画像生成タスクにおけるエージェント性の調査

松岡 竜輝[1,a)]　今井 倫太[1,b)]

概要：画像生成モデルを用いた画像生成インタラクションでは，ユーザは，自然言語を用いてモデルに画像生成の命令文 (プロンプト) を入力することで画像を生成することができる．このとき，ユーザは画像生成モデルに入力するプロンプトの作成に苦慮し，生成画像と意図のギャップに苦しむことが多い．このような脳内イメージの言語化プロセスにおけるギャップの問題の原因として，人がモデルに対して，主観的な言葉は理解できないという意識を持っているということがある．本研究では，この問題を主観表現理解の期待欠如として，エージェントの存在による問題解決の可否について調査した．

1. はじめに

人のコミュニケーションには，指示を伝える「指示型インタラクション」と，お互いに相手の心の状態を推察し合う「共感型インタラクション」がある．従来の AI エージェントは道具として設計されており，人が客観的な指示を与えるだけの指示型インタラクションにとどまっており，stable diffusion[9] をはじめとする画像生成インタラクションも指示型インタラクションの一種である．

画像生成インタラクションは，ユーザが自身の意図を自然言語でモデルに伝えることで画像を生成するプロセスである．このプロセスでは，ユーザが脳内のイメージを言語化する際に，モデルが主観的な表現を理解できないという認識がギャップを引き起こす要因となっている．特に，生成画像が意図に合致しない場合，ユーザはプロンプトの作成や調整に多くの時間と労力を要することが課題である．

従来の画像生成モデルは道具としての役割に特化しており，ユーザが明確で客観的な指示を与えることを前提としている．しかし，画像生成タスクはしばしば共創的な性質を持ち，ユーザが曖昧かつ感覚的な表現を用いることが多い．このような場面では，エージェントが単なる指示の受け手ではなく，ユーザの意図を推察し，補完する役割を果たすことが求められる．

このようなエージェントの実現には，「主観表現理解の期待欠如を引き起こすエージェント性の欠如」という課題を解決する必要がある．これは，ユーザがエージェントに対してテキスト入力を行う際に，エージェントが感情を理解できないという見方を持つことによって，指示型インタラクションにみられるような客観的な指示文の作成を目指してしまうヒューマンファクタのことを指す．

従来の LLM（大規模言語モデル）を利用したプロンプト補助では，表面的にはユーザの意図を反映しているように見えるが，実際にはモデルはユーザの内部状態や主観的な意図を深く理解していない．このような状況では，ユーザがエージェントに対して「自分の感情を理解することはできない」という印象を抱きやすくなる．

Rostami らの研究 [12] によれば，感情的に知能を持つチャットボットはユーザに共感されている感覚を提供する一方で，その信頼性や感情理解の限界が，ユーザにおける期待を阻害する要因となることが指摘されている．また，Svikhnushina らの研究 [6] によれば，人はエージェントに対して感情の理解を求めているものの，人が求めるものには至っていないとの結果が出ている．一方で，Schaaff らの研究 [7] では，GPT3.5 は自閉症患者よりも人の感情認識能力に長けていることがわかり，エージェントが感情認識に対して，ある程度の能力をすでに有していることがわかっている．

一方で，坂本らの研究 [8] によると，エージェントの見た目の問題は重要であり，エージェントが人に近い見た目をしているほど，人がエージェントを人らしいと判断する．また，小室らの研究 [11] によれば，ユーザが対話ロボットに対して適切な態度を取るためには，対話ロボットが完全ではないことを理解し，適切に対応する必要があると指摘されている．

これらの研究を踏まえると，エージェントに対するユーザの態度や期待は，そのエージェントが感情を理解し，人間らしい振る舞いをするかどうかに大きく依存していると考えられる．本研究では，画像生成インタラクションに

