The Effects of Diversity in Algorithmic Recommendations on Digital Content Consumption and User Retention – A Field Experiment on a Music Streaming Platform

Guangying Chen[1], Tat Y. Chan[1], Dennis J. Zhang[1], Senmao Liu[2], Yuxiang Wu[2]

October 30, 2022

(Preliminary draft; please do not circulate.)

Abstract

Social media platforms such as TikTok and Facebook are criticized for trapping consumers in "filter bubbles" (Pariser, 2011) through personalized recommendations based on users' detailed individual information. Practitioners and regulators have been calling for platforms to tackle the problem by incorporating more diversified content in their recommender systems. We aim to study the causal impacts of more diversified personalized recommendations on users' behaviors in practice. By collaborating with NetEase Cloud Music, the world's third-largest music streaming service company, we developed a new recommender system with more content diversity based on their existing state-of-art recommender system. We then conducted a large-scale field experiment where hundreds of millions of users were randomly assigned to receive video recommendations either from the platform's current recommender algorithm or our modified algorithm with 2.33% higher topic diversity. Overall, our new algorithm did not affect users' consumption diversity but lowered their consumption level by 3.23%. However, for active users, our system shows that a 1% increase in recommendation diversity boosted their consumption diversity by 0.55% without hurting retention, consumption, or engagement. We further explore the possible mechanisms and provide guidance for platforms to improve their recommender systems.

Keywords: Personalized Recommender Systems, Filter Bubble, Social Media Platforms, Recommendation Diversification, Content Consumption, User Retention, Field Experiment

[1] Washington University in St. Louis, St. Louis, Missouri, USA
[2] NetEase Cloud Music Inc., Hangzhou, Zhejiang, China

1. Introduction

Social Media plays an important role in today's life. In 2022, 59% of people in the world use social media and spend on average 2.48 hours on them every day.[3] Almost all large social media platforms including TikTok and Facebook personalize content recommendations based on users' detailed historical individual information in order to attract users to spend more time and generate more revenue through digital advertisement. Such personalized recommender systems often trap consumers in their own "filter bubbles" (Pariser, 2011)— platforms recommend users only the content with opinions and information that conform to their existing beliefs. Filter bubbles can severely enhance individuals' opinion bias, foster social media addiction, and even hurt consumers' long-term health. Practitioners and regulators have been calling for social media platforms to tackle the filter-bubble problem by incorporating more diversified content in their recommender systems. For example, the U.S. Senate introduced a "Filter Bubble Transparency Act (S.2024)" in June 2021, requiring online platforms to allow their consumers to conveniently opt out from personalized recommendations using their individual historical data.[4] As the bill sponsor, John Thune, mentioned in his speech in November 2019, the act aims to provide consumers an option "to see information that has not been selected specifically for them" and help them expand consumption diversity.[5]

But will allowing users to disable personalized recommendations help solve the filter-bubble problem on social media platforms? Evidence shows it might not be the case. First, users may not choose to opt out from personalized recommendations. A global survey in the Reuters Institute's 2016 Digital News Report shows most people prefer to get news from personalized recommendations (36%) compared to editorial/journalistic recommendations (30%) or social recommendations (22%).[6] Second, even if people choose to accept non-personalized recommendations to improve content diversity, they may directly ignore the more diversified content recommended by platforms and be pushed into a narrower filter bubble. For example, a Facebook experiment in 2018 showed that disabling the personalized News Feed ranking algorithm led users to hide 50% more recommended posts, mostly due to their dislikes or lack of interest. More concerningly, users sharply increased their usage of Facebook Groups, which often contain more extreme content. In addition, the experiment found more diversified recommendations also significantly lowered users' engagement (i.e., comments between friends) by 20%.[7] That is mainly why platforms are reluctant to recommend more diversified content. Hence, it is important for policy makers and social media platforms to understand how recommendation diversity will affect consumers' behaviors and how to improve recommendation diversity.

---

[3] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
[4] https://www.congress.gov/bill/117th-congress/senate-bill/2024; https://www.makeuseof.com/filter-bubble-transparency-act-explained/
[5] https://www.thune.senate.gov/public/index.cfm/press-releases?ID=F0D3EA8C-D573-4A7F-89BB-027ABE2781F2
[6] https://www.digitalnewsreport.org/essays/2016/people-want-personalised-recommendations/
[7] https://bigtechnology.substack.com/p/facebook-removed-the-news-feed-algorithm?s=09

Given the importance of recommendation diversity to consumers, platforms, and society, our research tries to address the following questions: (1) How does diversifying a cutting-edge recommender algorithm change consumers' behaviors? (2) Do consumers differ in their reactions to a more diversified recommender system? (3) How can social media platforms improve their recommendation diversity without hurting short-term metrics, such as consumption, engagement, and retention?

To answer these questions, we partnered with NetEase Cloud Music (NCM), the world's third largest music streaming service company, and conducted a large-scale, randomized field experiment in the "Cloud Village" page of their music mobile app. Cloud Village builds a personalized recommender system to recommend various music-related videos for its visitors. Unlike the past literature where a currently employed state-of-art algorithm is compared with a more diverse and but practically infeasible algorithm (e.g., Claussen, Peukert, and Sen, 2019; Holtz et al., 2020), we directly modified the platform's current state-of-art recommender algorithm to increase video topic diversity. Doing so will allow us to minimize the negative impact of increasing diversity on customer engagement and will make our insights more generalizable to practical settings on other platforms. We then conducted a field experiment for 14 weeks and randomly sampled 10 million users for our analysis. During the experiment, users of the treatment group saw content recommended by our new and more diverse algorithm, and users of the control group were recommended by the original system. A manipulation check reveals that the new algorithm exposed the treatment group to 2.33% more recommendation diversity than the control group (measured by the Herfindahl–Hirschman Index (HHI) of recommended video topics).

We find that, overall, our new algorithm did not increase users' consumption diversity but lowered their consumption level (measured by weekly clicking frequency) by 3.23%. Instead of helping consumers break social media bubbles, more diversified recommendations mainly hurt their consumption level, confirming the concerns of most social media platforms. We then tested if the results hold across new users, inactive users (users who have been inactive for more than four weeks) and active users. We find no significant effects of the new algorithm on new users' short-term metrics or consumption diversity. For inactive users, the new algorithm not only had no impact on their consumption diversity, but also decreased their consumption level. In contrast, for active users, the new algorithm effectively boosted their consumption diversity without hurting retention, consumption, or engagement level. A 1% increase in recommendation diversity led to active users' 0.55% increase in their consumption diversity. Our findings suggest while the trade-off between consumption and diversity exists on average across all users, for active users, who contribute the most engagement and revenue to the platform, higher recommendation diversity can effectively mitigate their filter-bubble problem and encourage them to consume more diversified content without sacrificing short-term metrics. Based on our findings, NCM updated its original recommender algorithm by increasing the recommendation diversity to its active users. The updated algorithm significantly increased both active users' consumption level and their consumption diversity.

We further explore what drives the positive reactions of active users. Results reveal that the positive effects of recommendation diversity only happen on users who spend significant time on the platform *as well as* whose preference is well understood by the algorithm (measured by the number of videos a user has clicked). If an active user only views a limited number of very long videos, she could spend a lot of time on the platform, but the recommender system does not understand her well because of the lack of her click data. Interestingly, for these users, the increased recommendation diversity did not increase their consumption diversity and hurt their short-term metrics such as clicking frequency, view time, and number of likes left. This result implies both a good understanding of individual consumption preferences and users' sufficiently high valuation of the platform are crucial for benefiting from increased recommendation diversity.

Our research builds on marketing and computer science literature that studies the social and economic impacts of algorithmic recommendations, such as ideological affiliation (e.g., Bakshy, Messing, and Adamic 2015; Ribeiro et al. 2020; Huszár et al. 2022), product purchase (e.g., Fleder and Hosanagar 2009; Ghose, Ipeirotis, and Li 2014; Lee and Hosanagar 2019), and media consumption (e.g., Zhou, Khemmarat, and Gao 2010; Hosanagar et al. 2014; Berman and Katona, 2020; Holtz et al., 2020; Moehring 2022). Our paper contributes to an emerging topic on how personalized algorithmic recommendations, which utilize users' individual historical data, affect the level and diversity of user consumption.

Previous studies mainly compare the cutting-edge deep-learning recommender systems in a company with less efficient recommender algorithms that are seldomly used in practice, such as popularity-based (e.g., Holtz et al., 2020), time-based (e.g., Dujeancourt et al. 2022), and human recommendations (e.g., Claussen, Peukert, and Sen, 2019). These comparisons have primarily documented that personalized algorithmic recommendations lead to a higher consumption level but lower consumption diversity. For example, using observational data from an online music service, Hosanagar et al. (2014) found consumers purchased more songs after receiving personalized recommendations. Claussen, Peukert, and Sen (2019) conducted a field experiment on a major German news website and found that personalized algorithmic recommendations generated more user clicks than blanket recommendations from human editors. Holtz et al. (2020) also proved causally through an online field experiment on Spotify that consumers significantly increased podcast streams after the platform replaced popularity-based recommendations with personalized recommendations. At the same time, this existing literature also shows personalized recommendations can polarize user consumption and trap users in their "filter bubbles" compared to non-personalized recommendations. For example, Claussen, Peukert, and Sen (2019) and Holtz et al. (2020) found personalization decreased users' consumption diversity measured by HHI and Shannon entropy of consumed topics. Claussen, Peukert, and Sen (2019) further demonstrated that the declining effect could spread to users' news consumption in other non-personalized sections. Anderson et. al (2020) also ran a

field experiment on Spotify and found personalized recommendations performed better for users with lower consumption diversity.

Although the relationship between personalized recommendation and user consumption has attracted much attention, little research has studied the causal impact of changing the diversity level of a cutting-edge personalized recommender algorithm. Ultimately, platforms who want to improve their recommendation diversity are more likely to adjust their existing algorithms instead of using popularity-based or human-editing ranking. By contrast, our study focuses on changing a cutting-edge personalized recommender system that is currently used in practice directly to incorporate more diversity and quantifying the causal impact of such changes on user consumption, engagement, and retention. Contrary to the prior literature, we demonstrate that making a state-of-art recommender system more diversified could help improve both active users' consumption diversity and consumption.

Our research also contributes to the literature on recommendation diversification in computer science. Since Bradley and Smyth (2001) first introduced diversification into recommender systems, a large body of research on recommendation diversity has emerged in computer science (e.g., Kaminskas and Bridge 2016; Kunaver and Požrl 2017). Most studies focus on measuring the diversity of recommendations (e.g., Clarke et al. 2008; Fleder and Hosanagar 2009; Vargas and Castells 2011; Vargas et al. 2014), evaluating the impacts of diversification on recommendation accuracy and consumer satisfaction (e.g., Adomavicius and Kwon 2011a; Hurley and Zhang 2011; Ekstrand et al. 2014; Javari and Jalili 2015), and improving diversification algorithms (e.g., Ziegler et al., 2005; Adomavicius and Kwon 2011b; Vaishnavi, Jayanthi, and Karthik 2013). We extend this literature about diversification effects in two ways. First, unlike the previous simulation or survey studies, we quantify the causal effects of algorithm diversification on consumer behaviors through a real-world field experiment. Second, we propose two possible underlying mechanisms for researchers and platforms to better understand consumers' heterogeneous responses to recommendation diversification.

