視覚言語モデルの主観的評価に基づく仮想生物進化

宮﨑翔太[1] 有田隆也[1] 鈴木麗璽[1]
[1]名古屋大学　大学院情報学研究科
miyazaki.shouta.a4@s.mail.nagoya-u.ac.jp

概要

本研究では，遺伝的アルゴリズムの適応度評価・選択に視覚言語モデル（VLM）による主観的評価を導入する枠組みを提案する．仮想ソフトロボットを対象に，2 個体のシーケンス画像を VLM に提示し，「可愛らしい」「奇妙である」等の評価語に基づくペアワイズ比較により選択・進化させる．実験の結果，VLM による選択は評価語に応じた形態・動作を進化させた．また，VLM は評価語を内的基準に分解して判断し，その基準の類似性が進化結果の類似性を決定することが示唆された．本研究では，主観的言語表現が身体性を伴う表現型へ写像される過程を可視化し，VLM の主観的評価に基づく判断構造の理解をするための枠組みを提供しうることも示された．

1　はじめに

多くの工学的・社会的課題は，評価関数の最小化，最大化として定式化できるため，これまでに多様な最適化手法が提案されてきた．一方で，そのような定式化が困難な問題も存在する．例えば，人間の主観や感情に依存した評価に基づく最適化は個人差や文脈依存性が大きく，評価関数を定義し最適化することが困難である．

そのような主観に基づく評価関数を明示的に設計することを避ける枠組みとして，対話型進化計算（IEC）が提案されている．IEC は，進化計算における親個体の選択に人間を組み込み，利用者の主観的選好に基づく選択を通じて探索を進める枠組みである．Dawkins の Biomorph は，観察者の選好という主観的な選択圧が形態多様性を生み得ることを示し，主観を評価に基づく探索の可能性を示した[1]．一方で IEC は，多数世代にわたり多数個体を評価する必要があるため，評価者疲労が大きなボトルネックとなる．この疲労問題に対し，評価回数の削減[2]や，web を用いることによる評価入力の効率化・多人数化[3]などが検討されてきた．しかし，これらの改善をふまえても，人間による評価に要するコストが大きいという根本的な問題の解決には至っていない．

近年，大規模言語モデル（LLM）の発展に伴い，評価者を LLM で代替する LLM-as-a-judge が提案されている．LLM-as-a-judge は，人手評価が高コストでスケールしにくい状況に対して，文脈依存な基準に基づく比較評価を自動化し得る点で重要である．実際に，強力な LLM が人間選好と高い一致を示し得ることや，評価に伴うバイアスの存在とその検討が報告されている[4]．さらに，LLM と進化計算の接点も拡大している．LLM を突然変異のオペレーターとして導入した手法の提案[5]や，LLM を用いた進化的なプロンプトの最適化手法[6]が提案されている．

また，視覚言語モデル（VLM）を用いれば，形や動きといった視覚情報を含む対象に対しても，比較評価を自動化できる可能性がある．LLM が持つ認知能力における視覚と言語の結びつきの理解のため，VLM に画像を選択させる課題を用いて VLM が人間に観察される音象徴（kiki/bouba）に対応する傾向を示す結果[7]や，音象徴のみならず色彩感覚に関しても人間と感覚が共通する結果[8]が示されている．視覚と言語を結び付けた探索の応用では，基盤モデルを用いた人工生命モデル探索の自動化手法が提案され[9]，VLM を用いた最適化手法の提案とその理解が重要となっている．

本研究は，VLM による主観的評価を進化計算に組み込む枠組みを提案し，その枠組みによって生じる進化過程と表現型を分析することで，主観的な言語表現が具体的な表現にどのように結び付けられるかを明らかにすることを目的とする．具体的には，遺伝的アルゴリズムにおける適応度評価および親個体の選択に，VLM による主観的評価に基づくペアワイズ比較を導入し，その結果を用いて選択を行う．実験では，進化の対象として繊細な動き方が表現可能な柔軟な形態を持つ仮想生物（ソフトロボット）を採用する．その動きを，adorably（可愛らしく），weirdly（奇妙に）等の計５種類の主観性を含む評価語に基づき VLM が同時に提示される複数個体から

