3 次元点群間の差異説明のための比較アーキテクチャ

梁 誠[1,a)]　内海 ゆづ子[1,b)]　岩村 雅一[1,c)]

概要

近年，物体の 3 次元形状を大規模言語モデルを用いて説明する分野が発展している．しかし，複数の物体の違いを自然言語で表現する問題は扱われていない．そこで本研究では，物体の 3 次元形状の差異を自然言語で説明できる点群比較アーキテクチャを提案する．人体モデルである Skinned Multi-Person Linear Model (SMPL) から生成したデータセットを用いた実験の結果，差異のある部位を認識し，それを自然言語で表現できることを確認できた．

1. はじめに

近年，3D Gaussian Splatting [5] 等の 3 次元再構成技術や，仮想現実，拡張現実といった技術の普及により物体の 3 次元形状を扱う分野が進展している．これに伴い，大規模言語モデル（Large Language Model: LLM）を 3 次元データに対応させたマルチモーダル大規模言語モデル（Multimodal Large Language Model: MLLM）の研究も盛んに行われており，PointLLM [11] や ShapeLLM [9] といったモデルにより，3D オブジェクトの形状や外観を言語化することが可能となった．

一方で，物体の 3 次元形状の差異を検出することに関しては，3 次元シーンの差分検出にとどまっており，3 次元物体の差異を言語化する研究は我々が知る限り存在しない．例えば植物学においては，遺伝子と形態的特徴との対応関係を解明するため，遺伝子の一部を破壊した変異体を作成し，その成長過程や最終的な形状を遺伝子を破壊する前の植物（野生体）と比較観察が行われている．この作業を機械により客観的に行えれば，植物学の発展に大きく寄与できる．また，形状比較の結果は最終的に言語を介して議論・共有されることを踏まえると，差異を自然言語で直接記述できることには実用上の意義がある．

そこで本研究では，2 つの物体の 3 次元形状を比較し，それらの差異を自然言語で表現する点群比較アーキテクチャを提案する．3D データに対応した既存の MLLM は単一の点群入力を前提として設計されており，このアーキテクチャを維持したままでは前述の目的を達成できない．そのため，2 つの点群を入力して，その差異を抽出し，自然言語として出力するアーキテクチャを提案する．提案手法の検証には，3 次元形状の変化と言語記述が対応づけられた大量の 3D データセットが必要である．しかし，そのような既存データセットは存在しない．そこで本研究では，植物形態比較をはじめとした多様なドメインへの応用を最終目標としつつ，タスクの実現可能性を検証する第一段階として，人体モデルである Skinned Multi-Person Linear Model (SMPL) [7] を採用する．SMPL はパラメータ制御によって異なるポーズを生成可能であり，変化量を定量的に取得できる．これを用いて多様な人体ポーズのペアを生成し，Vision Language Model (VLM) を用いて自然な比較記述を付与した合成データセットを構築する．

構築したデータセットを用いて提案モデルを評価した結果，提案モデルは 2 つの点群間の差異のある部位を特定し，それを自然言語で表現することができた．

2. 関連研究

3 次元点群データの変化検出手法として，Siamese KP-Conv がある [3]．この手法は，各点群で局所的な特徴を抽出し，その特徴の差分に基づいて変化領域を特定することで，都市点群などの大規模データの変化を検出する．しかし，点群データの変化を言語で説明できない．

MLLM の推論能力を 3 次元データに適用することで 3D データを言語化する MLLM が提案されている．これらは 3 次元情報の扱い方によって，(1) 2 次元画像を介する手法と (2) 点群データを直接扱う手法の 2 つに大別できる．(1) 画像を介する手法では，3 次元オブジェクトを複数の視点からレンダリングして 2 次元画像群に変換し，既存の VLM を活用して自然言語による説明を出力する．Luo らの Cap3D [8] は，3D オブジェクトのマルチビュー画像から BLIP-2 (Bootstrapping Language-Image Pre-training) [6] や GPT-4 を用いてテキストラベルを生成し，大規模な 3 次元データセットに対してキャプションを自動付与するフレームワークを構築した．これらの手法は，2 次元画像処理で高精度なモデルを 3 次元データに使用できる利点があ

[1] 大阪公立大学 大学院情報学研究科
[a)] sz22932s@st.omu.ac.jp
[b)] yuzuko@omu.ac.jp
[c)] masa.i@omu.ac.jp