In addition, our paper provides important practical value for practitioners. First, our results guarantee external validity through a large-scale, randomized field experiment involving hundreds of millions of platform users. Second, since our experiments were implemented in a personalized recommender system currently used in practice, other digital platforms can easily modify the design and incorporate the change into their own recommender system. Third, we provide suggestions for platforms on how to lift users' consumption diversity without hurting their short-term metrics. Platforms can customize the level of diversification for users in different stages. Specifically, platforms should increase the recommendation diversity for active users, especially heavy users whose preferences are well understood by recommender algorithms. On the one hand, these users are more loyal to the platform and more open to interest exploration. On the other hand, the algorithm learns their preferences better through their historical click behaviors and thus can more accurately predict what new content they will like. By

contrast, for inactive users, the platform should be cautious on exploration and set a low level of recommendation diversity. With little knowledge about these users, adopting a conservative recommendation strategy could help platforms better attract potential users and more quickly collect information about their preferences.

2. Field Setting and Experimental Design

To study the causal impact of content diversity of recommendations, we collaborated with NCM and conducted a randomized field experiment. NCM has over 800 million users till 2019 and 181.9 million monthly active users in the first half of 2022.[8] Almost all users access NCM services through its mobile app. The mobile app has a main tab called "Cloud Village", where users can watch various music-related videos recommended by the platform's algorithm. These videos are created by users of the platform and NCM categorizes each video into around 80 topics including music sharing, film mashup, instrumental performance, dancing, etc. Similar to other social media platforms like Facebook, users who click to watch a video can like, comment, and share the video (see Figure 1). Our experiment was implemented in this Cloud Village. In this section, we will first introduce the personalized recommender system used in our setting and then explain how our experiment changed their current recommender system to affect the diversity of recommended contents.

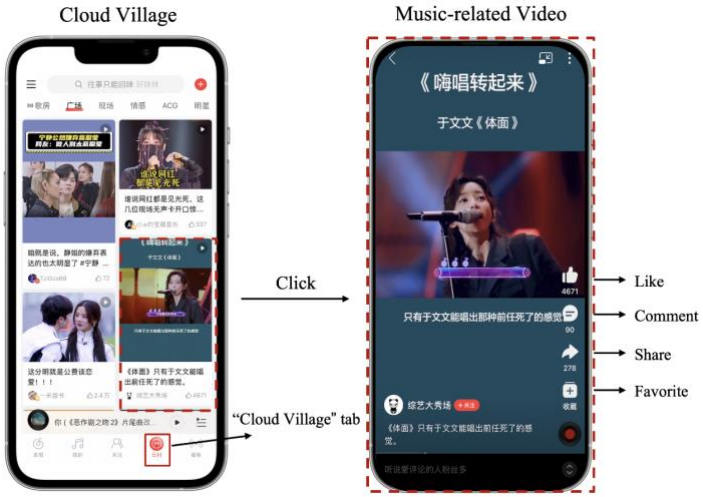


Figure 1 Cloud Village Tab (left) and Music-related Videos (right)

2.1. The Current Personalized Recommender System

NCM designs a personalized deep-learning recommender system to recommend videos to Cloud Village visitors. This system is standard in the industry, and many other social media companies are using similar architectures. Every time when a user visits the tab, the system receives a query and selects Z videos from

[8] https://news.mydrivers.com/1/640/640143.htm, https://www1.hkexnews.hk/listedco/listconews/sehk/2022/0922/2022092200788_c.pdf

its video pool to recommend to the user. For platforms such as TikTok and NCM, their video pools typically contain millions of videos. The selection algorithm consists of three stages: (1) Retrieval, (2) Ranking, and (3) Re-ranking. Next, we will explain the three stages in detail (see Figure 2 for an illustrative example).

(1) **Retrieval**: For user i, the platform uses multiple strategies to retrieve in total X (around 500 - 10,000, $X \gg Z$) videos from the video pool. The most common strategy is the user-to-item (u2i) model, which accounts for about 40% of video retrievals. Each video candidate and user i are represented by an embedding pre-trained based on the user's historical activities on the platform. The model calculates the distance between user i and each video by passing a sigmoid function of the dot product of user and video embeddings and then selects a list of videos with close distances. Such videos are usually from the topics the user has clicked on or the creator the user has liked. The other strategies account less for the video retrievals, including cold-start (10%), item-to-item model (8%), video popularity (5%), hot events (5%), etc.

(2) **Ranking**: A single algorithm then ranks all these X videos by their predicted match value Y_i to user i and passes a list of top X' (around 100 - 300, $Z \leq X' \leq X$) videos to the next stage. In order to predict a video's match value to a user, the algorithm first trains several deep neural networks to predict a user's actions towards a video, such as the probabilities to click or share a video and how long the user is watching the video, using the user's and the video's features. The algorithm then calculates the video-user pair's match value as a function of these predictions. Note that it is possible this algorithm ranks a very popular video from a topic that the user has not seen before very high and recommends the video to her.

(3) **Re-ranking**: An algorithm then re-ranks the top X' videos on the list from the Ranking stage and decides which Z (around 10 - 20) videos will be recommended and their listing orders. First, within the small set of X' videos, an algorithm re-trains the above deep neural networks to predict each video-user pair's match value Y_i' by adding more complex user-video features. The X' videos are then re-ranked based on Y_i'. This step aims to improve the algorithm's prediction accuracy while saving computational resources by focusing only on a small set of high-value videos.

Second, other re-ranking strategies are considered in this stage, such as adding advertisements on certain blocks (like every 7th video), list optimization, and the focus of our paper, *content diversification*. The algorithm can effectively adjust the diversity level of videos recommended to a user through a parameter called window size (denoted as S). The larger the window size S is, the more likely the user will be recommended more diversified videos.[9]

[9] In Appendix A, to illustrate why setting a larger window size can effectively increase the content diversity of recommendations, we randomly select a Cloud Village user and simulate the algorithm's recommendations to her request. We ask the algorithm to retrieve 400 videos, select 50 top-ranked videos, and re-rank the 50 videos using window size 5 or 30. Then we compare the topic diversity of final recommendations under window size 5 versus 30.

Specifically, the algorithm first picks the video with the highest Y_i' as the first recommended video. Then, for the second recommended video, the algorithm chooses, among the remaining top S videos on the list, the most different video compared with the first recommended video (regarding topics, creators, and other video characteristics). Similarly, for the third recommended video, the algorithm chooses, among the remaining top S videos on the list, the most different video compared with both the first and the second recommended videos. The Re-ranking stage will follow this procedure to select Z videos, which are then recommended to the user in order. Our experiment directly changed the window size S to affect the content diversity of recommendations.

When a user further swipes up the screen for more videos, the algorithm will remove all previously recommended videos from its video pool and repeat the above procedure.

2.2. Algorithm Change and Field Experiment Design

We conducted a large-scale field experiment for 14 weeks from December 22, 2021, to March 29, 2022, and randomly sampled 10 million users who visited the Cloud Village tab at least once during the experiment. Upon arrival of each user, we randomly assigned around 3% of them to the treatment group and the remaining 97% to the control group based on a hash function of their user ids.

Users of the control group were recommended by the original recommender system in which the window size is 5 if users having clicks in the past 30 days and 15 otherwise. For users of the treatment group, we modified the original recommender system and increased the window size in the Re-ranking stage to 30, which directly increased the likelihood of recommending contents of topics different from what are normally recommended to them. We choose to increase the window size in re-ranking to increase recommendation diversity because this is the most strategy-agnostic way. Other ways of increasing recommendation diversity may include adding additional more diverse retrieval sources or changing the ranking equation whose implementations may depend on each company's individual recommendation strategies.

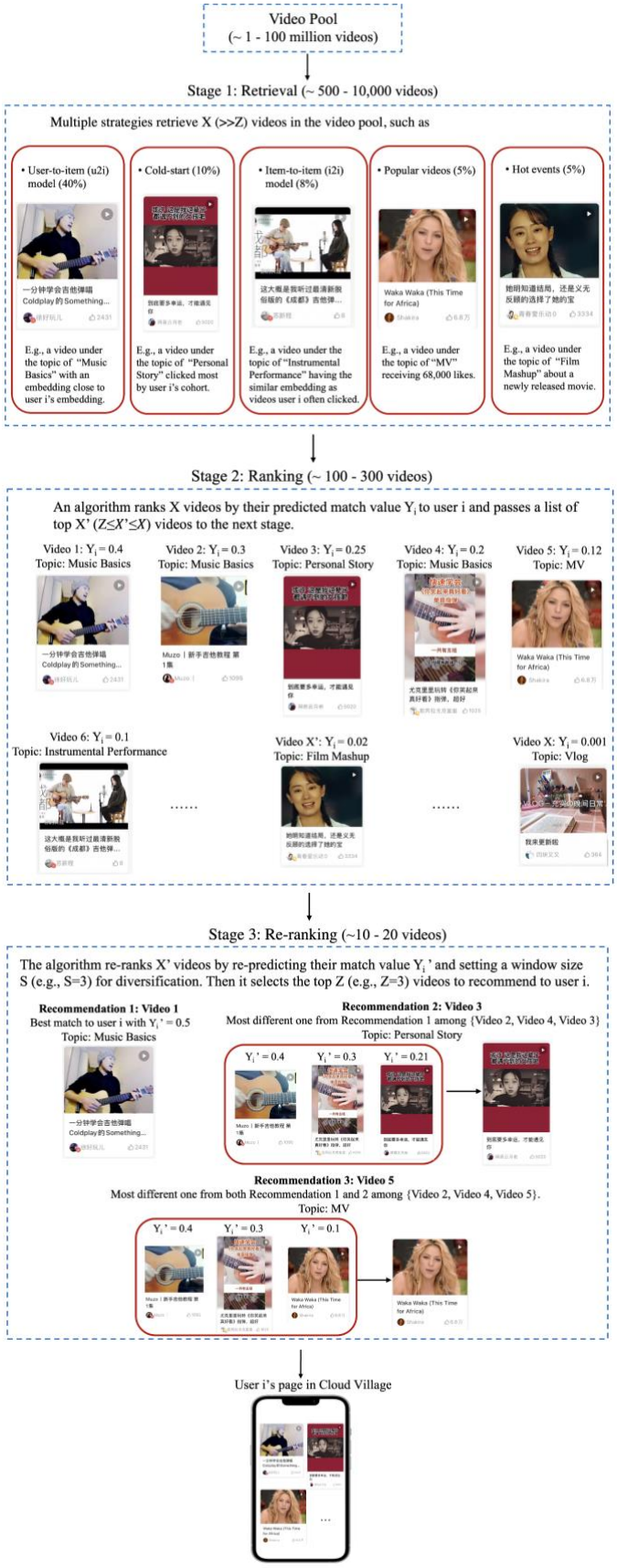


Figure 2 The Current Personalized Recommender System of Cloud Village: An Illustrative Example

3. Data

In this section, we introduce our data and conduct both randomization and manipulation checks to test the validity of our field experiment. Our data is divided into three time periods: (1) the pre-experiment period, including four weeks before the experiment (November 19 to December 16, 2021), (2) the experiment implementation period (December 17 to December 21, 2021),[10] and (3) the experiment period, totaling fourteen weeks after the experiment started (December 22, 2021, to March 29, 2022).[11] We use the pre-experiment data to conduct randomization check and classify users, and the 14-week experiment data to conduct manipulation check and estimate the causal effects of recommendation diversity.

