Crosslingual Visual Prompt に基づく テキスト付き画像からの日常物体検索

Scene Text Aware Multimodal Retrieval for Everyday Objects Based on Crosslingual Visual Prompts

戸倉 健登 Kento Tokura　是方 諒介 Ryosuke Korekata　小松 拓実 Takumi Komatsu　今井 悠人 Yuto Imai　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

This study explores a task where a robot searches for images containing target objects based on user language queries from a large set of images captured in diverse indoor and outdoor environments. Both images with and without scene text are considered. For example, when searching with the query, "Pass me the red container of Sun-Maid raisins on the kitchen counter," the model ranks images containing a container labeled "Sun-Maid raisins" on the kitchen counter higher. However, linking visual semantics with scene text is challenging. Additionally, multimodal search requires large-scale, high-speed inference, making it impractical to rely solely on a multimodal large language model (MLLM). To address this, we introduce a Scene Text Visual Encoder, integrating an Aligned Representation with a narrative representation obtained using an MLLM based on Crosslingual Visual Prompting. Incorporating OCR results into the prompt further reduces hallucination. Experiments show that the proposed method outperforms multimodal foundation models across multiple benchmarks in standard evaluation metrics for ranking-based learning.

1. はじめに

屋内外の画像に対してユーザが自然言語で検索可能なマルチモーダル検索は，公共空間における物体検索や生活支援ロボットによる物体操作などの幅広い応用を持つ．特に，商品ラベルや標識などに含まれる scene text（画像中の文字情報）を考慮したマルチモーダル検索は有用性が高い．

本研究では，ロボットが屋内外の広範な環境で撮影した画像群から，ユーザによる言語クエリに基づき対象となる物体を含む画像を検索するタスクを扱う．本タスクでは，クエリは物体操作や移動に関する多様な自然言語で記述される．また，scene text を含む画像と含まない画像の両方を検索対象として扱う．本研究のユースケースとしては，ロボットによる物体検索および物体操作などが挙げられる．例えば，ユーザが “Pass me the Sun-Maid raisins from the counter.” と指示すると，ロボットが事前に収集した環境画像からキッチンカウンターにある “Sun-Maid” と書かれた容器を検索し，物体操作を行う．

本タスクは，物体の視覚的特徴および空間的関係に加え scene text を考慮した統合的な理解をする必要があるため，困難である．また，マルチモーダル検索では大規模かつ高速な推論が要求されるため，multimodal large language model (MLLM) のみを用いる手法は現実的ではない．既存手法 [Radford 21, Wang 23] はマルチモーダル検索において良好な結果を得ているが，scene text とその他のマルチモーダル情報との統合が不十分であり， 本タスクにおける性能は十分でない．

そこで，本研究ではロボットが屋内外の広範な環境で撮影した scene text を含む画像群から，ユーザによる言語クエリに基づき対象となる物体を含む画像を検索するモデルを提案する．本手法では，言語とアラインされた画像特徴量と scene text に関する特徴量を MLLM による画像説明を介して統合することで，scene text を捉えた画像特徴量群を対象とした検索を可能とする．既存手法群と異なる点は，Crosslingual Visual Prompt (CVP) に基づく MLLM を用いた narrative representation を導入し，scene text を捉えた画像特徴量を獲得することである．これにより，scene text を明示的に強調した narrative representation を得ることが期待される．また，MLLM を用いた画像説明ではしばしばハルシネーションを生じることが問題とされており [Matsuda 24]，OCR で得られた scene text をプロンプトに含むことで，ハルシネーションが抑制されることが期待される．

図 1: 本タスクの入力例

提案手法の新規性は次の通りである．

- CVP に基づく MLLM を用いた narrative representation を，言語とアラインされた aligned representation と統合する Scene Text Visual Encoder(STVE) を導入する．
- OCR の結果をプロンプトに組み込むことで，ハルシネーションを抑制した narrative representation を得る．