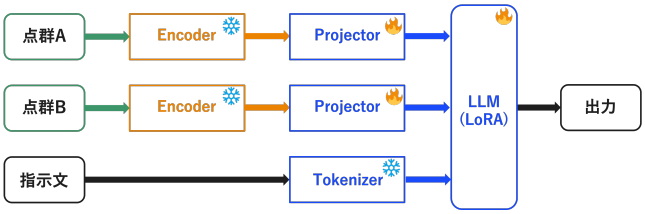

図 1: 提案する点群比較アーキテクチャ

る．(2) 点群データを直接扱う手法では，3 次元座標を持つ点群データを直接エンコーダに入力し，特徴抽出したのち，言語記述を出力する．3 次元点群を入力としてキャプションを出力する手法に，Point-BERT [14]，PointLLM [11]，PointLLM-V2 [12] がある．ShapeLLM [9] は，点群情報に加えてマルチビュー画像の情報をエンコーダに蒸留することで 3D 表現の質をさらに向上させた．ULIP-2 [13] では，マルチビュー画像・言語・点群の 3 つのモダリティを用いて分類タスクやキャプション生成を実現している．しかし，(1) と (2) のいずれの既存手法もすべて単一の入力を前提としているため，本研究の目的である 2 つの 3 次元物体の形状の差異を記述する能力は持ち合わせていない．

3. 提案手法

本研究では，2 つの 3 次元点群を入力として受け取り，それらの差異を自然言語で記述するマルチモーダル大規模言語モデルアーキテクチャを提案する．提案モデルのアーキテクチャを図 1 に示す．入力である 2 つの点群データは，それぞれ N 点で構成される．2 つの点群を，それぞれ同一の事前学習済みエンコーダに入力する．エンコーダにより抽出された点群特徴量は，線形層である Projector 層を通して LLM が解釈可能な埋め込み空間へ射影する．射影された 2 つの点群トークンは指示文のテキストトークンと結合され，LLM デコーダに入力される．この構成により，点群同士の比較キャプションの生成を行う．本提案手法では，損失関数として言語モデルの学習で標準的に用いられるクロスエントロピー損失を採用した．損失 L は，T を生成されるトークン列の長さ，y_t を t 番目の正解トークン，$y_{<t}$ をそれ以前のトークン列，P_A, P_B を入力点群トークン，I を指示文とすると，次式で表される．

$$L = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, P_A, P_B, I) \qquad (1)$$

4. データセット

3 次元空間における差異を正確に記述するためには，構造の差異を正確に記述した教師データが必要となる．しかし，3 次元空間における構造変化と言語記述が対になった大規模データセットは存在しない．そこで本研究では，構造の差異の正解データを明確に付与できる合成データセットを作成して，提案手法の性能を検証する．

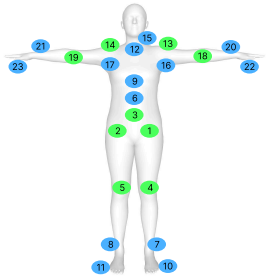

図 2: SMPL の 3D モデルと，23 個の関節．本研究で採用した制御対象関節を緑で示す．

4.1 SMPL による点群ペアの生成

図 2 に示すように，SMPL は，人体の体型を表すパラメータ β と関節の角度を表すポーズパラメータ θ を独立して制御可能な 3 次元モデルである．本実験では，体型差ではなくポーズの差異に焦点を当てるため，体型パラメータ β を固定する．そして，ポーズパラメータ θ は，パラメータの操作対象を視覚的にも認識しやすい股関節，膝関節，肩関節，肘関節，脊椎の計 5 つの関節における特定の 13 パラメータに限定し，2 つの 3 次元モデルのペア（Pose A と Pose B）を生成する．

各ペアの生成においては，13 パラメータのうち，3–5 個を選択し，それぞれの角度を決定する．データ生成の際には，以下の 3 点を条件とした．(1) 構造の視認性や自然言語記述の明確さを考慮し，角度は連続値ではなく 10° 刻みの離散値としてサンプリングする．(2) 動作とパラメータの対応関係を一意に定めるため，1 つの関節に対しては一度に 1 つの回転軸のみを操作する．(3) 同一ポーズのペアが生成されることを防ぐため，Pose A と Pose B の間で，少なくとも 1 つの関節において 10° 以上の角度差を設定する．

