生活支援ロボットにおける マルチモーダル検索に基づく 移動マニピュレーション

慶應義塾大学 理工学研究科 杉浦孔明研究室
是方諒介

Kaneda, K., Nagashima, S., Korekata, R., Kambara, M., Sugiura, K.
"Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine." IEEE RA-L presented at IEEE/RSJ IROS 2024.

背景：言語指示に基づく実世界検索は有用

ユースケース

1. **家庭内**でロボットに物体移動を依頼
   例：「洗面所からタオルを持ってきて」

2. **公共空間**で物体を検索
   例：ショッピングモール，展示会場

https://www.toyota.com/usa/toyota-effect/romy-robot.html

デモ：Dubai Mallにおける物体検索 (2025/10)

- “I’m looking for Christmas ornaments”

- “Are there any scary skull toys?”

問題設定：Learning-To-Rank Physical Objects

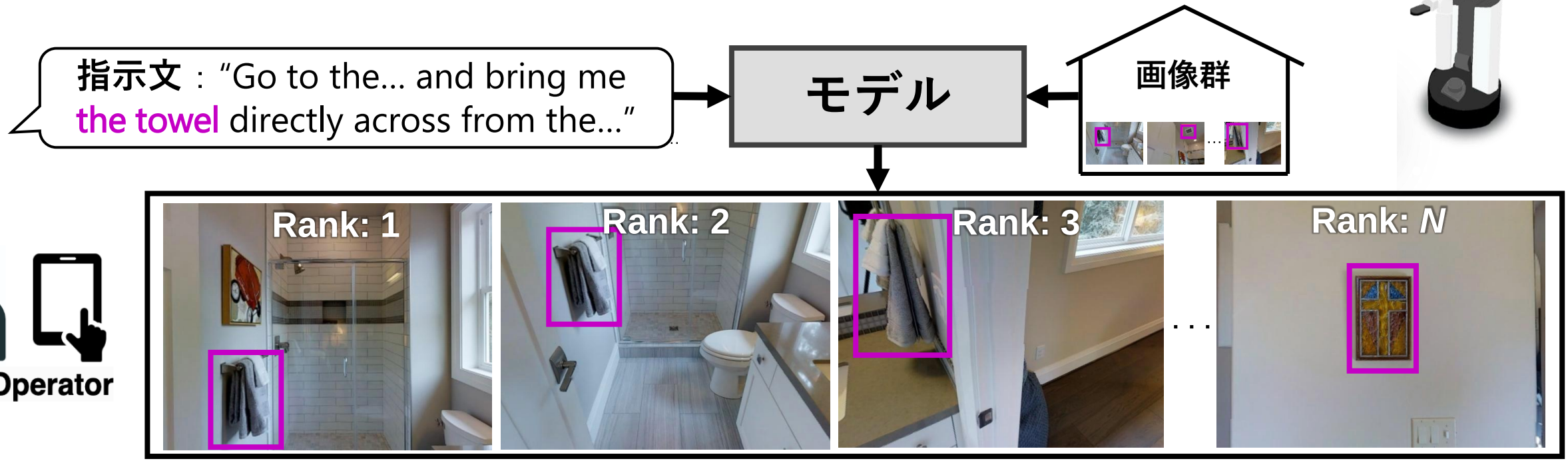
入力

- 参照表現を含む指示文
- 収集済みの画像群

出力

- 対象物体候補をランク付けした画像群

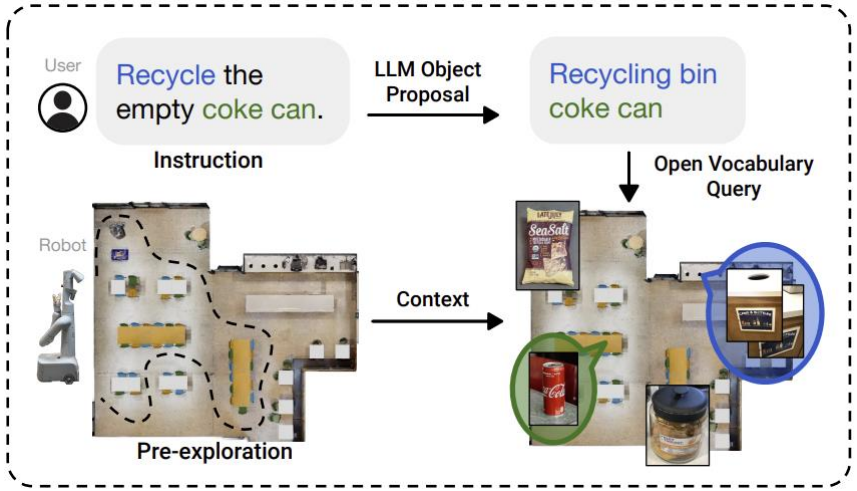
⇨ **適切な画像が上位に表示**される
ことが望ましい



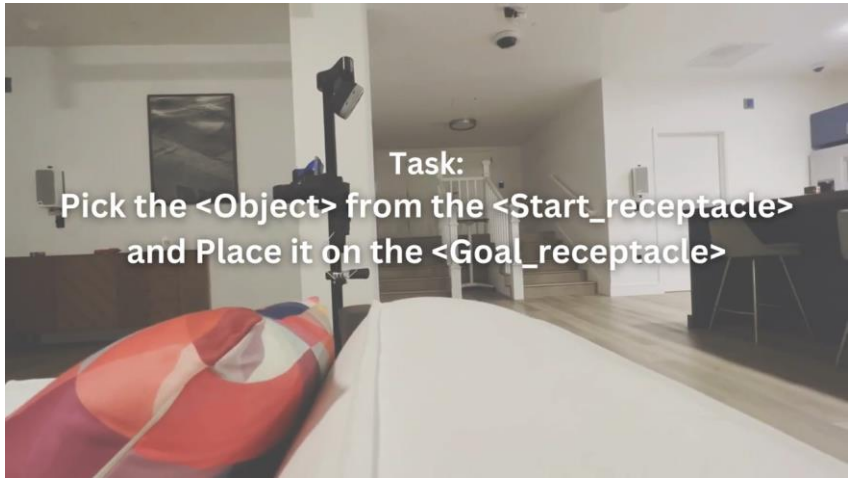
関連研究：検索設定の手法は限定的

NLMap [Chen+, ICRA23]	収集済み環境画像の検索に基づくmobile manipulation ☹ **小規模**環境のみが焦点
RREx-BoT [Sigurdsson+, IROS23]	事前探索に基づくVision-and-Language Navigation ☹ 物体操作タスクを扱うことが難しい
OVMM [Yenamandra+, CoRL23]	Open-Vocabulary Mobile Manipulationタスク ☹ 最先端手法でも成功率10%程度

