マルチモーダル LLM を活用した現実世界の状況認識に基づく運動モデリング

鈴木慎平 [1]　吉岡真治 [2]
[1] 北海道大学大学院情報科学院　[2] 北海道大学大学院情報科学研究院
suzuki.shinpei.h8@elms.hokudai.ac.jp　yoshioka@ist.hokudai.ac.jp

概要

近年のマルチモーダル大規模言語モデル (MLLM) の発展に伴い、大学入試センター試験レベルの物理の問題について、模式化された絵から得られる情報なども活用しながら物理法則に基づいた適切な式を立てて説明を行うことが可能となっている．しかし、MLLM は人間と異なり、現実の状況を示す写真と模式化された絵の対応関係を認識することは簡単ではない．そのため、写真による状況説明では、模式化した絵と異なり、適切な立式による運動の分析ができないことがある．本論文では、簡単なばねや斜面の摩擦のある運動の初期状態についての画像から、MLLM がどのように状況を理解するかを分析するとともに、ファインチューニングにより適切に状況を認識した説明を生成できるかについて検証した．

1　はじめに

日本の生産年齢人口の減少が世界の中でも著しく進んでいる．2032 年には 7000 万人を割り、2070 年には 4535 万人となり、ピークの 5 割の水準まで減少すると予測されている [1]．特に製造業は日本の国内総生産の 2 割を占めている主要産業であるが [2]、2030 年には 38 万人の人手が不足すると言われている [3]．人手不足・技能継承などの問題に対して、効率的な物理シミュレーションの活用などによる迅速な対応が求められており、様々なデジタル技術の導入が進められている [4]．物理シミュレーションの一例として故障診断が挙げられるが [5]、従来のシンボリックなシステムでは、想定外の現象に対応するために事前に全ての知識を用意することが困難であった．

そこで、MLLM を応用させることが期待される．従来は、大学入試センター試験物理のような必要な情報が与えられる模式的な問題に対して、「ロボットは東大に入れるか」プロジェクト（東ロボ）のように人手を介入させて回答を生成する必要があった [6]．一方で、MLLM は人手を介入させることなく概ね正しい説明を生成することが可能になってきている．実際に、故障診断を例に MLLM を用いてモデルを構築している例が複数存在している [7, 8]．

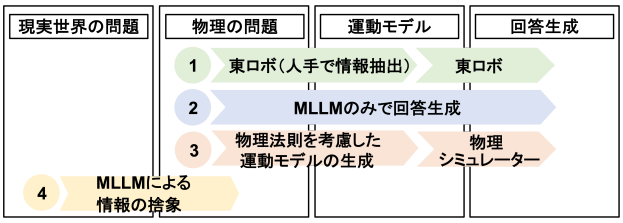

図 1: 運動モデルの生成（本稿は 4 について研究した内容である）

しかし、MLLM の物理世界とその法則を理解する能力には限界があり、物理法則を適切に組み合わせて正しく物体の挙動推定を行うことは容易ではない．[9] は、我々が無意識に理解しているものは重力で落ちるなどの直感的物理の理解が MLLM にとって困難であることを示している．一方で、現実世界の空間的な関係性と物理的な法則に関する情報を含む追加データを用いて学習させるフィジカル AI の研究が進められており [10]、現実の物理的な世界を認識、理解して複雑な行動ができるようになることが期待されている．

しかし、我々はニューロシンボリック AI の考え方に基づき [11]、MLLM を用いて推論に必要な情報を上手く抽出し、シンボリックなシステムと組み合わせて網羅的かつ汎用的なモデルを構築する方針が効果的であると考えている．我々は、MLLM に問題に関連する物理法則を列挙させるとともに、東ロボで用いられていた Modelica のような物理シミュレーターの実行可能なコードを生成させる手法を提案している [12]．しかし、現実世界の物理現象の問題を扱うためには、画像などのマルチモーダルな情

