HOI-R1：マルチモーダル大規模言語モデルによる HOI 検出

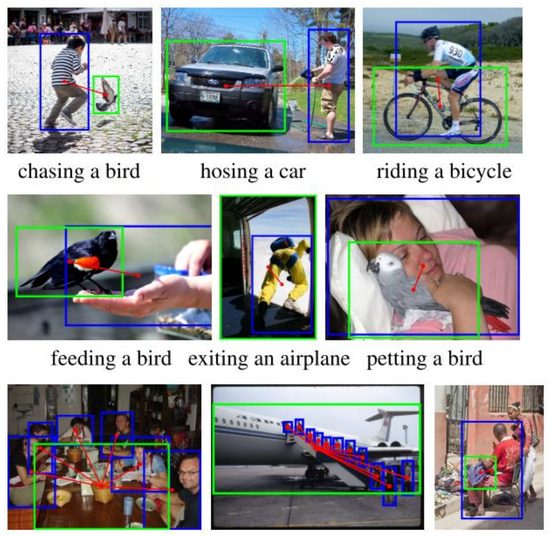
陳　俊文、熊　培林、柳井　啓司
電気通信大学

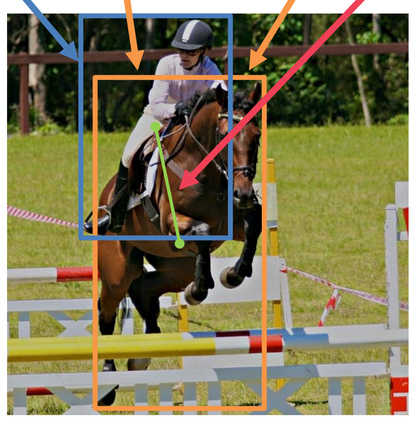
IS1-101

HOI 検出

$$\left\{\left[x_1^{\text{human}}, y_1^{\text{human}}, x_2^{\text{human}}, y_2^{\text{human}}\right], \left[x_1^{\text{obj}}, y_1^{\text{obj}}, x_2^{\text{obj}}, y_2^{\text{obj}}\right], c_{\text{HOI}}\right\}$$

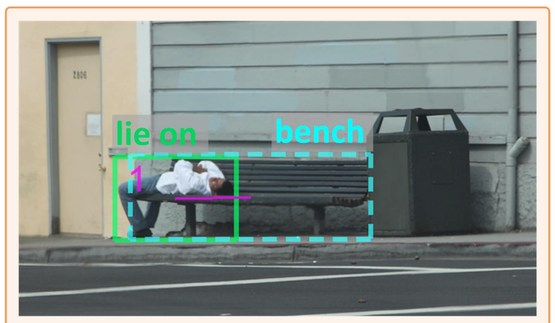
$c_{\text{HOI}} : \left[c_{\text{obj}}, c_{\text{action}}\right]$

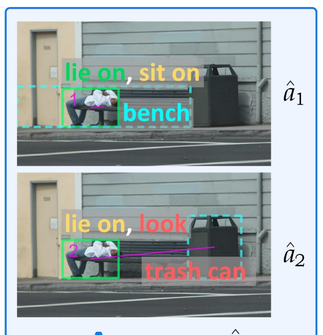


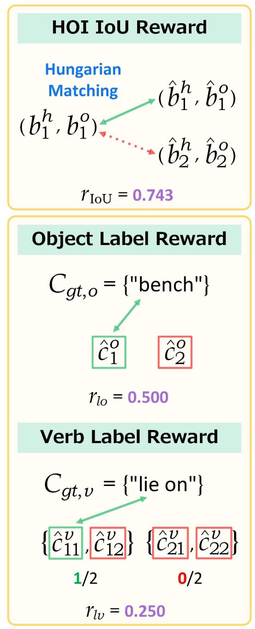


- 近年、人間と物体のインタラクション検出、HOI 検出は、大きな応用可能性を持つ分野として注目されている
- 47,776 枚の画像（トレーニングセット 38,118 枚、テストセット 9,658 枚）を含む HICO-DET は HOI 検出に最もよく使われるデータセット
  - HOI クラスは、117 個の動詞クラスと 80 個の物体クラスから構成される 600 種類がある

HOI-R1 の報酬関数







Format HOI Key Reward

$r_{\text{tag}} : 1$　$r_{ko} : 1 + 0$
$r_b : 2$　$r_{kv} : 2/2 + 1/2$
$w_{\text{tag}} = w_b = w_{kv} = w_{ko} = 0.2$
$r_{\text{format}} = 0.75$

