VLM における Attribution OoD の座標軸設計

d-hacks B3 omote 親:ryokawa さん

概 要

Vision-Language Model（VLM）に対する従来の分布外検知は，単一スカラーによる ID/OoD 二値判定にとどまり，運用側が本当に必要とする「なぜ誤ったのか（どの種類の崩れか）」という情報が欠落している．

本研究では，VLM の推論ログから得られる差分特徴に基づき，不確実性の増大軸 z_u と確信度低下軸 z_c を含む Attribution 型 OoD 座標系を，再現性・説明性・運用接続性の観点から固定する手続きを考える．

具体的には，Food101 を ID，CIFAR-100 および ImageNet-R を OoD とする合計 20,000 サンプルに対して，OpenCLIP ViT-B/32 の出力から confidence，エントロピー，energy，複合 OoD スコア等の drop/gain 特徴を構成し，分散閾値処理の後に SparsePCA を適用する．

軸数 k と正則化係数 α は CV 再構成誤差に基づき 1SE ルールで選択し，判定器側の情報を用いない形で最小複雑なモデル ($k = 4, \alpha = 8.0$) を決定する．

その結果として得られた軸群から，不確実性寄与が最大の成分を z_u，確信度低下寄与が最大の成分を z_c として自動的に割り当て，さらに乱数 seed を変えた再実行でも軸方向のコサイン類似度がほぼ 1.0 となる高い安定性を確認した．

本稿の成果は，後段の閾値最適化や判定器設計と独立に利用可能な固定軸定義とその安定性評価を与え,「理由付き OoD 検知」を実現するための前提条件として，VLM 時代の Attribution OoD における軸設計の恣意性・不安定性・手作業依存の問題を軽減する点にある．

1 背景

深層ニューラルネットワークを用いた画像認識は，多数のクラスに対して高精度な分類を実現しており，実運用システムにも広く組み込まれている．しかし，学習時に観測していない分布外（Out-of-Distribution, OoD）入力に対しては，モデルが高い信頼度で誤ったクラスを出力してしまうことが知られており，実環境での安全な適用において重要な課題となっている．この問題に対処するため，最大ソフトマックス確率（Maximum Softmax Probability, MSP）やエネルギーベース手法など，出力スコアに基づいて ID/OoD を二値判定する各種 OoD 検知手法が提案されてきた．これらの手法では，入力ごとに 1 次元の OoD スコアを算出し，しきい値処理によって未知入力を検知する枠組みが一般的である．

一方で，近年はテキストと画像を共通の埋め込み空間に写像する Vision-Language Model（VLM）が登場し，ゼロショット画像分類やプロンプトベースの推論が可能になっている．

VLM は，クラスラベルだけでなくテキストプロンプトの設計やテンプレートに応じて挙動が変化するため，従来の純粋な画像分類モデルとは異なる形で不確実性や分布外挙動が現れることが想定される．このような VLM を対象とした OoD 検知では，単一のスカラー値に還元された OoD スコアだけでなく，どのような観点から分布外とみなされているのかをより詳細に把握したいというニーズが高まっている．

こうした背景のもと，複数のスコアや特徴量を組み合わせて，モデルの失敗モードを「意味的なずれ」「外観の変化」「信頼度の低下」などの解釈可能な軸に分解しようとする Attribution 型 OoD の考え方が注目されつつある．従来は主に ResNet をバックボーンとする画像分類器に対して，MSP や Mahalanobis 距離，エネルギーなど複数の証拠スコアから，あらかじめ定義した線形結合により OoD の「軸」を構成し，入力ごとの OoD 挙動を多次元的に記述する試みが行われてきた．本研究もこの流れに位置づけられ，VLM の推論ログから得られる差分特徴に基づいて，多軸的な Attribution OoD 座標系の構成を目指すものである．

2 問題点

従来の OoD 検知手法の多くは，最大ソフトマックス確率（Maximum Softmax Probability, MSP）や energy などから入力ごとに 1 次元の OoD スコアを算出し，しきい値処理によって ID/OoD を判定する枠組みにとどまっている．この枠組みでは，スコアの大小から「分布外らしさ」の強弱を評価することはできるものの，なぜその入力が分布外と判断されたのか，

どのような種類の崩れ（意味的なミスマッチ，外観の劣化，モデル知識の不足 etc...）が支配的であるのかといった内訳が失われてしまう．実運用においては，検知結果に応じて対処方針（撮影条件の見直し，ラベル設計の変更，モデル更新の要否など）を決める必要があり，単一スカラーによる ID/OoD 判定だけでは意思決定に必要な情報が不足している．

本研究で扱う Attribution 型 OoD では，まず，モデルの推論ログから得られる複数の証拠スコア（例えば MSP, energy, Mahalanobis 距離，logit マージン，既存 OoD スコアの増減量など）を並べたベクトル

$$\mathbf{s}(x) = \big(s_1(x), s_2(x), \ldots, s_M(x)\big)^\top \tag{1}$$

を考える．ここでいう「軸」とは，この証拠ベクトルに対する 1 次元の射影

$$z^{(j)}(x) = \sum_{m=1}^{M} w_m^{(j)} s_m(x) = \mathbf{w}^{(j)\top}\mathbf{s}(x) \tag{2}$$

として定義されるスカラー量 $z^{(j)}(x)$ のことであり，重みベクトル $\mathbf{w}^{(j)}$ に応じて，ある特定の観点から見た「OoD の度合い」を表す座標成分に相当する．

例えば，意味的なミスマッチを強く反映する重み付けを用いれば $z^{(j)}(x)$ は「semantics 軸」として解釈され，外観の変化やノイズの影響を強く反映する重み付けを用いれば「covariates 軸」として解釈される．複数の軸 $\{z^{(1)}, \ldots, z^{(K)}\}$ を組み合わせることで，1 つの入力 x に対して「どの方向にどれだけ分布外であるか」を K 次元の座標として表現することができる．

ResNet をバックボーンとした従来の Attribution 型 OoD では，MSP，Mahalanobis 距離，energy などのスコアから，人手で定めた線形結合により「semantic」「appearance」などと名付けた軸 $z^{(j)}(x)$ を構成し，これを OoD 証拠空間の座標軸として用いる設計が多い．

このとき，どのスコアをどの比率で混合するか（すなわち $\mathbf{w}^{(j)}$ をどう選ぶか），何本の軸 K を採用するかといった判断は，研究者ごとの経験則や試行錯誤に依存しやすく，軸の本数や重みが恣意的になりやすい．また，サンプル数やデータ分割，乱数 seed が変わると最適な線形結合が変動し得るにもかかわらず，こうした「軸」の安定性や再現性を体系的に評価した報告は限られている．

さらに，これらの人手設計の「固定軸」は主に ResNet 系の純粋な画像分類器を前提として構成されており，テキストプロンプトと画像の両方に依存して出力が変化する Vision-Language Model（VLM）に対して，そのまま適用してよいかどうかは明らかではない．

VLM の場合，プロンプトテンプレートやクラス記述文の違いが信頼度スコアや energy の挙動に直接影響するため，ResNet 前提で設計された $\mathbf{w}^{(j)}$ が，VLM においても同じ意味で「semantic」や「appearance」の変化を表しているとは限らない．

したがって，VLM の推論ログに基づく Attribution 型 OoD を考える上では，証拠ベクトル $\mathbf{s}(x)$ に対する軸の本数 K や重みベクトル $\mathbf{w}^{(j)}$ をデータ駆動的かつ再現性のある手続きで決定し，ResNet 時代の人手設計の固定軸から脱却することが求められている．

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

本稿で解こうとする問題は,「どのような OoD 判定器を作るか」ではなく,

VLM の推論ログから得られる複数の証拠スコアに対して，どのように多軸の Attribution OoD 座標系を構成し，その軸を再現性のある形で固定するか?

