APPENDIX

A DETAIL OF THE COLLECTED YOUTUBE DATASET

Our dataset comprises a comprehensive set of 4,004 content creators, each with a unique influence level, as indicated by their follower counts varying from 10 to 12.32 K. Additionally, we included 1.97 M users and 3.97 M comments in total—spanning 0.19 M distinct items of 14 genres. Due to the inability to collect explicit preference behaviors from specific users (e.g., clicks and likes), we consider user comments as an indication of interest in the item. Out of privacy protection concerns, we mask specific sensitive information of users and channels.

Due to resource limitations, the number of creators and users simulated in this article is limited. Therefore, we specifically randomly sampled from the complete YouTube dataset (i.e., Big_Youtube) to construct a more densely populated Small_YouTube dataset. To better present the details of the YouTube dataset we collected, we present the statistics of the dataset in Table 1. The dataset has been shared at https://anonymous.4open.science/r/CreAgent-9B4B.

Table 1: Statistics of the collected YouTube Dataset

Dataset	Big_Youtube	Small_Youtube
#Interactions	3,970,123	40,479
#Users	1,967,066	1,571
#Items	186,164	64,300
#Creators	4,004	643
#Genres	14	14
Inter. Per User	2.02	25.77
Item Per Creator	46.49	100

B ABLATION STUDY ON DIFFERENT LLMS

We employ two additional LLMs as base models to carry out experiments (that is, Mistral-7B and Qwen2-7B), to verify the capability and effectiveness of CreAgent under these LLMs and explore some new findings. Due to the time and resource constraints during the rebuttal phase, the number of LLMs we could test is limited. We sincerely apologize for this. In the future, we will explore and evaluate CreAgent with more additional LLMs.

B.1 Real-world data alignment evaluation

Result.. We observe varying performance when using different LLMs as the base model for our method. In terms of the categories of items created by the simulated creators, the results consistently aligned closely with real-world data. However, in the aspect of creative diversity, some models, such as Mistral-7B, exhibited weaker performance. We hypothesize that this is due to the fact that our prompts were fine-tuned specifically on LLama, leading to potential inconsistencies when applied to other models. Despite this variation, it is notable that while these models may underperform compared to LLama, CreAgent still outperformed the baselines (e.g., CFD[1], LBR[2]).

Table 2: Comparison of the divergence between the simulated and real-world distributions using Jensen-Shannon divergence [2], with genre-level for preference evaluation and individual-level for diversity.

Simulation Method	Preference	Diversity
Creator Feature Dynamics [1, 4]	0.2537	0.7204
Local Better Response [3]	0.2833	0.6284
SimuLine [6]	0.3175	0.6949
CreAgent(Mistral-7B)	0.2045	0.4012
CreAgent(Qwen2-7B)	0.1917	0.3979
CreAgent(LLama3-8B)	0.1667	0.3014

Simulation Method	Accumulated Reward
Creator Feature Dynamics [1, 4]	3.08
Local Better Response [3]	2.51
SimuLine [6]	3.04
CreAgent(Mistral-7B)	7.32
CreAgent(Qwen2-7B)	7.98
CreAgent(LLama3-8B)	8.11

B.2 Strategic behavior alignment evaluation

Result. For the strategic behavior of creator agents, we first conducted experiments to evaluate CreAgent's reward acquisition capability using different LLMs as base models. All rewards were normalized against the random strategy, following the settings in the paper. As shown in the table, CreAgent consistently demonstrated superior analytical decision-making and creative abilities across almost all base models, achieving higher user rewards. However, we observe that not all base models achieved rewards comparable to LLama3-8B. For instance, Mistral-7B may have limitations in its post-pretraining capabilities, making it less effective at analyzing current user feedback.

Simulation Method	Very Low	Low	Medium	High	Very High
Random	0.0431	0.0343	0.0356	0.0547	0.052
Creator Feature Dynamics [1, 4]	0.8959	0.9867	0.9746	1.0000	1.0000
Local Better Response [3]	0.4069	0.7009	0.6778	0.6102	0.4583
SimuLine [6]	0.9104	0.9712	0.9556	0.9476	1.0000
CreAgent(LLama3-8B)	0.6138	0.8953	0.9220	0.9250	0.9498
CreAgent(Qwen2-7B)	0.7272	0.8057	0.7848	0.8170	0.8333
CreAgent(Mistral-7B)	0.4822	0.7500	0.8471	0.8261	0.8235

Result. We conducted experiments to evaluate the exploration-exploitation balance of creator agents under varying reward levels, assessing whether their behavior aligns with prospect theory. The table highlights the proportion of exploitation actions taken after receiving different levels of user feedback on newly-created items. While different LLMs exhibited varying exploration-exploitation levels under different rewards, they all displayed patterns resembling human behavior**. Specifically, agents showed a strong inclination to explore under low rewards and a remarkable tendency to exploit under high rewards, reflecting loss-seeking under low returns and risk aversion under high returns. This behavior sharply contrasts with the random strategy and traditional embedding-based baselines. For instance, CFD shifted to 100% exploitation at

high reward levels, while LBR paradoxically reduced exploitation proportions under high rewards.

