Attribution-OoD Axes for VLMs

推論ログに基づく二次元 Attribution-OoD 軸と OoD 検知性能の解析

d-hacks B2 omote 親:ryokawa さん

概 要

本稿では，Vision–Language Model (VLM) の推論ログから得られる信頼度指標に基づき，Out-of-Distribution (OoD) 検知の「失敗要因」を二次元平面上で解析する Fixed Attribution-OoD Axes を構成する．OpenCLIP ViT-B を用い，Food-101 を ID，CIFAR-100 および ImageNet-R を OoD とする設定のもと，softmax 信頼度・正規化エントロピー・エネルギー・単純 OoD スコアの四つを ID 平均からの差分 $[d\text{conf\_drop}, d\text{entropy\_gain}, d\text{energy\_gain}, d\text{oodscore\_gain}]$ として抽出し，SparsePCA と 1SE ルールにより軸数とスパース性をデータ駆動に決定する．その結果，得られた二軸は

$$z_u(x) = d\text{entropy\_gain}(x), \quad z_c(x) \approx \tfrac{1}{\sqrt{2}}\{d\text{conf\_drop}(x) + d\text{oodscore\_gain}(x)\}$$

という解釈容易な線形式に収束し，seed・プロンプト・バックボーン（ViT-B/32, ViT-B/16）の変更に対しても軸方向の余弦類似度がほぼ 1.0 であることから，高いロバスト性を持つ固定軸として機能することを示す．さらに，これら二軸上に線形分類器を構成すると，AUROC 0.94，TNR@95TPR 0.69 を達成し，softmax や単一軸ベースラインを大きく上回りつつ，エネルギーベース OoD スコアに近い検知性能を維持することを確認した．加えて，z_u–z_c 平面の四象限ごとの誤検知分布と効果量解析により，OoD 失敗パターンを「主に不確実性が増大する領域」と「主に確信度が崩れる領域」に分解可能であることを示す．

1 背景

深層ニューラルネットワークを用いた画像認識システムは，学習時に観測していない分布外入力（Out-of-Distribution; OoD）に対しても高い信頼度で誤分類し得ることが知られている．このため，実運用においては，入力が学習分布の範囲内か否かを判定する OoD 検知機構が必須となる．従来は softmax 出力の最大値を反転させた MSP や，logit に温度スケーリングや摂動を施す手法，エネルギーベースのスコア [1] など，単一のスカラー量を用いて ID/OoD を二値判定する設計が主流であった．

しかし，単一スコアによる OoD 検知は,「どのような理由でその入力が OoD と判断されたか」という失敗要因の内訳を与えない．運用者の観点からは，例えば「意味的にまったく別物のクラスに迷い込んでいるのか」,「本来のクラスのまわりで信頼度だけが不安定になっているのか」といった情報が，アラートの優先度付けやデータ収集方針の決定に直結する．このギャップを埋めるため，最近では複数の証拠スコアを組み合わせ，意味的変化やドメイン変化といった異なる OoD 要因を軸として可視化しようとする試みが提案されている．

一方で，既存の多軸的な OoD 解析は，特定のバックボーン（例えば ResNet）を前提とした特徴空間の幾何学に依存していたり，研究者が経験的に決めた重み付き和により軸を定義している場合が多い．このような設計では，学習データの分割や乱数 seed，あるいはモデル・プロンプトの変更に対して軸の向きや解釈が揺らぎやすく，再現性や汎用性の観点から問題が残る．さらに，Vision–Language Model (VLM) のように，テキストと画像の両方から特徴を生成するモデルでは，softmax 信頼度・エントロピー・エネルギーなど多様なスコアが得られるにもかかわらず，それらを統一的な低次元座標系として整理する枠組みは十分に確立されていない．

本研究では，この状況を踏まえ，VLM の推論ログから得られる複数の信頼度指標を出発点として,「不確実性」と「確信度の崩れ」という OoD の代表的な要因を表す二軸を，できるだけ少ないパラメータ自由度でデータ駆動に抽出することを目的とする．ここで得られる軸は，具体的な OoD 判定器（しきい値や分類境界）の設計から切り離された「固定座標系」として機能し，異なる seed・プロンプト・バックボーン間で安定した比較や失敗分析を行うための基盤となることを目指す．

2 問題点

前節で述べたように，従来の OoD 検知は単一スカラーのスコアに依存しており，入力ごとの「どのような要因で失敗しているか」を直接には与えない．これに対して，本研究が対象とする VLM では softmax 信頼度・エントロピー・エネルギー・単純な OoD スコアなど複数の指標が得られるにもかかわらず，それらを体系的に整理し，少数の解釈しやすい軸として固定する枠組みは十分に整備されていない．特に，以下の点が問題となる．

- **(P1) 軸設計の恣意性と再現性の欠如**
  複数の指標を用いて OoD の要因を分解しようとする既存の試みでは，研究者が経験的に決めた重み付き和によって「不確実性軸」「確信度軸」といった軸を定義することが多い．しかし，どの指標をどの比率で混ぜるかは明確な規準に基づかないことが多く，学習データの分割や乱数 seed，さらにはモデルやプロンプトを変えるたびに軸の向きや解釈が変動し得る．このため，異なる条件で得られた結果を同一の座標系で比較することが難しい．

- **(P2) 指標間の冗長性と有効次元の不透明さ**
  VLM の推論ログから得られる信頼度指標の中には，softmax 信頼度と MSP（$1 - \max p$）のように実質的に同一の情報を含むものや，強く相関するものが含まれる．このような冗長な特徴をそのまま用いて多次元の軸を構成すると，形式的な次元数と実効的な自由度がずれ，どれだけの軸が本質的に必要なのかが不明瞭になる．結果として,「何次元の Attribution-OoD 軸を採用すべきか」という設計判断も曖昧になってしまう．
- **(P3) 軸の解釈可能性と検知性能のトレードオフ**
  単純なエネルギースコアや MSP は，しばしば非常に高い AUROC と TNR@95TPR を達成する一方で，それ単体では失敗の内訳を与えない．逆に，多軸的な可視化を重視して複雑な線形結合や非線形変換を導入すると，軸の意味付けが難しくなり，実用上重要な検知性能（例：エネルギーベース OoD に匹敵する性能）をどこまで維持できるかが不明確である．現状では,「解釈しやすい少数軸」と「実用的な検知性能」の両立を，定量的に検証した例は限られている．
- **(P4) VLM 環境に特化した評価とロバスト性検証の不足**
  従来の固定軸設計は，ResNet 系の特徴空間や単一のバックボーンに依存して定義されていることが多く，VLM にそのまま適用できるかは検証されていない．また，seed・プロンプト・バックボーンの変更に対して軸がどの程度安定に保たれるか（軸方向の余弦類似度や選択される軸数の一貫性）を体系的に調べた報告も少ない．このため，得られた軸が「特定条件でのみ通用する座標系」にとどまっている可能性が残る．

本研究では，これらの問題に対して，(i) VLM の信頼度指標から冗長性を抑えた差分特徴ベクトルを構成し，(ii) SparsePCA と 1SE ルールに基づき軸数とスパース性をデータ駆動に決定しつつ，(iii) 不確実性軸 z_u と確信度軸 z_c を高いロバスト性と検知性能を保ったまま固定座標系として抽出することを目指す．

3 先行研究

画像分類における OoD 検知は，ID/OoD の分離スコア設計と評価設定の両面で発展してきた．スコア設計では，ソフトマックス確信度の過信を避けるためにエネルギースコアを用いる枠組みが提案され [1]，内部活性の異常値を抑えて過信を減らす ReAct[2]，特徴空間と logit を統合する ViM[3] などが代表的である．評価面では，設定差による不公平を是正するため統合ベンチマーク OpenOOD が整備された [4]．一方，Vision-Language Model（VLM）は大規模画像–テキスト対で学習した CLIP が提示され [5]，プロンプト学習（CoCoOp[6]）や学習不要アダプタ（Tip-Adapter[7]）により下流適応が容易になった．VLM を OoD に用いる研究として，テキスト概念との最大一致を用いる MCM[8]，適応で OoD 一般化を保つ CLIPood[9]，few-shot OoD を志向した LoCoOp[10] や負のプロンプト [11]，不確実性に基づく自己較正チューニング SCT[12] が報告されている．本稿の「2 軸化」は，疎な線形軸を得る SparsePCA[13] で寄与解釈を担保し，軸の再現性は帰属の安定性保証研究 [14] にならって検証する．

4 何を解くか

本研究が扱う対象は，Vision–Language Model (VLM) の推論ログから得られる複数の信頼度指標を入力とし，それらを

$$x \mapsto s(x) \in R^4$$

（softmax 信頼度・正規化エントロピー・エネルギー・単純 OoD スコアから構成した差分特徴ベクトル）として受け取り，

$$s(x) \mapsto z(x) = (z_u(x), z_c(x)) \in R^2$$

という二次元の Attribution-OoD 軸へ写像する表現学習の問題である．ここで z_u は「不確実性の増大」を主に表す軸，z_c は「確信度の崩れ」を主に表す軸として解釈したい．さらに，この二軸は単なる可視化にとどまらず，線形分類器を組み合わせることで OoD 検知器としても実用的な性能を発揮し，かつ seed・プロンプト・バックボーンが変わっても安定して再現される「固定座標系」として機能することが求められる．

4.1 問題設定

具体的には，ID データ集合 $\mathcal{D}_{\text{id}}$（本研究では Food-101），OoD データ集合 $\mathcal{D}_{\text{ood}}$（CIFAR-100 と ImageNet-R）を用意し，VLM による推論から各サンプル x に対して

- softmax 信頼度 $\text{conf}(x)$,
- 正規化エントロピー $\text{entropy\_norm}(x)$,
- エネルギー $\text{energy}(x)$,
- 単純 OoD スコア $\text{ood\_score}(x)$

を計算する．ID サンプルに対するこれらの平均値 $\mu_{\text{conf}}, \mu_{\text{entropy}}, \mu_{\text{energy}}, \mu_{\text{ood}}$ を基準とし，

$$d\text{conf\_drop}(x) = \mu_{\text{conf}} - \text{conf}(x), \tag{1}$$

$$d\text{entropy\_gain}(x) = \text{entropy\_norm}(x) - \mu_{\text{entropy}}, \tag{2}$$

$$d\text{energy\_gain}(x) = \text{energy}(x) - \mu_{\text{energy}}, \tag{3}$$

$$d\text{oodscore\_gain}(x) = \text{ood\_score}(x) - \mu_{\text{ood}} \tag{4}$$

からなる 4 次元差分特徴ベクトル

$$s(x) = \big(d\text{conf\_drop}(x), d\text{entropy\_gain}(x), d\text{energy\_gain}(x), d\text{oodscore\_gain}(x)\big)^\top \in R^4.$$

