H-015

学習データ外事象の説明が可能な Zero-shot 差分キャプショニング技術の検討
Zero-shot Change Captioning for Unseen Events

佐藤 拓杜 ‡　大橋 洋輝 ‡
Takuto Sato　Hiroki Ohashi

1. はじめに

差分キャプショニングは入力された 2 枚の画像データ間の差分を抽出し，その差分情報の説明を自然言語で生成する技術である。作業現場には巡回点検における設備の目視確認や作業前後の復旧確認など，差異を確認して報告するといった工程が多い。こういった作業を自動化するための技術として差分キャプショニングが有望視されている。

しかしながら文献[1]に代表される従来の差分キャプショニング技術は，検出機構と文章生成機構が End-to-End で学習されるため，学習データに含まれていない物体は検出できない上，学習データに含まれていない単語を含む文章も生成不可能といった制約が発生する。現場の確認作業では何が飛んでくるか予測不能な飛来物の対処なども含まれるため，そういった作業に対して差分キャプショニングを使用するには，事前に起こり得ると予想される限りの事象をデータとして収集し，学習させたモデルを作成しておくなどの困難さが生じる。

そこで本研究では，事前に学習されていない事象の差分に関しても検出・キャプショニング可能な能力を持つ，Zero-shot 差分キャプショニング技術について基礎検討を行った。技術構成のうち 2 枚の画像から差分を検出する部分については，大規模な差分合成画像データセットにより学習されることで強力な汎化性能を持つことが示された差分検出手法[2]を用いることとし，本研究では特に差分検出結果から Zero-shot で差分説明文の生成を行うデコーダ機構の実現性に関しての検討を行った。最も単純な差分説明文は{変化検出物を表す名詞句}と{変化種別を示す動詞句}の 2 句から構成されること，また各句が多少抽象的な表現であっても人間は事象を想起可能であることに着目し，抽象的な差分説明文の生成手段を検討した。未知事象のみで構成された評価データセットにおける比較実験を行い，生成文章に対する種々の評価指標で提案手法が従来手法よりも優れた性能を示すことを確認した。

2. 関連研究

2.1 Change Captioning

文献[1]は，差分検出を行う画像エンコーダおよび差分キャプションを生成する自然言語デコーダの組み合わせという本分野での基本的なアーキテクチャと，CLEVR-Change という人工的に生成された差分画像・差分キャプションデータセットによる評価方法を提案した。2023 年までに画像間の強い視点変化への対応[3]や，画像内に複数の変化を含むケースへの対応[4]など，Change Captioning の様々な課題に対する発展的手法が提案されている。

これまで本分野での手法の評価方法は，文献[1]などで提供された公開データセットを学習・テストへ分割して実施されてきた。従って学習とテストのドメインが異なる状況，

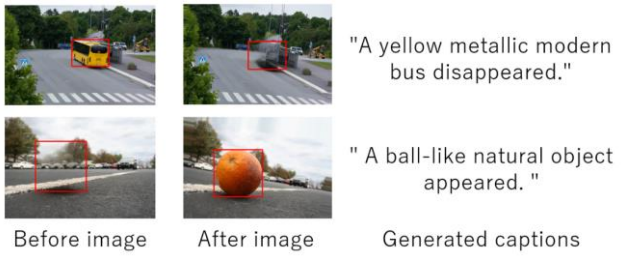


図 1　Zero-shot Change Captioning の生成例

つまりテストデータにおける差分検出対象や対象を表す語彙が未知である Zero-shot 状況での評価は，知る限りでは実施されていない。本稿の評価実験により，そういった Zero-shot 状況においては，特定のデータセットで学習された差分キャプショニング手法は検出・記述の両能力に問題が発生することを確認した。本研究の目的はそういった Zero-shot 状況でも機能する差分キャプショニング技術を構築することである。

2.2 Zero-shot Image Captioning

単一画像に対する Zero-shot Image Captioning の研究事例が存在する。BLIP[5]は，約 1.3 億件の画像-キャプションペアデータセットから学習された Vision-Language モデルである。ほかにも，画像-キャプションペアのデータで学習することを行わず，高い汎化能力が実証されている Vision-Language モデルである CLIP[6]を転用する形で Zero-shot Image Captioning を実現する研究もある。たとえば ZeroCap[7]は，CLIP の画像エンコーダから得た勾配情報を利用して言語モデルへの入力を操作することで Zero-Shot Image Captioning を行う方法について提案している。また ClipCap[8]は，CLIP の画像エンコーダの出力を言語モデルに入力する Token へ変換する Mapping Network を利用する方法を提案している。