という軸設計の問題である．

まず，VLM による推論から得られる各入力 i について，信頼度や不確実性に関する差分特徴ベクトル

$$\mathbf{x}_i = \big(d\text{conf\_drop}_i,\, d\text{msp\_drop}_i,\, d\text{entropy\_gain}_i,\, d\text{energy\_gain}_i,\, d\text{oodscore\_gain}_i\big)^\top \in R^M, \quad M = 5. \tag{3}$$

を構成する．これらを縦に並べた行列

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in R^{N \times M} \tag{4}$$

が，本稿で扱う「証拠空間」の生データである．ここでいう「軸」とは，この M 次元空間における一次元方向，すなわち重みベクトル $\mathbf{w}^{(j)} \in R^M$ による射影

$$z^{(j)}(x_i) = \mathbf{w}^{(j)\top}\mathbf{x}_i \quad (j = 1, \ldots, k) \tag{5}$$

として定義されるスカラー量であり，$\{z^{(1)}(x_i), \ldots, z^{(k)}(x_i)\}$ が入力 i の k 次元 Attribution OoD 座標となる．

本稿の中心的なタスクは，行列 X に対して

- 少数の軸数 k（本稿では $k \in \{1, 2, 3, 4\}$）を選び，
- 各軸の重みベクトル $\mathbf{w}^{(j)}$ と，対応する潜在座標 $\mathbf{z}_i = (z^{(1)}(x_i), \ldots, z^{(k)}(x_i))^\top$

を求めることである．

これを，本稿では SparsePCA による行列表現

$$X \approx ZW^\top, \quad Z \in R^{N \times k}, W = \big[\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\big] \in R^{M \times k} \tag{6}$$

として定式化し，再構成誤差と重みのスパース性のトレードオフを制御する正則化係数 α（本研究では $\alpha \in \{0.5, 1.0, 2.0, 4.0, 8.0\}$）を含めて，

$$\min_{Z,W} \frac{1}{2} \|X - ZW^\top\|_F^2 + \alpha \|W\|_1 \tag{7}$$

に類する目的関数を最小化する問題として解く．

したがって，本稿が具体的に解いているのは，次の 3 点からなる軸設計問題である．

1. 差分特徴行列 X に対して，SparsePCA により少数軸 k と重み行列 W を推定し，再構成誤差とスパース性の観点から「最小限かつ十分な」軸数 k と正則化係数 α を選ぶこと．
2. 得られた各軸について，$d$entropy_gain や $d$energy_gain といった不確実性寄与が卓越する軸を「不確実性軸」 z_u，$d$conf_drop や $d$msp_drop など確信度低下に関する寄与が卓越する軸を「確信度軸」 z_c として同定し，人間が解釈可能なラベルを与えること．
3. 乱数 seed やクロスバリデーション分割を変えても，軸方向 $\mathbf{w}^{(j)}$ がコサイン類似度の意味でほぼ一致するかどうかを検証し，後段の OoD 判定器設計とは独立に利用できる「固定軸」としての再現性を確認すること．

このように，本稿は OoD 検知そのものの性能最適化ではなく，VLM の推論ログから構成される差分特徴行列 X に対して，少数の Attribution OoD 軸 $\{z^{(1)}, \ldots, z^{(k)}\}$ をデータ駆動的かつ安定に抽出する問題を解くことを目的としている．

5 実行フロー

本節では，本研究における軸作成処理の実行フローを示す．入力は VLM の推論ログから構成した差分特徴であり，出力は再現可能に固定された軸群 $\{z_j\}_{j=1}^k$ と，運用上用いる 2 軸 (z_u, z_c) である．処理手順を以下の 8 段階に整理する．

1. **データ取得**：TensorFlow Datasets を用いて，ID データセットとして Food101，OoD データセットとして CIFAR-100 および ImageNet-R を所定件数だけ読み込む．これらを結合し，ID/OoD 合計 N サンプルからなる評価用集合を構成する．
2. **VLM 推論**：OpenCLIP ViT-B/32 に対して，あらかじめ決めたプロンプトテンプレートとクラス記述文を与え，各サンプルのクラス確率分布 $\mathbf{p}_i$ を計算する．同時に，最大確率 conf_i，正規化エントロピー $\tilde{H}_i$，energy，複合 OoD スコア s_i^{ood} などのスカラー指標を推論ログとして保存する．
3. **差分特徴構成**：ID サンプル集合に対して，各指標の ID 平均を計算し，それを基準とした drop/gain 特徴を定義する．具体的には，確信度低下，不確実性増大，energy の増加，複合 OoD スコアの増加といった量をまとめて，各サンプル i ごとに M 次元の差分特徴ベクトル $\mathbf{x}_i \in R^M$ を構成する．これらを縦に並べて差分特徴行列 $X \in R^{N \times M}$ を得る．

4. **前処理**：各特徴次元 j について標本分散 $\mathrm{Var}(x_j)$ を計算し，所定の閾値 τ_{var} 未満となるほぼ定数な特徴は軸学習から除外する．削除された特徴名は実験ログに記録し，どの指標が寄与していないかを後から追跡可能にする．必要に応じて，各特徴を平均 0・分散 1 に標準化し，SparsePCA の入力となる行列 X を準備する．

5. **モデル選択用学習**：候補とする軸数 k と正則化係数 α のグリッドをあらかじめ定める．各 (k, α) について F 分割クロスバリデーションを行い，検証用サンプル集合 V_f に対する再構成誤差

$$\mathrm{MSE}_f(k, \alpha)$$

を計算する．その平均 $\mathrm{MSE}(k, \alpha)$ と標準誤差 $\mathrm{SE}(k, \alpha)$ を集計し，候補ごとの性能表を得る．

6. **1SE ルールによる選択**：検証誤差が最小となる組 (k^*, α^*) を求め，閾値

$$\tau_{\mathrm{1SE}} = \mathrm{MSE}(k^*, \alpha^*) + \mathrm{SE}(k^*, \alpha^*)$$

を定義する．その上で，$\mathrm{MSE}(k, \alpha) \le \tau_{\mathrm{1SE}}$ を満たす候補の中から，軸数 k が最も小さい設定を「最小複雑モデル」として採用し，軸の本数と正則化強度を決定する．

7. **最終軸の固定**：選択された (k, α) を用いて，全サンプル N を使った SparsePCA を再学習する．得られた成分行列 $W = [\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}]$ と，サンプル側の潜在スコア行列 Z を保存し，以降の実験ではこれらを変更せず「固定軸」として用いる．

