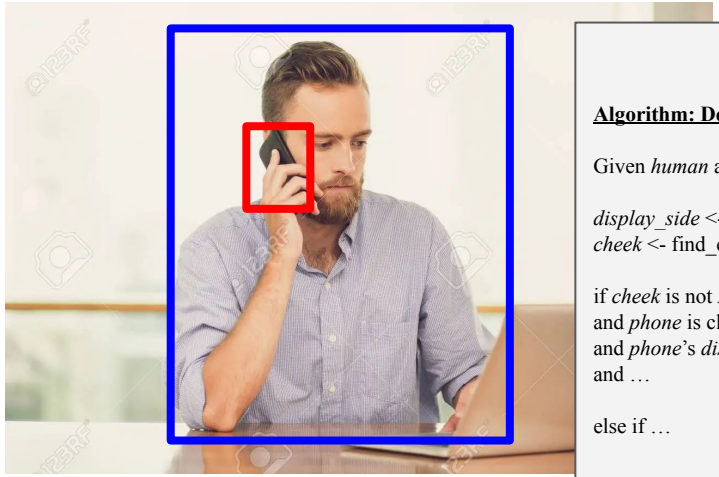


Figure. Man talking on the phone

Algorithm: Determine If Calling

Given *human* and *phone*

display_side <- find_phone_display_side(phone)
cheek <- find_cheek()

if *cheek* is not *None*
and *phone* is close to *cheek*
and *phone*'s *display_side* is toward *cheek*
and …

else if …

gure. Man holding the phone

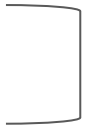
Image sources:
https://www.123rf.com/photo_77373715_serious-man-calling-on-phone-and-working-on-laptop.html
https://www.dreamstime.com/young-man-holding-up-his-cell-phone-vest-handsome-standing-gray-blue-shirt-bowtie-isolated-white-background-image132759432

Introduction to

Human Interaction Recognition

In this presentation, I present

- some benchmarks those are similar but different tasks
- each benchmark`s SOTA approach
- high-level explanation for general approaches



Benchmark section

General Approaches section

Introduction to

Human Interaction Recognition

In this presentation, I present

- some benchmarks those are similar but different tasks
- each benchmark`s SOTA approach
- high-level explanation for general approaches

Benchmark section

General Approaches section

I want you to think

“what is HIR”,

and “how others have solved it?”

Benchmarks

Benchmarks

Benchmarks represent research field`s objective.

Benchmarks

Benchmarks represent research field`s objective.

In this section, I introduce benchmarks along with their SOTA methods.

Benchmarks

Benchmarks represent research field`s objective.

In this section, I introduce benchmarks along with their SOTA methods.

I prepared 3 major benchmarks. Let’s look into it.

- HICO
- HICO-DET
- V-COCO

Benchmarks

Benchmarks represent research field`s objective.

In this section, I introduce benchmarks along with their SOTA methods.

I prepared 3 major benchmarks. Let’s look into it.

- HICO
- HICO-DET
- V-COCO

These three benchmarks are about human interaction understanding with common objects

HICO

Human Interacting with Common Objects

HICO: A Benchmark for Recognizing Human-Object Interactions in Images, Yu-Wei (UMICH) et al., ICCV2015

HICO

Human Interacting with Common Objects

HICO: A Benchmark for Recognizing Human-Object Interactions in Images, Yu-Wei (UMICH) et al., ICCV2015

three highlight **key features**

HICO
Human Interacting with Common Objects

HICO: A Benchmark for Recognizing Human-Object Interactions in Images, Yu-Wei (UMICH) et al., ICCV2015

three highlight **key features**

- diverse interactions

Dataset	#images	#actions	Sense	Clean
Sports event dataset [18]	1579	8	Y	Y
Ikizler *et al.* [11]	467	6	Y	Y
Ikizler-Cinbis *et al.* [12]	1727	5	Y	Y
The sports dataset [9]	300	6	Y	Y
Pascal VOC 2010 [6]	454	9	Y	Y
Pascal VOC 2011 [6]	2424	10	Y	Y
Pascal VOC 2012 [6]	4588	10	Y	Y
PPMI [33]	4800	12	Y	Y
Willow dataset [3]	968	7	Y	Y
Stanford 40 Actions [35]	9532	40	Y	Y
TBH dataset [23]	341	3	Y	Y
HICO (ours)	**47774**	**600**	**Y**	**Y**
89 action dataset [16]	2038	89	N	Y
TUHOI [17]	10805	2974	N	Y
MPII Human Pose [1]	40522	410	Y	Y
Google Image Search [24]	102830	2938	N	N

Table 2: Comparison of existing image datasets on action recognition. "Sense" means whether the category list is based on senses instead of words. "Clean" means whether the dataset is human verified.

HICO
Human Interacting with Common Objects

HICO: A Benchmark for Recognizing Human-Object Interactions in Images, Yu-Wei (UMICH) et al., ICCV2015

three highlight **key features**

- diverse interactions
- *sense* based Human-Object Interaction (HOI) categories
  - not *word*-based
  - e.g., “repair a bike”, “fix a bicycle” are the same

Dataset	#images	#actions	Sense	Clean
Sports event dataset [18]	1579	8	Y	Y
Ikizler *et al.* [11]	467	6	Y	Y
Ikizler-Cinbis *et al.* [12]	1727	5	Y	Y
The sports dataset [9]	300	6	Y	Y
Pascal VOC 2010 [6]	454	9	Y	Y
Pascal VOC 2011 [6]	2424	10	Y	Y
Pascal VOC 2012 [6]	4588	10	Y	Y
PPMI [33]	4800	12	Y	Y
Willow dataset [3]	968	7	Y	Y
Stanford 40 Actions [35]	9532	40	Y	Y
TBH dataset [23]	341	3	Y	Y
HICO (ours)	**47774**	**600**	**Y**	**Y**
89 action dataset [16]	2038	89	N	Y
TUHOI [17]	10805	2974	N	Y
MPII Human Pose [1]	40522	410	Y	Y
Google Image Search [24]	102830	2938	N	N

Table 2: Comparison of existing image datasets on action recognition. “Sense” means whether the category list is based on senses instead of words. “Clean” means whether the dataset is human verified.

HICO

Human Interacting with Common Objects

HICO: A Benchmark for Recognizing Human-Object Interactions in Images, Yu-Wei (UMICH) et al., ICCV2015

three highlight **key features**

- diverse interacti
- *sense* based HC
  - not *word-*
  - e.g., “repa
- multilabeled!

Figure 1: The “Humans Interacting with Common Objects” (HICO) dataset.

HICO

Human Interacting with Common Objects

HICO: A Benchmark for Recognizing Human-Object Interactions in Images, Yu-Wei (UMICH) et al., ICCV2015

three highlight **key features**

- diverse interacti
- *sense* based HC
  - not *word-*
  - e.g., "repa
- multilabeled!

Ternary labeled
yes, no, not sure

Figure 1: The "Humans Interacting with Common Objects" (HICO) dataset.

HICO

Task description

Figure 1: The “Humans Interacting with Common Objects” (HICO) dataset.

The task is multi-label classification.
E.g., categories are ‘hold-bicycle’, ‘ride-bicycle’, ‘repair cell phone’, …

HICO
HOI category

	#action	#HOI	#object	#action/object
MPII Human Pose [1]	410	102	66	1.55
HICO (ours)	**520**	**520**	**80**	**6.50**

Table 5: Comparison of action/HOI categories between MPII Human Pose [1] and our dataset (excluding "no interaction" classes).

