多言語シーンテキストを考慮した深層状態空間モデルに基づく実世界検索エンジン

○西牧宙輝，八島大地，戸倉健登，杉浦孔明（慶應義塾大学）

本研究では，言語クエリに基づき，対象となる日常物体を含む画像を検索するタスクを扱う．検索対象の画像は多言語のシーンテキストを含む場合があり，既存手法では特定の言語で記述されるクエリと多言語のシーンテキストの統合において不整合が生じてしまう．そこで本研究では，多言語のシーンテキストを正規化・拡張したテキスト情報および Crosslingual Visual Prompt を利用した視覚情報との関係をモデル化するマルチモーダル検索手法を提案する．また，深層状態空間モデルを用い，シーンテキストを含む視覚情報および複数階層の言語情報を統合する．新規構築したベンチマークを含む計 5 つのベンチマークで評価を行った結果，検索タスクの標準的な評価尺度において，提案手法がベースライン手法を上回った．

1. はじめに

高齢化および労働力不足が進行する現代社会において，物体の運搬や人間との協働作業が可能なロボットの重要性は高まっている．特に，家庭内での生活支援および商業施設における商品運搬補助はその代表的な応用例である．このような場面では，ユーザが自然言語を用いて対象物体を指示できれば利便性が高い．多くの商品や物体には文字が記載されているため，指示文の中でそれらを使用することは自然である．画像中に含まれる文字情報 (シーンテキスト) を考慮したマルチモーダル検索は広く研究が行われているが，現状では性能が十分ではない．

本研究では，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．ここで，検索対象の画像には多言語のシーンテキストを含む画像および含まない画像の両方が含まれる．

本タスクのユースケースとして，自由形式の指示文に基づいて対象となる物体を含む画像を検索し，その画像に基づいてロボットが物体操作を行う場面が挙げられる．図 1 に本タスクの代表例を示す．例えば，“Please take the clean and antibacterial sponge that’s to the right of the sheet-type sponge.” というクエリが与えられる．モデルは図中に示されるような多言語のシーンテキストを考慮し，事前に収集した画像群から対象画像を上位に順位付けすることが求められる．

本タスクは，物体の視覚的情報，物体間の空間的関係，および多言語のシーンテキスト情報の統合的な理解が必要であるため，困難である．また，大規模なデータに対して高速な推論が要求されるため，マルチモーダル LLM (MLLM) を直接用いる手法は非現実的である．既存手法 [1,2] は，マルチモーダル検索タスクにおいて良好な結果を得ているが，シーンテキストを明示的に考慮していないため，本タスクにおける性能は十分ではない．この課題に対し，戸倉ら [3] はシーンテキスト情報と視覚情報を統合し，本タスクにおいて良好な結果を得ている．しかし，同手法では光学文字認識 (OCR) で検出されたシーンテキストを直接用いているため，特定の言語で記述されるクエリと多言語のシーンテキストの統合において不整合が生じる．

そこで本研究では，多言語のシーンテキストを正規化・拡張したテキスト情報および Crosslingual Visual Prompt (CVP) を利用した視覚情報との関係をモデル化するマルチモーダル検索手法を提案する．これにより，元の言語に依存しないシーンテキスト情報の取得が可能となり，特定の言語で記述されるクエリと多言語のシーンテキストの統合における不整合を解消することが期待される．提案手法の新規性は以下である．

図 1　本タスクの代表例

- 多言語のシーンテキストを正規化・拡張したテキスト情報および CVP を利用した視覚情報との関係をモデル化する手法を導入する．
- 深層状態空間モデルを用い，シーンテキストを含む視覚情報および複数階層の言語情報を統合する M-STVE モジュールおよび M-FQE モジュールを導入する．これにより，MLLM が困難とするシーンテキスト情報の明示的な活用を可能とする．

2. 問題設定

本タスクでは，クエリとの関連度が高い画像が検索結果の上位に順位付けされることが望ましい．本タスクの入力は，クエリおよび検索対象の画像群である．ただし，対象画像は多言語のシーンテキストを含む場合と含まない場合がある．また，本タスクの出力はクエリとの関連度に基づいて順位付けされた画像のリストである．本論文では，クエリ，対象物体，対象画像，シーンテキスト，およびテキストの正規化・拡張をそれぞれ物体操作や移動に関する自由形式の指示文，クエリに対応する物体，対象物体を含む画像，画像に含まれるテキスト，およびテキストを特定の言語に翻訳または特定の言語で説明することと定義する．