<think> In the image, there is one person laying down on ... </think>
<answer> [
{"human": $\hat{b}_1^h$, "object": $\hat{b}_1^o$, "object class": $\hat{c}_1^o$, "verb class": $\{\hat{c}_{11}^v, \hat{c}_{12}^v\}$},
{"human": $\hat{b}_2^h$, "object": $\hat{b}_2^o$, "object class": $\hat{c}_2^o$, "verb class": $\{\hat{c}_{21}^v, \hat{c}_{22}^v\}$}
] </answer>

C_o = {"person", "bicycle", "car", "motorcycle", "airplane", "bus", "train", "truck", ...}
C_v = {"adjust", "assemble", "block", "blow", "board", "break", "brush with", "buy", ...}

- HOI-R1 の報酬関数の例：HOI の**書式**・**ラベル**・**IoU** の報酬により、出力の構造的・意味的・幾何的整合性を確保する

$$r_{\text{format}} = w_{\text{tag}} r_{\text{tag}} + \frac{\sum_{i=1}^{\hat{N}_a} \alpha_i [w_b r_{b_i} + w_{ko} r_{ko_i} + w_{kv} r_{kv_i}]}{\max(N_a, \hat{N}_a)}$$

$$r_{\text{IoU}} = \frac{1}{N_a} \sum_{(i,j) \in \mathcal{M}^*} s_{ij}$$

$$s_{ij} = \frac{1}{2}[\text{IoU}(\hat{b}_i^h, b_j^h) + \text{IoU}(\hat{b}_i^o, b_j^o)]$$

$$r_{lo} = \frac{\sum_{i=1}^{\hat{N}_a} \alpha_i \mathbf{1}[\hat{c}_i^o \in C_{gt,o}^{(i-1)}]}{\max(N_a, \hat{N}_a)}$$

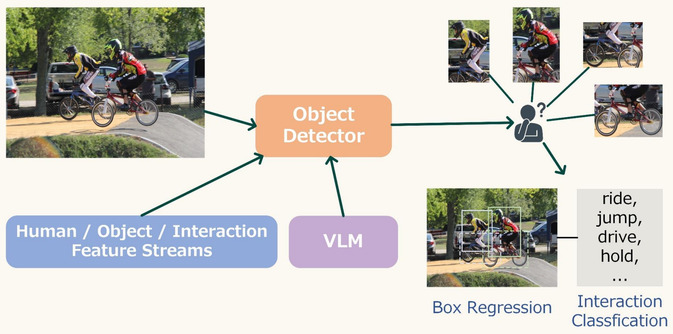
$$r_{lv} = \frac{\sum_{i=1}^{\hat{N}_a} \frac{\alpha_i}{|\{\hat{c}_i^v\}|} \sum_{j=1}^{|\{\hat{c}_i^v\}|} \mathbf{1}[\hat{c}_{ij}^v \in C_{gt,v}^{(i-1)}]}{\max(N_a, \hat{N}_a)}$$

$$r = r_{\text{format}} + r_{lo} + r_{lv} + r_{\text{IoU}}$$

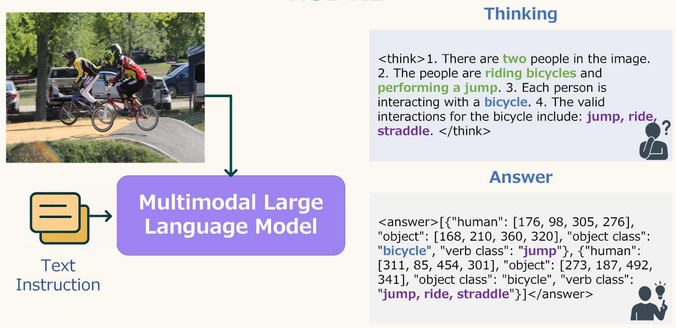
モチベーション

From Detection-Recognition Pipelines To MLLM-Based HOI-R1

Traditional Methods



HOI-R1



従来 HOID 手法と提案手法 HOI-R1 の枠組みの比較

- 従来手法は物体検出器に依存する
- VLM 事前知識を導入するため、複雑なモジュール設計
- HOI-R1 は MLLM による**自然言語で HOI を直接解釈**する