[1] 慶應義塾大学
[a)] matsuoka@ailab.ics.keio.ac.jp
[b)] michita@ailab.ics.keio.ac.jp

おけるエージェントによるユーザのシステム認識変化を，ユーザの入力文を分析することによって調査することを目的とする．

2. 関連研究

本研究の関連研究として，画像生成プロンプトの修正補助の研究及び，人同士のデザイン相互作用について述べる．

2.1 画像生成プロンプトの修正補助の研究

画像生成プロンプトの修正補助の研究としては，PromptCharm[13] や PromptPaint[2] のようにビジュアルイメージによる画像修正を混ぜ合わせる手法や，Promptify[3] のように複数画像を提示する方法，stable walk[5] にように，stable diffusion の画像空間を視覚的に表示することでプロンプティングを支援する手法といった研究があるが，これらの研究はユーザのプロンプト修正を支援することが目的である．また，Wen et al. のように，ハードプロンプトを学習させる [4] があるが，こちらもプロンプト学習を行うことで効果的なプロンプトを生成するアプローチである．これらの研究では，UI の設計を行うことによって，指示型インタラクションの質の向上を目指している．

2.2 人同士のデザイン相互作用

人間同士のデザインプロセスでは，反復的な改良が創造的なワークフローの自然な一部を成している．Mace ら [1] は，この創造的プロセスを 4 つの主要なフェーズに分けている．フェーズ 1 では作品の概念化が行われ，フェーズ 2 ではアイデアの展開に焦点が当てられる．フェーズ 3 では制作と改良が行われ，フェーズ 4 で作品の完成に至る．フェーズ 2 とフェーズ 3，さらにフェーズ 3 とフェーズ 4 の移行においては，視覚的な質や意図されたコンセプトへの適合性を多次元的に評価し，アーティストが自身の作品を反復的に向上させるプロセスが含まれている．本研究のアプローチは，この人間的な反復改良プロセスを模倣することを目指し，主観的および客観的なフィードバックを生成プロセスに組み込むことで，AI が人間の知覚や創造性を反映した形で出力を改善できるようにするものである．

2.3 仮説

本研究では，画像生成時のエージェント性の有無とユーザ入力文の質に着目し，以下の仮説を立てた．

- エージェントがいると，主観的な文章が発生する．
- 主観的な文章を入れると，ユーザの意図と生成画像ギャップが少なくなる．

3. 設計

3.1 システム UI

本研究では，図 1 のようなシステム Unity を用いて設計した．このシステムでは，左上にエージェントとして，響 [10] の Live2D モデルを掲示し，すぐ下のボックス（セリフ掲示部）に，セリフが掲示される．真ん中の列は，上部に生成した画像を掲示するスペースがあり，下部はユーザの文章入力を行うスペースである．右側には，ユーザの入力履歴とエージェントが生成した画像がチャット形式で出力される．

3.2 内部システム

内部システムは，松岡ら [14] のシステムをベースに，人の入力を文脈情報として保持し，文脈情報を人の入力として扱うようにした対話システムである．松岡らのシステムは，ユーザの入力を受けとると，画像を生成し，その画像に対して主観的な画像記述を付与する．付与した記述とユーザの入力を比較することで，主観情報を陽に扱うことができる．本研究では，ユーザの入力を単一のものではなく，会話履歴を要約して与える様にすることにより，ユーザのコンテキスト情報を持ったインタラクションを可能にしている．

3.3 インタラクションの流れ

システムが起動すると，エージェントは「ようこそ！ 画像生成デモへ 作ってほしい画像について、私に説明してね。」と発話する．なお，この発話は 3.1 で述べた UI のセリフ掲示部に掲示される．ユーザが画像説明文を入力すると，エージェントが「了解！　今から作るからちょっと待っててね」と発話し，内部システムの処理が開始される．内部システムでは，図 2 の「テキストの意味類似判定部」において，「NO」と出力された場合，すなわち出力画像がユーザ入力を満たさないと判定された場合には，エージェントは，「うーん... ちょっと違うからもう一回作り直すね。」と発話し，生成した画像を提示した上でもう一度画像生成を行う．一方，「YES」と判定された場合には，「画像が生成されました！」とエージェントが発話し，最終出力画像が提示される．もし，最終出力画像に対して，ユーザが満足しなかった場合には，再度入力欄に画像説明を入力することで画像生成を再度行う．

