The Monetization Agent: A Deployed POMDP for Maximizing Lifetime In-App Advertising Revenue

Extended Abstract

Jiacheng Chang
The University of Hong Kong
Hong Kong, China
jc5506@connect.hku.hk

Xiao Lei
The University of Hong Kong
Hong Kong, China
xlei@hku.hk

Zhixi Wan
The University of Hong Kong
Hong Kong, China
zhixiwan@hku.hk

Lei Huang
Tencent
shanghai, China
leihuang@tencent.com

Qi He
Tencent
shanghai, China
nickyhe@tencent.com

ABSTRACT

The mobile gaming industry is increasingly reliant on in-app advertising (IAA) revenue, with rewarded ads as a common model: players voluntarily watch an ad in exchange for an in-game item (e.g., lowering difficulty). This creates a key intertemporal trade-off—raising difficulty can drive more ad views and short-term revenue, but may also frustrate players, increase churn, and reduce lifetime value (LTV). The challenge is compounded by strong player heterogeneity and very high early churn, making slow-learning methods ineffective. This paper presents *The Monetization Agent*, a deployed decision-making system that maximizes lifetime rewarded-ad revenue via dynamic difficulty adjustment (DDA). We formulate DDA as a partially observable Markov decision process (POMDP) with latent player types capturing heterogeneous skill and ad tolerance. To enable data-efficient learning and interpretability, we instantiate the POMDP with a structural econometric model of player behavior that links difficulty to win probability, ad-watching, and churn. We develop a practical estimation pipeline and an offline Monte Carlo planning approach that yields a compact policy suitable for real-time deployment. In two large-scale online A/B tests on production games, the agent increases lifetime ad impressions by 23.6% and 7.0%, while improving mid-to-late lifecycle retention (survival at level 300) by up to 204% and 297%.[1]

KEYWORDS

POMDP; Digital Advertising; Dynamic Difficulty Adjustment; Structural Econometrics; Real-World Deployment

ACM Reference Format:
Jiacheng Chang, Xiao Lei, Zhixi Wan, Lei Huang, and Qi He. 2026. The Monetization Agent: A Deployed POMDP for Maximizing Lifetime In-App Advertising Revenue: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 2 pages. https://doi.org/10.65109/WOQF1299

[1]The full paper can be found at https://ssrn.com/abstract=6137709

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/WOQF1299

1 PROBLEM INTRODUCTION

The mobile gaming industry increasingly relies on Rewarded In-App Advertising (IAA) as a primary revenue stream. In this model, players voluntarily watch advertisements to gain in-game advantages, such as extra moves or retries. This mechanic introduces a fundamental intertemporal trade-off for experience management: increasing game difficulty can induce immediate ad impressions (short-term reward), but excessive challenge frustrates players, accelerating churn and destroying Lifetime Value (LTV) [7].

While Dynamic Difficulty Adjustment (DDA) has been extensively studied to maximize player engagement or maintain a state of "flow" [2, 3, 5, 7–10], optimizing for LTV presents unique challenges. First, this objective is fundamentally more complex because it introduces a sharper trade-off: the action needed for revenue (increasing difficulty) is also the primary driver of churn. Second, the player base exhibits extreme heterogeneity in skill and ad-tolerance, which are latent traits unobservable at registration. Third, the free-to-play market is characterized by severe early-stage churn—often exceeding 80% within the first two days. In this high-stakes, short-horizon environment, standard Reinforcement Learning (RL) methods [1, 4, 6] are often too sample-inefficient, and Multi-Armed Bandits (MAB) incur exploration costs that result in unacceptable user loss.

2 POMDP FORMULATION AND SOLUTION

To address these challenges, we propose a monetization agent that models the interaction with a player as a POMDP. The hidden state includes a latent player type (capturing skill and ad tolerance), while the observable state tracks game progress (current level and trial count). At each trial, the agent selects a difficulty action corresponding to a random seed from a finite action set. The agent then observes the trial outcome (win/loss), the rewarded-ad decision, and whether the player churns (terminal transition). The immediate reward is the ad impression indicator, and the objective is to maximize expected total reward, i.e., lifetime ad impressions.

A key challenge is specifying and estimating an accurate, deployable model of how actions influence outcomes under limited

Metric	Game	Level 1	Level 300	Level 1000
Total ads	A	-11.4%	+18.9%	+23.6%
	B	-77.0%	+3.1%	+7.0%
Retention (survival)	A	+5.4%	+204.9%	N/A
	B	+32.8%	+297.3%	N/A

Table 1: Treatment vs. control improvement rates at key milestones, computed as (Treatment/Control −1).

data. We address this by integrating a structural econometric model into the POMDP transition and reward components. For each latent type, we parameterize (via logistic models) the probability of winning, watching an ad, and churning as functions of various independent variables. This structure provides two practical benefits: (i) interpretability of behavioral mechanisms and (ii) data efficiency, enabling estimation from moderately sized experimental datasets.