報を利用して、先述の物理の問題に対応づける必要がある．

本研究では、ばねの減衰運動や斜面上を物体が滑り落ちる運動のような、比較的単純な物理現象を題材として、MLLM がどのように状況を理解するかなどについて実験を行い、課題を整理した．さらに、MLLM にファインチューニングを施し、物体の運動方向を認識して適切に物理現象の説明を生成させることがどの程度可能になるかについて議論した．

2 予備実験

本節では、MLLM が現実世界の状況をどのように認識できるのかについて行った実験に関して述べる．[1)]

gpt-5 に指示文と画像を含むプロンプトを入力して、生成される出力の内容を分析した．temperature はランダム性を持たせるために 1.0 に設定した．先端に物体が取り付けられているばねがテーブルの上に置いてあり、摩擦を引き起こす布・紙・クリアファイルが様々な場所に配置してある状況（ばねの運動方向線上、ばねの運動方向よりも手前にずれた位置、ばねの運動方向よりも奥側にずれた位置、ばねが届かない可能性のあるほどばねと離れた位置）において、ばねを引っ張って離すと、摩擦を引き起こす物体に接触するかどうかを認識した上でばねの減衰運動を説明し、Modelica のシミュレーションコードを生成することを、gpt-5 がどの程度正しく行うことができるのかに関して検証した．

結果、画像内の物体や関係する物理法則を列挙することは十分に可能であったが、運動への影響を考慮した物体の位置関係の認識には課題があった．ばねが摩擦を引き起こす物体に接触することを考慮していなかったり、接触しないはずなのに接触するという内容の説明を生成することが見られた．生成した Modelica のコードに関しては、摩擦を考慮しない場合のコードは正しく動作したが、摩擦を考慮する場合のコードは正しく動作しなかった．

gpt-5 のような MLLM は、人間が無意識に行っているように上手く物体の位置関係を認識して正しく情報を捨象し、適切に推論を行うには課題があることがわかった．MLLM に入力するプロンプトの工夫次第で所望する説明を生成することができる場合もあるが、基本的に生成する説明の質が不安定であることが理解できる．

1) 本予備実験については、第 20 回言語処理若手シンポジウム (YANS2025) [S3-P51] 生成 AI を活用した物体の運動に関するモデリング ～現実世界の状況認識に基づく運動モデリングに向けて～鈴木慎平, 吉岡真治で発表済み

3 実験

2 節で述べた様に、人間なら無意識に行っている様な物体の運動を考慮した位置関係の把握が MLLM には困難であることが確認された．この問題が本質的に MLLM にとって難しい問題なのか、MLLM は判断に有用な情報を有しているが、学習が不足しているために答えられないのかを議論する必要がある．そのため、データセットを用意し、MLLM に対してファインチューニングを施し、物体の位置関係を認識して運動への影響を考慮した物理現象の説明を生成できるようになるかを調査する実験を行うことにした．本節では、作成したデータセットの説明、行った実験の内容とその結果・考察を述べる．

3.1 データセットの作成

本研究で使用するために作成したデータセットについて説明する．題材は 2 つあり、運動に対して本質的に関係のある情報と無い情報を混合させた内容となっている．1 つ目はばねの減衰運動に関する問題（spring）、2 つ目は斜面を物体が滑り落ちる運動に関する問題（slope）である．2 つともテキストは英語で用意し、画像 1 枚と入力文、出力文をペアとする形式となっている．各データセット内のデータの順番はシャッフルされている．実際の画像は A.2 節の付録を参照されたい．

(1) spring-h：摩擦を引き起こす布が配置してあるテーブルの上で、先端に物体が取り付けられているばねが横方向に行う減衰運動に関するデータセットである．布の配置パターンは次の 5 つである．(i) 振動するばねと接触する、(ii) ばねの運動方向よりも手前側に配置されていて接触しない、(iii) ばねの運動方向よりも奥側に配置されていて接触しない、(iv) ばねの運動方向よりも手前側と奥側の 2 箇所に配置されていて接触しない、(v) ばねが届かない程遠くに配置されていて接触しない．各パターンに対して、布の色が異なるものを 45 件と、ばねに取り付けられている物体が異なるものを 63 件用意しており、全 108 件である．