We collected users' demographic information, including their app registration time, gender, and age. To quantify users' retention, consumption, and engagement level, we tracked their weekly activities in the Cloud Village, such as whether they visited the tab or clicked on a video, how long they watched a video, and whether they liked, commented, or shared a video. Moreover, we also tracked what videos have been recommended to the users in addition to their consumption. Consistent with the literature (e.g., Claussen, Peukert, and Sen, 2019; Holtz et al., 2020), we collected the topic of each video and used three indices to measure the content diversity of videos recommended to or consumed by users: the number of different video topics (num_topic_{it}), the HHI of video topics (HHI_topic_{it}), and the Shannon entropy of video topics ($entropy\_topic_{it}$). The latter two metrics are defined as follows:

$$HHI\_topic_{it} = \sum_{j=1}^{num\_topic_{it}} s_j^2 \tag{1}$$

$$Entropy\_topics_{it} = -\sum_{j=1}^{num\_topic_{it}} s_j \cdot ln\left(s_j\right) \tag{2}$$

where num_topic_{it} indicates the number of different topics user i was recommended to or consumed in week t. $s_j \in (0,1]$ refers to the share of videos from topic j. The content diversity increases as the number of topics increases, HHI decreases, and Shannon entropy increases. Based on van Dam (2019), the concept of diversity includes not only the "variety" dimension (i.e., the total number of categories in the taxonomy) but also the "balance" dimension (i.e., the distribution of elements across categories). Thus, HHI and Shannon entropy are more comprehensive indicators of content diversity than the number of video topics. The latter reflects only the variety of recommended videos, while the first two also consider how evenly distributed the recommended videos are across different topics. We present the variable description and summary statistics in Table 1.

[10] NCM first increased the window size of treatment group users to 15 on December 17, 2021, and further increased the window size to 30 on December 21, 2021.

[11] In Appendix B, we include the experiment implementation period into the experiment period and report the qualitatively consistent results.

Table 1: Variable Description and Summary Statistics

Variables	Description	Number of Observations	Mean	St. Dev.	Min	Max
Treatment	1 if a user was assigned into the treatment group, 0	10,000,000	0.0319	0.1758	0	1
New_user	1 if a user registered on the music app after the pre-experiment period, 0	10,000,000	0.0510	0.2199	0	1
Num_registered month	Number of months a user had been registered on the app by the end of the experiment	10,000,000	40.6003	22.0691	0.0000	108.5333
Male	1 if a user was predicted to be male, 0	9,886,059	0.5344	0.4988	0	1
Age	A user's predicted age	9,734,933	23.0331	5.8964	11	45
Visit	1 if a user visited the Cloud Village in a week, 0	140,000,000	0.1931	0.3948	0	1
Freq_click	Frequency of days per week that a user watched at least one video for no less than five seconds	140,000,000	0.0022	0.0242	0.0000	1.0000
Num_click	Number of clicks per week with more-than-5-second view time	140,000,000	0.2569	8.8395	0	3,655
View_min	Minutes spent watching videos per week	140,000,000	0.2088	6.9775	0.0000	4,091.2270
Num_like	Number of likes left per week	140,000,000	0.0090	0.9371	0	4,090
Num_comment	Number of comments left per week	140,000,000	0.0002	0.0399	0	167
Num_share	Number of shares left per week	140,000,000	0.0004	0.0487	0	171
Num_recommended_topic	Number of different topics recommended to a user per week when visiting the Cloud Village	27,039,411	8.1998	5.2659	1	80
HHI_recommended_topic	HHI of video topics recommended to a user per week when visiting the Cloud Village	27,039,411	0.1758	0.0741	0.0295	1.0000
Entropy_recommended_topic	Shannon entropy of video topics recommended to a user per week when visiting the Cloud Village	27,039,411	1.8912	0.4329	0.0000	3.7470
Num_clicked_topic	Number of different topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,631,675	5.5862	7.6677	1	66
HHI_clicked_topic	HHI of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,631,675	0.5810	0.3706	0.0391	1.0000
Entropy_clicked_topic	Shannon entropy of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,631,675	0.9193	0.9314	0.0000	3.4383

Notes. Summary statistics for 10 million randomly sampled Cloud Village visitors during the experiment. Users' age and gender were predicted by a supervised machine learning model built by NCM. Among the sampled users, 1.14% (113,941) did not have predicted gender, and 1.83% (182,526) did not have predicted age. To remove age outliers, we keep only users in the $0.5\% - 99.5\%$ quantile of the age distribution ($11 - 45$ years old).

As a randomization check, we compare the demographics, pre-experiment recommendation diversity, and pre-experiment activities between treatment and control group users (see Table 2 and Table 3; Regression specifications are reported in Appendix C). The results confirm no significant difference between the two groups.

Table 2: User Demographics at the End of the Experiment Between Control and Treatment Groups

	Dependent variable:			
	new user	*num_registered month*	*male*	*age*
Treatment	−0.00002	−0.0211	0.0001	0.0123
	(0.0004)	(0.0397)	(0.0009)	(0.0107)
Constant	0.0510***	40.6009***	0.5344***	23.0327***
	(0.0001)	(0.0071)	(0.0002)	(0.0019)
Observations	10,000,000	10,000,000	9,886,059	9,734,933
R^2	0.0000	0.000000	0.0000	0.000000
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3: Users' Retention, Consumption, Engagement, and Diversity of Recommendations and Consumption During the Pre-experiment Period

(a) Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.00001	0.00003	0.0130	0.0075	−0.0008	−0.00001	−0.00002
	(0.0004)	(0.00004)	(0.0124)	(0.0097)	(0.0005)	(0.00002)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	37,824,167	37,824,167	37,824,167	37,824,167	37,824,167	37,824,167	37,824,167
R^2	0.00004	0.000002	0.000001	0.000001	0.000000	0.000000	0.000000

(b) Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.0167	−0.0002	0.0011
	(0.0136)	(0.0002)	(0.0012)
Week FE	Yes	Yes	Yes
Observations	6,386,497	6,386,497	6,386,497
R^2	0.0008	0.0015	0.0018

(c) Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1411	−0.0035	0.0123
	(0.0928)	(0.0040)	(0.0105)
Week FE	Yes	Yes	Yes
Observations	383,371	383,371	383,371
R^2	0.0007	0.0001	0.0002

Notes. We exclude new users who registered on the music app after the pre-experiment period. Standard errors clustered at individual level in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

To ensure our experiment recommended significantly more diversified videos to the treatment group (i.e., manipulation check), we compare the recommendation diversity between treatment and control visitors during the experiment period. The regression model is specified as follows:

$$Recommendation\ Diversity_{it} = \alpha_1 \cdot Treatment_i + v_t + \epsilon_{it}, \quad (3)$$

where $Treatment_i$ is a binary variable that equals 1 if user *i* was in the treatment condition and 0 otherwise. v_t represents the week fixed effects controlling the weekly shocks across all users. We cluster the error term ϵ_{it} at the individual level. Since users receive recommendations only when they visit the Cloud Village, this regression only applies to weeks users visited. As we mentioned above, the content diversity of recommendations for user *i* visiting in week *t* ($Recommendation\ Diversity_{it}$) is quantified by the number ($num\_recommended\_topic_{it}$), HHI ($HHI\_recommended\_topic_{it}$), and entropy ($entropy\_recommended\_topic_{it}$) of recommended video topics.

We summarize the results in Table 4. Every week when users visited the Cloud Village, treatment users were recommended 0.120 more topics compared to control users (8.196 topics), amounting to a 1.47% increase. Similarly, the HHI (entropy) of recommended videos for treatment users also significantly decreased (increased) by 2.33% (1.02%) on a base of 0.176 (1.891) (all p-values < 0.0001). Therefore, our experiment successfully increased the diversity of videos recommended to treatment users.

Table 4: Content Diversity of Recommendations Between Treatment and Control Visitors During the Experiment

	Dependent variable:		
	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1203***	−0.0041***	0.0192***
	(0.0112)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	27,039,411	27,039,411	27,039,411
R^2	0.0030	0.0097	0.0076
Note:			*p<0.1; **p<0.05; ***p<0.01

4. Results

4.1. The Effects of Recommendation Diversity

To identify the causal effects of algorithmic recommendation diversity, we estimate the following regression model:

$$Outcome\ Variable_{it} = \beta_1 \cdot Treatment_i + u_t + e_{it}, \quad (4)$$

where $Outcome\ Variable_{it} \in \{visit_{it}, freq\_visit_{it}, num\_click_{it}, view\_min_{it}, num\_clicked\_topic_{it}, HHI\_clicked\_topic_{it}, entropy\_clicked\_topic_{it}\}$ is detailed later. $Treatment_i$ is defined the same as Equation (3). u_t represents the week fixed effects, and the error term e_{it} is clustered at the individual level.

For user *i* in week *t*, we examine her retention by whether visiting the tab ($visit_{it}$). The user's week *t* consumption level is quantified by the frequency of days when user *i* had clicked at least once ($freq\_visit_{it}$), the number of total clicks (num_click_{it}), and the number of minutes user *i* spent on watching videos ($view\_min_{it}$). Here we determine that user *i* clicked on a video if she watched the video for at least five seconds. The definition is consistent with NCM's measurement and more accurately indicates users' consumption intention by avoiding the situation where users mistakenly click on a video without the intention to watch it. User i's engagement level is measured by the number of likes (num_like_{it}), comments ($num\_comment_{it}$), and shares (num_share_{it}) she left in week t. Table 5(a) shows that the higher recommendation diversity for treatment users significantly lowered their consumption level, despite no significant effects on their retention or engagement level. Specifically, treatment users, on average, significantly reduced their weekly clicking frequency by 3.23% (p = 0.0137) on a base of 0.0022, which translates to the elasticity that a 1% decrease in the recommendation HHI to a user leads to a 1.39% drop in her clicking frequency.

To estimate the effects of recommendation diversity on user *i*'s consumption diversity, we calculated the number ($num\_clicked\_topic_{it}$), HHI ($HHI\_clicked\_topic_{it}$), and entropy ($entropy\_clicked\_topic_{it}$) of all video topics she clicked on in week *t*.[12] Table 5(b) indicates that the

[12] When estimating the effects on users' consumption diversity, we only consider the weeks when user i had at least clicked once in the Cloud Village, which involves 1,084,228 (10.84%) of the sampled 10 million users.

higher recommendation diversity for treatment users encouraged them to click on 0.181 more topics (3.24%, p = 0.0223) but did not significantly shift their consumption HHI (p = 0.3040) or entropy (p = 0.0551). It suggests that although treatment users were encouraged to try more topics, they still consumed disproportionally on their familiar topics and spent only a very limited amount of time on new topics.

The average treatment effects imply that the new and more diversified recommender system not only has a limited effect in increasing all users' consumption diversity to help them break social filter bubbles, but also significantly hurts their consumption level, which confirms the concerns of most social media platforms.