本研究では，掃除用品や玩具等の操作可能な日常物体や，建物や家具等の移動目標となりうるランドマークに対象物体を限定する．また，検索対象の画像群が事前に収集済みであることを前提とする．

3. 提案手法

本研究では，Scene Text Aware Text-Image Retrieval for Everyday Objects (STARE) [3] を拡張し，多言語のシーンテキストを考慮したマルチモーダル検索手法を提案する．提案手法では，マルチモーダル検索の際に考慮するシーンテキストを英語に限定せず，多

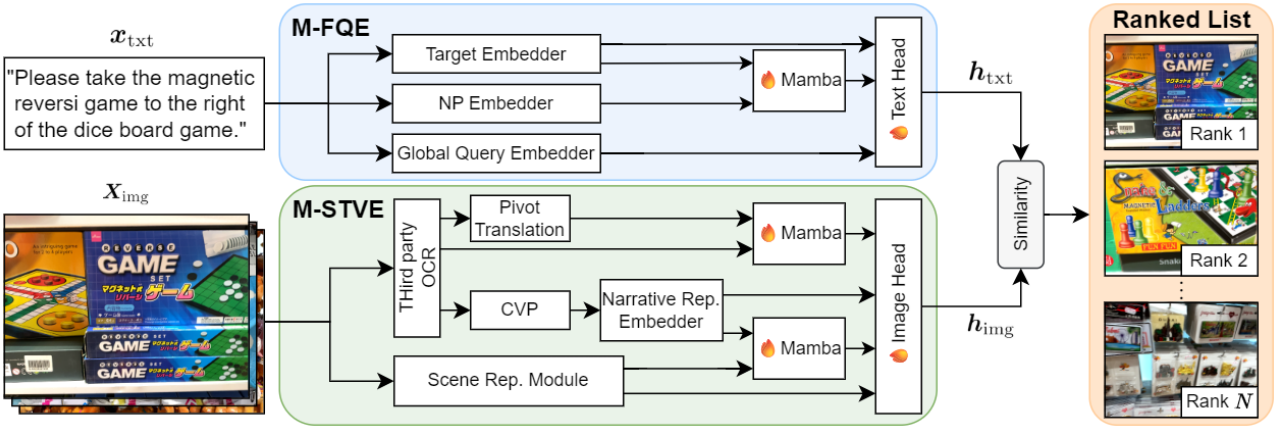

図 2　提案手法のモデル構造

言語に拡張する．本拡張は検索タスクに限らず，多言語のシーンテキストを含む画像を対象とした画像キャプショニングや視覚的質問応答などのタスクに適用可能である．

図 2 に提案手法のモデル構造を示す．本手法の主要モジュールは，Multilingual Scene Text Visual Encoder (M-STVE) モジュールおよび Multi-Form Query Encoder (M-FQE) モジュールである．本モデルへの入力 $\boldsymbol{x}$ を $\boldsymbol{x} = \{\boldsymbol{X}_{\text{img}}, \boldsymbol{x}_{\text{txt}}\}$, $\boldsymbol{X}_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)} \mid i = 1, \ldots, N_{\text{img}}\}$ と定義する．ここで，$\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V \times L}$ および $\boldsymbol{x}_{\text{img}}^{(i)} \in \mathbb{R}^{3 \times W \times H}$ はそれぞれクエリおよび画像を表す．また，V, L, i, N_{img}, W, および H はそれぞれ語彙サイズ，最大トークン長，各画像のインデックス，検索対象とする画像数，画像の幅，および画像の高さを表す．

3.1　M-STVE

M-STVE モジュールでは，多言語のシーンテキスト情報を考慮した画像特徴量を取得するため，多言語 MLLM を用いたシーンテキストの言語正規化を導入する．STARE では OCR によって検出されたシーンテキストを直接言語エンコーダに入力する．一般に，クエリは特定の言語であるため，多言語のシーンテキストとクエリの統合において不整合が生じる．そこで，本モジュールでは多言語 MLLM を用い，多言語のシーンテキストを一度クエリで使用される言語に翻訳したうえで言語エンコーダに入力する，言語正規化を導入する．これにより，元の言語に依存しないシーンテキスト特徴量の取得を可能とする．