2. 問題設定

本タスクでは，クエリとの関連度の高い画像が上位にランキングされたリストを出力することが望ましい．図 1 に本タスクの入力例を示す．本タスクでは，例えば “Please pass me the Banana Nut Crunch” などのクエリが入力される．モデルは

連絡先: 戸倉健登，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，tkento1985@keio.jp

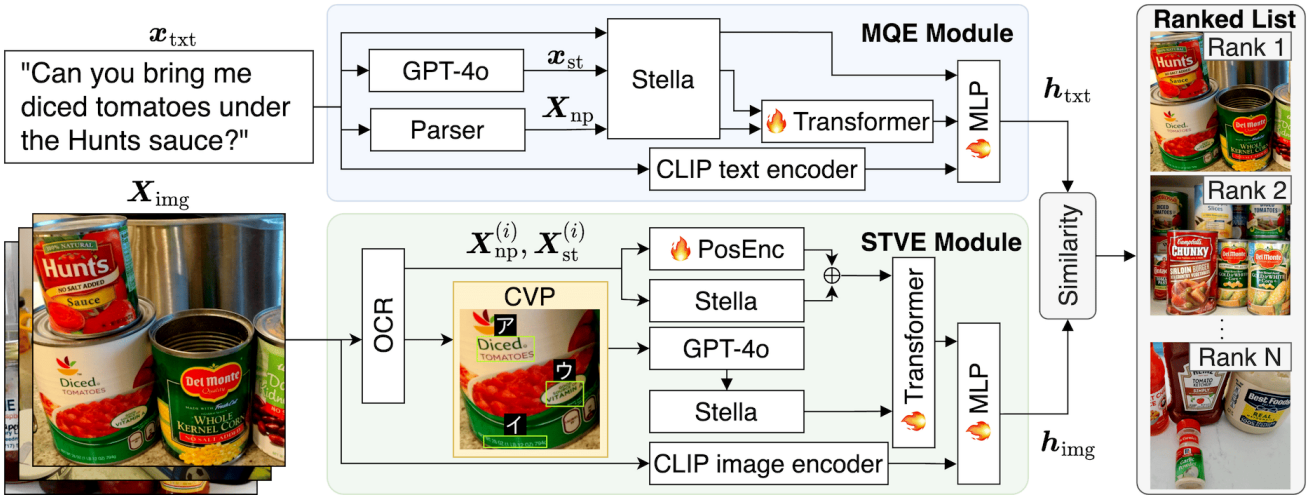

図 2: 提案手法のモデル構造．“MLP”，“PosEnc”，および $\oplus$ はそれぞれ多層パーセプトロン，位置埋め込み，および和を示す．

緑枠で示されるような scene text を考慮し，事前に収集した画像群から，対象物体である “Banana Nut Crunch” を含む画像を上位にランキングすることが求められる．

本タスクの入力は，クエリおよび事前に収集された画像群である．ただし，対象画像は scene text を含む場合も含まない場合もあり得る．出力は，クエリとの関連度に基づきランキングされた画像のリストである．本研究で使用する用語を以下のように定義する．

- **クエリ:** 参照表現を含む多様な自然言語の物体操作指示文
- **対象物体:** クエリに対応する物体
- **対象画像:** 対象物体を含む画像
- **scene text:** 画像内に存在するテキスト

本研究では，調理器具やボトル等の操作が可能な日常物体や，建物や標識等の移動目標となり得る物体が対象画像に含まれることを前提とする．

3. 提案手法

本研究では，マルチモーダル検索手法 [Kaneda 24, Korekata 25, Yashima 25, Chen 23] を拡張し，scene text を含む画像群から自然言語クエリに基づき対象画像を検索する Scene Text Aware Text-Image Retrieval for Everyday Objects (STARE) を提案する．提案手法では，CVP に基づく MLLM を用いた narrative representation を導入する．CVP により，MLLM が scene text を効果的に用いて画像の詳細を説明することが可能になると期待される．そのため，本拡張は scene text を含む画像に加え，scene text を含まない画像を対象とするマルチモーダル検索タスクにおいても広く適用可能である．