- 80 objects (MSCOCO setting)
  - E.g., 'hold-**bicycle**', 'ride-**bicycle**', 'repair **cell phone**', …
- 520 HOI categories
  - E.g., **'hold-bicycle'**, **'ride-bicycle'**, **'repair cell phone'**, …

SOTA on HICO

DEtection FRee (DEFR)

The Overlooked Classifier in Human-Object Interaction Recognition, Ying et al., 2021

SOTA on HICO

DEtection FRee (DEFR)

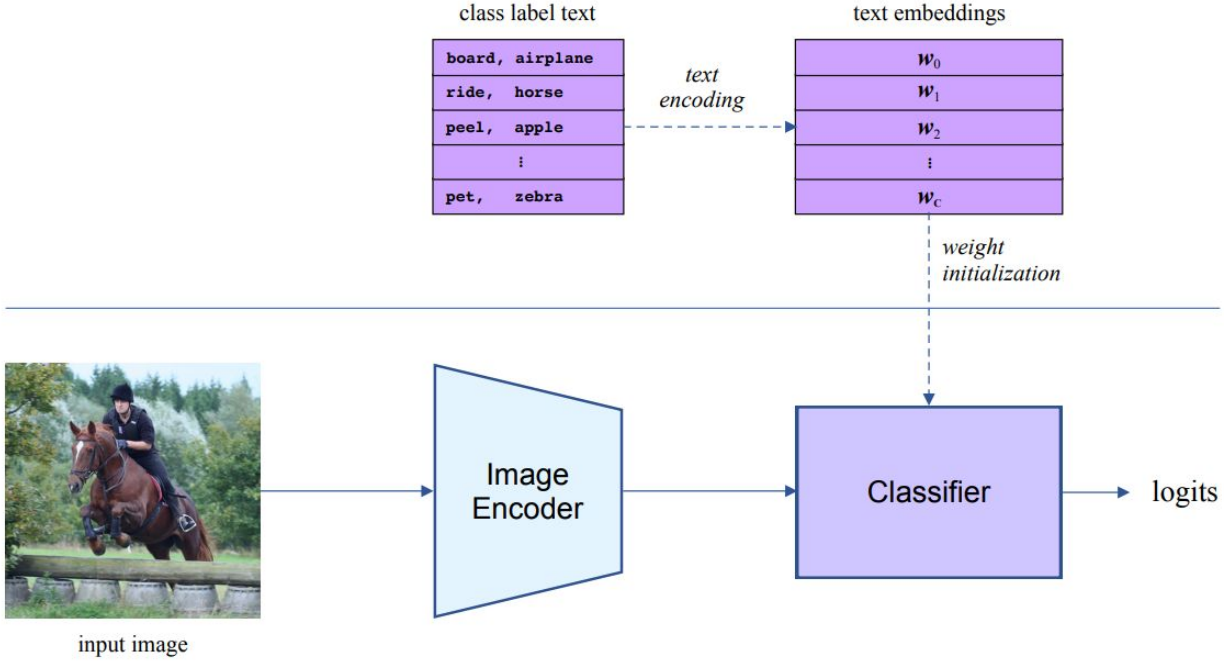
The Overlooked Classifier in Human-Object Interaction Recognition, Ying et al., 2021

Rank	Model	mAP↑	Extra Training Data	Pape
1	**DEFR**	65.6	×	The Rec
2	**HAKE**	47.1	✓	HA
3	**HAKE**	46.3	✓	Pa
4	**Pairwise-Part**	39.9	×	Pai Ob

SOTA on HICO

DEtection FRee (DEFR)

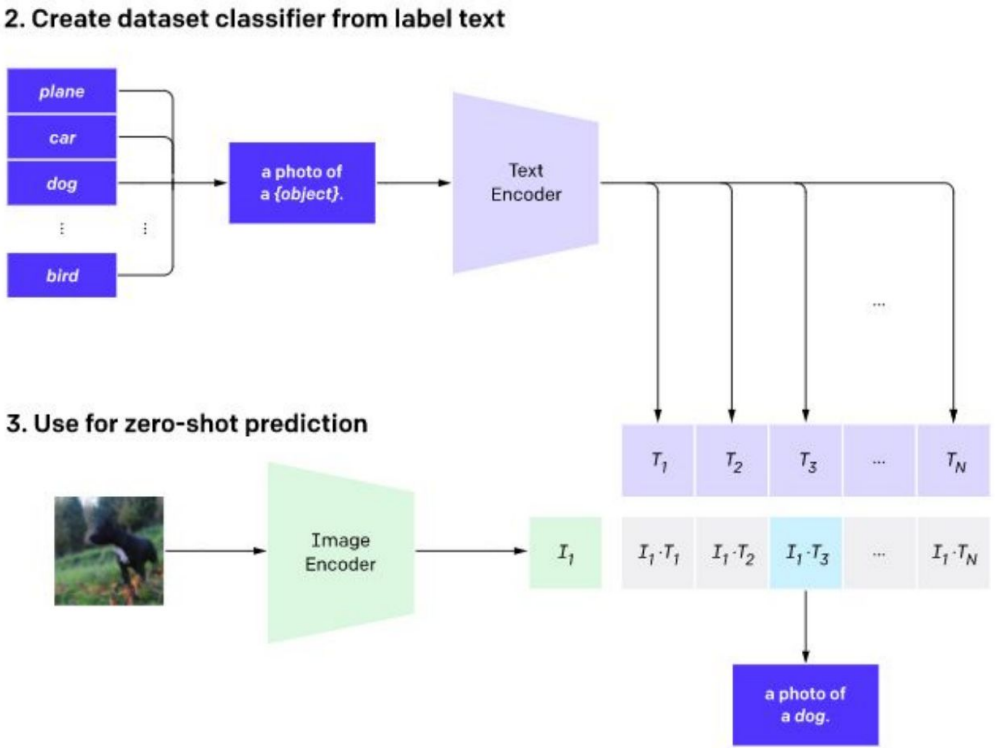
The Overlooked Classifier in Human-Object Interaction Recognition, Ying et al., 2021



SOTA on HICO

DEtection FRee (DEFR)

The Overlooked Classifier in Human-Object Interaction Recognition, Ying et al., 2021



I.e.,
DEFR =
CLIP
+ their own loss (for multi-labeled classification)

SOTA on HICO

DEtection FRee (DEFR)

The Overlooked Classifier in Human-Object Interaction Recognition, Ying et al., 2021

In contrast to previous approaches,
DEFR brought multi-modal (NLP knowledge)
concept and got overwhelming performance.

		mAP↑	Extra Training Data	Pape
1	DEFR	65.6	×	The Rec
2	HAKE	47.1	✓	HA
3	HAKE	46.3	✓	Pas
4	Pairwise-Part	39.9	×	Pai Ob

HICO-DET

Human Interacting with Common Objects Detection

Learning to Detect Human-Object Interactions, Yu-Wei (UMICH) et al., WACV2018

HICO-DET
Example

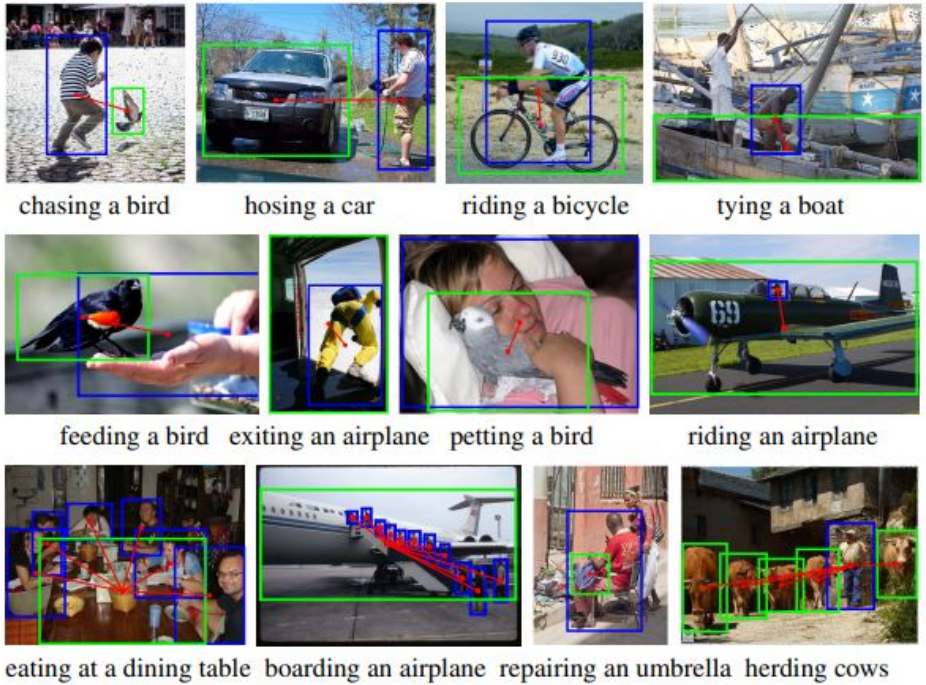


Figure 6: Sample annotations of our HICO-DET.