本モジュールの入力は $\boldsymbol{x}_{\text{img}}^{(i)}$ である．はじめに，$\boldsymbol{x}_{\text{img}}^{(i)}$ に対して OCR を実行し，シーンテキスト $\boldsymbol{X}_{\text{st}}^{(i)} = \{\boldsymbol{x}_{\text{st}}^{(i,j)} \mid j = 1, \ldots, N_{\text{ocr}}^{(i)}\}$ を取得する．ここで，$N_{\text{ocr}}^{(i)}$ は $\boldsymbol{x}_{\text{img}}^{(i)}$ における OCR の検出結果数を表す．言語正規化では，多言語 MLLM を用いて $\boldsymbol{x}_{\text{st}}^{(i,j)}$ を言語正規化した結果を取得し，Stella [4] により言語特徴量 $\boldsymbol{V}_{\text{pstl}}^{(i)} = \{\boldsymbol{v}_{\text{pstl}}^{(i,j)} \mid j = 1, \ldots, N_{\text{st}}\}$, $\boldsymbol{v}_{\text{pstl}}^{(i,j)} \in \mathbb{R}^{d_{\text{stl}}}$ を得る．ここで，N_{st} および d_{stl} はそれぞれ考慮するシーンテキストの数および Stella の出力次元数を表し，本研究では $N_{\text{st}} = 50$ とする．$N_{\text{ocr}}^{(i)} < N_{\text{st}}$ の場合は，d_{stl} 次元のゼロベクトルによるパディングを行い，$N_{\text{ocr}}^{(i)} > N_{\text{st}}$ の場合は $\boldsymbol{X}_{\text{st}}^{(i)}$ を文字数の降順に並び替え，上位 N_{st} 個のシーンテキストを使用する．また，多言語 MLLM には Pangea [5] を使用し，$\boldsymbol{x}_{\text{st}}^{(i,j)}$ が固有名詞の場合はそれをクエリで使用される言語で説明し，それ以外の場合はクエリで使用される言語に翻訳するように指示する．同様に，Stella を用いて $\boldsymbol{X}_{\text{st}}^{(i)}$ の言語特徴量 $\boldsymbol{V}_{\text{stl}}^{(i)} = \{\boldsymbol{v}_{\text{stl}}^{(i,j)} \mid j = 1, \ldots, N_{\text{st}}\}$, $\boldsymbol{v}_{\text{stl}}^{(i,j)} \in \mathbb{R}^{d_{\text{stl}}}$ を取得する．その後，$\boldsymbol{v}_{\text{pstl}}^{(i,j)}$ および $\boldsymbol{v}_{\text{stl}}^{(i,j)}$ に位置埋め込みを付与し，深層状態空間モデルを用いてこれらを統合することで，シーンテキスト特徴量 $\boldsymbol{V}_{\text{st}}^{(i)} = \{\boldsymbol{v}_{\text{st}}^{(i,j)} \mid j = 1, \ldots, N_{\text{st}}\}$, $\boldsymbol{v}_{\text{st}}^{(i,j)} \in \mathbb{R}^{d_{\text{ssm}}}$ を得る．ここで，d_{ssm} は深層状態空間モデルに基づく系列エンコーダの出力次元数を表す．深層状態空間モデルには様々なタスク (例: 時系列予測や言語モデリング) において Transformer と同等以上の性能が報告されている Mamba [6] を採用した．

