Optimal Hybrid Feedback-Driven Learning for Wireless Interactive Panoramic Scene Delivery

Xiaoyi Wu* Juaren Steiger* Bin Li* R. Srikant†
*Department of EE, The Pennsylvania State University, University Park, PA, USA
†Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT

Immersive technologies, such as virtual and augmented reality, demand high framerate, low latency, and precise synchronization between real and virtual environments. To meet these requirements, an edge server typically needs to perform high-quality rendering, and must predict user head motion and transmit a portion of the rendered panoramic scene that is large enough to cover the user's viewport, yet small enough to satisfy bandwidth constraints. Each portion yields two feedback signals: prediction feedback, indicating whether the selected portion covers the actual viewport, and transmission feedback, indicating whether all data packets are successfully delivered. While prior work models this setting as a multi-armed bandit with two-level bandit feedback, it overlooks that prediction feedback can be retrospectively computed for all possible portions, thus providing full-information feedback. In this work, we introduce a new two-level feedback model that combines full-information feedback with bandit feedback, and we formulate the portion selection problem as an online learning task under this hybrid setting. We derive an instance-dependent regret lower bound for this new hybrid feedback setting, and we propose AdaPort, a hybrid learning algorithm that leverages both the full-information feedback and bandit feedback to improve learning efficiency. We then show that the instance-dependent regret upper bound for AdaPort matches the lower bound asymptotically, proving its asymptotic optimality. Simulations using synthetic data and real-world traces demonstrate that AdaPort consistently outperforms state-of-the-art baselines, validating the benefits of exploiting the hybrid feedback structure.

CCS CONCEPTS

• **Theory of computation** → **Multi-armed bandit**; • **Networks** → **Network algorithms**;

KEYWORDS

Multi-Armed Bandit, Virtual Reality, Panoramic Scene Delivery, Full-Information Feedback, Two-Level Feedback

MobiHoc '25, October 27–30, 2025, Houston, TX, USA

ACM ISBN 979-8-4007-1353-8/25/10…$15.00
https://doi.org/10.1145/3704413.3764474

ACM Reference Format:
Xiaoyi Wu* Juaren Steiger* Bin Li* R. Srikant†, *Department of EE, The Pennsylvania State University, University Park, PA, USA, †Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL, USA, . 2025. Optimal Hybrid Feedback-Driven Learning for Wireless Interactive Panoramic Scene Delivery. In *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '25), October 27–30, 2025, Houston, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3704413.3764474

1 INTRODUCTION

The evolution of immersive technologies, including extended reality (XR), which includes both virtual reality (VR) and augmented reality (AR), holographic displays, and volumetric capture, is redefining digital interaction across multiple domains, such as gaming, telepresence, remote collaboration, and immersive training. These technologies enable users to experience highly interactive and dynamic virtual environments that adapt to their movements and perspectives in real time. To fully immerse the user in the experience, panoramic scenes are typically delivered to the user wirelessly from an edge server via a head-mounted display (HMD), such as a VR headset. This is because HMDs need to be lightweight and therefore are not typically equipped with the hardware necessary to render high-quality 3D images. Unlike conventional video streaming, when experiencing a panoramic scene, there is a possibility that the user experiences disorientation and motion sickness. To avoid this, perfect synchronization between the user's *head pose* (position and orientation) in the real world and the virtual world is required. Therefore, instead of using its last received update of the user's head pose, the edge server must *predict* the user's current head pose. The edge server must then deliver a portion of the panoramic scene large enough to cover the user's *viewport* (the portion of the scene visible to the user), while accounting for the motion prediction error.

Naively, the edge server would deliver a portion covering the entire 360° scene. However, the wireless channel and real-time streaming requirements impose additional constraints that make this infeasible. Panoramic video streaming requires bandwidth 4 to 6 times greater than that of traditional video streaming [1] and therefore delivering the entire scene may result in significant congestion. Additionally, a user interacting with an immersive panoramic scene results in a more dynamic wireless channel than a stationary user, leading to increased packet loss. The real-time constraint is also significant. For example, Meta recommends that developers aim for 60 frames-per-second (FPS) for media applications, but emphasize that interactive applications must achieve a minimum of 72 FPS on their Meta Quest headset [2]. Wang et al. [3] observe through user studies that 120 FPS represents a threshold where users tend

to feel a significant reduction in virtual reality sickness. At such high framerates, the edge server often does not have enough time to retransmit lost packets. Therefore, the edge server must select a portion to deliver that is large enough to cover the user's viewport, but small enough to be successfully delivered over the wireless channel in a single frame.

If the dynamics of the user's head motion and the wireless channel were known in advance, the edge server could calculate the optimal delivery portion. However, in practice, these statistics are unknown and must be learned in real-time. A natural approach is to model the problem as a multi-armed bandit, where each arm is a delivery portion and a Bernoulli reward is accrued in a timeslot if the chosen delivery portion results in the user successfully seeing their entire viewport. In general, the rewards are non-i.i.d. due to the nonstationary user head motion and panoramic content. However, stochastic bandit algorithms, such as KL-UCB, are shown to empirically converge in real-world panoramic scene delivery experiments [4]. Beyond this simple bandit formulation, Chen et al. [5] make the observation that the aforementioned reward is composed of two feedback signals: (1) the prediction outcome, indicating whether the chosen delivery portion covers the user's actual viewport, and (2) the transmission outcome, indicating whether all packets constituting this chosen delivery portion are successfully transmitted over the wireless channel. In particular, they treat the prediction and transmission as separate bandit feedback observations, and therefore refer to this as *"two-level bandit feedback"*. Throughout this paper, we refer to the two-level bandit feedback using the notation 2/B/B, and the usual (single-level) bandit feedback as 1/B. [1] When there are two underlying bandit feedback signals, 1/B feedback refers to the case when only the product of two bandit feedback signals is observable.

However, what prior authors [4, 5] did not notice is that after receiving the user's head pose, the edge server can calculate whether a delivery portion would have covered the user's viewport *for each delivery portion*, not just the one it selected in the previous timeslot. Therefore, the prediction outcome is actually *full-information feedback*, and the problem exhibits two-level feedback where one level is full-information feedback and the other level is bandit feedback. We refer to this new type of two-level feedback using the notation 2/F/B. In this work, we focus on the advantages of exploiting the full structure of the portion selection problem for wireless interactive panoramic scene delivery by leveraging 2/F/B feedback over 2/B/B and 1/B feedback. Our main contributions are as follows:

(1) We formulate the problem of maximizing cumulative successful viewport delivery of an interactive panoramic scene as an online learning problem with two-level full-information and bandit (2/F/B) feedback. To our knowledge, this formulation involving mixed feedback types is novel and has not been explored previously in a theoretical or practical setting in the literature.
(2) We derive an instance-dependent lower bound on the regret under 2/F/B feedback (Theorem 5.1), which is shown to be significantly smaller than the corresponding lower bounds for 2/B/B and 1/B feedback for some problem instances. This suggests that incorporating full-information feedback can significantly improve learning efficiency.
(3) We design the *Hybrid Feedback-Driven Learning for Adaptive Portion Selection (AdaPort)* algorithm: a novel online learning algorithm that leverages 2/F/B feedback by using the empirical mean estimate for the full-information feedback and the Thompson sample for the bandit feedback.
(4) We derive an instance-dependent upper bound (Theorem 5.2) for AdaPort that matches the lower bound asymptotically, thus showing that the algorithm is asymptotically optimal. We also verify the improved theoretical performance compared to prior algorithms using 2/B/B and 1/B feedback by conducting numerical simulations.
(5) To validate the practical effectiveness of AdaPort, we perform simulations using a real-world data trace consisting of user head motion and panoramic video content. Our results show that AdaPort consistently outperforms state-of-the-art stochastic bandit portion selection algorithms that use 2/B/B and 1/B feedback. It is also shown to outperform the EXP3 algorithm with 1/B feedback for adversarial bandits.

The remainder of this paper is organized as follows. In Section 2, we review the related work on interactive panoramic scene delivery and on online learning with full-information and bandit feedback. In Section 3, we formulate the portion selection problem for interactive panoramic scene delivery as an online learning problem with 2/F/B feedback. In Section 4, we present our algorithm AdaPort. In Section 5, we derive an instance-dependent lower bound for the problem, as well as a matching instance-dependent upper bound for AdaPort asymptotically. In Section 6, we conduct synthetic and trace-based simulations to verify our theoretical results. Finally, we conclude the paper and suggest future research directions.

