同じ画像，異なる意味：文脈依存画像検索タスクの提案

堤 歩斗 [1] 小比田 涼介 [2]
[1] 東京都立大学 [2] サイバーエージェント
tsutsumi-ayuto@ed.tmu.ac.jp kohita_ryosuke@cyberagent.co.jp

概要

画像の意味は画像内に閉じていない．「夜の窓辺に立つ人影」という画像があったとき，ロマンス映画では「家族の帰りを待つ温かな姿」を，サスペンス映画では「誰かに監視されている不気味さ」を意味しうる．つまり，同じ画像であっても文脈によって意味が大きく変わりうる．このような文脈に依存した意味の変容はさまざまな場面で観察されるものであるが，画像意味の文脈依存性を扱う研究は少なかった．意味の文脈依存性の考慮は，特に画像の意味検索を深めていくにあたって重要な方向性である．本稿では，文脈を考慮した画像の意味検索の実現に向けて，問題設定の定式化とデータセットの構築を行い，文脈依存画像検索というタスクを提案する．文脈による意味変化の要素を抽象度の異なる 4 段階のカテゴリに整理し，それぞれに対応した評価用データセットを構築した．実験の結果，文脈による意味変化は抽象度の高いカテゴリで特に顕著であることを確認し，文脈を考慮しない既存モデルは物体や行動といったオブジェクト認識は得意であるが，雰囲気や感情といった抽象的な意味の捉え方には課題があることを示した．一方で VLM は文脈を与えることによって抽象的な意味とその変化をある程度捉えられることを確認し，本研究は文脈を考慮した画像理解という新たな研究領域を開拓するものである．

1 はじめに

画像の意味は画像内に閉じていない．図 1 に示すように，同一の画像であっても，ロマンスの文脈では「家族の帰りを待つ」温かい場面を，サスペンスの文脈では「ターゲットを監視する」不穏な場面を表現しうる．つまり，同じ画像であっても文脈によって意味が大きく変わりうる．特に，「雰囲気」や「心理的効果」といった抽象度の高い意味ほど，文脈によって変化しやすい．この現象は認知科学や映画理論では古くから知られており，タスクによって注視パターンが変化すること [1]，文脈によって同じ顔から読み取る感情が変化すること [2] が示されてきた．しかし，画像の意味理解や検索の分野ではこのような文脈依存の意味は体系的に扱われてこなかった．現在の CLIP や BLIP といった画像・言語モデルは，画像内に閉じた具体的な意味の捉え方には優れている一方で，文脈に依存して変容しやすい抽象的な意味の扱いは困難である．

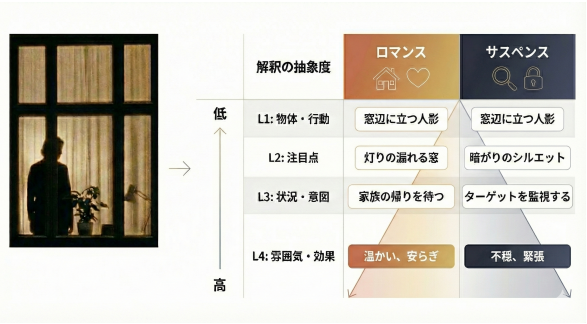

図 1 同一画像でもコンテキストにより意味が変化する例．窓辺に立つ人影という同じ画像（左）が，ロマンス文脈では「家族の帰りを待つ」温かい場面として，サスペンス文脈では「ターゲットを監視する」不穏な場面として解釈される（右）．L1（物体・行動）は文脈によらず共通だが，抽象度が高くなるほど解釈が分岐する．

本研究では，画像意味の文脈依存性という問題に着目し，文脈依存画像検索というタスクを提案する．画像意味を具体的な「物体・行動」から抽象的な「雰囲気・効果」まで 4 段階（L1–L4）に整理することで，文脈による意味変化を体系的に捉える枠組みを構築した．この分析枠組みに基づく評価用データセットを構築し，既存手法の適用可能性を検証した結果，抽象的意味ほど文脈による振れ幅が大きくなること，および VLM へのコンテキスト注入が抽象的意味の検索に有効であることを明らかにした．

2 関連研究

CLIP と固定埋め込みの限界 CLIP [3] は画像とテキストを共通空間に写像する強力なモデルであ

り，物体や行動など具体的な視覚内容の検索に優れる．しかし，CLIP の画像エンコーダは画像のみから表現を生成するため，文脈に応じて同一画像の表現を変えることができない．