図 2 に提案手法のモデル構造を示す．提案手法は，STVE モジュールおよび Multiform Query Encoder (MQE) モジュールの 2 つの主要モジュールで構成される．提案手法の入力 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = \{\boldsymbol{X}_{\text{img}}, \boldsymbol{x}_{\text{txt}}\}, \quad \boldsymbol{X}_{\text{img}} = \{\boldsymbol{x}^{(i)}_{\text{img}}\}^{N_{\text{img}}}_{i} \tag{1}$$

ここで，$\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V\times L}$ および $\boldsymbol{x}^{(i)}_{\text{img}} \in \mathbb{R}^{3\times W\times H}$ はそれぞれクエリおよび画像を表す．また，N_{img}，H，W，V，および L は，それぞれ検索対象とする画像の数，画像の高さ，幅，語彙サイズ，および最大トークン長を表す．

3.1 Scene Text Visual Encoder

STVE モジュールでは，scene text を捉えた画像特徴量を得るために，CVP に基づく MLLM を用いた narrative representation および言語とアラインされた aligned representation を統合する．既存手法 [Korekata 25, Yashima 25] では，scene text とこれらの画像特徴量の統合が不十分である．しかし，日常物体には scene text が含まれる場合が多いため，画像に含まれる scene text を考慮した画像特徴量の抽出は重要である．そこで，本モジュールでは画像を入力とし，OCR の結果から画像内での位置情報を含む scene text 特徴量および CVP に基づく MLLM を用いた narrative representation を抽出する．この際，OCR 結果に基づいてカタカナで記述されるマークを画像に重畳する CVP および scene text を用いたプロンプトを用いることで，ハルシネーションを抑制するとともに，画像内の scene text を捉えた narrative representation を得る．これらを言語とアラインされた aligned representation と統合することで，scene text を捉えた画像特徴量を得る．

本モジュールへの入力は $\boldsymbol{x}^{(i)}_{\text{img}}$ であり，OCR を用いて scene text $\boldsymbol{X}^{(i)}_{\text{st}} = \{\boldsymbol{x}^{(i,j)}_{\text{st}}\}^{N_{\text{ocr}}}_{j=1}$ およびその領域 $\boldsymbol{X}^{(i)}_{\text{bbox}} = \{\boldsymbol{x}^{(i,j)}_{\text{bbox}}\}^{N_{\text{ocr}}}_{j=1}$ を得る．ここで，N_{ocr} は OCR の検出結果数を示す．ここで，d_{sl} は Stella の出力次元数を表す．次に，画像内での位置情報を考慮した scene text 特徴量 $\boldsymbol{V}^{(i)}_{\text{lst}} = \{\boldsymbol{v}^{(i,j)}_{\text{lst}}\}^{N_{\text{st}}}_{j=1}$ を以下のように得る．

$$\boldsymbol{v}^{(i,j)}_{\text{lst}} = \text{Transformer}\left(\boldsymbol{v}^{(i,j)}_{\text{sl}} + \text{PosEnc}\left(\boldsymbol{x}^{(i,j)}_{\text{bbox}}\right)\right) \tag{2}$$

ここで，N_{st}，Transformer$(\cdot)$，および PosEnc$(\cdot)$ は，それぞれ scene text の数，transformer エンコーダ，および MLP による位置埋め込みを示す．なお，本研究では $N_{\text{st}} = 50$ とする．

次に，以下のように CVP および scene text を用いて，narrative representation を得る．

1. OCR によって検出された領域を対象に，面積の 20%以上が重なる領域を統合する．
2. 領域の面積が大きいものから P_{pmt}(=10 個) の領域を選択し，画像内にマークを重畳したものを CVP とする．本研究ではマークとして日本語のカタカナを用いて，領域の x 座標が小さい順にそれぞれ「ア」～「コ」を領域の中央上部に重畳する．
3. さらに，CVP とともに各マークに対応する scene text のリストをプロンプトに記述することで，scene text に関