Note on Notation: We use bold and script font of a variable to denote a vector and a set, respectively. We use $a \wedge b$ to denote the $\min\{a, b\}$. We use $[N] \triangleq \{1, 2, \ldots, N\}$ to denote the set of the first N positive integers. We use $f(x) = o(g(x))$ to denote that $\lim_{x\to\infty} f(x)/g(x) = 0$, and $f(x) = O(g(x))$ to denote that $\limsup_{x\to\infty} f(x)/g(x) < \infty$, for positive functions f and g. We write $f(x) = \Theta(g(x))$ if there exist constants $c_1, c_2 > 0$ and x_0 such that $c_1 g(x) \le f(x) \le c_2 g(x)$ for all $x \ge x_0$. $F^B_{n,p}(\cdot)$ denotes the cumulative distribution function (CDF) of the binomial distribution with n trials and success probability p. We use $\mathrm{Beta}(\alpha, \beta)$ to denote the Beta distribution with parameters α and β and $F^{\mathrm{Beta}}_{\alpha,\beta}(\cdot)$ to denote the CDF of this Beta distribution. $d(p_1, p_2) \triangleq p_1 \log\left(\frac{p_1}{p_2}\right) + (1 - p_1) \log\left(\frac{1-p_1}{1-p_2}\right)$ is the KL-divergence between the Bernoulli(p_1) and Bernoulli(p_2) distributions. 1/B refers to the standard (one-level) bandit feedback, 2/B/B refers to two-level bandit feedback, and 2/F/B refers to two-level feedback where one level is full-information, and the other level is bandit feedback.

2 RELATED WORK

In this section, we review the related work on interactive panoramic scene delivery, and on online learning under full-information feedback and under bandit feedback.

[1] This $n/x/x$ notation, where n indicates the number of different feedback signals, and the remaining x's indicate the type of each feedback signal, is inspired by Kendall's notation from queueing theory.

2.1 Interactive Panoramic Scene Delivery

Panoramic scene delivery places significantly greater demands on network bandwidth than traditional video streaming. To address these challenges and enhance user experience, numerous studies have explored strategies for optimizing panoramic content transmission (e.g., [6, 7]). For example, Qian et al. [7] introduced an adaptive scheme that transmits only the portion of the panoramic scene aligned with the predicted user viewport, effectively reducing bandwidth consumption. However, these early approaches primarily relied on heuristic methods without providing theoretical performance guarantees. In response, subsequent research has introduced multi-armed bandit frameworks to the problem of panoramic scene delivery, aiming to provide a more principled and theoretically grounded understanding. Notably, recent studies have begun to explore how to effectively utilize the two distinct feedback signals inherent in this setting: motion prediction and wireless transmission. For example, Chen et al. [5] proposed the Two-Level Thompson Sampling algorithm to optimize viewport selection and maximize system throughput, while Gupta et al. [4] developed a Two-Level KL-UCB approach for the same objective. In parallel, other works have addressed multi-user scenarios (e.g., [8–10]) and multi-objective learning (e.g. [11]) in panoramic scene delivery. Nevertheless, prior studies have generally overlooked a key observation: the prediction outcome provides a full-information feedback signal, while the wireless transmission outcome yields bandit feedback. Thus, the problem naturally exhibits a two-level hybrid feedback structure, combining both full-information and bandit feedback, which has not been explicitly modeled or exploited in existing work.

2.2 Full-information and Bandit Feedback

In the online learning setting, bandit feedback refers to the scenario where the player observes the reward only for the arm selected at each timeslot, while the rewards of all other actions remain unobserved. In contrast, full-information feedback provides the player with the rewards of all available arms at each timeslot, regardless of the action selected. This richer feedback enables more informed learning strategies. In the context of bandit feedback, several classical algorithms have been extensively studied, including UCB [12], KL-UCB [13], and Thompson Sampling [14]. While most foundational work focuses on this bandit feedback setting, there also exist a number of studies exploring full-information feedback in both stochastic and non-stochastic bandit frameworks (e.g., [15–20]). For instance, Zhao et al. [15] examined a one-sided full-information stochastic bandit problem, where selecting an arm yields a reward from an unknown distribution while also revealing feedback from all arms located on one side of the selected arm. Alon et al. [16] introduced a partial-information model for online learning that generalizes and interpolates between the classical full-information and bandit settings. However, to the best of our knowledge, existing literature does not address the hybrid feedback setting where the player receives both full-information feedback and bandit feedback simultaneously within an online learning setting.

3 SYSTEM MODEL

We consider a system in which a single user interacts with an immersive panoramic scene that is delivered wirelessly from an edge server. The panoramic scene can be conceptualized as video content projected onto the inner surface of a virtual sphere surrounding the user's head. The user's viewport, i.e., the visible portion of the scene, which typically constitutes around 20% of the sphere's surface, is a rectangle centered at a fixed origin point corresponding to the user's view direction. The content on the sphere's surface changes as the user moves through the scene and rotates their head, and also changes according to the dynamics of the virtual environment. Using a standard map projection technique, such as equirectangular projection, the spherical panoramic content is transformed into a two-dimensional finite grid of tiles [2]. A *delivery portion* is a rectangular formation of tiles centered at and with dimensions larger than the user's viewport. We assume there are N fixed delivery portions that the edge server can choose from in each timeslot. Note that each timeslot $t = 1, 2, \ldots$ corresponds to a video frame and the system operates at 60-120 timeslots per second depending on the application's framerate requirement.

3.1 Online Learning under 2/F/B Feedback

After the edge server selects the delivery portion $i(t) \in [N]$ in timeslot t, the content corresponding to that delivery portion is sent to the user as a sequence of packets. Note that in general, the content may not be predetermined and may be updated in real-time according to the user's inputs (e.g., consider a VR video game). Therefore, we can only use a limited playback buffer because the content in timeslot t may become immediately irrelevant in timeslot $t+1$. Therefore, the content occupying the user's viewport in timeslot t must be sent from the server and received by the user by the end of timeslot t. Additionally, due to the short timeslot length, we do not consider any packet retransmissions. However, we assume a packet sent from the user to the edge server, called the *ACK packet*, containing the transmission acknowledgment, head pose, and other inputs from the previous timeslot is small enough that it can be reasonably assumed to be reliably delivered.

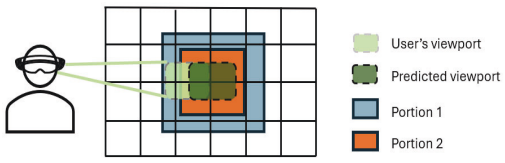


Figure 1: Relationship between the user's viewport, the predicted viewport, and delivery portions.

3.1.1 Prediction Outcome as Full-Information Feedback. At the beginning of timeslot $t+1$, the server will have received the user's ACK packet from timeslot t. Therefore, the server knows the user's predicted head pose, as well as their actual head pose from timeslot t, and can calculate whether or not each portion $i \in [N]$ would have covered their viewport. We use $X_i(t) = 1$ to indicate that portion i can cover the user's viewport in timeslot t and $X_i(t) = 0$ otherwise. As illustrated in Fig.1, given the user's actual viewport (light green), portion 1 fully covers it, i.e., $X_1(t) = 1$, whereas portion 2

[2] A tile represents the atomic unit for image encoding/decoding.

does not, i.e., $X_2(t) = 0$. Then the server observes the entire vector $\mathbf{X}(t) \triangleq (X_i(t))_{i=1}^N$ by the beginning of timeslot $t+1$, i.e. the server receives full-information feedback for the prediction outcome. We assume $\mathbf{X}(t)$ is i.i.d. over time and each component has unknown rate $\alpha_i \triangleq \mathbb{E}[X_i(t)]$. However, note that $\mathbf{X}(t)$ is component-wise dependent, i.e. for a fixed timeslot t and two portions $i \neq j$, $X_i(t)$ and $X_j(t)$ are not independent. As previously mentioned, they are actually both functions of two random variables, namely, the edge server's prediction of the user's head pose from timeslot t and the user's actual head pose from timeslot t.

3.1.2 Transmission Outcome as Bandit Feedback. At the beginning of timeslot $t+1$, after receiving the user's ACK packet from timeslot t, the server knows whether or not the transmission of delivery portion $i(t)$ was successful. We use $Y_i(t) = 1$ to indicate a successful transmission when delivery portion $i(t) = i$ is selected, and $Y_i(t) = 0$ otherwise. Then the server observes the transmission outcome $Y_{i(t)}(t)$ as bandit feedback. In this paper, we consider the transmission to have failed if any packet constituting the delivery portion $i(t)$ is lost. However, the problem can easily be extended to consider a failed transmission to occur only when a minimum portion of the packets are lost. As with the prediction outcome, we assume $\mathbf{Y}(t) \triangleq (Y_i(t))_{i=1}^N$ is i.i.d. over time, but unlike the prediction outcome, was assume it is component-wise independent, with each component having an unknown rate $\beta_i \triangleq \mathbb{E}[Y_i(t)]$. Note that, as illustrated in Fig.1, all candidate portions encompass the predicted viewport (dark green), being centered on it and expanded outward to varying extents. We also assume that for a fixed portion i and timeslot t, $X_i(t)$ and $Y_i(t)$ are independent.