Task requires

1. **detection humans and objects**
2. **matching human-object**
3. **classifying the interaction**

HICO-DET
Example

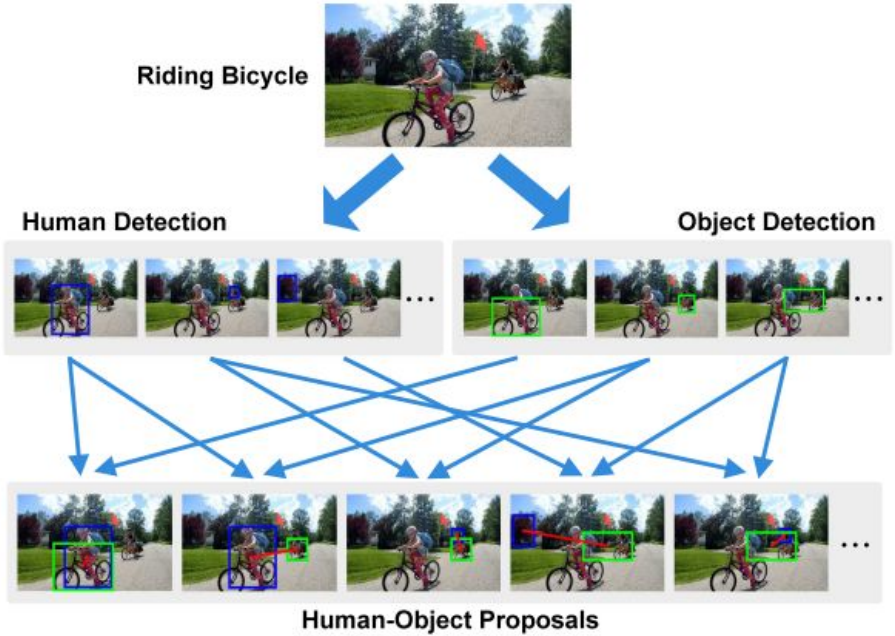


Figure 2: Generating human-object proposals from human and object detections.

This figure explains general approaches.

1. two-stage approach
    a. first, detect
    b. second, match and classify
2. one-stage approach
    a. detect and classify at once
    b. like detection transformer (DETR)

HICO-DET

Human Interacting with Common Objects Detection

Learning to Detect Human-Object Interactions, Yu-Wei (UMICH) et al., WACV2018

Contributions

HICO-DET

Human Interacting with Common Objects Detection

Learning to Detect Human-Object Interactions, Yu-Wei (UMICH) et al., WACV2018

Contributions

- The first large benchmark for HOI detection
    - by augmenting HICO classification with instance annotations

HICO-DET

Human Interacting with Common Objects Detection

Learning to Detect Human-Object Interactions, Yu-Wei (UMICH) et al., WACV2018

Contributions

- The first large benchmark for HOI detection
  - by augmenting HICO classification with instance annotations
-

	HICO-DET			
	#image	#positive	#instance	#bounding box
Train	38118	70373	117871 (1.67/pos)	199733 (2.84/pos)
Test	9658	20268	33405 (1.65/pos)	56939 (2.81/pos)
Total	47776	90641	151276 (1.67/pos)	256672 (2.83/pos)

Table 1: Statistics of annotations in our HICO-DET.

SOTA on HICO-DET

QAHOI

QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021

SOTA on HICO-DET

QAHOI

QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021

Rank	Model	mAP ↑	Time Per Frame (ms)	Extra Training Data
1	**QAHOI**	35.78		×
2	**UPT-R101-DC5**	32.62	124	×
3	**DEFR**	32.35		×
4	**UPT-R101**	32.31	61	×
5	**CDN** (ResNet101)	32.07		×
6	**UPT-R50**	31.66	42	×

SOTA on HICO-DET

QAHOI

QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021

Rank	Model	mAP ↑	Time Per Frame (ms)	Extra Training Data
1	**QAHOI**	35.78		×
2	**UPT-R101-DC5**	32.62	124	×
3	**DEFR**	32.35		×
4	**UPT-R101**	32.31	61	×
5	**CDN** (ResNet101)	32.07		×
6	**UPT-R50**	31.66	42	×

One-stage approach (transformer)

SOTA on HICO-DET

QAHOI

QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021

Rank	Model	mAP ↑	Time Per Frame (ms)	Extra Training Data
1	**QAHOI**	35.78		×
2	**UPT-R101-DC5**	32.62	124	×
3	**DEFR**	32.35		×
4	**UPT-R101**	32.31	61	×
5	**CDN** (ResNet101)	32.07		×
6	**UPT-R50**	31.66	42	×

One-stage approach (transformer)

Two-stage approach

SOTA on HICO-DET

QAHOI

QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021

Rank	Model	mAP ↑	Time Per Frame (ms)	Extra Training Data
1	**QAHOI**	35.78		×
2	**UPT-R101-DC5**	32.62	124	×
3	**DEFR**	32.35		×
4	**UPT-R101**	32.31	61	×
5	**CDN** (ResNet101)	32.07		×
6	**UPT-R50**	31.66	42	×