Table 5: The Average Treatment Effects of Recommendation Diversity

(a) Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00007**	0.0142	0.0124	0.0008	0.0001	0.00002
	(0.0003)	(0.00003)	(0.0117)	(0.0095)	(0.0010)	(0.0001)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	140,000,000	140,000,000	140,000,000	140,000,000	140,000,000	140,000,000	140,000,000
R^2	0.0016	0.0001	0.00001	0.00001	0.000002	0.000001	0.000001

(b) Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1809**	−0.0026	0.0138*
	(0.0791)	(0.0025)	(0.0072)
Week FE	Yes	Yes	Yes
Observations	1,631,675	1,631,675	1,631,675
R^2	0.0013	0.0008	0.0012

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

4.2. The Heterogeneous Effects of Recommendation Diversity for New, Inactive, and Active Users

Given the underwhelming average treatment effects, we naturally ask: will the effects be the same across different users, and will some user segments increase their consumption diversity without reducing short-term metrics? Following this logic, we divide all users into new users (5.10%), inactive users (91.42%), and active users (3.48%) based on how they were familiar with and active on the platform before the experiment. We define *new users* as those having not registered on the NCM app before the experiment implementation period and old users otherwise. During the experiment, new users contribute to 3.42% of the total view time of all platform users. Since new users just began to get familiar with the app and the Cloud Village tab after the experiment started, their expectations for the platform could be very different from old users and thus respond differently from old users. For old users, we further divide them into *inactive users*, who did not view any content during the pre-experiment period, and *active users*, who had viewed videos during the pre-experiment period. Despite having the lowest user percentage, active users accounted for the most view time (54.66%) during the experiment. In contrast, inactive users contributed only 41.92% of the total view time. Compared to new or inactive users, active users spending more time in

the Cloud Village may have been satiated with the videos of their familiar topics and be more open to recommendations of unfamiliar topics. Therefore, in this section, we will explore how higher recommendation diversity affects the three different user groups.

4.2.1. New Users

Table 6(a) tests the manipulation effect on new users as Equation (3) specifies. The results confirm that our new algorithm significantly increased the recommendation diversity to new users of the treatment group. Every week when they visited the Cloud Village, they were recommended 0.156 more topics ($p = 0.0037$, a 1.81% increase) than control users (8.657 topics) and faced a 2.21% ($p < 0.0001$) decrease in the HHI of recommended video topics on a base of 0.172[13].

Table 6(b) and 6(c) display regression results from Equation (4) on new users. The results indicate such recommendation diversity increase has no significant impacts on new users' short-term metrics or consumption diversity. Given new users signing up for the platform after the experiment began, they could be still in a stage exploring the functions of Cloud Village. Most contents are new to them, thus more diversified recommendations did not affect their retention, consumption, or engagement. At the same time, since the platform has little information about their preferences, new users may find the recommended new topics are far from their interest and choose to ignore them, leading to the null effects on their consumption diversity.

[13] The entropy of recommended topics for treatment users also increased by 1.02% ($p<0.0001$) on a base of 1.921.

Table 6: The Manipulation Check and Treatment Effects of Recommendation Diversity for New Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1564***	−0.0038***	0.0195***
	(0.0539)	(0.0005)	(0.0034)
Week Fixed Effects	Yes	Yes	Yes
Observations	976,467	976,467	976,467
R^2	0.0024	0.0135	0.0074

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0004	0.00002	0.0276	0.0290	0.0134	0.0007	0.0002
	(0.0009)	(0.0001)	(0.0346)	(0.0326)	(0.0097)	(0.0008)	(0.0001)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,133,294	7,133,294	7,133,294	7,133,294	7,133,294	7,133,294	7,133,294
R^2	0.0097	0.0007	0.0001	0.0001	0.00002	0.00002	0.00001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.2879	−0.0108	0.0297
	(0.2979)	(0.0101)	(0.0277)
Week FE	Yes	Yes	Yes
Observations	69,231	69,231	69,231
R^2	0.0021	0.0021	0.0025

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

4.2.2. Inactive Users

For inactive users, Table 7(a) confirms that our treatment effectively increased recommendation diversity by 1.29% in the topic number (on a base of 7.896, $p < 0.0001$) and 2.15% in HHI (on a base of 0.177, $p < 0.0001$).[14] The treatment effects indicate that inactive users responded negatively to higher recommendation diversity. As Table 7(b) and 7(c) show, treatment users significantly reduced their weekly clicking frequency by 5.59% (on a base of 0.0015, $p < 0.0001$) without consuming more diversely. A 1% decrease in recommendation HHI translates into a 2.60% drop in inactive users' clicking frequency. The significant consumption drop implies that the recommendation accuracy is crucial for keeping these inactive users engaged. Instead of exploring new topics, they might still wait for the recommender algorithm to learn their preference and recommend more videos matching their interests. Thus, more diversified but less accurate recommendations could easily annoy them and hurt their consumption. Compared with new users, inactive users might have explored the Cloud Village for a while and had a lower valuation of the tab with less uncertainty. Correspondingly, they would be less tolerant of the algorithm change.

[14] The entropy of recommended video topics for the treatment group also increased by 0.94% ($p < 0.0001$) on a base of 1.877.

Table 7: The Manipulation Check and Treatment Effects of Recommendation Diversity for Inactive Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1022***	−0.0038***	0.0177***
	(0.0082)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	24,458,918	24,458,918	24,458,918
R^2	0.0038	0.0101	0.0081

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00008***	−0.0052	−0.0016	−0.0003	0.00003	0.00003
	(0.0003)	(0.00001)	(0.0051)	(0.0048)	(0.0005)	(0.00003)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	127,992,298	127,992,298	127,992,298	127,992,298	127,992,298	127,992,298	127,992,298
R^2	0.0018	0.0001	0.00002	0.00002	0.000003	0.000001	0.000001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.0034	0.0035	−0.0037
	(0.0553)	(0.0024)	(0.0063)
Week FE	Yes	Yes	Yes
Observations	1,154,409	1,154,409	1,154,409
R^2	0.0022	0.0018	0.0021

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

4.2.3. Active Users

Unlike new users or inactive users, active users reacted positively to our new and more diversified recommender system. As Table 8(a) shows, treated active users were recommended 0.330 ($p = 0.0047$) more topics each week than those in the control group (12.486 topics), equivalent to a 2.64% increase. The experiment also significantly decreased the HHI of recommendations by 5.17% ($p < 0.0001$) for treatment users on a base of 0.161.[15] Regarding user reactions, treated active users significantly increased their consumption diversity without reducing retention, consumption, or engagement level (see Table 5(b) and 5(c)). Table 5(c) shows the significant consumption diversity elasticity of recommendation diversity is 2.10 in topic number and 0.55 in HHI.[16] The results imply that active users not only expand their topic variety but also balance more between their familiar and unfamiliar topics when facing more diversified recommendations.

[15] The entropy of recommended topics consistently increased by 1.94% ($p < 0.0001$) for treatment users on a base of 2.084.

[16] For active users of the treatment, the number of clicked topics increased by 5.54% from 9.497 ($p = 0.0164$), and the HHI of clicked topics decreased by 2.85% from 0.415 ($p = 0.0100$). Consistently, treatment users also increased consumption entropy by 3.04% (on a base of 1.388, $p = 0.0060$) facing a 1.94% increase in recommendation entropy, which translates into an elasticity of 1.56.

Table 8: The Manipulation Check and Treatment Effects of Recommendation Diversity for Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.3295***	−0.0083***	0.0405***
	(0.1167)	(0.0005)	(0.0044)
Week Fixed Effects	Yes	Yes	Yes
Observations	1,604,026	1,604,026	1,604,026
R^2	0.0023	0.0110	0.0098

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00001	0.4591	0.3203	0.0088	−0.0004	−0.0004
	(0.0022)	(0.0007)	(0.2976)	(0.2327)	(0.0206)	(0.0005)	(0.0006)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408
R^2	0.0075	0.0009	0.0003	0.0003	0.00004	0.00001	0.00003

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.5263**	−0.0118**	0.0422***
	(0.2192)	(0.0046)	(0.0154)
Week FE	Yes	Yes	Yes
Observations	408,035	408,035	408,035
R^2	0.0019	0.0010	0.0015

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

To summarize, we find that more diversified recommendations boosted active users' consumption diversity without sacrificing their retention, consumption, or engagement. In contrast, it did not affect new or inactive users' consumption diversity but hurt inactive users' consumption level.

4.3. Field Implementation Inspired by Our Results

Our findings suggest that, by recommending more diversified videos, the platform can effectively increase active users' consumption diversity without hurting their short-term metrics. Inspired by our results, NCM updated its original recommender system by increasing the recommendation diversity to active users. Such change significantly boosted active users' watching time per week by 7.27% (33 seconds, p = 0.0293) in the first two weeks before they expanded the treatment to every user. Moreover, it also encouraged active users to consume 4.83% more topics per week (p = 0.0001) and increased their consumption diversity in HHI by 2.19% (p = 0.0052). Both evidence from this implementation further verifies our empirical results (see Appendix D for more details about the implementation).

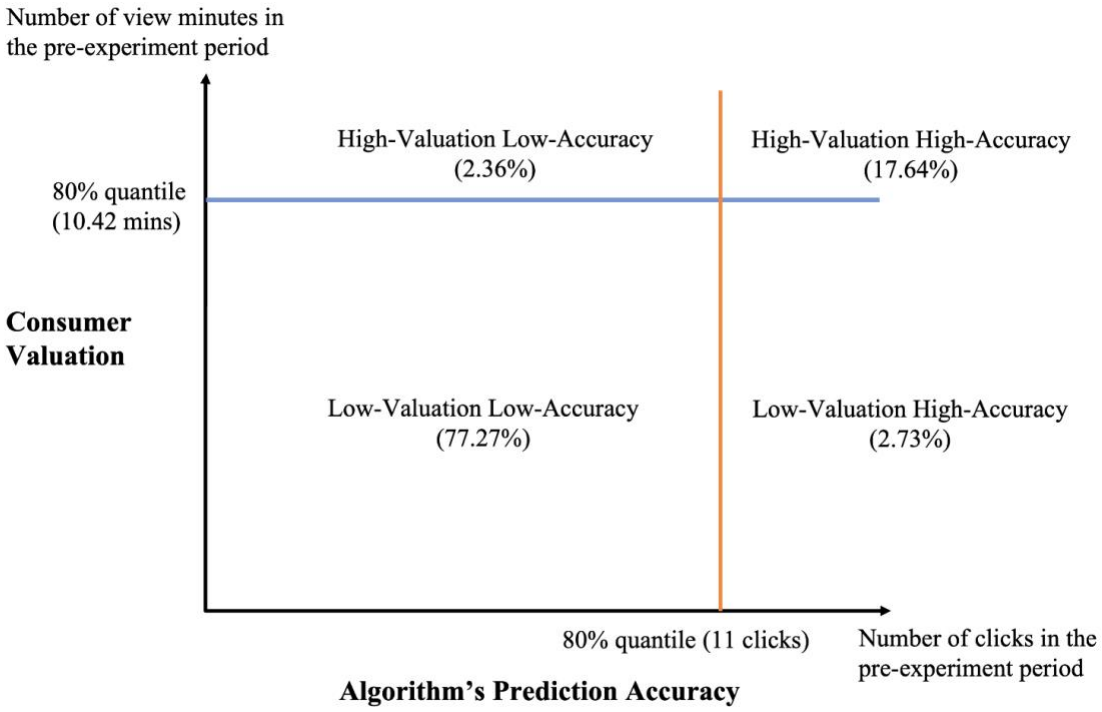
5. Possible Mechanisms

To explain why the effects of recommendation diversity are positive on active users but negative on inactive users, we propose two possible mechanisms. The first mechanism is customer valuation: when a user's valuation for the videos in Cloud Village is not high enough, she will be intolerant of unfamiliar or less

interesting recommendations. Higher recommendation diversity would only affect a user's consumption diversity and positively impacts her consumption when the user could tolerate the less interesting videos introduced by the more diversified algorithm at the beginning. Therefore, we could see our more diversified algorithm has a positive impact on active users who have higher valuation of the service and negative impact on inactive users who have lower valuation. The second mechanism is the algorithm's ability to predict a user's preferences: if the recommender system cannot understand a user's preference correctly due to lack of data, the user will be less tolerant of the algorithm recommending more diverse content. Active users tend to have more data on the platform and in turn the algorithm tends to understand them better. This could also be the reason that active users can tolerate our more diverse algorithms and increase their consumption diversity while inactive users cannot.