NLMap



OVMM



課題：複雑な参照表現を含む指示文理解

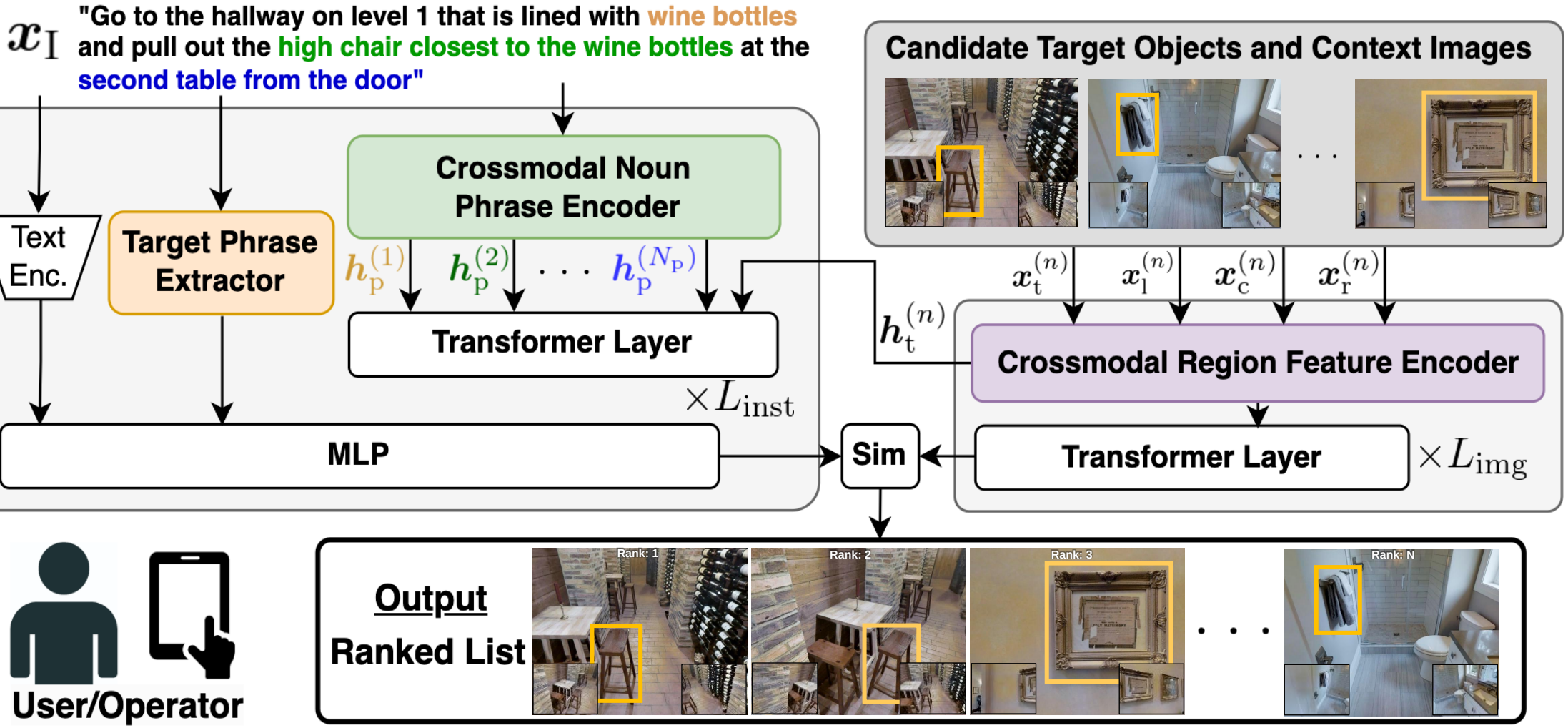
■ 例：“Please polish the black round table with a vase underneath a black