を構成する．
本研究の主問題は，この $s(x)$ の集合に対して

1. どのような規準で有効次元数 k（潜在軸の本数）とスパース性を決めるか,
2. どのようにして「不確実性軸 z_u」と「確信度軸 z_c」をデータ駆動に割り当てるか,
3. 得られた二軸が OoD 検知と失敗分析にどの程度有用か,
4. その性質が seed・プロンプト・バックボーンの変更に対してどこまで保たれるか

を一貫した枠組みで定式化し，評価することである．

4.2 設計要件

上記の問題に対し，本研究では次のような設計要件を満たす二軸 z_u, z_c を構成することを目的とする．

要件 (R1): 軸数のデータ駆動な決定 4 次元の差分特徴から SparsePCA により k 本の潜在軸を学習し，交差検証の再構成誤差と 1SE ルールを用いて「過学習を避けつつ最も単純なモデル」を自動的に選択する．これにより,「何本の Attribution-OoD 軸が本質的に必要か」という設計判断を，経験則ではなく統計的規準に基づき決める．

要件 (R2): 軸ラベル付けの自動化と解釈性 学習された k 本の軸のうち，エントロピー増大・エネルギー増大・OoD スコア増大に最も強く寄与する軸を「不確実性軸 z_u」，信頼度低下に最も強く寄与する軸を「確信度軸 z_c」としてデータ駆動に割り当てる．このとき，軸の符号も「不確実性が増える方向」「確信度が下がる方向」が正方向になるよう自動的に揃え，研究者の恣意的な回転・命名に依存しない解釈可能な座標系とする．

要件 (R3): OoD 判定器から独立した固定座標系 軸学習の段階では，ID/OoD ラベルを用いた AUROC などの検知性能指標を最適化対象とせず，あくまで差分特徴の再構成誤差のみを目的とする．これにより，後段でどのような判定器（一次元しきい値，2 次元線形分類器など）を組み合わせても再利用可能な,「表現としての固定軸」を得ることを目指す．

要件 (R4): 実用的な OoD 検知性能 得られた二軸 (z_u, z_c) 上に単純な線形分類器（一次元合成 $z_u + z_c$，線形 SVM，ロジスティック回帰など）を構成したとき，MSP や OoD スコアといった単一指標ベースラインよりも高い AUROC および TNR@95TPR を達成し，かつエネルギーベース OoD スコアに近い性能を維持できることを確認する．また，bootstrap による 95% 信頼区間を算出し，性能差が偶然ではないことを統計的に評価する．

要件 (R5): 失敗内訳の説明力 z_u–z_c 平面上で四象限を定義し，各象限における ID/OoD 比率や偽陽性 (FP)・偽陰性 (FN) の集中度を定量化する．これにより，ある入力が OoD と判断されたときに「主に不確実性が増大しているのか」「主に確信度が崩れているのか」といった失敗要因の内訳を，座標位置として説明できることを目指す．さらに，ID ／ OoD の分布差をマンホイットニーの U 検定や KS 検定，効果量（Cohen の d，delta）で評価し，二軸が統計的にも意味のある分離方向になっているかを確認する．

要件 (R6): seed・プロンプト・モデルに対するロバスト性 乱数 seed，プロンプトテンプレート（例：“a photo of {name}”, “a close-up photo of {name}”），およびバックボーン（ViT-B/32, ViT-B/16）を変えながら同一パイプラインを実行し，各条件で得られた z_u, z_c の負荷ベクトル同士の余弦類似度を評価する．特に，基準条件に対して $|\cos(z_u, z_u^{\mathrm{ref}})|$，$|\cos(z_c, z_c^{\mathrm{ref}})|$ がほぼ 1.0 となることを示し,「条件が変わってもほぼ同じ向きの軸が再現される」という意味での固定性を検証する．

4.3 本研究で解く具体的な問い

以上の設計要件を踏まえ，本研究が解く具体的な問いを整理すると，次の三点にまとめられる．

- **問い Q1：VLM の信頼度指標から，データ駆動に決まる少数の Attribution-OoD 軸を構成できるか．**
  4 つの差分特徴から SparsePCA と 1SE ルールを用いて，最小限の軸数 k とスパースな負荷ベクトルを決定し，そのうち二軸を z_u, z_c として自然に解釈できるかを検証する．
- **問い Q2：得られた二軸は，従来の単一指標に匹敵する OoD 検知性能を維持しつつ，失敗要因の内訳を説明できるか．**
  (z_u, z_c) を入力とした一次元合成および線形分類器の AUROC, TNR@95TPR を，MSP・エネルギースコアなどのベースラインと比較し，その性能と信頼区間を評価する．同時に，四象限別の FP/FN 分布や効果量解析を通じて，二軸が OoD 失敗パターンの解釈にどの程度寄与するかを明らかにする．
- **問い Q3：構成された軸は，seed・プロンプト・バックボーンが変わっても固定座標系として再現されるか．**
  seed，プロンプト，モデルを変えた複数の条件で軸学習を繰り返し，それぞれの z_u, z_c の負荷ベクトルが基準条件と高い余弦類似度を保つかどうかを調べる．これにより，本稿で提案する Attribution-OoD 軸が，特定条件に依存しないロバストな座標系として利用可能かを評価する．

本稿では，これら Q1–Q3 に対する実験的な回答を与えることで，VLM の推論ログに対する「データ駆動型 Fixed Attribution-OoD Axes」の構成とその有用性を示す．

5 実行フロー

本節では，本研究で実際に構築したパイプラインの処理手順を示す．

5.1 データセットと VLM 設定

まず，ID ／ OoD の構成と VLM の設定を固定する．

- **ID データ**：TensorFlow Datasets (TFDS) に含まれる Food-101 を用い，訓練分割から 10,000 枚を ID サンプルとする．
- **OoD データ**：CIFAR-100 の訓練集合から 5,000 枚，ImageNet-R のテスト集合から 5,000 枚を OoD サンプルとして用いる．いずれも画像をそのまま入力し，ラベルは ID/OoD の二値のみで利用する．
- **VLM バックボーン**：OpenCLIP 実装の ViT-B/32 を基本設定とし，事前学習重みには `laion2b_s34b_b79k` を採用する．ロバスト性実験では ViT-B/16（`laion2b_s34b_b88k`）も併用する．
- **プロンプトテンプレート**：クラス名 {name} に対して`"a photo of {name}"` を基本とし，ロバスト性実験では`"a close-up photo of {name}"` も併用する．

Food-101 のクラス名からテキストプロンプト列を生成し，OpenCLIP のテキストエンコーダで埋め込みを計算する．画像側は同一 VLM の image encoder で埋め込みを得て，両者の内積からクラスごとの logit と softmax 確率を求める．

5.2 信頼度指標の計算と ID 平均の取得

各画像 x に対して，VLM の出力から次の信頼度指標を計算する．

- softmax 最大値（信頼度）$\mathrm{conf}(x) = \max_k p_\theta(k \mid x)$
- 正規化エントロピー $\mathrm{entropy\_norm}(x)$：クラス数で割ったエントロピー
- エネルギースコア $\mathrm{energy}(x) = -\log \sum_k \exp(\mathrm{logit}_k(x))$
- 単純 OoD スコア $\mathrm{ood\_score}(x) = 1 - \mathrm{conf}(x)$

ID サンプル集合 $\mathcal{D}_{\mathrm{id}}$ に対して，これら 4 指標の平均値

$$\mu_{\mathrm{conf}},\quad \mu_{\mathrm{entropy}},\quad \mu_{\mathrm{energy}},\quad \mu_{\mathrm{ood}}$$

を計算し，後続の差分特徴の基準値として用いる．この段階で，すべてのサンプルについて ID/OoD フラグ，予測ラベル，正誤，4 種のスコアを含むテーブル `sample_metrics.csv` を出力する．

5.3 差分特徴ベクトルの構成

次に，ID 平均からの差分として 4 次元の特徴ベクトル $s(x)$ を構成する．具体的には，

$$d\mathrm{conf\_drop}(x) = \mu_{\mathrm{conf}} - \mathrm{conf}(x), \tag{5}$$
$$d\mathrm{entropy\_gain}(x) = \mathrm{entropy\_norm}(x) - \mu_{\mathrm{entropy}}, \tag{6}$$
$$d\mathrm{energy\_gain}(x) = \mathrm{energy}(x) - \mu_{\mathrm{energy}}, \tag{7}$$
$$d\mathrm{oodscore\_gain}(x) = \mathrm{ood\_score}(x) - \mu_{\mathrm{ood}} \tag{8}$$

を定義し，

$$s(x) = \big(d\mathrm{conf\_drop}(x), d\mathrm{entropy\_gain}(x), d\mathrm{energy\_gain}(x), d\mathrm{oodscore\_gain}(x)\big)^\top \in R^4$$

とする．softmax 信頼度と MSP は実質同一情報であるため，本研究では $\mathrm{conf}(x)$ のみを用い，MSP は別特徴としては持たない．

全サンプルの $s(x)$ を `axis_features_raw.csv` に保存し，分散が閾値以下の特徴を落とす簡易な分散スクリーニングを適用する．今回の設定では 4 変数すべてが閾値を上回り，4 次元のまま後続処理に渡される．

5.4 SparsePCA と 1SE ルールによる軸学習

差分特徴集合 $\{s(x)\}$ に対して，SparsePCA による低次元表現学習を行う．

1. 特徴ベクトル $s(x)$ を標準化し，平均 0・分散 1 のスケールに正規化する．
2. 潜在軸数 k を 1〜4，スパース性ハイパーパラメータ α を $\{0.5, 1.0, 2.0, 4.0, 8.0\}$ のグリッドで走査する．
3. ID/OoD ラベルに基づく層化 K 分割交差検証（本研究では 5 分割）を行い，各 (k, α) に対して再構成誤差 $\mathrm{MSE} = \|X_{\mathrm{val}} - \hat{X}_{\mathrm{val}}\|_2^2$ の平均と標準誤差を計算する．
4. 交差検証で最小の平均 MSE を達成した点 $(k_{\min}, \alpha_{\min})$ とその標準誤差 $\mathrm{SE}_{\min}$ から $\mathrm{MSE}_{\min} + \mathrm{SE}_{\min}$ を 1SE 閾値とし，その閾値以下の候補の中から最小の k を持つ設定を採用する（1SE ルール）．

