主観的空間解釈の理解に向けた 場所アフォーダンスと行為意味の乖離の分析

渡辺 圭貴[1] 志村 魁哉[1] 出口 秀輝[1] 浦野 健太[1] 米澤 拓郎[1] 河口 信夫[1,2]

概要：ある空間で行われる行為が，その場の典型的な行為である場所アフォーダンスからどれだけ逸脱しているかの計算的な定量化は，行動認識・映像異常検知等の諸分野で共通の重要な課題である．本論文では，このような場所アフォーダンスからの逸脱を意味的逸脱と呼ぶ．例えば「キッチンで料理する」は典型的だが「キッチンで運動する」は典型から外れる，といった判断を画像から計算的に取り出す手法が求められる．本研究は，画像と言語等を共通の意味空間に埋め込むマルチモーダル埋め込みモデルを用いて，この問題に取り組む．意味的逸脱の検出には，背景画像から場所アフォーダンスを記述する能力と，観察画像中の場所アフォーダンスから逸脱した行為成分を抽出する能力の 2 つが必要となる．本研究は，意味的逸脱検出を直接達成するものではなく，その実現に必要となるこれら 2 つの基礎能力をマルチモーダル埋め込みモデル ImageBind を対象として評価したものである．まず ImageBind の場所アフォーダンス記述能力を体系的に評価し，モデルが場所と行為の典型/非典型をゼロショットで識別しうる点を確認した．次に，観察画像と背景画像の埋め込みベクトルの差分により，場所アフォーダンスから逸脱した行為成分を取り出す差分ベクトル指標を提案した．行為と背景の典型性を制御した合成画像を用いた行為単位の相対順位の比較により，非典型条件において差分演算が場所アフォーダンスから逸脱した行為成分を相対順位の上昇として取り出しうる点を示した．

キーワード：アフォーダンス，マルチモーダル埋め込み表現，空間解釈

1. はじめに

「ある行為が当該空間において意味的に適切かどうか」を評価する能力は，行動認識・映像異常検知をはじめとする複数分野で共通の重要な課題である．Choi ら [1] は，行動認識の主要な課題として，シーン文脈の変化に頑健な認識の難しさを指摘した．Chen ら [2] は，「人がビーチで走るのは正常だが車道で走るのは異常」のように同じ行動でも場所によって正常か異常かが変わる判断が，既存の異常行動検知手法では難しい点を指摘した．本研究はこの共通課題に対し，「場所と人物行為の意味的整合性を計算的に定量化する」枠組みの提供により取り組む．

この枠組みを構築するためには，当該場所に対して多数の人が想起する典型的行為の集合的傾向を記述する必要がある．本論文ではこれを場所アフォーダンスと呼ぶ．この概念の理論的背景として，Gibson が導入したアフォーダンス概念 [3] を，Rietveld と Kiverstein による rich landscape of affordances の議論 [4] に基づき，社会文化的実践に埋め込まれた行為可能性として場所レベルへ展開する系譜が挙げられる．すなわち場所は，その物理的構造と社会的実践の両者を通じて，特定の行為レパートリーを規範的に成立させる．例えばキッチンは「調理する・食器を洗う」を，図書館は「読む・学習する」を適切な行為として成立させる．人物の行為がこの場所アフォーダンスから乖離する程度を，本論文では「意味的逸脱」と定義する．

意味的逸脱の定量化は，冒頭に挙げた行動認識や異常検知といった個別応用にとどまらず，個人の主観的なあり方の計算的理解においても基礎的な意義を持つ．我々が提唱した Internet of Realities（IoR）[5], [6] は，物理的な事実だけでなく個人の主観的なあり方そのものを計算的に表現し，それを介して人と人とのより深い相互理解を支えようとする情報通信パラダイムである．その基盤として求められるのは，個人の内面状態を外部から推定する手段の獲得であり，観察可能な行動はその有力な情報源である．場所アフォーダンスに沿った典型的な行為は場所自体に誘発される側面が大きく行為主体の固有性を反映しにくいのに対し，場所アフォーダンスから逸脱した行為には行為主体の意図や状況が現れやすい．したがって，人物の行為が置かれた場所アフォーダンスとどう整合するかという情報は，

[1] 名古屋大学大学院 工学研究科
Graduate School of Engineering, Nagoya University
[2] 名古屋大学 未来社会創造機構
Institutes of Innovation for Future Society, Nagoya University

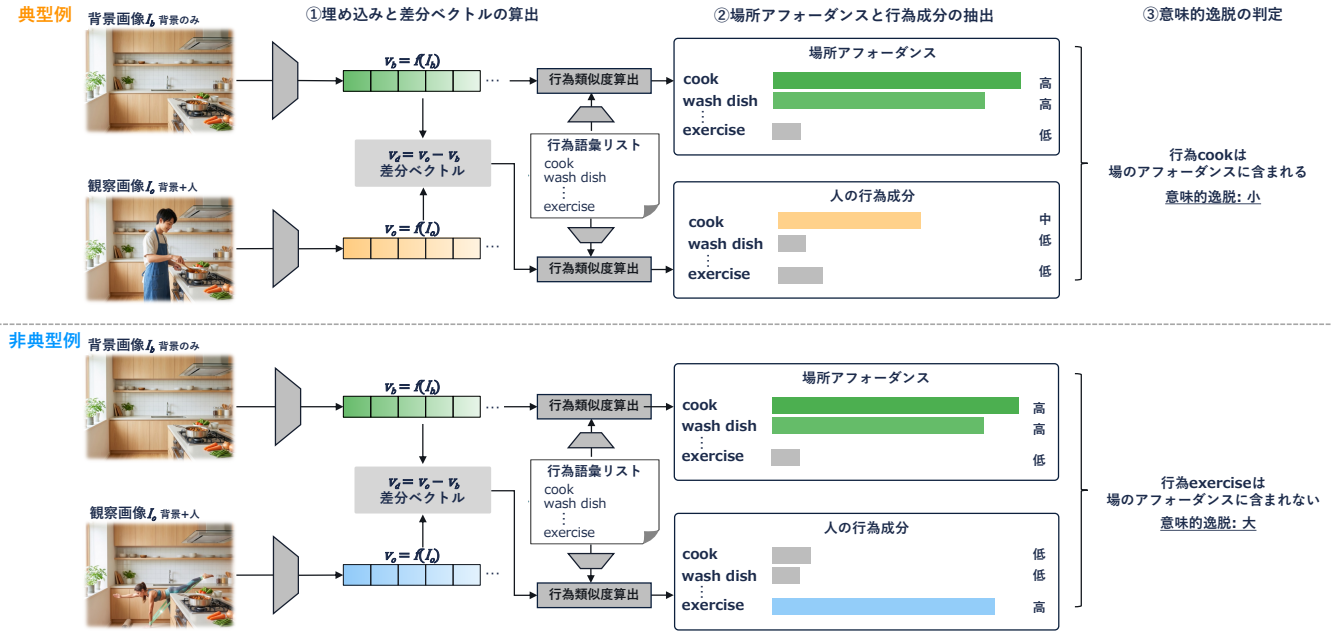

図 1: 場所アフォーダンスと行為成分の照合に基づく意味的逸脱判定の概念図

その人物の状況や意図を外部から読み取る手がかりとなり，こうした個人の理解に資する基礎的な観察指標となりうる．本研究における意味的逸脱は，主観的解釈の基礎となる「場所と行為の意味的整合性」を計算的に記述する第一段階として位置づけられる．