これらの手法の能力は膨大な学習データに支えられているものの，大半が Web から収集されたデータであるため，キャプショニング可能な画像や語彙にはドメインの偏りがある。たとえば CLIP は ImageNet に出現するクラスのような一般物体への対応力は非常に高いが，産業の現場で使われる工具などを正確に記述できないケースが多いことを予備実験で確認している。Zero-shot Change Captioning の方向性としては，こういった得意不得意の偏りなく，多くの事象をまんべんなく説明できる能力の実現が望ましい。

ここで人間の能力に立ち返ってみると，人間は具体的な名称を知らずとも，「黒くて丸い物体」「茶色い光沢のある液体」「黄色くて背の高い動物」といった抽象的な表現で説明をすることができる。説明の聞き手にとっても，与えられた説明と背景知識と組み合わせて具体的な物体を想起することは不可能ではない。そこで本研究の Zero-shot

Change Captioning の方向性としては，事象を具体的に説明する能力を備えるのではなく，説明の受け手の中で想起が起きうる程度に抽象的な説明を生成することを目指した。

3. 提案手法

3.1 全体構成と検討要素

提案手法の全体構成を図 2 に示す。提案手法は一般的な差分キャプショニング技術と同様に，2 つの画像から差分領域を抽出するエンコーダと，差分領域に関する説明文を自然言語で出力するデコーダから構成される。1 章で前述した通り，エンコーダに関しては本稿では手法[2]を利用する。デコーダへの入力は，2 枚の入力画像および各入力画像上で検出された変化領域を囲う Bounding Box 情報である。

差分説明文の最も単純な構成は，{変化検出物の名詞句}と{変化種別を示す動詞句}の 2 句からなる文章である。ここに画像内の位置関係や検出物の所有関係を示す句を追加していくことで説明の情報量は増していくが，まずは不可欠構成な要素である名詞句・動詞句を推定することを本稿の検討対象とする。

名詞句は，句の主役たる名詞と，それを限定・修飾する複数の句から成る。修飾句の例としては，Zero-shot Recognition の分野でしばしば用いられてきた色・形状・材質といった Attribute 表現や，輝いている・腐っているといった State 情報があげられる。名詞そのものが Thing や Object といった抽象的な単語でも，適切な修飾によって情報量を増すことで想起現象の発生確率を上げられると考えられる。名詞句の構成に関する課題は，名詞そのものや Attribute・State 表現の獲得，そしてそれらの適切な組み合わせ方の求め方である。その推定方法を 3.2 節で述べる。

一方で動詞句については，高抽象度の変化分類として ADD・DELETE・MOVE・REPLACE の 4 つが使われることが多く，本検討でもこれらを踏襲する。ADD は対象が出現する変化，DELETE は対象が消滅する変化，MOVE は対象が画像内を移動する変化で，REPLACE は対象が別のなにかに置き換わる変化である。課題となるのは，説明対象がこれら 4 つの変化のいずれに当てはまるか推定する方法である。この変化種別の推定結果が誤っていると，仮に完璧に正しい名詞節の構築ができていたとしても変化を説明する文章として全く不適な内容になりえる。従来の変化キャプショニング手法では，変化種別の推定能力はデータからの学習によって獲得していたうえ、学習・テストデータの背景は必ず灰色領域であるという単純な内容であった。一方で目的である Zero-shot 状況においては，背景の様子が未知かつ多様になるため，均一で簡単な背景条件下で学習された変化種別の識別能力は機能しない。本課題に対する解決方法として近年急速に発展してきた画像の inpainting 処理に基づく変化種別推定方法を 3.3 節で説明する。