To test these two mechanisms, we divide the active users into four segments based on their valuation and the algorithm's ability to predict their preferences. Considering the small sample size of active users (348,172) among the 10 million sample, we randomly re-sampled 2 million active users who had visited the Cloud Village during the experiment for our analysis to guarantee the statistical power. As Figure 3 shows, we use the total view time during the pre-experiment period to measure active users' valuation of the Cloud Village and apply the 80/20 rule to separate them into high-valuation and low-valuation users. Users who spent at least 10.42 minutes (80% quantile among active users) watching videos in the tab during the pre-experiment period are defined as high-valuation users, and the others are defined as low-valuation users. To measure how accurate the algorithm is able to predict a user's preference, we use the total number of clicks a user had in the pre-experiment period.[17] As a user clicks more in the tab, the algorithm gathers more information about what videos the user likes and learns better about her preference. We classify users having at least 11 clicks (80% quantile among active users) in the pre-experiment period as high-accuracy users and the others as low-accuracy users. Thus, the active users are grouped into four segments: high-valuation high-accuracy users (17.64%), high-valuation low-accuracy users (2.36%), low-valuation high-accuracy users (2.73%), and low-valuation low-accuracy users (77.27%).

[17] We also consider using the total number of impressions a user had received in the pre-experiment period to measure the algorithm's prediction accuracy. In Appendix E, we use both users' click number and impression number during the pre-experiment period to separately predict their next-week click-through rates (= the number of clicks/the number of impressions). We find the click number has a stronger predicting power in terms of R square (0.1390 vs. 0.0164). Besides, based on the experience of NCM, the total number of impressions is a noisier indicator for users' preferences because users may unintentionally click into the Cloud Village tab and receive impressions recommended by the algorithm. Thus, we choose the click number to measure the prediction accuracy of the recommender algorithm.



Note. The percentage of high-accuracy users are not exactly 20% since the click number is discrete.

Figure 3 Segmenting Active Users Based on Consumer Valuation and Algorithm's Prediction Accuracy

Table 9 summarizes the estimation results using the same specifications for the four segments of active users.[18] Facing more diversified recommendations, only high-valuation high-accuracy users significantly increased their consumption diversity without reducing retention or consumption. A 1% decrease in recommendation HHI led to a 0.29% ($p < 0.0001$) decrease in consumption HHI. In contrast, the increased recommendation diversity significantly reduced low-valuation users' retention or consumption. For low-valuation low-accuracy users, a 1% decrease in recommendation HHI resulted in a 0.69% ($p = 0.0190$) decrease in clicking frequency. Although low-valuation high-accuracy users were encouraged to consume marginally more diversified videos with an elasticity of 0.59 in HHI ($p = 0.0715$), they reduced their visiting probability by 0.79% ($p = 0.0036$) when facing a 1% decrease in recommendation HHI. As for high-valuation low-accuracy users, the increased recommendation diversity did not affect their consumption diversity but lowered their consumption and engagement level. When the recommendation HHI decreased by 1%, high-valuation low-accuracy users clicked 2.14% less frequently ($p = 0.0060$), spent 3.30% less time watching videos ($p = 0.0412$), and left 8.67% fewer likes ($p < 0.0001$) and 10.49% fewer comments ($p < 0.0001$). The above results imply that both a sufficiently high level of customer valuation and the algorithm's sufficiently high prediction accuracy are necessary to prevent users from reducing short-term metrics when facing increased recommendation diversity. More importantly, to

[18] Appendix F listed the complete regression results for the four segments.

benefit from the increased recommendation diversity and encourage users to consume more diversely, the platform's recommender algorithm needs to have a good understanding of users' consumption preferences.

Table 9: The Manipulation Check and Treatment Effects of Recommendation Diversity for Four Segments of Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
High-Valuation High-Accuracy User	0.6180***	−0.0134***	0.0690***
	(0.1387)	(0.0004)	(0.0043)
High-Valuation Low-Accuracy User	−0.0216	−0.0102***	0.0410***
	(0.1999)	(0.0014)	(0.0102)
Low-Valuation High-Accuracy User	0.4985**	−0.0092***	0.0485***
	(0.2425)	(0.0013)	(0.0106)
Low-Valuation Low-Accuracy User	0.2042***	−0.0069***	0.0322***
	(0.0306)	(0.0002)	(0.0017)

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
High-Valuation High-Accuracy User	0.0013	0.0008	0.7386	0.6262	0.0178	−0.0014	−0.0027*
	(0.0027)	(0.0013)	(0.5690)	(0.4594)	(0.0434)	(0.0015)	(0.0014)
High-Valuation Low-Accuracy User	−0.0002	−0.0022***	−0.1894	−0.3205**	−0.0261***	−0.0015***	0.0018
	(0.0059)	(0.0008)	(0.1840)	(0.1570)	(0.0046)	(0.0003)	(0.0018)
Low-Valuation High-Accuracy User	−0.0163***	−0.0002	0.4060	0.3240	−0.0189	−0.0010*	−0.0008
	(0.0056)	(0.0012)	(0.4118)	(0.3256)	(0.0189)	(0.0006)	(0.0011)
Low-Valuation Low-Accuracy User	0.0003	−0.0002**	0.0217	0.0151	−0.0022	0.0002	−0.00002
	(0.0010)	(0.0001)	(0.0323)	(0.0251)	(0.0018)	(0.0002)	(0.0001)

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
High-Valuation High-Accuracy User	0.3830***	−0.0080***	0.0342***
	(0.1313)	(0.0020)	(0.0076)
High-Valuation Low-Accuracy User	−0.2510	0.0174	−0.0407
	(0.2767)	(0.0110)	(0.0313)
Low-Valuation High-Accuracy User	0.4706	−0.0146*	0.0463*
	(0.3030)	(0.0081)	(0.0256)
Low-Valuation Low-Accuracy User	0.0779	0.0022	0.0015
	(0.0714)	(0.0028)	(0.0077)

Notes. Estimates for the difference between treatment and control groups after controlling week fixed effects. Standard errors clustered at individual level in parentheses. There are 352,756 high-valuation high-accuracy users, 47,246 high-valuation low-accuracy users, 54,574 low-valuation high-accuracy users, and 1,545,424 low-valuation low-accuracy users. *p<0.1; **p<0.05; ***p<0.01.

Other mechanisms may also explain the above heterogeneous treatment effects. For example, as documented by Hu and Pu (2011), users' perceived diversity can differ from the algorithm's designed diversity level. Compared to active users, new and inactive users may be less sensitive to the increase in recommendation diversity. However, given that our diversity-enhanced algorithm significantly increased the diversity of recommendations for new and inactive users, they should have consumed more diversely if they did not perceive the algorithm change. Instead, neither of them significantly changed their consumption diversity during the experiment. Therefore, we can rule out the mechanism of perceived diversity in our context.

6. Conclusion

We conduct a field experiment with NCM to evaluate the effects of content diversity in algorithmic recommendations on social media users' retention, consumption, engagement as well as their consumption diversity. We show that the increased recommendation diversity, on average, does not affect the platform

users' consumption diversity but hurts their consumption level. More importantly, we find the impacts differ significantly across users. New users are not affected by the recommendation diversity change. For inactive users, increasing recommendation diversity only lowers their clicking frequency without encouraging them to consume more diversely. In contrast, active users, who contribute most to the platform's profit, not only maintain their retention, consumption, and engagement level, but also increase their consumption diversity when facing more diversified recommendations. We further explore two possible mechanisms for active users' positive responses and find that, only when users have a sufficiently high valuation of the platform's contents and their preferences are well understood by the platform, can a more diversified recommender system bring positive effects.

Our findings suggest social media platforms like NCM should customize the levels of recommendation diversity for different users. For inactive users, platforms should set a low level of recommendation diversity to first guarantee the recommendation accuracy. By keeping these users visiting and consuming contents, the recommender system can quickly learn their preferences from their activities on the platform. After these users grow into high-valuation users and the recommender algorithm has well learned their preferences, platforms can increase the diversity of recommendations. This way, while helping users get out of social media bubbles, platforms can also explore users' interests to benefit long-term development without losing short-term profits.

Our research also has several important implications for policymakers. First, we demonstrate that it is possible for platforms to achieve both more diverse content and higher engagement, which motivates policymakers to push platforms for more diversified content recommendations. Second, our research also sheds light on the intersection between public policies about user data privacy and more diversified content. Our research shows that, to encourage users to consume more diversified content on digital platforms, platforms need to well understand their users' consumption preferences. Or else users would directly ignore the new contents recommended by these platforms and stay in their social media bubbles. Therefore, completely limiting the use of individual data for digital platforms might instead block the way for users to embrace more diversity, which points to more complicated privacy regulations without completely blocking data access, such as differential privacy.

In addition, our paper has several limitations that bring interesting future research opportunities. First, we conducted our experiment on a video-based social media platform and it would be interesting to extend our results to other text-based or photo-based social media platforms. Second, our experiment was built directly on the platform's cutting-edge recommender system, which could underestimate the treatment effect due to spillovers between algorithms. Specifically, when we increased the recommendation diversity to treatment users, those niche and less popular videos would get more impressions and usually receive more clicks and likes. The treatment algorithm could pass this information to the control algorithm through their shared DNN model and induce the control algorithm to increase the recommendation probability for

these videos as well. Since only 3% of the platform users were assigned to the treatment group, we believe the spillover effects should be slight, and our results still hold qualitatively. Third, while we focus on a strategy-agnostic way of improving recommendation diversity, it is also interesting to study how other ways of modifying existing recommender systems to increase diversity could affect users' behaviors. Last, consumer valuation and the algorithm's prediction accuracy are only two possible explanations for users' heterogeneous responses toward a diversity-enhanced recommender system. Other important mechanisms could also be explored in the future.

References

Adomavicius, Gediminas, and YoungOk Kwon. "Improving aggregate recommendation diversity using ranking-based techniques." *IEEE Transactions on Knowledge and Data Engineering* 24.5 (2011a): 896-911.

Adomavicius, Gediminas, and YoungOk Kwon. "Maximizing aggregate recommendation diversity: A graph-theoretic approach." *Proc. of the 1st International Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011)*. 2011b.

Anderson, Ashton, et al. "Algorithmic effects on the diversity of consumption on spotify." *Proceedings of The Web Conference 2020*. 2020.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. "Exposure to ideologically diverse news and opinion on Facebook." Science 348.6239 (2015): 1130-1132.

Berman, Ron, and Zsolt Katona. "Curation algorithms and filter bubbles in social networks." Marketing Science 39.2 (2020): 296-316.

Bradley, Keith, and Barry Smyth. "Improving recommendation diversity." *Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science, Maynooth, Ireland*. Vol. 85. 2001.

Clarke, Charles LA, et al. "Novelty and diversity in information retrieval evaluation." *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008.

Claussen, Jörg, Christian Peukert, and Ananya Sen. "The editor vs. the algorithm: Targeting, data and externalities in online news." *Data and Externalities in Online News (June 5, 2019)* (2019).

Dujeancourt, Erwan, et al. The Effects of Algorithmic Content Selection on User Engagement with News on Twitter. Working Paper, 2021.