chandelier in the dining room”

Ground Truth

CLIP [Radford+, ICML21]

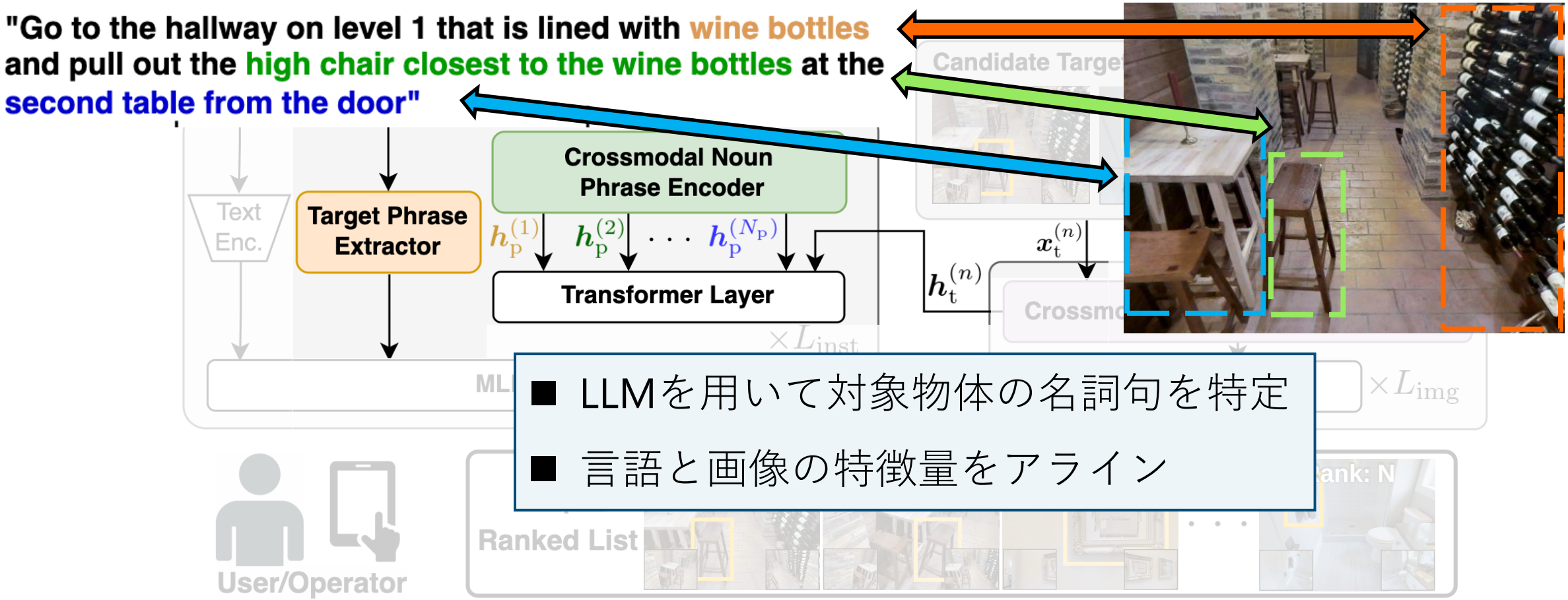
平均文長：**18.78 語**
> G-Ref [Mao+, CVPR16]：8.4 語