We encourage future research to utilize CreAgent and our simulation platform to explore the capabilities and limitations of LLMs in simulating human behavior.

C USER ALIGNMENT EVALUATION

In this section, we conduct experiments on the user agent employed in our simulation environment to validate how effectively it aligns with real-world user preferences and behaviors.

C.1 User Item Preference Alignment

Table 3: User Item Preference Alignment

1:m	Accuracy	Precision	Recall	F1 Score
1:1	$0.630_{\pm 0.031}$	$\mathbf{0.658}_{\pm 0.023}$	$0.603_{\pm 0.070}$	$\mathbf{0.589}_{\pm 0.054}$
1:2	$0.598_{\pm 0.037}$	$0.461_{\pm 0.042}$	$0.523_{\pm 0.059}$	$0.460_{\pm 0.047}$
1:3	$0.622_{\pm 0.015}$	$0.373_{\pm 0.022}$	$0.520_{\pm 0.031}$	$0.404_{\pm 0.020}$
1:9	$\mathbf{0.653}_{\pm 0.040}$	$0.276_{\pm 0.036}$	$\mathbf{0.740}_{\pm 0.063}$	$0.358_{\pm 0.039}$

Motivation. To validate how well generative agents align with the real-world preferences, we utilze the user agents to differentiate between items that actual users have engaged with and those they have not. Specifically, a total of 200 agents will each be randomly assigned 20 items. Among these, the ratio between items the user has interacted with (i.e., $\mathbf{Y}_{u,i}(0) = 1$) but was not utilized for profile initialization and items the user has not interacted with (i.e., $\mathbf{Y}_{u,i}(0) = 0$) is set as 1:m, with $m \in \{1, 2, 3, 9\}$. Under this setting, user agent responses (i.e., $\mathbf{Y}_{u,i}(n) = 1, n \in [1, 2, \cdots, N]$) to recommended items are considered binary discrimination, taking values between 0 and 1. Then, we compute the accuracy, precision, recall, and f1-score metric to show its performance.

Results. Table 3 reports the empirical discrimination results across various metrics. The best performance for each metric is highlighted in bold and marked with an asterisk. We observe that: The generative user agents consistently identify items that align well with user preferences, maintaining around 65.3% accuracy and 74% recall even when faced with 18 (i.e., $1 : m = 1 : 9$) distracting items. This high performance is attributed to the personalized profiles that accurately reflect users' true interests, demonstrating the agents' ability to encapsulate real preferences and highlighting the viability of LLM-powered generative agents in recommendation systems.

In our item-by-item recommendation setting, the user agents do not tend to click on a certain number of items in the recommendation list, as mentioned in [5]. However, our user agent can ensure a high level of Recall (above 0.5) and accuracy (nearly above 0.6) when 1:m decreases. This indicates that our user agent, however, maintains a certain tendency towards identifying positive examples even when the proportion of similar items decreases, which may result in some negative items being identified as positive. We attribute this failure to LLM's inherent hallucinations that agents tend to consistently pick a set number of items. However, we emphasize that in the subsequent simulation results with recommendation algorithms, the recommendation list length is set to 5, hence a substantial proportion of recommended items align with user preferences, thereby endorsing high trustworthiness in those simulation outcomes.

C.2 User Genre Preference Alignment

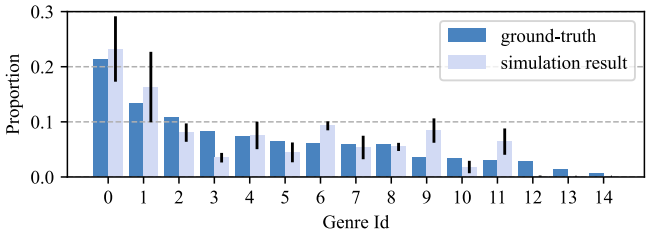

Figure 1: Comparison between the distributions of ground-truth and agent-simulated genre preference.

Motivation. In a real-world RS, users have unique interests in different genres of items. Since these interests drive user actions such as viewing, clicking, and liking, it is crucial for our user agents to align with the preferences of real-world users. Specifically, we aim to align the interest distribution of our user agents with the distribution observed in the real-world dataset we have collected.
Results. To verify the consistency between the interest distribution of our user agent and that of real-world users, we conducted experiments using our CreAgent framework. We initialized user and creator profiles with the YouTube dataset we collected, utilizing DIN as the recommendation model and simulating 100 time steps. Figure 1 (a) shows the interest distribution of users from the real-world dataset, where users who clicked more than five times on a category are considered to favor that category. The x-axis represents item genres, sorted from highest to lowest proportionally. Figure 1 (b) illustrates the interest distribution of our user agent, where we compute the proportion of clicks for each genre. From the comparison, we can see that our user agent ultimately achieved an interest distribution similar to that of real-world users. However, we are unable to perfectly replicate the relative differences across certain genres. For example, in the *Howto & Style* (H&S) genre, the user agent exhibited a higher preference than real users, while extremely low preference is observed in *Sports* (s). We attribute this to the LLM's extensive prior knowledge of genres, which causes the agent to exhibit a stronger preference for certain genres of items.

