Published: 27 October 2025

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

RESEARCH-ARTICLE

Progressive Learning with Human Feedback for Personalized Adaptive Video Streaming

ZHAOHUI JIANG, Shanghai Jiao Tong University, Shanghai, China

XUENING FENG, Shanghai Jiao Tong University, Shanghai, China

TIANCHI HUANG, Tsinghua University, Beijing, China

RUIXIAO ZHANG, ByteDance Ltd., Beijing, China

PAUL WENG, Duke Kunshan University, Kunshan, Jiangsu, China

YIFEI ZHU, Shanghai Jiao Tong University, Shanghai, China

Open Access Support provided by:

Tsinghua University

ByteDance Ltd.

Duke Kunshan University

Shanghai Jiao Tong University

MM '25: Proceedings of the 33rd ACM International Conference on Multimedia (October 2025)
https://doi.org/10.1145/3746027.3755323
ISBN: 9798400720352

Progressive Learning with Human Feedback for Personalized Adaptive Video Streaming

Zhaohui Jiang
Shanghai Jiao Tong University
Shanghai, China
jiangzhaohui@sjtu.edu.cn

Xuening Feng
Shanghai Jiao Tong University
Shanghai, China
cindyfeng2019@sjtu.edu.cn

Tianchi Huang
Tsinghua University
Beijing, China
mythkast@hotmail.com

Ruixiao Zhang
ByteDance Inc.
San Diego, United States
ruixiao.cs.zhang@gmail.com

Paul Weng
Duke Kunshan University
Kunshan, China
paul.weng@dukekunshan.edu.cn

Yifei Zhu*
Shanghai Jiao Tong University
Shanghai, China
yifei.zhu@sjtu.edu.cn

Abstract

Existing quality of experience (QoE)-driven adaptive bitrate (ABR) algorithms either fail to consider personalized QoE or rely on oversimplified QoE models, all resulting in unsatisfactory streaming experiences. Recognizing the wide existence of user feedback schemes in existing streaming applications, we introduce `Q+`, a framework leveraging progressively gathered personal user opinion scores from multiple interaction sessions for enhanced user-system alignment. `Q+` first innovates QoE modeling by incorporating both pairwise ordinal and cardinal preferences constructed from scores. The capturing of both preferences ensures reliable and robust preference representation. Moreover, we design a monotonic neural network as the QoE model to capture the inherent monotonicity property in ABR services, improving model expressivity and generalization ability even with limited human feedback. To align the policy with the progressively updated QoE, we then develop a value-based reinforcement learning (RL) algorithm for bitrate control that integrates reward relabeling and calibrated prioritized experience replay. Extensive experiments reveal that `Q+` consistently surpasses state-of-the-art rule-based, control-based, and RL-based baselines within only three sessions, improving QoE by 5.69% to 29.39% across diverse network conditions.

CCS Concepts

• **Information systems → Multimedia streaming**.

Keywords

Adaptive Video Streaming, Personalized Quality of Experience, Reinforcement Learning from Human Feedback, Online Learning

*Corresponding author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755323

ACM Reference Format:
Zhaohui Jiang, Xuening Feng, Tianchi Huang, Ruixiao Zhang, Paul Weng, and Yifei Zhu. 2025. Progressive Learning with Human Feedback for Personalized Adaptive Video Streaming. In *Proceedings of Proceedings of the 33rd ACM International Conference on Multimedia (MM '25).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755323

1 Introduction

Video streaming services, constituting about 60% of the Internet traffic [37], have become an indispensable part of modern life. In this context, Dynamic Adaptive Streaming over HTTP (DASH) [42] has emerged as a leading standard, segmenting videos into small chunks on the server side and encoding them at various bitrates. Adaptive bitrate (ABR) algorithms then select appropriate bitrates to ensure high-quality video experiences in fluctuating networks.

While Quality of Experience (QoE)-driven ABR algorithms aim to enhance user satisfaction by maximizing QoE [17, 30, 43, 48], most of them rely on oversimplified *general QoE* general models, which aggregate preferences of a wide population. However, studies [12, 21, 33] reveal significant individual variability in user expectations, highlighting the need for a *personalized QoE*-driven ABR framework that encompasses personalized QoE modeling and corresponding policy learning. Existing personalized approaches [19, 27, 34, 48] rely on manual user configuration and only depend on limited QoE parameters (usually fewer than three). This leads to narrow modeling spaces that fail to model the QoE accurately, resulting in the derived policies being misaligned with actual user experience. Furthermore, current methods for accurate QoE modeling [11, 17, 35, 45] rely on months of dataset preparation, which are too time-consuming and costly to encourage widespread user participation for personalized QoE modeling.