We use $Z_i(t) = 1$ to indicate that the user can successfully view the desired content when portion $i(t) = i$ is selected in timeslot t, and $Z_i(t) = 0$ otherwise. We also refer to $Z_i(t)$ as the *"throughput"* or *"reward"* for portion i in timeslot t. Note that $Z_i(t) = 1$ only when both the prediction and the transmission are successful, i.e. $Z_i(t) = X_i(t)Y_i(t)$. Our objective is to design an adaptive portion selection algorithm that maximizes the expected cumulative throughput $\sum_{t=1}^T \mathbb{E}\left[Z_{i(t)}(t)\right]$ up to a time horizon T when the prediction and transmission rates $(\alpha_i, \beta_i)_{i=1}^N$ are unknown. If the parameters were known, the optimal policy would be to always play the arm $i^* \in \arg\max_{i\in[N]} \alpha_i\beta_i$. Therefore, the regret is given by

$$R(T) \triangleq \sum_{t=1}^{T} \mathbb{E}\left[Z_{i^*}(t) - Z_{i(t)}(t)\right] \tag{1}$$

$$= T\alpha_{i^*}\beta_{i^*} - \sum_{t=1}^{T} \mathbb{E}\left[X_{i(t)}(t)Y_{i(t)}(t)\right] = \sum_{i\neq i^*} \Delta_i \mathbb{E}\left[n_i(T)\right],$$

where $\Delta_i \triangleq \alpha_{i^*}\beta_{i^*} - \alpha_i\beta_i$ denotes the reward gap for portion i and $n_i(T) \triangleq \sum_{t=1}^T \mathbb{1}\{i(t) = i\}$ denotes the number of plays of portion i up to timeslot T.

To summarize, adaptive portion selection for wireless interactive panoramic scene delivery under 2/F/B feedback proceeds as follows: In each timeslot t,

① The server decides the delivery portion $i(t)$ for timeslot t according to its history of observations $(\mathbf{X}(\tau), Y_{i(\tau)}(\tau), i(\tau))_{\tau=1}^{t-1}$.

② The server predicts the user's head pose for timeslot t and sends the content occupying the chosen delivery portion $i(t)$ centered on the user's predicted head pose to the user over the wireless channel.

③ The server receives the ACK packet for timeslot t and observes the user's actual head pose and the transmission outcome $Y_{i(t)}(t)$.

④ The server calculates the prediction outcome $X_i(t)$ for each portion i from the user's actual head pose (received in the ACK packet) and the server's predicted head pose.

3.2 Comparison with 1/B and 2/B/B Feedback

Using the language of the multi-armed bandit literature, we sometimes call a delivery portion $i \in [N]$ an *"arm"*. For a bandit feedback signal, balancing *exploration* and *exploitation* is crucial. That is, exploring underplayed arms vs. exploiting the current most promising arm. The prior work on portion selection [4, 5] improved on the 1/B feedback model, where the edge server only receives the throughput $Z_{i(t)}(t)$ as feedback, by considering the 2/B/B feedback model, where the edge server receives both $X_{i(t)}(t)$ and $Y_{i(t)}(t)$ as feedback. While 2/B/B gathers more information per timeslot, both feedback signals are bandit, and are therefore subject to the exploration/exploitation dilemma. The major improvement we introduce with the 2/F/B feedback model is to completely remove the necessity for exploration in one of the feedback signals. Also note that 2/F/B feedback yields $N-1$ more pieces of information compared to 2/B/B feedback.

3.3 Discussion and Limitation

In this subsection, we discuss the limitations of the assumptions in our system model. Specifically, we assume that $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ are i.i.d. over time, and that $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ are independent. These assumptions, also adopted in [4, 5], facilitate analytical tractability. In practice, however, $\mathbf{X}(t)$ may exhibit temporal dependencies due to user head movements, violating the i.i.d. assumption, while $\mathbf{Y}(t)$ may deviate from i.i.d. behavior due to variations in the panoramic scene content. Furthermore, $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ may be dependent because recent head poses can be related to the scene content. Recall that we also assumed $\mathbf{Y}(t)$ is component-wise independent. However, in reality, if we successfully deliver the chosen portion, then we know that we could have also successfully delivered a smaller portion, and if we fail to deliver the chosen portion, then we know that we would have also failed to deliver a larger portion. Despite these modeling assumptions, it is worth noting that our real-world trace-based evaluations, which *do not* rely on any i.i.d. or independence assumption, show that the proposed algorithm consistently outperforms baseline methods. Finally, we note that the 2/F/B feedback model is of independent theoretical interest beyond the 2/B/B model studied in [4, 5].

4 ALGORITHM DESIGN

Recall the step-by-step summary of the sequence of events for the adaptive portion selection problem under 2/F/B feedback presented in Section 3. In this section, we focus on step ①. Specifically, how to design an adaptive portion selection algorithm to select the delivery portion $i(t)$ in timeslot t given the history of observations

$(\mathbf{X}(\tau), Y_{i(\tau)}(\tau), i(\tau))_{\tau=1}^{t-1}$. Recall that the optimal policy always selects $i^* \in \arg\max_{i\in[N]} \alpha_i\beta_i$ in each timeslot t. Since $(\alpha_i, \beta_i)_{i=1}^N$ are unknown, we instead choose

$$i(t) \in \arg\max_{i\in[N]} \overline{\alpha}_i(t)\, \theta_{\beta,i}(t). \tag{2}$$

where each $\overline{\alpha}_i(t)$ and $\theta_{\beta,i}(t)$ are our best estimates of α_i and β_i respectively, estimated from the history of observations. The details of how these estimates are chosen are given in the following subsections, and our algorithm, AdaPort is given in Algorithm 1.

4.1 Prediction Rate Estimate

Recall that the prediction outcome is a full-information feedback signal. Then to estimate α_i for each portion $i \in [N]$, we simply use the empirical mean $\overline{\alpha}_i(t) \triangleq \frac{1}{t-1}\sum_{\tau=1}^{t-1} X_i(\tau)$ for timeslots $t > 1$ and $\overline{\alpha}_i(1) = 0$ (the value of $\overline{\alpha}_i(1)$ is arbitrary under full-information). This estimate can be updated iteratively using the recurrence

$$\overline{\alpha}_i(t+1) = \overline{\alpha}_i(t) + \frac{1}{t}\left(X_i(t) - \overline{\alpha}_i(t)\right). \tag{3}$$

4.2 Transmission Rate Estimate

Unlike the prediction outcome signal, the transmission outcome is a bandit feedback signal. Therefore, we need to use an estimate of each β_i that is able to balance exploration and exploitation. Prior work [4] used the KL-UCB estimate for this signal, which accomplishes this by adding an exploration bonus to the empirical mean. However, a Bayesian approach called *Thompson sampling* typically performs better in numerical experiments compared to UCB-type algorithms (see Section 6). In Thompson sampling, the estimate $\theta_{\beta,i}(t)$ is drawn from a prior distribution representing our current belief about the unknown parameter β_i. For Bernoulli rewards, we draw the Thompson sample $\theta_{\beta,i}(t) \sim \text{Beta}(S_{\beta,i}(t)+1, F_{\beta,i}(t)+1)$, where $S_{\beta,i}(t)$ and $F_{\beta,i}(t)$ denote the number of observed transmission successes and failures respectively. These are initialized as $S_{\beta,i}(1) = 0$ and $F_{\beta,i}(1) = 0$ and updated as

$$\begin{aligned} S_{\beta,i}(t+1) &= S_{\beta,i}(t) + \mathbb{1}\{i(t) = i\}\, Y_{i(t)}(t), \text{ and} \\ F_{\beta,i}(t+1) &= F_{\beta,i}(t) + \mathbb{1}\{i(t) = i\}\left(1 - Y_{i(t)}(t)\right). \end{aligned} \tag{4}$$

Exploration of each portion i is accomplished in Thompson sampling due to the variance in the Beta distribution, which shrinks as the number of plays $n_i(t)$ of portion i increases.

5 PERFORMANCE ANALYSIS

In this section, we begin by deriving the regret lower bound for the online learning problem with 2/F/B feedback. We then compare this lower bound against the known lower bounds for 2/B/B feedback derived by Gupta et al. [4] and for 1/B feedback derived by Lai and Robbins [21]. To further support our motivation, we present simple illustrative examples that highlight the significant improvement in the lower bound under the 2/F/B feedback setting. Finally, we establish the regret upper bound of AdaPort (Algorithm 1), which matches the lower bound asymptotically.

[3] Refer to steps (3) and (4) in the step-by-step summary of events in Section 3.

Algorithm 1 Hybrid Feedback-Driven Learning for Adaptive Portion Selection Algorithm (AdaPort)

Initialization: Set $S_{\beta,i}(1) = F_{\beta,i}(1) = \overline{\alpha}_i(1) = 0 \quad \forall\, i \in [N]$.
for each $t = 1, 2, \ldots$ **do**
  for each portion i **do**
    Draw $\theta_{\beta,i}(t) \sim \text{Beta}\left(S_{\beta,i}(t)+1, F_{\beta,i}(t)+1\right)$.
  end for
  Send the delivery portion $i(t) \in \arg\max_i \overline{\alpha}_i(t)\, \theta_{\beta,i}(t)$.
  Receive the feedback $(\mathbf{X}(t), Y_{i(t)}(t))$. [3]
  for each portion i **do**
    Update $\overline{\alpha}_i(t)$ according to (3).
    Update $S_{\beta,i}(t)$ and $F_{\beta,i}(t)$ according to (4).
  end for
end for

5.1 Regret Lower Bound under 2/F/B Feedback

In the following theorem, we present our asymptotic instance-dependent regret lower bound for online learning under 2/F/B feedback.