決定されたパラメータに基づき SMPL からメッシュを生成した後，各モデルに対して Farthest Point Sampling (FPS) アルゴリズムを適用し，$N = 8,192$ 点の点群データをサンプリングした．各点は，3 次元座標 (x, y, z) および色情報 (r, g, b) を持ち，計 6 次元のデータとして保存される．また，後述する VLM によるキャプション生成のために，各ポーズを前後左右の 4 視点からレンダリングした画像群（計 8 枚）も併せて生成した．本プロセスでは，学習用データとして 500,000 ペアを作成した．また，提案モデルを検証するため，テストデータは学習データには含まれない点群を生成し 1,000 ペア作成した．

4.2 VLM を用いたキャプション生成

生成された点群ペアに形状的特徴と差異を記述するテキストデータを付与するため，Qwen3-VL-8B-Instruct [1] を用いた自動キャプションパイプラインを構築し，キャプ

ションを生成した．

既存の MLLM は高い画像認識能力を持つ一方で，単にレンダリング画像を入力するだけでは，関節角度の定量的な変化を正確に認識することは困難である．一方で，パラメータの数値差分のみをルールベースで自然言語化した場合，「右肘が 40° 変化した」といった機械的な記述に留まり，「腕を大きく振り上げている」といった視覚的なニュアンスや自然言語らしさ，全体的なバランスや動作の意図等ポーズの印象が欠落する．そこで本研究では，SMPL の生成パラメータから算出したパラメータの角度の差分と，レンダリング画像から得られる視覚的な情報を統合するハイブリッドなキャプション生成を行った．具体的には，ポーズ単体の記述を行った個別キャプションと，ポーズ間の差異を記述する比較キャプションの 2 種類のキャプションを生成・付与した．

5. 実験

提案手法が入力された 2 つの点群の形状を言語で表現する能力と，両者の差異を言語化する能力を有しているかを生成したデータセットを用いて検証した．

5.1 実験設定

提案手法のベースモデルとして 3 次元点群からキャプション生成を行う PointLLM-7B v1.1 [11] を採用した．本モデルの点群エンコーダには ULIP-2 [13] で事前学習された Point-BERT [14] を採用しており，入力された点群から特徴を抽出する．Projector 層は 3 層の多層パーセプトロン（Multilayer Perceptron: MLP）で構成され，エンコーダが出力した特徴量を LLM が解釈可能な埋め込み空間への射影を行う．LLM のバックボーンは Vicuna-7B v1.5 [10] を使用し，入力トークンから自然言語の回答を生成した．Point-BERT エンコーダおよび Vicuna-7B のベースパラメータはすべて固定し，学習による更新は行わない．一方，Projector 層に関しては，全パラメータを学習対象とした．また，LLM に対しては Low-Rank Adaptation (LoRA) [4] を適用し，Attention 層に追加されたアダプタのパラメータのみを学習させることで，計算コストを抑えつつファインチューニングを行った．LoRA のランクは $r = 8$，スケーリング係数は $\alpha = 16$ とした．

表 1: 比較キャプション生成の精度評価結果

Level	Precision	Recall	F1
Lv1 部位検出	0.888	0.938	0.912
Lv2 動作+方向	0.861	0.910	0.885
Lv3 角度閾値 0°	0.615	0.650	0.632
Lv3 角度閾値 10°	0.834	0.881	0.857
Lv3 角度閾値 20°	0.856	0.904	0.880

テストでは，Pose A と Pose B をそれぞれ説明するキャプションの生成と，Pose A, B の比較キャプションを生成するよう，指示文を入力した．そして，その出力に関して精度を評価した．以降，Pose A と Pose B をそれぞれ説明するキャプションの生成を「個別タスク」，Pose A と Pose B の比較キャプションを生成するタスクを「比較タスク」と呼ぶ．

学習，テストには，4 節にて述べたパイプラインにより生成された 500,000 ペア，1000 ペアのデータをそれぞれ使用した．学習および評価は，NVIDIA H100 PCIe GPU を搭載したサーバー上で実施した．