HOI-R1 フレームワーク

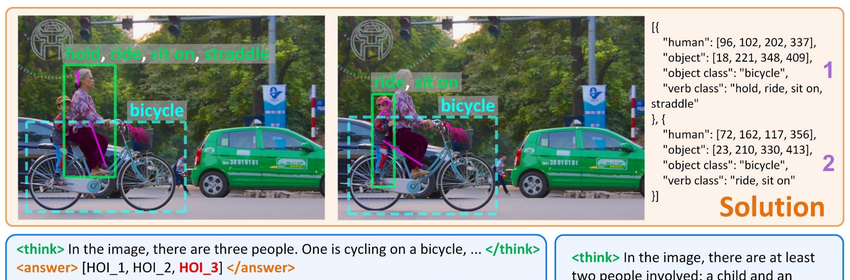
You are an HOI detection model. Follow this exact structure:

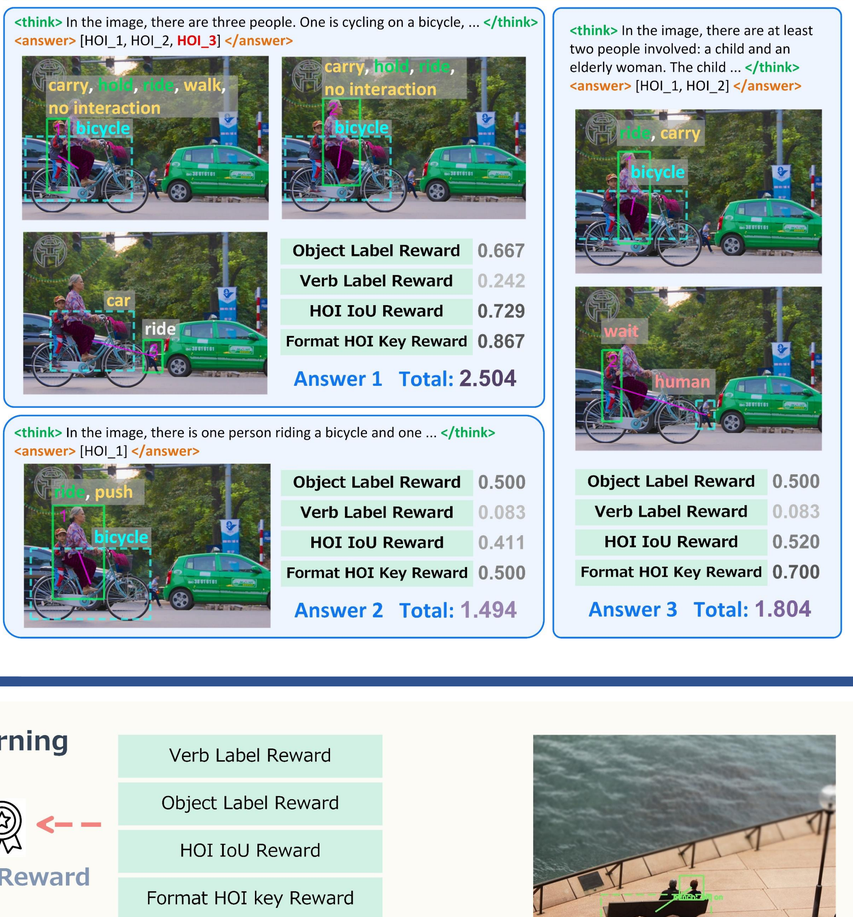
<VALID OBJECT CLASSES>:
<object class>:person, bicycle, car, motorcycle, airplane ...

<VALID INTERACTIONS> (object: allowed_verbs):
<object class>: <verb class>, <verb class>, ⋯
airplane: board, direct, exit, fly, inspect, load, ride, sit on, wash, no interaction
bicycle: carry, hold, inspect, jump, hop on, park, push, repair, ride, sit on, straddle, walk, wash, no interaction
...

