画像情報と画像表現が VLM エージェントのユーザー行動模倣に及ぼす影響の分析

畑中希栞　宮里龍平　岡本一志　軽部幸起　柴田淳司　原田慧
電気通信大学
k.hatanaka@uec.ac.jp

概要

大規模言語モデル（LLM）/ 視覚言語モデル（VLM）エージェントは，人間の行動を再現できる可能性が示されており，推薦システムの分野でも，A/B テストのようなオンライン評価のユーザーシミュレーションへの活用が進められている．しかし，エージェントがユーザーの行動を模倣する精度について，画像を考慮した検証は十分に行われていない．本研究では，推薦アイテムの提示情報に画像情報，ペルソナに画像表現（base64 エンコード）を追加し，実ユーザーとの 1 対 1 対応による模倣精度を評価した．その結果，視聴履歴をペルソナとして与えた条件では，画像情報や画像表現が模倣精度向上に寄与することが示された．

1　はじめに

近年，記憶検索や内省，計画などの機構を備えた大規模言語モデル（LLM）エージェントに人間の行動を再現させる研究が様々な分野で行われている [1, 2, 3, 4, 5]．これらのエージェントは，人間の属性や嗜好を表すペルソナをプロンプトで与えることで，個人に対応した行動模倣を可能としている．

推薦システムの分野では，A/B テストのようなユーザー反応を用いて性能を評価するオンライン評価のシミュレーションに，その活用が進められている [2, 4]．オンライン評価は，推薦システムの性能を最も直接的に測定できる評価方法であるが，評価環境構築やデータ収集に要する時間やコストが大きく [6]，容易に実施することが難しい [7]．Zhang [2] らは実ユーザーの行動をエージェントが模倣することでオンライン評価を再現する手法 Agent4Rec を提案し，これらの課題にアプローチしている．

本研究では，画像を扱える視覚言語モデル（VLM）をエージェントに用いて，推薦システムにおけるユーザーシミュレーションに画像情報を加えた場合に，エージェントが実ユーザーの行動をより忠実に再現できるかを検証することを目的とする．以下にリサーチクエスチョンを示す．

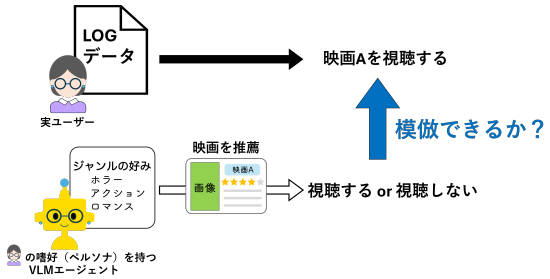

図 1　実験の概要図

RQ1：アイテムの提示情報に画像を加えることでエージェントの模倣精度は向上するか？

RQ2：ペルソナのシステムプロンプトに画像表現を加えることでエージェントの模倣精度は向上するか？

本研究では，画像表現の有無が異なるペルソナを用い，画像あり・なしの推薦におけるエージェントと実ユーザーの意思決定一致度を評価する（図 1）．なお，RQ2 では，実験環境を先行研究と揃えるため，システムプロンプトにペルソナ情報を入力した．この際，画像は直接入力できないため，base64 にエンコードした画像表現をテキストとして用いた．

2　関連研究

LLM を用いて様々な分野における人間の行動をシミュレートする研究が近年活発に行われている [2, 3, 5, 6, 8]．Park [1] らは，記憶検索，内省，計画からなる機構を LLM に組み込むことで，仮想環境内において人間らしい日常行動や社会的相互作用を自律的に生成できる Generative Agents を提案し，LLM エージェントによる人間行動シミュレーションの基盤的枠組みを提示した．Sun [5] らは LLM