提案手法：MultiRankIt
ランキング学習に基づき対象物体を特定



提案手法 (1/2)：
Target Phrase Extractor & Crossmodal Noun Phrase Encoder

■ 指示文から抽出された**参照表現を含む句と物体領域との関係**をモデル化

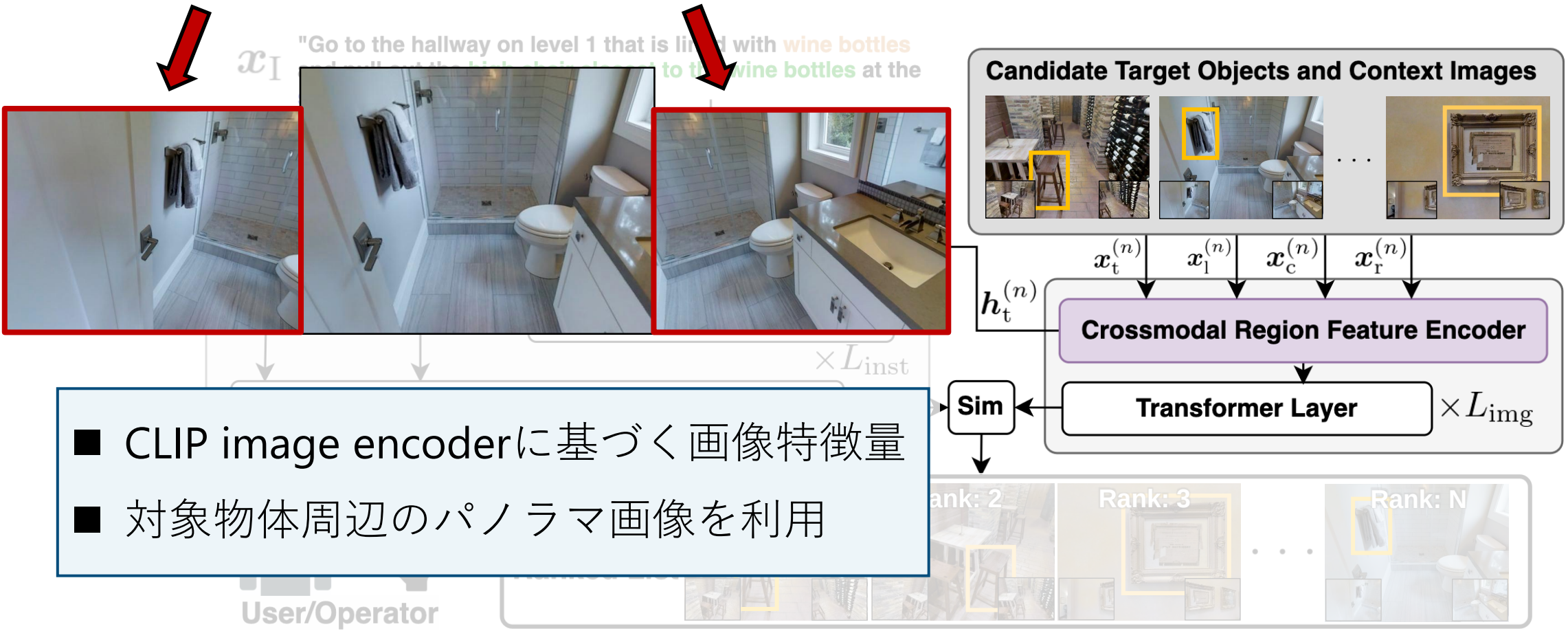


■ LLMを用いて対象物体の名詞句を特定

■ 言語と画像の特徴量をアライン

提案手法 (2/2)：Crossmodal Region Feature Encoder

- **画角外**の視覚的コンテキストと物体領域との関係をモデル化



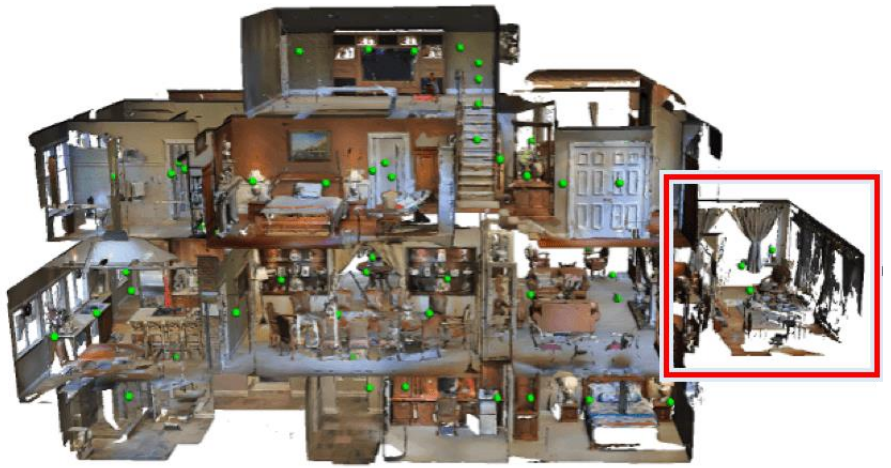
- CLIP image encoderに基づく画像特徴量
- 対象物体周辺のパノラマ画像を利用

実験設定：
指示文・実画像を含む大規模屋内環境データセットを構築

- REVERIE [Qi+, CVPR20], MP3D [Chang+, 3DV17] に準拠

環境数	**58**
物体数	**4,352**
語彙サイズ	**53,118**
指示文数	5,501
平均文長	18.78