(2) slope：茶色とオレンジ色の 2 つの木製ブロックが配置されている木製板斜面上を、斜面上部中央から茶色い木製のブロックを真下に滑り落とす運動に関するデータセットである．オレンジ色の木製ブ

ロックの配置パターンは次の 3 つである．(i) 斜面中央に配置されており茶色の木製ブロックと接触する、(ii) 斜面左部に配置されており茶色の木製ブロックと接触しない、(iii) 斜面右部に配置されており茶色の木製ブロックと接触しない．各 42 件で合計 126 件用意した．なお、オレンジ色のブロックの色は様々な色の画像を用意している．

(3) spring-h-slope：(1) と (2) を合わせたもので、合計 234 件用意した．

(4) spring-hv：(1) の spring-h の各データに関して時計回りに 90 度回転させて縦方向の減衰運動を行うデータを追加したもので、合計 216 件用意した．

(5) spring-hv-slope：(3) の spring の各データに対して時計回りに 90 度回転させて縦方向の減衰運動を行うデータを追加したもので、合計 342 件用意した．

(6) spring-mini：(1) の spring-h の各データに対して、(iv)、(v) の配置パターンのデータを削除し、(i) (iii) の配置パターンのデータのみを抽出してそれぞれ 90 度、180 度、270 度回転させたものも追加したものであり、合計 280 件用意した．

3.2 実験設定

本実験では、Apple M3 Ultra チップ（32 コア：24 performance cores、8 efficiency cores、 512GB のユニファイドメモリ）の Apple Mac Studio を使用して MLLM のファインチューニングを行った．MLLM は huggingface の mlx-community/Llama-3.2-11B-Vision-Instruct-4bit を使用し、これに対して、MLX-VLM[2] を用いて 3.1 節で説明した 6 つのデータセットをそれぞれ別々に用いて QLoRA[13] を行った．ハイパーパラメータについては MLX-VLM のデフォルトのままに設定した（learning-rate：1e-5、batch-size：1、lora-rank：10、lora-alpha：0.1、lora-dropout：0.1、steps：0）．エポック数について、spring-h、slope、spring-h-slope それぞれを用いた学習は 5 エポック、spring-hv、spring-hv-slope それぞれを用いた学習は 2 エポックの学習を行った．これは、それぞれのデータセットのデータ数を考慮して、学習データ数 * エポック数に差が出ないようして結果を比較するためである．なお、spring-mini は最もデータ数が多いが、直接的に他の学習結果との比較を行わず、5 エポック学習を行った．テストデータは spring の問題に関するものを合計 30 件用意した．ばねとの接触を問う物体として青い布、白い紙、透明なクリアファイルがそれぞれ配置された画像について、それぞれ 5 つの配置パターンを想定した合計 15 件と、それらを 90 度時計回りに回転させたものの 15 件を合わせて合計 30 件を使用した（spring-mini で学習した場合のみ、3 つの配置パターンの合計 9 件をテストデータとして使用した）．ばねに取り付けられている物体はオレンジ色の消しゴムのみに統一してある．評価指標については、ばねが布や紙、クリアファイルに接触する場合を True、接触しない場合を False として、適合率（Precision）、再現率 (Recall)、F1 値を用いて評価した．また、学習前のモデルでテストデータに対して 1 回推論させた結果についても評価した．なお、本実験では全て temperature を 0.0 に設定して推論を行った．

2） https://github.com/Blaizzy/mlx-vlm

3.3 結果・考察

学習結果は表 1 および表 2 のようになった．表 1 に関して、Precision を P、Recall を R、F1 値を F1 と表記しており、各指標について分母の値が 0 になった場合は 0 として扱った．また、表 2 についてはテストデータを次の種類に分類して集計した（「接触」：ばねと布や紙、クリアファイルが接触する問題．「方向」：布や紙、クリアファイルがばねの運動方向線上に配置されておらず接触しない問題．「距離」：ばねと布や紙、クリアファイルとの距離が大きく，接触しない可能性が大きい問題）．