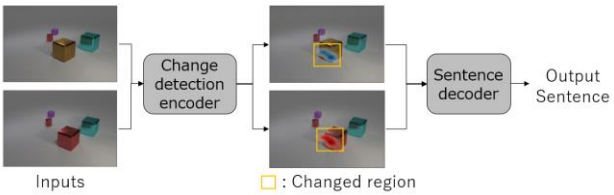


図 2　提案手法の構成

3.2 抽象的名詞句の獲得

3.2.1 Attribute・State の抽出方法と各構成要素の定義

本研究における画像から Attribute と State を抽出する手段については、CLIP を使った Concept Prompting と呼ばれる手法[9]を採用する。この手法の概要を図 3 に示す。まず、CLIP の text エンコーダに入力する prompt を抽出対象の Attribute・State ごとに作成する。そして prompt と抽出元の画像を CLIP の画像・text エンコーダを通じて生成した特徴ベクトルに変換したのち、最も画像特徴ベクトルと cosine 類似度が高い prompt に含まれていた Attribute・State を最終的な抽出結果とする。この手法の利点は，学習データを Attribute・State ごとに事前収集する必要がなく、抽出したい Attribute・State の再定義を CLIP に与える prompt の変更で実現できる点である。可能な限り広く多様な Attribute・State を抽出するに当たってこの手法の利便性は高い。なおこの抽出操作は，後述する色・形状・材質・質感など Attribute・State のカテゴリごとに実施し，各カテゴリで 1 つずつの代表単語を抽出する。

次に名詞句の各構成の定義を与える。まず名詞に関しては，広く一般的な事象を表現するために，一般生活で使用される名詞を抽象的すぎず具体的すぎない範囲で含むことが理想的である。そこで本研究では，あらゆるカテゴリの

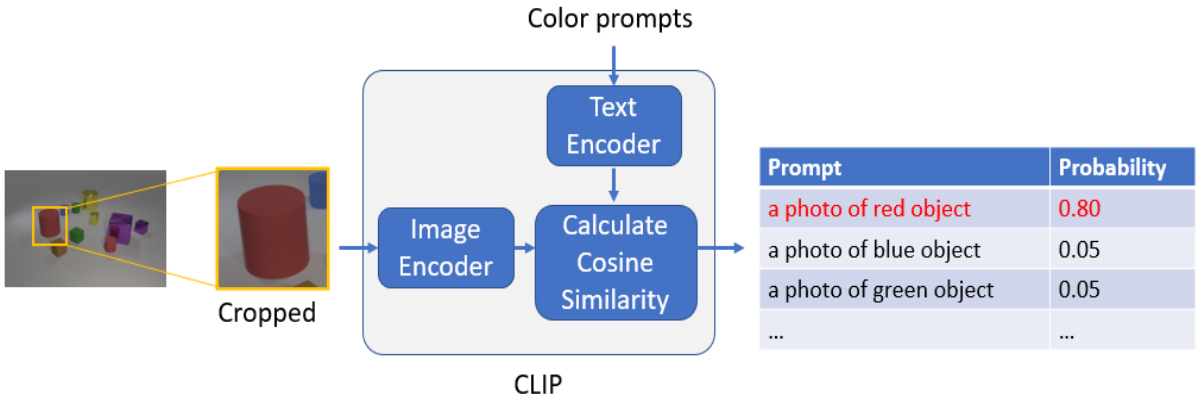


図 3　Concept Prompting の概要

代表的な集合としてのデータセット構築を目指した MS-COCO[10]/COCO-Stuff dataset[11]における名詞ラベルを踏襲する。この名詞ラベルは WordNet[12]を元に階層的に作成されている。たとえば，最も抽象度の高い階層に”thing”があり，その 1 階層下に outdoor/indoor，そのさらに下に outdoor であれば animal や vehicle，indoor であれば appliance や food などの単語が収まっている。この階層的名詞ラベルリストから前述の Concept Prompting により最も画像を表現するに適した単語を抽出するが，その際に things や stuff という抽象度の高い単語として常に候補に含むことで，語彙を知らない未知の対象においても全く関係のない名詞が選択されることを防ぐ。

次に Attribute と State の定義について述べる。Attribute は性質に関する修飾を，State は状態に関する修飾を行う単語とする。Attribute については色，形，材質，質感に関する言葉をそれぞれ保持する。State については画像から各種状態を推定する検討を行った文献[13]での単語リストを参考に，新古，清潔さ，温度，新鮮さなど広い名詞に当てはまる状態を指す単語を利用する。