著者らの知る限り，既存研究ではこの意味的逸脱の定量評価に対する体系的な枠組みを与えていない．この枠組み構築に対する有望なアプローチとして，画像・言語・音声などの異種モダリティを共通潜在空間に埋め込むマルチモーダル埋め込みモデル [7], [8], [9], [10] が挙げられる．これらのモデルは，画像と任意のテキストプロンプトとの意味的類似度をゼロショットで計算できるため，原理的には任意の場所カテゴリと任意の行為記述の間の関連を追加学習なしに評価できる可能性を持つ．本研究では，マルチモーダル埋め込みモデルを用い，背景画像から得られる場所アフォーダンスと，観察画像と背景画像の埋め込み差分から得られる行為成分との照合により意味的逸脱を判定する枠組みを提案する（図 1）．

意味的逸脱を計算的に検出するためには，2 つの能力が必要となる．第一は，背景画像から場所アフォーダンス，すなわち各行為が当該場所においてどの程度典型的であるかを記述する能力である．第二は，観察画像と背景画像の埋め込み差分により，場所アフォーダンスから逸脱した行為成分を抽出する能力である．本研究では，マルチモーダル埋め込みモデル ImageBind [8] を対象として，これら 2 つの能力に対応する以下の問いを設定する．

問い 1　マルチモーダル埋め込みモデルは，背景画像から場所アフォーダンス，すなわち行為の典型/非典型をゼロショットで記述しうるか．

問い 2　観察画像と背景画像の埋め込み差分に基づく差分ベクトル指標は，場所アフォーダンスから逸脱した行為成分を，非典型条件において取り出しうるか．

問い 1 に答えるため，背景画像と，行為を表すテキストとの間でマルチモーダル埋め込みモデルが与える意味的類似度を，各行為の典型性スコアとして用いる．モデルの埋め込み空間が場所と行為の対応関係を適切に表現していれば，典型行為は高い類似度を，非典型行為は低い類似度を示すと期待される．ただし，この期待が実際にどの程度満たされるかは事前学習データの構成に依存し，自明ではない．これを検証するため実験 1 では，397 種類の場所カテゴリからなる大規模シーン認識データセット SUN397 [11], [12] を対象に，ImageBind が背景画像から場所アフォーダンスを記述しうるかを評価する．評価の基準には SUN397 の各場所カテゴリで典型的に行われる行為を被験者の自由記述による回答に基づいて集計した SUN Action [13] のアノテーションを用いる．加えて，4 種類のプロンプト形式間での識別性能の比較を行う．

問い 2 に答えるため，背景画像から得られる埋め込みベクトル $\mathbf{v}_b$ と観察画像から得られる埋め込みベクトル $\mathbf{v}_o$ の差分 $\mathbf{v}_d = \mathbf{v}_o - \mathbf{v}_b$ を差分ベクトル指標として用いる．意味的逸脱の検出において本来関心があるのは，観察画像中の行為のうち場所アフォーダンスから逸脱した成分である．ここで行為は人物単体ではなく，人物とその行為に関与する物体（道具や対象）の組として観察画像に現れるため，本論文ではこの組を「人物-物体」と呼ぶ．差分ベクトル指標は，観察画像の埋め込みから背景画像の埋め込みを差し引き，背景由来の典型成分を相殺し，観察画像中の場所アフォーダンスから逸脱した行為成分を取り出す．こ

れにより当該場所が規範的に喚起する典型行為からの意味的逸脱が捉えうる．本指標は特定のモデルに依存せず，異種モダリティを結びつける任意のマルチモーダル埋め込みモデルへの適用可能性を持ち，任意の場所に対して追加学習を要さず即時適応できると期待される点に特徴がある．これを検証するため実験 2 では，日常的な人間行為 410 種類を網羅し各画像に行為ラベルが付与された人体姿勢推定データセット MPII Human Pose Dataset [14] から抽出した人物-物体領域を SUN397 データセットの典型背景および非典型背景に合成して観察画像を生成し，多数の行為候補のうち観察画像中の人物が行っている行為がどの順位に位置するかを，背景画像のみを用いた場合と比較する．

本研究は，意味的逸脱検出を直接達成するものではなく，その実現に必要となる 2 つの基礎能力である「場所アフォーダンスの記述」と「観察画像中の，場所アフォーダンスから逸脱した行為成分の抽出」を，ImageBind を対象として検証したものである．主な貢献は以下の通りである．

(1) 意味的逸脱検出を「場所アフォーダンスの記述」と「逸脱した行為成分の抽出」という 2 つの基礎能力の問題として定式化し，それぞれの評価枠組みを提案した．
(2) マルチモーダル埋め込みモデル ImageBind の場所アフォーダンス記述能力を SUN Action データセットの人間アノテーションを基準として 397 種類の場所カテゴリにわたって定量的に評価し，高い識別性能を確認した．
(3) 観察画像と背景画像の埋め込み差分に基づく差分ベクトル指標を提案し，MPII Human Pose Dataset の 24 行為クラスを対象とした合成画像実験により，非典型条件において場所アフォーダンスから逸脱した行為成分を相対順位の上昇として取り出しうる点を示した．

本論文の構成は以下の通りである．第 2 章では関連研究を整理する．第 3 章では本研究の提案フレームワークを定式化する．第 4 章および第 5 章では，二つの問いを検証する実験について，それぞれの方法と結果を述べる．第 6 章では総合的な議論と本研究の限界，今後の展望を述べ，第 7 章で結論をまとめる．

2. 関連研究

2.1 空間アフォーダンスの計算的記述

Gibson は，環境が動物に提供する行為可能性をアフォーダンスと定義した [3]．Gibson 自身はアフォーダンスを物体のみならず表面・場所・環境全体が提供するものとして論じていたが，コンピュータビジョン分野では主に物体レベルの行為可能性（「椅子は座る行為を示唆する」「ハンドルは把持を示唆する」）として扱われてきた [15], [16]．