https://yuankaiqi.github.io/REVERIE_Challenge/static/img/demo.gif



建物規模の屋内環境

定量的結果：
画像検索における標準的な評価指標で既存手法を上回った

[%]	MRR ↑	Recall@5 ↑	Recall@10 ↑	Recall@20 ↑
CLIP (ext.) [Radford+, ICML21]	41.5±0.9	45.3±1.7	63.8±2.5	80.8±2.0
Ours	50.1±0.8	52.2±1.4	69.8±1.5	83.8±0.6
Ours (ext.)	56.3±1.3	58.7±1.1	77.7±1.1	90.0±0.5

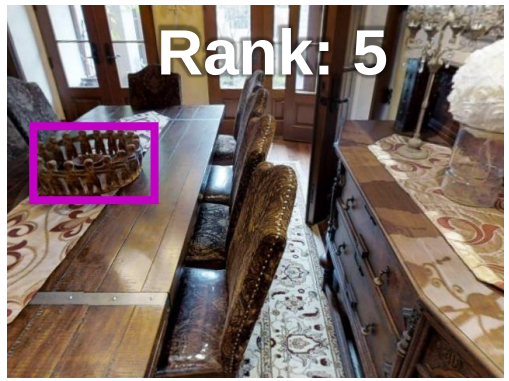
+14.8 +13.4 +13.9 +9.2

- 評価指標
  1. Mean Reciprocal Rank (MRR)
  2. Recall@K (K=1,5,10,20)

定性的結果：複雑な参照表現を理解

"Go to the hallway on level 1 that is lined with wine bottles and pull out the high chair closest to the wine bottles at the second table from the door"