3.2.2 抽象的名詞表現のための構成要素決定方法

Attribute や State は多くのカテゴリを有するため，抽出したすべての単語を使って名詞を修飾すると，たとえば red-round-natural-shiny-fresh-apple といったように冗長な表現となる。また，対象名詞の修飾に向いていないカテゴリからも必ず 1 つの単語が抽出されるため，単純に抽出単語を使って修飾すると説明の受け手を混乱させる可能性が高い。したがって，抽出単語の中から適切なものを選抜することが重要となる。

本検討では，この課題についても前述の Concept Prompting の応用で解決を試みる。たとえば先の red-round-natural-shiny-fresh-apple の例においては，名詞である apple をベースに fresh-apple, shiny-fresh-apple, red-shiny-apple といった具合に全単語の組み合わせを表現候補とし，それぞれを prompt テンプレート”a photo of <>”に当てはめて prompt 化する。そして Attribute・State 抽出時と同様に，CLIP の text エンコーダを通じて特徴ベクトルに変換する。そして最終的には，変化領域の画像の特徴ベクトルと各 prompt の特徴ベクトルの cosine 距離を指標に最も高い類似度を計測した prompt に採用されていた候補を最良の修飾表現の組み合わせであったとし，最終的な抽象的名詞節として採用する。この操作は，大規模に学習された VL モデルのドメインの偏りは抽象表現に大きな影響を与えないという仮定のもとで行うものであるが，実際に期待される修飾語選別能力が pretrained の VL モデルに備わっているかは実験の章で評価する。

3.3 変化種別の推定

変化部分として検出された画像領域の情報から，その変化が代表的種別である ADD・DELETE・MOVE・REPLACE のいずれかを推定する方法について説明する。

推定の鍵になるのは，変化領域の画像が背景のみを映したものなのか否かである。背景判定の情報があれば ADD か DELETE かそれ以外かが確定でき，残る MOVE と REPLACE については，変化領域から抽出する各 Attribute・State の情報が変化前後の画像で一致しているかで識別が可能である。そのため，いかにして変化領域の背景判定を実施するかが検討のポイントとなる。

直感的に考えれば，変化領域の画像が背景のみを映したものであるのかの判定は，変化領域とその領域外との連続性・一貫性の観点で行える。つまり，変化領域の画像の構成や意味情報が領域外と一定以上合致するならば背景，合致しないならば変化領域内部に特異な情報として物体を含んでいるとみなす，という判定処理である。

本研究では，上記の考えに基づく判定処理を mask 領域の Inpainting を利用して行う方法について考案した。変化検出領域を mask 領域とし，手法[14]を用いて領域外の情報をもって mask 領域内を Inpainting する。Inpainting 処理は mask 領域外と mask 領域が違和感なく調和するように mask 領域内の内容生成を行うので，領域外に広がる背景情報が mask 領域内に復元される可能性が高い。この復元された