場所・環境レベルへのアフォーダンス概念の拡張は，生態心理学の文脈で独自に発展してきた．Rietveld と Kiverstein は，アフォーダンスを「物質的環境の側面と，ある生活様式において利用可能な能力との関係」として再定式化し，個別の物体ではなく空間全体が一群の目的的行為を喚起するという「rich landscape of affordances」の概念を提唱した [4]．本研究が定義する場所アフォーダンスは，この系譜に立脚している．

コンピュータビジョン分野には，シーン画像と人物行為の対応関係を扱った研究がある．Vu らは SUN Action データセットを構築し，397 種類の場所カテゴリに対して典型的な人間行為をクラウドソーシングでアノテーションした [13]．この研究はシーン画像から典型行為を予測するタスクを定式化した先駆的な取り組みであるが，教師あり学習を前提としており，未知の場所カテゴリへのゼロショット適用は想定されていない．Chuang らは行為-物体関係をアノテートした ADE-Affordance データセットと，シーン全体の文脈に基づき各物体に対する行為の適切性を Graph Neural Network で推論するモデルを提案した [17]．Chuang らは行為-物体関係を物理的障害・機能的不適合・社会的不適切・社会的禁止・危険性等に分類した点で本研究と問題意識を共有するが，人手アノテーションと教師あり学習を必要とし，物体インスタンス単位の推論に留まる．近年，OpenScene[18] は CLIP 特徴の 3D 共埋め込みにより，AffordanceCLIP[19] は CLIP のみで，それぞれゼロショットなアフォーダンス推論を実現するが，前者は 3D 点群入力を要し，後者は物体レベルでの領域特定に留まる．本研究は任意の場所画像に対して追加学習なしに場所アフォーダンスをスコア分布として記述する点でこれらと異なり，加えて観察画像と背景画像の埋め込み差分を用いた逸脱行為成分の抽出機構を提案する点に独自性がある．

2.2 マルチモーダル埋め込みモデルとその系統的バイアス

Radford らが提案した CLIP は，Web 上から収集された大規模な画像-テキストペアに対する対照学習により，画像とテキストを共通の意味空間に埋め込むモデルである [7]．CLIP は事前学習時に明示的なラベル付け対象でないクラスについても，テキストプロンプトとのコサイン類似度を介してゼロショットで識別可能であり，物体認識・シーン分類・行為認識など多様な視覚タスクに適用されている．

CLIP の枠組みは画像-言語以外のモダリティへも拡張されている．Guzhov らによる AudioCLIP[10] は音声を，Ni らによる X-CLIP[9] および Xu らによる VideoCLIP[20] は動画を，それぞれ CLIP の埋め込み空間と整合的に統合する手法を提案している．Girdhar らによる ImageBind[8] はこの方向性をさらに推し進め，画像・テキスト・音声・深度・熱画像・IMU の 6 モダリティを単一の埋め込み空間に統合する．また，特定タスクへの適応として，Wang らによる ActionCLIP[21] は人物・行為認識向けに CLIP をファインチューニングする派生モデルを提案している．

これらマルチモーダル埋め込みモデルは原理的には任意

の空間カテゴリと任意の行為記述の関連を追加学習なしに評価しうるが，その埋め込み空間における系統的なバイアスの存在が指摘されている．Yuksekgonul らは CLIP 系モデルにおける物体間の関係や属性の結合に対する表現能力の限界を示し [22]，Thrush らは Winoground ベンチマークにより視覚-言語モデルの構成的推論能力の限界を定量化した [23]．これらの研究は VLM の表現が「個別概念の検出」に強く「概念間の関係構造の弁別」に弱い点を示唆する．本研究は，こうした VLM の表現が「場所と行為の関係」をどの程度記述しうるかを ImageBind を対象に検証するとともに，背景画像との差分演算により観察画像中の場所アフォーダンスから逸脱した行為成分を抽出する手法を提案する点に位置づけられる．

本研究では ImageBind を実装に用いる．画像と行為記述テキストのコサイン類似度をゼロショットで計算可能であり，かつ画像の埋め込みが行為概念と整合する空間構造が期待されるためである．ただし，モデルが埋め込み空間においてどの程度の精度で場所アフォーダンスを記述できるかは事前学習データの構成に依存し，自明ではない．この点は本研究の実験 1 において検証する．

2.3 シーン文脈と行為認識

ある行為が当該空間において意味的に適切かどうかという問題は，行動認識および映像異常検知の文脈で繰り返し言及されてきたが，その関係を計算的に記述する体系的な枠組みは確立していない．行動認識の分野では，モデルが空間的手がかりに過度に依存するシーンバイアスが長年の課題として認識されている．Choi らは，行動認識モデルが「バスケットコートにいる→バスケットボールをしている」のように空間情報に依拠する傾向を実証し，adversarial loss を用いたバイアス除去手法を提案している [1]．Choi らは空間と行為の意味的相関を「除去すべき誤差」とみなす立場をとるが，本研究はこの相関そのものを規範的アフォーダンスの根拠として活用しつつ，背景由来の典型成分を演算的に相殺し，場所アフォーダンスから逸脱した行為成分を取り出す立場をとる．

映像異常検知 (video anomaly detection; VAD) の分野では，UCF-Crime[24] や ShanghaiTech[25]，CUHK Avenue[26] などのデータセットを対象とした教師あり・弱教師あり・教師なし手法が多数提案されている．近年では大規模言語モデル (large language model; LLM) を活用したゼロショット手法も提案されており [27], [28]，アノテーションコストを削減する方向の研究が進められている．これらの手法は判定基準として LLM の汎用知識やドメイン固有の学習データを前提としており，当該空間の視覚的特徴そのものを規範の根拠として組み込む設計にはなっていない．Chen らによる DecoAD[2] は，シーンと行為の関係を知識グラフ上で切り離し学習する手法を提案しており，本研究と問題意識を共有する．ただし DecoAD は知識グラフの構築と弱教師あり学習を要する．

これら既存研究は，空間と行為の関係を扱う重要性を示している一方で，関係そのものを計算的に記述する枠組みは提示していない．本研究ではこの課題に対し，マルチモーダル埋め込みモデルが場所アフォーダンスを記述しうるかを検証するとともに，背景由来の典型成分を相殺して観察画像中の場所アフォーダンスから逸脱した行為成分を取り出す差分ベクトル指標を提示する．

3. 差分ベクトル指標

本研究では，ある空間における人物の行為が当該場所からどれだけ意味的に逸脱しているかを，マルチモーダル埋め込みモデルの埋め込み空間上のベクトル演算によって計算的に記述する手法を提案する（図 1）．本手法は Encoder Model $f(\cdot)$ を抽象化した設計として定式化されており，本稿では ImageBind [8] を用いた実装によりその有効性を検証する．本手法は，観察画像から得られる埋め込みベクトルと当該画像中の背景画像から得られる埋め込みベクトルの差分により，背景由来の典型成分を相殺し，場所アフォーダンスから逸脱した行為成分を取り出しうる，という考え方に基づく．

