CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

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[1,a)]　熊 培林[1,b)]　柳井 啓司[1,c)]

概要

近年の Human-Object Interaction Detection（HOID）手法は，物体検出器に視覚言語モデル（VLM）の事前知識を結合する複雑な枠組みに大きく依存しており，拡張や応用が困難となっている．一方で，マルチモーダル大規模言語モデル（MLLM）が本来備える推論能力を HOID へ直接適用する潜在能力に期待が寄せられている．本研究では，物体検出器を使わず，MLLM が自然言語推論のみによって HOID を解くフレームワーク **HOI-R1** を提案する．構造化プロンプトによる思考蒸留型の教師あり微調整（SFT）と，構成・物体ラベル・動詞ラベル・IoU の 4 種の報酬を用いた強化学習（GRPO）を組み合わせることで，HICO-DET において Qwen2.5-VL ベースラインの精度を約 2 倍に向上させ，Qwen3-VL および Rex-Omni に対しても一貫した性能改善を確認した．ソースコードは `https://github.com/cjw2021/HOI-R1` で公開している．

1. はじめに

HOID は，入力画像から $\{B_h, B_o,$ **物体カテゴリ**, **動作カテゴリ** $\}$ で表される全ての HOI インスタンスを推定するタスクである．細粒度の関係認識と位置特定を同時に要求する難問であり，現在の主流は物体検出器に基づくパイプラインである [10], [18], [20]．近年の高性能手法 [4], [6], [14] は CLIP[15]，BLIP[12] や BLIP2[12] に代表される VLM の事前知識をデコーダに導入してロングテール分布に対処しているが，検出器とインタラクションヘッドの結合や追加モジュール，目的関数の導入により全体が複雑化し，拡張や再利用が難しいという欠点を抱えている．

一方，Qwen2.5-VL[3] や Qwen3-VL[2]，Rex-Omni[9] に代表される現代の MLLM は，細粒度の視覚理解，位置特定，マルチモーダル推論において強力な能力を示している．Qwen2.5-VL は動的解像度視覚エンコードと動画モデリングにより複雑な画像と長時間映像に対する構造的理解を実現し，Rex-Omni は物体検出を離散座標トークン予測として再定式化し，強力な空間位置事前知識を備える．さらに DeepSeek-R1[7] は，教師あり微調整を介さず強化学習（RL）手法，GRPO（Group Relative Policy Optimization）が高度な推論挙動を誘導することを示し，Vision-R1[8]，Reason-RFT[19]，VLM-R1[17] などの後続研究が RL による視覚推論能力の整合化の有効性を裏付けている．しかし，標準的な評価指標で HOI インスタンスを予測する MLLM の能力は，これまで検討されていない．

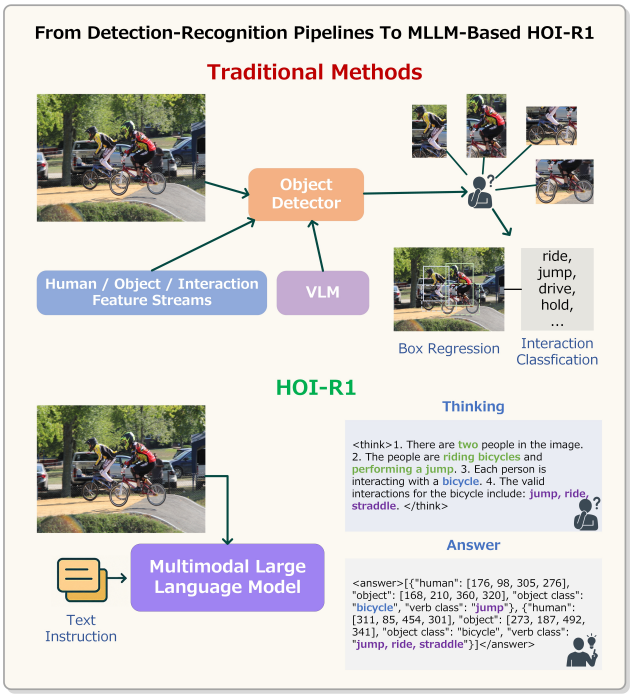


図 1: 従来 HOID 手法と提案手法 HOI-R1 の枠組みの比較．従来手法は物体検出器と HOI 埋め込み抽出に依存するが，HOI-R1 は MLLM による自然言語推論で相互作用を直接解釈する．