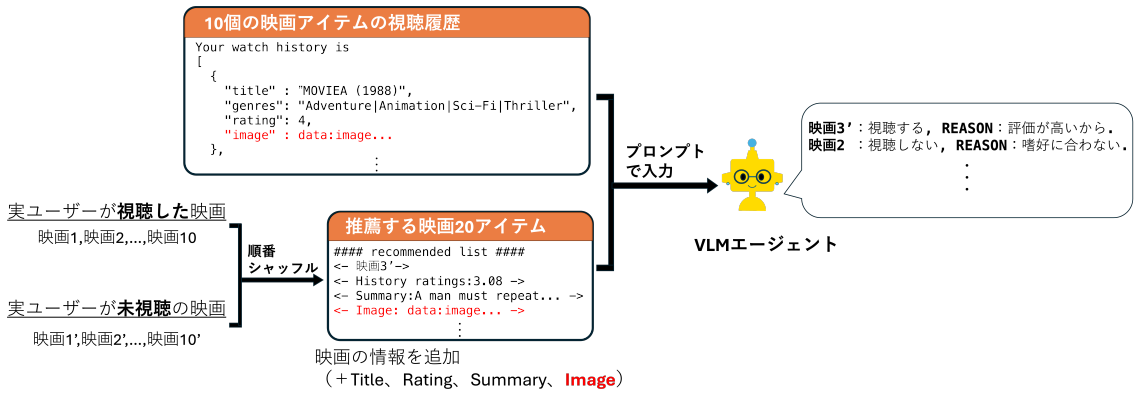

図 2　実験タスクの概要図

エージェントに，AI アシスタントを用いたオンラインショッピングの購買行動やその評価を通してユーザーの行動を模倣させた．

推薦システムの分野では，Zhang [2] らにより，LLM エージェントが人間のユーザー行動を模倣し，推薦システムのオンライン評価を再現する手法である Agent4Rec が提案されている．

しかし，Agent4Rec は人間の意思決定に影響を与える画像情報をシミュレーションに考慮していない．近年では，画像を考慮した研究 [4] も見られるが，個々のユーザーとエージェントを対応付け，画像情報がユーザー行動を忠実に再現するために影響するかの検証は筆者の知る限り行われていない．

本研究では，実ユーザーとエージェントを 1 対 1 で対応付け，推薦アイテムの提示情報に画像情報，ペルソナ構築に画像表現を導入した際の，ユーザー行動の模倣精度の変化を定量的に評価する．

3　実験手法

3.1　実験の概要

目的　推薦アイテムの提示情報，およびエージェントのペルソナとして入力する情報に画像情報あるいは画像表現を加えることで，エージェントによる実ユーザー行動の模倣精度がどのように変化するかを検証する．実験手法は Agent4Rec [2] を基盤とする．データセットには MovieLens-100K [9] を用い，その中から OMDb API でポスター画像を取得可能な映画とその視聴履歴を対象とする．

ペルソナとエージェント　Agent4Rec では視聴履歴を LLM によりジャンル嗜好に要約した文章をペルソナとして用いているが，画像を要約するのは困難なため，本研究では視聴履歴を入力し，比較として要約したペルソナを用いた実験も行う．また，Agent4Rec とペルソナ入力形式を揃えるため，画像を base64 にエンコードした画像表現をシステムプロンプトに入力した．（例を付録 A，図 3）に示す）．本研究でのエージェントは，VLM である GPT-4o-mini にペルソナをプロンプトとして与えたものとし，データセットからランダムに抽出した 100 ユーザーに対応するエージェントを作成した．

タスクと評価方法　タスクの概要を図 2 に示す．

- エージェントに，ジャンル嗜好の要約または視聴履歴をペルソナとしてシステムプロンプトに入力する．
- 20 の映画をユーザープロンプトに入力し，エージェント推薦する（図 4）．これらの映画は，実ユーザーが視聴したアイテムと未視聴のアイテムを 1:1 の割合で含む．
- エージェントは推薦された 20 個の映画すべてについて，視聴するか否かを「yes/no」で出力し，その判断理由を「reason」として出力する．
- 実ユーザーが視聴したアイテムに「yes」を出力できたかを，Precision，Recall，Accuracy，F1 score で算出し，本実験では F1 score をエージェントの模倣精度とする．

各条件で 5 回実験を行い，結果は平均値とする．

3.2　実験 1

推薦アイテムの提示情報にポスター画像を加えることが，エージェントの模倣精度に与える影響を検証する．推薦アイテムの提示情報として，「タイトル，サマリー，評価値」をエージェントに与える条件と，これらの情報に加えて「ポスター画像」を提示する条件を設定し，両条件におけるエージェントの模倣精度を比較する．

表 1　実験 1，実験 2 の結果

Condition	Precision	Recall	Accuracy	F1 score
Summary	0.6648	0.5418	0.6310	0.5810
Summary +I	0.6267	0.5212	0.6053	0.5567
History	0.6907	0.5658	0.6553	0.6123
History(high)	0.7249	0.4526	0.6377	0.5467
History +I	0.6400	0.6526	0.6420	0.6410
History(high) +I	0.6652	0.6058	0.6456	0.6249
History +H	0.6197	0.6796	0.6339	0.6430
History(high) +H	0.6484	0.6734	0.6518	0.6534
History +H +I	0.6216	0.7176	0.6386	0.6615
History(high) +H +I	0.6452	**0.7290**	**0.6647**	**0.6800**