3.1 定式化

$f(\cdot)$ を Encoder Model，すなわち画像あるいはテキストを d 次元の埋め込み空間へ写像する関数とする：

$$f : \mathcal{X} \to \mathbb{R}^d \tag{1}$$

人物の行為を含む観察画像 I_o および当該画像中の背景領域のみの画像 I_b に対して，それぞれの埋め込みベクトルを以下のように定義する：

$$\mathbf{v}_o = f(I_o) \in \mathbb{R}^d \tag{2}$$

$$\mathbf{v}_b = f(I_b) \in \mathbb{R}^d \tag{3}$$

ここで $\mathbf{v}_b$ は当該場所を記述する 場所アフォーダンスベクトル，$\mathbf{v}_o$ は当該場所に置かれた人物の行為を含む観察全体を記述するベクトルである．

次に，意味的逸脱を捉える 差分ベクトル $\mathbf{v}_d$ を以下のように定義する：

$$\mathbf{v}_d = \mathbf{v}_o - \mathbf{v}_b \in \mathbb{R}^d \tag{4}$$

$\mathbf{v}_d$ は，観察画像から背景由来の典型成分を取り除いた残差ベクトルである．背景が当該行為に対して典型的である場合，$\mathbf{v}_o$ と $\mathbf{v}_b$ はいずれもその行為方向の成分を強く持つため，人物-物体由来の行為成分は背景由来の典型成分に対して相対的に小さくなり，差分 $\mathbf{v}_d$ における行為方向の成分は弱く現れる．一方，背景が非典型的な場合，$\mathbf{v}_b$ は当

該行為方向の信号をほとんど持たないため，観察画像中の人物-物体由来の行為成分が差分 $\mathbf{v}_d$ において強く現れる．この成分は，場所アフォーダンスから逸脱した行為成分に相当する．

意味的逸脱の大きさをスカラ指標として取り出す方法は，$\mathbf{v}_d$ のノルム $\|\mathbf{v}_d\|$ や特定の行為テキスト埋め込みとの内積など複数の選択肢があり得る．本研究で提案する差分ベクトル指標は具体的なスコア化を実装に委ねたフレームワーク水準の提案であるが，本論文では行為テキスト埋め込み $f(t_A)$ との内積

$$s_d(A) = \cos\big(\mathbf{v}_d, f(t_A)\big) \tag{5}$$

を用いた形式の有効性を 5 章の実験 2 において検証する．

本指標は特定のモデルに依存せず，異種モダリティを結びつける任意のマルチモーダル埋め込みモデルに原理的に適用可能であり，任意の場所に対して追加学習を要さず即時適応できる点に特徴がある．

3.2 検証の流れ

意味的逸脱の検出には 2 つの能力が必要となる．第一は，背景画像から場所アフォーダンスを記述する能力，すなわち各行為が当該場所においてどの程度典型的であるかを評価する能力である．これは $\mathbf{v}_b$ が場所アフォーダンスを意味的に記述するという定式化（式 3）の妥当性に対応する．第二は，観察画像と背景画像の埋め込み差分から，場所アフォーダンスから逸脱した行為成分を抽出する能力である．これは差分演算（式 4）が場所アフォーダンスと整合する典型成分を相殺し，逸脱成分を取り出せるかという問題に対応する．1 章で提示した問い 1，問い 2 はそれぞれこの 2 点に対応する．本研究では問い 1 を 4 章の実験 1 で，問い 2 を 5 章の実験 2 で検証する．

両実験における評価指標は問いの性質に応じて使い分ける．実験 1 では，コサイン類似度が典型ペアと非典型ペアをどれだけ分離できるかを ROC-AUC により評価する．実験 2 では，観察画像中の人物が行っている行為を他の候補行為からどれだけ識別できるかという画像内での識別能力を評価するため，当該行為が候補行為集合の中でどの位置にランクされるかを表す相対順位を比較指標として用いる．具体的な定式化はそれぞれ 4 章および 5 章で示す．

4. 実験 1: マルチモーダル埋め込みモデルによる場所アフォーダンス記述

本章では問い 1 として提示した「マルチモーダル埋め込みモデルは場所アフォーダンスをゼロショットで記述しうるか」を，ImageBind [8] を対象として検証する．具体的には，場所と行為の典型/非典型のペアに対する識別性能を，4 種類のプロンプト形式について評価する．

4.1 データセットと評価基準

画像データセットには SUN397 [11], [12] を用いる．場所アフォーダンスの人手による基準としては SUN Action [13] のアノテーションを用いる．SUN Action は，SUN397 の各シーンカテゴリに対して被験者がそのシーンで想起しうる行為を回答した結果を集計したものである．これは，本論文で定義する場所アフォーダンス（当該場所に対して多数の人が想起する典型的行為の集合的傾向）と概念的に対応する．なお，本研究で用いるアノテーションは，SUN Action の著者が公開している被験者ごとの生回答データから，原論文の集計手続きを参考に自ら再構築したものである．

SUN Action は屋外シーンと屋内シーンで異なる行為集合を採用する設計となっている．「海（屋外）で皿を洗う（屋内）」のように属性が明らかに不整合な行為を評価対象から除くためである．具体的には，屋外用 38 行為，屋内用 23 行為（うち eat や sleep など 9 行為は両方に共通）のリストに票数が付与される．行為語彙の構成を表 1 に示す．本実験では SUN Action のこの設計に従い，シーンの Indoor/Outdoor 属性に応じて対応する行為集合を評価対象とする．評価対象となる画像数は計 94,111 枚である．

4.2 評価指標とプロンプト

各 (シーン画像 I, 行為 a) のペアについて，ImageBind の画像埋め込み $f(I)$ と行為プロンプトのテキスト埋め込み $f(t_a)$ とのコサイン類似度を当該ペアの典型性スコアとする．

$$S(I, a) = \cos(f(I), f(t_a)) \tag{6}$$

人手アノテーションの正解ラベルは，場所と行為の整合性が明確な極端事例を評価の基準とするため，SUN Action 原論文に倣い，票数に基づく以下の 2 値で定義する．被験者の大多数が想起した行為を典型として Human-Typical，誰も想起しなかった行為を非典型として Human-Atypical と表記する．

- **Human-Typical** (人間が典型と判定): 票数 ≥ 20
- **Human-Atypical** (人間が非典型と判定): 票数 $= 0$

票数が中間的なペアは，整合性の解釈が分かれる可能性があるため本評価の対象外とした．

Human-Typical を陽性，Human-Atypical を陰性とした 2 値分類問題における ImageBind のコサイン類似度の ROC-AUC により，典型/非典型を分ける識別性能を評価する．ROC-AUC は 2 値分類器の閾値非依存な識別能力を表す指標であり，値域は 0 から 1，ランダムな分類器で 0.5，完全に正しく分類する分類器で 1.0 となる．値が大きいほど典型/非典型を区別する能力が高い．