図 1 に示すように，本研究では検出器を自然言語推論に置き換える抜本的な転換を行う HOI-R1 を提案する．純粋に言語推論のみで HOID を解くには，複数の境界ボックスを同時に予測し，物体と相互作用を正しく対応付け，関係を正確に認識する必要があり，これら全てを構造化された出力として生成しなければならない．HOI-R1 は，体系的なプロンプト設計に基づく Chain-of-Thought（CoT）と思考蒸留型 SFT によって HOI 知識を導入し，さらに HOID

1 電気通信大学
a) chen-j@mm.inf.uec.ac.jp
b) xiong-p@mm.inf.uec.ac.jp
c) yanai@mm.inf.uec.ac.jp

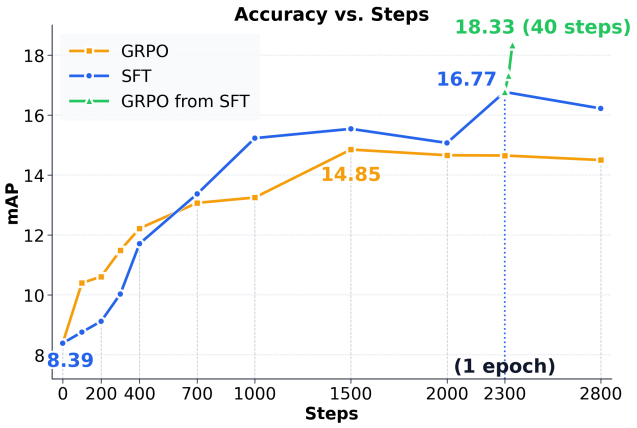


図 2: HICO-DET 上での HOI-R1（Qwen2.5-VL-3B）の学習曲線．1 エポック SFT と 40 ステップ RL のみで 2 倍以上の性能向上が得られる．

評価指標に整合した 4 種類の報酬関数による RL で方策モデルを整合化する．図 2 に示すように，HOI-R1 はわずか 1 エポックの SFT と 40 ステップの RL のみでベースラインの 2 倍を超える mAP を達成する．本研究の貢献を以下にまとめる．

- 物体検出器を使わず，自然言語のみで HOID を解く初の MLLM フレームワーク HOI-R1 を提案する．
- 思考蒸留型 SFT と，HOID 固有の 4 種類の報酬関数に基づく強化学習を組み合わせる学習方法を設計する．
- Qwen2.5-VL，Qwen3-VL，Rex-Omni といったオープンソース MLLM 上で一貫した性能改善を達成し，MLLM ベースの HOID に新たな基準を提示する．

2. 関連研究

視覚推論における RL 強化 MLLM． Vision-R1[8] はコールドスタート初期化と RL を組合せた視覚推論 MLLM を構築し，Reason-RFT[19] は CoT を活性化する SFT と GRPO[16] を二段階で組み合わせ，視覚推論の汎化性能を向上させる．VLM-R1[17] は参照表現理解やオープン語彙物体検出において，正解を用いた安定な報酬計算と「物体検出における aha moment」現象を報告している．本研究はこれらの知見を，より複雑な HOID タスクへ拡張し，入力指示・学習方法・報酬関数を新たに設計する点で従来研究と異なる．

3. 提案手法: HOI-R1

3.1 言語ベースの HOID 予測

従来手法のように境界ボックス回帰と分類ヘッドを設けるのではなく，本研究では MLLM の構造を変更せず，全ての HOI インスタンスを自然言語形式で直接出力する新たなパラダイムを提案する．モデルの HOID 能力を引き出すために，図 4 に示す質問テンプレートを以下の 3 要素から構成した．