3.3　実験 2

ペルソナに画像表現を含めることが，エージェントの模倣精度に与える影響を検証する．ペルソナに画像表現を入力するため，実ユーザーの視聴履歴をペルソナとした．視聴履歴には「タイトル，評価値，ジャンル」を含め，これらの情報のみを入力した条件と，base64 にエンコードした画像表現を加えて入力した条件の模倣精度を比較する．

視聴履歴は入力トークン数の制約から最大 10 アイテムとし，この制約下での評価値分布の偏りを考慮して，高評価（評価値 4 または 5）のみの視聴履歴と，低評価から高評価までが満遍なく含まれる視聴履歴の 2 条件を設定した．

4　結果と分析

実験 1, 2 の結果を表 1 に示す．ペルソナの入力が，ジャンル嗜好の要約文を「Summary」，視聴履歴を「History」とし，高評価のみ場合は「History(high)」と表記する．「+I」アイテム提示に画像，「+H」はペルソナに画像表現を付与した条件を示す．

4.1　RQ1：推薦アイテム提示情報への画像付与の影響

4.1.1　結果

嗜好を要約したペルソナでは，推薦アイテムの提示情報にポスター画像を追加すると，F1 score は 0.581 から 0.557 へと低下した．一方で，視聴履歴を用いたペルソナでは，ポスター画像を追加することで F1 score の向上が確認された．

4.1.2　考察

アイテム提示情報にポスター画像を追加すると，ペルソナにジャンル嗜好の要約文を用いた条件ではエージェントの模倣精度は低下し，視聴履歴を用いた条件では模倣精度が向上した．

また，視聴履歴を用いた多くの条件で，要約文を用いた条件より高い模倣精度を示した．生の行動履歴はノイズが多いため [10]，既存研究では，行動履歴を要約したペルソナを一般的に用いている [2, 4]．一方で，本結果は視聴履歴をペルソナとして用いる有効性を示唆する．これは，エージェントは視聴履歴からユーザー特性を把握し，その上でアイテムの画像情報を意思決定に活用できる可能性が考えられる．

4.2　RQ2：ペルソナへの画像表現付与の影響

4.2.1　結果

視聴履歴に画像表現を追加した条件（+H）では，F1 score の向上が示され，Recall が大きく増加していることも確認された．表 1 より，高評価のみの視聴履歴に対して，アイテム情報に画像，ペルソナに画像表現を加えた条件では，Accuracy 0.6647，F1 score 0.6800 と最も高い模倣精度を示した．

4.2.2　分析

ペルソナに画像表現を含む視聴履歴を用いた条件（+H）で Recall が向上した要因が，単にエージェントの yes 回答の増加によるものかを検証するため，yes の回答率を分析した．結果を表 2 に示す．エージェントの中には 20 個すべての映画に対して視聴可否を出力しないものも含まれるため，本分析では

表 2　エージェントの「yes」の回答率の統計量

Condition	Mean	Std	Min	Count
Summary	0.4284	0.1126	0.11	84
Summary +I	0.4255	0.1021	0.15	80
History	0.4115	0.0695	0.22	99
History(high)	0.3161	0.0802	0.12	98
History +I	0.5094	0.0833	0.28	95
History(high) +I	0.4576	0.0843	0.25	91
History +H	0.5460	0.0869	0.31	99
History(high) +H	0.5199	0.0730	0.36	100
History +H +I	0.5771	0.0688	0.38	96
History(high) +H +I	0.5642	0.0714	0.33	94

5 回の実験のうち 3 回以上，全アイテムに対して出力を行ったエージェントのみを有効とした．有効なエージェントの数は Count 列に示している．

その結果，視聴履歴に画像表現を追加した条件では，no ばかりを回答するエージェントが減少し，yes の回答率が増加する傾向が確認された．

次に Recall の向上が単に yes の出力数増加によるものかを検証するため，視聴履歴ペルソナに画像表現を付与していない条件（-H）をベースラインとし，ペルソナ画像表現ありの条件（+H）の Recall に達するまでランダムに yes の出力数のみを増加させる試行を 3000 回行った．その結果，Recall を同程度に再現した試行で，Precision や Accuracy が+H で観測された値を上回る試行が一定割合で存在した．そのため，ペルソナへの画像表現追加による性能変化が判断能力の向上によるものとは断定できなかった．