One-stage approach

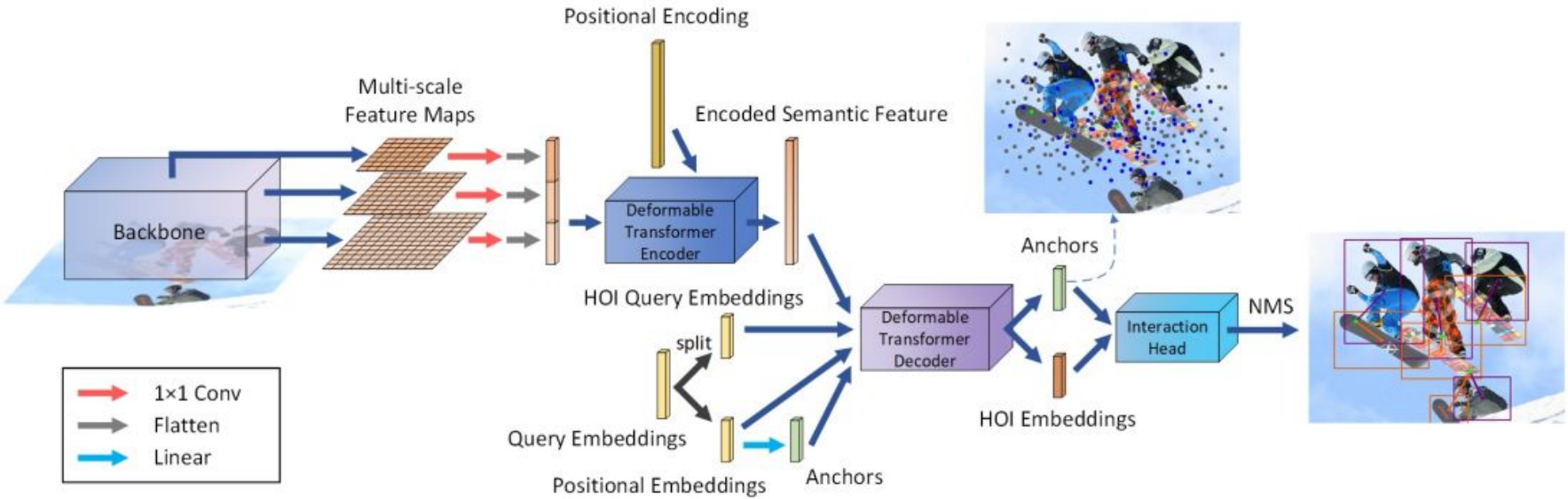
Two-stage approach

CLIP based two-stage approach

SOTA on HICO-DET

QAHOI

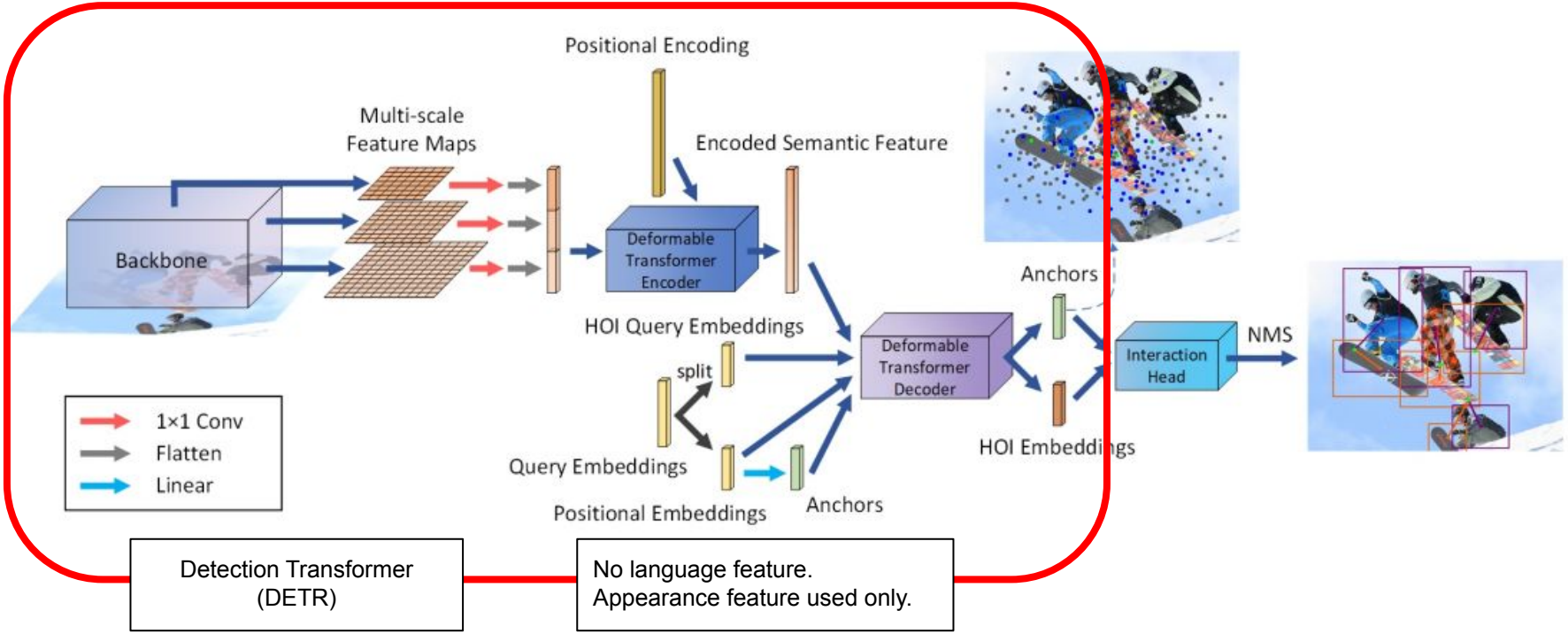
QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021



SOTA on HICO-DET

QAHOI

QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021



Detection Transformer (DETR)

No language feature.
Appearance feature used only.

SOTA on HICO-DET

QAHOI

QAHOI: Query-Based Anchors for Human-Object Interaction Detection, Junwen and Keiji, 2021

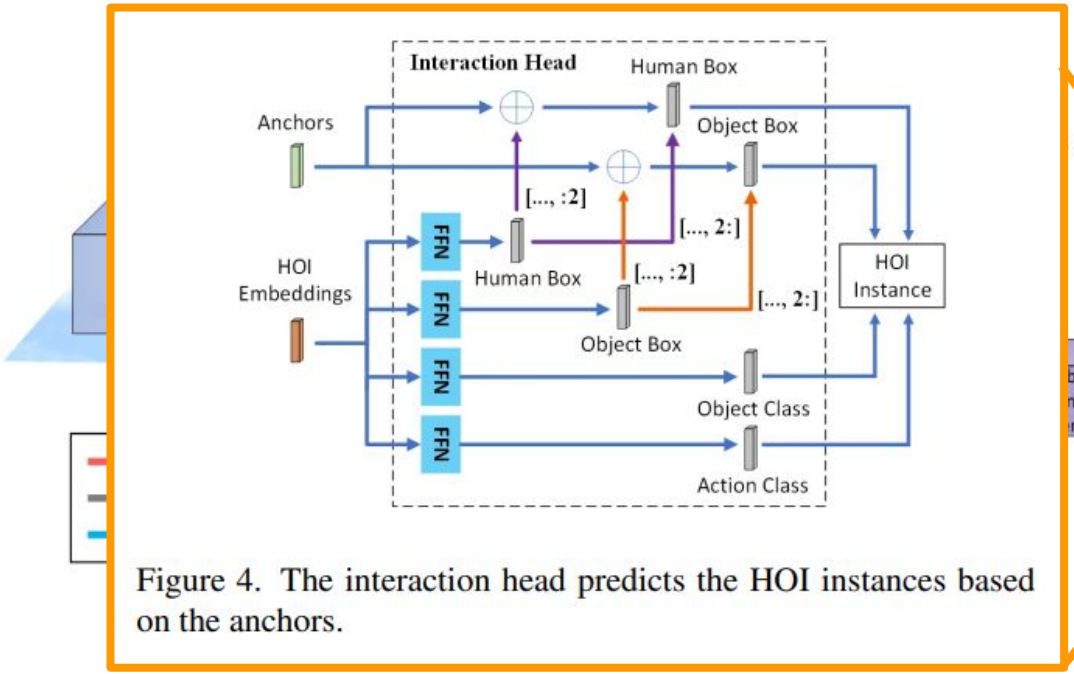
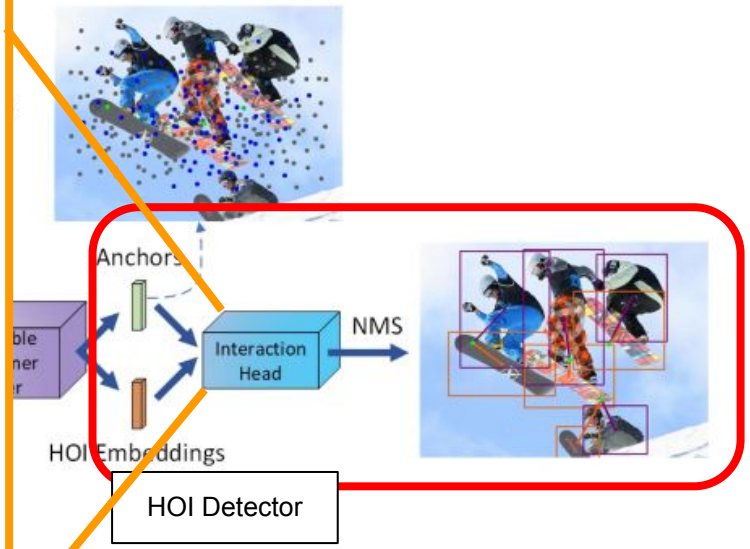


Figure 4. The interaction head predicts the HOI instances based on the anchors.