8. **運用軸への割当**：各軸 j に対して，確信度低下系の特徴と不確実性増大量の特徴に関する寄与を集約し，不確実性寄与が最大となる軸を「不確実性軸」 z_u，確信度低下寄与が最大となる軸を「確信度軸」 z_c として固定する．以後，後段の OoD 判定器設計や可視化では，この (z_u, z_c) を基本となる 2 軸として利用する．

以上のフローにより，後段の判定器設計とは独立に，VLM の推論ログに対する多軸 Attribution OoD 座標系を再現可能な形で構成できる．

段階	処理名	概要
1	データ取得	Food101 を ID，CIFAR-100 と ImageNet-R を OoD として読み込み，ID/OoD を結合した評価用集合を構成する．
2	VLM 推論	OpenCLIP ViT-B/32 に対してプロンプトテンプレートとクラス記述文を与え，クラス確率分布や confidence，entropy，energy，OoD スコアなどの指標を推論ログとして取得する．
3	差分特徴構成	ID 集合の指標平均を基準として，確信度低下，不確実性増大，energy 増加，OoD スコア増加などの drop/gain 特徴を計算し，各サンプルごとに差分特徴ベクトル $\mathbf{x}_i$ を構成して行列 X を得る．
4	前処理	各特徴次元の分散を評価し，所定の閾値未満のほぼ定数な特徴を削除する．必要に応じて標準化を行い，SparsePCA の入力行列として整形する．
5	モデル選択用学習	軸数 k と正則化係数 α のグリッドについて F 分割クロスバリデーションを行い，再構成誤差の平均と標準誤差を集計する．
6	1SE ルールによる選択	検証誤差が最小の組 (k^*, α^*) を基準に 1SE ルールを適用し，許容誤差内で最も軸数の小さい組を「最小複雑モデル」として採用する．
7	最終軸の固定	選択された (k, α) で全サンプルを用いて SparsePCA を再学習し，成分行列 W と潜在スコア行列 Z を保存して，後段ではこれらを変更せず固定軸として用いる．
8	運用軸への割当	各軸の重みを解析し，不確実性指標への寄与が最大の軸を不確実性軸 z_u，確信度低下指標への寄与が最大の軸を確信度軸 z_c として割り当て，運用上の基本 2 軸とする．

6 数理的な定義

ここでは，それぞれの言葉・指標をどう定義したかを数学的に記した．その後，なぜこのようにしたかを述べる．

差分特徴空間の定義

ID データ集合と OoD データ集合をそれぞれ

$$\mathcal{D}^{\mathrm{ID}} = \{x_i^{\mathrm{ID}}\}_{i=1}^{N_{\mathrm{ID}}}, \quad \mathcal{D}^{\mathrm{OOD}} = \{x_i^{\mathrm{OOD}}\}_{i=1}^{N_{\mathrm{OOD}}} \tag{8}$$

とし，$N = N_{\mathrm{ID}} + N_{\mathrm{OOD}}$ とおく．以下では，両者をまとめた列

$$\{x_i\}_{i=1}^{N} = \mathcal{D}^{\mathrm{ID}} \cup \mathcal{D}^{\mathrm{OOD}} \tag{9}$$

に対して定義を行う．

Vision-Language Model（VLM）の出力を

$$\mathbf{p}_\theta(x) = \big(p_\theta(1 \mid x), \ldots, p_\theta(C \mid x)\big)^\top \tag{10}$$

とし，そこから得られるいくつかのスカラー指標を

$$\mathrm{conf}(x) = \max_c p_\theta(c \mid x), \tag{11}$$

$$\mathrm{msp}(x) = \mathrm{conf}(x), \tag{12}$$

$$\mathrm{entropy}(x) = -\sum_{c=1}^{C} p_\theta(c \mid x)\, \log p_\theta(c \mid x), \tag{13}$$

$$\mathrm{energy}(x) = -\log \sum_{c=1}^{C} \exp f_\theta^{(c)}(x), \tag{14}$$

$$\mathrm{oodscore}(x) = g\big(\mathbf{p}_\theta(x), f_\theta(x)\big) \tag{15}$$

と定義する．ここで，$f_\theta^{(c)}(x)$ はクラス c に対応する logit であり，$g(\cdot)$ は既存 OoD スコア（energy など）から構成したスカラー関数である．

ID 集合における各指標の基準値を

$$\mu_{\mathrm{conf}} = \frac{1}{N_{\mathrm{ID}}} \sum_{x \in \mathcal{D}^{\mathrm{ID}}} \mathrm{conf}(x) \quad \text{など} \tag{16}$$

と定義し，同様に μ_{msp}，μ_{entropy}，μ_{energy}，μ_{oodscore} を定める．

各サンプル x_i に対して，ID 基準からの drop/gain 特徴を

$$d\mathrm{conf\_drop}_i = \mu_{\mathrm{conf}} - \mathrm{conf}(x_i), \tag{17}$$

$$d\mathrm{msp\_drop}_i = \mu_{\mathrm{msp}} - \mathrm{msp}(x_i), \tag{18}$$

$$d\mathrm{entropy\_gain}_i = \mathrm{entropy}(x_i) - \mu_{\mathrm{entropy}}, \tag{19}$$

$$d\mathrm{energy\_gain}_i = \mathrm{energy}(x_i) - \mu_{\mathrm{energy}}, \tag{20}$$

$$d\mathrm{oodscore\_gain}_i = \mathrm{oodscore}(x_i) - \mu_{\mathrm{oodscore}} \tag{21}$$

と定義する．これらをまとめて

$$\begin{aligned}\mathbf{x}_i = \big(&d\mathrm{conf\_drop}_i,\, d\mathrm{msp\_drop}_i,\, d\mathrm{entropy\_gain}_i,\\ &d\mathrm{energy\_gain}_i,\, d\mathrm{oodscore\_gain}_i\big)^\top \in R^M, \quad M = 5\end{aligned} \tag{22}$$

とおき，これを縦に並べた行列

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in R^{N \times M} \tag{23}$$

を，本稿で扱う差分特徴行列とする．

SparsePCA による軸抽出

本稿でいう「軸」とは，差分特徴空間における一次元方向であり，重みベクトル $\mathbf{w}^{(j)} \in R^M$ による射影

$$z^{(j)}(x_i) = \mathbf{w}^{(j)\top}\mathbf{x}_i \quad (j = 1, \ldots, k) \tag{24}$$

として定義されるスカラー量である．k 本の軸を列に並べた行列

$$W = \left[\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\right] \in R^{M\times k} \tag{25}$$

と，サンプル側の潜在表現

$$Z = \begin{bmatrix} \mathbf{z}_1^\top \\ \mathbf{z}_2^\top \\ \vdots \\ \mathbf{z}_N^\top \end{bmatrix} \in R^{N\times k}, \quad \mathbf{z}_i = \left(z^{(1)}(x_i), \ldots, z^{(k)}(x_i)\right)^\top \tag{26}$$