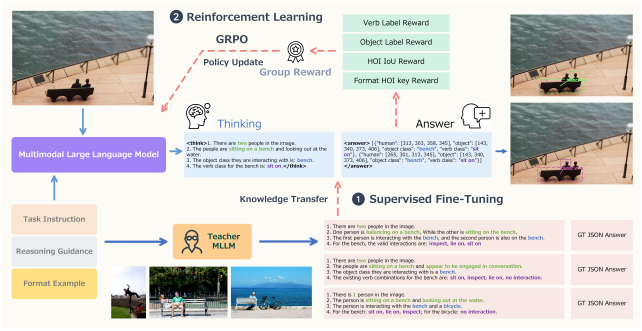


図 3: HOI-R1 の全体像．画像とテキストを入力とし，質問文はタスク指示・推論ガイド・出力書式例から構成される．教師 MLLM が生成した推論列を用いて SFT を行い，その後，4 種類の報酬信号による GRPO でポリシーモデルを更新する．

(1) タスク指示． モデルの役割を「あなたは HOI 検出モデル」と定義し，物体・相互作用クラスの全語彙<VALID OBJECT CLASSES>と<VALID INTERACTIONS>を明示することで，幻覚的出力を抑え標準ベンチマークと整合する出力空間を強制する．
(2) 推論ガイド． 人物発見 → 動作分析 → 相互作用判定という人間の認知過程に倣った逐次推論ステップを与え，HOI の構成的性質を体系的に扱えるようにする．
(3) 出力形式例． <think>タグ内の思考過程と<answer>タグ内の JSON 形式予測列から成る出力を提示する．近年の MLLM は座標トークンによる境界ボックス表現に対応しているため，空間座標を直接言語出力に含めることができる．

3.2 思考蒸留による SFT

学習対象である生徒 MLLM にタスク固有の知識を導入するため，教師-生徒蒸留型の SFT 段階を設計する．教師には GPT4o-mini[1] を採用し，HICO-DET[5] 訓練画像と推論ガイドプロンプトを与え，<think>タグに包まれた段階的推論列を生成させる．生徒モデルはこの推論列と正解 HOI 三項組の双方を教師信号として学習する．画像 x，質問 q，教師生成推論 r，正解応答 a に対する自己回帰的学習目的は，

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

となる．ここで π_θ はパラメータ θ を持つ生徒 MLLM である．本 SFT 段階は，後続する RL 段階のための強力な基礎を確立する．

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>

図 4: HOI-R1 の入力質問テンプレート．タスク指示，推論ガイド，出力書式例の 3 要素から構成される．

3.3 HOID 強化学習

SFT で基礎能力を確立した後，HOID 固有の構造的・意味的・幾何的整合性を強化するために RL を適用する．本研究では GRPO[16] を採用し，入力 (x, q) に対し旧ポリシー $\pi_{\theta_{old}}$ から G 個の出力 $\{o_i\}_{i=1}^{G}$ をグループ化し，

$$\mathcal{J}_{\text{GRPO}} = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left\{\min[s_1\hat{A}_i, s_2\hat{A}_i] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta\|\pi_{\text{ref}}]\right\}\right] \tag{2}$$

を最適化する．s_1, s_2 はクリッピング項，$\hat{A}_i = (r_i - \text{mean}(\{r_k\}))/\text{std}(\{r_k\})$ はグループ内標準化された優位度，r_i は i 番目の出力に対する総報酬である．HOI インスタンスは複数の要素（人物・物体ボックス，物体クラス，動詞クラス）から成るため，図 5 に示す要素別の報酬を設計する．

HOI キー形式報酬． テキスト出力の形式正確性を強化する報酬である．`<answer>`タグが欠落した場合，全報酬を 0 とする．総形式報酬は，`<think>`タグ報酬 r_{tag}，第 i 番目 HOI インスタンステキスト $\hat{a}_i$ に対する人物・物体ボックス形式報酬 r_{b_i}，物体ラベルキー報酬 r_{ko_i}，動詞ラベルキー報酬 r_{kv_i} から構成される．例えばボックス形式報酬は，

$$r_{b_i} = \mathbf{1}\left[\{\texttt{"human"}, \texttt{"object"}\} \subseteq \text{keys}(\hat{a}_i) \wedge \tilde{B}_i \text{条件}\right] \tag{3}$$