Theorem 5.1. *(Lower bound) Consider an online learning algorithm under 2/F/B feedback that achieves $R(T) = o(T^\delta) \quad \forall\, \delta > 0$. Then the algorithm is subject to the following regret lower bound:*

$$\liminf_{T\to\infty} \frac{R(T)}{\log T} \geq \sum_{i\neq i^*\,:\,\alpha_i > \alpha_{i^*}\beta_{i^*}} \frac{\Delta_i}{d\left(\beta_i, \frac{\alpha_{i^*}\beta_{i^*}}{\alpha_i}\right)}. \tag{5}$$

Proof Sketch. Recall from (1) that the regret can be decomposed as $R(T) = \sum_{i\neq i^*} \Delta_i \mathbb{E}\left[n_i(T)\right]$. Then it suffices to show that $\liminf_{T\to\infty} \frac{n_i(T)}{\log T} \geq \Delta_i / d\left(\beta_i, \frac{\alpha_{i^*}\beta_{i^*}}{\alpha_i}\right)$ for some fixed arm $j \neq i^*$ with $\alpha_j > \alpha_{i^*}\beta_{i^*}$. Given arm j, fix $\lambda \in (\Delta_j, \alpha_j(1-\beta_j))$. We construct a new problem instance given by a second set of parameters $(\alpha_i', \beta_i')_{i=1}^N$ where each $\alpha_i' = \alpha_i$, $\beta_j' = \beta_j + \lambda/\alpha_j$, and $\beta_i' = \beta_i$ for all $i \neq j$. The remainder of the proof is similar to the usual instance-dependent lower bound for stochastic bandits (see e.g. Theorem 16.2 in [22]), with some slight modifications to account for the two levels of feedback. The full proof can be found in Appendix A. □

The right-hand side of (5) is called the *"lower bound constant"*. A key observation about the bound in Theorem 5.1 is that arms with $\alpha_i \leq \alpha_{i^*}\beta_{i^*}$ have zero contribution. This shows that the full-information feedback is able to completely eliminate some arms from play, whereas 2/B/B and 1/B feedback require a continual $\log T$ number of plays of each suboptimal arm. Intuitively, when $\alpha_i \leq \alpha_{i^*}\beta_{i^*}$, even if the transmission rate of the suboptimal arm i is 1, it is unlikely that this arm i can compete with the optimal arm. Consistent with this intuition, in the upper bound proof (see (20)), only a constant number of plays is sufficient to identify such an arm i as suboptimal. Consequently, these arms do not contribute to the lower bound constant.

Remark 1 (lower bound constant comparison).

$$\overbrace{\sum_{\substack{i\neq i^*:\\ \alpha_i>\alpha_{i^*}\beta_{i^*}}} \frac{\Delta_i}{d\left(\beta_i, \frac{\alpha_{i^*}\beta_{i^*}}{\alpha_i}\right)}}^{\text{2/F/B lower bound constant}} \overset{(i)}{\leq} \sum_{i\neq i^*} \frac{\Delta_i}{d\left(\beta_i, \frac{\alpha_{i^*}\beta_{i^*}}{\alpha_i} \wedge 1\right)}$$

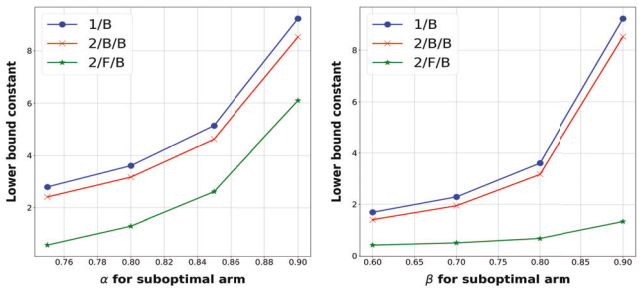
$$= \sum_{i\neq i^*} \frac{\Delta_i}{d(\alpha_i, x_i) + d(\beta_i, y_i)}, \quad \begin{matrix} x_i = \alpha_i, \\ y_i = \frac{\alpha_{i^*}\beta_{i^*}}{\alpha_i} \wedge 1 \end{matrix}$$

$$\overset{(ii)}{\leq} \underbrace{\sum_{i\neq i^*} \frac{\Delta_i}{\min\limits_{\substack{0\leq x_i, y_i\leq 1 \\ x_i y_i \geq \alpha_{i^*}\beta_{i^*}}} d(\alpha_i, x_i) + d(\beta_i, y_i)}}_{\text{2/B/B lower bound constant derived by [4]}}$$

$$\overset{(iii)}{\leq} \underbrace{\sum_{i\neq i^*} \frac{\Delta_i}{\min\limits_{\substack{0\leq x_i, y_i\leq 1 \\ x_i y_i \geq \alpha_{i^*}\beta_{i^*}}} d(\alpha_i\beta_i, x_i y_i)}}_{\text{1/B lower bound constant derived by [21]}} \tag{6}$$

The above shows that the lower bound constant under 2/F/B given in Theorem 5.1 is smaller than the lower bound constant under 2/B/B derived by Gupta et al. [4] on two levels: (i), the lower bound constant under 2/F/B does not count all suboptimal arms, as does the lower bound constant under 2/B/B, and (ii), the lower bound constant under 2/B/B minimizes the denominator for each suboptimal arm, unlike the lower bound constant under 2/F/B. The relationship (iii) between the lower bound constants under 2/B/B feedback and 1/B feedback is proven in Theorem 2 of [4].

We further illustrate the size comparison between the lower bound constants under 2/F/B, 2/B/B, and 1/B feedback shown in Remark 1 by numerical simulations with two arms. Specially, we set the optimal arm's prediction and transmission rates as $\alpha_{i^*} = 0.8$ and $\beta_{i^*} = 0.9$, respectively. In the first example (Figure 2a), we fix the suboptimal arm's transmission rate at 0.8 and vary its prediction rate across the set $[0.75, 0.8, 0.85, 0.9]$. In the second example (Figure 2b), we fix the prediction rate at 0.75 and vary the transmission rate across $[0.6, 0.7, 0.8, 0.9]$. In Figure 2, we observe that the lower bound constants corresponding to 1/B and 2/B/B settings exhibit only marginal differences. In contrast, the lower bound constant of 2/F/B is significantly smaller, indicating that a learning algorithm may substantially reduce its regret by exploiting the additional full-information feedback.



(a) Fixing β and varying α. **(b) Fixing α and varying β.**

Figure 2: Lower bound constant comparison for two arms.

5.2 Regret Upper Bound of AdaPort

Finally, in the following theorem, we validate that AdaPort (Algorithm 1) achieves an upper bound constant equal to the lower bound constant in Theorem 5.1, demonstrating its asymptotical optimality.

Theorem 5.2. *(Upper bound) AdaPort (Algorithm 1) achieves an asymptotic regret upper bound of*

$$\limsup_{T\to\infty} \frac{R(T)}{\log T} \leq \sum_{i\neq i^*:\ \alpha_i>\alpha_{i^*}\beta_{i^*}} \frac{\Delta_i}{d\left(\beta_i, \frac{\alpha_{i^*}\beta_{i^*}}{\alpha_i}\right)}. \tag{7}$$

Proof Sketch. Recall from (1) that the regret can be decomposed as $R(T) = \sum_{i\neq i^*} \Delta_i \mathbb{E}[n_i(T)] = \sum_{i\neq i^*} \Delta_i \sum_{t=1}^{T} \Pr(i(t) = i)$. Fix a suboptimal arm $i \neq i^*$ and a corresponding $\epsilon > 0$. We decompose each $\Pr(i(t) = i)$ into the probabilities of the following events:

(i) $i(t) = i$ when $\overline{\alpha}_i(t)\theta_{\beta,i}(t) \leq \alpha_{i^*}\beta_{i^*} - \epsilon$, $|\overline{\alpha}_i(t) - \alpha_i| \leq \epsilon, |\overline{\beta}_i(t) - \beta_i| \leq \epsilon$ and $|\overline{\alpha}_{i^*}(t) - \alpha_{i^*}| \leq \epsilon$,
(ii) $i(t) = i$ when $\overline{\alpha}_i(t)\theta_{\beta,i}(t) > \alpha_{i^*}\beta_{i^*} - \epsilon$ and $|\overline{\alpha}_i(t) - \alpha_i| \leq \epsilon, |\overline{\beta}_i(t) - \beta_i| \leq \epsilon$,
(iii) $i(t) = i$ when either $|\overline{\alpha}_i(t) - \alpha_i| > \epsilon$ or $|\overline{\beta}_i(t) - \beta_i| > \epsilon$,
(iv) $|\overline{\alpha}_{i^*}(t) - \alpha_{i^*}| > \epsilon$.