この結果，本研究の設定では $k = 3$ が選択され，対応する α は 1.0 付近に収束する．すべての (k, α) に対する CV 結果は `sparsepca_cv_table.csv` に保存し，図**??**1sefig:cv_1se ルールによる選択過程を可視化する．

5.5 Attribution-OoD 軸 z_u, z_c の確定

SparsePCA により得られた 3 本の潜在軸の負荷行列を $W \in R^{3\times 4}$ とする．行 W_i は第 i 軸の 4 特徴に対する線形結合係数を表す．ここから

- エントロピー増大・エネルギー増大・OoD スコア増大に対する寄与が大きい軸を「不確実性軸候補」として選び，
- 信頼度低下に対する寄与が大きい軸を「確信度軸候補」として選ぶ．

より形式的には，$|W_{i,j}|$ を用いて各軸の特徴ごとの寄与度を評価し，$\{d\mathrm{entropy\_gain}, d\mathrm{oodscore\_gain}, d\mathrm{energy\_gain}\}$ への寄与が最大となる軸を z_u，$d\mathrm{conf\_drop}$ への寄与が最大となる軸を z_c として割り当てる．二軸が一致してしまう場合は，次点の軸を採用する．また，「不確実性が増える方向」「確信度が落ちる方向」が正になるよう，該当特徴に対する係数が負であれば軸全体の符号を反転する．

本研究の 4 変数設定では，最終的に

$$z_u(x) = d\mathrm{entropy\_gain}(x),$$
$$z_c(x) \approx \tfrac{1}{\sqrt{2}}\big\{d\mathrm{conf\_drop}(x) + d\mathrm{oodscore\_gain}(x)\big\}.$$

という非常に単純で解釈しやすい形に収束した．各サンプルに対して計算された $(z_u(x), z_c(x))$ は `axis_scores.csv` に保存し，図**??**zc_densityfig:zu_zc_densityOoD の分布を可視化する．

5.6 二軸上での OoD 判定器の学習と評価

得られた座標 (z_u, z_c) 上で OoD 判定器を構成し，従来の単一スコアベースラインと比較する．まず，全サンプルを ID/OoD ラベルに基づき層化して 7:3 に分割し，訓練インデックス $\mathcal{I}_{\mathrm{train}}$，テストインデックス $\mathcal{I}_{\mathrm{test}}$ を作成する．

- **単一スコアベースライン**：MSP（$1 - \mathrm{conf}$），エネルギースコア energy をそのまま OoD スコアとみなし，AUROC と TNR@95TPR を評価する．
- **一次元合成**：標準化した z_u, z_c を用いて $z_{\mathrm{sum}} = z_u^{\mathrm{norm}} + z_c^{\mathrm{norm}}$ を定義し，OoD スコアとして評価する．
- **二次元線形分類器**：特徴ベクトル (z_u, z_c) を入力として，標準化＋ロジスティック回帰，標準化＋線形 SVM を学習し，それぞれの決定関数・確率を OoD スコアとみなして評価する．

各モデルに対してテスト分割上の AUROC および TNR@95TPR を算出し，さらにブートストラップ再標本化（反復数 $B = 1000$）による 95% 信頼区間を求める．結果は `detector_eval_with_ci.csv` に保存し，図??cifig:perf_ci で可視化する．

5.7 z_u–z_c 平面における誤りパターン解析

次に，二軸平面上で OoD 判定の失敗パターンを解析する．ロジスティック回帰により学習した判定器を固定し，テスト分割上の予測に対して FP（ID を OoD と誤検知），FN（OoD を ID と見逃し）をラベル付けする．

- **四象限の定義**：z_u–z_c 平面を $\mathrm{Q1}(+u,+c), \mathrm{Q2}(-u,+c), \mathrm{Q3}(-u,-c), \mathrm{Q4}(+u,-c)$ の四象限に分割する．
- **象限ごとの集計**：各象限ごとに，サンプル数，OoD 比率，FP 数，FN 数を集計し，どの領域にどの種の誤りが集中しているかを調べる（`quadrant_summary_testsplit.csv`）．
- **代表事例の抽出**：全サンプル版では，象限ごとに FP ／ FN の上位例をスコア順に抽出し，後続の定性的なエラービジュアル分析に利用できるようにする．

これにより,「不確実性が大きいが確信度も崩れている領域」（例：Q1）や,「不確実性は小さいが確信度のみ崩れている領域」（例：Q4）など，OoD 失敗パターンの内訳を座標として記述できるようになる．

5.8 統計解析とエフェクトサイズの評価

二軸および元の信頼度指標が，ID と OoD の間で統計的に有意な差を持つかどうかを検証するため，マンホイットニーの U 検定および Kolmogorov–Smirnov 検定を指標ごとに実施する．さらに，分布差の大きさを把握するために，Cliff に類似した delta 指標および Cohen の d を算出し，指標ごとに effect_delta を比較する（`significance_id_vs_ood.csv`, 図??sizesfig:effect_sizesseed・プロンプト・モデルに対するロバスト性検証最後に，Attribution-OoD 軸が条件変更に対してどの程度安定かを検証する．

- **条件の組み合わせ**：seed $\in \{42, 43, 44\}$，プロンプト $\in$ {`"a photo of {name}"`, `"a close-up photo of {name}"`}, バックボーン $\in \{\mathrm{ViT\text{-}B/32}, \mathrm{ViT\text{-}B/16}\}$ の全組み合わせに対して，同一の軸構築パイプラインを実行する．
- **参照軸の設定**：seed=42，プロンプト `"a photo of {name}"`, ViT-B/32 の結果を基準とし，その z_u, z_c の負荷ベクトルを ref とする．
- **余弦類似度による比較**：各条件で得られた z_u, z_c の負荷ベクトルと参照軸との絶対余弦類似度 $|\cos(\cdot,\cdot)|$ を計算し，seed ／プロンプト固定での平均，さらにモデルごとの平均を算出する（`robustness_cosine_summary_multimodel.csv`, `robustness_cosine_by_model.csv`）．

実験の結果，すべての条件で $|\cos(z_u, z_u^{\mathrm{ref}})| \approx 1.0$, $|\cos(z_c, z_c^{\mathrm{ref}})| \approx 1.0$ が得られ，seed・プロンプト・バックボーンを変更してもほぼ同一方向の軸が再構成されることを確認した．このことは，本研究で構成した二軸が VLM の推論ログに対する「固定的な Attribution-OoD 座標系」として機能し得ることを示している．

6 数理的な定義

本節では，本研究で用いる記号と量を形式的に定義し，Attribution-OoD 軸 (z_u, z_c) がどのような写像として構成されているかを数理的に整理する．可能な箇所については，実験設定における具体的な値も併記する．

6.1 データ分布とラベルの定義

ID データ分布を P_{id}，OoD データ分布を P_{ood} とし，それぞれからサンプルされる入力画像を

$$x \sim P_{\mathrm{id}}, \qquad x \sim P_{\mathrm{ood}}$$

と書く．本研究では，具体的に次の 3 つの TFDS データセットを用いる：

- Food-101（訓練分割）を ID 分布の近似 P_{id} とみなし，$N_{\mathrm{id}} = 10{,}000$ 枚の画像をサンプリングする．
- CIFAR-100（訓練分割）から $N_{\mathrm{cifar}} = 5{,}000$ 枚，ImageNet-R（テスト分割）から $N_{\mathrm{imr}} = 5{,}000$ 枚の画像を取り出し，これらを OoD 分布の近似 P_{ood} とみなす．

全体のサンプル数は

$$N = N_{\mathrm{id}} + N_{\mathrm{cifar}} + N_{\mathrm{imr}} = 20{,}000$$

である．各サンプル x_i に対し，ID/OoD の二値ラベル

$$y_i = \begin{cases} 0, & x_i \sim P_{\mathrm{id}} \text{ (Food-101)} \\ 1, & x_i \sim P_{\mathrm{ood}} \text{ (CIFAR-100 または ImageNet-R)} \end{cases}$$

を付与する．

Food-101 のクラス数を K とおくと，Food-101 は $K = 101$ クラスからなる．クラスインデックス集合を

$$\mathcal{C} = \{1, 2, \ldots, K\}$$

と書き，クラス $c \in \mathcal{C}$ のクラス名（英語ラベル）を $\mathrm{name}(c)$ と表す．

6.2 VLM と基本スコアの定義

OpenCLIP による Vision–Language Model を $(f_{\mathrm{img}}, f_{\mathrm{txt}})$ と書く．画像エンコーダ f_{img} は画像 x を埋め込みベクトル $\mathbf{v}(x) \in R^d$ に写像し，テキストエンコーダ f_{txt} はクラスプロンプト τ_c を埋め込みベクトル $\mathbf{u}_c \in R^d$ に写像する：

$$\mathbf{v}(x) = \frac{f_{\mathrm{img}}(x)}{\|f_{\mathrm{img}}(x)\|_2}, \tag{9}$$

$$\mathbf{u}_c = \frac{f_{\mathrm{txt}}(\tau_c)}{\|f_{\mathrm{txt}}(\tau_c)\|_2}, \tag{10}$$

ただし $\|\cdot\|_2$ はユークリッドノルムを表す．

プロンプトテンプレートは

$$\tau_c = \texttt{"a photo of \{name\}"}\ (\mathrm{name}(c))$$

を基本とし，ロバスト性実験では `"a close-up photo of {name}"` も用いる．

CLIP 型モデルのロジットは，学習済みスカラー $\beta > 0$（`logit_scale` の指数）を用いて

$$\ell_c(x) = \beta\, \mathbf{v}(x)^\top \mathbf{u}_c \quad (c \in \mathcal{C}) \tag{11}$$

と定義する．これに softmax を適用して，クラス事後確率（予測分布）

$$p_\theta(c \mid x) = \frac{\exp(\ell_c(x))}{\sum_{k \in \mathcal{C}} \exp(\ell_k(x))} \quad (c \in \mathcal{C}) \tag{12}$$