Therefore, to obtain personal user-aligned ABR policies in an accurate and scalable manner, we introduce a progressive personalized QoE-policy refinement process that aligns ABR policies with evolving personalized QoE models iteratively, where in each *training iteration*, the personalized QoE is first updated with new user feedback, and then the policy is fine-tuned to optimize with the revised QoE rewards. Such a progressive learning scheme can be seamlessly integrated into modern streaming platforms (e.g., Netflix [3], YouTube [5], Amazon Prime Video [2]) by leveraging user registration/login and explicit feedback (e.g., star ratings [2], pop-up surveys [5]). In these applications, during each *interaction session*, which is a distinct period of continuous user engagement with the platform, users may rate the streaming quality for viewed video experiences at their discretion as new feedback.

For personalized QoE modeling, two challenges are: (i) learning reliable QoE from *cross-session feedback* (i.e., feedback from multiple interaction sessions), and (ii) ensuring model expressivity while avoiding misleading predictions with limited feedback. For the first challenge, our analysis reveals that traditional regression-based methods [11, 18, 33, 35] are unreliable for cross-session feedback due to score distortions. In contrast, pairwise ordinal and cardinal comparisons constructed from these scores demonstrate greater consistency, where the two kinds of comparisons capture preferences between video experiences and the strength of these preferences, respectively. We utilize both as learning sources to fully exploit the information inside the human feedback for reliable QoE modeling. For the second challenge, while MLP-based QoE models excel in expressivity [17, 46], they risk generating unreasonable predictions for unseen data, destabilizing policy performance if used as rewards [17]. To address this, we employ monotonic neural networks [36], embedding prior knowledge of monotonic relationships (e.g., rebuffering time vs. perceptual quality) to avoid unreasonable predictions and improve generalization for reward modeling.

For personalized QoE-driven policy learning, the challenge lies in ensuring effective and efficient training. To guarantee that repeating training iterations leads to better policies aligning with personal human expectations, we employ an *online* Reinforcement Learning from Human Feedback (RLHF) framework [9, 14, 16], where feedback is provided on video experiences rendered by the online ABR policy being trained. For sample-efficient fine-tuning under progressive QoE updates, we propose a value-based deep reinforcement learning (DRL) algorithm with two key designs: (i) rewards relabeling for historical trajectories in the replay buffer for reuse, and (ii) calibrated prioritized replay to speed up training.

Our main contributions can be summarized as follows:

- We introduce a user-friendly interaction scheme capturing personal preferences for ABR by gathering user feedback progressively across multiple interaction sessions (Section 3).
- We develop a reliable and efficient QoE modeling method using pairwise ordinal and cardinal preferences to learn from cross-session feedback, enhanced by monotonicity constraints for generalizable results (Section 4).
- We develop an efficient value-based DRL approach that fine-tunes policies along with personalized QoE model updates, progressively aligning with true user preferences (Section 5).
- Integrating our QoE modeling with the personalized QoE-driven ABR policy learning results in a whole framework, named **Q**-learning-based **P**ersonalized **L**earning for **U**ser-centric ABR **S**ystem (Q-PLUS, `Q+`), outperforming state-of-the-art rule-based, MPC-based, and DRL-based baselines across diverse network conditions (Section 6).

2 Background and Motivation

Scenarios. We focus on a common scenario where video platforms track user identities, typically through registration or login, enabling a progressive feedback collection to build datasets with *cross-session feedback* gathered from different interaction sessions.