次に，CVP では STARE と同様に $\boldsymbol{x}_{\text{img}}^{(i)}$ に対して OCR 検出領域をマークと共に重畳する．その後，MLLM および Stella を用いて Narrative Representation の言語特徴量 $\boldsymbol{v}_{\text{nr}}^{(i)} \in \mathbb{R}^{d_{\text{stl}}}$ を取得する．なお，本研究では MLLM に GPT-4o を使用する．Scene Representation Module では，SigLIP2 [1] 画像エンコーダを用い，$\boldsymbol{x}_{\text{img}}^{(i)}$ および重複パッチに基づくマルチモーダル画像特徴量 $\boldsymbol{v}_{\text{sig}}^{(i)} \in \mathbb{R}^{2 \times d_{\text{sig}}}$ を取得する．ここで，d_{sig} は SigLIP2 画像エンコーダの出力次元数を表す．また，$\boldsymbol{x}_{\text{img}}^{(i)}$ の異なるレベルの視覚特徴を取得するため，$\boldsymbol{x}_{\text{img}}^{(i)}$ を DINOv2 [7] に入力し，複数の中間層における特徴量を取得する．次に，これらの中間層の視覚特徴と $\boldsymbol{v}_{\text{st}}^{(i)}$ を Mamba により統合することで視覚特徴量 $\boldsymbol{v}_{\text{mdi}}^{(i)} \in \mathbb{R}^{N_{\text{layer}} \times d_{\text{ssm}}}$ を取得する．ここで，N_{layer} は使用した DINOv2 の層数を表し，本研究では $N_{\text{layer}} = 4$ とする．また，DINOv2 の全 24 層のうち，第 6 層，第 12 層，第 18 層，および第 24 層を使用した．

M-STVE モジュールの最終的な出力 $\boldsymbol{h}_{\text{img}}^{(i)} \in \mathbb{R}^{d_{\text{img}}}$ は以下のように得られる．

$$\boldsymbol{h}_{\text{img}}^{(i)} = \text{ImageHead}\left(\left[\boldsymbol{V}_{\text{st}}^{(i)}; \boldsymbol{v}_{\text{nr}}^{(i)}; \boldsymbol{v}_{\text{sig}}^{(i)}; \boldsymbol{v}_{\text{mdi}}^{(i)}\right]\right) \quad (1)$$

ここで，d_{img} および ImageHead$(\cdot)$ はそれぞれ M-STVE モジュールの出力次元数および非線形変換を行う FFN を表す．

3.2　M-FQE

M-FQE モジュールでは複数の形式の $\boldsymbol{x}_{\text{txt}}$ から対象物体を捉えるために，$\boldsymbol{x}_{\text{txt}}$ 全体，対象物体，および名詞句などの複数の粒度で言語特徴量を取得し，統合する．本タスクにおいて，$\boldsymbol{x}_{\text{txt}}$ は対象物体に関する名詞句もしくは物体操作や移動に関する指示文などの多様な形式である．また，$\boldsymbol{x}_{\text{txt}}$ には複雑な参照表現が含ま

れる場合があり，それらに基づいて対象物体を特定する必要がある．そこで，本モジュールでは入力 $\boldsymbol{x}_{\text{txt}}$ を複数粒度で分解し，それぞれの言語特徴量を統合する．

はじめに，Target Embedder では LLM を用いて $\boldsymbol{x}_{\text{txt}}$ 中の対象物体に関する名詞句 $\boldsymbol{x}_{\text{targ}}$ を取得し，Stella を用いて言語特徴量 $\boldsymbol{l}_{\text{targ}} \in \mathbb{R}^{d_{\text{stl}}}$ を得る．ここで，LLM には GPT-4o を使用した．次に，NP Embedder では一般的な構文解析器を用いて $\boldsymbol{x}_{\text{txt}}$ に含まれる全名詞句を取得し，Stella を用いて言語特徴量 $\boldsymbol{L}_{\text{np}} = \{\boldsymbol{l}_{\text{np}}^{(i)} \mid i = 1, \ldots, N_{\text{np}}\}$, $\boldsymbol{l}_{\text{np}}^{(i)} \in \mathbb{R}^{d_{\text{stl}}}$ を得る．ここで，$\boldsymbol{l}_{\text{np}}^{(i)}$ および N_{np} はそれぞれ名詞句の言語特徴量および $\boldsymbol{x}_{\text{txt}}$ に含まれる名詞句の数を表す．さらに，$\boldsymbol{l}_{\text{targ}}$ および $\boldsymbol{L}_{\text{np}}$ を Mamba を用いて統合し，$\boldsymbol{l}_{\text{ssm}} \in \mathbb{R}^{d_{\text{ssm}}}$ を得る．