を得る．

この予測分布から，以下の 4 つの基本スコアを定義する．

softmax 信頼度

$$\mathrm{conf}(x) = \max_{c \in \mathcal{C}} p_\theta(c \mid x). \tag{13}$$

正規化エントロピー シャノンエントロピー $H(x) = -\sum_{c \in \mathcal{C}} p_\theta(c \mid x) \log p_\theta(c \mid x)$ をクラス数で正規化した

$$\mathrm{entropy\_norm}(x) = \frac{H(x)}{\log K} = -\frac{1}{\log K} \sum_{c \in \mathcal{C}} p_\theta(c \mid x) \log p_\theta(c \mid x) \tag{14}$$

を用いる．$0 \le \mathrm{entropy\_norm}(x) \le 1$ となるようにスケールされている．

エネルギースコア

$$\mathrm{energy}(x) = -\log \sum_{c \in \mathcal{C}} \exp\bigl(\ell_c(x)\bigr). \tag{15}$$

OoD 入力に対しては $\mathrm{energy}(x)$ が大きくなる（すなわち低エネルギー領域から外れる）ことが期待される．

単純 OoD スコア 本研究の 4 変数版では，MSP 型スコアと同一の

$$\mathrm{ood\_score}(x) = 1 - \mathrm{conf}(x) \tag{16}$$

を単純 OoD スコアとして採用する．

6.3 ID 平均と差分特徴ベクトル

ID サンプル集合 $\mathcal{D}_{\mathrm{id}} = \{x_i : y_i = 0\}$ に対して，それぞれのスコアの経験平均を

$$\hat{\mu}_{\mathrm{conf}} = \frac{1}{N_{\mathrm{id}}} \sum_{i : y_i = 0} \mathrm{conf}(x_i), \tag{17}$$

$$\hat{\mu}_{\mathrm{entropy}} = \frac{1}{N_{\mathrm{id}}} \sum_{i : y_i = 0} \mathrm{entropy\_norm}(x_i), \tag{18}$$

$$\hat{\mu}_{\mathrm{energy}} = \frac{1}{N_{\mathrm{id}}} \sum_{i : y_i = 0} \mathrm{energy}(x_i), \tag{19}$$

$$\hat{\mu}_{\mathrm{ood}} = \frac{1}{N_{\mathrm{id}}} \sum_{i : y_i = 0} \mathrm{ood\_score}(x_i) \tag{20}$$

と定義する．実験設定（ViT-B/32, `laion2b_s34b_b79k`, `"a photo of {name}"`, seed=42）における具体的な値は

$$\hat{\mu}_{\mathrm{conf}} \approx 0.8373, \tag{21}$$

$$\hat{\mu}_{\mathrm{entropy}} \approx 0.1112, \tag{22}$$

$$\hat{\mu}_{\mathrm{energy}} \approx -32.1942, \tag{23}$$

$$\hat{\mu}_{\mathrm{ood}} \approx 0.1627 \tag{24}$$

であった（小数第 5 位を四捨五入）．

これらを基準として，各サンプルに対し 4 次元の差分特徴を定義する：

$$d\mathrm{conf\_drop}(x) = \hat{\mu}_{\mathrm{conf}} - \mathrm{conf}(x), \tag{25}$$

$$d\mathrm{entropy\_gain}(x) = \mathrm{entropy\_norm}(x) - \hat{\mu}_{\mathrm{entropy}}, \tag{26}$$

$$d\mathrm{energy\_gain}(x) = \mathrm{energy}(x) - \hat{\mu}_{\mathrm{energy}}, \tag{27}$$

$$d\mathrm{oodscore\_gain}(x) = \mathrm{ood\_score}(x) - \hat{\mu}_{\mathrm{ood}}. \tag{28}$$

これをまとめて

$$s(x) = \begin{pmatrix} d\mathrm{conf\_drop}(x) \\ d\mathrm{entropy\_gain}(x) \\ d\mathrm{energy\_gain}(x) \\ d\mathrm{oodscore\_gain}(x) \end{pmatrix} \in R^4 \tag{29}$$

と書く．全サンプル分を並べた行列を

$$S = \begin{pmatrix} s(x_1)^\top \\ \vdots \\ s(x_N)^\top \end{pmatrix} \in R^{N \times 4} \tag{30}$$

とおく．

特徴ごとの分散 $\mathrm{Var}[S_{\cdot,j}]$ が極端に小さい場合は情報量が乏しいとみなし，閾値 $\tau_{\mathrm{var}} = 10^{-12}$ 未満の特徴を削除する（本実験では 4 変数すべてが閾値を上回り，削除される特徴はなかった）．

6.4 標準化と SparsePCA の入力行列

分散しきい値を通過した特徴の集合を $\mathcal{F} = \{1, 2, 3, 4\}$ とみなし，各列の平均 $\boldsymbol{\mu} \in R^4$ と標準偏差 $\boldsymbol{\sigma} \in R^4$ を

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} S_{i,j}, \tag{31}$$

$$\sigma_j^2 = \frac{1}{N} \sum_{i=1}^{N} (S_{i,j} - \mu_j)^2 \tag{32}$$

として計算する（$\sigma_j < 10^{-12}$ のときは $\sigma_j = 1$ とみなす）．

標準化行列 $X \in R^{N \times 4}$ を

$$X_{i,j} = \frac{S_{i,j} - \mu_j}{\sigma_j}, \quad 1 \le i \le N,\ 1 \le j \le 4 \tag{33}$$

として定義する．この X を SparsePCA の入力とする．

6.5 SparsePCA による潜在軸の学習

潜在軸数を $k \in \{1,2,3,4\}$，スパース性ハイパーパラメータを $\alpha \in \{0.5, 1.0, 2.0, 4.0, 8.0\}$ とし，SparsePCA による線形生成モデル

$$X \approx ZW \tag{34}$$

を学習する．ここで

$$Z \in R^{N\times k}, \quad W \in R^{k\times 4}$$

はそれぞれ潜在表現（スコア）と負荷行列である．scikit-learn の SparsePCA が解く目的関数は

$$\min_{Z,W} \frac{1}{2N}\|X - ZW\|_F^2 + \alpha \sum_{r=1}^{k} \|W_{r,\cdot}\|_1, \tag{35}$$

に相当するとみなせる（$\|\cdot\|_F$ は Frobenius ノルム）．第 2 項の ℓ_1 正則化により，各軸の負荷 $W_{r,\cdot}$ のスパース性が促され，どの差分特徴がどの軸に効いているかを解釈しやすくする．

本研究では，ID/OoD ラベルに基づく層化 K 分割（$K = 5$）交差検証を用いて (k, α) を選択する．層化分割を $(\mathcal{I}_{\text{train}}^{(\ell)}, \mathcal{I}_{\text{val}}^{(\ell)})$（$\ell = 1, \ldots, K$）とし，fold ℓ における訓練・検証行列を

$$X_{\text{train}}^{(\ell)} = X_{\mathcal{I}_{\text{train}}^{(\ell)},\cdot}, \tag{36}$$

$$X_{\text{val}}^{(\ell)} = X_{\mathcal{I}_{\text{val}}^{(\ell)},\cdot} \tag{37}$$

とする．各 (k, α) について，(35) を訓練集合で最適化し，得られた $(Z^{(\ell)}, W^{(\ell)})$ から検証集合の再構成誤差

$$\text{MSE}^{(\ell)}(k,\alpha) = \frac{1}{|\mathcal{I}_{\text{val}}^{(\ell)}|} \left\| X_{\text{val}}^{(\ell)} - Z_{\text{val}}^{(\ell)} W^{(\ell)} \right\|_F^2 \tag{38}$$

を計算する．これを fold 全体で平均した

$$\overline{\text{MSE}}(k,\alpha) = \frac{1}{K}\sum_{\ell=1}^{K} \text{MSE}^{(\ell)}(k,\alpha), \tag{39}$$

$$\text{SE}(k,\alpha) = \sqrt{\frac{1}{K(K-1)}\sum_{\ell=1}^{K}\left(\text{MSE}^{(\ell)}(k,\alpha) - \overline{\text{MSE}}(k,\alpha)\right)^2} \tag{40}$$

を交差検証指標とする．

1SE ルールによるモデル選択 $\overline{\text{MSE}}(k,\alpha)$ を最小にする点を

$$(k_*, \alpha_*) = \arg\min_{k,\alpha} \overline{\text{MSE}}(k,\alpha)$$

とおき，$\overline{\text{MSE}}_* = \overline{\text{MSE}}(k_*, \alpha_*)$，$\text{SE}_* = \text{SE}(k_*, \alpha_*)$ とする．1SE ルールでは，

$$T_{\text{1SE}} = \overline{\text{MSE}}_* + \text{SE}_*$$

を閾値とし，$\overline{\text{MSE}}(k,\alpha) \le T_{\text{1SE}}$ を満たす候補のうち，最小の k を持つものを採用する：

$$k_{\text{sel}} = \min\left\{k \;\middle|\; \exists\alpha,\ \overline{\text{MSE}}(k,\alpha) \le T_{\text{1SE}}\right\}. \tag{41}$$

実験設定では，交差検証の結果として

$$k_{\text{sel}} = 3, \qquad \alpha_{\text{sel}} = 1.0$$

が選択された．このときの最小平均 MSE および標準誤差はそれぞれ

$$\overline{\text{MSE}}_* \approx 9.80 \times 10^{-5}, \quad \text{SE}_* \approx 5.01 \times 10^{-7}$$

であり，$T_{\text{1SE}} \approx 9.85 \times 10^{-5}$ となる．

6.6 Attribution-OoD 軸の定義と具体式

選択された $(k_{\text{sel}}, \alpha_{\text{sel}})$ を用いて全データ X に SparsePCA を適用し，負荷行列 $W \in R^{3\times 4}$ とスコア行列 $Z \in R^{N\times 3}$ を得る．軸 $r \in \{1,2,3\}$ に対応する負荷ベクトルを

$$W_r = (w_{r,1}, w_{r,2}, w_{r,3}, w_{r,4}) \in R^4$$

と書く（添字 1〜4 はそれぞれ $d$conf_drop, $d$entropy_gain, $d$energy_gain, $d$oodscore_gain に対応）．すると，各サンプルの軸座標は

$$z_r(x_i) = (Z_{i,r}) = \sum_{j=1}^{4} w_{r,j}\,\frac{S_{i,j} - \mu_j}{\sigma_j} \tag{42}$$