4. ケーススタディ

4.1 ケーススタディの手法

ケーススタディでは，20 代の男女 3 人に図 3 に示した画像を提示し，以下のように指示をした．

> この画像に似た画像をシステムに作ってもらいます．システムの指示に従って，画像を作成してもらってください．
> システムのピンク色の欄に，文章の入力を行い．
> 終わったら送信ボタンを押してください．
> 生成された画像が思ったものと違った場合，違う

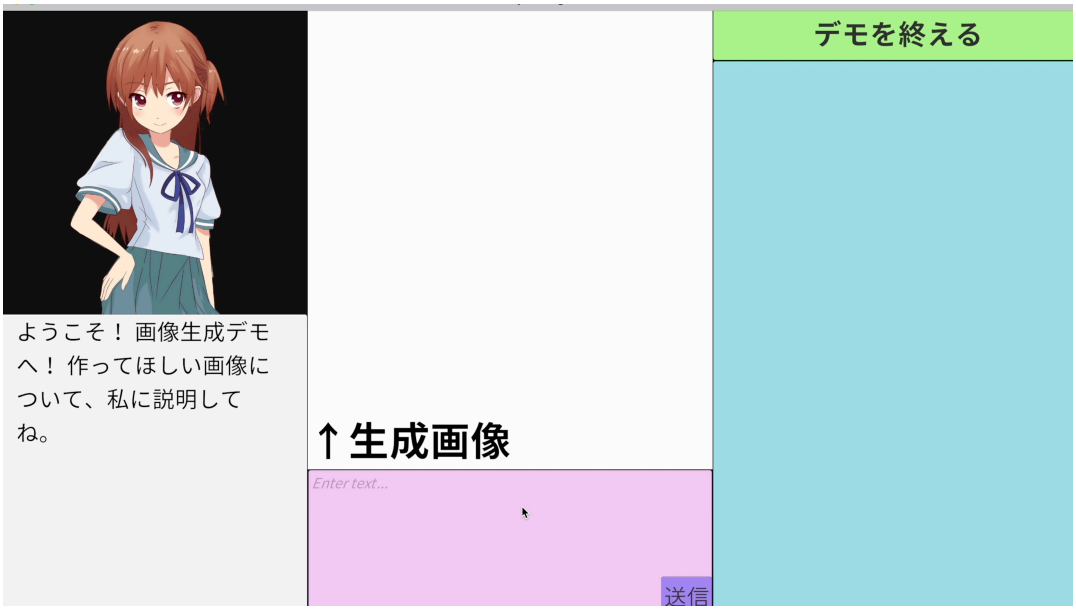

図 1 システム UI

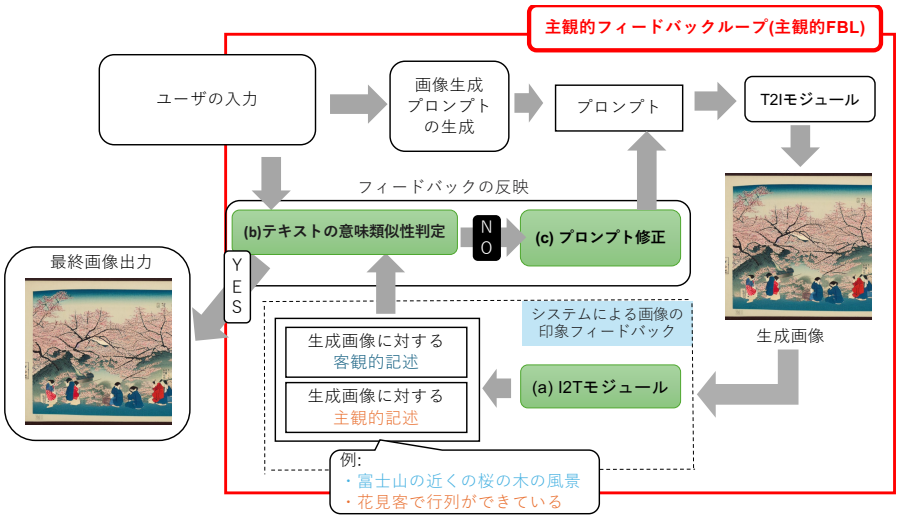

図 2 内部システムのアーキテクチャ

点を指摘してください．

4〜5 回入力を行っても満足のいく画像が出なかった場合，その時点で終了してください．

ユーザはこの指示に従い画像タスクに取り組んだ．その後，できた画像に対する以下のアンケート項目に回答した．

(1) システムに説明文を入力するときに苦労した点はどんなところでしたか？

(2) システムの出力とあなたの意図にどれくらいギャップがありましたか？

(3) システムにフラストレーションが溜まることはありましたか．具体的に教えてください．

4.2 ケーススタディの結果

ケーススタディでは，ユーザは全員が 4 回で入力を終えた．本研究では，ユーザの入力文を分析することで，画像生成インタラクションにおけるエージェントの課題と可能性を明らかにした．以下に主な考察を示す．以下では，3 人のユーザについて，User1，User2，User3 とし，各人の記述について分析を行った．

4.2.1 主観的な表現の利用状況

ユーザの入力文には，感覚的や抽象的な表現が多く含まれていた．例えば，「神秘的な雰囲気」「小学生が描いたような絵」など，主観的な評価が頻繁に用いられた．特に，User2 の入力では，芸術的なスタイルや感覚的なタッチへの言及が多く，主観的なニュアンスが重要視されているこ

図 3 ケーススタディで提示した画像

とが示唆された．

一方で，User3 の入力は簡潔で具体的な内容が多いものの，「抽象的」といった主観的な要素も含まれていた．これにより，主観的な表現はユーザごとに程度の差があるものの，画像生成インタラクションにおいて重要な役割を果たしていることが分かった．