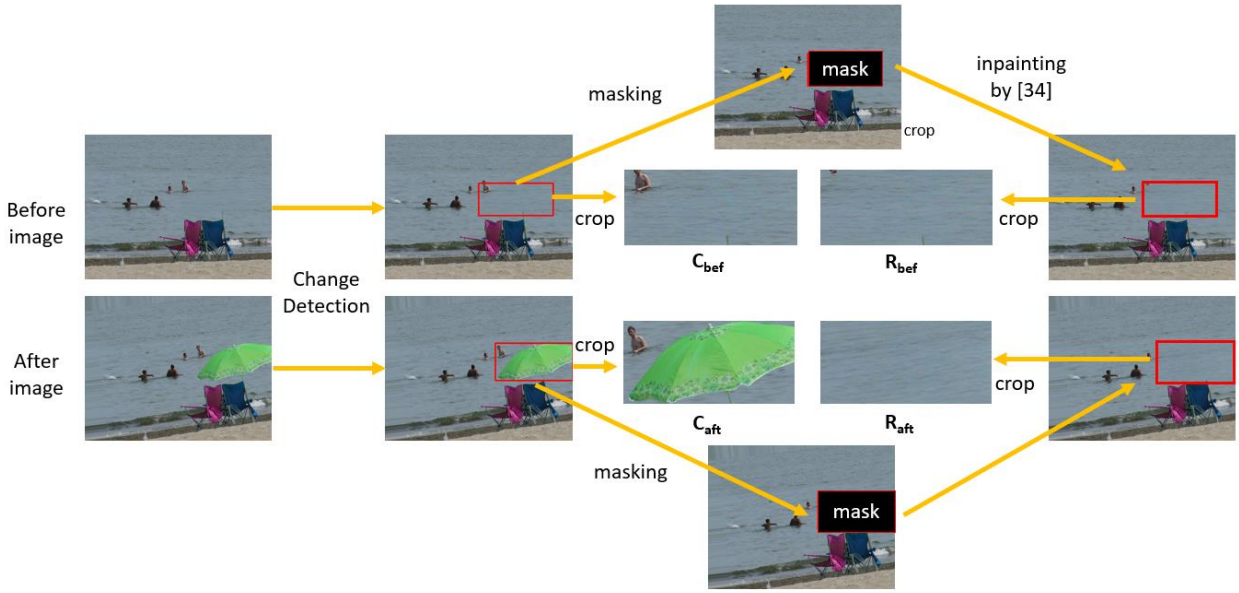


図 4　変化種別推定のための変化領域画像(C_{bef}, C_{aft}, R_{bef}, R_{aft})の生成

mask 領域画像を代表的背景画像とし，もともとの変化検出領域画像と内容を比較することで，変化検出領域が背景画像なのか否かを判定することができる。この処理の流れを図 4 に示す。内容比較は各画像を CLIP の画像エンコーダを通じて特徴ベクトルに変換し，その Cosine 類似度が一定の閾値以上であれば一致，閾値を下回れば不一致として行う。

背景判別手法を使った変化種別の推定方法を表 1 に整理する。変数として，取得時刻が若いほうの画像から切り出した変化検出領域を C_{bef}，あとの画像から切り出した変化検出領域を C_{aft}，Inpainting 処理によって復元した各変化検出領域を R_{bef}，R_{aft} としている。また，同様の表記を Attribute に適用し，取得時刻が若いほうの画像から抽出した Attribute 一式を A_{bef}, あとの画像から抽出した Attribute 一式を A_{aft} とする。

表 1: 変化種別の推定ルール

Change Type	Rule between C and R	Rule at extracted attribute
ADD	(Cbef ＝ Rbef) and (Caft ≠ Raft)	-
DELETE	(Cbef ≠ Rbef) and (Caft ＝ Raft)	-
MOVE	(Cbef ≠ Rbef) and (Caft≠ Raft)	Abef ＝ Aaft
REPLACE	(Cbef ≠ Rbef) and (Caft ≠ Raft)	Abef ≠ Aaft
NO CHANGE	(Cbef ＝ Rbef) and (Caft ＝ Raft)	-

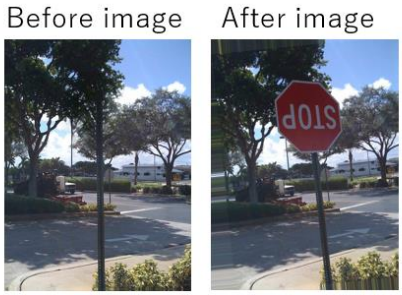
4. 評価実験

4.1 実験設定

本評価では従来手法[1]と提案手法について，どちらの手法の学習データにも含まれていない未知のデータに対する Zero-shot 状況での Change Captioning 能力について評価する。

本研究では，Change Detection の評価用に手法[2]で作成された COCO-inpainted データセットのテストデータセットの一部を利用し，ADD・DELETE・MOVE・REPLACE の 4 変化を新たに合成した上で，各変化キャプションをアノテーションすることで新たな Change Captioning 用評価データセットを構築した。画像ペア間にはランダムな affine 変換を適用し，画像間の疑似的な視点変化を生じさせる。なお基礎検討であることを鑑み，1 つの画像ペアには 1 つの変化のみを含めた。そして本実験において 1 つの変化画像ペアから複数の変化対応関係が検出された場合は，最も確信度が高い変化に限定してキャプションを生成する。複数の同時変化への対応は Future Work とする。

図 5 は STOP という標識を inpainting することにより ADD の変化ペア画像を生成した例である。各画像ペアの正解キャプションに関しては，具体的・少し抽象的・最も抽象的な表現という具合に 3 種類ずつ抽象度を変えてアノテーションを施した。4 種類の変化に対して，各 30 画像ペア 90 キャプション生成し，合計 120 枚 360 キャプションからなるデータセットを構築した。



(i) “A sign marked STOP has emerged.“
(ii) “The red octagonal sign is added.”
(iii) “The red object appeared.”