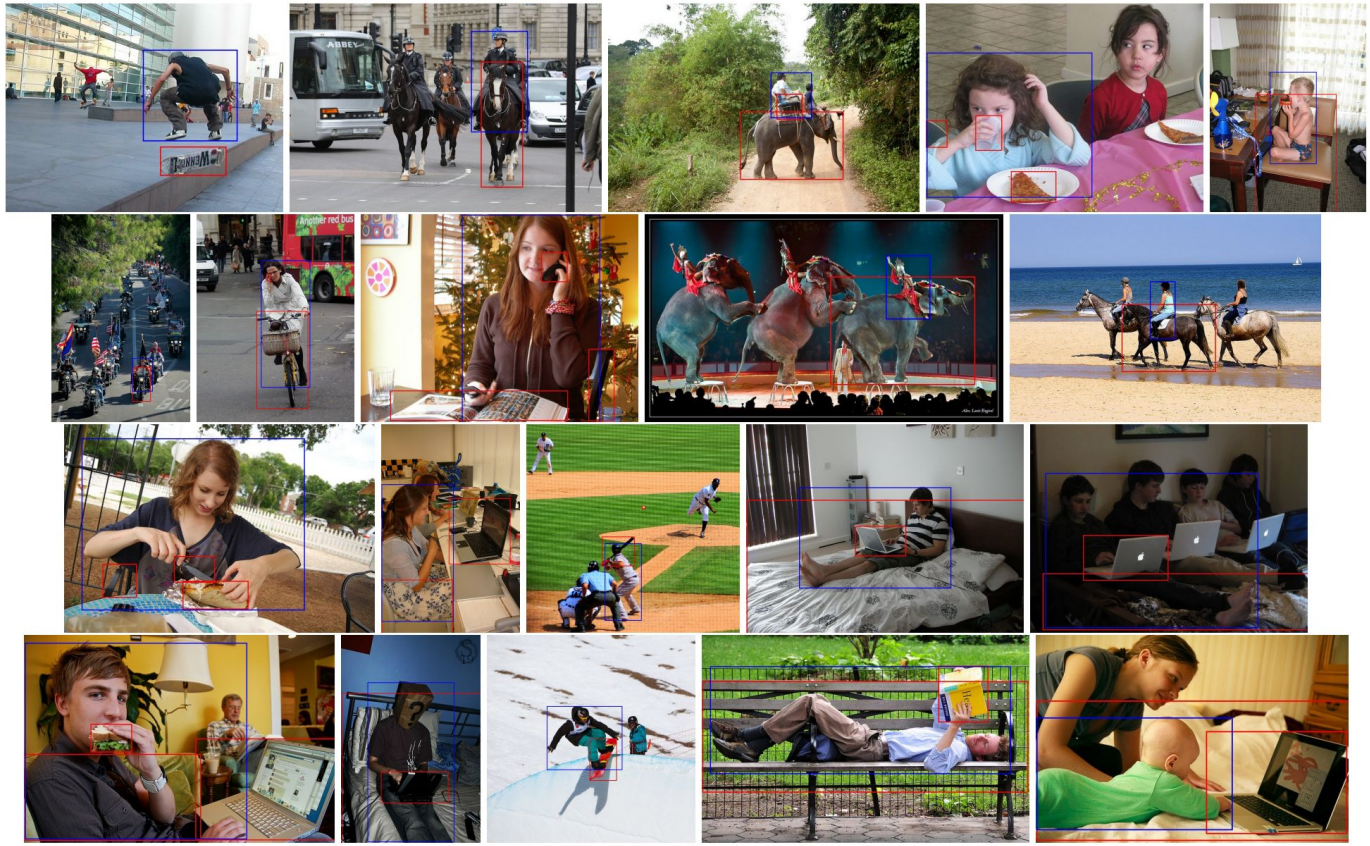
V-COCO
Verbs in COCO

Visual Semantic Role Labeling, Saurabh and Jitendra, 2015

V-COCO

Verbs in C

Visual Semar



V-COCO
Verbs in COCO

Visual Semantic Role Labeling, Saurabh and Jitendra, 2015

Very similar to HICO-DET, but less number of images, annotations and action categories.

V-COCO
Verbs in COCO

Visual Semantic Role Labeling, S

Very similar to HICO-DET, but les ories.

Rank	Model	AP(S1)↑	AP(S2)	Time Per Frame(ms)
1	**OCN** (ResNet101)	65.3	67.1	
2	**OCN** (ResNet50)	64.2	66.3	43
3	**CDN** (ResNet101)	63.91	65.89	
4	**UPT-R101-DC5**	61.3	67.1	131

V-COCO
Verbs in COCO

Visual Semantic Role Labeling, S

Very similar to HICO-DET, but les ories.

Rank	Model	AP(S1)↑	AP(S2)	Time Per Frame(ms)
1	**OCN** (ResNet101)	65.3	67.1	
2	**OCN** (ResNet50)	64.2	66.3	43
3	**CDN** (ResNet101)	63.91	65.89	
4	**UPT-R101-DC5**	61.3	67.1	131

SOTA on V-COCO

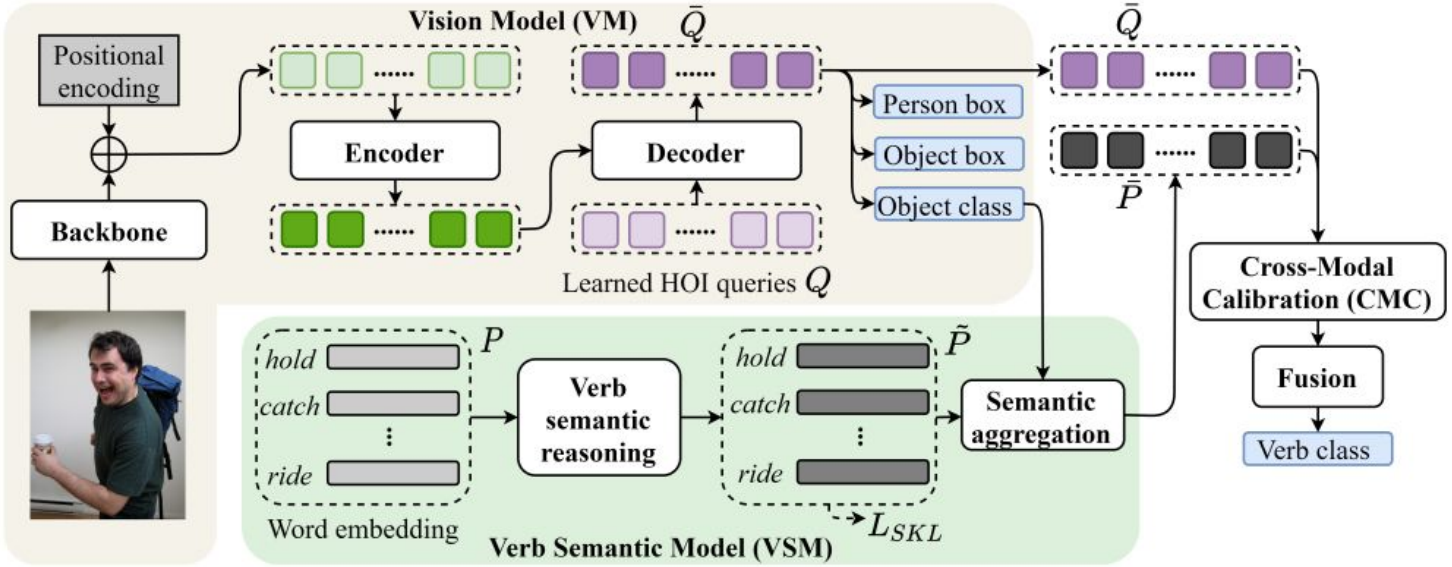
Object-guided Cross-modal Calibration Network (OCN)

Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics, Hangjie et al., 2022

SOTA on V-COCO

Object-guided Cross-modal Calibration Network (OCN)