ここで $\tilde{B}_i$ は既出予測ボックスと IoU≤ 0.5 である条件で，重複ボックスによる報酬ハッキングを防ぐ．重複キーに対するペナルティ $\alpha_i = N_k/(N_k + |\hat{N}_{k_i} - N_k|)$（$N_k = 4$）を導入し，全要素を重み付け平均して，

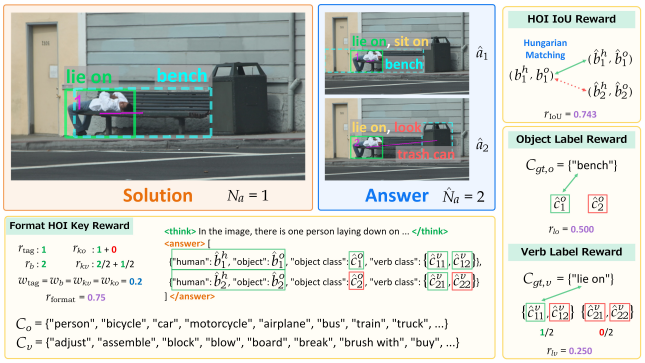


図 5: HOI-R1 の報酬関数の例．書式・ラベル・IoU の 3 種の報酬により，出力の構造的・意味的・幾何的整合性を確保する．

$$r_{\text{format}} = w_{\text{tag}} r_{\text{tag}} + \frac{\sum_{i=1}^{\hat{N}_a} \alpha_i [w_b r_{b_i} + w_{ko} r_{ko_i} + w_{kv} r_{kv_i}]}{\max(N_a, \hat{N}_a)} \tag{4}$$

と定める．

物体・動詞ラベル報酬． 形式とは独立に，正解集合 C_{gt} と予測ラベルを順次照合する「マッチ消去（drop-on-match）」方式で物体ラベル報酬 r_{lo} および動詞ラベル報酬 r_{lv} を計算し，正解に現れる HOI をより正確に予測することを促進する．

HOI IoU 報酬． 予測 HOI 対と正解 HOI 対の空間整合のため，Hungarian アルゴリズム [11] による一対一マッチングを行う．コストは

$$C_{ij} = 1 - s_{ij}, \quad s_{ij} = \tfrac{1}{2}[\text{IoU}(\hat{b}_i^h, b_j^h) + \text{IoU}(\hat{b}_i^o, b_j^o)] \tag{5}$$

で，最適マッチング $\mathcal{M}^*$ に対して $r_{\text{IoU}} = \frac{1}{N_a}\sum_{(i,j)\in\mathcal{M}^*} s_{ij}$ と定義する．最後に，総合報酬は $r = r_{\text{format}} + r_{lo} + r_{lv} + r_{\text{IoU}}$ である．

4. 実験

データセットと評価指標． HICO-DET[5] を用いる．訓練画像 38,118 枚，テスト画像 9,658 枚，80 物体クラスと 117 動詞クラスから成る 600 HOI カテゴリで構成される．評価指標は mAP で，IoU 閾値 0.5 を用い，物体・動詞ラベルの一致を要求する．カテゴリは出現頻度に応じて *Full*/*Rare*/*Non-Rare* に，評価設定は *Default*/*Known Object* に分けられる．

実装の詳細． 主な実験は Qwen2.5-VL-3B-Instruct[3] で行い，Qwen3-VL-4B-Instruct[2] と Rex-Omni-3B[9] でも有効性を検証した．SFT 段階では GPT4o-mini[1] で思考過程を生成し，HICO-DET の正解アノテーションを出力形式に変換する．AdamW[13]（学習率 1×10^{-6}，cosine スケジューラ）を用いて，全モデルをバッチサイズ 8 で 1 エポック学習する．RL 段階ではバッチサイズ 16，グループサイズ $G = 4$，40 ステップで学習し，形式報酬の重みは全て 0.2 とする．

表 1: HICO-DET における比較結果（mAP）.