として与えられる（実装では標準化済み X を直接入力としている）．

軸候補のスコアリング 不確実性関連特徴の添字集合を

$$\mathcal{F}_u = \{2,3,4\} \quad \text{(entropy\_gain, energy\_gain, oodscore\_gain)},$$

確信度関連特徴の添字を

$$\mathcal{F}_c = \{1\} \quad \text{(conf\_drop)}$$

とする．各軸 r に対し

$$s_u(r) = \sum_{j\in\mathcal{F}_u} |w_{r,j}|, \tag{43}$$

$$s_c(r) = \sum_{j\in\mathcal{F}_c} |w_{r,j}| = |w_{r,1}| \tag{44}$$

を定義し，

$$r_u = \arg\max_r s_u(r), \qquad r_c = \arg\max_{r\neq r_u} s_c(r)$$

をそれぞれ「不確実性軸候補」「確信度軸候補」として選ぶ（もし $r_u = r_c$ となる場合は $s_c(r)$ が 2 番目に大きい軸を r_c とする）．

符号の整合性 軸の正方向を「不確実性が増大する方向」「確信度が低下する方向」と揃えるため，

$$w_{r_u,2} < 0 \;\Rightarrow\; W_{r_u,\cdot} \leftarrow -W_{r_u,\cdot},\ Z_{\cdot,r_u} \leftarrow -Z_{\cdot,r_u},$$

$$w_{r_c,1} < 0 \;\Rightarrow\; W_{r_c,\cdot} \leftarrow -W_{r_c,\cdot},\ Z_{\cdot,r_c} \leftarrow -Z_{\cdot,r_c}$$

という符号反転を行う．

Attribution-OoD　軸の定義　以上を踏まえ，Attribution-OoD 軸 (z_u, z_c) を

$$z_u(x_i) = Z_{i,r_u}, \tag{45}$$
$$z_c(x_i) = Z_{i,r_c} \tag{46}$$

として定義する．

実験設定では，最終的な軸の具体式は差分特徴空間で

$$z_u(x) = d\text{entropy\_gain}(x), \tag{47}$$
$$z_c(x) \approx \frac{1}{\sqrt{2}}\, d\text{conf\_drop}(x) + \frac{1}{\sqrt{2}}\, d\text{oodscore\_gain}(x) \tag{48}$$

にほぼ一致した．実装上は数値丸めの結果として

$$z_u(x) = 1.0000\, d\text{entropy\_gain}(x),$$
$$z_c(x) = 0.7071\, d\text{conf\_drop}(x) + 0.7071\, d\text{oodscore\_gain}(x).$$

と保存されている（$0.7071 \approx 1/\sqrt{2}$）．

6.7　OoD スコアと評価指標の定義

Attribution-OoD 軸を用いた OoD 判定器およびベースラインを，スコア関数

$$s : \mathcal{X} \to R$$

として定義する．

ベースラインスコア

- MSP 型スコア：

$$s_{\text{msp}}(x) = 1 - \text{conf}(x).$$

- エネルギースコア：

$$s_{\text{energy}}(x) = \text{energy}(x).$$

Attribution-OoD 軸に基づくスコア

- 一次元合成：訓練分割に対して $\tilde{z}_u, \tilde{z}_c$ を標準化し，

$$s_{\text{zsum}}(x) = \tilde{z}_u(x) + \tilde{z}_c(x)$$

を OoD スコアとみなす．

- 2 次元ロジスティック回帰：特徴 (z_u, z_c) に対しロジスティック回帰を学習し，OoD クラス（$y = 1$）に属する確率

$$s_{\text{logistic}}(x) = \Pr(y = 1 \mid z_u(x), z_c(x))$$

をスコアとする．

- 2 次元線形 SVM：特徴 (z_u, z_c) に対し線形 SVM を学習し，決定関数値

$$s_{\text{svm}}(x) = f_{\text{svm}}(z_u(x), z_c(x))$$

をスコアとする（符号は OoD 側が大きくなるように揃える）．

ROC 曲線・AUROC・TNR@95TPR の定義　スコア関数 $s(x)$ に対し，閾値 t ごとの真陽性率（TPR）と偽陽性率（FPR）を

$$\text{TPR}(t) = \Pr\big(s(x) \geq t \mid y = 1\big), \tag{49}$$
$$\text{FPR}(t) = \Pr\big(s(x) \geq t \mid y = 0\big) \tag{50}$$

と定義する．t を変化させた $(\text{FPR}(t), \text{TPR}(t))$ の軌跡が ROC 曲線となる．

ROC 曲線下面積（AUROC）は

$$\text{AUROC} = \int_0^1 \text{TPR}\big(\text{FPR}^{-1}(u)\big)\, du \tag{51}$$

として定義される（実装では離散近似）．

さらに，本研究では実運用を意識した指標として，$\text{TPR} \geq 0.95$ を達成する最小 FPR を求め

$$\text{TNR@95TPR} = 1 - \min_{t:\, \text{TPR}(t) \geq 0.95} \text{FPR}(t) \tag{52}$$

を用いる．

ブートストラップによる信頼区間　テスト分割のラベル列 $\mathbf{y} \in \{0,1\}^{n_{\text{test}}}$ とスコア列 $\mathbf{s} \in R^{n_{\text{test}}}$ に対し，評価関数 $m(\mathbf{y}, \mathbf{s})$（AUROC または TNR@95TPR）についてブートストラップ再標本化（反復数 $B = 1000$）により 95% 信頼区間を推定する．

1. 各反復 $b = 1, \ldots, B$ で，インデックス集合 $\{1, \ldots, n_{\text{test}}\}$ から大きさ n_{test} のブートストラップサンプル $\mathcal{J}^{(b)}$ を一様復元抽出する．
2. $\mathbf{y}^{(b)} = (y_j)_{j \in \mathcal{J}^{(b)}}, \mathbf{s}^{(b)} = (s_j)_{j \in \mathcal{J}^{(b)}}$ を作り，$m^{(b)} = m(\mathbf{y}^{(b)}, \mathbf{s}^{(b)})$ を計算する．
3. $\{m^{(b)}\}_{b=1}^{B}$ の経験分布から 2.5 パーセンタイルと 97.5 パーセンタイルを取り，それぞれ下側・上側 95% 信頼限界とする．

この手順により，各モデルの AUROC と TNR@95TPR について (point estimate, 95% CI) を一貫した方法で報告できる．

6.8　ID / OoD 間の分布差と効果量の定義

Attribution-OoD 軸および元スコアが，ID と OoD の間で統計的にどの程度異なるかを評価するため，各指標 $m(x)$ に対して以下を定義する．

マンホイットニーの U 検定　ID 集合 $\{m(x_i) \mid y_i = 0\}$ と OoD 集合 $\{m(x_i) \mid y_i = 1\}$ を比較するノンパラメトリック検定として，マンホイットニーの U 検定を用いる．検定統計量 U は，OoD サンプルが ID サンプルより大きい順位をどの程度取るかを表し，p 値 p_U から有意差を判断する．

さらに，順位に基づく効果量として，ID・OoD の標本サイズをそれぞれ $n_{\text{id}}, n_{\text{ood}}$ とすると

$$\delta_{\text{rank}} = 2\frac{U}{n_{\text{id}} n_{\text{ood}}} - 1 \tag{53}$$

を用いる．これは「ランダムに 1 つずつ OoD と ID をとったとき，OoD の方が大きい確率」と解釈できる量の線形変換に相当し，$\delta_{\text{rank}} > 0$ なら OoD の方が大きい方向にずれていることを意味する．

Kolmogorov–Smirnov 検定　分布全体の形状差を捉えるため，累積分布関数の最大差

$$D = \sup_t |F_{\text{ood}}(t) - F_{\text{id}}(t)| \tag{54}$$

を用いる Kolmogorov–Smirnov 検定を併用し，p 値 p_KS から分布差の有意性を評価する．

Cohen の d　平均値の差を分散で正規化した効果量として，ID・OoD の標本平均と標本分散をそれぞれ $\bar{m}_{\text{id}}, \bar{m}_{\text{ood}}, s^2_{\text{id}}, s^2_{\text{ood}}$ とすると，

$$s^2_{\text{pool}} = \frac{1}{2}\left(s^2_{\text{id}} + s^2_{\text{ood}}\right), \tag{55}$$

$$d_{\text{cohen}} = \frac{\bar{m}_{\text{ood}} - \bar{m}_{\text{id}}}{\sqrt{s^2_{\text{pool}} + 10^{-12}}} \tag{56}$$

を Cohen の d として用いる．$d_{\text{cohen}} > 0$ なら OoD の平均が ID より大きい方向にシフトしていることを意味する．

これらの指標を指標ごとにまとめたものが，ID と OoD の分布差を定量化した表 `significance_id_vs_ood.csv` に対応する．

以上が，本研究における VLM の信頼度指標，差分特徴，SparsePCA による潜在軸学習，Attribution-OoD 軸 (z_u, z_c)，および OoD 検知性能指標・統計的効果量の数理的な定義である．

7　実装結果

本節では，前節までに定義した 4 次元差分特徴 $(d\text{conf\_drop}, d\text{entropy\_gain}, d\text{energy\_gain}, d\text{oodscore\_gain})$ から構成した Attribution-OoD 軸 (z_u, z_c) の具体的な学習結果と，それを用いた OoD 検知性能，誤分類パターン，および seed・プロンプト・バックボーンに対するロバスト性を報告する．

7.1　SparsePCA による軸構築の結果

まず，Food-101，CIFAR-100，ImageNet-R の計 $N = 20{,}000$ サンプルから得られた 4 次元差分特徴行列 $X \in R^{N\times 4}$ に対し，潜在次元 $k \in \{1,2,3,4\}$，スパース性ハイパーパラメータ $\alpha \in \{0.5, 1.0, 2.0, 4.0, 8.0\}$ を格子探索した SparsePCA を適用した．ID/OoD ラベルに基づく層化 5 分割交差検証により，各 (k,α) について検証再構成誤差 $\overline{\text{MSE}}(k,\alpha)$ と標準誤差 $\text{SE}(k,\alpha)$ を算出し，最小点 (k_*, α_*) とその 1SE しきい値 $\overline{\text{MSE}}_* + \text{SE}_*$ を用いる 1SE ルールでモデルを選択した（**図 1**）．