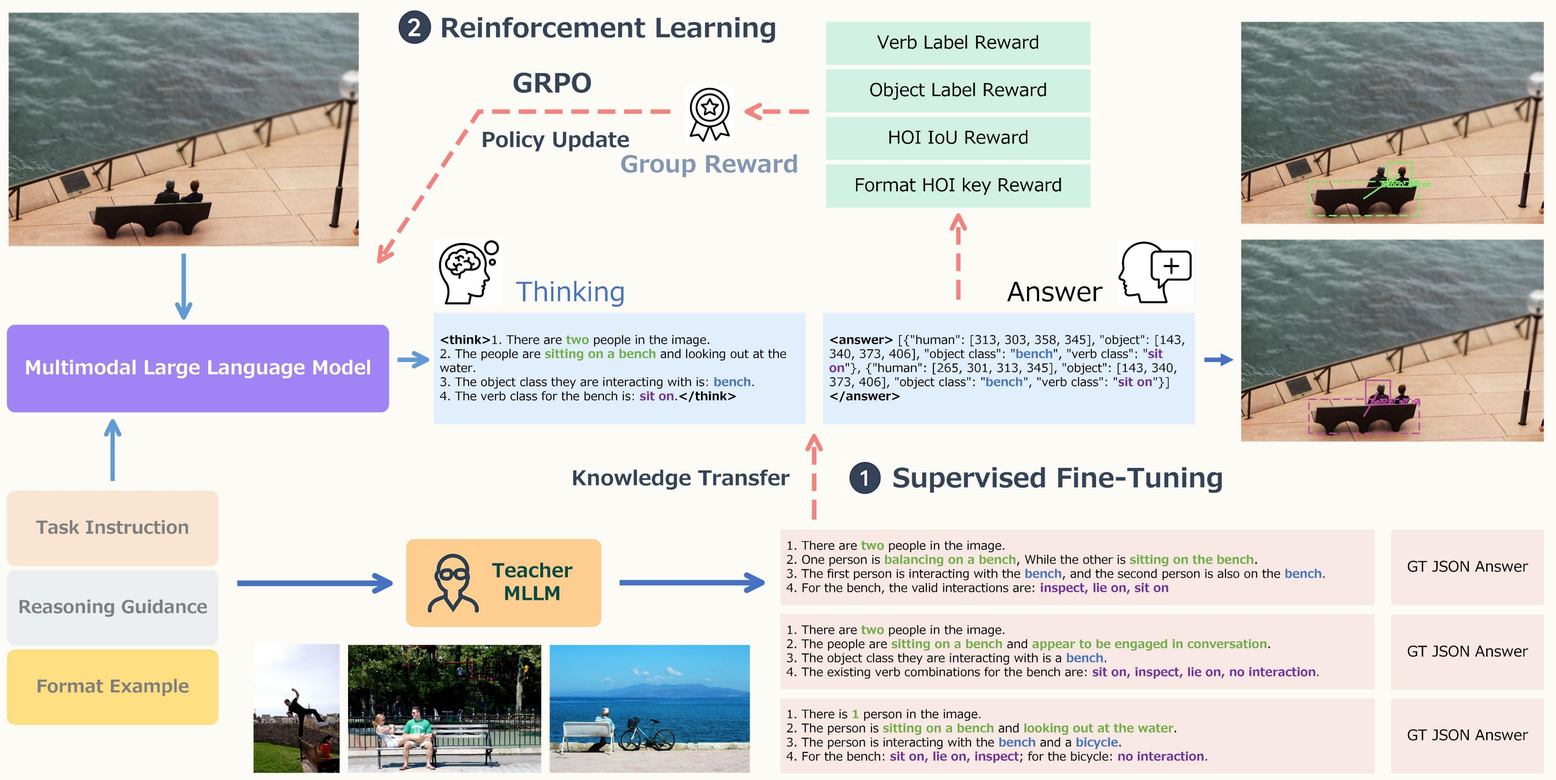
Thinking Process:
1. How many people are in the image?
2. What are those people doing?
3. What <object class> in <VALID OBJECT CLASSES> are each people interacting with?
4. Find existing <verb class> combinations from <VALID INTERACTIONS> for each <object class>.

EXAMPLE OUTPUT FORMAT:
<think>
reasoning process here
</think>
<answer>
[
  {
    "human": [x1, y1, x2, y2],
    "object": [x1, y1, x2, y2],
    "object class": "<object class>",
    "verb class": "<verb class>, <verb class>, ... "
  },
  ...
]
</answer>