学習前のモデルについては全ての評価指標において 0.000 となり、ばねと布や紙との接触について考慮していない結果となった．また、基準として全ての問題に対して接触するという内容の回答を生成した場合、Precision は 0.200、Recall は 1.000、F1 値は 0.333 となった．これをベースラインとする．

まず、spring の問題のみを学習した場合の結果について考察する．spring-h で横方向のばねの運動に関する問題を学習した結果、横方向のばねの運動に関する問題に対する F1 値がベースラインを上回り、学習によってばねと布や紙との接触を考慮するようになったと考えられる．しかし、縦方向のばねの運動に関する問題に対しては F1 値が低いことから、学習した問題のばねの運動方向と一致しない問題に対する推論性能が低いと考えられる．一方、縦方向のばねの運動に関する問題も含む spring-hv を学習した結果、縦方向・横方向の問題の両方に対する各評価指標について spring-h 以上のスコアとなっ

表 1: ばねの問題に対する実験結果

学習データ	ばね横方向 P	R	F1	ばね縦方向 P	R	F1
学習前	0.000	0.000	0.000	0.000	0.000	0.000
全て接触	0.200	1.000	0.333	0.200	1.000	0.333
spring-h	0.286	0.667	0.400	0.250	0.333	0.286
spring-hv	0.333	0.667	0.444	0.286	0.667	0.400
slope	0.000	0.000	0.000	0.000	0.000	0.000
spring-h-slope	0.200	0.667	0.308	0.500	0.667	0.571
spring-hv-slope	0.273	1.000	0.429	0.200	0.667	0.308
spring-mini	0.400	0.667	0.500	0.250	0.333	0.286

表 2: ばねの問題に対する問題別正解率

学習データ	ばね横方向 接触	方向	距離	ばね縦方向 接触	方向	距離
spring-h	0.667	0.333	0.000	0.333	0.778	0.000
spring-hv	0.667	0.556	0.333	0.667	0.667	0.000
slope	0.000	0.000	1.000	0.000	0.000	1.000
spring-h-slope	0.667	0.111	0.000	0.667	0.778	0.000
spring-hv-slope	1.000	0.000	0.000	0.667	0.000	0.000
slope-mini	0.667	0.222	-	0.333	0.222	-

た．縦方向の運動に関するデータも学習させることで横方向、縦方向の運動に対する推論能力が向上したと考えられる．そのため、運動方向に関する推論能力は学習したデータセットの問題の運動方向の影響を受けていると捉えることができる．このことからも、MLLM が与えられた問題に対して類似した事例をもとに回答を生成しており、情報を適切に捨象して物理法則をもとに物理現象を推論することに対して課題があると考えられる．

次に、slope の問題も学習した場合の結果について考察する．slope のみで学習した結果、各評価指標のスコアは著しく低かった．実際にばねが振動するという単一の内容が多く出力されており、接触に関して特に考慮できているとは考えられなかった．よって、ばねと斜面の運動のように、テストデータで扱う運動とは異なる運動について学習させただけでは十分な効果が期待できないと考えられる．一方、slope に加えてばねの横方向の運動の問題を含む spring-h-slope を学習した結果、縦方向・横方向のばねの運動の問題に対して共にスコアが向上した．特に、縦方向の問題に対して spring-h と spring-hv のスコアよりも向上している結果となった．斜面上を滑り落ちる木製ブロックの運動方向は縦方向であったため、横方向のばねの問題と合わせて学習することで、両方向に関する問題に対しても接触を上手く認識することができるようになったと考えられる．さらに縦方向のばねの運動の問題も含まれる spring-hv-slope を学習した結果、spring-h-slope の結果と比較して縦方向の問題に対するスコアが悪化したが、横方向の問題に対するスコアが向上し、両スコア間の差が縮まった結果となった．