ImageBind の出力はテキストプロンプトの記述に依存するため，場所アフォーダンス記述に最も適したプロンプト

表 1: SUN Action における行為語彙の構成

屋外シーン専用 (29 種類)	屋内シーン専用 (14 種類)	両方に共通 (9 種類)
cheer, climb, cross bridge, dig, drive, explore, farm, fly, garden, go fishing, harvest, have a picnic, hike, live, open door, park, play, play baseball, play golf, ride, ride train, run, sail, sightsee, skate, ski, smell flower, take a walk, take picture	bowl, cook, dance, do laundry, exercise, gamble, listen, play game, play pool, read, study, wait, wash dish, watch	eat, learn, pray, relax, shop, sit, sleep, swim, work

表 2: プロンプト形式の比較対象．A は行為動詞を表す．

プロンプト名	形式
`word`	A
`photo`	a photo of A
`photo_people`	a photo of a place where people A
`people`	people A

形式を探るべく，行為の表現粒度の異なる 4 種類のプロンプト形式を比較対象とする．各プロンプトの形式を表 2 に示す．A には対象行為の動詞句が挿入される (例: `climb`)．

4.3 結果と考察

人間が典型と判定したペア (Human-Typical) と非典型と判定したペア (Human-Atypical) を対象とした 2 値識別の ROC 曲線を図 2 に示す．4 種類のプロンプト形式すべてで AUC が 0.7 を超える識別性能を示し，最高性能はシンプルな人物中心プロンプト (`people`, AUC = 0.822) で達成された．

- `word`: AUC = 0.733
- `photo`: AUC = 0.759
- `photo_people`: AUC = 0.786
- `people`: **AUC = 0.822** (最高)

プロンプト形式の選択により AUC は 0.089 程度の差を示したが，全プロンプトで AUC が 0.7 を超える水準は維持された．

この結果は，ImageBind が場所アフォーダンスをゼロショットで記述しうる点を示している．Web スケールの事前学習データに含まれる場所と典型行為の結びつきが，特定のプロンプト形式に強く依存せず ImageBind の埋め込み空間にある程度反映されているといえる．これにより，問い 1 に対して肯定的な答えが得られた．ただし，意味的逸脱の検出には，場所アフォーダンスの記述だけでなく，観察画像中の場所アフォーダンスから逸脱した行為成分を抽出する能力が必要となる．次章では問い 2 として，観察画像と背景画像の埋め込み差分により場所アフォーダンスから逸脱した行為成分を抽出する差分ベクトル指標の有効性を検証する．

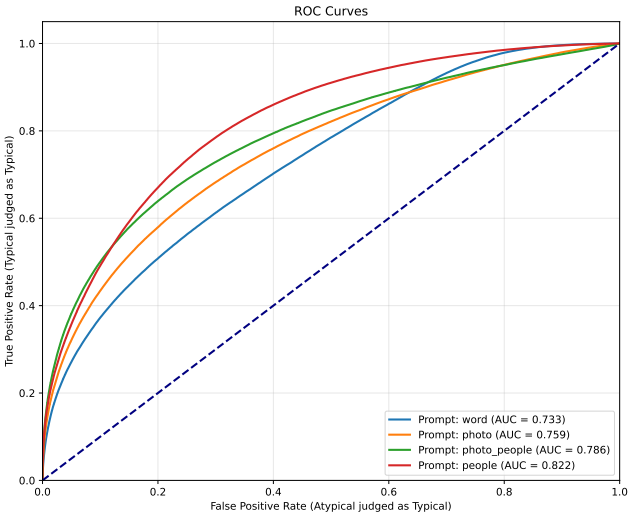

図 2: 4 種類のプロンプト形式における典型/非典型識別性能の ROC 曲線

5. 実験 2: 差分ベクトル指標による逸脱行為成分の抽出

本章では問い 2 として，観察画像と背景画像の埋め込み差分に基づく差分ベクトル指標が，観察画像中の場所アフォーダンスから逸脱した行為成分を取り出しうるかを検証する．意味的逸脱の検出という用途上，特に関心があるのは行為と背景の組み合わせが非典型な場合である．評価では，多数の行為候補のうち観察画像中の人物が行っている行為（以下，正解行為）が差分ベクトル指標においてどの順位に位置するかを，背景画像のみを用いた場合と比較する．

5.1 合成画像の生成

問い 2 の検証には，行為と背景の典型性を制御した観察画像が必要である．本研究では，MPII Human Pose Dataset [14] から抽出した人物-物体領域を SUN397 [11], [12] の異なるシーン画像へ合成して観察画像を生成する．

MPII 行為と SUN Action 行為の間に厳密な一対一対応が成立するものを抽出し，目視選別の結果，最終的に画像サンプル数が 20 件以上残った 25 行為が候補となった．対応リストと各行為の画像数を表 3 に示す．このうち dance については，SUN Action において票数 0 となる SUN397 カテゴリが存在せず非典型背景を構成できないため，評

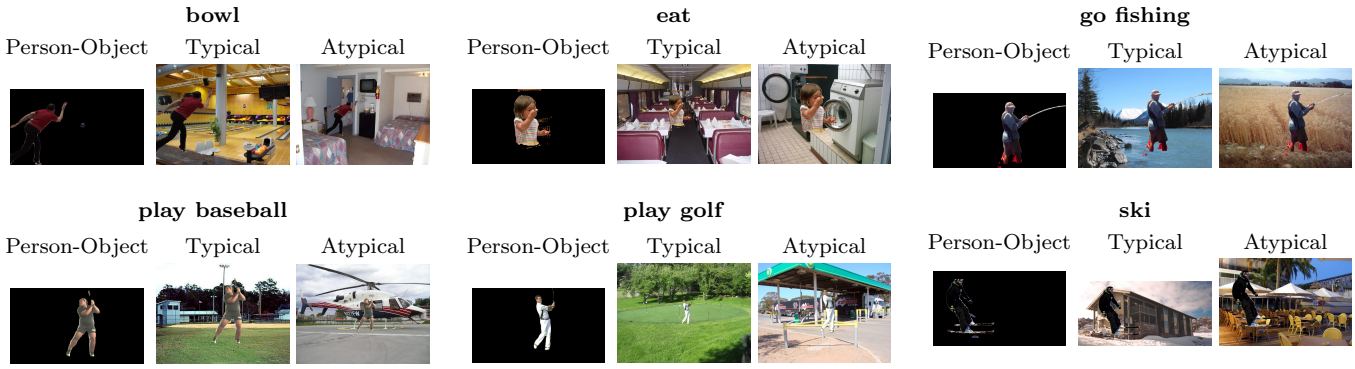

図 3: 人物-物体領域と合成観察画像の例 (Person-Object: 人物-物体領域 / Typical: 典型背景 / Atypical: 非典型背景)