Global Query Embedder では，2 つの言語エンコーダを用いて $\boldsymbol{x}_{\text{txt}}$ 全体に関する言語特徴量 $\boldsymbol{l}_{\text{g}} \in \mathbb{R}^{d_{\text{tenc}}}$ を得る．ここで，言語エンコーダには Stella および SigLIP2 言語エンコーダを使用し，ユニモーダルおよびマルチモーダル言語特徴量を取得する．また，d_{tenc} はそれらの出力次元数の和を表す．M-FQE モジュールの最終的な出力 $\boldsymbol{h}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$ は，以下のように得られる．

$$\boldsymbol{h}_{\text{txt}} = \text{TextHead}\left([\boldsymbol{l}_{\text{ssm}}; \boldsymbol{l}_{g}]\right) \tag{2}$$

ここで，d_{txt} および TextHead$(\cdot)$ はそれぞれ MFQE モジュールの出力次元数および線形層である． 本モデルは，上記で得られた $\boldsymbol{h}_{\text{txt}}$ および $\boldsymbol{h}_{\text{img}}^{(i)}$ のコサイン類似度に基づき，順位付けされた画像リスト $\hat{\boldsymbol{Y}}$ を出力する．損失関数には，Double Relaxed Contrastive (DRC) 損失 [8] を用いた．

4. 実験設定

本研究で扱う多言語のシーンテキストを考慮したマルチモーダル検索タスクにおいて，標準的なベンチマークは我々の知る限り存在しない．本タスクで扱うベンチマークは，多言語のシーンテキストを含む画像およびそれらを考慮したクエリで構成されることが望ましい．そのため，本研究では新たに Multilingual Scene Text Aware Retrieval (M-STAR) ベンチマークを構築した．M-STAR ベンチマークの画像には日本語，アラビア語，および英語などの多言語のシーンテキストが含まれていた．また，各画像には多言語のシーンテキストを考慮したクエリが，平均約 5 個ずつ付与された．

英語のシーンテキストのみを考慮したマルチモーダル検索のベンチマークとして，GoGetIt (RefText) ベンチマーク [3]，GoGetIt (Instruction) ベンチマーク [3]，および TextCaps-test ベンチマーク [3] を使用した．これらのベンチマークはシーンテキストを含む画像，英語のシーンテキストを用いた対象物体の名詞句，および物体操作指示文もしくは移動指示文から構成されるため使用した．また，シーンテキストを含まない画像群を検索対象とした場合のモデルを評価するため，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 [9] を使用した．

M-STAR ベンチマークの画像群は，ドバイに所在する日本の小売店舗において店内の商品を撮影して収集され，画像のサイズは 640 × 480 であった．各画像に対して OCR を実行し，画像のフィルタリングを OCR 検出結果の文字数および文字サイズを基に行った．文字数によるフィルタリングは，OCR において単語を構成する文字が切り離されて検出されてしまう場合があり，それらを排除する目的で行った．文字サイズに関するフィルタリングは，画像サイズに対してテキスト領域が小さい場合に OCR による文字の誤認識が起きる可能性が高いため，それらを排除する目的で行った．

文字数に関するフィルタリングの条件は $C \geq \theta$ であり，文字サイズに関するフィルタリングの条件は $w_{\text{bbox}} \geq kC$ である．ここで， C，θ，w_{bbox}，および k はそれぞれ OCR 検出結果の文字数，文字数の閾値，OCR 検出結果の矩形領域幅，および 1 文字あたりの最低幅を表し，w_{bbox} は画像サイズを 224 × 224 にリサイズした時の値である．英語と非英語で単語を構成する平均文字数および 1 文字あたりの幅が異なるため，それぞれの場合で以下のように定数を設定した．OCR 検出結果が英語の場合は $\theta = 3$ および $k = 1.5$ とし，非英語の場合は $\theta = 2$ および $k = 1.1$ とした．

その後，クラウドソーシングにより各画像にクエリを付与した．アノテーション画面には，対象物体を示す矩形領域を重畳した画像，OCR で検出されたシーンテキスト，およびアノテータへの指示を表示した．ここで，対象物体を示す矩形領域は物体検出モデル [10] を用いて付与した．アノテータにはシーンテキストを考慮した物体操作や移動に関する自由形式の指示文を英語で回答するように指示した．また，非英語の一般名詞および固有名詞のシーンテキストは，それぞれ英語に翻訳および英語で説明して使用するように指示した．