最後に spring の問題のみに注目し、布の配置パターンを単純化した spring-mini を学習した結果について考察する．横方向の運動の問題に対するスコアが高くなり、縦方向のスコアとの差が開いた結果となった．ばねが布や紙に届かない問題と、ばねや布が 2 枚ある問題を除外したが、これらの問題はいずれも特に正しく接触を認識できた数が少ない問題であったため、今回のスコアが向上した可能性がある．しかし、縦方向、横方向に加えて、さらに 180 度、270 度時計回りに回転させた運動方向の問題について学習したにも関わらず、なぜ横方向と縦方向でスコアの差が開いたのか、さらに検証する余地がある．また、問題のパターン数を減らしたにも関わらず、接触するという内容の出力を多く生成し、接触しない問題と区別することが難しかったことについて、さらなる実験により検証したい．同時に、エポック数や学習データの内容、量などの学習方法についても改善を検討する．

4 おわりに

本研究では、様々なデータセットを用意して MLLM にファインチューニングを施して比較実験を行うことで、MLLM の物体の運動方向に関する情報の捨象する能力にどのような課題があるのかを示した．MLLM に学習させる物理現象の種類や物体の運動方向などの特徴によって、物体の運動方向を認識して物体の接触を考慮した説明を生成する能力が変化することを確認できた．本研究を受けて、実験手法の改善やさらなる学習データセットの構築により比較を進めて、MLLM が現実世界における多様な状況を認識して情報の捨象を適切に行えるよう、さらなる発見につながることを期待する．

参考文献

[1] 日本経済新聞. 生産年齢人口とは 日本は 5 割台後半、g7 最低, 2025. https://www.nikkei.com/article/DGXZQOUA145KM0U5A410C2000000/ 最終アクセス: 2026-01-01.

[2] 内閣府経済社会総合研究所. 2024 年度（令和 6 年度）国民経済計算年次推計（フロー編）, 2025. https://www.esri.cao.go.jp/jp/sna/data/data_list/kakuhou/files/2024/sankou/pdf/point_flow20251223.pdf 最終アクセス: 2026-01-01.
[3] パーソル総合研究所. 労働市場の未来推計 2030, 2019. https://rc.persol-group.co.jp/thinktank/spe/roudou2030/ 最終アクセス: 2026-01-01.
[4] 厚生労働省. 2022 年版 ものづくり白書 概要. Technical report, 2022. https://www.mhlw.go.jp/content/000944612.pdf 最終アクセス: 2026-01-01.
[5] 來村徳信, 西原稔人, 植田正彦, 池田満, 小堀聡, 角所収, 溝口理一郎. 故障オントロジーの考察に基づく故障診断方式 : 網羅的故障仮説生成. 人工知能, Vol. 14, No. 5, pp. 838–847, 1999.
[6] Hikaru Yokono and Tetsunari Inamura. A framework of recognizing physical situation in text description with physics simulation. In **2014 International Conference on Information Science, Electronics and Electrical Engineering**, Vol. 2, pp. 1090–1094, 2014.
[7] Akshay J Dave, Tat Nghia Nguyen, and Richard B Vilim. Integrating llms for explainable fault diagnosis in complex systems. **arXiv preprint arXiv:2402.06695**, 2024.
[8] Linus Kohl, Sarah Eschenbacher, Philipp Besinger, and Fazel Ansari. Large language model-based chatbot for improving human-centricity in maintenance planning and operations. In **PHM Society European Conference**, Vol. 8, pp. 12–12, 2024.
[9] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. **Nature Machine Intelligence**, Vol. 7, No. 1, pp. 96–106, 2025.
[10] Vahid Salehi. Fundamentals of physical ai. **Journal of Intelligent System of Systems Lifecycle Management**, Vol. 2, , 2025.
[11] Pascal Hitzler and Md Kamruzzaman Sarker. **Neuro-symbolic artificial intelligence: The state of the art**. 2022.
[12] 鈴木慎平, 吉岡真治. 大規模言語モデルの物理法則・原理の認識結果に基づく定性推論のためのモデル構築手法. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 3Xin227–3Xin227, 2024.
[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. **ICLR**, Vol. 1, No. 2, p. 3, 2022.