表 3: SUN Action 行為と MPII Human Pose Dataset の MPII 行為との一対一マッピング

SUN Action	MPII 代表行為	画像数
bowl	bowling	25
climb	rock climbing	102
cook	cooking or food preparation	226
dance†	ballet, modern, or jazz	115
dig	digging, spading garden	44
do laundry	laundry (home activities)	55
eat	eating, sitting	53
exercise	yoga	554
farm	farming, rice, planting	76
garden	planting, potting seedlings	159
go fishing	fishing, general	275
harvest	picking fruits/vegetables	33
hike	backpacking	67
play baseball	softball, general	79
play golf	golf	104
play pool	billiards	26
ride	horseback riding	85
run	running	98
sail	sailing	21
shop	food shopping, standing or walking	22
skate	skateboarding	207
ski	skiing, downhill	210
swim	swimming, general	62
take a walk	walking, general	42
wash dish	wash dishes	37
合計（25 候補）		2,777

† 非典型条件（票数 0 の SUN397 カテゴリ）を満たす背景が存在しないため評価対象から除外．評価対象は dance を除く 24 行為．

価対象から除外する．結果として，最終的な評価対象は dance を除く 24 行為となる．

各行為について 20 枚の画像をランダムに選定し，それぞれから抽出した人物-物体領域に対して以下の手順で合成画像を生成した．

(1) 典型背景の選定: 当該行為 A が SUN Action において票数 ≥ 20 で出現する SUN397 カテゴリの中からランダムに 1 つを選び，そのカテゴリの画像から 1 枚をランダムに選定する．

(2) 非典型背景の選定: 当該行為 A の票数が 0 の SUN397 カテゴリの中からランダムに 1 つを選び，そのカテゴリの画像から 1 枚をランダムに選定する．

(3) 合成: 抽出した人物-物体領域マスクを用いて，人物-物体領域を典型背景画像および非典型背景画像にそれぞれ重畳合成し，2 種類の観察画像を生成する．

人物-物体領域の抽出には先行研究 [29] の手法を用いる．背景画像 I_b には合成に用いた SUN397 画像をそのまま用いる．24 行為 $\times$ 各 20 サンプルにより 480 個の人物-物体領域を得て，それぞれ典型背景・非典型背景に合成を行い，非典型ペア・典型ペア各 480 サンプル，計 960 サンプルが評価対象となる．抽出した人物-物体領域と，それを典型背景および非典型背景に合成した観察画像の例を図 3 に示す．

5.2 評価指標とプロンプト

実験 2 では，(画像 I, 行為 a) のペアの典型性スコアを計算する 2 つの方法を比較する．観察画像の埋め込みを $\mathbf{v}_o$，背景画像の埋め込みを $\mathbf{v}_b$，行為プロンプトのテキスト埋め込みを $\mathbf{t}_a$ と表記する．

- **Background Only**：背景画像の埋め込みと行為プロンプトのコサイン類似度

$$S_{\mathrm{BG}}(I, a) = \cos(\mathbf{v}_b, \mathbf{t}_a) \tag{7}$$

- **Difference**：差分ベクトル $\mathbf{v}_d = \mathbf{v}_o - \mathbf{v}_b$ と行為プロンプトのコサイン類似度

$$S_{\mathrm{Diff}}(I, a) = \cos(\mathbf{v}_d, \mathbf{t}_a) \tag{8}$$

Difference は本研究で提案する差分ベクトル指標であり，Background Only は差分演算の効果を考察するための比較対象として併せて評価する．

本実験で評価したいのは，観察画像中で正解行為を他の候補行為からどれだけ識別できるかという画像内での識別能力である．これは cos 類似度の絶対値ではなく，正解行為が他の候補行為と比較してどの位置にランクされるかと

表 4: 差分演算による正解行為の順位上昇

	順位上昇行為数 ($\Delta\text{Rank} > 0$)		平均 ΔRank	
プロンプト	非典型	典型	非典型	典型
`word`	17/24 (70.8%)	7/24 (29.2%)	+0.293	−0.160
`photo`	20/24 (83.3%)	9/24 (37.5%)	+0.350	−0.077
`people`	20/24 (83.3%)	8/24 (33.3%)	+0.384	−0.041
`photo_people`	19/24 (79.2%)	8/24 (33.3%)	+0.333	−0.164

いう相対的な量で直接的に捉えられる．そこで本実験では，各方法において正解行為が候補行為集合の中で何位に位置するかを正規化した相対順位を比較指標として用いる．

観察画像 I と正解行為 a_{target} の組について，行為集合 $\mathcal{A}$（実験 1 と同じく SUN Action の行為語彙）内の各行為 a_i に対するスコア $S(I, a_i)$ を用いて，正解行為 a_{target} の相対順位を以下で算出する．

$$\text{RelRank}(a_{\text{target}}) = \frac{|\{a_i \in \mathcal{A} : S(I, a_i) < S(I, a_{\text{target}})\}|}{|\mathcal{A}| - 1} \tag{9}$$

RelRank = 1.0 は，スコア関数 S が a_{target} を当該行為集合の最上位と評価した結果に，0.0 は最下位と評価した結果に対応する．正解行為 a_{target} は合成画像の生成に用いた行為を設定する．評価対象は表 3 の 24 行為である．

差分演算が場所アフォーダンスから逸脱した行為成分を取り出していれば，Difference の相対順位は Background Only を上回ると予想される．そこで本実験では，行為単位で両条件の相対順位を比較する．具体的には，各行為について Difference 条件の平均相対順位から Background Only 条件の平均相対順位を引いた値を ΔRank として算出する．$\Delta\text{Rank} > 0$ は差分演算による正解行為の順位上昇を意味する．集約指標として，$\Delta\text{Rank} > 0$ となる行為数（差分演算が順位上昇をもたらした行為の割合）と，全行為にわたる ΔRank の平均値を用いる．行為テキストプロンプトとして，実験 1 と同様に `word`, `photo`, `photo_people`, `people` の 4 種類を評価対象とする．

5.3 結果と考察

集約結果を表 4 に示す．非典型ペアでは 4 種類のプロンプトすべてで多数派の行為 (24 行為中 17–20 行為，70.8%–83.3%) において差分演算が相対順位を上昇させ，平均 ΔRank は +0.293 から +0.384 の上昇幅を示した．行為単位での内訳を図 4 に示す．ここでは，意味的逸脱検出の主たる関心対象である非典型ペアについて，4 種類のプロンプトのうち平均 ΔRank が最大であった `people` プロンプトの結果を代表として示す．Background Only 条件で相対順位が極めて低い行為（図の左側）において，差分演算後に大幅な順位上昇が確認できる．これは観察画像中の場所アフォーダンスから逸脱した行為成分が，差分演算により上位に取り出される挙動を示している．