注目点の切替と参照表現理解　「画像中のどこに注目するか」が指示で変化する問題は，参照表現理解や phrase grounding として研究されてきた．RefCOCO [4] や GuessWhat?! [5] は，自然言語から対象領域を同定するベンチマークを提供する．これらの多くは「正しい参照先は一つ」という前提で曖昧性を解消することを目的とする．

状況・意図の推論　「何が起きていて，登場人物は何を意図しているか」という推論は，物語理解や心の理論として研究されてきた．VisualCOMET [6] は画像から意図や因果関係を推論するデータセットを提供する．

雰囲気・効果の推定　雰囲気・感情・美的印象といった抽象的評価は，Affective Image Content Analysis [7] や visual sentiment analysis として体系化されている．SentiBank [8] や ArtEmis [9] は感情反応のモデル化を進めてきた．ただしこれらの多くは「画像が内在的に持つ印象」を推定する設定である．

VLM の埋め込み応用　E5-V [10] や VLM2Vec [11] は VLM の隠れ状態を埋め込みとして利用し，文脈条件付きの表現抽出を可能にしつつある．これらは画像と文脈を統合的に処理して文脈条件付きの埋め込みを得るための技術的基盤であるが，既存の検証は主に画像内のオブジェクト認識に留まっており，文脈依存的な意味理解への適用可能性は十分に検証されていない．

本研究の位置づけ　既存研究は主に画像内に閉じた意味理解を扱ってきたが，本稿で扱うような文脈依存的な意味変容を扱う枠組みは整備されていない．上述の各研究領域—客観的内容の認識，注目点の切替，状況・意図の推論，雰囲気・効果の推定—を抽象度の階層として統合し，同一画像の解釈変化を体系的に計測するフレームワークを以下で提示する．

3　タスク定義

本稿では検索場面を取り上げて，文脈依存的に変容する画像の意味理解を扱う．本節ではまず基本の問題設定の定式化を行い，次節にて具体的な分析フレームワークやデータセット構築方法，そして，文脈を加味した検索手法の提案を行う．

問題設定　画像集合 $\mathcal{I} = \{I_1, I_2, ..., I_n\}$ と各画像に紐づくコンテキスト C_I が与えられたとき，テキストクエリ q に対して最も適切な画像 I^* を選択する：

$$I^* = \arg\max_{I \in \mathcal{I}} f(I, q, C_I) \tag{1}$$

ここで f はコンテキスト C_I を加味してクエリ q に対する画像 I の適合度を返すスコア関数である．

データセットの要件　本タスクを評価するためのデータセットを以下のように定義する．画像集合 $\mathcal{I}$，コンテキスト集合 $\mathcal{C}$，クエリ集合 $\mathcal{Q}$ を用意し，各画像に対応するコンテキストを返す写像を $\pi : \mathcal{I} \to \mathcal{C}$ とする（$C_I = \pi(I)$）．1 つのコンテキストに複数の画像が紐づきうる．評価用の正解データはタプル $(I, C, q) \in \mathcal{I} \times \mathcal{C} \times \mathcal{Q}$ の集合 $\mathcal{D}$ として定義され，各タプルは $C = \pi(I)$ を満たす．

4　提案手法

4.1　L1-L4 抽象度フレームワーク

先行研究は客観的内容の認識，注目点の切替，状況・意図の推論，雰囲気・効果の推定といったさまざまな観点から画像意味を扱ってきた．本研究ではこれらを抽象度の異なる 4 段階（L1-L4）に整理する．L1 は画像から直接読み取れる客観的事実（物体・行動）であり，コンテキストに依存しない．L2 は視覚的フォーカス（注目点）であり，同一画像でも文脈により着目する要素が変わる．L3 は状況と意図であり，画像単体からは一意に定まらない文脈に基づく解釈である．L4 は雰囲気と効果であり，画像の受け手が持つ印象や知覚である．図 1 の画像を例にとると，L1 は「窓辺に立つ人影」と文脈によらず共通である．一方，ロマンス文脈では L2 が「灯りの漏れる窓」，L3 が「家族の帰りを待つ」，L4 が「温かい，安らぎ」となるのに対し，サスペンス文脈ではそれぞれ「暗がりのシルエット」「ターゲットを監視する」「不穏，緊張」となる．