を用いれば，差分特徴行列 X に対する近似

$$X \approx ZW^\top \tag{27}$$

を考えることができる．

本稿では，W をスパースに保ちながら X を再構成する SparsePCA 型の最適化問題

$$\min_{Z,W} \frac{1}{2}\left\|X - ZW^\top\right\|_F^2 + \alpha\left\|W\right\|_1 \tag{28}$$

を解くことで，軸数 k 本の重みベクトル $\mathbf{w}^{(j)}$ と潜在スコア $\mathbf{z}_i$ を同時に推定する．ここで $\|\cdot\|_F$ は Frobenius ノルム，$\|\cdot\|_1$ は要素ごとの L_1 ノルムである．

モデル選択と 1SE ルール

軸数 k と正則化係数 α の候補集合を

$$\mathcal{K} = \{1, 2, 3, 4\}, \quad \mathcal{A} = \{0.5, 1.0, 2.0, 4.0, 8.0\} \tag{29}$$

とし，各組 $(k, \alpha) \in \mathcal{K} \times \mathcal{A}$ について F 分割クロスバリデーションを行う．第 f 分割における検証集合 V_f に対する再構成誤差を

$$\mathrm{MSE}_f(k, \alpha) = \frac{1}{|V_f|}\sum_{i\in V_f}\left\|\mathbf{x}_i - \hat{\mathbf{z}}_i^{(f)}(k, \alpha)\,\hat{W}^{(f)}(k, \alpha)^\top\right\|_2^2 \tag{30}$$

と定義し，その平均と標準誤差を

$$\mathrm{MSE}(k, \alpha) = \frac{1}{F}\sum_{f=1}^{F}\mathrm{MSE}_f(k, \alpha), \tag{31}$$

$$\mathrm{SE}(k, \alpha) = \sqrt{\frac{1}{F(F-1)}\sum_{f=1}^{F}\left(\mathrm{MSE}_f(k, \alpha) - \mathrm{MSE}(k, \alpha)\right)^2} \tag{32}$$

として求める．

最良モデル (k^*, α^*) を

$$(k^*, \alpha^*) = \arg\min_{(k,\alpha)} \mathrm{MSE}(k, \alpha) \tag{33}$$

とし，その 1SE 閾値

$$\tau_{\mathrm{1SE}} = \mathrm{MSE}(k^*, \alpha^*) + \mathrm{SE}(k^*, \alpha^*) \tag{34}$$

を定義する．このとき，

$$\mathcal{S}_{\mathrm{1SE}} = \left\{ (k, \alpha) \;\middle|\; \mathrm{MSE}(k, \alpha) \le \tau_{\mathrm{1SE}} \right\} \tag{35}$$

を満たす候補集合のうち，軸数 k が最小の組

$$(\hat{k}, \hat{\alpha}) = \arg \min_{(k,\alpha) \in \mathcal{S}_{\mathrm{1SE}}} k \tag{36}$$

を「最小複雑モデル」として採用し，最終的な軸数と正則化係数を決定する．

軸ラベル付けと安定性指標

最終学習では，選択された $(\hat{k}, \hat{\alpha})$ のもとで，全サンプル N を用いて式 (28) を解き，成分行列

$$\hat{W} = \left[\hat{\mathbf{w}}^{(1)}, \ldots, \hat{\mathbf{w}}^{(\hat{k})} \right] \tag{37}$$

を得る．

指標ごとのグループを

$$\mathcal{G}_{\mathrm{conf}} = \{\mathrm{conf\_drop}, \mathrm{msp\_drop}\}, \tag{38}$$

$$\mathcal{G}_{\mathrm{uncert}} = \{\mathrm{entropy\_gain}, \mathrm{energy\_gain}, \mathrm{oodscore\_gain}\} \tag{39}$$

とし，軸 j に対する寄与量を

$$C_{\mathrm{conf}}(j) = \sum_{m \in \mathcal{G}_{\mathrm{conf}}} \left| \hat{w}_m^{(j)} \right|, \tag{40}$$

$$C_{\mathrm{uncert}}(j) = \sum_{m \in \mathcal{G}_{\mathrm{uncert}}} \left| \hat{w}_m^{(j)} \right| \tag{41}$$

と定義する．このとき，不確実性寄与が最大の軸と確信度低下寄与が最大の軸を

$$j_u = \arg \max_{1 \le j \le \hat{k}} C_{\mathrm{uncert}}(j), \tag{42}$$

$$j_c = \arg \max_{1 \le j \le \hat{k}} C_{\mathrm{conf}}(j) \tag{43}$$

として求め，それぞれ

$$z_u(x_i) = \hat{\mathbf{w}}^{(j_u)\top} \mathbf{x}_i, \quad z_c(x_i) = \hat{\mathbf{w}}^{(j_c)\top} \mathbf{x}_i \tag{44}$$

を「不確実性軸」「確信度軸」として定義する．

さらに，乱数 seed を変えて R 回の学習を行い，各回 r で得られた軸行列を

$$\hat{W}^{(r)} = \left[\hat{\mathbf{w}}^{(1,r)}, \ldots, \hat{\mathbf{w}}^{(\hat{k},r)} \right] \tag{45}$$

とおく．符号反転の不定性を補正したうえで，対応する軸どうしのコサイン類似度

$$\mathrm{cos\_sim}_{r,r'}^{(j)} = \frac{\hat{\mathbf{w}}^{(j,r)\top} \hat{\mathbf{w}}^{(j,r')}}{\left\| \hat{\mathbf{w}}^{(j,r)} \right\|_2 \left\| \hat{\mathbf{w}}^{(j,r')} \right\|_2} \tag{46}$$

を計算し，その平均

$$\overline{\mathrm{cos\_sim}}^{(j)} = \frac{2}{R(R-1)} \sum_{1 \le r < r' \le R} \mathrm{cos\_sim}_{r,r'}^{(j)} \tag{47}$$

を軸 j の安定性指標とする．

ここで導入した数式は，VLM の推論ログから得られる複数の指標を，(1) ID 基準からの差分として正規化し，(2) 線形かつスパースな低次元表現に写像し，(3) モデル選択とラベル付け，安定性検証までを一貫して扱うための枠組みである．以下では，各ブロックが何を意図しているかを整理する．

まず，差分特徴ベクトル $\mathbf{x}_i$ の定義式は，ID 集合における各指標の平均値を基準として，「どの程度 ID から崩れているか」を方向つきの量として表現するために用いたものである．$\mathrm{conf}(x)$ や $\mathrm{msp}(x)$ に対しては ID 平均からの減少量 $d\mathrm{conf\_drop}_i, d\mathrm{msp\_drop}_i$ を，$\mathrm{entropy}(x)$ や $\mathrm{energy}(x)$，$\mathrm{oodscore}(x)$ に対しては ID 平均からの増加量 $d\mathrm{entropy\_gain}_i, d\mathrm{energy\_gain}_i, d\mathrm{oodscore\_gain}_i$ を定義することで,「確信度の低下」と「不確実性・エネルギーの増大」が正のスケールで比較可能になるよう揃えている．このような差分表現を採用することで，絶対値そのものよりも「ID からのズレ」に着目した軸学習が可能となり，ID/OoD やプロンプト条件の違いに対して一貫した解釈を与えやすくしている．