するハルシネーションに頑健な画像説明を得る．

4. このようにして得られる画像説明を，Stella エンコーダを用いて特徴抽出することで narrative representation $\boldsymbol{v}_{\mathrm{nr}}^{(i)} \in \mathbb{R}^{d_{\mathrm{sl}}}$ を得る．

ここで，日本語のカタカナは画像内に含まれる主要な言語とは異なる言語かつ箇条書きに適した文字であるため，visual prompt として有効であると考える．

また，クエリの言語特徴量とのアラインするために，CLIP 画像エンコーダ [Radford 21] を用いて $\boldsymbol{x}_{\mathrm{img}}^{(i)}$ から aligned representation $\boldsymbol{v}_{\mathrm{cl}}^{(i)} \in \mathbb{R}^{d_{\mathrm{cl}}}$ を抽出した．ここで，d_{cl} は CLIP の出力次元を示す．STVE モジュールの最終的な出力 $\boldsymbol{h}_{\mathrm{img}}^{(i)} \in \mathbb{R}^{d_{\mathrm{img}}}$ は，以下のように得られる．

$$\boldsymbol{h}_{\mathrm{img}}^{(i)} = \mathrm{MLP}\left(\left[\mathrm{Transformer}\left(\left[\boldsymbol{V}_{\mathrm{lst}}^{(i)}, \boldsymbol{v}_{\mathrm{nr}}^{(i)}\right]\right), \boldsymbol{v}_{\mathrm{cl}}^{(i)}\right]\right) \quad (3)$$

ここで，d_{img} および MLP$(\cdot)$ は，STVE モジュールの出力次元および多層パーセプトロンを示す．

3.2 Multiform Query Encoder

MQE モジュールでは，自然言語で記述された多様な形式のクエリを用いた検索を可能とするため，対象物体および名詞句など複数粒度でクエリの分解を行い，各言語特徴を統合する．本タスクで扱うクエリは，対象物体名のみを含む名詞句，物体操作や移動に関する指示文，および対象物体に関する質問など，多様な形式が存在する．これらのクエリには複雑な参照表現が含まれる場合があり，対象物体と参照表現の関係性を捉えることが困難である．そこで，本モジュールではクエリを複数粒度で分解および統合することにより言語特徴量を獲得する．

本モジュールへの入力は $\boldsymbol{x}_{\mathrm{txt}}$ であり，複数の言語エンコーダ（Stella，CLIP）を用いてクエリ全体に関する言語特徴量 $\boldsymbol{l}_{\mathrm{q}} \in \mathbb{R}^{d_{\mathrm{tenc}}}$ を得る．ここで，d_{tenc} は言語エンコーダの出力次元を示す．また，参照表現に関する特徴抽出を行うため，クエリ中に含まれる対象物体に関するフレーズ $\boldsymbol{x}_{\mathrm{targ}}$ および名詞句 $\boldsymbol{X}_{\mathrm{np}} = [\boldsymbol{x}_{\mathrm{np}}^{(1)}, \ldots, \boldsymbol{x}_{\mathrm{np}}^{(N_{\mathrm{np}})}]$ を獲得し，transformer を用いて統合することで，対象物体および参照表現を考慮した言語特徴量を獲得する．ここで，$\boldsymbol{x}_{\mathrm{targ}}$ は GPT-4o [OpenAI 24]，$\boldsymbol{X}_{\mathrm{np}}$ は構文解析器を用いて取得した．また，N_{np} は取得された名詞句の数を示す．Stella エンコーダを用いて，$\boldsymbol{x}_{\mathrm{targ}}$ および $\boldsymbol{X}_{\mathrm{np}}$ から言語特徴量 $\boldsymbol{L}_{\mathrm{np}} = [\boldsymbol{l}_{\mathrm{np}}^{(1)}, \ldots, \boldsymbol{l}_{\mathrm{np}}^{(N_{\mathrm{np}})}], \boldsymbol{l}_{\mathrm{np}}^{(i)} \in \mathbb{R}^{d_{\mathrm{sl}}}$ を得る．MQE モジュールの最終的な出力 $\boldsymbol{h}_{\mathrm{txt}} \in \mathbb{R}^{d_{\mathrm{txt}}}$ は，以下のように得られる．

