第 29 回日本バーチャルリアリティ学会大会論文集（2024 年 9 月）

リアルタイム画像生成 AI により作成したアバタを用いた身体ゲームの提案

村留一舞，脇坂崇平，Mark Armstrong，南澤孝太
Kazuma MURATOME，Sohei WAKISAKA，Mark ARMSTRONG and Kouta MINAMIZAWA

慶應義塾大学大学院メディアデザイン研究科
（〒 223-8526 神奈川県横浜市港北区日吉 4-1-1，{muratome.kaz888，wakisaka，mark, kouta}@kmd.keio.ac.jp）

概要: 生成 AI の発展により，ユーザ周囲の現実環境をリアルタイムに視覚変換する現実変容体験への可能性が開けつつある．だがそういった体験においては，文脈に沿わないイメージの出現が本質的に不可避であり，それは本来的にはエラーとして扱われてきた．本研究ではその様なイメージをエラーではなく，新たな表象の創発として捉え，現実変容を同時に体験するユーザ間のコミュニケーションにおいてどの様に記号使用されていくかを調べる．

キーワード： 生成 AI，エラー，アバタ，コミュニケーション

1. はじめに

2000 年代初頭からスタートした深層学習 AI のブームは，text-to-text 生成 AI である大規模言語モデル (LLM)[1] や text-to-image 生成 AI である潜在拡散モデル (LDM)[2] の急速な発展を経て，現在日常的なツールとしていたるところに浸透をし始めており，その傾向は今後も続くことが容易に予想される．今では，LLM や LDM を筆頭に，多様な AI アルゴリズムがシームレスに統合され，いわば any-to-any 生成の様相を呈してきた．

そのような生成 AI の普及のなかで，生成 AI が起こす「エラー」をどう取り扱うか，というきわめてクリティカルな課題に，我々は直面している．LLM においては，事実でないものを事実として出力する事象を Hallucination と呼ぶ（存在しない文献情報を出力する，など）．また LDM においても，文脈やユーザの意図から乖離した生成画像は，修正すべき「エラー」として，扱われることが多い．よくある例として，目の欠如，7 本の指，物理的にはありえない構造を有す，などがある．上述したように，any-to-any 生成モデルにおいては，Hallucination やエラーが複雑に立ち現れる．Hallucination を減少させる様々な効果的な手法が存在するが [3]，これらは本質的に消去できるものではない．例えば，ある文化では特定の色やシンボルが吉兆とされる一方で，別の文化ではそれが悪い兆しと見なされるであったり，“ 骨を折る ” や “ 手のひらを返す ” のように特定言語特有の言い回しであるため誤解を生んだりする，などの事態は常に潜在してるのである．

2. 生成 AI により変容する現実

観測情報から環境を予測生成する世界モデル [4] や，さらには VR 技術と組み合わせることにより，生成 AI 研究は「現実生成 AI」の実現という局面に足を踏み入れつつある．そのような，生成 AI が現実体験の基盤，あるいは主要なファクターとなる状況においては，生成 AI におけるエラー（あるいはイレギュラー）と非エラーの境界はますます曖昧となっていくだろう．そしてそれはもはや排除する対象ではなく，「現実で生じる現象」としての性質を獲得し，ユーザは現象とインタラクションするように，エラーとインタラクションするようになるのではないだろうか（エラーの意味論的転回）．生成 AI，現実環境，ユーザ（の行為）が独立ではなく相互に影響を与えあう三項関係の中で，従来はエラーとみなされていた事物がどのような記号性をもち，コミュニケーションの中で扱われていくのか—これが本研究の射程となる．

3. 提案手法およびシステム構成

ユーザと現実環境を生成ＡＩが媒体する状況を考えよう．ここでは，ユーザの姿勢・ジェスチャにより，ユーザの容姿や現実の風景が多様に変容する状況を取り扱う．そのような状況においては，どのようなコミュニケーションが創発するだろうか．上記の問いをヒューリスティックに探るためにまず，以下の仕組みを作成した (図 1)．