その結果，本実験条件（ViT-B/32, `laion2b_s34b_b79k`, `"a photo of {name}"`, seed=42）では

$$k_{\text{sel}} = 3, \qquad \alpha_{\text{sel}} = 1.0$$

が選択された．このときの平均検証 MSE は $\overline{\text{MSE}}_* \approx 9.80 \times 10^{-5}$，標準誤差は $\text{SE}_* \approx 5.01 \times 10^{-7}$ であり，しきい値は

$$T_{\text{1SE}} = \overline{\text{MSE}}_* + \text{SE}_* \approx 9.85 \times 10^{-5}$$

となる．1SE ルールにより，より大きな k を選べばわずかに MSE を下げられるものの，$k = 3$ で十分に再構成精度

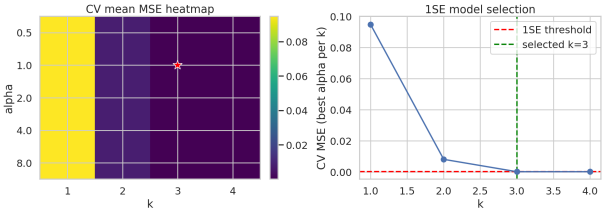

図 1:　SparsePCA の次元数 k とスパース係数 α に対する再構成誤差の交差検証結果と，1SE ルールに基づくモデル選択．

が飽和していることが確認できる（**図 1** の黒線と赤破線の関係）．

選択された SparsePCA モデルの負荷行列 $W \in R^{3\times 4}$ から，「不確実性寄与の大きい軸」（entropy/energy/oodscore 方向の係数が大きい軸）を r_u，「確信度寄与の大きい軸」（conf_drop 方向の係数が大きい軸）を r_c として抽出し，それぞれを z_u, z_c と定義した．符号を entropy_gain（不確実性）および conf_drop（確信度低下）の正方向と揃えるために，必要に応じて軸方向の符号反転を行っている．

その結果，4 変数版では軸の具体式が差分特徴空間において

$$z_u(x) = d\text{entropy\_gain}(x), \tag{57}$$

$$z_c(x) \approx \frac{1}{\sqrt{2}}\, d\text{conf\_drop}(x) + \frac{1}{\sqrt{2}}\, d\text{oodscore\_gain}(x) \tag{58}$$

とほぼ一致することがわかった．実際に保存された係数は

$$z_u(x) = +1.0000 \cdot d\text{entropy\_gain}(x),$$

$$z_c(x) = +0.7071 \cdot d\text{conf\_drop}(x) + 0.7071 \cdot d\text{oodscore\_gain}(x)$$

であり（$0.7071 \approx 1/\sqrt{2}$），$z_u$ は純粋に entropy の増加方向を表す軸，z_c は「softmax 信頼度の低下」と「MSP 型 OoD スコアの増加」を等重みで合成した軸として解釈できる（**図 ??**）．

7.2　OoD 検知性能の比較

構築した軸 (z_u, z_c) を用いて，以下の 5 種類の OoD スコアを比較した：

- **msp_single**：$s_{\text{msp}}(x) = 1 - \text{conf}(x)$.
- **energy_single**：$s_{\text{energy}}(x) = \text{energy}(x)$.
- **zsum_1d**：訓練分割に対して標準化した $\tilde{z}_u(x), \tilde{z}_c(x)$ の和 $s_{\text{zsum}}(x) = \tilde{z}_u(x) + \tilde{z}_c(x)$.
- **logistic_2d**：特徴 (z_u, z_c) に対して学習した 2 次元ロジスティック回帰の $\Pr(y = 1 \mid z_u, z_c)$.
- **linear_svm**:特徴 (z_u, z_c) に対して学習した線形 SVM の決定関数値 $f_{\text{svm}}(z_u, z_c)$（符号は OoD 側が大きくなるように揃える）．

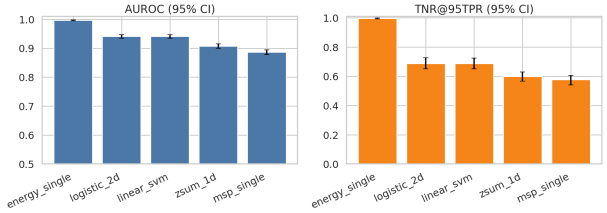

図 2: 各検出器における AUROC および TNR@95TPR の比較（ブートストラップに基づく 95% 信頼区間付き）.

全データを ID/OoD で層化したうえで，70% を訓練，30% をテストとする分割を 1 回固定し，テスト分割に対して ROC 曲線と AUROC，および TPR ≥ 0.95 を満たす点での TNR@95TPR を評価した．さらにブートストラップ再標本化（反復数 1,000）により 95% 信頼区間を推定した結果を**図 2** および **図 3** に示す.

代表的な指標値は以下のとおりである：

- **energy_single**：AUROC $\approx$ 0.996 （95% CI: $[0.995, 0.997]$），TNR@95TPR $\approx$ 0.996 （95% CI: $[0.993, 0.998]$）．非常に高い分離性能を示し，energy スコアがこの設定ではほぼ上限に近い OoD 判定性能を持つことが確認できる.
- **logistic_2d**（Attribution-OoD）：AUROC $\approx$ 0.941 （95% CI: $[0.935, 0.946]$），TNR@95TPR $\approx$ 0.687 （95% CI: $[0.649, 0.726]$）.
- **linear_svm**：AUROC および TNR@95TPR は logistic_2d とほぼ同等であり，線形境界のクラス分離が 2 次元平面上でほぼ飽和していることを示唆する.
- **zsum_1d**：AUROC $\approx$ 0.907 （95% CI: $[0.900, 0.914]$），TNR@95TPR $\approx$ 0.599 （95% CI: $[0.567, 0.633]$）．単純な直線射影による 1 次元スコアであっても，MSP 単独より一段高い OoD 検知性能を達成している.
- **msp_single**：AUROC $\approx$ 0.886 （95% CI: $[0.878, 0.894]$），TNR@95TPR $\approx$ 0.578 （95% CI: $[0.541, 0.607]$）.

以上より，Attribution-OoD 軸を用いた 2 次元線形判別（logistic_2d, linear_svm）は，MSP 単独よりも一貫して高い AUROC／TNR@95TPR を達成しており，energy ベースの強力なベースラインには及ばないものの，「どの方向の変化で OoD と判定されたのか」を説明可能な 2 軸表現を保ったまま，実用的な判定性能を維持できていることがわかる.

7.3 z_u–z_c 平面上の挙動と誤りパターン

Attribution-OoD 軸の幾何学的な挙動を確認するため，全サンプルの一部（最大 12,000 枚）を抽出し，ID/OoD を色分けした z_u–z_c 平面上のカーネル密度等高線を描画した（**図 5**）．同図から，概ね以下のような構造が観測された：

- $z_u \geq 0, z_c \geq 0$ の第 1 象限（Q1）は OoD サンプルが高密度で，ID サンプルは少ない．ここでは entropy の増大と confidence の低下が同時に起きているケースが主となる.

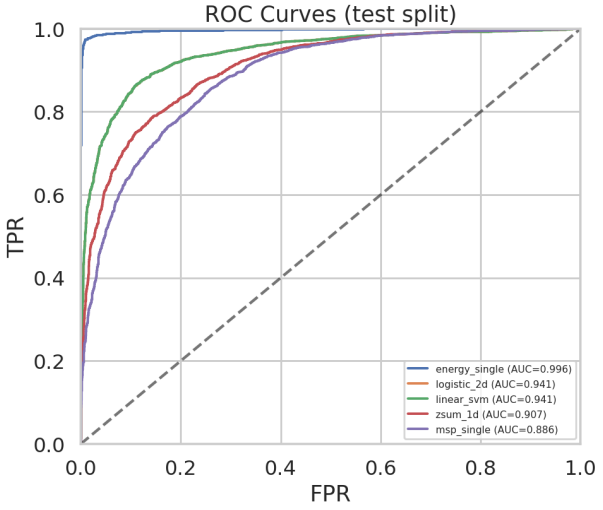

図 3: 各検出器の ROC 曲線（テスト分割）.

- $z_u < 0, z_c < 0$ の第 3 象限（Q3）は ID サンプルが支配的であり，entropy が低く，confidence も高い「典型的な ID 領域」として機能している.
- 第 2 象限（Q2）および第 4 象限（Q4）は，片方の軸のみが大きく変化しているケースを表し，決定境界近傍に ID/OoD が混在する「曖昧な領域」となっている.

さらに，(z_u, z_c) を特徴とするロジスティック回帰（logistic_2d）を用いたときのテスト分割における誤判定を象限別に集計すると，第 3 象限（ID 優勢領域）の内部に OoD の偽陰性（FN）が，第 1 象限および第 4 象限（OoD 優勢領域）の境界付近に ID の偽陽性（FP）が集中する傾向が見られた（**図 8**）．これは，「entropy は上がっていないが confidence が微妙に落ちている ID」や，「entropy は大きく増加しているが，もともと類似したクラス間での揺らぎである OoD」など，VLM の振る舞いとして解釈しやすい誤りパターンを 2 次元平面上で可視化できていることを示している.

7.4 seed・プロンプト・バックボーンに対するロバスト性

次に，Attribution-OoD 軸が学習の乱数 seed，テキストプロンプト，および CLIP バックボーンの違いに対してどの程度安定であるかを検証した.

具体的には，

- seed $\in \{42, 43, 44\}$,
- prompt $\in$ {`"a photo of {name}"`, `"a close-up photo of {name}"`},
- model $\in$ {`ViT-B-32/laion2b_s34b_b79k`, `ViT-B-16/laion2b_s34b_b88k`}

の直積 12 通りの条件それぞれについて，同じ 4 変数差分特徴から SparsePCA + 1SE ルールによる軸構築を行った.