この結果は，差分ベクトル指標が ImageBind の埋め込み空間において，観察画像中の場所アフォーダンスから逸脱した行為成分を識別しうる可能性を示唆する．非典型背景は当該行為について場所アフォーダンスを共有しないため，Background Only 条件では正解行為が候補集合中の下位に留まる傾向がある．これに対し，差分演算後には正解行為が候補集合中の上位に位置するようになる．この挙動は，差分演算が場所アフォーダンスから逸脱した行為成分を正解行為として識別可能な形に取り出している点と整合する．これは意味的逸脱の検出，すなわち「場所と矛盾する行為」が観察画像に含まれるかの判定に向けた基礎指標として，本指標が機能しうる点を示唆する．

一方，非典型ペアの中にも差分演算の効果が限定的な行為群（例として `take a walk`, `harvest`, `garden` 等）が存在する．これらの行為では，Background Only 条件における正解行為の相対順位が既に高い点が観察された（図 4 の右側）．「非典型」とラベル付けされた背景画像でも ImageBind が当該行為を相応の類似度で評価しており，差分演算による順位上昇の余地が小さくなる．これは差分演算の問題ではなく，差分演算の前提となる「当該背景画像が正解行為に対して非典型である」という個別ケースが ImageBind の埋め込み空間において必ずしも成立しないためである．SUN Action の人間アノテーションでは非典型と判定された組み合わせでも，視覚的な手がかりやモデルの認識上の類似性により背景画像が正解行為の手がかりを提供してしまう場合があると考えられる．

続いて典型ペアでの結果を見る．これは差分演算の効果が逸脱の有無に応じてどう変化するかを示す，補完的な検証である．典型ペアでは差分演算の効果は限定的であり，順位上昇行為数は 24 行為中 7–9 行為（29.2%–37.5%）にとどまり，平均 Δ Rank もわずかに負の値（-0.041 から -0.164）を示した．これは典型背景が当該行為と関連付くため Background Only 条件でも正解行為の相対順位が比較的高く，差分演算による背景情報の除去がむしろ正解行為の順位を下げる方向に働くためと考えられる．意味的逸脱検出の用途を考えると，関心の中心は「行為と場所アフォーダンスが乖離した非典型条件で逸脱した行為成分を取り出せるか」にあり，典型条件で差分演算の効果が限定的である点はむしろ本指標の用途と整合する．

以上から，差分ベクトル指標は ImageBind の埋め込み空間において，観察画像中で場所アフォーダンスから逸脱した行為成分を正解行為として識別する能力を非典型条件において向上させうる点が示唆された．本指標が機能しない行為群は ImageBind 自体の場所アフォーダンス記述能力の限界に対応しており，差分演算の限界というよりは元のモデルの限界に由来する．差分演算アプローチは意味的逸脱検出の基礎指標として予備的な有効性が示唆された．

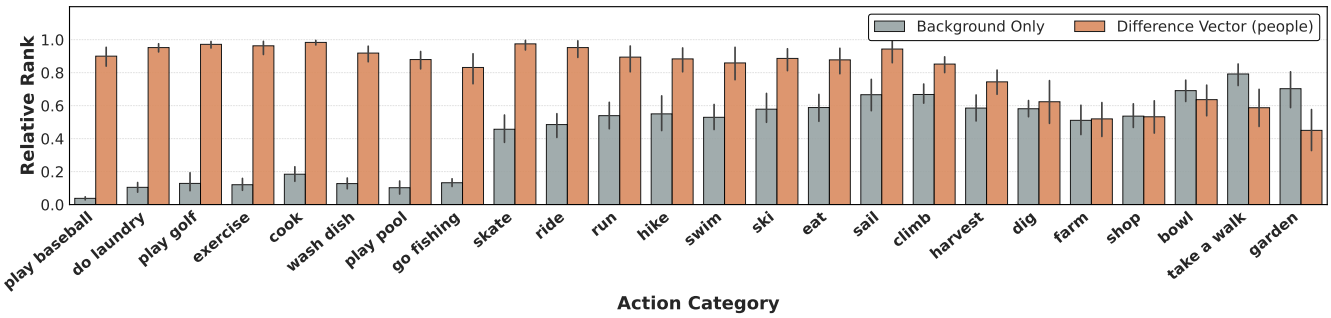

図 4: 非典型ペアにおける行為単位の相対順位の比較 (people プロンプト)

6. 議論と今後の課題

4 章および 5 章では，意味的逸脱検出に必要な 2 つの基礎能力を ImageBind を対象に実験的に検証した．本章では，両実験から得られた知見を統合的に解釈し，本研究の限界と今後の展望を整理する．

6.1 二つの問いへの統合的解釈

実験 1 では，ImageBind の場所アフォーダンス記述能力を SUN397 の 397 種類の場所カテゴリ・SUN Action の人手アノテーションを基準として評価した．4 種類のプロンプト形式すべてで AUC が 0.733–0.822 と 0.7 を超える識別性能が確認され，問い 1 であるマルチモーダル埋め込みモデルが背景画像から場所アフォーダンスをゼロショットで記述しうるかに対しては肯定的な答えが得られた．

実験 2 では，観察画像と背景画像の埋め込み差分に基づく差分ベクトル指標を提案した．MPII Human Pose Dataset の 24 行為を SUN397 の典型/非典型背景に合成した実験により，差分演算が背景単独以上に正解行為を上位に取り出すかを，行為単位の相対順位の比較により評価した．非典型ペアでは，4 種類のプロンプトすべてで多数派の行為 (24 行為中 17–20 行為，70.8–83.3%) において差分演算が相対順位を上昇させ，平均 ΔRank は +0.293 から +0.384 の上昇幅を示した．問い 2 として設定した「差分ベクトル指標が場所アフォーダンスから逸脱した行為成分を非典型条件において取り出しうるか」に対しては肯定的な答えが得られた．

これら二つの実験を通じて，意味的逸脱検出に必要な 2 つの能力について，ImageBind を対象とした基礎的な検証を行った．第一の能力は背景画像からの場所アフォーダンスの記述であり，第二の能力は差分演算による場所アフォーダンスから逸脱した行為成分の抽出である．ImageBind は「場所と典型行為の対応関係」を Web スケールの事前学習を通じてある程度獲得しており，場所アフォーダンスを記述する基礎能力を備えうる点が示された．加えて，差分ベクトル指標は，意味的逸脱検出の用途で関心の中心となる非典型条件において，場所アフォーダンスから逸脱した行為成分を相対順位の上昇として取り出しうる点が示された．効果は完全ではないが，意味的逸脱検出の基礎指標として差分演算アプローチの予備的な有効性が示唆された．