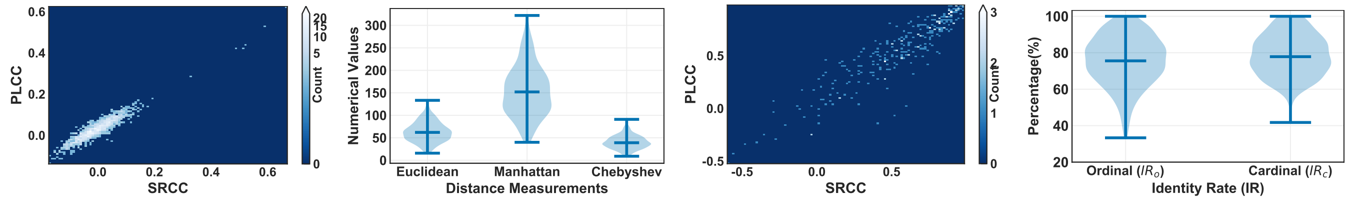
Dataset. Our subsequent analysis employs `SQoE-IV` [12], the latest published subjective QoE dataset, which is well-organized and comprises 1,350 video experiences, each rated by 31 human evaluators on a 1-100 scale. The 4.5-hour scoring process is split into three *testing sessions* for each evaluator to reduce human fatigue. To mitigate session-specific biases (e.g., mood or cognitive fluctuations) by cross-session score normalization, a *calibration phase* with 10 *overlapping video experiences* is included in each session, with each overlapping video experience presented in two sessions and yielding two scores. The structure of `SQoE-IV` mirrors the streaming scenarios where *evaluators* as *users*, and *testing sessions* as *interaction sessions*.

Necessity of personalized QoE. While the cardinal aspect of *user heterogeneity*, where scores from different users show significantly varied numerical distributions [12, 17, 33, 46], is widely recognized, our analysis reveals its ordinal aspect as well: users exhibit diverse preference rankings. This is evidenced by Figure 1a, showing low SRCC and PLCC values (mostly <0.2) for different users' scores, indicating weak rank and numerical correlations between different users. Thus, QoEs derived from population-level cardinal scores (i.e., mean opinion score) [11, 18, 33, 35] or ordinal comparisons [17] all fail to align with personal perceptions, underscoring the need for personalized QoE modeling from individual feedback.

Necessity to learn personalized QoE progressively. Subjective score is an ideal form of human feedback for accurate QoE modeling. Still, existing methods [11, 17, 35, 45] rely on datasets that being pre-collected systematically (e.g., `SQoE-IV`), requiring at least 50 video experiences with more than 25 minutes of time cost per user [33], making them impractical for scalable personalized QoE modeling for large populations. This motivates our *progressive* interaction scheme, where users can provide feedback gradually during many interaction sessions, enabling user-friendly feedback acquisition, therefore scalable for large populations.

Necessity to avoid learning directly from numerical scores. Cross-session subjective scores have significant numerical distortion, as indicated by Figure 1b via various distance measurements on overlapping video experiences's cross-session scores. It reveals that even the same video experience can receive vastly different scores across sessions, suggesting a higher score in one session does not necessarily indicate stronger preferences than lower scores in another. Such distortion is hard to mitigate in real-world scenarios, due to the impracticality of calibration phases. Therefore, traditional methods [11, 17, 18, 35] that model QoE by directly approximating numerical scores are fundamentally flawed with cross-session datasets.

Necessity to learn from pairwise preferences. Despite numerical distortions, there are strong rank and numerical correlations between overlapping video experiences's cross-session scores as shown in Figure 1c, evidenced by high SRCC and PLCC values (mostly >0.6). This motivates the use of cross-session feedback via rank-related preferences for reliable QoE modeling, leveraging preference-based learning techniques that learn from pairwise preferences [9, 13, 25]. Inspired by Huang et al. [17], instead of



(a) PLCCs and SRCCs for pairs-of-evaluators' scores on all video experiences. **(b) Distances measurements for overlapping video experiences' cross-session scores.** **(c) PLCCs and SRCCs for overlapping video experiences' cross-session scores.** **(d) Ratio of identity ordinal and cardinal preferences across sessions for overlapping video experiences.**

Figure 1: Data analysis of SQoE-IV's subjective scores from all evaluators. Figures 1a and 1c are 2D histograms with colors indicating counts. Figures 1b and 1d are violin plots with bars representing the max, min, and median values, and shadows indicating estimated distribution. Mathematical descriptions can be checked in Appendix A [4].