4.2　データセット構築

タスク定義の要件を満たすデータセット $\mathcal{D}$ を以下の手順で構築する．まず，画像集合 $\mathcal{I}$ から複数の画像をグループ化し，各グループに対して LLM を用いて複数の異なる物語（コンテキスト C）を生成する．次に，各画像とコンテキストの組に対して，L1-L4 の定義に基づきクエリ q を生成する．同一画像に対して異なるコンテキストを付与することで，

タイトル： ヴァンス事件
ストーリー： 保険調査員のジェフ・キンケイドは，失踪した実業家と，その魅惑的な未亡人デライラが渦巻く不透明な世界へと引き込まれていく．調査を進めるうちに嘘と欺瞞を暴いていくジェフだったが，いつしか彼自身も，強欲と策謀が張り巡らされた網の目に取り込まれていく．

	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5
L1	背後から見た二人の男，オフィスの扉．	女性，二人の男，ソファ，タバコ．女性が座っている．	女性，ポーター，駅のホーム．女性が列車に乗り込む．	建物内の男女．向かい合い，彼女が彼の腕に手を伸ばす．	二人の男．一人はタバコ，もう一人はライターを持つ．
L2	出口のドアに向き合うジェフ．	デライラの優雅なたたずまいと計算された眼差し．	列車のステップを上がるデライラ．	デライラの必死とした眼差しと伸ばされた手．	ジェフのタバコに火を灯すサリー．
L3	ハリソンが指示を出し，キンケイドは冷笑的に立ち去る．	デライラは潔白を装い，反抗的に受け流す．	彼女は逃亡し，欺瞞を続けることを目論む．	ジェフは求めた真実を，デライラは追及を逃れようとする．	サリーが支えを申し出，ジェフは苦い慰めを見出す．
L4	冷笑的，不穏，フィルム・ノワール．	緊迫，神秘的，魅惑的．	緊張，回避的，クライマックス．	緊張，疑念，ドラマチック．	陰鬱，冷笑的，倦怠．

表 1　Scene 1-5 の画像と生成ラベル（L1-L4）

コンテキスト依存性を検証可能なタプル (I, C, q) の集合が得られる．

本研究では，画像ソースとして LSMDC [12]（映画シーン）と MS COCO [13]（一般画像）を使用し，各 50 グループ（1 グループ 5 画像，計 500 画像）を選定した．ストーリーおよびクエリの生成には Gemini 2.5 Flash を用いた．表 1 に実際の生成例を示す．

作成されたデータセットを用いて，文脈の切り替えによるクエリの変化を測定した．全 500 画像について 2 つのストーリー間でのクエリの意味的距離を算出した（図 2）．L1 の平均距離は 0.18–0.25 と低く，ストーリーに関わらずほぼ同じクエリが生成された．一方，L3 では 0.48–0.52 と最も高い分岐を示し，同一画像でもストーリーによって全く異なる「状況・意図」が記述されることが確認された．この傾向は，抽象度が高いほど文脈依存性が高まるという想定と一致しており，L1-L4 フレームワークの妥当性を支持する結果である．

4.3　検索手法

本研究では埋め込みベースの手法を用い，スコア関数を以下で実現する：

$$f(I, q, C_I) = \mathrm{sim}(\phi(I, C_I), \psi(q, C_I)) \qquad (2)$$

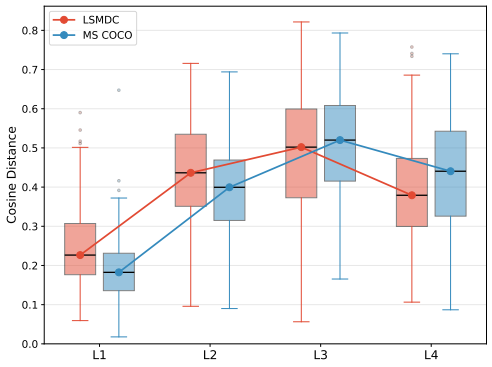

図 2　抽象度レベル別のストーリー間ラベル距離．L1 はストーリーに依存せず安定，L3-L4 は大きく分岐する．

ここで $\phi(I, C_I)$ はコンテキストを考慮した画像埋め込み，$\psi(q, C_I)$ はクエリ埋め込み，sim はコサイン類似度である．手法によっては C_I を用いない場合もある．本タスクに対して現状利用可能な手法を検証するため，以下の手法を比較する．