Ground Truth

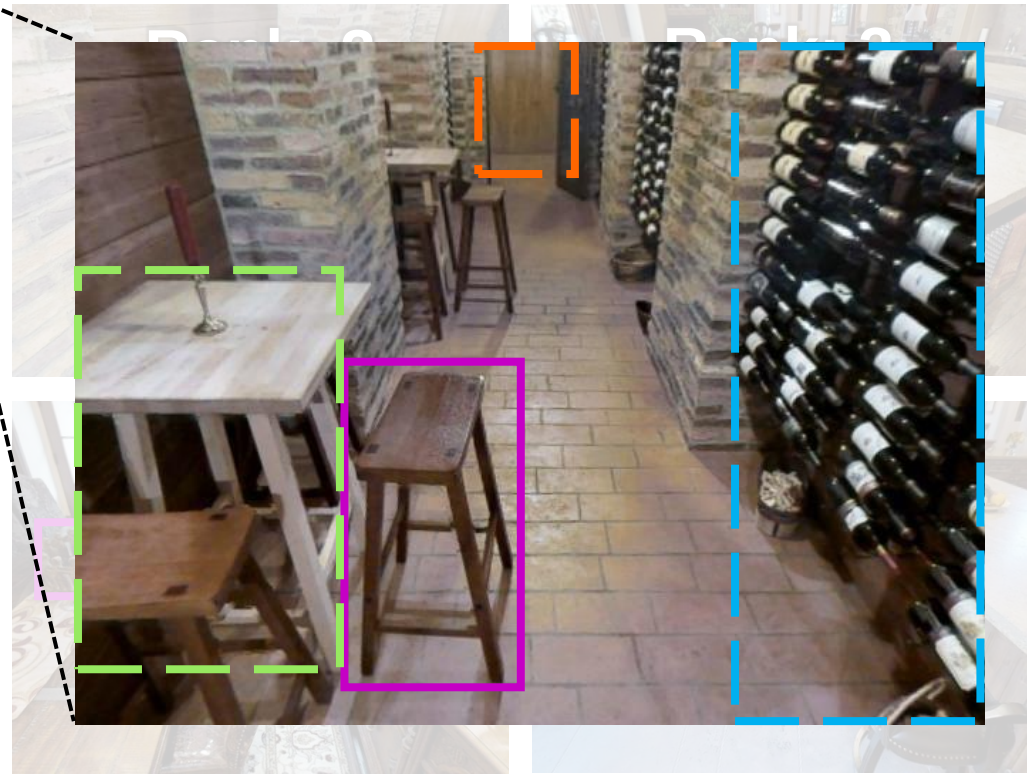


...

定性的結果：複雑な参照表現を理解

"Go to the hallway on level 1 that is lined with wine bottles and pull out the high chair closest to the wine bottles at the second table from the door"

Ground Truth



実験設定（実機）：ゼロショット転移条件

- 環境：WRS 2020 Partner Robot Challengeの**標準環境**に準拠
- 実機：HSR by Toyota
- 物体：YCB
- 評価指標：**SR** [%]

実験手順（実機）：
実世界検索に基づくOpen-Vocabulary Mobile Manipulation

Step 2: 指示文入力

Instruction:
“Could you bring me a green cup?”

Step 3: 画像選択

定量的結果（実機）：既存手法を成功率において上回った

☺ **ゼロショット転移**条件において**成功率80%**を達成

Method	SR ↑ [%]
CLIP (ext.) [Radford+, ICML21]	60 (12/20)
NLMap (reprod.) [Chen+, ICRA23]	70 (14/20)
Ours (ext.)	80 (16/20)