図 5　COCO-inpainted への Annotation 例

提案手法の中で利用する既存手法に関して，差分検出手法[1]および inpainting 手法[14]，CLIP[6]は公開されている pretrained モデルを用いる。CLIP のアーキテクチャは ViT-B/16 で，変化種別推定で利用する CLIP 特徴量の Cosine 類似度閾値は 0.85 に設定する。

評価指標としては METEOR[15]，CIDEr[16]，SPICE[17] の 3 つを利用する。なお本評価においてデータセットにはそれぞれ抽象度の異なる正解アノテーション文が 3 つずつ含まれているため，最大の評価値を得たアノテーション文に対する結果を採用する。

4.2 実験結果

提案手法の結果サンプルを図 6 に示す。上段は正確にキャプショニングできたと思われるもの，中段は一部間違っていて一部合っているものであり，下段は変化種別の推定に失敗した例を示している。

表 2 は各評価指標において生成キャプションの精度を従来手法[1]と提案手法で比較した結果である。それぞれの指標で提案手法の評価値が上回っていることが確認できる。また，表 3 は手法前段の差分検出処理において正しい差分領域の抽出が行えた割合を示している。従来手法はどの変化タイプにおいても 5 割前後の検出成功率となっている一方で，提案手法は MOVE 以外の変化種別で高い精度で差分検出が行えている。

図 7 は変化種別推定の混合行列である。従来手法のほうは，真の変化種別が何であっても変化なしと誤認識する傾向が強い。一方で提案手法のほうは，差分検出の傾向と同様に MOVE 以外の変化種別で高い精度の認識ができている。