CLIP　CLIP は画像とテキストを共通の埋め込み空間に写像する．Dual-Encoder 構造により画像埋め込みは $\phi(I)$ として固定され，コンテキストを考慮できないため，本タスクのベースラインとして位置づける．

VLM-Caption　VLM で画像に対するキャプションを生成し，そのテキストをテキストエンコーダで埋め込みに変換する手法である．

VLM-Embed　CLIP の dual-encoder 構造とは異なり，VLM は画像とテキストを統合的に処理できる．E5-V [10] と同様に，VLM の最終層における最終トークン位置の隠れ状態を埋め込みベクトルとして抽出し，L2 正規化を適用した．

VLM-Caption および VLM-Embed では，キャプション生成時や埋め込み生成時にコンテキストをプロンプトとして与えた．これにより画像埋め込みがコンテキストに条件付けられ，同じ画像でも異なるコンテキストで異なる埋め込み・キャプションが生成される．

5　実験

実験設定　評価指標として Recall@1 を使用し，LSMDC と MS COCO の各データセット 250 画像の候補プールから正解画像を検索するタスクで評価した．CLIP として EVA-CLIP-8B/18B [14], SigLIP [15], OpenAI CLIP [3], OpenCLIP-H14 を使用した．VLM-Embed には Gemma3-4B/12B [16], Mistral3-3B/8B，Qwen3-VL-2B/4B/8B [17] を使用した．VLM-

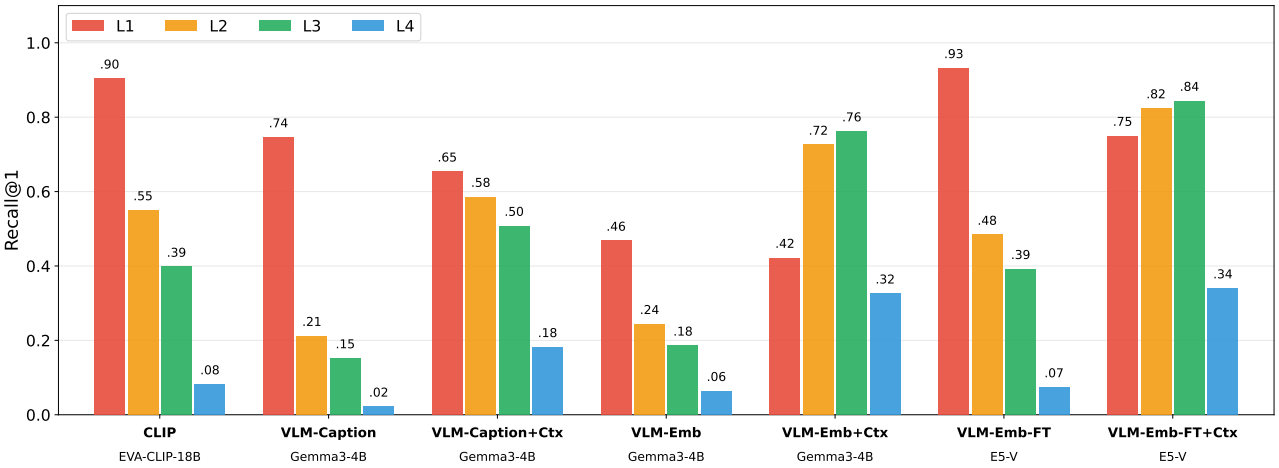

図 3　各手法の最高性能モデルにおける抽象度レベル別検索性能（LSMDC, Recall@1）．CLIP は L1 で優れるが高抽象度で劣化．VLM-Emb+Ctx が L2-L4 で最高性能．

Caption には Gemma3-4B/12B，Qwen3-VL-2B/8B をキャプショナーとして，Qwen3-Embedding-0.6B をテキストエンコーダーとして使用した．VLM-Embed では Gemma3 と Mistral3 のベースモデルと Instruct モデルの両方を評価したが，ベースモデルがより良い結果を示したため，以降ではベースモデルの結果を報告する．画像-テキスト埋め込み用にファインチューニングされた VLM（VLM-Emb-FT）である E5-V [10]，VLM2Vec-2B/7B/v2 [11] も評価した．VLM-Caption，VLM-Embed，VLM-Emb-FT については，コンテキストあり/なしの両条件を評価し，コンテキスト注入の効果を検証した．