4.2.2 アンケート結果からの課題

アンケート結果から，以下の課題が明らかになった．

- 語彙力不足による表現の難しさ: ユーザは自身のイメージを言語化する際に，適切な語彙を見つけることに苦労していた．
- システムとの意図のギャップ: 特に「タッチ」や「雰囲気」といった感覚的要素の伝達が難しく，結果として生成物が期待に合致しないことがあった．
- フラストレーションの原因: システムが以前の入力内容を忘れることや，意図した変更が反映されないことが，不満の原因となっていた．

4.2.3 仮説検証

本研究では，2.3 に記述したように，以下の仮説について検証を行った．

(1) エージェントがいると，主観的な文章が発生する．
(2) 主観的な文章を入れると，ユーザの意図と生成画像のギャップが少なくなる．

仮説 1 については，エージェントがユーザの意図を解釈しようとする際に，ユーザが感覚的なニュアンスを伝えるため，主観的な表現が増加する傾向が観察された．特に，芸術的なスタイルや雰囲気に関する表現が多く使われていたことが，エージェントの存在による主観的表現の促進を示唆している．

仮説 2 については，主観的な表現を含むプロンプトは，エージェントがユーザの意図をより正確に推察し，生成画像とのギャップを減少させることに寄与する可能性が示唆された．一方で，主観的な表現の曖昧さが，エージェントによる解釈にばらつきをもたらすケースも見られた．このため，主観的表現を補完する形で，具体的な情報を追加することが重要であると考えられる．実際，User2 ははじめは主観的な発言が多かったものの，徐々に客観的な発言が増えていったことから，客観的記述によるディティールの説明も重要であることがわかる．

5. まとめと展望

本研究では，画像生成インタラクションにおけるユーザの主観的な表現と，それに対応するエージェントの課題を明らかにした．ユーザの入力文分析およびアンケート結果から，主観的な意図の伝達が画像生成において重要であり，それに伴う課題として，語彙力の不足，意図とのギャップ，フラストレーションが確認された．