To estimate the model parameters. We show that direct joint MLE is non-convex, so we use a two-step procedure that exploits convexity once key hyperparameters are fixed. We first fit segmented win models and recover seed-level difficulty via EM on win/loss outcomes; conditioning on the estimated difficulties, we then fit ad-watching and churn using weighted (convex) logistic regressions. Hyperparameters (e.g., reference-difficulty dynamics) are chosen by grid search maximizing the joint likelihood, with model size controlled by information criteria and some business constraints.

Solving the resulting POMDP online is infeasible under strict time constraints and a large state space. We therefore compute the policy offline using Partially Observable Monte Carlo Planning (POMCP) with the learned model. We evaluate optimal actions at a discretized set of belief points over latent types and store the resulting mapping as a compact lookup table. The final policy requires only fast belief updates and table lookup at runtime, and fits within a few megabytes, enabling on-device deployment

3 ONLINE EXPERIMENTS AND RESULTS

We validate the Monetization Agent in two large randomized controlled field experiments with industrial partners (Game A and Game B). New users are randomly assigned to a control group (Original difficulty design) or a treatment group (agent-managed seeds for the early part of gameplay; afterwards reverting to the same mechanism as control). The main results are shown in Table 1. As demonstrated, the agent learns a conservative early strategy that reduces immediate ad pressure to mitigate churn and to infer player type in both games; this can temporarily decrease early ad impressions. Nevertheless, the long-run effect is strongly positive: lifetime total ad impressions increase by 23.6% in Game A and 7.0% in Game B at long-run milestones, while retention improvements compound over time, reaching up to 204% and 297% higher survival at level 300, respectively.

The learned policy can be understood as a two-stage strategy. Under high initial uncertainty and severe early-churn penalties, the agent "invests" by assigning easier difficulties to keep users engaged and gather informative behavioral signals. As beliefs concentrate, the policy shifts to "cultivation": it increases difficulty for ad-receptive players to induce rewarded ads with controlled churn risk, while protecting ad-averse or frustration-sensitive users from excessive difficulty spikes. This belief-conditioned personalization reconciles monetization and experience goals in a principled sequential framework.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Tencent Marketing Solution Rhino-Bird Focused Research Program. We also extend our sincere gratitude to their industry partners, Shenzhen Yuren Technology Co., Ltd (the developer of Game A) and See Game (the developer of Game B). Their collaboration was instrumental to this project, providing the live game environments for our large-scale experiments and offering invaluable support throughout the research process. We also thank Canbin Hong, Peng Tan and Ling Yu from Tencent for their support to this project.

REFERENCES

[1] Himanshu Gupta, Bradley Hayes, and Zachary Sunberg. 2022. Intention-Aware Navigation in Crowds with Extended-Space POMDP Planning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 562–570.

[2] Johan Hagelback and Stefan J Johansson. 2009. Measuring player experience on runtime dynamic difficulty scaling in an RTS game. In *2009 IEEE Symposium on Computational Intelligence and Games*. IEEE, 46–52.

[3] Martin Jennings-Teats, Gillian Smith, and Noah Wardrip-Fruin. 2010. Polymorph: dynamic difficulty adjustment through level generation. In *Proceedings of the 2010 Workshop on Procedural Content Generation in Games*. 1–4.

[4] Saaduddin Mahmud, Marcell Vazquez-Chanlatte, Stefan Witwicki, and Shlomo Zilberstein. 2024. Explaining the behavior of pomdp-based agents through the impact of counterfactual information. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 1346–1354.

[5] Elinga Pagalyte, Maurizio Mancini, and Laura Climent. 2020. Go with the flow: Reinforcement learning in turn-based battle video games. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*. 1–8.

[6] Aditya Shinde, Prashant Doshi, and Omid Setayeshfar. 2021. Cyber attack intent recognition and active deception using factored interactive pomdps. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 1200–1208.

[7] Su Xue, Meng Wu, John Kolen, Navid Aghdaie, and Kazi A Zaman. 2017. Dynamic difficulty adjustment for maximized engagement in digital games. In *Proceedings of the 26th international conference on world wide web companion*. 465–471.

[8] Yaqian Zhang and Wooi-Boon Goh. 2021. Personalized task difficulty adaptation based on reinforcement learning. *User Modeling and User-Adapted Interaction* 31, 4 (2021), 753–784.

[9] Alexander Zook, Stephen Lee-Urban, Michael R Drinkwater, and Mark O Riedl. 2012. Skill-based mission generation: A data-driven temporal player modeling approach. In *Proceedings of the The third workshop on Procedural Content Generation in Games*. 1–8.

[10] Alexander Zook and Mark Riedl. 2012. A temporal data-driven player model for dynamic difficulty adjustment. In *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment*, Vol. 8. 93–98.