結果　図 3 に各手法の最高性能モデルの結果を示す．CLIP は抽象度が上がるにつれて性能が急激に低下し，L1 の 0.90 から L4 では 0.08 まで 91%低下した．一方，VLM-Caption と VLM-Embed はコンテキスト注入により，L2-L4 で大幅な改善を達成した．特に VLM-Embed の Gemma3-4B では，L3 の Recall@1 が 0.19 から 0.76 へと 301%改善した．映画単位でのクラスタブートストラップ（n = 10,000）による検定の結果，L2-L4 での改善は全て統計的に有意であった（$p < 0.001$）．一方，L1 ではコンテキスト注入による有意な差は見られなかった（$p = 0.17$）．この結果は，コンテキストが抽象的意味（L2-L4）の検索に選択的に効果を発揮することを示している．なお，Recall@5 や他のモデルでも同様の傾向を確認した．MS COCO 由来のデータセットでも同様である．

VLM-Caption もコンテキスト付与で改善するが，VLM-Embed が一貫して上回る．この差の理由は，キャプション生成時に視覚情報が言語に変換される際の損失がないことと，コンテキスト情報を中間表現を介さず埋め込み空間で直接反映できることであると考えられる．

表 2　コンテキスト注入パターンの比較（Gemma3-4B, LSMDC, Recall@1）

パターン	画像	クエリ	L1	L2	L3	L4
No-Ctx	$\phi(I)$	$\psi(q)$	.464	.248	.190	.054
Query-Ctx	$\phi(I)$	$\psi(q, C)$	**.504**	.400	.372	.310
Image-Ctx	$\phi(I, C)$	$\psi(q)$	.422	**.726**	**.762**	**.326**
Both-Ctx	$\phi(I, C)$	$\psi(q, C)$	.522	.430	.388	.310

5.1　コンテキスト注入パターンの分析

VLM-Embed におけるコンテキストの注入方法として 4 パターンを比較した（表 2）．Image-Ctx が最も効果的であり，他のモデルでも同様の傾向であった．これはクエリが明確な意図を持つ一方，画像の「意味」は本質的に多義的であり，コンテキストは曖昧な側（画像）に注入すべきためと考えられる．

6　結論

本研究では，文脈依存画像検索という新しいタスクを提案した画像の意味が文脈によって変化するという課題を定式化し，意味の抽象度に着目したフレームワークを導入した．構築したデータセットを用いた分析により，抽象度が高くなるほど文脈による意味の振れ幅が大きくなることを明らかにした．現状の手法の本タスクへの適用可能性を検証し，コンテキスト注入によって抽象的意味の検索が改善しうることを示した．本研究は文脈を考慮した画像理解という新たな研究領域の基盤を提供する．

参考文献

[1] Alfred L Yarbus. **Eye movements and vision**. Springer, 2013.

[2] Dean Mobbs, Nikolaus Weiskopf, Hakwan C Lau, Eric Featherstone, Ray J Dolan, and Chris D Frith. The kuleshov effect: the influence of contextual framing on emotional attributions. **Social cognitive and affective neuroscience**, Vol. 1, No. 2, pp. 95–106, 2006.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PmLR, 2021.

[4] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In **European conference on computer vision**, pp. 69–85. Springer, 2016.

[5] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 5503–5512, 2017.

[6] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In **European Conference on Computer Vision**, pp. 508–524. Springer, 2020.

[7] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn Schuller, and Kurt Keutzer. Affective image content analysis: A comprehensive survey. 2018.

[8] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In **Proceedings of the 21st ACM international conference on Multimedia**, pp. 459–460, 2013.

[9] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 11569–11579, 2021.

[10] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. **arXiv preprint arXiv:2407.12580**, 2024.

[11] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. **arXiv preprint arXiv:2410.05160**, 2024.

[12] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. **International Journal of Computer Vision**, Vol. 123, No. 1, pp. 94–120, 2017.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In **European conference on computer vision**, pp. 740–755. Springer, 2014.

[14] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. **arXiv preprint arXiv:2303.15389**, 2023.

[15] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In **Proceedings of the IEEE/CVF international conference on computer vision**, pp. 11975–11986, 2023.

[16] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. **arXiv preprint arXiv:2503.19786**, 2025.

[17] Shuai Bai, et al. Qwen3-vl technical report. **arXiv preprint arXiv:2511.21631**, 2025.