$$\boldsymbol{h}_{\mathrm{txt}} = \mathrm{MLP}\left(\left[\mathrm{Transformer}\left(\left[\boldsymbol{l}_{\mathrm{targ}}, \boldsymbol{L}_{\mathrm{np}}\right]\right), \boldsymbol{l}_{\mathrm{q}}\right]\right) \quad (4)$$

ここで，d_{txt} は MQE モジュールの出力次元を示す．上記で得られた $\boldsymbol{h}_{\mathrm{txt}}$ および $\boldsymbol{h}_{\mathrm{img}}^{(i)}$ を用いて，$\boldsymbol{x}_{\mathrm{txt}}$ および $\boldsymbol{x}_{\mathrm{img}}^{(i)}$ の類似度 s$(\cdot,\cdot)$ を以下のように定義する．

$$\mathrm{s}\left(\boldsymbol{x}_{\mathrm{txt}}, \boldsymbol{x}_{\mathrm{img}}^{(i)}\right) = \frac{\boldsymbol{h}_{\mathrm{txt}} \cdot \boldsymbol{h}_{\mathrm{img}}^{(i)}}{\|\boldsymbol{h}_{\mathrm{txt}}\| \|\boldsymbol{h}_{\mathrm{img}}^{(i)}\|} \quad (5)$$

本モデルは，類似度に基づきランキングした画像リスト $\hat{\boldsymbol{Y}}$ を出力する．本研究では，positive ペアのコサイン類似度を最大化しつつ，unlabeled positive ペアおよび negative ペアの対称性を緩和し，モデルを最適化する double relaxed contrastive 損失 [Yashima 25] を使用する．

4. 実験設定

本タスクのための標準データセットは我々の知る限り存在しないため，本研究では新たに RefText-R および RefText-RM データセットを構築した．本タスクで扱うデータセットは，scene text を含む画像で構成されることが望ましい．さらに，操作可能な日常物体や移動目標となり得る物体を対象とする，scene text を考慮した指示文形式のクエリで構成されることが望ましい．

本研究では RefText データセットを基に，RefText-R データセットを構築した．RefText データセット [Bu 23] は，屋内外の多様な環境における scene text を含む画像を扱ったデータセットである．RefText-R データセットは，RefText データセットに含まれる画像およびクエリによって構成される．さらに，物体操作に関する指示文形式のクエリを用いて評価を行うために，RefText-RM データセットを構築した．RefText-RM データセットでは，RefText-R データセットを構成する画像に対し，指示文形式のクエリを新たに付与した．また，scene text を含まない屋内環境を対象としたマルチモーダル検索タスクのデータセットである LTRRIE データセット [Kaneda 24] を使用し，モデルの性能を検証した．

RefText-R データセットおよび RefText-RM データセットの構築手順は次の通りである．RefText データセットは，単一の画像内から対象物体を特定するタスクを対象としたデータセットであるため，RefText データセットを本タスクで直接使用することはできない．そのため，RefText データセットにおけるクエリと対象物体の矩形領域のペアで構成されるサンプルを，クエリと対象物体の矩形領域を含む対象画像のペアで構成されるサンプルに変更した．また，RefText データセットのサブセットのうち “sport” に属する画像は対象画像として適さないため，検索対象の画像群から除外した．RefText-RM データセットでは，RefText-R データセットを構成する画像に対し，クラウドソーシングを用いて参照表現を用いた物体操作に関する指示文形式のクエリ（例：“Pass me the frizz ease from the top shelf.”）を付与した．アノテータには対象画像および対象物体を示す矩形領域を提示し，ロボットが対象物体に関する物体操作や移動を行うための英語のクエリを回答するように指示した．この際，scene text もアノテータに提示することで，scene text を適切に用いたクエリ作成を指示した．