A　付録 (Appendix)

A.1　評価指標

本実験で使用した評価指標の定義は次のとおりである．

TP : True Positive（正例を正しく正例と予測）

FP : False Positive（負例を誤って正例と予測）

FN : False Negative（正例を誤って負例と予測）

TN : True Negative（負例を正しく負例と予測）

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

A.2　データセットの画像

本実験で使用した画像について示す．ばねの問題については図 2 のような状況を示す画像を用いた．上段の左から右に向かって順に、布が振動するばねと接触する、ばねの運動方向よりも手前側に配置されていて接触しない、ばねの運動方向よりも奥側に配置されていて接触しない、ばねの運動方向よりも手前側と奥側の 2 箇所に配置されていて接触しない、下段の左から右に向かって順に、ばねが届かない程遠くに配置されていて接触しない、時計回りに 90 度回転させて縦方向に振動するようにしたもの、ばねに取り付けられている物体を変更したもの、時計回りに 90 度回転させて縦方向に振動するようにしたものである．

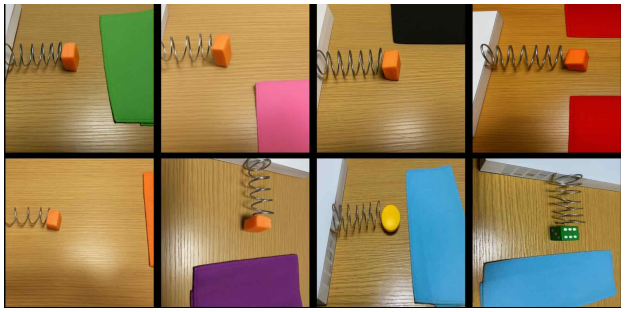

図 2: 使用した spring の画像

斜面の問題については図 3 のような状況を示す画像を用いた．左から順に、木製ブロックが斜面中央に配置されており滑り落ちる茶色の木製ブロックと接触する、(ii) 斜面右部に配置されており滑り落ちる茶色の木製ブロックと接触しない、(iii) 斜面左部に配置されており滑り落ちる茶色の木製ブロックと接触しない．

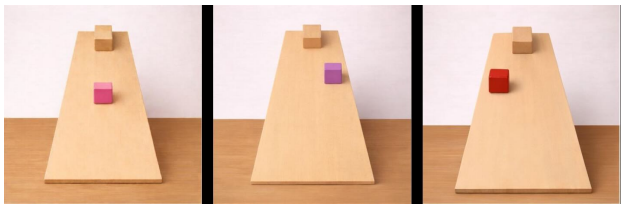

図 3: 使用した slope の画像

A.3　データセットの入出力文

spring のデータセットについて、入力文は次のようになっている．"What kind of motion will begin if you pull the spring shown in this image and release it?"

出力文は以下のようになっている．なお、布の色は画像ごとに対応させて文に記載してある．

- 接触する場合："Simple harmonic motion will begin. Since it will touch the blue cloth midway, friction is likely to occur and the motion will gradually be damped."
- ばねの運動方向線上に布や紙が無いため接触しない場合："Simple harmonic motion will begin. Although there is a yellow cloth placed nearby, it is not on the line of the spring’s motion, so it will not affect the spring’s motion."
- ばねが届かない程遠くに配置されていて接触しない場合："Simple harmonic motion will begin. Although there is a white cloth placed nearby, the spring will not reach the cloth, so it will not affect the spring’s motion.""

slope のデータセットについて、入力文は次のようになっている．"What kind of motion begins when you slide the brown wooden block placed at the top of this slope downward?"

出力文は以下のようになっている．なお、ブロックの色は画像ごとに対応させて文に記載してある．

- 接触する場合："The brown wooden block slides downward and then collides with the white block."
- 接触しない場合："The brown wooden block slides downward and then continues to fall smoothly without colliding with the white block."