5.2 学習方法

本研究では，モデルに個別の点群の言語記述と点群間の比較を段階的に習得させるため，2 段階に分けて学習した．1 段階目の学習では，モデルが入力された 2 つの点群を個別に記述する能力の獲得を目的とした．学習データとして，それぞれの点群を説明するキャプションのみを使用した．2 段階目の学習では，本研究の目的である比較タスクを学習する．ただし，比較キャプションのみを学習させると，1 段階目で獲得した個別の点群認識能力を喪失するリスクや，タスク間の混同が生じる可能性がある．そこで，比較キャプションデータを中心としつつ，個別キャプションデータを少量再度学習データに混合する方式を採用した．具体的には，1 エポックあたりの学習データ比率を比較キャプションと個別キャプションを 8 : 2 に設定した．

5.3 評価指標

生成されたキャプションの精度評価に，自然言語処理および画像キャプション生成分野で標準的に用いられる METEOR [2] を採用した．加えて，比較タスクで Pose A と Pose B で変化がある部位を認識できているかを確認するため，比較キャプションの生成結果に変化がある部位への言及の有無の精度を評価した．生成された比較キャプションと正解のキャプションから身体部位の変化記述を抽出し，これらのマッチング精度を評価した．マッチングは 3 段階で評価した．Level 1 は部位のみの一致，Level 2 は部位・動作種別・方向の一致，Level 3 は Level 2 に加えて角度差が閾値以内かを判定する．閾値は 0°（完全一致），10°，20° の 3 種を設けた．

精度は，全テストペアでの生成キャプション内の正しい部位数の総数を true positive (TP) とし，生成キャプション内の部位の総数，正解キャプションの部位の総数をそれぞれ分母として precision，recall，を計算したのち，これらをもとに F1 score を算出した．

5.4 実験結果と考察

単体タスクでの METEOR は 0.334 であった．提案手法

生成されたテキスト:
"pose_a_caption": "This is a dynamic running or lunging pose. The arms are extended straight out to the sides for balance, and one leg is forward with a bent knee while the other is back, also bent. The body is leaning forward, suggesting motion."

"pose_b_caption": "This is a dynamic, open-armed pose where one arm is raised high and the other is extended out to the side, with legs in a wide, slightly bent stance. The body is balanced and upright, suggesting a moment of reaching, stretching, or celebration."

"comparison_caption": "Compared to Pose A, Pose B has left leg moved 30 degrees outward from Pose A, right leg moved 40 degrees backward from Pose A, left knee bent 10 degrees from Pose A, right knee straightened 60 degrees from Pose A, left arm moved 20 degrees forward from Pose A, left arm raised 50 degrees from Pose A, and right arm moved 10 degrees forward from Pose A. Pose B differs from Pose A by introducing dynamic asymmetry and openness. While Pose A maintains a stable, symmetrical T-pose with arms level, Pose B features one arm raised high and the other extended, creating a reaching, expressive posture. This asymmetry shifts the center of gravity, suggesting motion or celebration, contrasting with Pose A's static, balanced stance."

正解データ:
"pose_a_caption": "The person has their arms outstretched horizontally to the sides, with straight, slightly bent elbows. Their legs are in a running or jumping motion, with one leg forward and bent, and the other back and extended, creating a dynamic, mid-air posture."

"pose_b_caption": "This is a dynamic, open-armed pose where one arm is raised high and the other is extended out to the side. The legs are spread apart in a wide stance with bent knees, giving the body a balanced, expressive, and slightly floating appearance."

"comparison_caption": "Compared to Pose A, Pose B has left leg moved 30 degrees outward from Pose A, right leg moved 20 degrees backward from Pose A, right knee straightened 60 degrees from Pose A, left arm moved 20 degrees forward from Pose A, left arm raised 50 degrees from Pose A, and right arm moved 10 degrees forward from Pose A. Pose B differs from Pose A by featuring a dynamic asymmetry: one arm is raised high while the other extends horizontally, creating a reaching or celebratory gesture. The legs are spread wide with bent knees, shifting the center of gravity lower and more outward, suggesting a stable, expressive stance. This contrasts with Pose A's more balanced, forward-moving posture, implying a focus on propulsion rather than open-ended expression."

