3次元点群データ間の差異に対する自然言語記述の生成

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出来る点群比較アーキテクチャを提案する．Skinned Multi-Person Linear Model (SMPL) 人体モデル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 [1] をはじめとする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 [2] や ShapeLLM [3] 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．この作業を機械の目で客観的に行えれば，植物学の発展に大きく寄与できる．

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 モデル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．提案手法の学習と検証には，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，タスクの実現可能性を検証する第一段階として，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，かつ変化量を定量的に取得できる Skinned Multi-Person Linear Model (SMPL) 人体モデル [4] を使用し，多様な人体ポーズのペアを生成し，MLLM 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．

[1] 大阪公立大学 大学院情報学研究科
Graduate School of Informatics, Osaka Metropolitan University
[a)] sz22932s@st.omu.ac.jp
[b)] yuzuko@omu.ac.jp
[c)] masa.i@omu.ac.jp

2. 関連研究

本章では，関連する研究領域として，3次元空間の変化を検出する 3D Change Detection と，単一の3次元データを入力として言語生成を行う MLLM について述べる．

2.1 3D Change Detection

3次元空間における変化の特定は，点群データを対象とした 3D Change Detection として研究されている．De らによる Siamese KPConv [5] は，重みを共有した2つの Kernel Point Convolution を提案した．各点群の局所的な特徴を抽出し，その特徴マップ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における変化を検出した．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，変化の内容や動作を自然言語で説明できない．

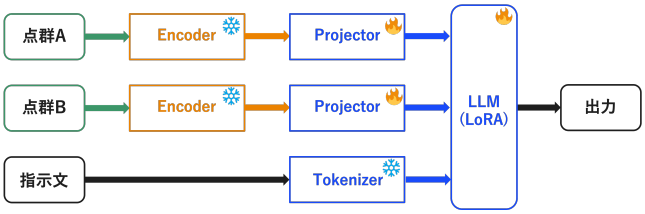

図 1 提案する点群比較アーキテクチャ

2.2 3D データに対応した MLLM

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 MLLM を活用して自然言語による説明を出力する．Luo らの Cap3D [6] 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) [7] 等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 [8]，PointLLM [2]，PointLLM-V2 [9] がある．ShapeLLM [3] 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 [10] 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．

(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．そのため，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．

3.1 点群比較アーキテクチャ

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 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 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．

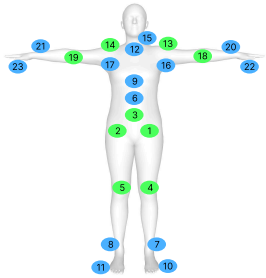

図 2 SMPL モデルの 3D モデルと，23 個の関節．本研究で採用した制御対象関節を緑で示す．

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

3.2 モデルの学習

本研究の提案モデルは，2 段階の学習方法を採用する．1 段階目は，点群の区別能力の獲得を目的とし，入力点群のいずれか一方のみを記述するタスクを学習させる．2 段階目は，1 段階目で得た能力を基盤とし，本研究の目的である 2 つの点群の差異を記述するタスクを学習させる．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混合して学習を行う．両段階において Projector はスクラッチ学習，LLM は LoRA [11] によるファインチューニングを行う．

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．

4.1 SMPL モデルによる点群ペアの生成

図 2 に示すように，Skinned Multi-Person Linear Model (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．学習データの生成では，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

表 1 データ生成に使用した制御パラメータと可動域制約

部位	動作内容	軸	可動域（°）
股関節	前後屈	X	$-90 \sim +40$
	開脚	Z	$0 \sim +30$
膝関節	屈曲	X	$0 \sim +90$
肩関節	前後	Y	$-90 \sim +90$
	上下	Z	$-90 \sim +90$
肘関節	屈曲	Y	$-90 \sim 0$
胴体	前屈	X	$0 \sim +30$

る．そして，ポーズパラメータ θ は，表 1 に示すように，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 モデル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

表 2 パラメータ差分から自然言語への変換ルール

部位	差分	生成される自然言語
股関節 (前後)	+	left/right leg moved backward
	−	left/right leg moved forward
股関節 (横)	+	left leg moved inward / right leg moved outward
	−	left leg moved outward / right leg moved inward
膝関節	+	left/right knee straightened
	−	left/right knee bent
肩関節 (前後)	+	left/right arm moved backward
	−	left/right arm moved forward
肩関節 (上下)	+	left/right arm lowered
	−	left/right arm raised
肘関節	+	left elbow bent / right elbow straightened
	−	left elbow straightened / right elbow bent
胴体	+	torso straightened
	−	torso leaned forward