HOI-R1 の入力質問テンプレート



- 画像とテキストを入力とし、質問文はタスク指示・推論ガイド・出力書式例から構成される
- 教師 MLLM が生成した推論例を用いて SFT を行う

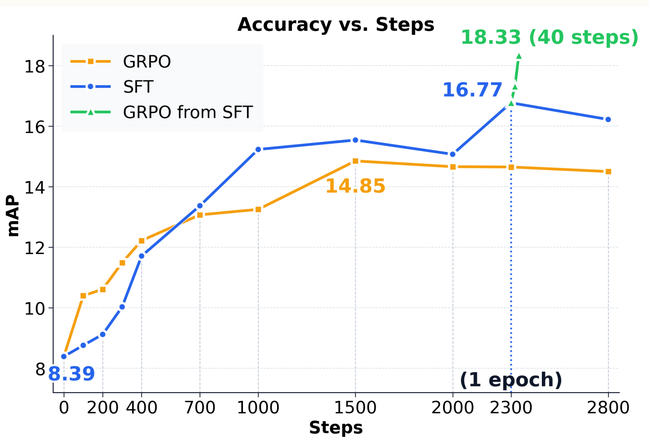
$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,q,r,a)\sim\mathcal{D}} \left[\underbrace{\sum_{t=1}^{T_r} \log \pi_\theta(r_t \mid x, q, r_{<t})}_{\text{reasoning supervision}} + \underbrace{\sum_{t=1}^{T_a} \log \pi_\theta(a_t \mid x, q, r, a_{<t})}_{\text{answer supervision}} \right]$$

- 4 種類の報酬信号による GRPO でポリシーモデルを更新する

$$\mathcal{J}_{\text{GRPO}} = -\mathbb{E}_{(x,q)\sim D, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|x,q)}$$

$$\frac{1}{G} \sum_{i=1}^{G} \left\{ \min[s_1 \cdot \hat{A}_i, s_2 \cdot \hat{A}_i] - \beta \mathbb{D}_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}] \right\}$$

- 1 エポック SFT と 40 ステップ RL のみで **2 倍以上の性能向上**



実験結果

Method	Training	Default			Known Obj.		
		Full	Rare	N-Rare	Full	Rare	N-Rare
— Traditional HOID Methods —							
HO-RCNN[5]	150k	7.81	5.37	8.54	10.41	8.94	10.85
HOTR[10]	100ep	25.10	17.34	27.42	-	-	-
HOI-Trans[20]	150ep	26.61	19.15	28.84	29.13	20.98	31.57
QPIC[18]	150ep	29.07	21.85	31.23	31.68	24.14	33.93
— MLLM Zero-Shot Baselines —							
Qwen2.5-VL-3B	-	8.39	9.60	8.03	8.96	9.83	8.70
Qwen2.5-VL-7B	-	10.46	14.30	9.31	11.01	14.63	9.93
Qwen2.5-VL-32B-AWQ	-	18.12	24.56	16.20	19.90	25.77	18.15
Qwen2.5-VL-72B-AWQ	-	20.71	29.62	18.05	22.93	31.87	20.26
Qwen3-VL-4B	-	16.16	22.31	14.32	17.55	22.87	15.96
Qwen3-VL-8B	-	17.41	21.96	16.05	19.17	23.39	17.91
Qwen3.5-4B	-	17.17	16.08	20.82	19.60	18.53	23.16
— SFT or RL Fine-Tuning —							
Qwen2.5-VL-3B-SFT	400	11.71	10.52	12.07	12.60	10.70	13.17
Qwen2.5-VL-3B-SFT	1ep	16.77	14.20	17.53	18.05	14.45	19.13
Qwen2.5-VL-3B-GRPO	400	12.22	14.56	11.52	13.12	14.84	12.60
Qwen2.5-VL-3B-GRPO	1ep	14.65	15.90	14.28	15.48	16.14	15.28
Qwen3-VL-4B-SFT	1ep	22.46	21.53	22.73	24.04	21.97	24.66
Rex-Omni-3B[9](HOI-SFT)	1ep	25.30	21.24	26.51	26.97	21.79	28.51
Qwen3.5-4B-SFT	1ep	29.36	29.03	**30.47**	31.22	**31.31**	30.92
— HOI-R1 (SFT+GRPO) —							
HOI-R1-Qwen2.5-VL-3B	1ep+40	18.33	16.03	19.02	19.83	16.25	20.90
HOI-R1-Qwen3-VL-4B	1ep+40	23.35	23.58	23.28	25.12	24.11	25.42
HOI-R1-Rex-Omni-3B	1ep+40	25.24	22.57	26.04	27.00	23.21	28.14
HOI-R1-Qwen3.5-4B	1ep+40	**29.79**	**30.39**	29.60	**31.61**	30.67	**31.89**

HICO-DET における比較結果（mAP）