Ekstrand, Michael D., et al. "User perception of differences in recommender algorithms." *Proceedings of the 8th ACM Conference on Recommender systems*. 2014.

Fleder, Daniel, and Kartik Hosanagar. "Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity." Management science 55.5 (2009): 697-712.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li. "Examining the impact of ranking on consumer behavior and search engine revenue." *Management Science* 60.7 (2014): 1632-1654.

Holtz, David, et al. "The engagement-diversity connection: Evidence from a field experiment on spotify." *Proceedings of the 21st ACM Conference on Economics and Computation*. 2020.

Hosanagar, Kartik, et al. "Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation." *Management Science* 60.4 (2014): 805-823.

Hu, Rong, and Pearl Pu. "Helping Users Perceive Recommendation Diversity." *DiveRS@ RecSys*. 2011.

Hurley, Neil, and Mi Zhang. "Novelty and diversity in top-n recommendation--analysis and evaluation." *ACM Transactions on Internet Technology (TOIT)* 10.4 (2011): 1-30.

Huszár, Ferenc, et al. "Algorithmic amplification of politics on Twitter." Proceedings of the National Academy of Sciences 119.1 (2022): e2025334119.

Javari, Amin, and Mahdi Jalili. "A probabilistic model to resolve diversity–accuracy challenge of recommendation systems." *Knowledge and Information Systems* 44.3 (2015): 609-627.

Kaminskas, Marius, and Derek Bridge. "Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.1 (2016): 1-42.

Kunaver, Matevž, and Tomaž Požrl. "Diversity in recommender systems–A survey." *Knowledge-based systems* 123 (2017): 154-162.

Lee, Dokyun, and Kartik Hosanagar. "How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment." Information Systems Research 30.1 (2019): 239-259.

Moehring, Alex. News Feeds and User Engagement: Evidence from the Reddit News Tab. Diss. Massachusetts Institute of Technology, 2022.

Pariser, Eli. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

Ribeiro, Manoel Horta, et al. "Auditing radicalization pathways on YouTube." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

Vaishnavi, S., A. Jayanthi, and S. Karthik. "Ranking technique to improve diversity in recommender systems." *International Journal of Computer Applications* 68.2 (2013).

van Dam, Alje. "Diversity and its decomposition into variety, balance and disparity." *Royal Society open science* 6.7 (2019): 190452.

Vargas, Saúl, and Pablo Castells. "Rank and relevance in novelty and diversity metrics for recommender systems." *Proceedings of the fifth ACM conference on Recommender systems*. 2011.

Vargas, Saúl, et al. "Coverage, redundancy and size-awareness in genre diversity for recommender systems." *Proceedings of the 8th ACM Conference on Recommender systems*. 2014.

Ziegler, Cai-Nicolas, et al. "Improving recommendation lists through topic diversification." *Proceedings of the 14th international conference on World Wide Web*. 2005.

Zhou, Renjie, Samamon Khemmarat, and Lixin Gao. "The impact of YouTube recommendation system on video views." Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. 2010.

Appendix A: An example of algorithmic recommendations under different window sizes

We randomly select a Cloud Village user and simulate the algorithm's recommendations for her one request. To illustrate why increasing window size can effectively increase the content diversity of recommendations, we ask the recommender algorithm to retrieve 400 videos, select 50 top-ranked videos based on their predicted match value to the user, and re-rank the 50 videos using window size 5 or 30. [19]

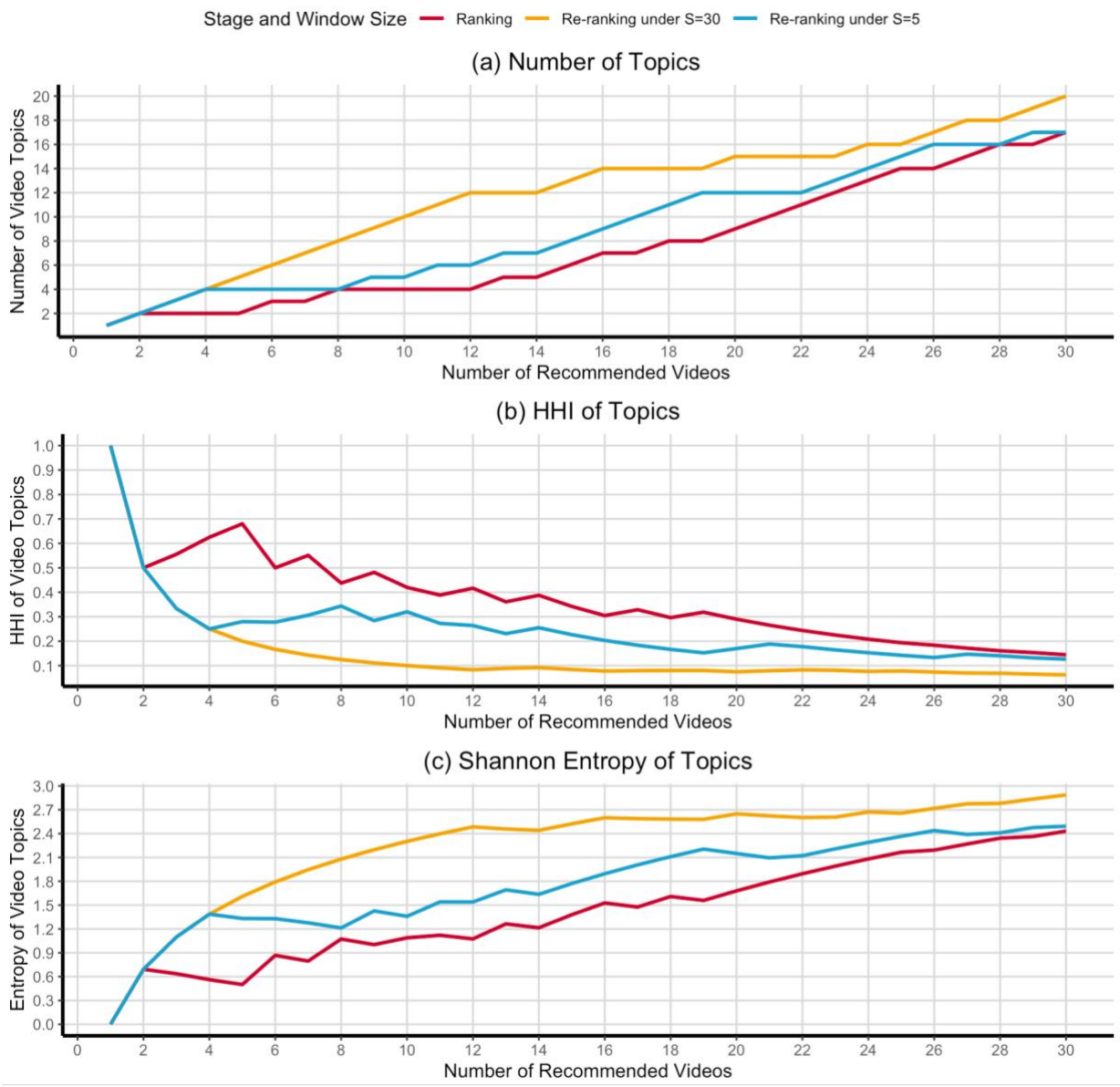
Table A1 shows the relationship between the number of recommended videos and the topic diversity of recommendations in the ranking stage and re-ranking stage with different window sizes.[20] First, in the ranking stage (red line), as the algorithm recommends more videos to the user, the content diversity of recommendations overall increases in terms of topic number, HHI of topics, and Shannon entropy of topics. It implies that the recommender system can retrieve more diverse and unfamiliar video topics for the user, despite giving these videos a lower ranking. Therefore, when we expand the window size, the re-ranking algorithm is more likely to include these more diverse videos in the consideration set and recommend them to the user. To further prove this point, we compare the algorithm's final recommendations after it re-ranks videos using the window size of 5 (blue line) and 30 (orange line). Conditional on recommending the same number of videos, the algorithm under window size 30 indeed selects more diverse videos.[21] For example, suppose the algorithm recommends 10 videos to a user per request. Under window size 5, the algorithm recommends in total 5 different topics with an HHI of 0.32 and Shannon entropy of 1.36. By contrast, under window size 30, the algorithm recommends in total 10 different topics with an HHI of 0.10 and Shannon entropy of 2.30. Therefore, we confirm a larger window size effectively increases the content diversity of recommendations in the Cloud Village.

[19] For simplicity, here we incorporate the re-prediction of user-video match values in the Ranking stage and emphasize only the window size change in the Re-ranking stage.

[20] We assume the algorithm recommends at most 30 videos per user request, which is consistent with the setting in NCM's recommender system.

[21] In this illustrative example, the recommendation diversity under two window sizes is the same if the algorithm recommends four or fewer videos to the user. In reality, the recommender algorithm recommends far more than 4 videos per user's query.

Table A1: The Diversity of Video Topics Under Different Recommendation Stages and Window Sizes



Appendix B: Treatment effect estimation by including the experiment implementation period in the experiment period

We show the estimation effects are qualitatively consistent if we include the experiment implementation period in the experiment period. Specifically, we split the data into two periods: (1) the pre-experiment period, including four weeks before the experiment started (November 19 to December 16, 2021), (2) the experiment period, totaling fourteen weeks after December 17, 2021 (December 18, 2021, to March 25, 2022). Among the sampled 10 million users, we filter 9,821,066 users who had visited the Cloud Village during the experiment period, including 485,157 (4.94%) new users, 8,989,773 (91.54%) inactive users, and 346,136 (3.52%) active users. Table B1 shows the summary statistics of regression variables.

Table B1: Variable Description and Summary Statistics

Variables	Description	Number of Observations	Mean	St. Dev.	Min	Max
Treatment	1 if a user was assigned into the treatment group, 0	9,821,066	0.0319	0.1758	0	1
New_user	1 if a user registered on the music app after the pre-experiment period, 0	9,821,066	0.0494	0.2167	0	1
Num_registered month	Number of months a user had been registered on the app by the end of the experiment	9,821,066	40.5598	22.0299	0.0000	108.4000
Male	1 if a user was predicted to be male, 0	9,719,427	0.5345	0.4988	0	1
Age	A user's predicted age	9,577,479	23.0285	5.8925	11	45
Visit	1 if a user visited the Cloud Village in a week, 0	137,494,924	0.1953	0.3964	0	1
Freq_click	Frequency of days per week that a user watched at least one video for no less than five seconds	137,494,924	0.0022	0.0244	0.0000	1.0000
Num_click	Number of clicks per week with more-than-5-second view time	137,494,924	0.2609	8.9135	0	4,219
View_min	Minutes spent watching videos per week	137,494,924	0.2120	7.0225	0.0000	4,730.8230
Num_like	Number of likes left per week	137,494,924	0.0091	0.9462	0	4,267
Num_comment	Number of comments left per week	137,494,924	0.0002	0.0398	0	142
Num_share	Number of shares left per week	137,494,924	0.0004	0.0486	0	195
Num_recommended_topic	Number of different topics recommended to a user per week when visiting the Cloud Village	26,846,235	8.2227	5.2693	1	79
HHI_recommended_topic	HHI of video topics recommended to a user per week when visiting the Cloud Village	26,846,235	0.1758	0.0749	0.0302	1.0000
Entropy_recommended_topic	Shannon entropy of video topics recommended to a user per week when visiting the Cloud Village	26,846,235	1.8927	0.4352	0.0000	3.7212
Num_clicked_topic	Number of different topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,617,598	5.6072	7.6738	1	65
HHI_clicked_topic	HHI of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,617,598	0.5797	0.3708	0.0386	1.0000
Entropy_clicked_topic	Shannon entropy of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,617,598	0.9227	0.9327	0.0000	3.4800

Notes. Summary statistics for 9,821,066 Cloud Village visitors during the experiment. Users' age and gender were predicted by a supervised machine learning model built by NCM. Among the sampled users, 1.03% (101,639) did not have predicted gender, and 1.65% (162,327) did not have predicted age. To remove age outliers, we keep only users in the $0.5\% - 99.5\%$ quantile of the age distribution ($11 - 45$ years old).