+10

まとめ

背景

- 言語指示に基づく実世界検索の応用性
  （例：**生活支援ロボット**）

提案：MultiRankIt

- **ランキング学習**に基づき対象物体を特定

結果

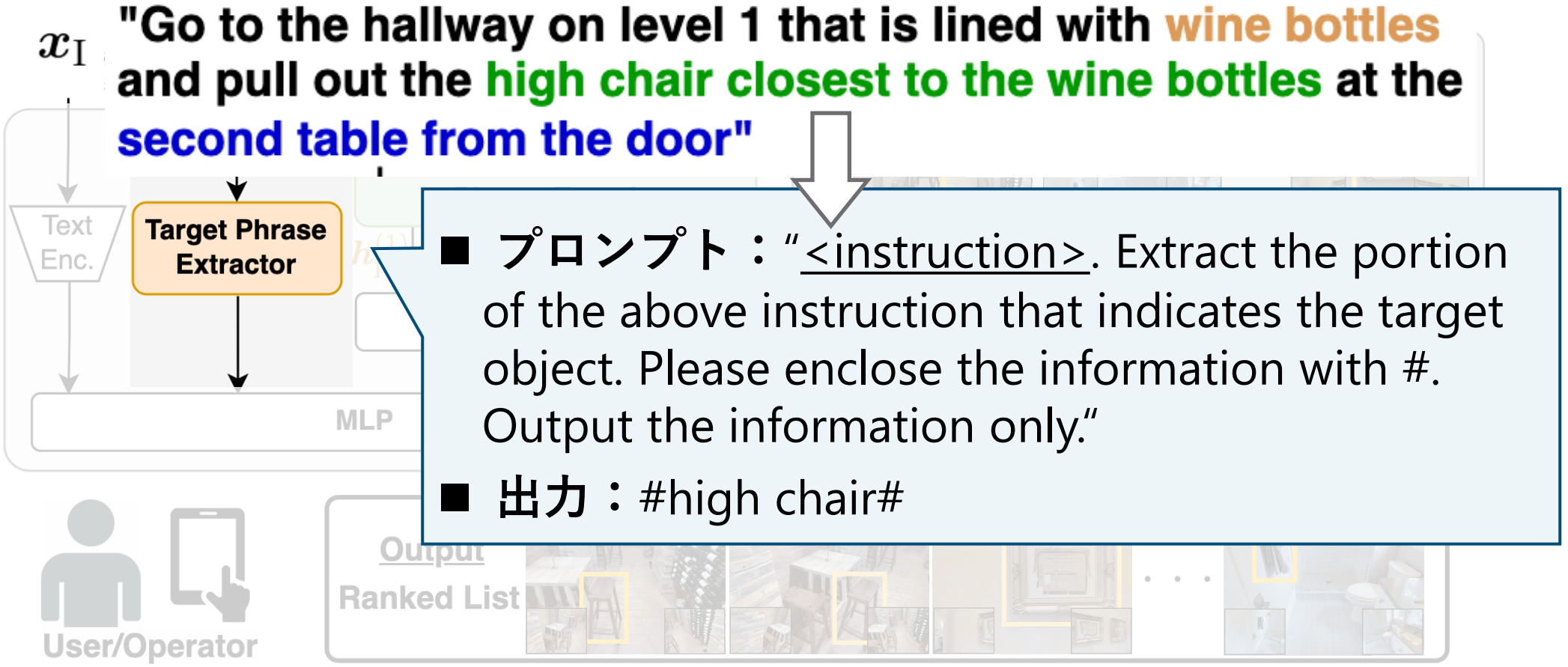
- 標準的な家庭環境で**成功率80%**を達成

Personal Website

Appendix

提案手法：Target Phrase Extractor

- LLMを用いて複雑な参照表現を含む指示文から**対象物体の名詞句**を特定



- **プロンプト：**"<instruction>. Extract the portion of the above instruction that indicates the target object. Please enclose the information with #. Output the information only."
- **出力：**#high chair#

Ablation Studies：CRFEの導入が最も性能向上に寄与

[%]	MRR ↑	Recall@1 ↑	Recall@5 ↑	Recall@10 ↑
w/o CRFE	37.3±1.5	12.1±0.5	39.6±1.4	56.1±1.1
w/o CNPE	42.6±0.4	14.6±0.4	45.3±0.5	66.1±1.7
Ours (ext.)	50.1±0.8	18.3±1.0	52.2±1.4	69.8±1.5