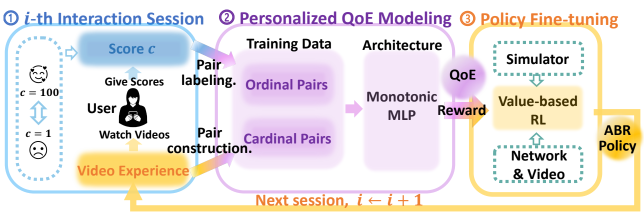


Figure 2: An overview of the proposed framework.

directly letting users provide preferences, a *preference* for a *pairwise comparison* is determined by numerical responses generated from subjective scores, where an object with a higher response is preferred over one with a lower response. Specifically, an *ordinal comparison* [9] is constructed by a pair of video experiences, with the video experience as the object and its score as the response to determine the *ordinal preference*, and a *cardinal comparison* [13] is constructed by a pair of ordinal comparisons, with the ordinal comparison as the object and its absolute score difference (i.e., the strength of an ordinal preference) as the response to determine the *cardinal preference*. Figure 1d shows that most ordinal and cardinal comparisons receive the same preference labels across different sessions, demonstrating strong consistency and making them appropriate learning sources within the progressive scheme.

3 Progressive Learning Scheme for Adaptive Video Streaming

As shown in Figure 2, Q+ progressively refines the QoE and ABR policy by iterating a three-phase cycle: (i) During each user-system interaction sessions, users are prompted to rate their video experiences, (ii) Personalized QoE modeling based on collected user feedback (see Section 4), and (iii) ABR policy fine-tuning with the updated personalized QoE as a reward model (see Section 5).

Concretely, the system receives new human feedback $\mathcal{D}^c_i$ during the i-th interaction session, consisting of video experiences $\{\tau\}$ generated from the latest ABR policy $\pi_{\theta_{i-1}}$ and corresponding scores $\{c\}$. The user's *training personalized QoE* R_ϕ is then updated to R_{ϕ_i} by training on feedback from all sessions $\mathcal{D}^c$ (Section 4), where $\mathcal{D}^c = \cup_{j\leq i}\mathcal{D}^c_j$. Since the *ground-truth personalized QoE* R^* is inaccessible due to the lack of a systematic subjective test, our

Algorithm 1 Overview: Q+ with Progressive Interaction Scheme

Initial policy: π_{θ_0}
Dataset saving user's feedback: $\mathcal{D}^c \leftarrow \emptyset$
Total number of interaction sessions: N_I
for $i \leftarrow 1$ to N_I **do** ▷ Repeat training iteraions.
  ▷ Feedback collection during interaction session.
  $\mathcal{D}^c_i \leftarrow \emptyset$
  while User e is using the video streaming service **do**
    Video experiences τ generated with $\pi_{\theta_{i-1}}$
    if User e provides a score c for τ **then**
      $\mathcal{D}^c_i \leftarrow \mathcal{D}^c_i \cup \{(\tau, c)\}$
    end if
  end while
  $\mathcal{D}^c \leftarrow \mathcal{D}^c \cup \mathcal{D}^c_i$ ▷ Cross-session feedback.
  ▷ QoE modeling (Algorithm 2 in [4].)
  Train personalized QoE R_{ϕ_i} for user e with $\mathcal{D}^c$
  ▷ QoE-driven policy fine-tuning (Algorithm 3 in [4].)
  Obtain π_{θ_i} by fine-tuning $\pi_{\theta_{i-1}}$ with R_{ϕ_i}
end for
return π_{N_I}

policy learning adaptively adjusts $\pi_{\theta_{i-1}}$ to π_{θ_i} by maximizing R_{ϕ_i}. Repeating such iterations as outlined in Algorithm 1, the policy is expected to align with user preferences gradually, ultimately enhancing user satisfaction.

4 Progressive Personalized QoE Modeling

Formulating a video experience τ with H chunks as $(s_0, a_0, s_1, a_1, \ldots, s_H, a_H)$, where s_t refers to the state at timestep t containing information for bitrate selection, and a_t is the target bitrate selected by an ABR policy π given s_t for downloading the t-th chunk. Our QoE modeling aims to learn a QoE model R_ϕ to predict scalar rewards for states such that $R_\phi(\tau) = \frac{1}{H}\sum_{t=0}^{H} R_\phi(s_t)$ can reflect user preferences for different video experiences. We explain our QoE modeling from the perspective of model training and model architecture design in Section 4.1 and Section 4.2, respectively.