- カメラで取得するユーザの容姿を，リアルタイムに意味レベルで多様に変化させる．
- 生成される内容は，ユーザの姿勢・ジェスチャにより確率的に決定される．
- 結果はディスプレイに表示され，ユーザはディスプレイを介して自身の容姿変容を観察する．
- 多くの場合，人の容姿へと変容するが，姿勢・ジェスチャを工夫することにより，人以外の生き物や，その他幻覚的要素の強いイメージへの変容も生じる．

LDM に画像を入力することにより，その画像をベースとし，prompt に設定されたテキスト内容を反映させた画像を出力するよう設定する．本研究では，もっとも普及している LDM の 1 つである Stable Diffusion の高速版 StreamDiffusion[5] を使用した．StreamDiffusion では，1 フレームごとに，prompt に即した画像変換が適用される．設定によっては，例えば 100fps，512x512pixel での動作も可能である．詳細は省くが，画像変換の結果は，学習モデル（SD model），LoRA，denoising のステップ数などのパラメータに強く依拠する．StreamDiffusion をビジュアルプログラミング環境 TouchDesigner(Derivative) で稼働させるツールキット[1]を使用し，PC に接続したカメラに映るユーザの容姿をリアルタイムに（意味レベルで）変容し，ディスプレイに映し出す図 1 に示すようなシステムを作成した．例を図 2 に示す．

4. 身体ジェスチャゲームワークショップ

上述のシステムを稼働させつつ，参加者 6 人を対象としたワークショップを実施した．テーマは,「見た目のイレギュラーな変容に注目したインタラクティブなゲームの提案」とした．その結果，身体ジェスチャを用いたしりとりや伝言ゲームなどのゲーム形式のコミュニケーションフォーマットが提案された．これら体験はディスプレイ上でも体験できるが，MR 空間での体験も可能であると考えられる．以下にそのうちの 1 つ「変身レースゲーム」を示す．

4.1 変身レースゲーム

姿勢とジェスチャを駆使し，体験者 2 人（P1 と P2）のどちらが早く，目的イメージの生成を行えるかを競ってもらう．勝負は 3 回，制限時間は 1 テーマにつき 1 分とする．下記に示すテーマ (A)(B)(C) における目的イメージの例と，実際に体験してもらった際の体験の様子をそれぞれ図 3，図 4 に示す．

(A) 顔のみ狼に変容させる．

(B) 全身を狼に変容させる．

(C) 3 面狼に変容させる．

prompt は以下のとおり:
((ocean wave:1.5)), solo man, (tree:1.2),
(green:1.2), 20years old, man, glasses, wolf perm

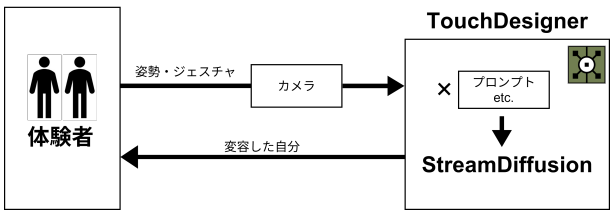

図 1: システム構成図

[1]https://www.patreon.com/dotsimulate/

図 2: ディスプレイに映し出される変容した容姿（SD model: stable-diffusion-v1-5, prompt: 20years old, woman, pink hair．体験者はカメラの前で腕を上げている．腕の位置を調整すると，時折腕が人と解釈され，3 人の人物が出現することがある．）

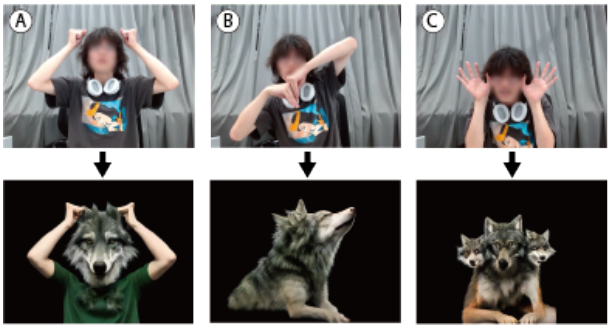

図 3: 目的イメージの例．(上) 体験者の様子 (下) ディスプレイ上に映し出される容姿

prompt の設定により，ディスプレイにはウルフパーマの 20 代男性が表示されやすくなっている．一方，ウルフパーマには wolf という単語が含まれているため，入力画像によっては狼が生成される．各種パラメータについては，ソフトウェアおよびシステム依存のため割愛するが，図 2 のような画像変換がなされるよう調整を行った．画像更新速度は約 15fps となった．