差分特徴行列 $X \in R^{N\times M}$ は，全サンプルの崩れ方を一括して表現するためのデータ行列であり，以降の SparsePCA による軸抽出の入力となる．ここで各特徴次元の分散を閾値でフィルタする前処理は，ほぼ定数で情報を持たない指標をあらかじめ除去し，ノイズ的な次元がスパース性推定を乱すことを防ぐ意図で導入している．

次に，射影

$$z^{(j)}(x_i) = \mathbf{w}^{(j)\top}\mathbf{x}_i$$

によって定義されるスカラー量 $z^{(j)}$ は，差分特徴空間における「1 本の軸」であり，各軸 $\mathbf{w}^{(j)}$ は「どの指標の組合せを，どの割合で見ると一貫した崩れ方として説明できるか」を表す重みベクトルである．すなわち，k 本の軸をまとめた行列 W と潜在表現 Z による近似 $X \approx ZW^\top$ は,「M 次元の差分特徴を k 次元の潜在座標でどこまで説明できるか」を明示的に書き下したものである．

式 (28) は，この近似を実現するための目的関数であり，Frobenius ノルムの二乗を用いた再構成誤差項 $\|X - ZW^\top\|_F^2$ に，重み行列 W に対する L_1 正則化項 $\alpha\|W\|_1$ を付加した形になっている．通常の PCA では L_2 ノルムに基づく低ランク近似を行うが，ここでは各軸が少数の指標にのみ依存する「解釈しやすい軸（ほとんどの成分が 0 の軸）」を得ることを優先して，L_1 正則化によりスパース性を強制している．このように定式化することで，従来 ResNet ベースの OoD 設計で人手により設定していた「semantic」「appearance」といった線形結合を，データ駆動で自動的に学習しつつ，各軸がどの指標から構成されているかが一目で分かる表現が得られる．

モデル選択のためのクロスバリデーションと $\mathrm{MSE}(k,\alpha), \mathrm{SE}(k,\alpha)$ の定義は，軸数 k と正則化係数 α の組ごとに,「どれだけ差分特徴行列 X をよく説明できているか」を再構成誤差という形で評価するために導入したものである．ここであえて ID/OoD ラベルに依存した指標（AUROC など）ではなく，ラベルフリーな再構成誤差を用いているのは，軸設計の段階を後段の OoD 判定器（しきい値最適化や分類器）から切り離し,「失敗モードをよく要約する座標系」を純粋に幾何学的な観点から決めたいという意図による．

1SE ルールに基づく $(\hat{k}, \hat{\alpha})$ の選択は，過学習を避けつつ軸数を最小限に抑えるための基準である．最良モデル (k^*, α^*) の検証誤差とその標準誤差から閾値 τ_{1SE} を定め，その範囲内で最も軸数 k の小さい組を採用することで,「ほぼ同等の再構成性能であれば，より少ない軸で説明する」方針を明文化している．この基準により，研究者が恣意的に軸数を決めるのではなく，データに基づいて「最小複雑モデル」を一意に選ぶことができる．

軸ラベル付けのための寄与量 $C_{\mathrm{conf}}(j)$ および $C_{\mathrm{uncert}}(j)$ は，各軸 $\mathbf{w}^{(j)}$ が「確信度低下系の指標」と「不確実性・エネルギー増大量の指標」のどちらを主に見ているかを定量化するために定義したものである．これにより，人手で軸を回転させたり，係数を微調整したりすることなく,「不確実性寄与が最大の軸を z_u，確信度低下寄与が最大の軸を z_c と呼ぶ」というルールだけで運用上の 2 軸を決定できる．すなわち，semantic / appearance などのラベルは重み行列から自動的に読み取られ，定義が研究者ごとに大きく変わってしまうリスクを抑えられる．

最後に，コサイン類似度に基づく安定性指標 $\overline{\mathrm{cos\_sim}}^{(j)}$ は，乱数 seed や分割を変えた再学習に対しても各軸 $\mathbf{w}^{(j)}$ がどの程度再現されているかを評価するために導入した．スパースな線形モデルでは，わずかなデータの違いで解が変動しやすいことが知られているため,「たまたま今回の実行で得られただけの軸」ではなく，再実行してもほぼ同じ方向が現れる「固定軸」になっているかどうかを確認する必要がある．こ

こでコサイン類似度を用いているのは，線形モデルにおける符号反転の不定性（$\mathbf{w}$ と $-\mathbf{w}$ が同じ軸を表す）を自然に吸収しつつ，軸方向の一致度を測ることができるためである．

以上のように，本節で導入した数式群は，(1) ID 基準からの差分として OoD 挙動を記述し，(2) SparsePCA によって少数のスパースな軸に要約し，(3) クロスバリデーションと 1SE ルールで軸数と正則化を決定し，(4) 寄与量とコサイン類似度により，人間可読なラベル付けと安定性検証を行う，という一連の設計方針を形式化したものである．この枠組みにより，後段の OoD 判定器とは独立に，VLM の推論ログに対する多軸 Attribution OoD 座標系を再現可能な形で固定することが可能となる．

7　実装結果

本節では，第**??**節で定義した差分特徴行列 X と SparsePCA に基づく軸作成手続きを，Food101 ／ CIFAR-100 ／ ImageNet-R を対象として実装した結果を報告する．まず ID 基準となる統計量を確認し，ついで 1SE ルールにより選択された軸数 k と正則化係数 α を示す．その後，得られた運用軸 (z_u, z_c) の具体的な形と，乱数 seed を変えた際の軸安定性，最後に図による可視化結果について述べる．

7.1　基礎統計

ID 集合 $\mathcal{D}^{\mathrm{ID}}$ に対して，式で定義した各指標の平均値を計算した結果，

$$\begin{aligned} \bar{\mathrm{conf}}_{\mathrm{ID}} &= 0.8259888770, \\ \bar{\tilde{H}}_{\mathrm{ID}} &= 0.1187297247, \\ \bar{s}^{\mathrm{ood}}_{\mathrm{ID}} &= 0.1463704238 \end{aligned} \tag{48}$$

であった．ここで $\bar{\mathrm{conf}}_{\mathrm{ID}}$ は ID データにおける平均確信度，$\bar{\tilde{H}}_{\mathrm{ID}}$ は正規化エントロピーの平均，$\bar{s}^{\mathrm{ood}}_{\mathrm{ID}}$ は既存 OoD スコアの平均を表す．

この値は，ID データに対して VLM が比較的高い確信度（約 0.83）を維持しつつ，エントロピーと OoD スコアが低めに抑えられていることを意味する．差分特徴 $\mathbf{x}_i$ はこれらの ID 平均からの drop/gain として定義されているため，以降の軸学習では，「ID に比べてどれだけ確信度が落ちたか」「どれだけ不確実性／エネルギーが増えたか」が直接の入力となる．このように ID 基準を明示しておくことで，異なるデータセットやプロンプト設定に対しても，差分特徴の解釈を一定に保つことができる．