選択することで集団が進化する．主観的な言語表現の違いが，創発する仮想生物の形態・動作にどのように反映されたかを議論する．

2　手法

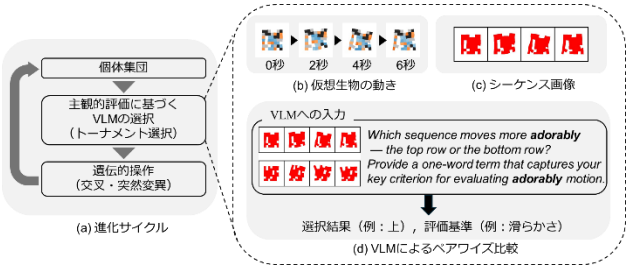

図 1　モデル概要

2.1　仮想生物

提案手法の概要を図 1 に示す．進化させる対象として，仮想ソフトロボットベンチマーク Evolution Gym[10]で実装した仮想生物を用い，仮想生物の形・動きを進化させる．進化の対象として仮想生物を用いる理由は，主観という抽象的な情報が，外形のみならず，その変形の過程といった時間的特徴としても表出し得る状況を扱うことができるためである．

仮想生物は$W \times W$のグリッドの中で voxel と呼ばれる正方形のブロックを組み合わせた身体を持つ（図 2a）. Voxel には静的なブロックである rigid（硬質）, soft（軟質）と, 動的なブロックである horizontal actuator（横伸縮），vertical actuator（縦伸縮）の計 4 種類があり（図 2b），actuator を伸縮させることで移動する．伸縮は各 voxel が時間ステップごとに辺の長さを変えることで行われ，その辺の長さ$L(t)$は以下で与えられる周期関数によって計算される．

$$L(t) = \frac{\sin\left(\frac{\pi t}{30} + \varphi\right) + 1}{2} + 0.6, \qquad (1)$$

ここで, tは仮想空間内の時間ステップ, φは$0 \le \varphi \le 2\pi$の範囲の値をとる位相である．仮想生物の遺伝子型は，voxel に対応した整数（empty: 0, rigid: 1, soft: 2, horizontal actuator: 3, vertical actuator: 4）と，voxel が actuator であった場合の伸縮タイミングを決定する位相に対応した小数からなる 1 次元のリストで定義される（図 2c）．

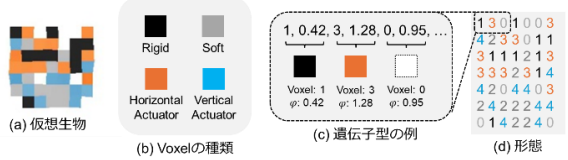

図 2　仮想生物

2.2　VLM による主観的評価に基づく選択

仮想生物の動きは VLM に同時に提示される複数個体のスナップショット画像に基づくペアワイズ比較によって評価される（図 1d）．この手法を考案した理由は 2 つある．一つは，仮想生物の動きのような抽象性の高い対象を主観的な基準で絶対評価するのは難しく, LLM に同時に提示し比較させることで繊細で一貫した評価が可能であるためである．もう一つは，進化計算への応用において問題となる繰り返し推論を行う計算コストの抑制である．動画を入力する代わりに複数個体のスナップショットをまとめた画像を入力することで，動きのような動的な特徴に対して画像一枚で評価が可能である．