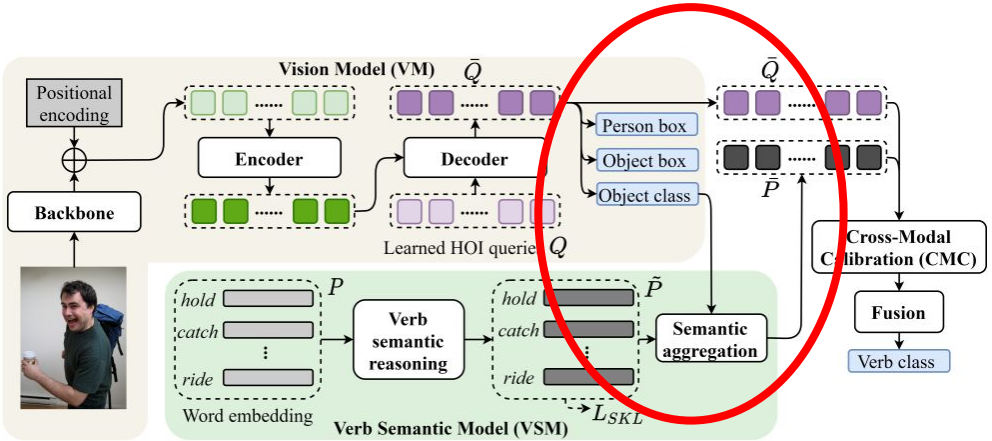
Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics, Hangjie et al., 2022



SOTA on V-COCO

Object-guided Cross-modal Calibration Network (OCN)

Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics, Hangjie et al., 2022

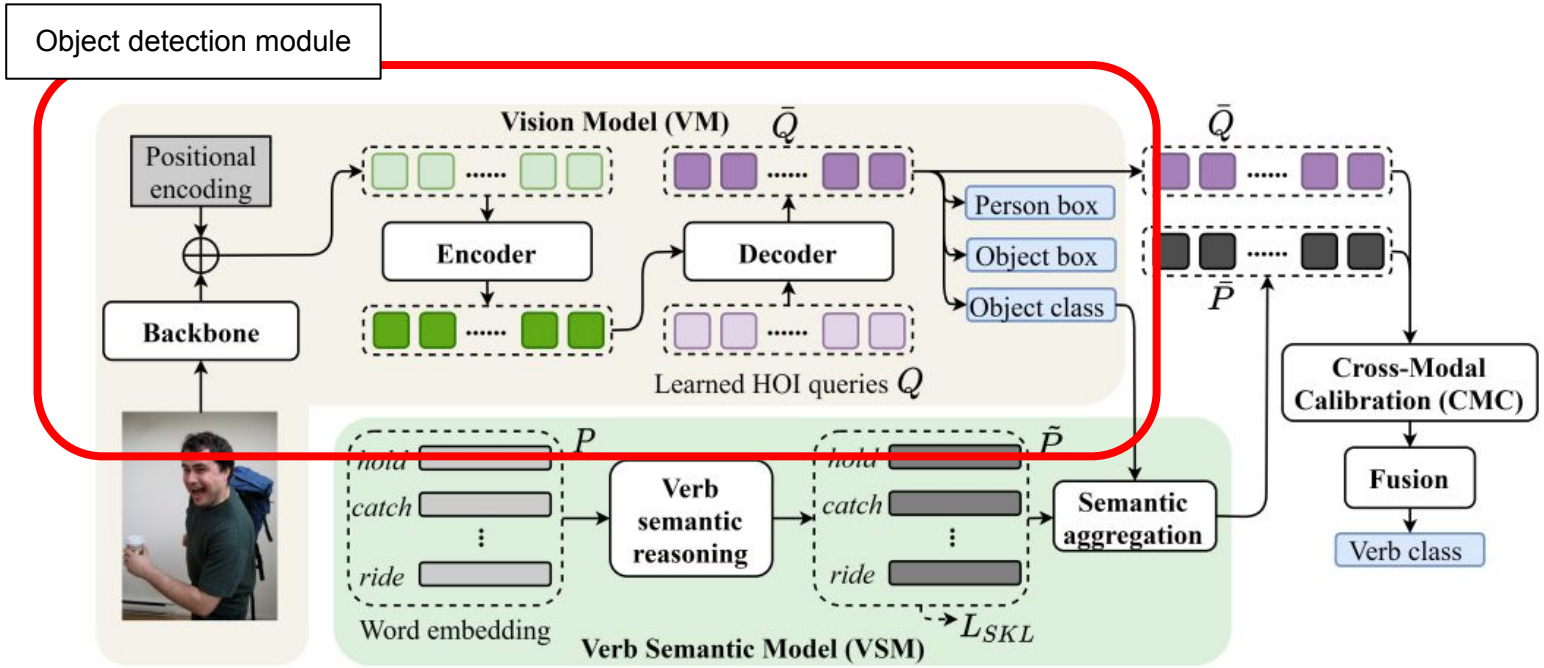


The main idea is to guide *action* by object.
I.e.,
object prediction effects action prediction

SOTA on V-COCO

Object-guided Cross-modal Calibration Network (OCN)

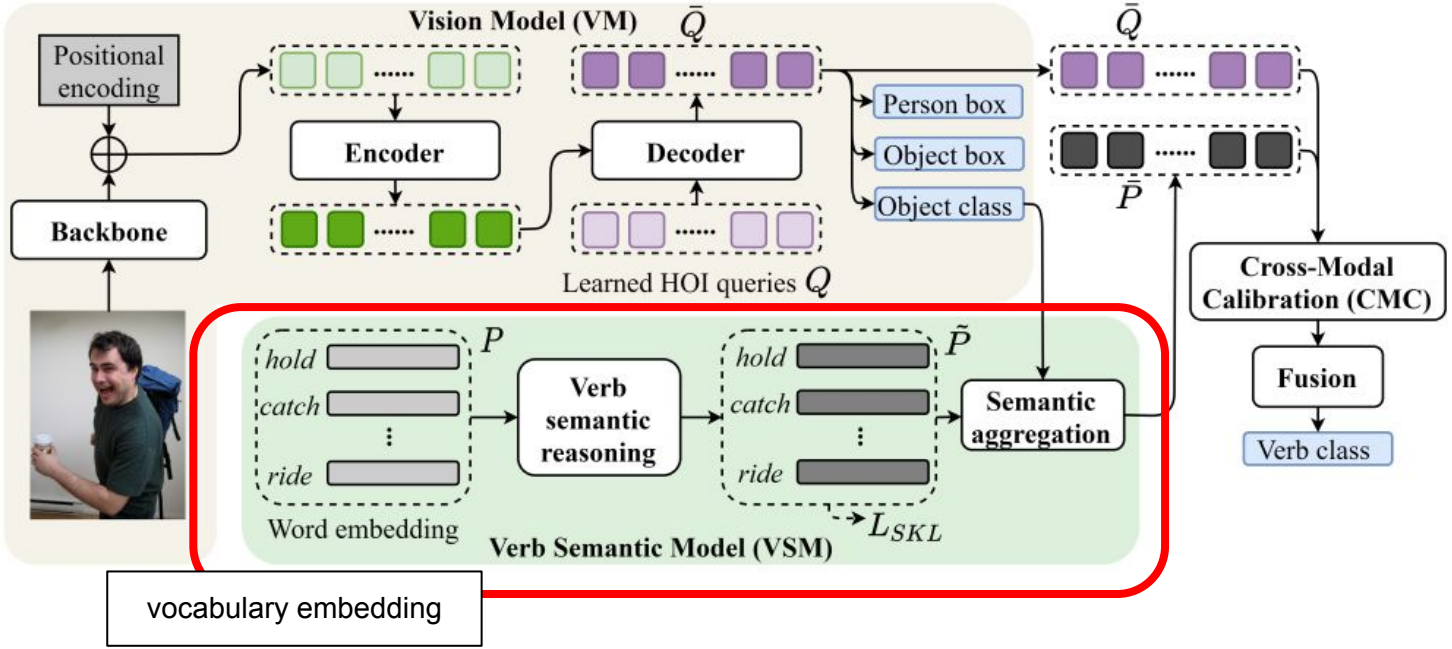
Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics, Hangjie et al., 2022