7.2　1SE ルールによる軸数・正則化の選択

軸数 $k \in \{1,2,3,4\}$，正則化係数 $\alpha \in \{0.5, 1.0, 2.0, 4.0, 8.0\}$ のグリッドに対して F 分割クロスバリデーションを行い，式で定義した再構成誤差 $\mathrm{MSE}(k,\alpha)$ と標準誤差 $\mathrm{SE}(k,\alpha)$ を評価した．その結果，最小の検証誤差を与える組は

$$(k^*, \alpha^*) = \arg\min_{k,\alpha} \mathrm{MSE}(k,\alpha) = (4, 8.0) \tag{49}$$

であり，

$$\begin{aligned} \mathrm{MSE}_{\min} &= 8.3656 \times 10^{-5}, \\ \mathrm{SE}_{\min} &= 4.7763 \times 10^{-7} \end{aligned} \tag{50}$$

であった．これらから 1SE 閾値

$$\begin{aligned} \tau_{\mathrm{1SE}} &= \mathrm{MSE}(k^*, \alpha^*) + \mathrm{SE}(k^*, \alpha^*) \\ &= 8.4134 \times 10^{-5} \end{aligned} \tag{51}$$

を定義し，$\mathrm{MSE}(k,\alpha) \le \tau_{\mathrm{1SE}}$ を満たす候補の中で最も軸数 k が小さい組を選んだ結果，

$$\hat{k} = 4, \quad \hat{\alpha} = 8.0 \tag{52}$$

が「最小複雑モデル」として採用された．

これは，「4 軸あれば差分特徴行列 X を十分に再構成できる」ことを意味している．また，$k = 3$ から $k = 4$ への増加で MSE はわずかに改善する一方，$k > 4$ としても 1SE 閾値の範囲内で有意な改善は得られなかったことから，4 軸以上に増やしても「説明力の割に複雑さが増えるだけ」と判断できる．このように，1SE ルールを用いることで，軸数の選択をヒューリスティックな「見た目」や単一ランの OoD 性能から切り離し，再構成誤差に基づく手続き的な基準で決定できる点に意義がある．

7.3　運用軸 (z_u, z_c) の具体形と寄与構造

選択された $(\hat{k}, \hat{\alpha}) = (4, 8.0)$ のもとで，全サンプル N を用いて SparsePCA を再学習し，4 本の軸 $\{\mathbf{w}^{(j)}\}_{j=1}^{4}$ を得た．そのうち，不確実性関連指標（entropy_gain, energy_gain, oodscore_gain）への寄与が最大の軸を z_u，確信度低下指標

(conf_drop, msp_drop) への寄与が最大の軸を z_c として割り当てた結果，次のような線形結合が得られた．

$$z_u(x) = 0.9187\, d\text{entropy\_gain}(x) + 0.3949\, d\text{oodscore\_gain}(x), \tag{53}$$

$$z_c(x) = 0.6546\, d\text{conf\_drop}(x) + 0.6546\, d\text{msp\_drop}(x) + 0.3783\, d\text{oodscore\_gain}(x). \tag{54}$$

実装上は msp = conf であるため，式 (54) は事実上

$$z_c(x) \simeq 1.3092\, d\text{conf\_drop}(x) + 0.3783\, d\text{oodscore\_gain}(x) \tag{55}$$

と等価である．この式は，z_u がほぼ「エントロピー増加＋ OoD スコア増加」を見る不確実性軸であるのに対し，z_c は主として「確信度の低下」に反応しつつ，一部に不確実性要素（$d$oodscore_gain）も混入していることを示している．

すなわち，本実験で得られた運用軸は，

- z_u：ほぼ純粋に「不確実性の増大」を測る軸，
- z_c：主に「確信度の低下」を測りつつ，OoD スコアの増加も若干取り込んだ軸

として解釈できる．特に z_c へ $d$oodscore_gain が有意に寄与している点は，現実の失敗モードが「確信度だけが落ちる」単純なパターンではなく，不確実性の増大と複合したモードとして現れる可能性を示唆している．このことは，「運用上，軸同士が完全に直交しているとは限らない」という前提を持つ方が安全であることを意味し，後段の判定器設計においても考慮すべき点である．

7.4 乱数 seed に対する軸安定性

軸の再現性を確認するため，乱数 seed を $42, 43, 44$ に変えて同一手続きを 3 回実行し，seed 42 で得られた軸を基準に，対応する運用軸どうしのコサイン類似度（絶対値）を評価した．その結果，いずれの seed においても選択された軸数は $k = 4$ であり，不確実性軸 z_u と確信度軸 z_c に対応する軸インデックスは全て $(\text{axis}_u, \text{axis}_c) = (3, 0)$ で一致した．コサイン類似度は

- seed 42：$(z_u, z_c) = (1.000000, 1.000000)$,
- seed 43：$(z_u, z_c) = (0.999998, 0.999999)$,
- seed 44：$(z_u, z_c) = (1.000000, 1.000000)$,

となり，3 seed 平均でも

$$\left(\overline{\cos\_\text{sim}}_{z_u},\ \overline{\cos\_\text{sim}}_{z_c}\right) = (0.99999948,\ 0.99999983) \tag{56}$$

であった．

この結果は，SparsePCA による軸方向 $\mathbf{w}^{(j)}$ が乱数 seed の違いにほとんど依存しないことを示している．すなわち，「今回の実行でたまたま得られた軸」ではなく，データ規模 $N = 20000$ 程度であれば，再実行してもほぼ同じ軸が復元される「固定軸」として扱えることが確認できた．後段の研究では，この固定軸を前提として，判定器側（しきい値最適化や非線形境界）の設計だけを比較すればよく，軸自体の揺らぎを気にする必要がない．

7.5 図による検証結果

以下の 4 種類の図を作成．

- 図 1：z_u–z_c 平面上で ID ／ OoD サンプルを色分けした散布図である．単一の OoD スコアではなく 2 軸で表現することにより，「主に不確実性が増大している領域」と「主に確信度が低下している領域」がどのように分布しているかを直感的に把握できる．
- 図 2：図 1 の散布図に対して見かけ密度を重ねた可視化である．ID ／ OoD の重なり領域と分離方向を明確にし，z_u–z_c 空間上でどの方向に閾値境界を引くと有効かを，視覚的に検討する材料を与える．
- 図 3：各軸に対する特徴寄与の棒グラフである．z_u が $d$entropy_gain と $d$oodscore_gain に強く依存し，z_c が $d$conf_drop と $d$msp_drop に強く依存していることを定量的に示し，上述の解釈（不確実性軸／確信度軸）を裏付ける．
- 図 4：軸数 k と正則化係数 α ごとのクロスバリデーション誤差と 1SE 閾値 τ_{1SE} をプロットした図である．本稿で用いた 1SE ルールにより，$(\hat{k}, \hat{\alpha}) = (4, 8.0)$ が「再構成誤差の観点から妥当な最小複雑モデル」であることを可視的に確認できる．

これらの実装結果と可視化により，第 6 章で定義した数理的枠組みが実データ上でも安定に機能し，VLM の推論ログから多軸 Attribution OoD 座標系 $(z_u, z_c, \dots)$ を再現可能な形で固定できることが示された．