表 2: 従来手法と提案手法のキャプション精度の比較結果

	METEOR[19]	CIDEr[20]	SPICE[21]
従来手法 [1]	0.07	0.02	0.01
提案手法	0.18	0.36	0.13

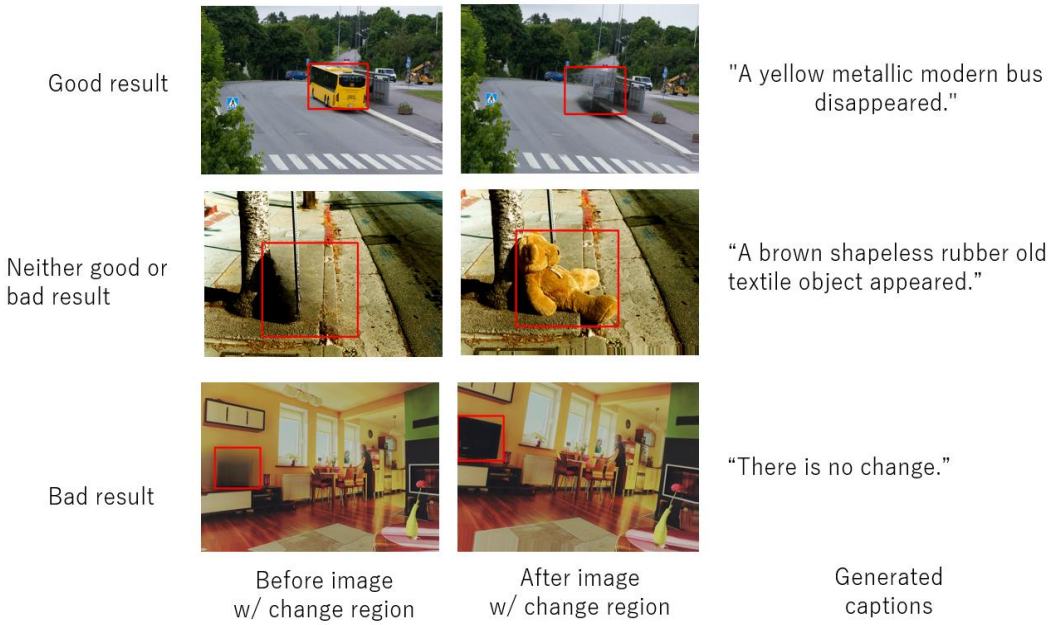


図 6　提案手法の結果例

表 3: 従来手法と提案手法の差分領域検出精度の比較結果

	ADD	DELETE	MOVE	REPLACE
従来手法 [1]	0.50 (15/30)	0.60 (18/30)	0.30 (9/30)	0.47 (14/30)
提案手法	1.00 (30/30)	0.97 (29/30)	0.50 (15/30)	0.93 (28/30)

4.3 考察

まず提案手法の差分検出精度および変化種別推定精度が MOVE だけ悪い件について考察する。MOVE の場合，差分領域を囲う Bounding Box が，before 画像では対象の移動前の位置に，after 画像では移動後の位置に出現することが理想の差分検出結果となる。しかし MOVE の本実験結果においては，移動前後のどちらかの位置にのみ Bounding Box が出現する事例を多く観測した。差分検出の段階で正しい Bounding Box の位置推定に失敗する場合，自動的に後段の変化種別推定にも失敗する。差分検出で正しい位置推定に失敗する原因に関しては，MOVE の変化と，ADD と DELETE の複合的同時変化の区別が本質的に難しいことがあげられる。また，提案手法で利用した差分検出手法[1]に関しては ADD・DELETE 変化のみの差分合成データセットで学習されており，MOVE 変化を ADD・DELETE と捉える傾向が強いことも要因と考えられる。なお MOVE の位置推定は合っているが変化種別推定で失敗するケースについては，REPLACE に誤認識する数が多い。これは MOVE 前後の背景の違いや MOVE 対象の向きの違いなどの影響により，before-after 画像間での抽出 Attribute・State の一致判定が誤るためである。今回は完全一致の場合のみ MOVE と判定する条件であり，より柔軟な判定方法が求められる。

次に MOVE 以外の変化種別の誤推定に関して考察する。変化種別の推定で重要な要素は 3.3 節で述べた背景判定の精度である。比較的多かった誤認識パターンとして，inpainting しきれず変化なしと誤検知したもの，inpainting が元画像に含まれないものを生み出した結果 REPLACE と誤検知したものがあげられる。図 6 の下段は変化なしの誤認識ケースの典型例であり，黒い TV の影響が壁として inpainting しきれず残った影響を受けている。一方で REPLACE のケースのように inpainting が逆に変化を作り出してしまうパターンは画像内容や切り出し位置の影響による不確実性が高く，回避が難しいと予想される。より安定的な背景判定手法の構築は今後の課題の 1 つである。

最後に Attribute や State の推定・選定に誤りがあったケースについて考察する。図 8(a)(b)は不適な例であり，(b)などは flexible と solid とほぼ逆の意味の単語が修飾語として含まれている。solid のほうは一緒の画像内に含まれている白い椅子の影響を受けた可能性があり，CLIP が画像内のどこから情報を拾ってきているかについては詳しい検討が必要である。また図 8(b)における matte や modern が該当するが，抽象的名詞表現の最終決定処理では全体的な傾向として CLIP は概ね意味が合っていそうな形容詞は排除せずそのまま残す傾向にある。たとえば室内の家具の画像に対して、高頻度で”dry”を残す。この観点で言えば，今回の検討で実施した CLIP による Attribute・State の選定は予想外の単語が使われる傾向があり，うまく機能しなかったといえる。この Attribute や State から変化事象を説明するにふさわしいものを選定する技術は改善の必要がある。

5. おわりに

本報告では，学習データに含まれていない未知の差分事象に対応可能な，Zero-shot 差分キャプショニング手法について検討した。特に差分領域の情報を受け取って説明を生成するデコーダについて，抽象的名詞表現と変化種別推定

からなる手法を検討し，評価を通じて種々の課題を発見した。Future Work としては，画像内に複数の変化が含まれているパターンへの対応，変化を示す動詞の拡張，微細な環境変動が含まれた実環境取得データでの検証があげられる。また，どの程度の抽象度であれば差分説明文から事象を想起しうるかに関しての主観評価も実施する予定である。