SOTA on V-COCO

Object-guided Cross-modal Calibration Network (OCN)

Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics, Hangjie et al., 2022

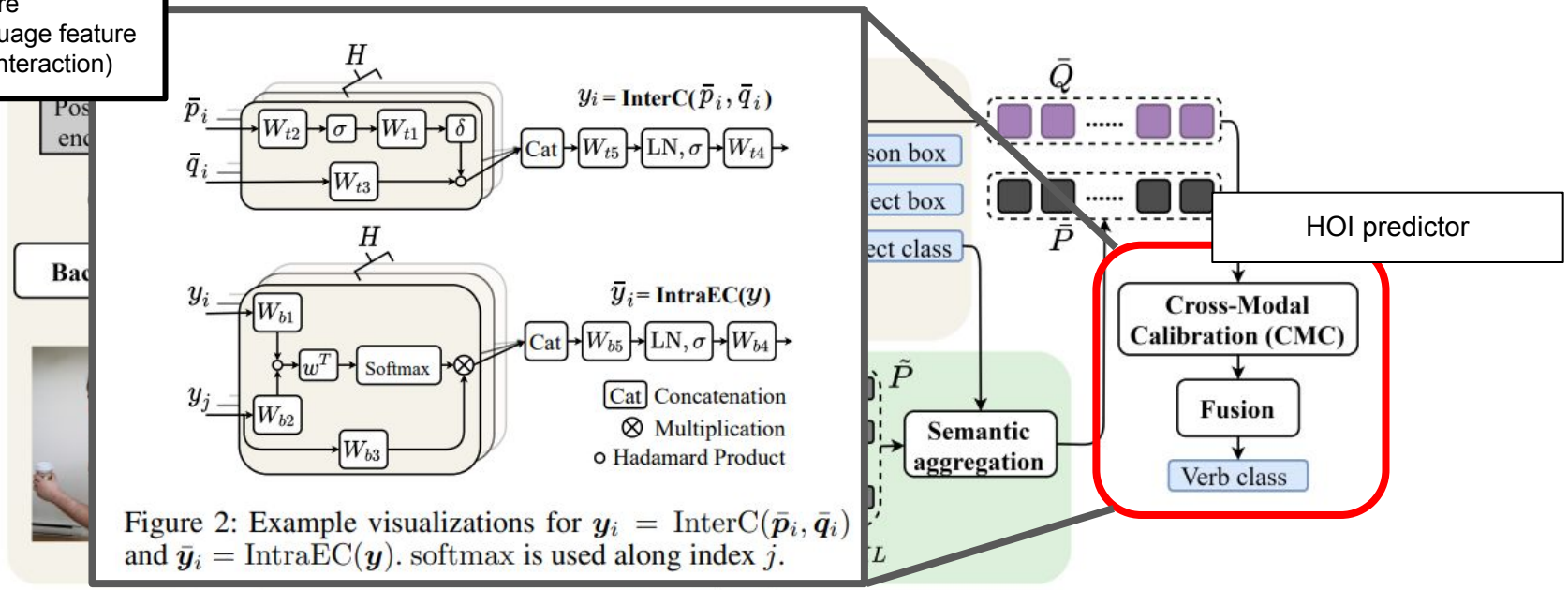


SOTA on V-COCO

Object-guided Cross-modal Calibration Network (OCN)

Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics, Hangjie et al., 2022

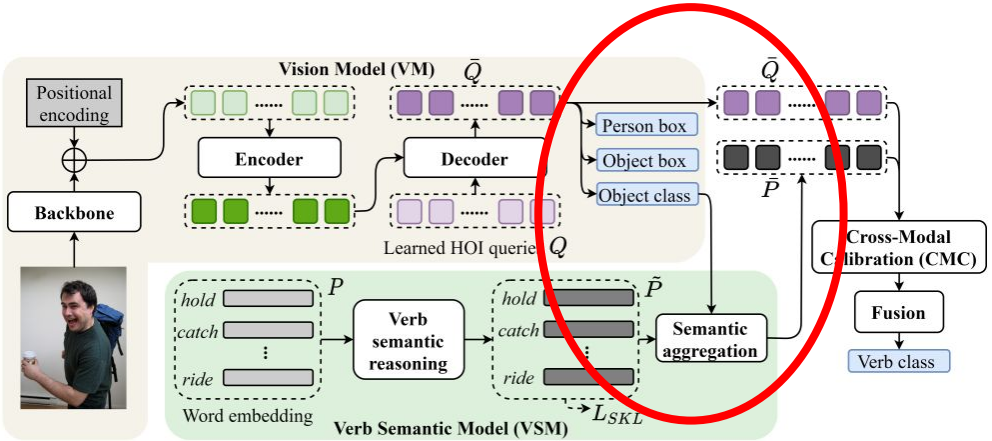
p := visual feature
q := visual+language feature
(about interaction)



SOTA on V-COCO

Object-guided Cross-modal Calibration Network (OCN)

Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics, Hangjie et al., 2022



Their main idea is to guide ***interaction prediction*** with ***object prediction***.

It consists of

- detection module
- action embedding module
- HOI predictor

General Approaches

High-level Explanations for HIR

General Approaches
High-level Explanations for HIR

There are two HIR tasks *dealing with common objects*.