D DETAILS OF THE PROMPTS

D.1 Prompts for Profile Summarization

Designed Prompt for Social Identity Summarization

Prompt: You are a content creator on {platform name} and your name is {creator name}. Here is the basic information about the content you have previously created.
Recent created content: title:<title>, genre:<genre>, description:<description>

Created content genre (the genres you have created in the past and their respective proportions): {created genre proportion}
Creation frequency (the average number of items you create each day): {creation time per day}
Please summarize your social identity in the following format: [Social Identity]: <the specific identity>. For example, [Social Identity]: movie enthusiast.

Designed Prompt for Intrinsic Motivation Summarization

Prompt: You are a content creator on {platform name} and your name is {creator name}. Here is the basic information about the content you have previously created.
Follower number: follower number
Average views per video: average views.
Recent created content: title:<title>, genre:<genre>, description:<description>
Recent interaction with users (your recent interaction records with the audience in the comments section.): {recent comments}
Creation frequency (the average number of items you create each day): {creation time per day}
Intrinsic motivation refers to whether your purpose for creating content is for profit or simply for sharing. Please summarize your intrinsic motivation in the following format: [Intrinsic Motivation]: <the specific motivation>. For example, [Intrinsic Motivation]: profit.

D.2 Prompts for Creation Module

Designed Prompt for Fast Thinker

Prompt: You are a content creator on YouTube and your nickname is {name}.
{profile text of P_c^m,and P_c^a}
Based on the analysis: {$\mathbf{A}_c^{\text{exp}}$}, please create ONE new content for {name} that fits user's interest.
You can refer to the creation history of {name}: {$f(\mathcal{M}_c^{\text{cre}})$}
Response in JSON dictionary format. Write "name": [item name], "genre": genre1|genre2|...., "tags": [tag1, tag2, tag3], "description": "item description text")

Designed Prompt for Slow Thinker

Prompt: You are a content creator on YouTube and your nickname is {name}.
{profile text of P_c^m,and P_c^a}
The average utility per item of each genre {name} has created is as below: {$\mathbf{B}_{\mathbf{c}}^{\text{aud}}$}. ([unknown] means the item genre {name} have not explored.
Recently, {name} created an item of genre {g_i}, and receives {$\mathbf{z}_i(n)$} utility.
Due to the statistical data, {name}'s profile and {name}'s familiarity on each genre: $\mathbf{B}_{\mathbf{c}}^{\text{skill}}$, {name} must choose one of the two actions below to obtain more user clicks:
(1) [EXPLORE] Create content in a new genre that has not been explored before, which means other genres may have a larger audience and more opportunities to profit. But it might not be {name}'s area of expertise and requires greater effort to create.
(2) [EXPLOIT] Sticking to creating content of a familiar genre, which means {name} will leverage his creative expertise to build a stable brand identity. But it might limit {name}'s audience reach and lead to insufficient income.
To explore a new genre, write: [EXPLORE]:: <genre name>. If so, give the specific genre name chosen from unknown cates.
To stick to familiar genres, write: [EXPLOIT]:: <genre name>. If so, give the specific genre name chosen from known cates.
Let's think step by step. Please answer concisely and strictly follow the output rules.
Responses Example of $\mathbf{A}_c^{\mathbf{exp}}$: [EXPLORE]: Entertainment

REFERENCES

[1] Tao Lin, Kun Jin, Andrew Estornell, Xiaoying Zhang, Yiling Chen, and Yang Liu. 2024. User-Creator Feature Dynamics in Recommender Systems with Dual Influence. *arXiv preprint arXiv:2407.14094* (2024).
[2] María Luisa Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute* 334, 2 (1997), 307–318.
[3] Fan Yao, Yiming Liao, Mingzhe Wu, Chuanhao Li, Yan Zhu, James Yang, Qifan Wang, Haifeng Xu, and Hongning Wang. 2024. User Welfare Optimization in Recommender Systems with Competing Content Creators. *arXiv preprint arXiv:2404.18319* (2024).
[4] Ruohan Zhan, Konstantina Christakopoulou, Ya Le, Jayden Ooi, Martin Mladenov, Alex Beutel, Craig Boutilier, Ed Chi, and Minmin Chen. 2021. Towards content provider aware recommender systems: A simulation study on the interplay between user and provider utilities. In *Proceedings of the Web Conference 2021*. 3872–3883.
[5] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1817.
[6] Guangping Zhang, Dongsheng Li, Hansu Gu, Tun Lu, Li Shang, and Ning Gu. 2023. Simulating news recommendation ecosystem for fun and profit. *arXiv preprint arXiv:2305.14103* (2023).