As Table B2 and B3 show, the randomization check also passes since there is no significant difference between treatment and control users in their demographics, pre-experiment behaviors, or pre-experiment recommendation diversity.

Table B2: User Demographics at the End of the Experiment Between Control and Treatment Groups

	Dependent variable:			
	new user	*num_registered month*	*male*	*age*
Treatment	0.00001	−0.0195	0.0003	0.0138
	(0.0004)	(0.0400)	(0.0009)	(0.0108)
Constant	0.0494***	40.5604***	0.5345***	23.0281***
	(0.0001)	(0.0071)	(0.0002)	(0.0019)
Observations	9,821,066	9,821,066	9,719,427	9,577,479
R^2	0.0000	0.000000	0.0000	0.000000
Note:				*p<0.1; **p<0.05; ***p<0.01

Table B3: Users' Retention, Consumption, Engagement, and Diversity of Recommendations and Consumption During the Pre-experiment Period

(a) Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0001	0.00002	0.0129	0.0074	−0.0008	−0.00001	−0.00002
	(0.0005)	(0.00004)	(0.0126)	(0.0099)	(0.0005)	(0.00002)	(0.00004)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	37,207,613	37,207,613	37,207,613	37,207,613	37,207,613	37,207,613	37,207,613
R^2	0.00004	0.000002	0.000001	0.000001	0.000000	0.000000	0.000000

(b) Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.0167	−0.0002	0.0012
	(0.0137)	(0.0002)	(0.0012)
Week FE	Yes	Yes	Yes
Observations	6,343,722	6,343,722	6,343,722
R^2	0.0008	0.0015	0.0018

(c) Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1370	−0.0031	0.0115
	(0.0932)	(0.0040)	(0.0106)
Week FE	Yes	Yes	Yes
Observations	381,491	381,491	381,491
R^2	0.0007	0.0001	0.0002

Notes. We exclude new users who registered on the music app after the pre-experiment period. Standard errors clustered at individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table B4 shows the average treatment effect of recommendation diversity all sampled users. Consistent with the results in Section 4.1, despite 2.27% ($p < 0.0001$) increased recommendation diversity in HHI, our new algorithm did not significantly affect an average user's consumption diversity in HHI ($p = 0.3371$) but reduced her clicking frequency by 3.02% ($p = 0.0175$).

Table B4: The Manipulation Check and Average Treatment Effects of Recommendation Diversity

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1191***	−0.0040***	0.0190***
	(0.0113)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	26,846,235	26,846,235	26,846,235
R^2	0.0019	0.0054	0.0044

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0002	−0.00007**	0.0139	0.0116	0.0008	0.0001	0.00002
	(0.0003)	(0.0001)	(0.0118)	(0.0095)	(0.0010)	(0.0001)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	137,494,924	137,494,924	137,494,924	137,494,924	137,494,924	137,494,924	137,494,924
R^2	0.0027	0.0001	0.00001	0.00001	0.000002	0.000001	0.000001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1770**	−0.0024	0.0136*
	(0.0795)	(0.0025)	(0.0073)
Week FE	Yes	Yes	Yes
Observations	1,617,598	1,617,598	1,617,598
R^2	0.0014	0.0010	0.0014

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Regarding the heterogeneous treatment effects of recommendation diversity, Table B5 confirms that more recommendation diversity did not affect new users' short-term metrics or consumption diversity. Table B6 show that, facing 2.14% increased recommendation diversity in HHI ($p < 0.0001$), inactive users significantly reduced their weekly clicking frequency by 5.51% ($p < 0.0001$) without consuming more diversely. A 1% decrease in recommendation HHI translates into a 2.57% drop in inactive users' clicking frequency. Table B7 also confirms that more recommendation diversity increased active users' consumption diversity without reducing their short-term metrics. A 5.00% ($p < 0.0001$) decrease in recommendation HHI translated into a 2.30% ($p = 0.0355$) decrease in their consumption HHI. Therefore, all results are qualitatively consistent with Section 4.2.

Table B5: The Manipulation Check and Treatment Effects of Recommendation Diversity for New Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1643***	−0.0036***	0.0191***
	(0.0547)	(0.0005)	(0.0035)
Week Fixed Effects	Yes	Yes	Yes
Observations	926,116	926,116	926,116
R^2	0.0018	0.0106	0.0054

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0009	0.00001	0.0236	0.0259	0.0129	0.0006	0.0002
	(0.0009)	(0.0001)	(0.0338)	(0.0310)	(0.0096)	(0.0007)	(0.0002)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,792,198	6,792,198	6,792,198	6,792,198	6,792,198	6,792,198	6,792,198
R^2	0.0141	0.0008	0.0001	0.0001	0.00002	0.00002	0.00002

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3152	−0.0089	0.0285
	(0.3026)	(0.0103)	(0.0285)
Week FE	Yes	Yes	Yes
Observations	65,478	65,478	65,478
R^2	0.0027	0.0026	0.0030

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Table B6: The Manipulation Check and Treatment Effects of Recommendation Diversity for Inactive Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1011***	−0.0038***	0.0174***
	(0.0082)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	24,296,402	24,296,402	24,296,402
R^2	0.0024	0.0057	0.0047

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00008***	−0.0056	−0.0020	−0.0003	0.00004	0.00002
	(0.0003)	(0.00001)	(0.0050)	(0.0047)	(0.0005)	(0.00003)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	125,856,822	125,856,822	125,856,822	125,856,822	125,856,822	125,856,822	125,856,822
R^2	0.0029	0.0002	0.00003	0.00003	0.000004	0.000002	0.000001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.0010	0.0030	−0.0024
	(0.0547)	(0.0024)	(0.0063)
Week FE	Yes	Yes	Yes
Observations	1,136,224	1,136,224	1,136,224
R^2	0.0027	0.0022	0.0025

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Table B7: The Manipulation Check and Treatment Effects of Recommendation Diversity for Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.3258***	−0.0080***	0.0398***
	(0.1154)	(0.0005)	(0.0044)
Week Fixed Effects	Yes	Yes	Yes
Observations	1,623,717	1,623,717	1,623,717
R^2	0.0020	0.0079	0.0078

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0001	0.00002	0.4649	0.3124	0.0090	−0.0004	−0.0004
	(0.0022)	(0.0007)	(0.2984)	(0.2324)	(0.0206)	(0.0005)	(0.0006)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,845,904	4,845,904	4,845,904	4,845,904	4,845,904	4,845,904	4,845,904
R^2	0.0083	0.0011	0.0003	0.0003	0.00004	0.00001	0.00003

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.4801**	−0.0095**	0.0361**
	(0.2164)	(0.0045)	(0.0153)
Week FE	Yes	Yes	Yes
Observations	415,896	415,896	415,896
R^2	0.0021	0.0011	0.0017

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Last, to confirm the results of mechanism analysis are also consistent with Section 5, we filter 1,988,668 active users who had visited the Cloud Village from December 18, 2021, to March 25, 2022, among the 2 million sampled active users. Following the main analysis, we classify these active users into four segments by their 80% quantile of view time (10.46 minutes) and clicking number (11) during the pre-experiment period: high-valuation high-accuracy users (17.66%), high-valuation low-accuracy users (2.34%), low-valuation high-accuracy users (2.75%), and low-valuation low-accuracy users (77.25%).

Table B8-B11 show the estimation results are qualitatively consistent with the main paper. Only for high-valuation high-accuracy users, more recommendation diversity encouraged them to consume more diversely without reducing short-term metrics. The 8.79% decreased recommendation HHI ($p < 0.0001$) led to 2.65% decreased consumption HHI ($p < 0.0001$), translating into an elasticity of 0.30. For high-valuation low-accuracy users and low-valuation low-accuracy users, more diversified recommendations did not affect their consumption diversity but significantly reduced their short-term metrics. Facing a 1% decrease in recommendation HHI, high-valuation low-accuracy users clicked 2.10% less frequently ($p = 0.0087$), viewed 3.43% less time ($p = 0.0398$), and left 8.64% fewer likes ($p < 0.0001$) and 11.01% fewer comments ($p = 0.0002$); low-valuation low-accuracy users clicked 0.60% less frequently ($p = 0.0455$). For low-valuation high-accuracy, although higher recommendation diversity marginally lifted their

consumption diversity with an elasticity of 0.61 in HHI (p = 0.0595), it significantly reduced their visiting probability with an elasticity of 0.80 in HHI (p = 0.0036). [22]

Table B8: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.6129***	−0.0129***	0.0673***
	(0.1369)	(0.0004)	(0.0042)
Week FE	Yes	Yes	Yes
Observations	2,157,064	2,157,064	2,157,064
R^2	0.0029	0.0162	0.0119

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0017	0.0010	0.8081	0.6647	0.0160	−0.0014	−0.0026*
	(0.0026)	(0.0013)	(0.5758)	(0.4637)	(0.0443)	(0.0015)	(0.0015)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,916,058	4,916,058	4,916,058	4,916,058	4,916,058	4,916,058	4,916,058
R^2	0.0114	0.0052	0.0021	0.0021	0.0003	0.00003	0.0001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3859***	−0.0079***	0.0350***
	(0.1299)	(0.0019)	(0.0075)
Week FE	Yes	Yes	Yes
Observations	1,246,870	1,246,870	1,246,870
R^2	0.0027	0.0020	0.0027

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Table B9: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	−0.0351	−0.0096***	0.0373***
	(0.1980)	(0.0013)	(0.0100)
Week FE	Yes	Yes	Yes
Observations	214,018	214,018	214,018
R^2	0.0018	0.0073	0.0072

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0005	−0.0021***	−0.1889	−0.3183**	−0.0251***	−0.0015***	0.0017
	(0.0060)	(0.0008)	(0.1828)	(0.1548)	(0.0048)	(0.0004)	(0.0018)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	652,218	652,218	652,218	652,218	652,218	652,218	652,218
R^2	0.0078	0.0009	0.0001	0.0001	0.00004	0.00002	0.0001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.2063	0.0120	−0.0284
	(0.2696)	(0.0107)	(0.0305)
Week FE	Yes	Yes	Yes
Observations	57,421	57,421	57,421
R^2	0.0041	0.0014	0.0023

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

[22] The experiment decreased the recommendation HHI by 5.75% for high-valuation low-accuracy users, 4.01% for low-valuation low-accuracy users, and 5.71% for low-valuation high-accuracy users (all p-values < 0.0001).