To bound the probabilities of events (i) and (ii), we note that

$$\Pr\left(\theta_{\beta,i}(t) > x \mid \mathcal{F}_{t-1}\right) = \Pr\left(\theta_{\beta,i}(t) > x \mid S_{\beta,i}(t), F_{\beta,i}(t)\right)$$

almost surely for some $x > 0$, where $\mathcal{F}_{t-1} \triangleq (\mathbf{X}(\tau), Y_{i(\tau)}(\tau), i(\tau))_{\tau=1}^{t-1}$ is the history of observations (see the proof of [22, Theorem 36.2]). This allows use to ignore the dependence on $\mathbf{X}(t)$ and proceed analogously to the analysis in [14]. Event (iii) and (iv) can be handled using standard concentration inequalities and are therefore straightforward. The full proof can be found in Appendix B. □

Remark 2. *The regret contribution from event (ii) is of order $O(\log T)$. Under the 2/B/B feedback model, even if the empirical estimates $\overline{\alpha}_i(t)$ and $\overline{\beta}_i(t)$ are close to their true means α_i and β_i, the arm selection remains uncertain due to the stochasticity of both $\theta_{\alpha,i}(t)$ and $\theta_{\beta,i}(t)$. These variables represent Thompson Sampling draws for the transmission and prediction rates, respectively, and are updated based on bandit feedback [4, 5]. Thus, uncertainty persists in both dimensions. In contrast, under the 2/F/B feedback model, the transmission rate α_i is estimated with full information, so $\overline{\alpha}_i(t)$ concentrates rapidly around α_i. As a result, the only remaining source of exploration-induced randomness is $\theta_{\beta,i}(t)$, which governs the uncertainty in the transmission rate. This reduction in uncertainty leads to more confident decision-making and lower regret in practice.*

6 EMPIRICAL EVALUATION

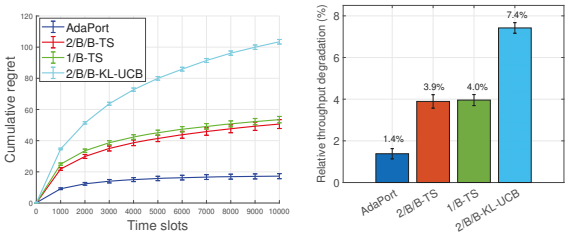
In this section, we evaluate our theoretical regret upper bound in both synthetic and real-world trace-based simulations using data traces collected from a user interacting with a panoramic video stream. We compare the performance of our algorithm against four baseline algorithms:

(1) **Thompson sampling under 1/B feedback** (1/B-TS): the standard Thompson sampling MAB algorithm.
(2) **Thompson sampling under 2/B/B feedback** (2/B/B-TS): two-level Thompson sampling (Chen et al. [5]).

(3) **KL-UCB under 2/B/B feedback** (2/B/B-KL-UCB): the two-level KL-UCB algorithm (Gupta et al. [4]).
(4) **EXP3 under 1/B feedback** (1/B-EXP3): an optimal algorithm for adversarial bandits (Auer et al. [23]).

(2) and (3) are state-of-the-art algorithms for the portion selection problem, while (4) is chosen to test our i.i.d. prediction and transmission assumptions. In the following subsections, we detail our findings for the synthetic and trace-based simulations. In addition to plotting the regret $R(T)$, in order to evaluate user experience, we also plot the *relative throughput degradation* compared to the optimal policy, which is given by:

$$\mathbb{E}\left[\frac{\sum_{t=1}^{T}(1-Z_{i(t)}(t))-\sum_{t=1}^{T}(1-Z_{i^*}(t))}{\sum_{t=1}^{T}(1-Z_{i^*}(t))}\right]. \tag{8}$$



(a) (b)

Figure 3: Synthetic simulation results. (a) shows the cumulative regret, and (b) shows the relative throughput degradation compared to the optimal policy.

6.1 Synthetic Simulation

For the synthetic simulation setup, we make a simplifying approximation by considering circular portions instead of rectangular ones. The prediction error is modeled as a standard Normal random variable $\delta(t) \sim \text{Normal}(0,1)$, and each portion i is associated with a radius r_i. Then the prediction outcome is given by $X_i(t) = \mathbb{1}\{|\delta(t)| \le r_i\}$, i.e. the prediction is successful if the portion is large enough to accommodate the prediction error. We choose the parameters $\mathbf{r} = [1.0, 1.15, 1.3, 1.6, 2.0]$, i.e. a larger portion i has a larger radius r_i. The transmission outcome for portion i is directly generated as a $Y_i(t) \sim \text{Bernoulli}(\beta_i)$ random variable. We choose the parameters $\boldsymbol{\beta} = [0.98, 0.97, 0.95, 0.93, 0.9]$, i.e. a smaller portion i has a higher transmission probability. To reduce the variance in our results, we average over 1000 parallel experiments, each consisting of 10^4 timeslots. We plot the mean and 95% confidence interval (error bars) of the regret in Figure 3a. The results demonstrate that AdaPort (Algorithm 1) outperforms Thompson sampling under both 1/B and 2/B/B feedback and KL-UCB under 2/B/B feedback, which is consistent with Theorem 5.1, Remark 1, and Theorem 5.2. The significant performance gain also agrees with out numerical experiments demonstrating the lower bound improvement in Figure 2. Our algorithm achieves a notable reduction in cumulative regret of more than 60% compared to its closest competitor, Thompson sampling under 2/B/B feedback.

Furthermore, although Remark 1 suggests that increased observational feedback can lead to a smaller regret lower bound, our experiments reveal that Thompson Sampling under 1/B feedback outperforms KL-UCB under 2/B/B feedback. This suggests that while additional feedback asymptotically reduces regret, it may take an impractically long time for an algorithm to converge, and that Thompson sampling is a much better choice than KL-UCB in practical scenarios.

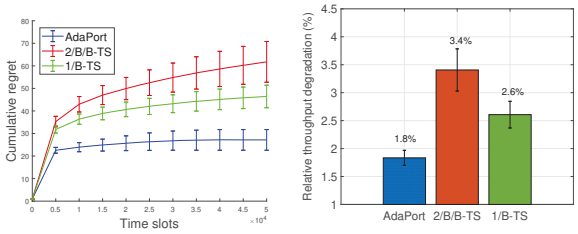
According to Figure 3b, AdaPort is highly effective at learning the optimal strategy, incurring only a modest 1.4% increase relative throughput degradation compared to the optimal policy. This indicates that AdaPort delivers a near-optimal user experience, especially when contrasted with other baseline algorithms that suffer from significantly higher relative throughput degradation.

6.2 Trace-Based Simulation

We collect a data trace from a user viewing a free educational panoramic video (see [24]). The trace captures a 3 DoF (orientation only) head motion over 3000 timeslots. We iterate over the dataset multiple times to achieve a total of $T = 5 \times 10^4$ timeslots for our experiments:

6.2.1 Delivery portions. At each timeslot t, we apply a linear regression model to the motion trace in order to predict the user's head pose for the subsequent timeslot $t+1$. The model is trained using motion trajectories from the preceding three timeslots. We set $N = 4$ different portions to select from: $100^\circ \times 90^\circ$ (minimum viewport), $102^\circ \times 91^\circ$, $108^\circ \times 94^\circ$, and $120^\circ \times 100^\circ$, where each pair gives the angles corresponding to the yaw and pitch axes, respectively. For each portion $i \in [N]$, we compute $X_i(t)$ from the user's actual head pose (received in the ACK packet) and the server's predicted head pose, as described in step (4) in Section 3.

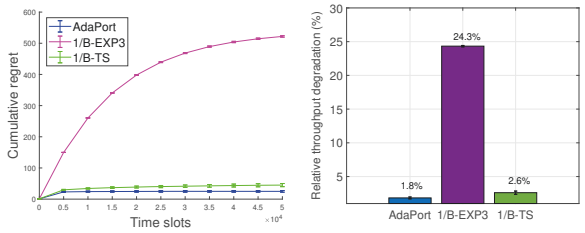
6.2.2 Wireless Transmissions. Let $\texttt{pkt}_{\texttt{net}}(t)$ denote the number of packets that can be successfully delivered over the wireless channel from the edge server to the user in timeslot t. We assume this follows a Poisson distribution: $\texttt{pkt}_{\texttt{net}}(t) \sim \text{Poisson}(\lambda_{\texttt{net}})$ with $\lambda_{\texttt{net}} = 170$. We use $\texttt{pkt}_i(t)$ to denote the number of packets required to transmit the delivery portion i in timeslot t. In each timeslot t, after selecting the portion $i(t)$ to transmit, we extract the corresponding number of bytes from the dataset and divide it up into $\texttt{pkt}_{i(t)}(t)$ UDP packets, each with a payload size of 1400 bytes. We observe a successful transmission if the required number of packets does not exceed the channel rate, i.e., $Y_{i(t)}(t) = \mathbb{1}\{\texttt{pkt}_{\texttt{net}}(t) \ge \texttt{pkt}_{i(t)}(t)\}$.



(a) (b)

Figure 4: Trace-based simulations: AdaPort compared against Thompson sampling under 2/B/B and 1/B feedback. (a) shows the cumulative regret, and (b) shows the relative throughput degradation compared to the optimal policy.