4.1 Preference-Based QoE Training

To maximize the utility of limited human feedback in the progressive learning scheme, we train our QoE model using both ordinal

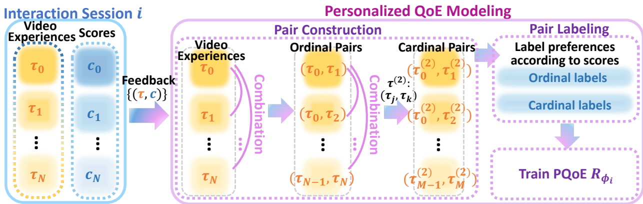


Figure 3: Illustration for the personalized QoE modeling.

and cardinal pairwise comparisons, leveraging the consistency of these preferences across sessions. Our approach is adapted from reward training in preference-based RLHF methods [9, 13], specifically tailored for progressive learning with cross-session datasets.

Pair construction and labeling. An *ordinal comparison* [9] is constructed by a pair of video experiences $\tau^{(2)} = (\tau_j, \tau_k)$, where j and k are used in subscript to distinguish different objects. The corresponding *ordinal preference* $y_o \in \{=, \succ, \prec\}$ describes which one of the video experiences is being preferred over another:

$$y_o(\tau_j, \tau_k) = \begin{cases} =, & \text{if } |c_j - c_k| < \delta_o, \\ \succ, & \text{if } c_j > c_k + \delta_o, \\ \prec, & \text{if } c_j < c_k - \delta_o, \end{cases} \tag{1}$$

where $\succ$ (resp. $\prec$) means τ_j (resp. τ_k) is preferred than τ_k (resp. τ_j), $=$ means equal preference, and δ_o is a positive constant indicating the allowed difference for equal ordinal preference to account for label noise. In addition, a *cardinal comparison* [13] is constructed by a pair of ordinal comparisons $(\tau_j^{(2)}, \tau_k^{(2)})$. The corresponding *cardinal preference* $y_c \in \{=, \succ, \prec\}$ describes which ordinal comparison has stronger preference strength based on absolute score differences $|c_j - c_k|$ (denoted as $\Delta_c(\tau^{(2)})$):

$$y_c(\tau_j^{(2)}, \tau_k^{(2)}) = \begin{cases} =, & \text{if } |\Delta_c(\tau_j^{(2)}) - \Delta_c(\tau_k^{(2)})| < \delta_c, \\ \succ, & \text{if } \Delta_c(\tau_j^{(2)}) > \Delta_c(\tau_k^{(2)}) + \delta_c, \\ \prec, & \text{if } \Delta_c(\tau_j^{(2)}) < \Delta_c(\tau_k^{(2)}) - \delta_c, \end{cases} \tag{2}$$

where δ_c is a constant indicating the allowed difference for equal cardinal preference. As illustrated in Figure 3, ordinal and cardinal comparisons are formed with video experiences within the same session, and preferences are labeled based on their associated scores.

Training method. Given preference labels, the *target ordinal (resp. cardinal) preference probability* P_o^* (resp. P_c^*) for an ordinal (resp. cardinal) comparison, representing the true probability of one object being preferred over another, is derived as Equation (3) (resp. Equation (4)):

$$P_o^*(\tau_j, \tau_k) = \begin{cases} (0.5, 0.5), & \text{if } \tau_j = \tau_k, \\ (1, 0), & \text{if } \tau_j \succ \tau_k, \\ (0, 1), & \text{if } \tau_j \prec \tau_k, \end{cases} \tag{3}$$

$$P_c^*(\tau_j^{(2)}, \tau_k^{(2)}) = \begin{cases} (0.5, 0.5), & \text{if } \tau_j^{(2)} = \tau_k^{(2)}, \\ (1, 0), & \text{if } \tau_j^{(2)} \succ \tau_k^{(2)}, \\ (0, 1), & \text{if } \tau_j^{(2)} \prec \tau_k^{(2)}. \end{cases} \tag{4}$$

Then, following Christiano et al. [9], the *modeled ordinal preference probability* is derived using the reward model R based on the Bradley-Terry model [8]:

$$\mathbb{P}(\tau_j \succ \tau_k; R) = \frac{\exp(R(\tau_j))}{\exp(R(\tau_j)) + \exp(R(\tau_k))}. \tag{5}$$

This defines the *ordinal loss* $\mathcal{L}_\phi^o$, which optimizes R_ϕ by minimizing the cross-entropy loss between $\mathbb{P}(\tau_j \succ \tau_k; R_\phi)$ and $P_o^*(\tau_j, \tau_k)$:

$$\mathcal{L}_\phi^o(\tau^{(2)}) = -\Big[P_o^*[0] \log \mathbb{P}(\tau_j \succ \tau_k; R_\phi) + P_o^*[1] \log \mathbb{P}(\tau_j \prec \tau_k; R_\phi)\Big], \tag{6}$$

where $P_o^*[0]$ and $P_o^*[1]$ refer to the two value of $P_o^*(\tau_j, \tau_k)$.