これらの課題を解決するためには，主観的表現を適切に解釈できる自然言語処理技術や文脈を考慮した一貫性のある応答生成が必要である．特に，エージェントが忘却することによるユーザのストレスが強いため，忘却しないことの重要性が示された．将来的な研究として，エージェントがある場合とない場合での比較実験を行い，その際のシステムの印象評価における有意差をとることにより，エージェントの有無によるユーザの印象の変化を調査するほか，形態素分析や言語ベクトル分析により，文章の散らばり具合の変化などの入力文評価について，被験者を増やして調査を行う．

謝辞 —
本研究は，JST，CREST，JPMJCR19A1 の支援を受けたものである．
—

参考文献

[1] Mace, Mary-Anne and Ward, Tony:Modeling the creative process: A grounded theory analysis of creativity in the domain of art making,Creativity research journal, 14, 2, 179–192, Taylor & Francis, (2002).
[2] Chung, John Joon Young and Adar, Eytan:Promptpaint: Steering text-to-image generation through paint medium-like interactions, Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–17,(2023).
[3] Brade, Stephen and Wang, Bryan and Sousa, Mauricio and Oore, Sageev and Grossman, Tovi: Promptify: Text-to-image generation through interactive prompt exploration with large language models,Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology,1–14, (2023).
[4] Wen, Yuxin and Jain, Neel and Kirchenbauer, John and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom: Hard prompts made easy: Gradient-based discrete

optimization for prompt tuning and discovery,Advances in Neural Information Processing Systems, 36, (2024).
[5] Mattias Rost and Sebastian Andreasson:Stable Walk: An interactive environment for exploring Stable Diffusion outputs 89-97, Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC,SHAI, SketchRec, SOCIALIZE co-located with the ACM International Conference on Intelligent User Interfaces (IUI 2023), Sydney, Australia, March 27-31, 2023,CEUR Workshop Proceedings,3359,pp.89–97,CEUR-WS.org, (2023), 入手先 ⟨https://ceur-ws.org/Vol-3359/paper10.pdf⟩.
[6] Ekaterina Svikhnushina and Pearl Pu: Social and Emotional Etiquette of Chatbots: A Qualitative Approach to Understanding User Needs and Expectations, (2020), 入手先 ⟨https://arxiv.org/abs/2006.13883⟩,
[7] Kristina Schaaff and Caroline Reinig and Tim Schlippe: Exploring ChatGPT's Empathic Abilities, (2023), 入手先 ⟨https://arxiv.org/abs/2308.03527⟩,
[8] 坂本大介，神田崇行，小野哲雄，石黒浩，萩田紀博：遠隔存在感メディアとしてのアンドロイド・ロボットの可能性，情報処理学会論文誌，Vol.48, No.12, pp.3729-3738(2007).
[9] Rombach, Robin and Blattmann, Andreas and Lorenz, Dominik and Esser, Patrick and Ommer, Björn: High-resolution image synthesis with latent diffusion models, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684–10695, (2022)
[10] Live2D：Live2D サンプルデータ集（無料配布），入手先 ⟨https://www.live2d.com/learn/sample/⟩ (2024.12.22).
[11] 小室 允人, 船越 孝太郎: 相互行為実践としての対話ロボットに対する態度, 人工知能学会論文誌, 2022, 37 巻, 1 号, p. A-L61_1-15, Online ISSN 1346-8030, Print ISSN 1346-0714,(2022/01/01) 入手先 ⟨https://doi.org/10.1527/tjsai.37-1_A-L61⟩.
[12] Rostami, M. and Navabinejad, S. 2023. Artificial Empathy: User Experiences with Emotionally Intelligent Chatbots. AI and Tech in Behavioral and Social Sciences. 1, 3 (Jul. 2023), 19–27. DOI:https://doi.org/10.61838/kman.aitech.1.3.4. (2011.09.15).
[13] Wang, Zhijie and Huang, Yuheng and Song, Da and Ma, Lei and Zhang, Tianyi:PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement,Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–21, (2024).
[14] 松岡竜輝, 熊野史朗, 今井倫太, 成松宏美: 意味的類似性にもとづく主観的印象テキストからの画像生成. 人工知能学会 言語・音声理解と対話処理研究会（SLUD）第 102 回研究会 (第 15 回対話システムシンポジウム), p.223-228, 2024.