6.2.3 Comparison with 1/B-TS and 2/B/B-TS. As shown in Figure 4, AdaPort clearly outperforms both 1/B-TS and 2/B/B-TS. Note that when $\lambda_{\text{net}} = 170$, the network throughput is approximately 60 Mbps, which represents a realistic and commonly encountered network rate. Under this setting, the observed transmission success rates across all portions lie within the range of approximately 0.94 to 0.99. The superior performance of our algorithm AdaPort aligns with our intuition that, under high-quality network conditions, transmission randomness is minimal and the main source of uncertainty stems from prediction. By utilizing more feedback in prediction, AdaPort effectively reduces prediction uncertainty and thus achieves improved performance. Consequently, under high-quality and stable network conditions, AdaPort demonstrates superior performance compared to state-of-the-art baselines. In addition, consistent with the results from the synthetic simulations shown in Figure 3, AdaPort incurs less than a 2% increase in failed deliveries sacrificed in learning, thereby providing users with a near-optimal experience. Moreover, when considered in conjunction with the theoretical lower bound results in Figure 2, we observe that augmenting with additional bandit feedback may only lead to marginal improvements in regret. In practice, this is further reflected in cases where 2/B/B-TS performs worse than 1/B-TS, indicating that additional feedback does not always translate to better performance in realistic settings.



(a) (b)

Figure 5: Trace-based simulations: AdaPort compared against Thompson sampling and EXP3 under 1/B feedback. (a) shows the cumulative regret, and (b) shows the relative throughput degradation compared to the optimal policy.

6.2.4 Comparison with 1/B-TS and 1/B-EXP3. Recall that our i.i.d. assumptions about the prediction and transmission outcomes in Section 3 may not be realistic in practice. Therefore, we compare AdaPort, as well as 1/B-TS against 1/B-EXP3, which is a well-established algorithm designed for adversarial multi-armed bandit scenarios. As illustrated in Figure 5, both AdaPort and 1/B-TS outperform 1/B-EXP3, suggesting that our i.i.d. assumptions are not that unreasonable in practice. Our simulations consider i.i.d. Poisson-distributed network conditions. In scenarios where this setting is violated, such as when the network rate varies significantly over time, 1/B-EXP3 may offer advantages due to its design for more adversarial environments.

7 CONCLUSION AND FUTURE WORK

In this paper, we investigated a hybrid feedback-driven online learning framework tailored to the interactive delivery of panoramic scenes over wireless networks. A critical observation is that, upon receiving the user's current head movement data, the system can determine whether each candidate portion successfully covers the user's viewport. This implies that the prediction feedback can be categorized as full-information feedback. We first establish a theoretical improvement in the regret lower bound for 2/F/B, as compared to 2/B/B and 1/B feedback. This highlights the potential benefit of leveraging full-information predictions to enhance regret performance. Building on this, we propose AdaPort, a hybrid learning algorithm that leverages both the full-information feedback and bandit feedback and demonstrate that the proposed algorithm is asymptotically optimal. Extensive evaluations, including both synthetic simulations and real-world trace-based experiments, validate the superior performance of our approach relative to state-of-the-art portion selection algorithms. In future work, we plan to investigate more realistic reward models that better capture the nuances of user experience. For example, we aim to refine the definition of delivery success by accounting for cases in which only a small portion of a segment fails to be delivered. Such partial failures may have minimal impact on the user's viewing experience and thus should not necessarily be treated as complete delivery failures. We also intend to explore alternative problem formulations, such as nonstationary bandits, to more accurately reflect the dynamics of real-world panoramic scene delivery.

8 ACKNOWLEDGMENT

The authors would like to thank Prof. Stratis Ioannidis for his valuable suggestions. This work was supported in part by NSF under the Grants CNS-2152610, CNS-2152658, and CNS-2106801, and the ARO grant W911NF-24-1-0103.

REFERENCES

[1] Y. Bao, H. Wu, T. Zhang, A. A. Ramli, and X. Liu, "Shooting a moving target: Motion-prediction-based transmission for 360-degree videos," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1161–1170.

[2] "Meta horizon - testing and performance analysis." [Online]. Available: https://developers.meta.com/horizon/documentation/unity/unity-perf/

[3] J. Wang, R. Shi, W. Zheng, W. Xie, D. Kao, and H.-N. Liang, "Effect of frame rate on user experience, performance, and simulator sickness in virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2478–2488, 2023.

[4] H. Gupta, J. Chen, B. Li, and R. Srikant, "Online learning-based rate selection for wireless interactive panoramic scene delivery," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1799–1808.

[5] J. Chen, B. Li, and R. Srikant, "Thompson-sampling-based wireless transmission for panoramic video streaming," in *2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*. IEEE, 2020, pp. 1–3.

[6] X. Liu, C. Vlachou, M. Yang, F. Qian, L. Zhou, C. Wang, L. Zhu, K.-H. Kim, G. Parmer, Q. Chen *et al.*, "Firefly: Untethered multi-user {VR} for commodity mobile devices," in *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, 2020, pp. 943–957.

[7] F. Qian, B. Han, Q. Xiao, and V. Gopalakrishnan, "Flare: Practical viewport-adaptive 360-degree video streaming for mobile devices," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 99–114.

[8] J. Chakareski, "Viewport-adaptive scalable multi-user virtual reality mobile-edge streaming," *IEEE Transactions on Image Processing*, vol. 29, pp. 6330–6342, 2020.

[9] J. Chen, X. Qin, G. Zhu, B. Ji, and B. Li, "Motion-prediction-based wireless scheduling for multi-user panoramic video streaming," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[10] X. Wu and B. Li, "Achieving regular and fair learning in combinatorial multi-armed bandit," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, 2024, pp. 361–370.

[11] J. Steiger, X. Wu, and B. Li, "Learning to wirelessly deliver consistent and high-quality interactive panoramic scenes," in *2025 23st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. IEEE, 2025.
[12] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, pp. 235–256, 2002.
[13] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th annual conference on learning theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 359–376.
[14] S. Agrawal and N. Goyal, "Near-optimal regret bounds for thompson sampling," *Journal of the ACM (JACM)*, vol. 64, no. 5, pp. 1–24, 2017.
[15] H. Zhao and W. Chen, "Stochastic one-sided full-information bandit," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 150–166.
[16] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir, "Nonstochastic multi-armed bandits with graph-structured feedback," *SIAM Journal on Computing*, vol. 46, no. 6, pp. 1785–1826, 2017.
[17] Y.-H. Yan, P. Zhao, and Z.-H. Zhou, "Online non-stochastic control with partial feedback," *Journal of Machine Learning Research*, vol. 24, no. 273, pp. 1–50, 2023.
[18] Q. Zhang, Z. Deng, Z. Chen, K. Zhou, H. Hu, and Y. Yang, "Online learning for non-monotone dr-submodular maximization: From full information to bandit feedback," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 3515–3537.
[19] F. Liu, S. Buccapatnam, and N. Shroff, "Information directed sampling for stochastic bandits with graph feedback," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[20] D. Cheng, X. Zhou, and B. Ji, "Understanding the role of feedback in online learning with switching costs," 2023. [Online]. Available: https://arxiv.org/abs/2306.09588
[21] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
[22] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
[23] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.
[24] N. Geographic, "First-ever 3d vr filmed in space." [Online]. Available: https://vuze.camera/vr-gallery/3d-360-video/first-ever-3d-vr-filmed-space
[25] "Optimal hybrid feedback-driven learning for wireless interactive panoramic scene delivery," https://sites.psu.edu/xiaoyiwu/files/2025/04/2025_mobihoc.pdf, 2025, technical Report.

A PROOF OF THEOREM 5.1

To prove Theorem 5.1, we require the following divergence decomposition lemma, which can be thought of as a version of Lemma 15.1 in [22] for 2/F/B feedback.

Lemma A.1. *Given an alternative set of system parameters $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$, and the corresponding probability measure $\Pr'$ under these parameters, under any 2/F/B online learning algorithm, we can decompose*

$$D_{\mathrm{KL}}\left(\Pr \parallel \Pr'\right) = T d^{(X)}\left(\boldsymbol{\alpha}, \boldsymbol{\alpha}'\right) + \sum_{i=1}^{N} d_i^{(Y)}\left(\beta_i, \beta_i'\right) \sum_{t=1}^{T} \Pr(i(t)=i),$$

where $d^{(X)}(\boldsymbol{\alpha}, \boldsymbol{\alpha}') \triangleq \sum_{\mathbf{x}} \Pr(\mathbf{X}(1)=\mathbf{x}) \log\left(\frac{\Pr(\mathbf{X}(1)=\mathbf{x})}{\Pr'(\mathbf{X}(1)=\mathbf{x})}\right)$ *and each* $d_i^{(Y)}\left(\beta_i, \beta_i'\right) \triangleq \sum_y \Pr\left(Y_i(1)=y\right) \log\left(\frac{\Pr(Y_i(1)=y)}{\Pr'(Y_i(1)=y)}\right)$.