その結果，いずれの条件でも

$$|\cos(z_u, z_u^{\mathrm{ref}})| \approx 1.0, \qquad |\cos(z_c, z_c^{\mathrm{ref}})| \approx 1.0$$

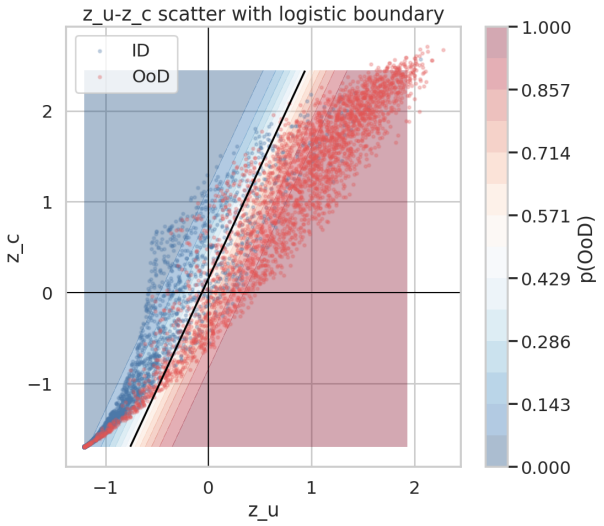

図 4: z_u–z_c 平面上での ID / OoD サンプルの散布図と，ロジスティック回帰による決定境界．

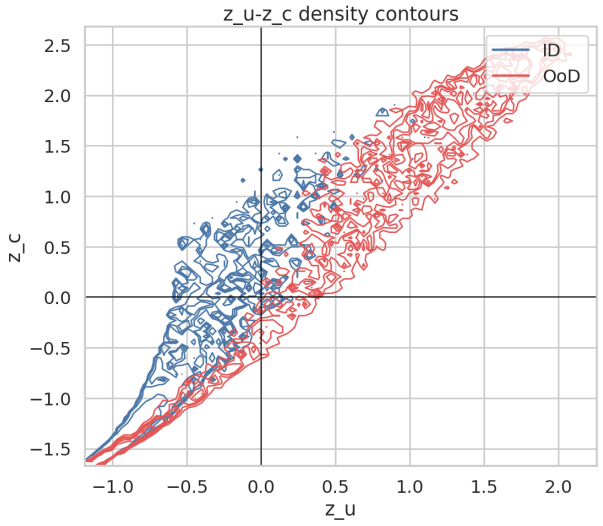

図 5: z_u–z_c 平面における ID / OoD のカーネル密度等高線．四象限ごとの分布の違いが可視化されている．

となり，平均値はモデルごと，またバックボーン横断でも有効数字 3 桁レベルで 1.0 に一致した．すなわち，本研究で導入した「entropy 由来の不確実性軸」z_u と「confidence/MSP 由来の確信度軸」z_c は，乱数初期化・プロンプト文言・ViT のパッチ解像度といった実装上の選択に対して極めて安定である．

この結果は，Attribution-OoD を VLM の「固定座標系」として扱うという設計意図が，少なくとも本実験条件の範囲では達成されていることを示している．異なる backbone や prompt を用いた場合でも，「どの方向が不確実性増大を表すか」「どの方向が確信度低下を表すか」がほぼ同じ線形結合として再現されるため，下流タスクにおいて軸そのものを再学習する必要がない可能性が高い．

7.5 ID/OoD 間の分布差と効果量

最後に，Attribution-OoD 軸および元スコアが ID と OoD の間でどの程度分離しているかを，統計的検定と効果量の観点から整理した．

各指標 $\{z_u, z_c, \mathrm{conf}, \mathrm{entropy\_norm}, \mathrm{energy}, \mathrm{ood\_score}\}$ について，ID 集合と OoD 集合の間でマンホイットニーの U 検定と Kolmogorov–Smirnov 検定を行ったところ，いずれの指標においても p 値は事実上 0 に近く，ID/OoD の分布差が統計的に有意であることが確認された．

また，順位に基づく効果量 δ_{rank} と Cohen の d を指標ごとに整理した結果を**図 10** に示す．おおまかには，以下の傾向がみられる：

- z_u および energy は，OoD 側で大きくなる指標として大きな正の効果量を持ち，ID/OoD の分布が大きくシフトしている．
- z_c や ood_score も正の効果量を示し，confidence/MSP 系の変化が OoD 側で一貫して大きいことを反映している．
- conf は ID 側で大きいため，効果量は負の方向に大きな値をとる．
- entropy_norm 単独よりも z_u の方が効果量がやや大きくなる傾向があり，entropy の増加分を ID 平均からの差分として扱うことが分離性の向上に寄与している可能性がある．

これらの結果から，Attribution-OoD 軸 (z_u, z_c) は，単に視覚的に解釈しやすい 2 次元表示を提供するだけでなく，統計的にも ID/OoD の分離度が高い指標となっていることが分かる．特に z_u は entropy 系の変化を集約した「不確実性軸」として OoD 側への強いシフトを示し，z_c は confidence/MSP の変化を統合した「確信度軸」として決定境界近傍の曖昧なケースを捉える役割を果たしている．

8 結論・考察・今後

8.1 結論

本研究では，Vision-Language Model（VLM）の推論ログから，スカラー不確実性指標の差分のみを用いて 2 つの Attribution-OoD 軸 (z_u, z_c) をデータ駆動で構成する手法を示した．Food-101 を ID，CIFAR-100 および ImageNet-R を OoD とする設定で OpenCLIP（ViT-B/32）を用いて実験した結果，以下の点が明らかになった．

- ID の平均からの差分として定義した $d\mathrm{conf\_drop}, d\mathrm{entropy\_gain}, d\mathrm{energy\_gain}, d\mathrm{oodscore\_gain}$ の 4 変数を入力とし，SparsePCA と 1SE ルールにより潜在次元 $k_{\mathrm{sel}} = 3$，$\alpha_{\mathrm{sel}} = 1.0$ を選択することで，再構成誤差 $\mathcal{O}(10^{-4})$ 程度の低い誤差で差分特徴を要約できることを確認した．
- SparsePCA の負荷行列から抽出した 2 軸は，実質的に

$$z_u(x) = d\mathrm{entropy\_gain}(x),$$
$$z_c(x) \approx \frac{1}{\sqrt{2}}\{d\mathrm{conf\_drop}(x) + d\mathrm{oodscore\_gain}(x)\}.$$

と同等となり,「ID 平均からの entropy 増加量」と「confidence 低下＋ MSP 型 OoD スコア増加」の合成という解釈しやすい 2 軸が自動的に得られることが分かった.

- 構成した (z_u, z_c) を用いたロジスティック回帰（logistic_2d）は，テスト分割上で AUROC $\approx$ 0.94, TNR@95TPR $\approx$ 0.69 を達成し，MSP 単独（AUROC ≈ 0.89, TNR@95TPR ≈ 0.58）を一貫して上回る OoD 検知性能を示した．単純な 1 次元射影 $z_u + z_c$ であっても MSP を上回ることから，ID 平均からの差分として 2 方向の変化を分離すること自体に一定の効果があると考えられる．
- 一方で，energy 単独は AUROC $\approx$ 0.996, TNR@95TPR $\approx$ 0.996 ときわめて高い性能を示し，本設定では依然として最も強力な OoD スコアであることが確認された．提案軸は energy に匹敵する性能を目指すものではなく，energy ほどの性能を失わずに「どの方向の変化によって OoD と判定されたのか」を説明可能にする補助座標として機能する．
- seed，テキストプロンプト，および ViT バックボーン（ViT-B/32 と ViT-B/16）の組合せ 12 条件すべてに対して同様の処理を行ったところ，得られた z_u, z_c の負荷ベクトルは基準条件との絶対コサイン類似度がほぼ 1.0 に一致した．すなわち，本研究で得られた「不確実性軸」 z_u と「確信度軸」 z_c は，実装上の選択に対して極めて安定な VLM の内部座標として振る舞うことが示された．

総じて，本研究は，VLM の推論ログから,「不確実性の増大」と「確信度の低下」という 2 つの観点で OoD 振る舞いを可視化・分析するための固定 2 次元軸を構成できることを示し，その軸が実用的な OoD 検知性能と高いロバスト性を兼ね備えていることを確認した．

8.2 考察

energy との関係と役割の違い 実験結果から，energy 単独が最も高い OoD 分離性能を示した一方で，提案した Attribution-OoD 軸は energy に比べて AUROC, TNR@95TPR の両面で劣ることが明らかになった．これは，energy が softmax 正規化前の logit 空間全体を利用した情報量の高いスコアであるのに対し，本研究の軸は $\mathrm{conf}, \mathrm{entropy}, \mathrm{ood\_score}$ といったスカラー指標の差分にのみ依存しているため，必然的に情報量が制限されていることに起因する．

しかし，energy は「なぜこのサンプルが OoD と判定されたのか」を構造的に分解することが難しいのに対し，本研究の軸は

- entropy 側の変化（ラベル分布の拡散）としての OoD,
- confidence 側の変化（最尤クラスへの信頼度低下）としての OoD

を明示的に分離して扱える点に本質的な価値がある．実際に z_u–z_c 平面上では，ID の典型領域（低 entropy・高 confidence）が第 3 象限に集まり，entropy と confidence が同時に悪化する OoD が第 1 象限に偏る一方，片方のみが変化する曖昧なサンプルが第 2 象限・第 4 象限に分布する構造が確認された．この幾何構造は,「どのタイプの不確実性が増大した結果として OoD 判定になったのか」を人間が理解するうえで有用であり，energy スコア単独では得にくい説明性を提供する．

差分表現の意義 本研究では，絶対値としての $\mathrm{conf}, \mathrm{entropy}$ を直接用いるのではなく，ID における平均からの差分 $(d\mathrm{conf\_drop}, d\mathrm{entropy\_gain}, \dots)$ を用いて特徴空間を定義した．統計的検定および効果量の分析から，これらの差分指標は ID/OoD 間で大きな分布差を持つことが確認された．特に z_u は，単純な entropy_norm よりもやや大きな効果量を示しており,「ID 基準からどれだけ外れたか」という相対値として扱うことが，異常度の測定に有利に働いている可能性がある．

また，差分を取ることにより，同一 VLM を異なるドメインに適用した場合でも「ID 側の基準」が自動的に再定義されるため，異なるデータセット間で軸の意味付けを揃えやすいという利点がある．本実験では Food-101 を固定の ID としたが，実運用においては「その場で観測された ID ログ」から都度 baseline を推定する形で同じ手順を適用できる．

ロバスト性評価の意味 seed・プロンプト・バックボーンを変えても z_u, z_c の負荷ベクトルがほぼ完全に一致したことは，本手法が単なる「一度きりの分解」ではなく，VLM の予測構造に内在する安定な 2 軸を抽出していることを示唆する．