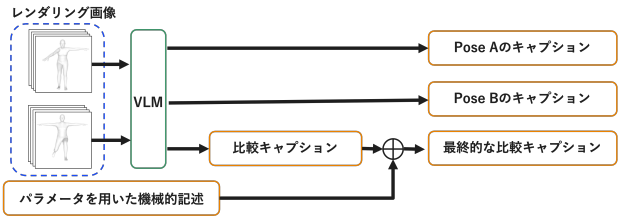

図 3 MLLM を用いた自動キャプションパイプライン

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 MLLM 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 4 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．さらに，提案手法が学習データと異なる体型に対してもロバストに動作するかを検証するため，体型を変化させたテストデータも作成した．具体的には，通常のテストデータから 100 ペアを抽出し，ポーズパラメータは元の値を保持したまま，SMPL の 10 次元の体型パラメータ β の各成分を独立にサンプリングした．Pose A と Pose B には互いに異なる β を与えることで，2 つの入力点群が異なる体型を持つペアを構築した．体型パラメータ以外の生成条件は学習データと同一である．

4.2 MLLM 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 [12] 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．自動キャプション生成のパイプラインを図 3 に示す．個別キャプションは Qwen3-VL に各ポーズの 4 視点画像をそれぞれ入力し，手足の位置関係，身体の向きなどその姿勢の視覚的な特徴を

記述させた．比較キャプションは，SMPL の生成パラメータを用いた機械的記述と MLLM による自然言語記述の 2 種から構成される．前者の機械的記述は，Pose A と Pose B のパラメータ差分を計算し，定量的な差異を記述する．表 2 に示すような変換ルールに基づき，パラメータの差異を自然言語表現へと変換する．ただし，生成される記述の末尾には，基準となるポーズを明示するために「from Pose A」が付与される．表内の差分は各パラメータにおける Pose A の値から Pose B の値を引いたものを指す．後者の自然言語記述は，各ポーズのレンダリング画像計 8 枚の画像を Qwen3-VL に入力し，比較文章を生成させる．こうして生成された 2 種の比較キャプションを繋げることで，定量的な正確性を担保しつつ，人間にとって自然で解釈しやすい比較キャプションを生成する．

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検証した．加えて，学習データ規模の影響と，学習時と異なる体型に対するロバスト性についても評価した．検証には，4 章で生成したデータセットを用いた．

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 [2] を採用した．本モデルの点群エンコーダには ULIP-2 [10] で事前学習された Point-BERT [8] 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 [13] を使用し，入力トークンから自然言語の回答を生成する．Point-BERT エンコーダおよび LLaMA-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 LoRA 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．

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．

表 3 比較キャプション生成の精度評価結果

	200,000 ペア			500,000 ペア		
Level	Precision	Recall	F1	Precision	Recall	F1
Lv1 部位検出	0.898	0.940	0.918	0.979	0.978	0.978
Lv2 動作+方向	0.868	0.913	0.890	0.977	0.975	0.976
Lv3 角度閾値 0°	0.613	0.644	0.628	0.913	0.912	0.913
Lv3 角度閾値 10°	0.838	0.881	0.859	0.976	0.975	0.975
Lv3 角度閾値 20°	0.861	0.905	0.882	0.976	0.975	0.976

学習には，4 章にて述べたパイプラインにより生成された 500,000 ペアのデータを使用した．学習は 3 章で述べた 2 段階学習に従い，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で 8 : 2 に設定した．LoRA のランクは $r = 8$，スケーリング係数は $\alpha = 16$ とした．学習データの規模が性能に与える影響を評価するため，200,000 ペアで学習したモデルとの比較も行った．テストには，学習データに含まれない 1,000 ペアを使用した．さらに，4 章で述べた体型を変化させた 100 ペアのテストデータを用いて，体型変化に対するロバスト性も評価した．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 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 [14] 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異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．マッチングの度合いには，以下の 3 レベルを用いた．

Level 1: 部位検出　変化が生じた部位の集合単位でマッチングを行う．当該部位に何らかの変化があることを検出できたかを評価する．例えば，「正解キャプション: 左脚が前方に動いた，生成キャプション: 左脚が外側に動いていた」といった場合でも左脚が動いたと同一視する．

Level 2: 定性的一致　部位・動作種別・方向の組み合わせ単位でマッチングを行う．変化の種類と方向まで正しいかを評価し，角度は問わない．

Level 3: 角度閾値　Level 2 のマッチングに加えて，角度差が閾値以内であるかを判定する．閾値は 0°（完全一致），10°，20° の 3 段階を設定する．

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．

5.3 実験結果と考察

本節では，まず提案手法の生成精度と学習データ規模の

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 4 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），と正解データ（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．

影響を評価し，次に学習データと異なる体型に対するロバスト性を検証する．

5.3.1 生成精度の評価

まず，500,000 ペアで学習したモデルの評価結果を述べる．単体タスクでの METEOR は 0.339 であった．提案手法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 0.426 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 3 に示す．Level 1 の部位のみの精度は F1 スコアが 0.978 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．Level 2 の動作方向までを考慮した場合でも F1 スコアは 0.976 を保ち，Level 3 の角度閾値 10°，20° でも同等の精度を維持している．角度閾値 0°（完全一致）では F1 スコアが 0.913 に低下するが，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．

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 4 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は正解データとは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．