具体的には，図 1b に示すように，決められた時点における仮想生物のスナップショット画像（voxel 固有の色が評価に及ぼす影響をなくすため，仮想生物の身体は赤色に着色）を 4 枚横に時系列順に並べ（図 1c）, それを比較する 2 個体分上下に並べた画像（以下，シーケンス画像）を作成する．これを VLM にプロンプト（付録 A 参照）とともに入力し，上下どちら側の個体が評価基準に即した動きをしているかを，理由となる評価の基準をつけて回答させる（図 1d）．プロンプトは，SLM と呼ばれるローカルな計算機環境でも実行可能な小規模な言語モデルを利用しており，そのようなモデルでも一貫した評価が可能なように工夫している．また，選択をより確実なものとするため，入力画像中の個体の上下を入れ替えて同様に評価を行い, 入れ替える前と逆の選択（つまり，同じ個体を選択）した場合のみ選択結果を採用する．

2.3　進化計算

仮想生物の進化は前述の選択手法を利用した遺伝的アルゴリズムによって行われる（図 1a）．個体数Nの集団を世代数がGに到達するまで進化を繰り返す．初期世代の個体はランダムに生成する．各世代において，集団からランダムに 2 個体を選択し，2.2 節で説明したペアワイズ比較を行い，勝者を親個体とする．この操作を，親個体の数が集団のサイズと同じになるまで繰り返す. 交叉は確率p_{cross}で適用し，交叉を行う場合には親個体間で一点交叉を行う．交叉を行う位置はランダムに決定する．突然変異は遺伝子座ごとに確率p_{mut}で適用し，変異が発生した遺伝子座は取りうる別の値へランダムに置換する．

Evolution Gym の仕様により，個体を構成する全 voxel は，少なくとも一辺が他の voxel と接している必要がある．交叉・突然変異によりこの条件を満たさない個体が生成された場合には，条件を満たすまで，遺伝子座をランダムに選択し，その値を別のランダムな値へ変化させる操作を繰り返す．

3　実験結果と考察

個体数$N = 30$，世代数$G = 50$，交叉確率$p_{cross} = 0.05$，突然変異確率$p_{mut} = 0.01$，仮想生物の最大縦横幅$W = 7$，時間ステップ$t = 600$で実験を行った．VLM には gemma-3-12b-it[11]の 4bit 量子化モデルを使用し，$Temperature = 0$に設定して実験を行った．推論には GEFORCE RTX 4080 SUPER を使用した．

進化に用いる主観的評価語として，それぞれ意味的な方向が異なる単語として adorably（可愛らしく），weirdly（奇妙に），solemnly（厳かに），dynamically（ダイナミックに），motionlessly（じっとして）の計５種類の条件で実験を行った．adorably，weirdly，solemnly は多様な解釈が可能な抽象度の高い表現として採用し，dynamically，motionlessly は運動の様態をより直接的に想起しうる表現として採用した．

3.1　VLM による選択の有効性

本研究で注目する主観的な言語表現を評価基準とした VLM による選択が，どの程度進化の方向や圧力をもたらすかは自明でない．そこで，まず集団内の多様性を定量化し，ランダムな選択で進化させた場合と比較した．具体的には，各世代で，各個体の遺伝子型と同世代のほか個体とのハミング距離を計算し，個体ごとの平均をとったうえで，その値をさらに平均して世代内の平均ハミング距離を計算した．

主観的評価語ごとの，各世代の平均ハミング距離の推移を図３に示す．点線は各試行の値，黒色の線は５試行の平均値である．５試行の平均値に注目すると，ランダムな選択によって進化させた場合には緩やかに一定の水準で減少し，最終世代では 33.9 まで減少した．これは，個体集団の収束が一定の速度で起きたことを示している．一方，他の条件ではランダムな選択の場合と比較して，初期世代から急激に減少し，最終世代におけるハミング距離はいずれの条件でも 20 未満であった．この結果は，VLM の主観的評価によって，ある特徴を持つ生物がより積極的に選択されたことを示している．