RefText-R データセットは，3,657 の対象画像と，21,481 の英語クエリを含む．語彙サイズは 12,303 語，全単語数は 149,547 語，平均文長は 6.96 語である．また，RefText-RM データセットは，RefText-R データセットに含まれる 1,721 の対象画像と，129 名のアノテータによって付与された参照表現を含む 3,957 の英語クエリで構成される．語彙サイズは 6,704 語，全単語数は 62,864 語，平均文長は 15.89 語である．RefText-R データセットおよび RefText-RM データセットにおける訓練集合，検証集合，およびテスト集合の分割は RefText データセットと同様の分割を行った．

提案手法における学習可能なパラメータ数は約 40M，積和演算数は約 5G であった．モデルの訓練には VRAM24GB の GeForce RTX 4090 および RAM64GB 搭載の Intel Core i9 13900KF を使用した．モデルの訓練時間は約 0.9 時間，推論時における 1 クエリと 100 枚の画像群間の計算は約 20.9ms を要した．各エポックで検証集合を用いて Recall@5 を計算し，Recall@5 が最大となったモデルを用いて，テスト集合における評価を行った．

5. 実験結果

5.1 定量的結果

表 1 に，RefText-R データセット，RefText-RM データセット，および LTRRIE データセット [Kaneda 24] におけるベースライン手法および提案手法の定量的比較結果を示す．ここ

表 1: ベースライン手法との定量的比較結果

[%]	手法	RefText-R			RefText-RM			LTRRIE		
		R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
(i)	[Radford 21]	40.8	63.2	72.8	51.2	74.3	83.3	19.2	56.1	71.0
(ii)	[Wang 23]	32.4	54.4	65.3	44.0	63.7	79.5	20.2	59.9	76.6
(iii)	提案手法	**50.4**	**79.6**	**87.2**	**55.1**	**79.5**	**86.4**	**24.0**	**67.6**	**82.3**

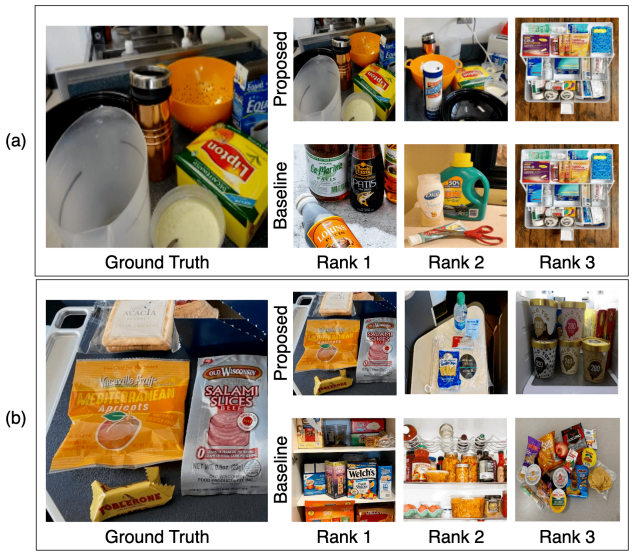

図 3: (a) RefText-R データセットおよび (b) RefText-RM データセットにおける提案手法およびベースライン手法 [Radford 21] の定性的結果