M-STAR ベンチマークのクエリ数，画像数，語彙サイズ，全単語数，および平均文長はそれぞれ 2,994 クエリ，580 枚，3,177 語，49,677 語，および 16.72 語であった．また，クエリは自由形式の指示文であり，シーンテキストの言語は多言語であった．M-STAR ベンチマークでは日本語のシーンテキストを扱うため，CVP のマークとしてカタカナではなくギリシャ文字を使用した．訓練集合はモデルの学習に使用し，検証集合はハイパーパラメータの調整に使用した．また，テスト集合はモデルの性能評価に使用した．訓練時における最適化手法，バッチサイズ，およびエポック数は AdamW, 128, および 6 であった．また，DRC 損失における α, λ, および γ はそれぞれ 0.8, 0.8, 0.5 であった．モデルの訓練には約 1 時間を要し，推論時における 1 クエリと 100 枚の画像群間の計算には約 29.9ms を要した．また，各エポックで検証集合を用いて recall@10 を計算し，recall@10 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 に提案手法とベースライン手法の定量的比較結果を示す．BLIP2 [11] および BEiT-3 [2] はゼロショット設定における実験結果を示す．ここで，事前学習済みモデルでは複数回の試行において一貫した結果が得られるため，各手法につき 1 回の実験結果を示す．他の手法については 5 回の実験を実施し，その結果の平均および標準偏差を示す．また，表中の太字は各評価指標における最も高い数値を表す．ベースライン手法には，SigLIP2 (ViT-So/14)，BLIP-2 (ViT-g)，BEiT-3 (large)，および STARE を用い，SigLIP2 は fine-tuning したモデルを用いた．SigLIP2，BLIP-2，および BEiT-3 はマルチモーダル検索タスクにおいて良好な結果を得ているため使用した．また，STARE は英語のシーンテキストを考慮したマルチモーダル検索タスクにおいて良好な結果を得ているため選択した．

評価尺度には recall@$K(K = 5, 10)$ を用い，主要評価尺度は recall@10 とした．Recall@K は，検索タスク

表 1　ベースライン手法との定量的結果

[%]	M-STAR		GoGet (RefText)		GoGet (Instruction)		Text-Caps-test		LTRRIE	
手法	R@5↑	R@10↑	R@5↑	R@10↑	R@5↑	R@10↑	R@5↑	R@10↑	R@5↑	R@10↑
BLIP2 [11]	54.9	71.5	-	-	-	-	76.8	86.3	66.6	81.7
BEiT-3 [2]	44.6	61.0	67.0	77.2	77.3	84.2	82.5	89.5	55.4	69.8
SigLIP2 [1]	**61.0** ±0.6	77.8 ±0.5	67.8 ±0.6	75.9 ±0.6	70.4 ±1.5	80.3 ±0.6	78.5 ±0.6	85.5 ±0.5	65.9 ±0.6	80.6 ±0.5
STARE [3]	58.2 ±1.2	76.9 ±0.7	78.7 ±0.6	86.8 ±0.3	79.7 ±0.6	86.7 ±0.5	81.5 ±1.1	87.6 ±0.7	71.9 ±0.7	86.2 ±0.2
提案手法	60.0 ±0.6	**81.0** ±1.0	**80.5** ±0.6	**88.2** ±0.5	**80.9** ±0.5	**88.1** ±0.5	**83.4** ±0.8	**90.6** ±0.8	**73.5** ±0.7	**86.6** ±0.8

図 3　提案手法およびベースライン手法の定性的結果

において標準的な評価尺度であるため使用した [12].