特に，ViT-B/32 と ViT-B/16 のようにパッチ解像度やパラメータ規模が異なるバックボーン間でも軸の向きが変化しなかった事実は，本研究で用いた 4 変数差分が「アーキテクチャ固有の細部」よりも「softmax 出力の統計構造」に強く依存していることを意味する．これは，将来的に異なる VLM 群に対しても同一の座標系で OoD 振る舞いを比較する，という応用に向けた重要な性質である．

限界と課題 一方で，本研究にはいくつかの明確な限界が存在する：

- ID/OoD の組み合わせが（Food-101, CIFAR-100, ImageNet-R）に固定されており，実世界に近い長尾分布や複雑な共変量シフト，言語側のノイズを含むタスクでの検証は行っていない．
- 差分特徴は 4 次元に限定しており，ロジット分布の高次モーメントや，テキストエンコーダ側の不確実性，注意分布などは利用していない．より豊かな特徴を含めれば，energy に近い性能と説明性の両立が期待できる一方，軸の解釈性を維持できるかは検討が必要である．
- 軸構築の内部で用いている SparsePCA は，あくまで線形モデルであり，非線形な構造が強い場合にはより柔軟な方法（例：スパースなオートエンコーダ）への拡張が有効となる可能性がある．

これらの点は，本手法を「VLM 一般の OoD 解析ツール」として位置づけるために今後解決すべき課題である．

8.3 今後の展望

本研究で示した Attribution-OoD 軸は，VLM の OoD 振る舞いを「不確実性」と「確信度」の 2 本の軸に射影するための基盤的な枠組みであり，今後，より実践的な文脈に拡張していく余地が大きい．具体的な展望として，以下を挙げる．

実タスクログへの適用　本研究では，ベンチマークデータセットに対する分類タスクを対象としたが，実運用を想定すると，

- 複数プロンプト・複数画像のバッチ推論ログ，
- 自然言語質問応答やキャプション生成タスクにおける内部スコア，
- フィールドで収集された長期間のログ

など，より多様な形態のログに適用する必要がある．Attribution-OoD 軸をそのまま適用し，「時間経過とともにどの象限の OoD が増えているか」，「特定のプロンプトや入力分布でどのタイプの異常が多いか」といったログ解析に用いることで，運用上のリスク評価やモニタリングへの応用が期待できる．

軸の拡張と意味づけの精緻化　本研究では 2 軸に絞ったが，Full-Spectrum 型の枠組みと組み合わせると，

- semantic global / semantic near，
- covariate（ドメイン）シフト，
- 動的ノイズ（動画・連続フレームにおける揺らぎ）

といった複数種類のシフトを，Attribution-OoD 空間上のサブスペースとして扱う設計も考えられる．例えば，z_u と z_c に加えて「ドメイン変化軸」や「時系列スパース性軸」を追加し，4 次元程度の Attribution-OoD 空間を構成することで，よりリッチなシフト分類と可視化が可能になるだろう．

また，差分特徴の候補として，

- 上位 k クラスの確率質量集中の変化，
- テキストエンコーダ側の類似度スコアの変化，
- マルチモーダル類似度（image–text）のギャップ

などを追加し，SparsePCA の入力空間を拡張することで，より多様な「Attribution 軸」を自動抽出する方向性も考えられる．

検知器との統合・学習的利用　現状の Attribution-OoD 軸は，(i) VLM 本体は固定，(ii) 軸は post-hoc に学習，(iii) そのうえで単純な線形判別器を重ねる，という構成である．これを発展させ，

- Attribution-OoD 軸上で OoD 分布がより明確に分離するように VLM 側を追加学習する，
- 軸上の位置に応じて損失関数や正則化項を切り替える（例：semantic shift 優位な領域では別種の正則化をかける），
- 軸を条件変数として用いる生成モデルと組み合わせ，「どのタイプの OoD がどの程度出現し得るか」を事前にシミュレーションする

といった形で，学習過程そのものに Attribution-OoD 軸を組み込む方向性も考えられる．

付録

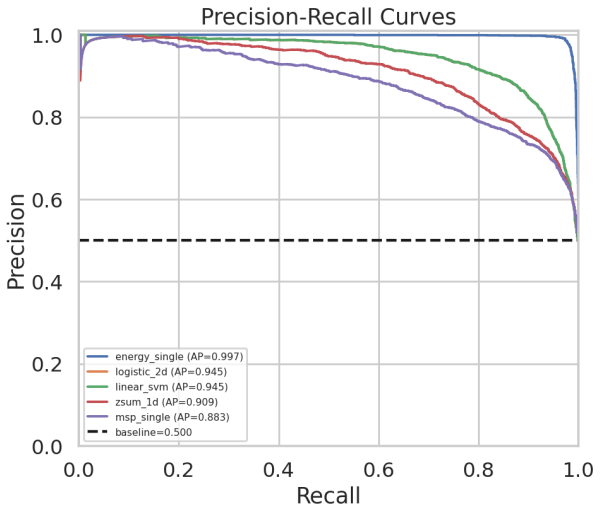

図 6:　各検出器の Precision–Recall 曲線（テスト分割）．

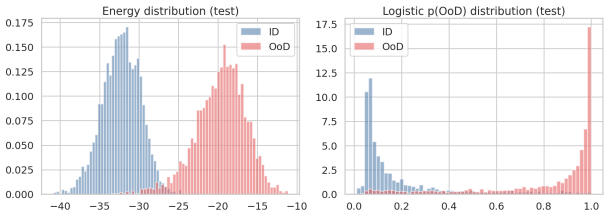

図 7:　Energy スコアおよび logistic_2d スコアにおける ID / OoD の一変量分布．

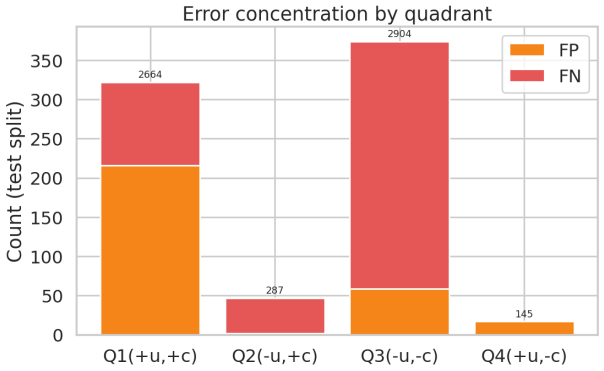

図 8:　z_u–z_c 平面の四象限ごとに集計した，誤検出（FP）および検出漏れ（FN）の件数（テスト分割に対する結果）．

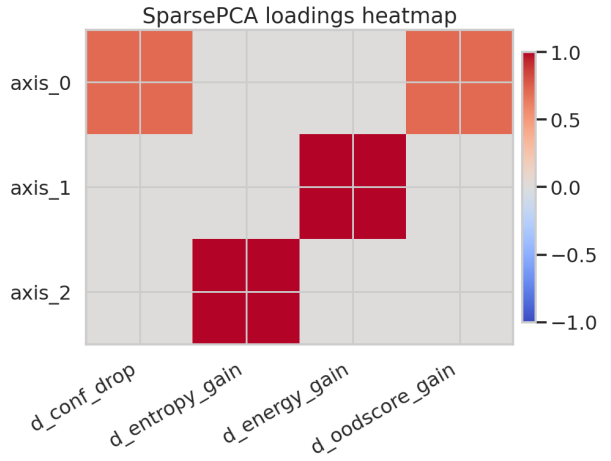

図 9: SparsePCA により得られた各軸のロードニング（特徴量 $\{d\text{conf\_drop}, d\text{entropy\_gain}, d\text{energy\_gain}, d\text{oodscore\_gain}\}$ と学習された軸との対応関係を示すヒートマップ）.

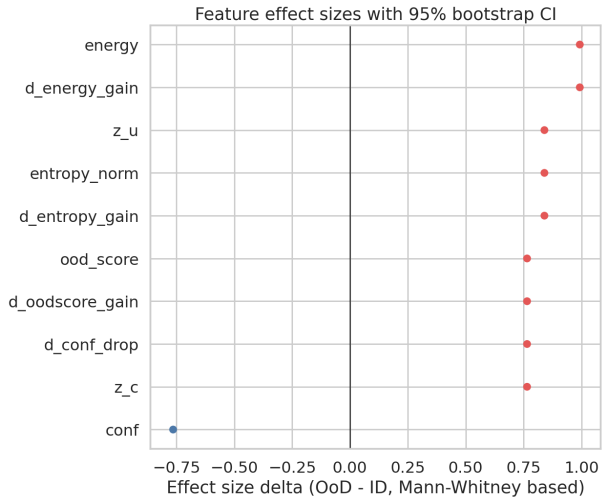

図 10: 各スコアについて ID / OoD 間の差を Mann–Whitney 検定に基づく effect delta で可視化した結果.

参考文献

[1] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 21464–21475. Curran Associates, Inc., 2020.

[2] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, Vol. 34, pp. 144–157. Curran Associates, Inc., 2021.

[3] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930, June 2022.

[4] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, Vol. 35, pp. 32598–32611. Curran Associates, Inc., 2022.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022.

[7] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Computer Vision – ECCV 2022*, Vol. 13695 of *Lecture Notes in Computer Science*, pp. 493–510. Springer, 2022.

[8] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, Vol. 35, pp. 35087–35102. Curran Associates, Inc., 2022.

[9] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. CLIPood: Generalizing CLIP to out-of-distributions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 31716–31731. PMLR, 2023.

[10] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *Advances in Neural Information Processing Systems*, Vol. 36, pp. 76298–76310. Curran Associates, Inc., 2023.

[11] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, pp. 1–20, 2024.

[12] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, Vol. 37, pp. 56322–56348. Curran Associates, Inc., 2024.

[13] N. Benjamin Erichson, Peng Zheng, Krithika Manohar, Steven L. Brunton, J. Nathan Kutz, and Aleksandr Y. Aravkin. Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics*, Vol. 80, No. 2, pp. 977–1002, 2020.

[14] Anton Xue, Rajeev Alur, and Eric Wong. Stability guarantees for feature attributions with multiplicative smoothing. In *Advances in Neural Information Processing Systems*, Vol. 36, pp. 62388–62413. Curran Associates, Inc., 2023.