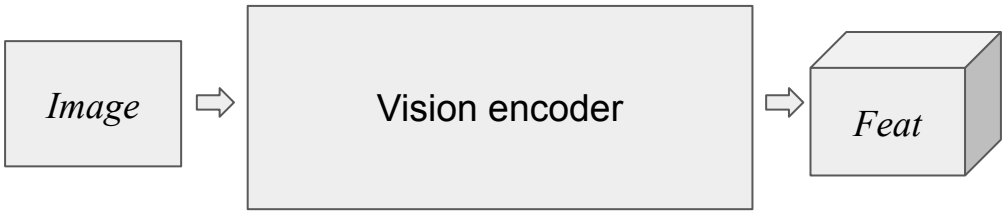
- HOI classification
- HOI detection

General Approaches
HOI classification

CLIP based approach

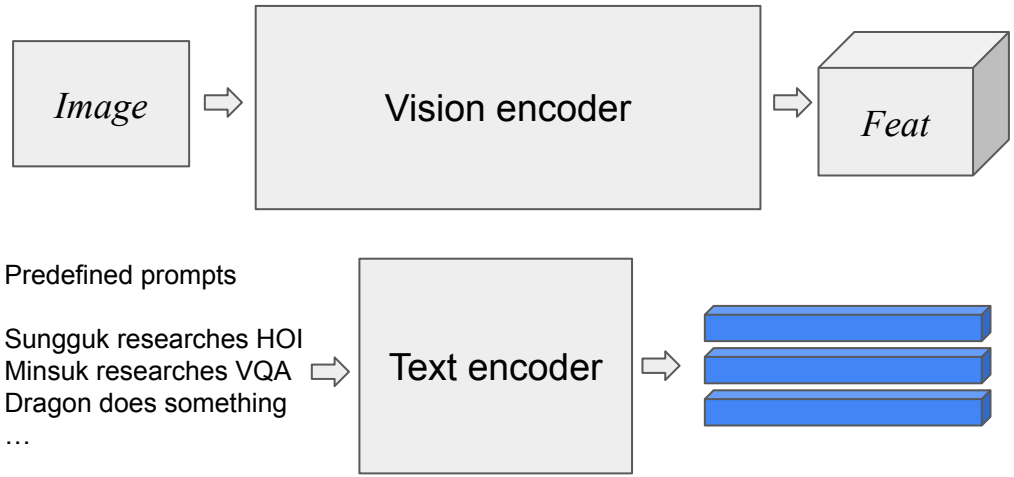
General Approaches
HOI classification

CLIP based approach



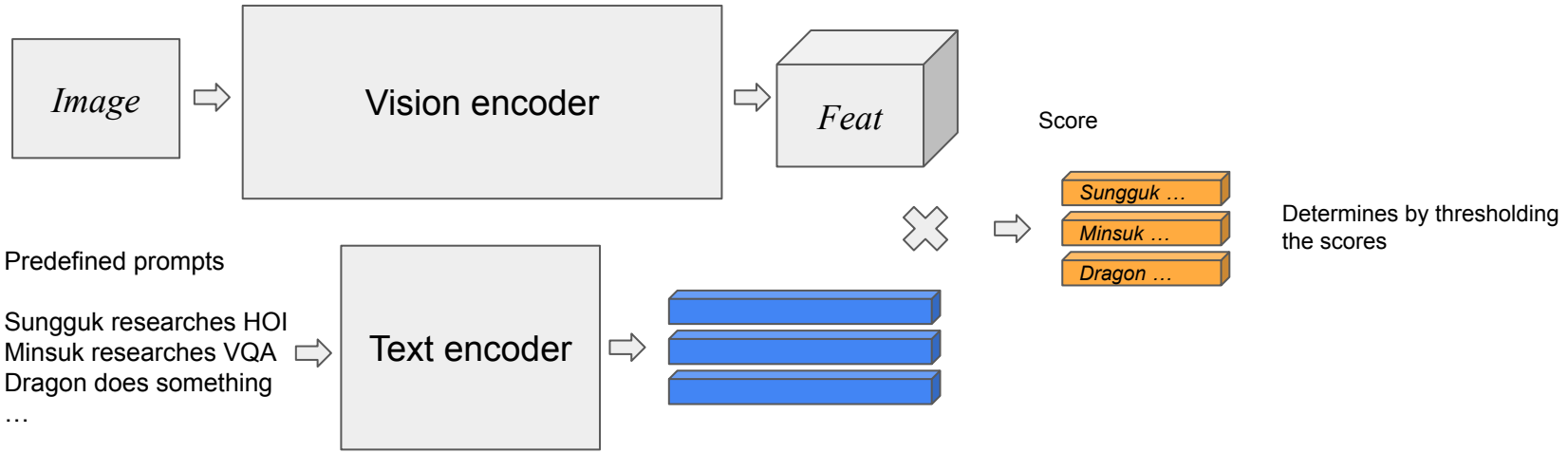
General Approaches
HOI classification

CLIP based approach



General Approaches
HOI classification

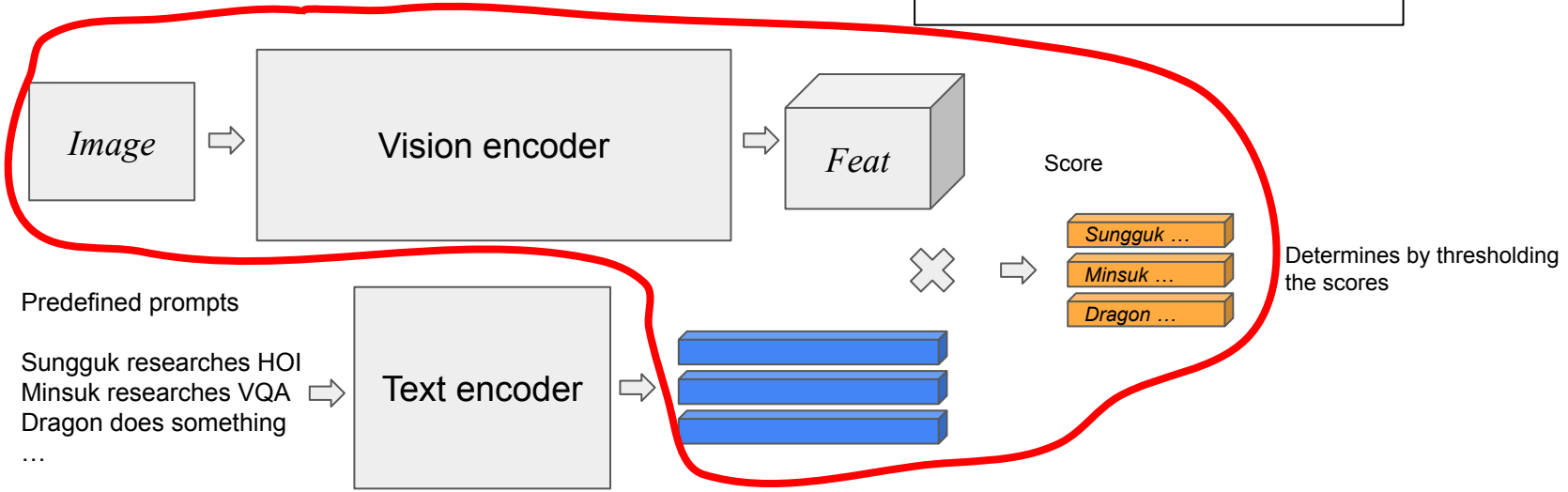
CLIP based approach



General Approaches
HOI classification

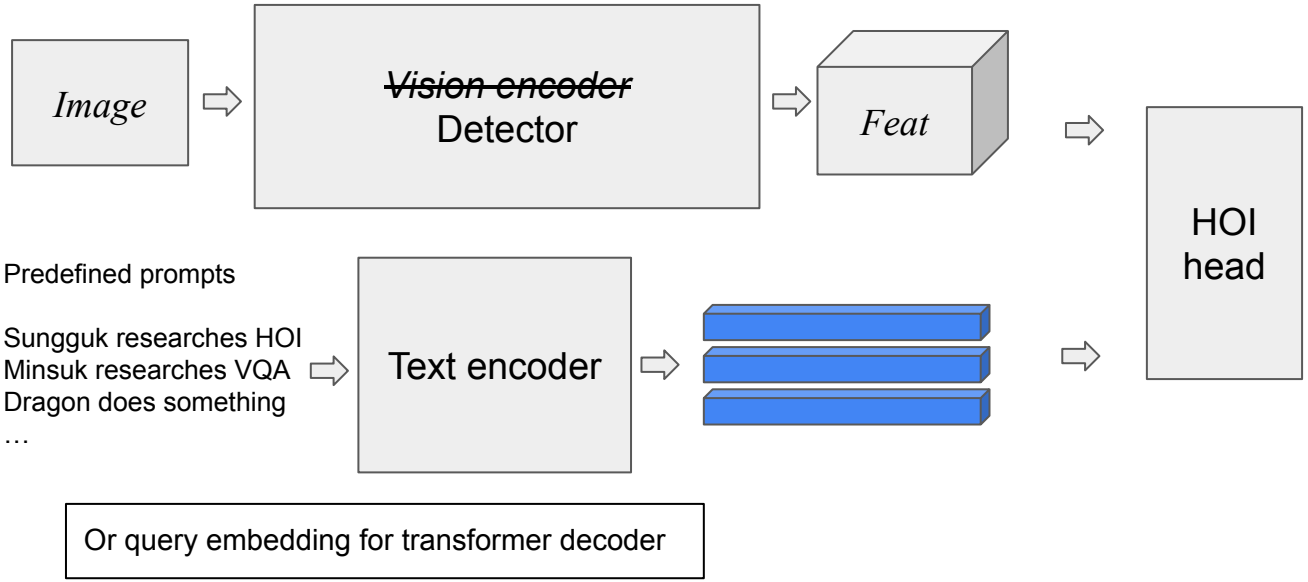
CLIP based approach

Actual model size can be as minimized as vision encoder.



General Approaches
HOI detection

General Approaches
HOI detection



detector and HOI head added.

TL; DR

- Introduce **human interaction recognition**.
- Visit **benchmarks**.
- Explain general approaches.