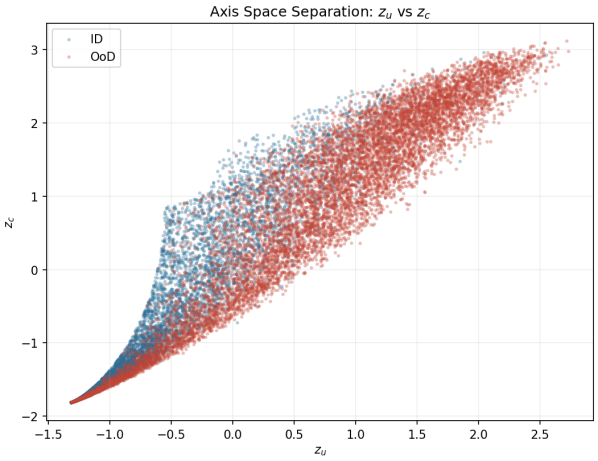

図 1: z_u–z_c 平面上での ID ／ OoD 散布図

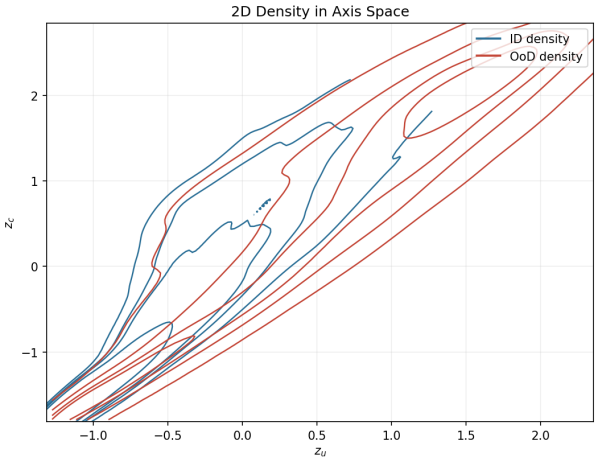

図 2: z_u–z_c 平面上での ID ／ OoD 密度可視化

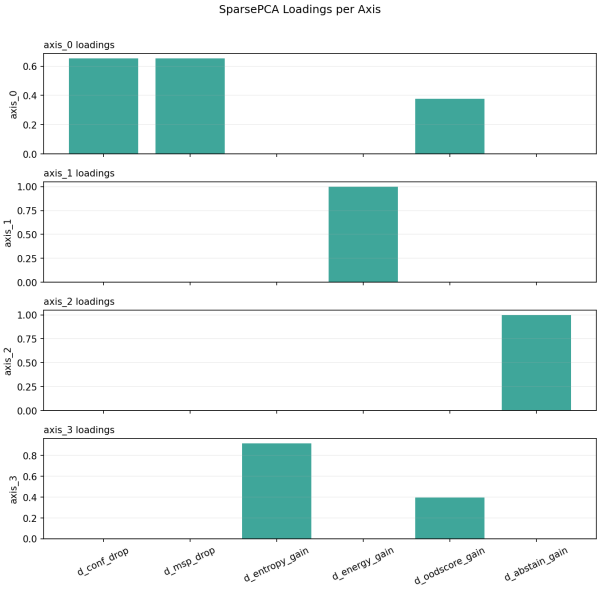

図 3: 各軸に対する差分特徴の寄与（ロードニング）

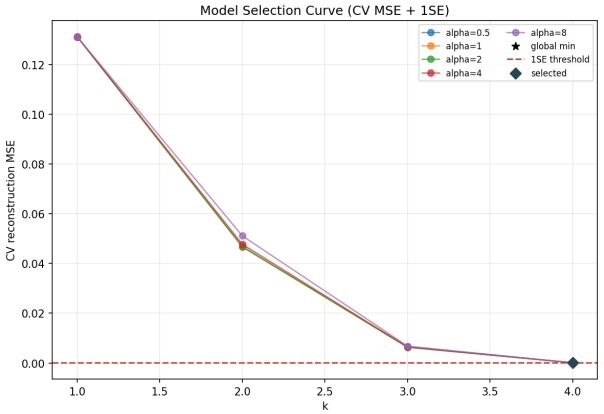

図 4: (k, α) ごとの CV 誤差と 1SE 閾値

8　結論・考察・今後の予定

本節では，本稿で提案した多軸 Attribution OoD 軸設計の枠組みについて，得られた結果を総括し，設計上の含意と限界を整理したうえで，今後の展開可能性について述べる．

8.1　結論

本稿では，従来 ResNet 前提で人手設計されていた OoD「固定軸」の問題点に対し，Vision-Language Model（VLM）の推論ログから得られる複数の指標に

基づき，データ駆動的かつ再現性のある形で OoD 軸を構成する手続き（N-Axis Attribution OoD）を定式化し，Food101 ／ CIFAR-100 ／ ImageNet-R を用いた実装により検証した．

まず，ID 集合に対する確信度・エントロピー・エネルギー・既存 OoD スコアの平均を基準として，ID からの drop/gain をそろえた差分特徴ベクトル $\mathbf{x}_i$ を定義し，全サンプルを並べた差分特徴行列 X を構成した．次に，X に対して SparsePCA による行列近似 $X \approx ZW^\top$ を行い，再構成誤差とスパース性のトレードオフを制御する目的関数 (28) を最小化することにより，線形かつスパースな軸集合 $\{\mathbf{w}^{(j)}\}_{j=1}^{k}$ を学習した．

軸数 k と正則化係数 α は，クロスバリデーション誤差 $\mathrm{MSE}(k,\alpha)$ と標準誤差 $\mathrm{SE}(k,\alpha)$ に基づき 1SE ルールで選択し，候補の中から「最小複雑モデル」として $k=4$，$\alpha=8.0$ を採用した．さらに，各軸における指標寄与量に基づき，不確実性指標（エントロピー増加，energy 増加，OoD スコア増加）への寄与が最大の軸を不確実性軸 z_u，確信度低下指標（conf_drop, msp_drop）への寄与が最大の軸を確信度軸 z_c として自動的にラベル付けした結果，z_u はほぼ純粋に不確実性の増大を，z_c は主として確信度低下を捉える，解釈しやすい線形結合となることが確認された（式 (53),(54)）．

最後に，乱数 seed を変えた複数回の学習に対して軸方向のコサイン類似度を評価したところ，z_u，z_c ともに平均類似度がほぼ 1.0 となり，軸数 $k=4$，$\alpha=8.0$ のもとでは軸方向が実質的に不変であることが分かった．これにより，本稿で構成した (z_u, z_c) を，後段の OoD 判定器設計とは独立に利用可能な「固定軸」として扱えることが示された．

以上の結果から，本稿の枠組みは，VLM 時代の Attribution OoD における「軸設計の恣意性・不安定性・手作業依存」という問題を軽減し，理由付き OoD 検知に向けた基盤となる座標系を与えるという点で一定の有効性を有すると結論づけられる．

8.2 考察

本稿の結果から，いくつかの含意と限界が得られる．

第一に，差分特徴として「ID 基準からの drop/gain」を採用したことにより，z_u と z_c がいずれも「ID からのズレの向きと大きさ」を表す軸として自然に解釈できる形になった．