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**

表 2: 報酬関数のアブレーション（Qwen2.5-VL-3B+GRPO 1000 ステップ）.

Method	Default			Known Obj.		
	Full	Rare	N-Rare	Full	Rare	N-Rare
HOI-R1	**13.25**	**15.22**	**12.66**	**14.18**	**15.60**	**13.75**
w/o label reward	12.49	14.78	11.80	13.54	15.26	13.02
w/o IoU reward	9.63	11.12	9.19	10.29	11.38	9.96

ベースラインおよび従来 HOID 手法との比較． 表 1 より，本研究で設計した HOID プロンプトを用いるだけで，未学習の Qwen2.5-VL-3B と Qwen3-VL-4B がそれぞれ 8.39 mAP と 16.16 mAP を達成し，HO-RCNN[5] を上回ることがわかる．特に Rare カテゴリの精度が Full/Non-Rare より高い傾向は，希少な HOI を含む画像はインスタンス数が少なく，MLLM が強い推論能力で焦点を当てやすいためと考えられる．32B/72B モデルは未学習でも QPIC[18] を上回っており，大規模 MLLM の強力な事前知識を示している．次に，3B モデルに対する追加学習結果から，初期段階では RL が SFT より効率的に精度を伸ばすことがわかる．本研究の SFT は教師モデルから推論知識を蒸留するため，1 エポック学習でベースラインの 2 倍となる 16.77 mAP に到達する．SFT と GRPO を組み合わせた HOI-R1 により，Rex-Omni-3B は Known Object/Full で 27.00 mAP に達し，Qwen2.5-VL-72B-AWQ を 17.7%上回る．Qwen3-VL-4B は Rare カテゴリで 23.58 mAP に達し，QPIC[18] を 7.9%上回る．さらに，最新の Qwen3.5-4B を生徒モデルとした場合，未学習でも 17.17 mAP を達成し，1 エポックの SFT 後は Default/Full で 29.36 mAP，Known Object/Rare で 31.31 mAP と従来の QPIC を含む全ての比較手法を上回る．これは，より新しい MLLM が HOI 関連の事前知識を持っていることを示している．特に，数百エポックを要する従来 HOID 手法と比べて大幅に少ない学習量で収束できる．MLLM が既に保持している潜在的な人物物体相互作用知識を，構造化出力として表出させるだけで十分であることを意味している．

アブレーション研究． 表 2 に報酬関数のドロップ実験を示す．ラベル報酬の除去で *Default/Full* が 0.76 mAP 低下し，正解 HOI に沿った予測の促進に寄与していることがわかる．IoU 報酬の除去では 3.62 mAP の大幅な低下が見られ，HOID タスクにおける正確な位置特定の重要性が示された．これは，形式とラベルが正しくとも境界ボックスが評価 IoU 閾値 0.5 を下回れば偽陽性として扱われるためであり，MLLM による HOID では空間整合の能力が最大のボトルネックであることを示している．また，プロンプト構成のアブレーションからは，思考過程の除去は Rare カテゴリ性能を最も損ない，CoT が希少相互作用の推論に貢献すること，さらにタスク記述の除去は全カテゴリで性能を大きく損なうことが確認された．特にタスク指示と語彙制約の除去後は Default/Full が 8.39 から 3.06 まで低下しており，制約された出力空間の明示が MLLM による HOID の動作スコアを表す必須要素であることを意味する．

HOI-R1-Qwen3.5-4B が RL 段階を経ず SFT のみで QPIC[18] を超える 29.36 mAP に到達した結果は，MLLM による HOID 性能の上限は，言語モデルに内在する常識，物体アフォーダンス，および行動理解の豊かさによって大きく左右されるため，これらの知識を視覚的な人・物体・相互作用の手がかりと整合的に活用することが重要である．本研究の貢献は，このような汎用知識を構造化プロンプトと軽量な報酬整合化のみで HOID の標準評価に橋渡しした点にあり，将来の高性能 MLLM にも自然に適用可能なフレームワークを提供する．一方で，HOI-R1 はテキスト出力に対する二値スコア（出現/非出現）のみを用いており，連続的な確信度に基づく mAP 計算とは本質的に異なる困難を抱えるため，今後はランキング指標との整合や，より大規模なモデルでの検証が課題となる．