Table B10: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.4626*	−0.0089***	0.0468***
	(0.2382)	(0.0012)	(0.0103)
Week FE	Yes	Yes	Yes
Observations	272,165	272,165	272,165
R^2	0.0016	0.0069	0.0065

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*				*engagement*	
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0163***	−0.0003	0.3525	0.2718	−0.0196	−0.0010*	−0.0009
	(0.0056)	(0.0012)	(0.3981)	(0.3105)	(0.0186)	(0.0006)	(0.0010)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	766,108	766,108	766,108	766,108	766,108	766,108	766,108
R^2	0.0084	0.0014	0.0002	0.0002	0.0001	0.00002	0.0001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3419	−0.0147*	0.0395
	(0.2865)	(0.0078)	(0.0246)
Week FE	Yes	Yes	Yes
Observations	95,250	95,250	95,250
R^2	0.0050	0.0009	0.0020

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Table B11: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1930***	−0.0066***	0.0306***
	(0.0301)	(0.0002)	(0.0017)
Week FE	Yes	Yes	Yes
Observations	6,675,641	6,675,641	6,675,641
R^2	0.0028	0.0072	0.0083

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*				*engagement*	
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0010	−0.0002**	0.0202	0.0139	−0.0019	0.0003	−0.00003
	(0.0010)	(0.0001)	(0.0313)	(0.0243)	(0.0018)	(0.0002)	(0.0001)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	21,506,968	21,506,968	21,506,968	21,506,968	21,506,968	21,506,968	21,506,968
R^2	0.0078	0.0003	0.00004	0.00003	0.00001	0.00001	0.000005

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.0917	0.0008	0.0049
	(0.0694)	(0.0028)	(0.0076)
Week FE	Yes	Yes	Yes
Observations	991,250	991,250	991,250
R^2	0.0037	0.0016	0.0024

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Appendix C: Regression specifications for randomization check

We compare treatment and control users' demographics using the following regression specification:

$$Demographics_i = \gamma_1 \cdot Treatment_i + \eta_i, \quad (C1)$$

where $Demographics_i \in \{new\ user_i, num\_registered\ month_i, male_i, age_i\}$ is detailed later. $Treatment_i$ is a binary variable that equals 1 if user *i* was in the treatment condition and 0 otherwise. η_i is the individual-specific error term. We define new users as those who registered on the app after the pre-experiment period ($new\ user_i$) and calculate each user i's number of registered months till the end of the experiment ($num\_registered\ month_i$). Each user i's age and gender are predicted by a well-trained supervised, deep learning model of NCM. $male_i$ is 1 if the user is predicted to be male and 0 otherwise.

For users who had registered before the experiment implementation period, we also checked if the treatment and control group behaved differently or were recommended videos of different diversity levels during the pre-experiment period. The regression model is specified as follows:

$$Outcome\ Variable_{it} = \lambda_1 \cdot Treatment_i + \theta_t + \tau_{it}, \quad (C1)$$

where $Outcome\ Variable_{it}$ for user i in week t is detailed later. $Treatment_i$ is a binary variable indicating whether user *i* was in the treatment (vs. control) condition. θ_t represents the week fixed effects and we cluster the error term τ_{it} at the individual level.

The outcome variables consist of three sets of measures: the topic diversity of recommended videos when user *i* visited Cloud Village in week *t* including the number ($num\_recommended\_topic_{it}$), the HHI ($HHI\_recommended\_topic_{it}$), and the Shannon entropy of recommended topics ($entropy\_recommended\_topic_{it}$); users' weekly retention, consumption, and engagement levels including whether they visited the Cloud Village ($visit_{it}$), the frequency of days having clicks ($freq\_click_{it}$), the number of clicks (num_click_{it}), the number of minutes spent on watching the videos ($view\_min_{it}$), and the likes (num_click_{it}), comments ($num\_comment_{it}$), and shares (num_share_{it}) left; user i's consumption diversity when she clicked on videos in week t including the number ($num\_clicked\_topic_{it}$), the HHI ($HHI\_clicked\_topic_{it}$), and the Shannon entropy of clicked topics ($entropy\_clicked\_topic_{it}$).

Appendix D: NCM's field implementation

NCM ran an extra field experiment to test the benefit of increasing active users' recommendation diversity. The experiment took two weeks, from March 17, 2022, to March 30, 2022. Users of the control group were recommended by the original recommender algorithm where the window size is 5 if users had clicks in the past 30 days and 15 otherwise. For the treatment group, NCM increased the window size to 15 for users having clicks in the past 30 days and decreased the window size to 1 for the other users. Consistent with our main analysis, we define the active users as those having viewed videos four weeks before the experiment (i.e., from February 16, 2022, to March 15, 2022). Thus, active users of the treatment group would face higher recommendation diversity than those of the control group. We randomly sample one million active users and show the experiment results in Table D1. The manipulation check results confirm that treated active users were recommended 1.35% more video topics ($p = 0.0126$). Their recommendation HHI decreased by 2.69% ($p < 0.0001$) and entropy increased by 0.94% ($p<0.0001$).

Table D1(b) and D1(c) show that the treatment significantly increased both active users' consumption level and their consumption diversity without hurting their retention or engagement. Specifically, treated active users clicked 0.74 more videos ($p = 0.0176$, an 8.07% increase) and spent 33 more seconds watching videos ($p = 0.0293$, an 7.27% increase) every week. It implies a 1% decrease in recommendation HHI resulted in 2.99% more clicks and 2.70% longer view time. More importantly, the treatment also encouraged active users to click 0.46 more topics ($p=0.0001$, a 4.83% increase), decrease their consumption HHI by 2.19% ($p = 0.0052$), and increase their consumption entropy by 2.47% ($p = 0.0005$). Equivalently, a 1% decrease in recommendation HHI led to a 0.81% decrease in consumption HHI.

Table D1: The Manipulation Check and Treatment Effects of Recommendation Diversity for Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1690**	−0.0046***	0.0192***
	(0.0677)	(0.0004)	(0.0030)
Week FE	Yes	Yes	Yes
Observations	1,369,357	1,369,357	1,369,357
R^2	0.0014	0.0003	0.0011

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0002	0.0003	0.7356**	0.5496**	0.0166	0.0015	0.0012
	(0.0014)	(0.0007)	(0.3100)	(0.2521)	(0.0238)	(0.0011)	(0.0019)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000
R^2	0.0016	0.0001	0.00005	0.00002	0.000001	0.000003	0.000000

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.4596***	−0.0092***	0.0339***
	(0.1197)	(0.0033)	(0.0098)
Week FE	Yes	Yes	Yes
Observations	381,626	381,626	381,626
R^2	0.0004	0.0002	0.0003

Notes. Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Appendix E: Comparing two measures of the recommender algorithm's prediction accuracy for user preference

We consider two possible indicators to measure the algorithm's familiarity with a user: the number of clicks and the number of impressions a user had during the pre-experiment period. To find the more accurate indicator, we test which one could better predict a user's future clicking-through rate. Specifically, among 2 million active users, we selected 203,263 high-valuation users in the control group, who spent at least 10.42 minutes watching videos in the pre-experiment period and visited the Cloud Village in the first week of the experiment. Then we predicted these users' click-through rate (i.e., the number of clicks/the number of impressions) in the first experiment week by their standardized click (or impression) number during the pre-experiment period. Table E1 shows that the click number has a higher correlation with users' future click-through rate than the impression number (0.0726 vs. 0.0250). Consistently, the click number better predicts future click-through rate in terms of R square (0.1390 vs. 0.0164). Thus, we use the click number to measure the algorithm's prediction accuracy.

Table E1: Correlation Between Click-through Rate and Indicators of Algorithm's Prediction Accuracy

	Dependent variable:		
	click-through rate in the first week of the experiment		
	(1)	(2)	(3)
standardized number of impressions in the pre-experiment period	0.0250***		−0.0266***
	(0.0004)		(0.0005)
standardized number of clicks in the pre-experiment period		0.0726***	0.0881***
		(0.0004)	(0.0005)
Constant	0.1439***	0.1439***	0.1439***
	(0.0004)	(0.0004)	(0.0004)
Observations	203,263	203,263	203,263
R^2	0.0164	0.1390	0.1512
Note:			*p<0.1; **p<0.05; ***p<0.01

Appendix F: Regression tables for each segment of active users

Table F1: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.6180***	−0.0134***	0.0690***
	(0.1387)	(0.0004)	(0.0043)
Week FE	Yes	Yes	Yes
Observations	2,126,826	2,126,826	2,126,826
R^2	0.0031	0.0164	0.0122

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0013	0.0008	0.7386	0.6262	0.0178	−0.0014	−0.0027*
	(0.0027)	(0.0013)	(0.5690)	(0.4594)	(0.0434)	(0.0015)	(0.0014)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,938,584	4,938,584	4,938,584	4,938,584	4,938,584	4,938,584	4,938,584
R^2	0.0102	0.0043	0.0020	0.0019	0.0003	0.00002	0.0001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3830***	−0.0080***	0.0342***
	(0.1313)	(0.0020)	(0.0076)
Week FE	Yes	Yes	Yes
Observations	1,219,555	1,219,555	1,219,555
R^2	0.0022	0.0014	0.0018

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Table F2: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	−0.0216	−0.0102***	0.0410***
	(0.1999)	(0.0014)	(0.0102)
Week FE	Yes	Yes	Yes
Observations	213,936	213,936	213,936
R^2	0.0025	0.0097	0.0093

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*			*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0002	−0.0022***	−0.1894	−0.3205**	−0.0261***	−0.0015***	0.0018
	(0.0059)	(0.0008)	(0.1840)	(0.1570)	(0.0046)	(0.0003)	(0.0018)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	661,444	661,444	661,444	661,444	661,444	661,444	661,444
R^2	0.0072	0.0007	0.0001	0.0001	0.00004	0.00003	0.0001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.2510	0.0174	−0.0407
	(0.2767)	(0.0110)	(0.0313)
Week FE	Yes	Yes	Yes
Observations	56,874	56,874	56,874
R^2	0.0033	0.0013	0.0020

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Table F3: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.4985**	−0.0092***	0.0485***
	(0.2425)	(0.0013)	(0.0106)
Week FE	Yes	Yes	Yes
Observations	266,290	266,290	266,290
R^2	0.0018	0.0089	0.0078

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*				*engagement*	
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0163***	−0.0002	0.4060	0.3240	−0.0189	−0.0010*	−0.0008
	(0.0056)	(0.0012)	(0.4118)	(0.3256)	(0.0189)	(0.0006)	(0.0011)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	764,036	764,036	764,036	764,036	764,036	764,036	764,036
R^2	0.0079	0.0012	0.0002	0.0002	0.0001	0.00002	0.0001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.4706	−0.0146*	0.0463*
	(0.3030)	(0.0081)	(0.0256)
Week FE	Yes	Yes	Yes
Observations	92,485	92,485	92,485
R^2	0.0038	0.0008	0.0016

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.

Table F4: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.2042***	−0.0069***	0.0322***
	(0.0306)	(0.0002)	(0.0017)
Week FE	Yes	Yes	Yes
Observations	6,596,126	6,596,126	6,596,126
R^2	0.0037	0.0111	0.0112

(b) Treatment Effect: Retention, Consumption, and Engagement

	retention	*consumption*				*engagement*	
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.0002**	0.0217	0.0151	−0.0022	0.0002	−0.00002
	(0.0010)	(0.0001)	(0.0323)	(0.0251)	(0.0018)	(0.0002)	(0.0001)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	21,635,936	21,635,936	21,635,936	21,635,936	21,635,936	21,635,936	21,635,936
R^2	0.0071	0.0003	0.00004	0.00003	0.00001	0.00001	0.00001

(c) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.0779	0.0022	0.0015
	(0.0714)	(0.0028)	(0.0077)
Week FE	Yes	Yes	Yes
Observations	975,996	975,996	975,996
R^2	0.0027	0.0013	0.0019

Notes: Standard errors clustered at individual level in parentheses; *p<0.1; **p<0.05; ***p<0.01.