Following Feng et al. [13], the *modeled cardinal preference probability* is also formulated with the Bradley-Terry model, based on the reward difference $|R(\tau_j) - R(\tau_k)|$ as $\Delta_R(\tau^{(2)})$:

$$\mathbb{P}(\tau_j^{(2)} \succ \tau_k^{(2)}; R) = \frac{\exp(\Delta_R(\tau_j^{(2)}))}{\exp(\Delta_R(\tau_j^{(2)})) + \exp(\Delta_R(\tau_k^{(2)}))}. \tag{7}$$

This leads to the *cardinal loss* $\mathcal{L}_\phi^c$ that optimizes R_ϕ by minimizing a cross entropy loss between $\mathbb{P}(\tau_j^{(2)} \succ \tau_k^{(2)}; R_\phi)$ and $P_c^*(\tau_j^{(2)}, \tau_k^{(2)})$:

$$\mathcal{L}_\phi^c(\tau_j^{(2)}, \tau_k^{(2)}) = -\Big[P_c^*[0] \log \mathbb{P}(\tau_j^{(2)} \succ \tau_k^{(2)}; R_\phi) + P_c^*[1] \log \mathbb{P}(\tau_j^{(2)} \prec \tau_k^{(2)}; R_\phi)\Big]. \tag{8}$$

To effectively combine $\mathcal{L}_\phi^o$ (Equation (6)) and $\mathcal{L}_\phi^c$ (Equation (8)) for cross-session feedback, we specifically design our final QoE loss $\mathcal{L}_\phi$ that enforces concurrent optimization of ordinal and cardinal preferences on sampled cardinal comparisons:

$$\mathcal{L}_\phi = \mathbb{E}_{(\tau_j^{(2)}, \tau_k^{(2)}) \sim \mathcal{D}^c}\Big[\mathcal{L}_\phi^o(\tau_j^{(2)}) + \mathcal{L}_\phi^o(\tau_k^{(2)}) + \mathcal{L}_\phi^c(\tau_j^{(2)}, \tau_k^{(2)})\Big]. \tag{9}$$

Pseudo-code for our QoE modeling can be checked in Appendix E.2 [4].

4.2 Monotonic MLP as QoE Architecture

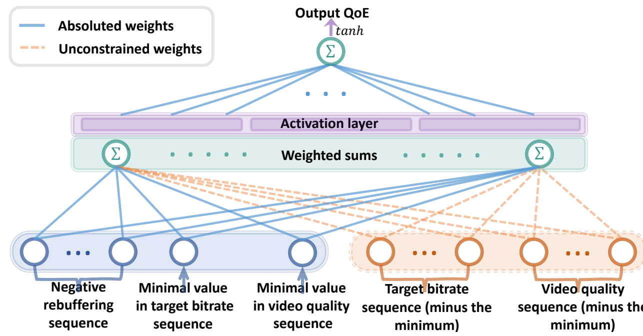


Figure 4: Monotonic MLP as the QoE model, illustrated with one hidden layer for clarity.

To obtain robust QoE outputs, we introduce *monotonicity constraints* to ensure monotonic relationships between input features and QoE predictions, preventing unreasonable QoE improvements under worsening conditions, thereby aligning with domain knowledge.

Model inputs. The input for our QoE model at timestep t, denoted as s_t, contains recent K chunks' rebuffering time $\boldsymbol{u}_t = \{u_{t-K+1}, \cdots, u_t\}$, target bitrate $\boldsymbol{q}_t = \{q_{t-K+1}, \cdots, q_t\}$, and video quality $\boldsymbol{v}_t = \{v_{t-K+1}, \cdots, v_t\}$ measuring by VMAF [1].