表 1 に示すように，M-STAR ベンチマーク，GoGetIt (RefText) ベンチマーク，GoGetIt (Instruction) ベンチマーク，TextCaps-test ベンチマーク，および LTRRIE データセットの recall@10 において，ベースライン手法で最も高いスコアはそれぞれ 77.8%，86.8%，86.7%，89.5%，および 86.2%であった．一方で，提案手法はそれぞれ 81.0%，88.2%，88.1%，90.6%，および 86.6%であり，ベースライン手法の最良値と比べて 3.2 ポイント，1.4 ポイント，1.4 ポイント，1.1 ポイント，および 0.4 ポイント上回った．また，M-STAR ベンチマークの recall@5 を除き，提案手法はベースライン手法を上回る結果であった．

5.2　定性的結果

図 3 に，GoGetIt (Instruction) ベンチマークにおける提案手法およびベースライン手法の定性的結果を示す．ここで，ベースライン手法には，5 つのデータセットにおける recall@10 の平均値が最も高かった STARE を採用した．ここでは，対象画像，対象物体，および各手法における検索結果の上位 3 枚の画像を示す．図 3 において与えられた $\boldsymbol{x}_{\text{txt}}$ は，“Please bring me the purple Lavender bottle on the right side of white thieves bottle.” であった．ベースライン手法は「thieves」や「Lavender」などのシーンテキストに基づいた順位付けができておらず，対象画像は上位 3 位以内に含まれなかった．一方，提案手法はこれらのシーンテキストに基づき対象画像を 1 位に順位付けした．これらの結果から，提案手法では多言語のシーンテキストに基づいて対象画像を上位に順位付けすることが可能であることが示唆される．

5.3　Ablation studies

表 2 に ablation study の結果を示す．各モデルについてそれぞれ 5 回の実験を実施し，M-STAR ベンチマークにおける結果の平均および標準偏差を示す．また，表中の太字は各評価指標における最も高い数値を表す．Ablation study として以下の 2 つの検証を行った．

Pivot translation ablation.　言語正規化の性能への寄与を調べるため，Pivot Translation ブロックを除いて実験を行った．モデル (iii) と比較してモデル (i) では，recall@10 が 3.2 ポイント低下した．この結果より，M-STAR ベンチマークのような多言語のシーンテキストを考慮したマルチモーダル検索において，言語正規化が性能に寄与していることが示唆される．

表 2　Ablation study の結果

[%]	Pivot	アーキテクチャ	M-STAR	
手法	Translation		R@5↑	R@10↑
(i)		Mamba	55.7 ±0.4	77.8 ±0.9
(ii)	✓	Transformer	58.8 ±1.0	79.6 ±0.4
(iii)	✓	Mamba	**60.0** ±0.6	**81.0** ±1.0

Mamba ablation.　Mamba による性能への寄与を調べるため，Transformer に置き換えて実験を行った．モデル (iii) と比較してモデル (ii) では，recall@10 が 1.4 ポイント低下した．この結果より，Mamba の選択的機構 (S6) が Transformer の注意機構に比べて効果的に働き，性能が向上することが示唆される．

6.　おわりに

本研究では，クエリに基づき，事前に収集した画像群から対象画像を検索するタスクを扱った．ここで，対象画像には多言語のシーンテキストを含む画像および含まない画像の両方を扱った．新規構築した M-STAR ベンチマークを含む計 5 つのベンチマークで評価を行った結果，検索タスクの標準的な評価尺度において提案手法がベースライン手法を上回った．

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

[1] M. Tschannen, et al., “SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features,” arXiv:2502.14786, 2025.

[2] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, et al., “Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks,” CVPR, pp.19175–19186, 2023.

[3] 戸倉健登他，“固有表現とシーンテキストを考慮したリランキング付きマルチモーダル検索,” 第 28 回画像の認識・理解シンポジウム，2025.

[4] D. Zhang, et al., “Jasper and Stella: distillation of SOTA embedding models,” arXiv preprint arXiv:2412.19048, 2024.

[5] X. Yue, Y. Song, A. Asai, et al., “Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages,” ICLR, 2025.

[6] A. Gu and T. Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” COLM, 2024.

[7] M. Oquab, T. Darcet, et al., “DINOv2: Learning Robust Visual Features without Supervision,” TMLR, pp.1–31, 2024.

[8] D. Yashima, et al., “Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning with Dense Labeling,” IEEE RA-L, vol.10, pp.1728–1735, 2024.

[9] K. Kaneda, 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.

[10] X. Zhou, et al., “Detecting Twenty-thousand Classes using Image-level Supervision,” ECCV, pp.350–368, 2022.

[11] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” ICML, pp.19730–19742, 2023.

[12] M. Cao, et al., “Image-text Retrieval: A Survey on Recent Research and Development,” IJCAI, pp.5410–5417, 2022.