次に，学習データ規模の影響を確認するため，200,000 ペアで学習したモデルとの比較を行う．単体タスクの METEOR は 0.326，比較タスクの METEOR は 0.400 となり，いずれのタスクでも 500,000 ペアモデルを下回った．また，部位検出の F1 スコアも Level 1 で 0.918，角度閾値 0° の Level 3 で 0.628 と，500,000 ペアモデルからそれぞれ 0.060，0.285 低下した．この結果から，学習データ規模の増加は特に角度の正確な推定に大きく寄与することが示された．

5.3.2 体型変化に対するロバスト性評価

4 章で述べた体型を変化させたテストデータ 100 ペアを用いて評価した．評価結果を表 4 に示す．比較タスクの

METEOR は 0.411 で，通常のテストデータ（0.426）と比べて 0.015 の低下に留まった．部位検出の Level 1 の F1 スコアは 0.914 であり，通常のテストデータ（0.978）に対して 0.064 の低下となった．一方，角度閾値 0° の Level 3 の F1 スコアは 0.599 となり，通常のテストデータ（0.913）から 0.314 と大きく低下した．

この結果から，「どの部位が変化したか」という定性的な検出は体型変化に対してロバストに行える一方，正確な角度の推定は体型の違いに影響を受けることが分かる．これは，モデルが標準体型の点群で学習した角度推定のパターンが，体型変化によって崩れるためと考えられる．ただし，角度閾値を ±10° に緩和すると F1 スコアは 0.844 となり，ある程度の誤差を許容すれば実用的な精度は維持される．応用を考える上では，前述の通り 3 次元データから直接関節角度を算出できるため，角度推定精度の低下は本手法の有効性を大きく損なうものではないと考えられる．部位の検出に関して体型変化下でも高い精度を維持していることは，今後，形状が多様な対象への適用を考える上で有望な結果と言える．

6. おわりに

近年の 2 次元画像処理から 3 次元空間への拡張の進展に伴い，3 次元データに対応した MLLM モデルが開発され，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．しかし，これらのモデルは単一の点群入力にのみ対応しており，2 つのデータを同時に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有していない．さらに，3 次元データにおける構造の差異を記述するための大規模なデータセットも存在しない．

そこで本研究は，2 つの 3 次元点群を入力として受け取り，それらの構造的な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，学習データの不足を補うため，SMPL モデルを活用した合成データセット構築手法を提案した．

実験では，提案手法が入力された 2 つの点群を個別に認識，区別する能力と，両者の構造的な差異を言語化する能力を有しているかを検証した．実験の結果，提案モデルは変化部位の検出精度の F1 スコアが 0.978 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．また，± 10° の許容範囲を設ければ角度の違いについても F1 スコア 0.975 と高精度で記述できることが確認された．一方で，学習データに含まれない体型の異なるペアについては，大まかな変化は認識できているものの，精密な関節角度の推定は困難であった．

本研究には，今後解決すべき 3 つの課題がある．第一に，比較キャプションの品質をより適切に評価する手法の確立である．本研究では，METEOR に加えて部位単位の Precision/Recall 評価を導入することで，変化部位の検出精度と角度誤差を定量的に分離して評価した．しかし，より細かな差分や姿勢の自然さまで含めて適切に評価する手法の確立は今後の課題として残されている．第二に，学習データの品質改善が必要である．学習データを VLM によるキャプションで生成しているものの，ハルシネーションや余計な文言が含まれており，学習の品質低下につながっていると考えられる．第三に，SMPL 以外の幅広いドメインへの対応である．植物の形態比較など，実世界のさまざまな応用場面に対応するため，より多様な 3 次元データへ適用可能なモデルへの拡張が求められる．これらの課題に取り組むことで，3 次元点群比較タスクの基盤が確立され，より高度な解析の実現に貢献できる．

表 4 体型変化したテストデータに対する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	0.870	0.962	0.914
Lv2 動作+方向	0.833	0.936	0.882
Lv3 角度閾値 0°	0.566	0.635	0.599
Lv3 角度閾値 10°	0.798	0.896	0.844
Lv3 角度閾値 20°	0.826	0.928	0.874

参考文献

[1] Kerbl, B., Kopanas, G., Leimkühler, T. and Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering., *ACM Trans. Graph.*, Vol. 42, No. 4, pp. 139: 1–14 (2023).
[2] 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).
[3] 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).
[4] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. and Black, M. J.: SMPL: A Skinned Multi-Person Linear Model, *ACM Trans. Graph.*, Vol. 34, No. 6, pp. 248: 1–16 (online), DOI: 10.1145/2816795.2818013 (2015).
[5] 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).
[6] 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).
[7] Li, J., Li, D., Savarese, S. and Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *International conference on machine learning*, pp. 19730–19742 (2023).
[8] 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, *Proceeding of 2022 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR), pp. 19291–19300 (2022).

[9] 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 (online), DOI: 10.1109/TPAMI.2025.3590784 (2025).

[10] 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).

[11] 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, pp. 1–13 (2022).

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

[13] 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).

[14] 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).