図 3: 比較キャプション生成例．入力した 3D データ（上）と，生成されたテキスト（中），正解データ（下）．生成されたテキスト中の黒文字の下線部は 3 次元データと不一致の記述を示し，赤色の下線部は正解データと一致しない記述を示す．

のベースモデルである PointLLM の一般物体でのキャプション生成が 0.1192 であり，本研究での単体タスクの対象が人物のみであることを考えると，妥当な精度であると考えられる．また，比較タスクの METEOR は 0.405 で，単体記述タスクを上回っている．これは，比較タスクの正解データが「Compared to Pose A, Pose B has...」という定型的な構文を多く含んでいるため，モデルがその文法構造や頻出フレーズを容易に学習できたためと考えられる．

続いて，比較タスクで生成されたキャプションにおける変化部位の検出精度を表 1 に示す．Level 1 の部位のみの精度は recall が 0.938 と高く，変化がある部位を取りこぼしなく識別できていることがわかる．また，Level 2, 3 の動作の向きや角度の違いまでを考慮すると，部位のみよりも精度が低下する．変化の向きや角度は，キャプション生成で精度が多少悪くても 3 次元データを直接比較することで正確に算出することができるため，本手法の有効性には大きな影響はないと考えられる．

最後に，テストデータの入力と生成されたキャプション，正解データの一例を図 3 に示す．Pose A と Pose B の単体のキャプション生成結果は，入力された 3D データとは異なる姿勢の表現が見られる．この生成キャプションが正解データと一致していることから，正解データのキャプションが正しく付与されていないことが間違いの原因である．後半の変化の言語的な説明に関しては，例のように「dynamic, athletic pose」「stable, upright posture」などにパターン化されている上，変化部位の様子の違いよりも全体の姿勢に言及している傾向が見られた．

6. 結論

近年，物体の 3 次元形状を自然言語で表現する研究が発展してはいるものの，3 次元形状の差異を自然言語で表現する研究は存在していない．そこで本研究では，2 つの 3 次元点群を入力として受け取り，それらの差異を自然言語で記述するためのモデルを提案した．また，SMPL を活用した合成データセットを構築し，提案手法の評価に用いた．実験の結果，提案モデルは変化部位の検出精度の recall が 0.938 となり，変化部位を取りこぼしなく認識し，自然言語で表現できた．

参考文献

[1] Bai, S. et al.: Qwen3-VL Technical Report, *arXiv preprint arXiv:2511.21631* (2025).

[2] Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72 (2005).

[3] De Gélis, I., Corpetti, T. and Lefèvre, S.: Change detection needs change information: Improving deep 3-D point cloud change detection, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp. 1–10 (2024).

[4] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al.: LoRA: Low-rank adaptation of large language models, *Proceedings of International Conference on Learning Representations*, pp. 1–13 (2022).

[5] Kerbl, B., Kopanas, G., Leimkühler, T. and Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering., *ACM Transaction of Graphics*, Vol. 42, No. 4, pp. 139–1 (2023).

[6] Li, J., Li, D., Savarese, S. and Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *Proceedings of International Conference on Machine Learning*, pp. 19730–19742 (2023).

[7] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. and Black, M. J.: SMPL: A Skinned Multi-Person Linear Model, *ACM Transactions on Graphics (TOG)*, Vol. 34, No. 6 (2015).

[8] Luo, T., Rockwell, C., Lee, H. and Johnson, J.: Scalable 3D Captioning with Pretrained Models, *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 1–31 (2023).

[9] Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L. and Ma, K.: ShapeLLM: Universal 3D Object Understanding for Embodied Interaction, *Proceedings of European Conference on Computer Vision*, pp. 214–238 (2024).

[10] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. et al.: Llama 2: Open foundation and fine-tuned chat models, *arXiv preprint arXiv:2307.09288* (2023).

[11] Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J. and Lin, D.: PointLLM: Empowering Large Language Models to Understand Point Clouds, *Proceedings of European Conference on Computer Vision*, pp. 131–147 (2025).

[12] Xu, R., Yang, S., Wang, X., Wang, T., Chen, Y., Pang, J. and Lin, D.: PointLLM-V2: Empowering Large Language Models to Better Understand Point Clouds, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15 (2025).

[13] Xue, L., Yu, N., Zhang, S., Panagopoulou, A., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C. and Savarese, S.: ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding, *Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27081–27091 (2024).

[14] Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J. and Lu, J.: Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling, *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19291–19300 (2022).