Proof Sketch. We first derive for all sample paths $(\mathbf{x}_t, y_t)_{t=1}^T$ that

$$\begin{aligned}&\Pr\left((\mathbf{X}(t), Y_{i(t)}(t))_{t=1}^T = (\mathbf{x}_t, y_t)_{t=1}^T\right)\\ &= \prod_{t=1}^{T} \Pr(\mathbf{X}(t)=\mathbf{x}_t) \Pr\left(Y_{i(t)}(t)=y_t \;\middle|\; \begin{matrix}(\mathbf{X}(\tau), Y_{i(\tau)}(\tau))_{\tau=1}^{t-1}\\ =(\mathbf{x}_\tau, y_\tau)_{\tau=1}^{t-1}\end{matrix}\right)\end{aligned} \tag{9}$$

by using the chain rule for probability and noting that $\mathbf{X}(t)$ is independent from $(\mathbf{X}(\tau), Y_{i(\tau)}(\tau))_{\tau=1}^{t-1}$ and $Y_{i(t)}(t)$. The same calculation can be made for $\Pr'$. The rest of the proof follows standard manipulations and we omit it due to space limitation. □

We now prove Theorem 5.1.

Proof. Fix an arm $j \neq i^*$ with $\alpha_j > \alpha_{i^*}\beta_{i^*}$ and fix $\lambda \in (\Delta_j, \alpha_j(1-\beta_j))$. We construct a new problem instance given by a second set of parameters $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}' = \boldsymbol{\alpha}$, $\beta_j' = \beta_j + \lambda/\alpha_j$, and $\beta_i' = \beta_i$ for all $i \neq j$. Note that setting $\lambda = \alpha_j\left(1-\beta_j\right)$ gives $\beta_j' = 1$, and $\Delta_j < \alpha_j\left(1-\beta_j\right)$ since $\alpha_j > \alpha_{i^*}\beta_{i^*}$. Then choosing any $\lambda \in \left(\Delta_j, \alpha_j\left(1-\beta_j\right)\right]$ indeed gives valid parameters $\alpha_i', \beta_i' \in [0,1]$ for all $i \in [N]$.

Observe that $\alpha_j'\beta_j' = \alpha_j\beta_j + \lambda > \alpha_{i^*}\beta_{i^*}$ and therefore $j \in \operatorname{argmax}_{i\in[N]} \alpha_i'\beta_i'$ is an optimal arm under the bandit setting with parameters $\boldsymbol{\alpha}', \boldsymbol{\beta}'$ and the reward gaps $\Delta_i' \triangleq \alpha_j'\beta_j' - \alpha_i\beta_i$ satisfy

$$\Delta_i' = \alpha_j\beta_j + \lambda - \alpha_i\beta_i = \lambda - \left(\alpha_i\beta_i - \alpha_j\beta_j\right) \geq \lambda - \Delta_j, \quad \forall i \neq j. \tag{10}$$

Let $R'(T) \triangleq T\alpha_j'\beta_j' - \sum_{t=1}^T \mathbb{E}'\left[Z_{i(t)}(t)\right]$ denote the cumulative regret under the bandit setting with parameters $\boldsymbol{\alpha}', \boldsymbol{\beta}'$. Then

$$\begin{aligned}R'(T) &= \sum_{i\neq j} \Delta_i' \mathbb{E}'\left[n_i(T)\right] \geq \left(\lambda - \Delta_j\right)\left(T - \mathbb{E}'\left[n_j(T)\right]\right),\\ R(T) &= \sum_{i\neq i^*} \Delta_i \mathbb{E}\left[n_i(T)\right] \geq \Delta_j \mathbb{E}\left[n_j(T)\right]\end{aligned} \tag{11}$$

Define the event $\mathcal{E} \triangleq \left\{n_j(T) < T/2\right\}$. Then from the above, we have $R'(T) \geq \frac{1}{2}\left(\lambda - \Delta_j\right) T \Pr'(\mathcal{E})$ and $R(T) \geq \frac{1}{2}\Delta_j T \Pr\left(\mathcal{E}^{\mathrm{c}}\right)$. Therefore

$$R'(T) + R(T) \geq \frac{1}{2} T \min\left\{\left(\lambda - \Delta_j\right), \Delta_j\right\}\left[\Pr\left(\mathcal{E}^{\mathrm{c}}\right) + \Pr'(\mathcal{E})\right]. \tag{12}$$

Combining the Bretagnolle-Huber inequality and an inequality due to Tsybakov gives a bound on the total variation distance:

$$D_{\mathrm{TV}}\left(\Pr, \Pr'\right) = \sup_A \left|\Pr(A) - \Pr'(A)\right| \leq 1 - \frac{1}{2} e^{-D_{\mathrm{KL}}\left(\Pr \parallel \Pr'\right)} \tag{13}$$

Using $\Pr(\mathcal{E}) - \Pr'(\mathcal{E}) \leq D_{\mathrm{TV}}\left(\Pr, \Pr'\right)$ with the above bound and rearranging gives

$$\Pr\left(\mathcal{E}^{\mathrm{c}}\right) + \Pr'(\mathcal{E}) \geq \frac{1}{2} e^{-D_{\mathrm{KL}}\left(\Pr \parallel \Pr'\right)} \overset{(a)}{=} \frac{1}{2} e^{-d\left(\beta_j, \beta_j'\right)\mathbb{E}\left[n_j(T)\right]}, \tag{14}$$

where (a) is from Lemma A.1 and noting that $d^{(X)}(\boldsymbol{\alpha}, \boldsymbol{\alpha}') = 0$ since $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ and each $d\left(\beta_i, \beta_i'\right) = 0$ since $\beta_i = \beta_i'$ for $i \neq j$. Substituting (14) into (12) and rearranging gives

$$e^{d\left(\beta_j, \beta_j'\right)\mathbb{E}\left[n_j(T)\right]} \geq \frac{\frac{1}{4} T \min\left\{\left(\lambda - \Delta_j\right), \Delta_j\right\}}{R'(T) + R(T)}. \tag{15}$$

Taking the logarithm of both sides, dividing by $\log T$ and rearranging gives

$$\begin{aligned}\frac{\mathbb{E}\left[n_j(T)\right]}{\log T} &\geq \frac{1}{d\left(\beta_j, \beta_j'\right)}\\ &\quad + \frac{\log\left(\frac{1}{4}\min\left\{\left(\lambda - \Delta_j\right), \Delta_j\right\}\right)}{d\left(\beta_j, \beta_j'\right)\log T} - \frac{\log\left(R'(T) + R(T)\right)}{d\left(\beta_j, \beta_j'\right)\log T}.\end{aligned} \tag{16}$$

Taking the limit inferior as $T \to \infty$ of both sides, we need only inspect the last term on the right-hand side. Recall that $R(T) =$

$o(T^\delta) \quad \forall\, \delta > 0$. Then $\forall \delta > 0$, there exists a constant $C_\delta > 0$ such that $R'(T) + R(T) \le C_\delta T^\delta$ and therefore

$$\limsup_{T\to\infty} \frac{\log\left(R'(T)+R(T)\right)}{\log T} \le \limsup_{T\to\infty} \frac{\log\left(C_\delta\right)+\delta\log T}{\log T} = \delta. \tag{17}$$

Since δ can be taken arbitrarily small and positive, it follows that

$$\limsup_{T\to\infty} \frac{\log\left(R'(T)+R(T)\right)}{\log T} = 0. \tag{18}$$

and so $\liminf_{T\to\infty} \frac{\mathbb{E}[n_j(T)]}{\log T} \ge \frac{1}{d\left(\beta_j,\beta'_j\right)}$. Since this holds for all $\lambda \in \left(\Delta_j, \alpha_j\left(1-\beta_j\right)\right]$, taking the infimum over λ gives the result. □

B PROOF OF THEOREM 5.2

Recall that $n_i(T)$ denotes the number of times that portion i is selected for transmission within the given time horizon T. We only need to bound $\mathbb{E}[n_i(T)]$ for each suboptimal arm $i \ne i^*$ and eventually bound the total cumulative regret using (1).

We consider a fixed suboptimal arm $i \ne i^*$ and bound $\mathbb{E}[n_i(T)]$ for this fixed arm $i \ne i^*$. We begin by selecting a positive constant $\epsilon_3 > 0$. The choice of this parameter depends on the relation between α_i and $\alpha_{i^*}\beta_{i^*}$ of this suboptimal arm $i \ne i^*$ as follows.