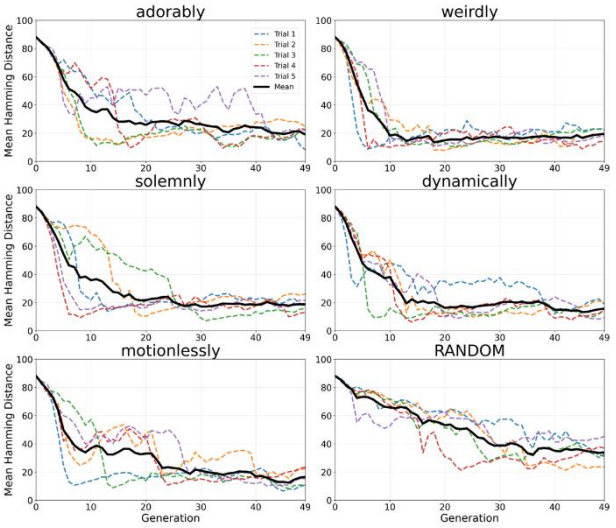

図 3　平均ハミング距離の推移

3.2　進化後の仮想生物

図４は，各試行における最終世代個体のうち，他個体の遺伝子型と自身の遺伝子型との平均ハミング距離が最小であった個体（すなわち，当該試行の最終世代集団において遺伝子型が最も集団中心に近い個体）のシーケンス画像を示したものである．また，図５は，全条件および全試行における各個体のシーケンス画像を CLIP[12]により高次元特徴ベクトルとして表現し，それらを UMAP[13]によって二次元空間へと射影したうちの，最終世代の個体のみをプロットしたものである．

Adorably 条件では，足状の構造を有し，歩くような個体（Trial 1, 2, 5）と，塊状で弾むように変形する個体（Trial 3, 4）か観察された．前者は小動物的な動作の可愛らしさ，後者は柔軟な質感の可愛らしさに対応すると解釈でき，試行ごとに異なる「可愛らしさ」を持つ生物が進化したことを示唆する．

Weirdly 条件では切れ目状の構造とその開閉･拡張に類する運動が観察された．身体が裂けるような不規則な変形が「奇妙さ」の印象を生んでいると解釈できる．Dynamically 条件は運動の様態を直接的に想起し得る表現として採用したが，身体を大きく広げる動きが weirdly 条件と共通して観察された．VLM が「ダイナミックさ」を動きの大きさや急激な変形として解釈した結果と考えられる．

[i] https://huggingface.co/lmstudio-community/gemma-3-12b-it-GGUF

Solemnly 条件では正方形に近い外形と，内部に空洞を有する構造が観察された．安定した形状と緩やかな変形が「厳かさ」の印象に対応すると解釈できる．Motionlessly 条件は運動の様態を直接的に想起し得る表現として採用したが，正方形に近い外形を示す個体が多く，solemnly 条件と類似した表現型へと収束した．UMAP 空間上でも両条件の個体が同一領域へ収束する試行（図 5 上部，Trial 2）が見られ，VLM が「静止」と「厳かさ」を類似した基準で評価したことを示唆する．

以上より，VLM による選択を通じて，各評価語の特徴を反映した形態・動作が進化することが確認された．ただし，抽象的な評価語の中でも，adoably のように試行間で異なる表現型へ分岐するものや，weirdly や solemnly のように進化の方向が比較的一貫するものが存在した．また，運動の様態を直接的に想起し得る表現として採用した dynamically と motionlessly は，それぞれ weirdly，solemnly と類似した表現型へと収束した．これらの結果は，VLM が評価語をそのまま適用するのではなく，独自の内的基準に分解して判断しており，その基準の類似性が進化結果の類似性を決定することを示唆する．