5. むすび

本研究では，物体検出器を排し，純粋に言語推論によって HOID を解く初の MLLM フレームワーク HOI-R1 を提案した．思考蒸留による SFT と HOI 固有の 4 種類の報酬関数による GRPO を組み合わせた軽量な学習レシピにより，複数の汎用 MLLM において一貫した大幅な性能改善を達成し，MLLM が HOI タスクに対して持つ潜在能力を実証した．検出器中心のアーキテクチャを構築する従来路線とは異なり，汎用推論モデルを構造化プロンプトと整合化学習により HOI タスクへ適応させる新たな方向性を提示する．今後は，動画や三人称視点を含む多様な HOI ベ

ンチマークへの展開，連続的な確信度推定との統合，さらにより大規模な MLLM における汎化性能の検証を通じて，本フレームワークの一般性と実用性を高めていきたい．

参考文献

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al.: Gpt-4 technical report, *arXiv preprint arXiv:2303.08774* (2023).

[2] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C. et al.: Qwen3-VL Technical Report, *arXiv preprint arXiv:2511.21631* (2025).

[3] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J. et al.: Qwen2.5-vl technical report, *arXiv preprint arXiv:2502.13923* (2025).

[4] Cao, Y., Tang, Q., Su, X., Chen, S., You, S., Lu, X. and Xu, C.: Detecting Any Human-Object Interaction Relationship: Universal HOI Detector with Spatial Prompt Learning on Foundation Models, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 739–751 (2023).

[5] Chao, Y.-W., Liu, Y., Liu, X., Zeng, H. and Deng, J.: Learning to Detect Human-Object Interactions, *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 381–389 (2018).

[6] Chen, J., Wang, Y. and Yanai, K.: Focusing on what to Decode and what to Train: SOV Decoding with Specific Target Guided DeNoising and Vision Language Advisor, *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 9416–9425 (2025).

[7] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, *arXiv preprint arXiv:2501.12948* (2025).

[8] Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y. and Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models, *arXiv preprint arXiv:2503.06749* (2025).

[9] Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J. and Zhang, L.: Detect anything via next point prediction, *arXiv preprint arXiv:2510.12798* (2025).

[10] Kim, B., Lee, J., Kang, J., Kim, E.-S. and Kim, H. J.: HOTR: End-to-end human-object interaction detection with transformers, *Proc. of IEEE Computer Vision and Pattern Recognition* (2021).

[11] Kuhn, H. W.: The Hungarian method for the assignment problem, *Naval Res. Logist. Quart*, pp. 83–97 (1955).

[12] Li, J., Li, D., Savarese, S. and Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *Proc. of the International Conference on Machine Learning*, pp. 19730–19742 (2023).

[13] Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, *Proc. of International Conference on Learning Representations* (2018).

[14] Ning, S., Qiu, L., Liu, Y. and He, X.: HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 23507–23517 (2023).

[15] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning Transferable Visual Models From Natural Language Supervision, *Proc. of the International Conference on Machine Learning*, pp. 8748–8763 (2021).

[16] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y. et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models, *arXiv preprint arXiv:2402.03300* (2024).

[17] Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q. et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model, *arXiv preprint arXiv:2504.07615* (2025).

[18] Tamura, M., Ohashi, H. and Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 10405–10414 (2021).

[19] Tan, H., Ji, Y., Hao, X., Lin, M., Wang, P., Wang, Z. and Zhang, S.: Reason-rft: Reinforcement fine-tuning for visual reasoning, *arXiv preprint arXiv:2503.20752* (2025).

[20] Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y. et al.: End-to-end human object interaction detection with hoi transformer, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 11820–11829 (2021).