Monotonicity constraints. Then, given two states s'_t (with features $\{\boldsymbol{u}_{t'}, \boldsymbol{q}_{t'}, \boldsymbol{v}_{t'}\}$) and s''_t (with features $\{\boldsymbol{u}_{t''}, \boldsymbol{q}_{t''}, \boldsymbol{v}_{t''}\}$):

(1) *Monotonicity in rebuffering time:* Given $\boldsymbol{q}_{t'} = \boldsymbol{q}_{t''}$, $\boldsymbol{v}_{t'} = \boldsymbol{v}_{t''}$, if $\exists j \in \{t-K+1, \cdots, t\}$ such that $\boldsymbol{u}_{t'}[j] > \boldsymbol{u}_{t''}[j]$ and $\forall k \neq j, \boldsymbol{u}_{t'}[k] = \boldsymbol{u}_{t''}[k]$, then $R_\phi(s'_t) \leq R_\phi(s''_t)$.
(2) *Monotonicity in target bitrate:* Given $\boldsymbol{u}_{t'} = \boldsymbol{u}_{t''}$, $\boldsymbol{v}_{t'} = \boldsymbol{v}_{t''}$, if $\min(\boldsymbol{q}_{t'}) < \min(\boldsymbol{q}_{t''})$ and $(\boldsymbol{q}_{t'} - \min(\boldsymbol{q}_{t'})) = (\boldsymbol{q}_{t''} - \min(\boldsymbol{q}_{t''}))$, then $R_\phi(s'_t) \leq R_\phi(s''_t)$.
(3) *Monotonicity in video quality:* Given $\boldsymbol{u}_{t'} = \boldsymbol{u}_{t''}$, $\boldsymbol{q}_{t'} = \boldsymbol{q}_{t''}$, if $\min(\boldsymbol{v}_{t'}) < \min(\boldsymbol{v}_{t''})$ and $(\boldsymbol{v}_{t'} - \min(\boldsymbol{v}_{t'})) = (\boldsymbol{v}_{t''} - \min(\boldsymbol{v}_{t''}))$, then $R_\phi(s'_t) \leq R_\phi(s''_t)$.

For rebuffering time, constraint (i) reflects that increasing it for any specific chunk has a non-positive monotonic impact (i.e., either degrade or maintain) on QoE, aligning with empirical evidence that even a single rebuffering incident significantly disrupts user experience [10, 24]. For target bitrate or video quality, increased values of a single chunk do not guarantee a monotonic QoE response, as they can affect playback smoothness and therefore perceptual continuity [7]. Therefore, constraints (ii) and (iii) enforce non-negative monotonicity only for the minimum values in the feature sequence, allowing users to express preferences regarding quality fluctuations.

Monotonicity implementation. Our QoE model, named MonMLP, employs a monotonic neural network [36], integrating the monotonicity constraints seamlessly into standard MLPs by restricting specific first-layer weights and all subsequent-layer weights to be non-negative, as depicted in Figure 4. Concretely, for a state $\{\boldsymbol{u}_t, \boldsymbol{q}_t, \boldsymbol{v}_t\}$, MonMLP assigns absolute non-negative trainable weights to $\{-\boldsymbol{u}_t, \min(\boldsymbol{q}_t), \min(\boldsymbol{v}_t)\}$, ensuring monotonicity. The remaining state information, $\{\boldsymbol{q}_t - \min(\boldsymbol{q}_t), \boldsymbol{v}_t - \min(\boldsymbol{v}_t)\}$, is also included as inputs but with unconstrained first-layer weights. The QoE predictions are normalized to $(-1, 1)$ using a tanh function. Further implementation details are provided in Appendix E.1 [4].

5 Progressive Policy Learning

With the learning scheme overviewed in Section 3, we develop an RL-based method to effectively adapt $\pi_{\theta_{i-1}}$ (optimized for $R_{\phi_{i-1}}$) into π_{θ_i} in each training iteration, maximizing the updated QoE R_{ϕ_i} by treating QoE as the reward function. Section 5.1 reviews the fundamental algorithm design, serving as the backbone, then Section 5.2 details two novel modifications to the backbone algorithm for efficient progressive fine-tuning across training iterations.