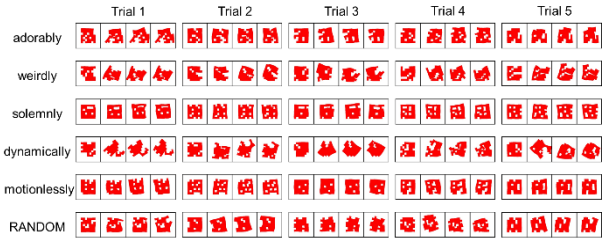

図 4　最終世代の仮想生物

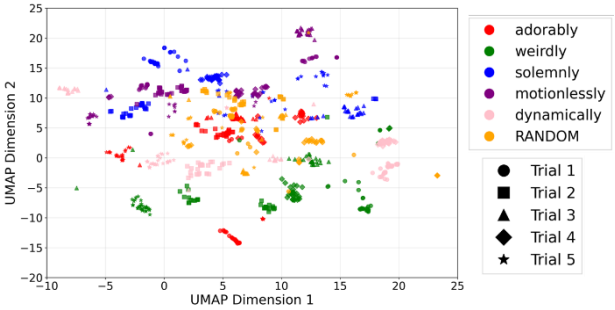

図 5　表現型空間における最終世代個体分布

3.3　評価の基準

VLM が主観的評価に基づいてペアワイズ比較を行う際の判断傾向を調べるため，全 5 試行における比較時の出力から，全世代・全個体の評価基準語を集計し，主観的評価ごとの出現割合を算出した．図 6 は Word Cloud により頻出語ほど単語を大きく可視化したものである．括弧内の数値は最頻語の出現割合を示す．

Adorably 条件では評価基準が fluidity に偏る傾向が見られた．Fluidity（滑らかさ）の解釈は文脈に依存しやすいため，結果として試行ごとに異なる方向へ進化したと考えられる．Weirdly 条件と dynamically 条件では unpredictability（予測不可能性）が共通して頻出した．大きな形状変化や不規則な変形を生みやすい構造と予測不可能性が結び付けられ，切れ目を持つ形態へと進化した可能性が考えられる．Solemnly 条件と motionlessly 条件では regularity（規則性）が共通して頻出した．これは，正方形状の安定した形態による動きと規則性が結び付けられ，選択圧に反映されたと考えられる．

以上より，進化した形態の類似性と，出現した評価の基準の頻度には共通する点が見られた．これは，前節で示唆された，VLM が判断の際に用いる内的基準が評価の基準として可視化されていることを示唆する結果である．

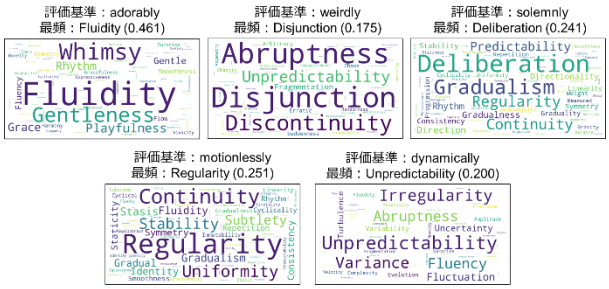

図 6　各主観的評価における評価基準

4　おわりに

本研究では，VLM による主観的評価を導入した進化計算の枠組みを提案した．この枠組みを用いた仮想生物進化実験の結果，主観的評価語が反映された形態・動きへと仮想生物が進化した．異なる主観的評価語で似た形態が進化したことから，VLM の主観的判断には独自の内部基準が存在することが示唆され，評価の基準がそれに類するものと考えられる．

本研究で提案した枠組みを用いることで，従来表現が難しかった主観的な評価を進化計算に盛り込むことが可能になり，また，これまでの人的なコスト問題によるボトルネックを解消し，その可能性を拡張することが期待される．

参考文献

[1] Richard Dawkins. *The Blind Watchmaker*, Longman, Essex, 1986.

[2] Xavier Llorà, Kumara Sastry, David E. Goldberg, Abhimanyu Gupta, and Lalitha Lakshmi. Combating user fatigue in iGAs: partial ordering, support vector machines, and synthetic fitness. *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, pp. 1363-1370, 2005.