で，RefText-R データセットおよび RefText-RM データセットでは “street”，“shelf”，“home”，“others”，“oov”，および “semantic information” サブセット毎に評価した平均値を示す．また，表中の太字は各評価尺度において最も高い数値を表す．本研究では，マルチモーダル検索タスクで良好な結果を得ている CLIP および BEiT-3 をベースライン手法とした．評価尺度として，ランキング学習において標準的な評価尺度である Recall@K(K=1, 5, 10) を用いた [Cao 22]．

表 1 より，RefText-R データセット，RefText-RM データセット，および LTRRIE データセットの Recall@10 において，提案手法はそれぞれ 87.2%，86.4%，および 82.3%であった．ベースライン手法の最良値と比べて 14.4 ポイント，3.1 ポイント，および 5.7 ポイント上回った．他の評価尺度でも同様に，提案手法がベースライン手法を上回った．

5.2 定性的結果

図 3-(a) で与えられた $\boldsymbol{x}_{\text{txt}}$ は “The container with white liquid in front of Lipton.” である．本例において，ベースライン手法は “Lipton” に関連のない画像を上位にランキングした．一方で，提案手法は対象物体である「“Lipton” の前にある白い液体の入った容器」を含む画像を 1 位および 2 位にランキングした．

図 3-(b) において，与えられた $\boldsymbol{x}_{\text{txt}}$ は “Pass me a yellow TOBLERONE in front of the orange Apricots.” であり，対象物体は “TOBLERONE” と言う文字列が記載されたチョコレートであった．ベースライン手法は “TOBLERONE” および “Apricots” に関連しない画像を上位にランキングし，対象画像を 4 位にランキングした．一方で，提案手法は “TOBLERONE” を含む対象画像を適切に 1 位にランキングした．したがって，提案手法では scene text や物体の位置関係に対応した検索が可能であることが示唆される．

6. おわりに

本研究では，屋内外の広範な環境で撮影した画像群から，言語クエリに基づき対象物体を含む画像を検索するタスクを扱った．本研究は scene text を含む画像と含まない画像の両方を検索対象として扱うことが特徴である．実験の結果，ランキング学習における標準的な評価尺度において，提案手法は複数のベンチマークでマルチモーダル基盤モデルを上回る結果を得た．

謝辞

本研究の一部は，JSPS 科研費 23K28168，JST ムーンショットの助成を受けて実施されたものである．

参考文献

[Bu 23] Bu, Y., et al.: Scene-Text Oriented Referring Expression Comprehension, *IEEE TMM*, Vol. 25, pp. 7208–7221 (2023)

[Cao 22] Cao, M., Li, S., Li, J., Nie, L., and Zhang, M.: Image-text Retrieval: A Survey on Recent Research and Development, in *IJCAI*, pp. 5376–5383 (2022)

[Chen 23] Chen, B., Xia, F., Ichter, B., Rao, K., et al.: Open-vocabulary Queryable Scene Representations for Real World Planning, in *ICRA*, pp. 11509–11522 (2023)

[Kaneda 24] Kaneda, K., et al.: Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine, *IEEE RA-L*, Vol. 9, No. 3, pp. 2088–2095 (2024)

[Korekata 25] Korekata, R., Kaneda, K., et al.: DM2RM: Dual-Mode Multimodal Ranking for Target Objects and Receptacles Based on Open-Vocabulary Instructions, *AR* (2025)

[Matsuda 24] Matsuda, K., Wada, Y., and Sugiura, K.: Deneb: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning, in *ACCV*, pp. 3570–3586 (2024)

[OpenAI 24] OpenAI, : GPT-4o: Optimized Generative Pre-trained Transformer 4, https://openai.com (2024), Accessed: Jan. 2025

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., et al.: Learning Transferable Visual Models From Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Wang 23] Wang, W., Bao, H., Dong, L., Bjorck, J., et al.: Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks, *CVPR*, pp. 19175–19186 (2023)

[Yashima 25] Yashima, D., Korekata, others R., and Sugiura, K.: Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning With Dense Labeling, *IEEE RA-L*, Vol. 10, No. 2, pp. 1728–1735 (2025)