さらに，条件ごとに，エージェントがアイテム選択時に着目している観点を調べるため，出力された reason 文を分析した．その結果，模倣精度が高い条件では評価値への言及が多く，高評価の映画に yes，低評価の映画に no を出力する傾向が見られた．詳細は付録 B,C,D に記載する．

4.3　アブレーションスタディ

実験 2 の結果から，ペルソナに視聴履歴を入力した条件では，アイテム提示情報に画像情報，ペルソナに画像表現を含めた場合に，最も高い模倣精度が得られることが確認された．一方で，画像を加えたことによる性能向上が，テキスト情報を除去した場合に生じる性能低下と比べて十分に大きいのかは明らかではない．本節では，テキスト情報および画像情報，画像表現を段階的に除去するアブレーションスタディを行い，画像の寄与を相対的に評価する．

表 3　アイテム提示情報に関するアブレーション

Condition	Precision	Recall	Accuracy	F1 score
	0.6452	0.7290	0.6647	0.6800
w/o title	0.6511	0.6854	0.6575	0.6623
w/o summary	0.6799	0.6842	0.6766	0.6764
w/o rating	0.6222	0.7536	0.6432	0.6763
w/o image	0.6484	0.6734	0.6518	**0.6534**

表 4　ペルソナに関するアブレーション

Condition	Precision	Recall	Accuracy	F1
	0.6452	0.7290	0.6647	0.6800
w/o title	0.6758	0.6846	0.6708	0.6735
w/o genre	0.6573	0.7092	0.6648	0.6780
w/o rating	0.6464	0.7418	0.6632	0.6870
w/o image	0.6652	**0.6058**	0.6456	**0.6249**

表 3 より，アイテム提示情報において画像情報を除去した条件（w/o image）では，F1 score が最も低下していることが確認できる．同様に，表 4 に示すように，ペルソナ情報から画像表現を除去した条件では，Recall および F1 score が大きく低下した．したがって，画像はエージェントの行動に影響を与える重要な要素であることが示唆される．

5　おわりに

本研究では，アイテムの提示情報およびペルソナに画像情報，画像表現を付与することが，エージェントによる実ユーザー行動の模倣精度に与える影響を検証した．その結果，画像情報，画像表現が模倣精度に影響を与える重要な要素となることが示された．本研究で示された知見は，サムネイルなどの画像最適化におけるユーザー反応をシミュレートする応用可能性を示している．

さらに，嗜好を要約したペルソナよりも，視聴履歴を直接入力したペルソナの方が高い模倣精度を示し，視聴履歴を直接ペルソナとして用いる有効性を示したのは，本研究の貢献である．

一方で，本研究は先行研究と揃えるためにペルソナに画像を画像表現として追加したため，画像として入力した場合においても同様の傾向が観察されるかを検討することが今後の課題である．

参考文献

[1] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In **Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology**, pp. 1–22, 2023.

[2] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 1807–1817, 2024.

[3] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. **arXiv preprint arXiv:2405.02957**, 2024.

[4] Nicolas Bougie and Narimawa Watanabe. Simuser: Simulating user behavior with large language models for recommender system evaluation. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)**, pp. 43–60, 2025.

[5] Lu Sun, Shihan Fu, Bingsheng Yao, Yuxuan Lu, Wenbo Li, Hansu Gu, Jiri Gesi, Jing Huang, Chen Luo, and Dakuo Wang. Llm agent meets agentic ai: Can llm agents simulate customers to evaluate agentic-ai-based shopping assistants? **arXiv preprint arXiv:2509.21501**, 2025.

[6] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In **Proceedings of the eleventh ACM international conference on web search and data mining**, pp. 198–206, 2018.

[7] Marco Rossetti, Fabio Stella, and Markus Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In **Proceedings of the 10th ACM conference on recommender systems**, pp. 31–34, 2016.

[8] Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Zheshen, Wang, Qi He, and Dakuo Wang. Can llm agents simulate multi-turn human behavior? evidence from real online customer behavior data, 2025.

[9] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. **ACM transactions on interactive intelligent systems (tiis)**, Vol. 5, No. 4, pp. 1–19, 2015.