Fix $\epsilon_3 \in (0, \epsilon_3^{\max})$ where $\epsilon_3^{\max} \triangleq \begin{cases}\Delta_i & \alpha_i \ge \alpha_{i^*}\beta_{i^*}\\ \alpha_{i^*}\beta_{i^*}-\alpha_i & \alpha_i < \alpha_{i^*}\beta_{i^*}.\end{cases}$

Then define the intervals $\Gamma_1 \triangleq \left(0, \epsilon_1^{\max}\right)$ and $\Gamma_2 \triangleq \left(0, \frac{\epsilon_3}{\beta_{i^*}}\right)$ where $\epsilon_1^{\max} \triangleq \frac{\sqrt{(\alpha_i+\beta_i)^2+4(\Delta_i-\epsilon_3)}-(\alpha_i+\beta_i)}{2}$ and fix $\epsilon_1 \in \Gamma_1$ and $\epsilon_2 \in \Gamma_2$. Note that $\epsilon_1^{\max}$ is a zero of the quadratic function $q_i(\epsilon_1) = \epsilon_1^2 + (\alpha_i+\beta_i)\,\epsilon_1 - \left(\Delta_i - \epsilon_3\right)$. Then since $\epsilon_1 < \epsilon_1^{\max}$, we have $q_i(\epsilon_1) < 0$ due to the convexity of $q_i(\epsilon_1)$. Rearranging gives $\epsilon_1^2 + (\alpha_i+\beta_i)\,\epsilon_1 < \Delta_i - \epsilon_3$, and adding $\alpha_i\beta_i$ to both sides and factoring the left-hand side gives $(\alpha_i+\epsilon_1)(\beta_i+\epsilon_1) < \alpha_{i^*}\beta_{i^*} - \epsilon_3$. Combining this lower bound with the upper bound $\alpha_{i^*}\beta_{i^*} - \epsilon_3 < (\alpha_{i^*}-\epsilon_2)\beta_{i^*}$, which follows directly from the definition of Γ_2 and the fact that $\epsilon_2 \in \Gamma_2$, gives

$$(\alpha_i+\epsilon_1)(\beta_i+\epsilon_1) < \alpha_{i^*}\beta_{i^*} - \epsilon_3 < (\alpha_{i^*}-\epsilon_2)\beta_{i^*}. \tag{19}$$

To facilitate the upper bound analysis of $\mathbb{E}\left[n_i(T)\right]$, we first define the following three events:

Definition B.1. (Events $E_i^\mu(t)$, $E_{i^*}^\mu(t)$, and $E_i^\theta(t)$)

Let $\epsilon_1, \epsilon_2, \epsilon_3$ be chosen as described above. For each time step t, we define the following events:

(1) $E_i^\mu(t) \triangleq \left\{|\overline{\alpha}_i(t) - \alpha_i| \le \epsilon_1,\ \left|\overline{\beta}_i(t) - \beta_i\right| \le \epsilon_1\right\}$, where

$$\overline{\beta}_i(t) = \begin{cases}\frac{1}{n_i(t-1)}\sum_{s=1}^{n_i(t-1)} Y_i(\tau_s), & \text{if } n_i(t-1) > 0\\ 0, & \text{otherwise}\end{cases},$$

and τ_s denotes the timeslot that the fixed suboptimal arm i was played for the s-th time.

(2) $E_{i^*}^\mu(t) \triangleq \left\{|\overline{\alpha}_{i^*}(t) - \alpha_{i^*}| \le \epsilon_2\right\}$.

(3) $E_i^\theta(t) \triangleq \left\{\overline{\alpha}_i(t)\,\theta_{\beta,i}(t) \le \alpha_{i^*}\beta_{i^*} - \epsilon_3\right\}$.

We next bound the expected number of plays of a fixed suboptimal arm i by analyzing the probability of selecting this arm at each timeslot. Consider the four events (i) $E_i^\mu(t) \cap E_i^\theta(t) \cap E_{i^*}^\mu(t)$, (ii) $E_i^\mu(t) \cap E_i^\theta(t) \cap \overline{E_{i^*}^\mu(t)}$, (iii) $E_i^\mu(t) \cap \overline{E_i^\theta(t)}$, and (iv) $\overline{E_i^\mu(t)}$. These events are mutually exclusive and collectively exhaustive, and moreover $E_i^\mu(t) \cap E_i^\theta(t) \cap \overline{E_{i^*}^\mu(t)} \subseteq \overline{E_{i^*}^\mu(t)}$. By the monotonicity of probability and the law of total probability, the probability of selecting arm i at each timeslot can thus be decomposed into cases (i)–(iv) and upper bounded separately. Building upon (19), we present the following lemma that provides the corresponding upper bounds.

Due to limited space, intermediate steps are omitted. Interested readers may refer to our technical report [25] for the full derivations.

Lemma B.2. *The cumulative probability of playing the suboptimal arm i across horizon T, decomposed according to events (i)–(iv), can be upper bounded as*

$$\sum_{t=1}^{T} \Pr\left(i(t)=i, E_i^\mu(t), E_i^\theta(t), E_{i^*}^\mu(t)\right) \le M_1,$$

$$\sum_{t=1}^{T} \Pr\left(i(t)=i, E_i^\mu(t), \overline{E_i^\theta(t)}\right) \le \begin{cases}\inf_{\epsilon_1\in\Gamma_1} \frac{\log T}{d(\beta_i+\epsilon_1, \frac{\alpha_{i^*}\beta_{i^*}-\epsilon_3}{\alpha_i+\epsilon_1})} + 1 & \alpha_i \ge \alpha_{i^*}\beta_{i^*}\\ M_2 & \alpha_i < \alpha_{i^*}\beta_{i^*}\end{cases},$$

$$\sum_{t=1}^{T} \Pr\left(i(t)=i, \overline{E_i^\mu(t)}\right) \le M_3, \qquad \sum_{t=1}^{T} \Pr\left(\overline{E_{i^*}^\mu(t)}\right) \le M_4,$$

where $M_1 \triangleq \inf_{\epsilon_2\in\Gamma_2}\left(\frac{24}{\Delta'^2} + \Theta\left(\frac{1}{\Delta'^2} + \frac{1}{\Delta'^2 D} + \frac{1}{\Delta'^4}\right)\right)$, $\Delta' = \beta_{i^*} - x, D \triangleq d\left(x, \beta_{i^*}\right) = x\log\frac{x}{\beta_{i^*}} + (1-x)\log\frac{1-x}{1-\beta_{i^*}}, x \triangleq \frac{\alpha_{i^*}\beta_{i^*}-\epsilon_3}{\alpha_{i^*}-\epsilon_2}, M_2 \triangleq \inf_{\epsilon_4\in\Gamma_4}\frac{1}{2\epsilon_4^2} + 1$, $\Gamma_4 \triangleq (0, \alpha_{i^*}\beta_{i^*} - \alpha_i - \epsilon_3)$, $M_3 \triangleq \inf_{\epsilon_1\in\Gamma_1}\frac{2}{\epsilon_1^2} + 1$, $M_4 \triangleq \inf_{\epsilon_2\in\Gamma_2}\frac{1}{\epsilon_2^2} + 1$.

Finally, we can bound $\mathbb{E}[n_i(T)]$ for the fixed arm i by incorporating lemma B.2 as follows.

$$\mathbb{E}[n_i(T)] \le \begin{cases}\inf\limits_{\epsilon_1\in\Gamma_1} \frac{\log T}{d\left(\beta_i+\epsilon_1, \frac{\alpha_{i^*}\beta_{i^*}-\epsilon_3}{\alpha_i+\epsilon_1}\right)} & \alpha_i \ge \alpha_{i^*}\beta_{i^*},\\ \quad + M_1 + M_3 + M_4 + 1 & \\ M_1 + M_2 + M_3 + M_4, & \alpha_i < \alpha_{i^*}\beta_{i^*}.\end{cases} \tag{20}$$

When $\alpha_i < \alpha_{i^*}\beta_{i^*}$, by noting that M_1, M_2, M_3, M_4 are some constants, it is easy to see after dividing $\log T$ and taking the limit superior to both sides of above gives $\limsup_{T\to\infty}\frac{\mathbb{E}[n_i(T)]}{\log T} = 0$.

When $\alpha_i \ge \alpha_{i^*}\beta_{i^*}$, dividing by $\log T$ and taking the limit superior on both sides of the above gives $\limsup_{T\to\infty}\frac{\mathbb{E}[n_i(T)]}{\log T} \le \limsup_{T\to\infty}\inf_{\epsilon_1\in\Gamma_1}\frac{1}{d(\beta_i+\epsilon_1, \frac{\alpha_{i^*}\beta_{i^*}-\epsilon_3}{\alpha_i+\epsilon_1})} = \frac{1}{d(\beta_i, \frac{\alpha_{i^*}\beta_{i^*}-\epsilon_3}{\alpha_i})}$.

Since ϵ_3 can be taken arbitrarily small and positive, we conclude that $\limsup_{T\to\infty}\frac{\mathbb{E}[n_i(T)]}{\log T} \le \frac{1}{d(\beta_i, \frac{\alpha_{i^*}\beta_{i^*}}{\alpha_i})}$.

Noting that we only do the summation over the suboptimal arms that $\alpha_i > \alpha_{i^*}\beta_{i^*}$ and not include $\alpha_i = \alpha_{i^*}\beta_{i^*}$ in (7), this is because for arm i satisfying $\alpha_i = \alpha_{i^*}\beta_{i^*}$, $\limsup_{T\to\infty}\frac{\mathbb{E}[n_i(T)]}{\log T} \le \limsup_{T\to\infty}\inf_{\epsilon_1\in\Gamma_1}\frac{1}{d(\beta_i+\epsilon_1, \frac{\alpha_{i^*}\beta_{i^*}-\epsilon_3}{\alpha_i+\epsilon_1})} = \frac{1}{d(\beta_i, \frac{\alpha_i-\epsilon_3}{\alpha_i})}$.

Similarly, since ϵ_3 can be taken arbitrarily small and positive, we conclude that $\limsup_{T\to\infty}\frac{\mathbb{E}[n_i(T)]}{\log T} = 0$.