参考文献

[1] Park, Dong Huk, Trevor Darrell, and Anna Rohrbach. "Robust change captioning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[2] Sachdeva, Ragav, and Andrew Zisserman. "The Change You Want to See." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023..
[3] Shi, Xiangxi, et al. "Finding it at another side: A viewpoint-adapted matching encoder for change captioning." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer International Publishing, 2020.
[4] Qiu, Yue, et al. "Describing and localizing multiple changes with transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021..
[5] Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." International Conference on Machine Learning. PMLR, 2022.
[6] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
[7] Tewel, Yoad, et al. "Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022..
[8] Mokady, Ron, Amir Hertz, and Amit H. Bermano. "Clipcap: Clip prefix for image captioning." arXiv preprint arXiv:2111.09734 (2021).
[9] Tian, Yun, et al. "Do Vision-Language Pretrained Models Learn Composable Primitive Concepts?" Transactions on Machine Learning Research, 2023.
[10] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014.
[11] Caesar, Holger, et al. "Coco-stuff: Thing and stuff classes in context." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[12] Princeton University "About WordNet." WordNet. Princeton University. 2010.
[13] Isola, Phillip, Joseph J. Lim, and Edward H. Adelson. "Discovering states and transformations in image collections." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[14] Suvorov, Roman, et al. "Resolution-robust large mask inpainting with fourier convolutions." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.
[15] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005..
[16] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[17] Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer International Publishing, 2016.

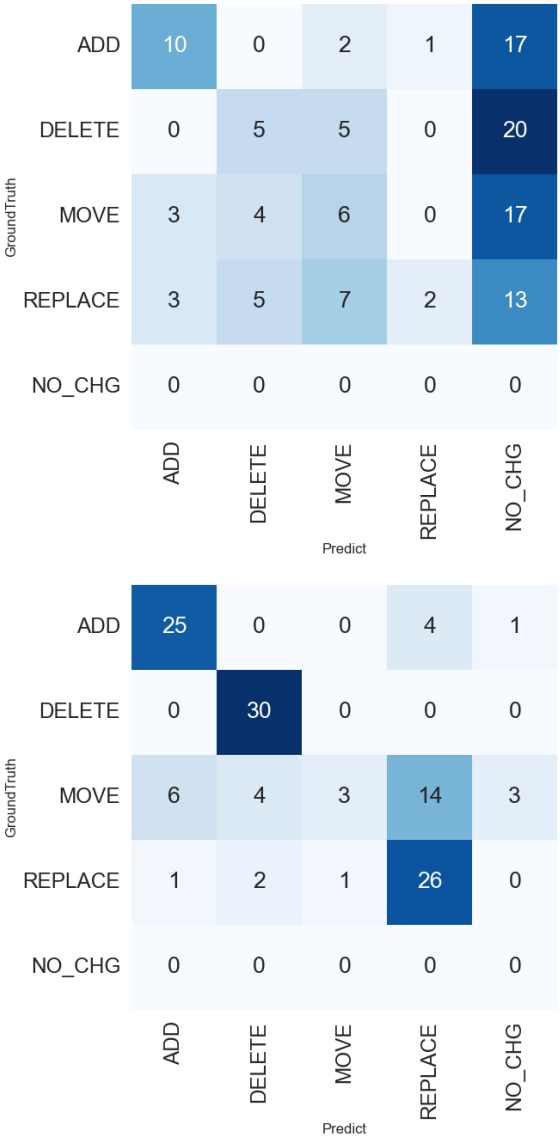


図 7　変化種別推定の混合行列
（上段：従来手法，下段：提案手法。
NO_CHG は変化なしを意味する）

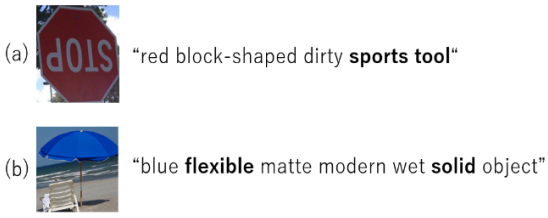


図 8　Attribute・State 選定の失敗例