[10] Jiaxing Wu, Lin Ning, Luyang Liu, Harrison Lee, Neo Wu, Chao Wang, Sushant Prakash, Shawn O'Banion, Bradley Green, and Jun Xie. Rlpf: Reinforcement learning from prediction feedback for user summarization with llms. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 39, pp. 25488–25496, 2025.

A　プロンプト例

ペルソナプロンプト（視聴履歴の場合）

```
    Assume you are a user browsing movie
    recommendation system who has the following
    characteristics: "
    Your watch history is
[
  {
    "title": "Akira (1988)",
    "genres": "Adventure|Animation|Sci-Fi|Thriller",
    "rating": 4
  },
                        中略
]
```

図 3　入力した視聴履歴のプロンプト例

アイテム推薦のプロンプト

```
#### recommended list ####
<- Billy Madison (1995) -> <- History ratings:3.08 ->
<- Summary:A man must repeat grades 1-12 in order to
inherit his father's company. ->

                        中略

Please choose movies in the ##recommended list## that
you want to watch and explain why. After watching the
movie, evaluate each movie based on your
characteristics, taste and historical ratings to give
a rating from 1 to 5.
You only watch movies which aligh with your taste.
Use this format: MOVIE: [movie name]; WATCH: [yes or
no]; REASON: [brief reason]
You must judge all the movies. If you don't want to
watch a movie, use WATCH: no; REASON: [brief reason]
Each response should be on one line. Do not include
any additional information or explanations and stay
grounded in reality.
```

図 4　入力した推薦アイテムのプロンプト例

B　REASON 文分析

条件ごとに模倣精度が異なる要因の一つとして，エージェントがアイテム選択の際に着目している観点が条件によって変化している可能性が考えられる．エージェントが出力する reason は，その判断根拠を反映した重要な手掛かりであるため，LDA（Latent Dirichlet Allocation）を用いて reason 文を 3 つのトピックに分類し，条件ごとのトピック分布を分析した．結果を図 5 に示す．

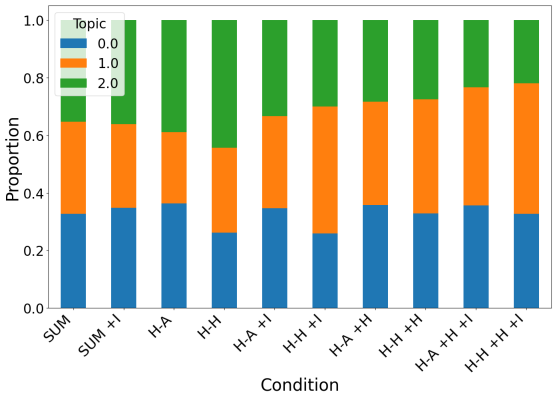

図 5　条件ごとの reason トピック分布

Topic 0 は「themes」「story」「unique」などの単語を多く含み，映画の内容に言及するトピック，Topic 1 は「rating」「historical」「low」など，映画の評価値に関する言及を中心としたトピック，Topic 2 は「interested」「taste」「aligned」など，自身の嗜好に合うか否かに着目したトピックであると解釈できる．

各条件のトピック割合と模倣精度の順位相関を調べたところ，Topic 2 は Precision と 0.56 の正の相関を，Topic 1 は Recall および F1 とそれぞれ 0.75 程度の相関を示した．

C　トピックごとの単語

トピックごとに出現した上位 5 個の単語を表 5 にまとめた．

表 5　LDA によるトピック分類での出現単語

Topic	単語
0	themes,taste,comedy,premise,story
1	rating,historical,low,strong,appealing
2	interested,interests,taste,aligned,align

D　映画の評価ごとの yes/no 回答分析

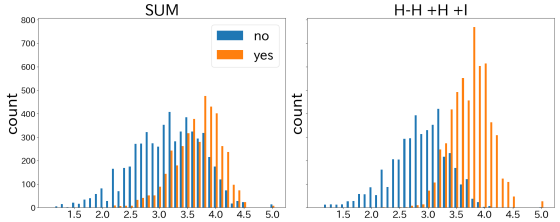

図 6　映画の評価ごとの yes/no 回答分布

映画の評価値ごとに yes/no の出力分布を分析した．図 6 より，F1 score が低い条件では高評価の映画でも no が多く出力される一方，F1 score が高い条件では高評価の映画には yes，低評価の映画には no が多く出力される傾向が確認できた．