+12.8　+6.2　+11.6　+13.7

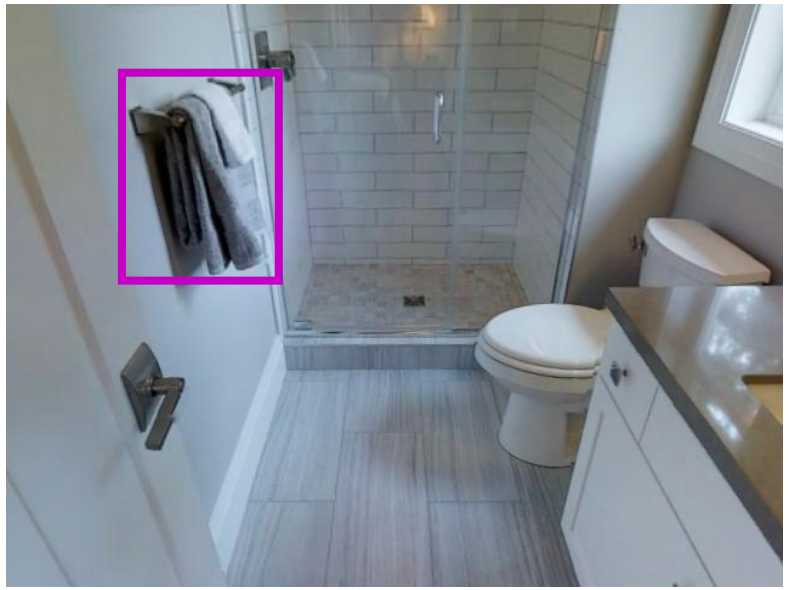
☺ すべての新規モジュールが有効

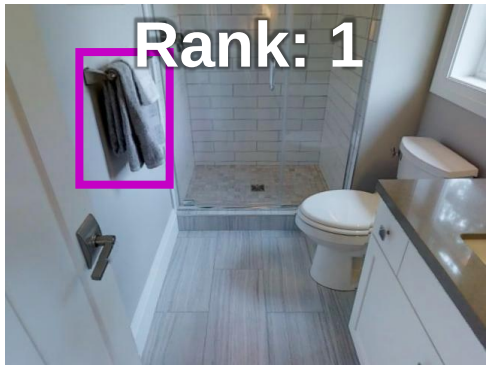
→ **画角外**の視覚的コンテキストを扱うCRFEの導入が最も性能向上に寄与

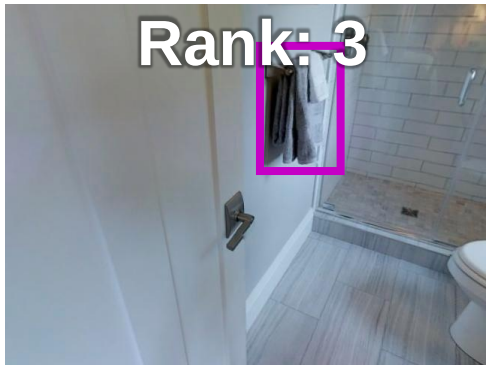
定性的結果：画角外の視覚的コンテキストを考慮

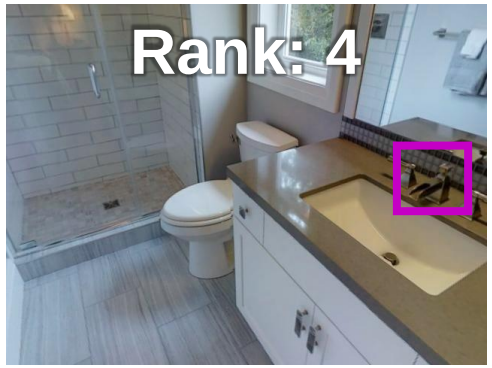
"Go to the bathroom with a picture of a wagon and bring me the towel directly across from the sink"

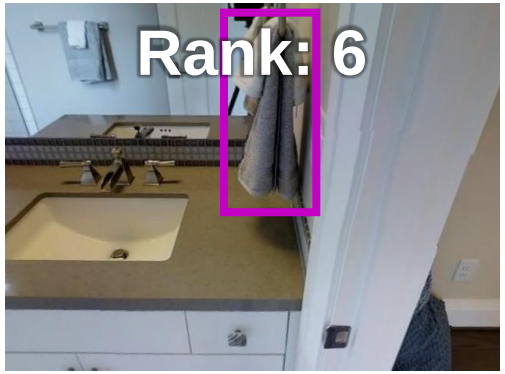
Ground Truth











...

定性的結果：画角外の視覚的コンテキストを考慮

"Go to the bathroom with a picture of a wagon and bring me the towel directly across from the sink"

エラー分析：
現状のボトルネックは参照表現理解に関する誤り

- 最もMRRが低かった20サンプルについて分析
  - ☹ 参照表現理解に関する誤りが最多
  - ⇨ 今後：**地図情報**を扱うモジュールの導入を検討

エラー内容	サンプル数
参照表現理解に関する誤り	6
目的語選択に関する誤り	5
Object Groundingに関する誤り	3
アノテーションに関する誤り	2
曖昧な指示文	2
その他	2
合計	20

指示文：" Proceed to the hallway on level 2 with the basketball and level painting above the open book"

Ground Truth

提案手法 (Rank 1)