6.2 限界

本研究で得られた知見にはいくつかの限界が伴う．以下では，本研究の枠組みと結論の解釈に関わる主要な論点を整理する．

- 評価対象を ImageBind に限定したため，本研究で観察された知見が CLIP 系列の事前学習構成に由来する一般的特性なのか，ImageBind 固有の特性なのかは未確認である．他のマルチモーダル埋め込みモデルでの再現検証が今後の課題となる．
- 実験 2 の検証には人物-物体領域を背景画像に合成して生成した観察画像を用いたため，撮影された実画像において行為が背景と乖離している自然な逸脱事例での有効性は別途検証が必要である．合成画像では人物-物体と背景の照明・遠近・視覚的整合性が完全ではなく，差分演算が捉えている信号に合成由来の不整合が混入している可能性が排除できない．本研究で観察された順位上昇が行為そのものに由来する信号と合成由来の人工的不整合のどちらに支えられているかの切り分けは，自然画像での検証によって初めて可能となる．
- 実験 2 の対象行為は MPII Human Pose Dataset と SUN Action が厳密に対応する 24 クラスに限定されており，より広範な日常行為への一般化は未確認である．本指標の適用範囲は行為クラスの性質と ImageBind の場所アフォーダンス記述能力の両者に依存し，後者の限界が現れる行為群（`take a walk`, `harvest`, `garden` 等）では差分演算の効果も限定される．

6.3 今後の展望

実験 1 で確認された ImageBind の場所アフォーダンス記述能力と実験 2 で示された差分ベクトル指標の予備的な有効性は，マルチモーダル埋め込みモデルにおける行為主体表現の強化が中心的な研究課題である点を示唆する．本枠組みの拡張として，以下のような方向が考えられる．

- 行為認識や動画理解に特化したマルチモーダル埋め込みモデルとの統合により，観察画像中の行為成分の抽出を強化する方向.
- 差分ベクトル指標をシーン・姿勢・物体といった情報源ごとに適用し統合する階層的アプローチの検討.

また, 本研究の問題設定は, 我々が提唱する IoR (Internet of Realities) [5], [6] の基盤課題に直接接続する．IoR が目指す個人の主観的なあり方の計算的表現に向けては，観察可能な行動から行為主体の意図や状況を読み取る手段が必要であり，場所アフォーダンスからの意味的逸脱はその有力な観察指標となりうる．本研究は，意味的逸脱を計算的に記述する第一段階としてマルチモーダル埋め込みモデルの基礎能力を検証したものであり，今後は検出された逸脱から行為主体の意図や状況を推論する段階へと枠組みを展開していく必要がある．本研究の結果は，こうした方向において必要となるマルチモーダル埋め込みモデルにおける行為主体表現の強化を示すものであり，空間と行為の意味的関係を計算的に扱う今後の研究の具体的な設計指針となる.

7. 結論

本研究は，行動認識・映像異常検知等で共通する課題である意味的逸脱の検出を見据え，その実現に必要となる二つの基礎能力である場所アフォーダンスの記述と，観察画像中の場所アフォーダンスから逸脱した行為成分の抽出を，マルチモーダル埋め込みモデル ImageBind を対象に評価した．実験 1 では，SUN397 と SUN Action の人手アノテーションを基準として ImageBind の場所アフォーダンス記述能力を評価した．4 種類のプロンプト形式すべてで一貫した識別性能が確認され，問い 1 として設定した「マルチモーダル埋め込みモデルは背景画像から場所アフォーダンスをゼロショットで記述しうるか」に対し肯定的な答えが得られた．実験 2 では，観察画像と背景画像の埋め込み差分に基づく差分ベクトル指標を提案した．MPII Human Pose Dataset の行為を SUN397 の典型/非典型背景に合成した評価により，非典型条件で多数派の行為で差分演算による相対順位の上昇を示した．問い 2 として設定した「差分ベクトル指標は場所アフォーダンスから逸脱した行為成分を非典型条件において取り出しうるか」に対し肯定的な答えが得られた.

以上から, マルチモーダル埋め込みモデルが場所アフォーダンスをゼロショットで記述しうる点，および差分演算が場所アフォーダンスから逸脱した行為成分を非典型条件において取り出しうる点を示した．本研究はこれらを意味的逸脱検出の基礎指標として位置づけ，マルチモーダル埋め込みモデルを意味的逸脱検出に応用する予備的な有効性を示したものである.

謝辞 本研究の一部は JST CREST (JPMJCR22M4) に支援いただいています.

参考文献

[1] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, Vol. 32, , 2019.

[2] Chenglizhao Chen, Xinyu Liu, Mengke Song, Luming Li, Shaojiang Yuan, Xu Yu, and Shanchen Pang. Unveiling context-related anomalies: Knowledge graph empowered decoupling of scene and action for human-related video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[3] James J Gibson. The theory of affordances. *Hilldale, USA*, Vol. 1, No. 2, pp. 67–82, 1977.

[4] Erik Rietveld and Julian Kiverstein. A rich landscape of affordances. *Ecological Psychology*, Vol. 26, No. 4, pp. 325–352, 2014.

[5] Takuro Yonezawa. Towards internet of realities - vision and challenges. In *2025 IEEE International Symposium on Emerging Metaverse (ISEMV)*, pp. 157–157, 2025.

[6] Takuro Yonezawa. Vision and challenges of internet of realities. https://internet-of-realities.org/vision/. Accessed: 2026-03-23.

[7] 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.

[8] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

[9] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pp. 1–18. Springer, 2022.

[10] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980. IEEE, 2022.

[11] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, Vol. 119, pp. 3–22, 2014.

[12] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

[13] T.H. VU, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. Predicting actions from static scenes. In *ECCV*, 2014.

[14] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[15] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR 2011*, pp. 1529–1536,

2011.

[16] Yun Jiang, Hema Koppula, and Ashutosh Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2993–3000, 2013.

[17] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. *arXiv preprint arXiv:1712.07576*, 2017.

[18] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

[19] Claudia Cuttano, Gabriele Rosi, Gabriele Trivigno, and Giuseppe Averta. What does clip know about peeling a banana? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2238–2247, June 2024.

[20] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6787–6800, 2021.

[21] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[22] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

[23] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

[24] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

[25] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - a new baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 2018.

[26] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pp. 2720–2727, 2013.

[27] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.

[28] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pp. 304–322. Springer, 2024.

[29] Kaiya Shimura, Kazuma Kano, Tahera Hossain, Shin Katayama, Kenta Urano, Shun Taguchi, Hideki Deguchi, Hiroyuki Sakai, Takuro Yonezawa, and Nobuo Kawaguchi. Real-time activity-aware video streaming while preserving privacy. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 236–240, 2025.