特に z_u はエントロピー増加と OoD スコア増加に強く依存し，「予測分布が平坦化し，従来の OoD スコアも上昇している状況」を一軸で表す不確実性軸として機能している．一方で z_c は確信度低下を主成分としながらも，OoD スコア増加の寄与も無視できないことから，「確信度が落ちているが，単なるノイズではなく，モデル側が OoD 的な挙動も同時に示している」失敗モードを捉える軸になっていると考えられる．

これは，現実の失敗が必ずしも「純粋に semantic だけ」「純粋に appearance だけ」といった形で分離されないことを反映しており，軸間の完全な直交性を前提としない多軸設計の妥当性を裏付けている．

第二に，軸数 $k=4$ が 1SE ルールのもとで最小複雑モデルとして選ばれたことは，本稿で採用した 5 次元の差分特徴のうち，実質的に 4 次元程度の線形結合で大部分のばらつきが説明できることを示唆している．これは「指標が冗長である」というより，「確信度低下系」と「不確実性増大系」が内部的に相関を持つために，完全に独立した軸としては分解されない構造を持っていると解釈できる．今後，特徴次元を増やしていく場合でも，同様の手続きにより「どのくらいの軸数で十分か」を客観的に評価できる点は，多軸 Attribution OoD をスケールさせるうえでの利点である．

第三に，本稿の軸設計は，意図的に ID/OoD ラベルやタスク性能を目的関数に含めず，再構成誤差に基づく幾何学的な圧縮だけで軸を決めている．これは「軸設計」と「判定器設計」をステップ分離するという方針に基づくものであり，得られた軸が特定の評価指標（例：AUROC）に最適化されていない一方で，一度決まった軸を様々な判定器やタスクに再利用できる柔軟性を持つ．ただしこの方針は，同じ差分特徴空間に対して，別途「タスク最適な軸」を求めた場合とは異なる座標系が得られる可能性があることも意味する．したがって，本稿で構成した軸は，あくまで「失敗モードの要約・説明」のための基盤座標であり，OoD 判定性能の最適化とは役割が異なることを明確にしておく必要がある．

一方で，本稿の枠組みには明確な限界も存在する．第一に，差分特徴として用いた指標は 5 種類に限定されており，テキスト側の埋め込み情報やクラスプロトタイプとの距離，画像側の局所的なパッチ情報などは含まれていない．そのため，本稿の軸は「確信度・不確実性・エネルギー」というスカラー量の範囲に閉じており，semantic global ／ semantic near ／ appearance

といったより細かい失敗モードの区別には十分ではない可能性がある．第二に，本稿の実装は OpenCLIP ViT-B/32 と Food101 ／ CIFAR-100 ／ ImageNet-R に限定されており，他の VLM（例：SigLIP 系やより大型の CLIP），他ドメインのデータセットに対する汎用性は未検証である．第三に，本稿では軸の安定性評価までは行ったものの，軸に基づく OoD 判定器（しきい値設計や非線形境界）の性能比較は行っておらず，「軸を固定したうえでどのような判定器が有効か」という問いは今後の課題として残されている．

“‘tex

8.3 今後の予定

本稿は，「VLM の推論ログから差分特徴を構成し，SparsePCA によって多軸の Attribution OoD 座標系 $(z_u, z_c, \dots)$ を安定に固定できる」ことを示した段階で止まっており，軸そのものの性質（安定性・解釈可能性）は評価できているものの，**その軸を用いることで *OoD* 検知として何がどれだけ改善されるのか** という検証は行っていない．したがって，研究としての完結性（Completeness）という観点からは，本稿は「軸設計フェーズ」のみを扱った前段階の位置づけにとどまっている．今後は，ここで構成した固定軸を前提に，OoD 判定性能・説明可能性・運用有用性の 3 つの観点から，軸の「役に立ち方」を定量的・定性的に検証することが必要である．

第一の方向性は，**OoD 判定性能に関する検証**である．具体的には，本稿で得られた (z_u, z_c) を入力とする単純なしきい値判定器（例：z_u または z_c に対する 1 次元閾値，あるいは z_u–z_c 平面上の線形境界）やロジスティック回帰を構成し，MSP や energy といった従来の 1 次元スコアを用いた判定器と，ID/OoD 識別性能（AUROC，TNR@TPR=0.95 など）を比較する．同一の VLM・同一の OoD 分割（Food101 ／ CIFAR-100 ／ ImageNet-R）を用いることで，「軸を導入すること自体が OoD 性能に寄与しているのか」「単に元のスコアを別表現にしただけなのか」を切り分けることができる．さらに，ResNet 前提で人手設計された固定軸を持つ既存手法と比較することで，本稿のデータ駆動な軸設計が，性能面でも少なくとも劣化しないことを確認することが望ましい．

第二の方向性は，**説明可能性・運用有用性の検証**である．ここでの主眼は，軸を導入することで「ID/OoD の成績がどれだけ上がったか」だけでなく，**どのような失敗モードがどの軸方向として可視化されるか** を評価することにある．例えば，

- z_u–z_c 平面上の四象限ごとに代表的な誤分類例を抽出し，「主に不確実性が高い OoD」「主に確信度が不自然に高い OoD」といった失敗クラスターを事例ベースで整理する．
- 特定の OoD ドメイン（例：ImageNet-R のスタイル変化，CIFAR-100 のクラス外れ）に対して，どの軸がどれだけ寄与しているかをプロットし，1 次元スコアでは見えなかった違いが軸空間で分解できているかを確認する．

といった分析を通して，「軸を見ればどの種類の崩れが支配的か一目で分かる」というレベルまで説明性を高められるかを検証したい．これにより，「軸を作りました」で終わらず，**運用者がどのような判断に軸を使えるのか** まで含めた形で，研究としての完結性に近づけることができる．

第三の方向性は，**汎用性とロバスト性の検証**である．本稿では OpenCLIP ViT-B/32 と特定の ID/OoD 組み合わせに限定して軸設計を行ったが，今後は同一の手続きを他の OoD 分割（例：ID を CIFAR-100，OoD を SVHN ／ Tiny-ImageNet とする設定）や，プロンプトテンプレートの異なる設定に適用し，得られる軸の構造・寄与パターン・安定性がどの程度共通しているかを確認する必要がある．もし異なる分割やプロンプトでも類似の (z_u, z_c) が再現されるなら，本稿の軸設計は「特定データセットに特化した座標系」ではなく，VLM に対するより一般的な失敗モード座標として位置づけられる．逆に大きく異なる軸が得られる場合には，「どの条件のときに軸を再学習し直すべきか」という運用上の指針を与える材料となる．

以上のステップを通じて，

(1) 固定軸が OoD 判定性能に与える影響，

(2) 固定軸が失敗モードの分析・報告に与える価値，

(3) 固定軸がデータセット・プロンプト・VLM の選択に対してどの程度ロバストか，

を体系的に評価することが，本研究を「軸設計の提案」にとどめず，

「軸を設計し，それが実際に役に立つことを示した」という一つの完結した研究 とするうえで，今後の最優先課題である．

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.