[3] Jimmy Secretan, Nicholas Beato, David B. D Ambrosio, Adelein Rodriguez, Adam Campbell, and Kenneth O. Stanley. Picbreeder: evolving pictures collaboratively online. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1759-1768, 2008.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, Vol. 36, pp. 46595-46623, 2023.

[5] Lehman Joel, Gordon Jonathan, Jain Shawn, Ndousse Kamal, Yeh Cathy, and Stanley Kenneth O. Evolution through large models. *Handbook of Evolutionary Machine Learning*. Singapore: Springer Nature Singapore, pp. 331-366, 2023.

[6] Tong Zeliang, Ding Zhuojun, and Wei, Wei. Evoprompt: Evolving prompts for enhanced zero-shot named entity recognition with large language models. *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5136-5153, 2025.

[7] Morris Alper, and Hadar Averbuch-Elor. Kiki or Bouba? Sound Symbolism in Vision-and-Language Models. *Advances in Neural Information Processing Systems*, Vol. 36, pp. 78347-78359, 2023.

[8] Ryoma Shinto and Hiroyuki Iizuka. Analyzing the sensibility of visual language models using an evolving image generation system: focusing on color impressions and sound symbolism. *Proceedings of the 2024 Artificial Life Conference*, 36, 2024.

[9] Akarsh Kumar, Chris Lu, Louis Kirsch, Yujin Tang, Kenneth O. Stanley, Phillip Isola and David Ha. Automating the search for artificial life with foundation models. *arXiv preprint arXiv:2412.17799*, 2024.

[10] Jagdeep Singh Bhatia, Holly Jackson, Yunsheng Tian, Jie Xu, and Wojciech Matusik. Evolution Gym: a large-scale benchmark for evolving soft robots. *Advances in Neural Information Processing Systems*, Vol. 34, pp. 2201-2214, 2021.

[11] Team Gemma, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[12] 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. *Proceedings of the 38th International Conference on Machine Learning, in Proceedings of Machine Learning Research*, pp. 8748-8763, 2021.

[13] Leland McInnes, John Healy, and James Melville. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

A 実験用プロンプト

以下は adorably 条件において使用したプロンプトである．

```
# Motion Pattern Evaluation Task

You are an impartial motion analysis specialist who evaluates motion patterns based solely on their intrinsic qualities.
You observe subtle dynamics, rhythm, and temporal flow in object transformations without bias toward spatial positioning.

## Image Description

In this image, there are two frame sequences (top and bottom rows), each containing 4 chronological frames (time steps: t, t+1, t+2, t+3) of tracked red objects from left to right:

- The red object is located in the center of each frame.
- The red object remains completely whole and intact at all times, never splitting or separating into multiple parts, regardless of the passage of time.
- The red object does not rotate.
- Each frame is enclosed by a black border.
- Focus only on how the object's shape changes over time, while maintaining constant size.

## Step-by-Step Analysis Process

### Objective Frame Description

- Top row:
  - Describe the shape in each frame (t, t+1, t+2, t+3).
  - Focus on outline, structure, and distinctive features.

- Bottom row:
  - Describe the shape in each frame (t, t+1, t+2, t+3).
  - Focus on outline, structure, and distinctive features.

### Change Pattern Analysis

- Top row:
  - Describe how the shape transforms between frames.

- Bottom row:
  - Describe how the shape transforms between frames.

### Quality Assessment for “adorably”

1. Define what “adorably” means in the context of motion/transformation.
2. Identify relevant characteristics from your observations that relate to this quality.
3. Evaluate which sequence better embodies these characteristics.

### Question

Which sequence moves more adorably — the top row or the bottom row?
Support your answer by explaining:
- (1) your definition of “adorably” in this context
- (2) which observed characteristics are relevant
- (3) how each sequence demonstrates these characteristics
Key Criterion: Provide a one-word term that captures your key criterion for evaluating “adorably” motion.

At the end of your response, answer with your choice (“top” or “bottom”) enclosed in quotation marks.
```