図 4: ゲーム体験の様子．(上) 実世界の体験者ペアの様子 (下) ディスプレイ上に映し出される容姿．どちらも目的イメージに到達できない場合は引き分けとなる．C は正解は 3 面狼だが出力は 3 体の狼となっている．

4.2 体験者からのフィードバック

10 名が変身レースゲームに参加した．上述の prompt と目的イメージによるゲームに加えて，他のバリエーションもいくつか試した．きわめてゲーム性，娯楽性が高いことが確認され，自身の形状が変化していくことに徐々になれていく様子も観察された．また，特定のジェスチャでないと目的イメージを実現できないという制約があり，実世界の本体がアバタが存在している CG 世界のルールに合わせに行っているというフィードバックがあった．これにより，流動的に変化し，かつタスク遂行に制約があった際，アバタに所有を感じるのではなく，アバタに合わせにいっている感覚になることが確認できた．

現在の設定では，複数の体験者はリアルタイムでイメージが移り変わることへの対応の難しさを感じていた．また，映像が約 15fps で瞬時に変容し小さな動きだけでもイメージが変わるため,「想定していたイメージを維持するのが難しい」「トラッキングがずれてしまった際の調整が難かしい」などの意見がみられた．

5. 考察と展望

生成 AI が環境を変容させる状況で，どのようなコミュニケーションが出現するか，という問いについてワークショップ等を通して検討を行った．ゲーム形式で，自身の動きに合わせて劇的に変化していく自己の姿を体験しつづけると，そのような変化になれる（自身の容姿は変化するものであり，想定外の容姿になっても，違和感を感じなくなっていく）可能性を示唆している．

今回は背景から抜き出した参加者の容姿を入力としたが，背景も入力とする場合や，その他の情報（体験者のパーソナリティ，趣向，生体生理情報など）も入力として取り扱うことも当然可能である．画像生成 AI のシステムデザインは無数にあり，今回はシンプルなデザインを第一歩としてまず試した段階ではあるが，今後自己容姿の受容の仕組み，変化願望など，多様な問題系に新しいアプローチを与えてくれることを期待している．

今回作成した体験システムでは，1 フレームごとに独立に処理をしている．時間的な映像の連続性を維持するアルゴリズムを入れていないため，ジェスチャの少しの変化により出力が大きく変わりえる．この点に難しさを感じるユーザーもいたが，連続性を維持するアルゴリズム（例えば [6]）を入れると，出力の劇的な変化は抑えられ，エラー的な挙動も現象する．今後，時間的連続性と挙動の多様性のトレードオフと体験内容の関係について，より詳しく調査していく予定である．なお現在没入型 VR 環境版を構築中だが，その場合 1 フレームごとに視覚情報が大きく変化してしまうと知覚認知的負荷が高すぎると考えられる．そのため VR 版については，時間的な映像の連続性の維持を必要条件として設定している．

変身レースゲームでは,「まれに出現するもの，あらかじめ決められた容姿」を目的に設定した．その意味では，エラー性は低いといえる．本研究で注目しているエラーの意味論的転回という観点においては，より強いエラーのリアルタイム創発，といったものを取り扱う必要があるだろう．それは今後の課題とする．

謝辞 本研究は JST ムーンショット型研究開発 Cybernetic being プロジェクト (JPMJMS2013) の支援を受けて行われた．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840–6851. Curran Associates, Inc., 2020.

[3] Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation, 2023.

[4] David Ha and Jürgen Schmidhuber. World models. 2018.

[5] Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, and Kurt Keutzer. Streamdiffusion: A pipeline-level solution for real-time interactive generation, 2023.

[6] Zhongjie Duan, Chengyu Wang, Cen Chen, Weining Qian, Jun Huang, and Mingyi Jin. FastBlend: a powerful Model-Free toolkit making video stylization easier. November 2023.