5.1 Value-Based RL as Backbone

Given the frequent policy fine-tuning in each training iteration, sample efficiency (i.e., minimizing environment timesteps to achieve target performance) is crucial for scalability across a large user base. Value-based methods, with their inherent advantage in sample efficiency through trajectory storage and reuse via replay buffers, have become the preferred choice for discrete action tasks [39]. Thus, our ABR policy learning backbone is value-based that trains

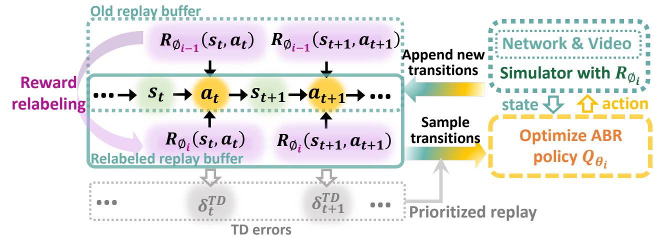


Figure 5: Illustration for the policy fine-tuning with personalized QoE model R_{ϕ_i} learned after the i-th interaction session.

a Q-network Q_θ by minimizing $\mathcal{L}_\theta$ based on TD errors δ^{TD} [31]:

$$\delta^{TD}(s_t, a_t, r_t, s_{t+1}) = r_t + \max_{a'} \gamma Q_{\bar{\theta}}(s_{t+1}, a') - Q_\theta(s_t, a_t), \quad (10)$$

$$\mathcal{L}_\theta = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}^\tau} \left[\left(\delta^{TD}(s_t, a_t, r_t, s_{t+1}) \right)^2 \right]. \quad (11)$$

Here, transitions (s_t, a_t, r_t, s_{t+1}) are sampled from a *replay buffer* $\mathcal{D}^\tau$ storing historical trajectories τ, where r_t is the reward for (s_t, a_t), and the target network $Q_{\bar{\theta}}$ is updated less frequently than the online network Q_θ. The policy π_θ is derived as $\pi_\theta(s_t) = \arg\max_a Q_\theta(s_t, \cdot)$.

5.2 Progressive Policy Fine-Tuning with Updated Reward

To avoid bad beginning performance, we employ a policy pretrained with general QoE as the initial policy π_{θ_0}. For the subsequent policy Q_{θ_i} optimizing the updated reward model R_{ϕ_i}, which is fine-tuned from $Q_{\theta_{i-1}}$ without reinitializing θ, we propose two key modifications[1] to adapt the backbone algorithm for progressive fine-tuning, ensuring effectiveness and efficiency (see Figure 5):

Reward relabeling. Unlike standard settings with fixed reward functions, our progressive training scheme dynamically updates the reward function. Directly reusing historical samples labeled with obsolete rewards $R_{\phi_{<i}}$ compromises policy performance due to misleading TD errors (Equation (10)) when optimizing Q_{θ_i}. On the other hand, cleaning $\mathcal{D}^\tau$ as an empty buffer wastes historical samples and reduces the sample diversity for future training, necessitating prolonged training to achieve target performance. To address these issues, we relabel historical samples in $\mathcal{D}^\tau$ by updating r_t to $R_{\phi_i}(s_t, a_t)$, which can be beneficial to both the policy performance and sample efficiency when fine-tuning the policy.

Calibrated prioritized replay. Building upon our reward relabeling approach, we implement priority relabeling [38], where transition priorities are updated alongside reward relabeling. This allows transitions (s_t, a_t, r_t) with larger TD errors (Equation (10)) under the new reward scheme to be sampled more frequently. Combined with our fine-tuning strategy that retains the old policy parameters θ_{i-1} at the beginning of the i-th training iteration, this calibrated prioritized replay significantly accelerates training.

6 Experiments

We first validate our personalized QoE modeling using `SQoE-IV` in Section 6.1, showing its superior alignment with human perception.

[1] Additional implementation details can be checked in Appendix F [4].

Table 1: Evaluation results for QoE modeling methods, averaged over all evaluators in the SQoE-IV dataset.

General QoE	Methods	Evaluation Metrics (mean±std) IR_o	IR_c	$SRCC$	$PLCC$
Intuitive	MPC [44]	$0.65_{\pm 0.04}$	$0.55_{\pm 0.04}$	$0.42_{\pm 0.11}$	$0.27_{\pm 0.10}$
	BOLA [41]	$0.70_{\pm 0.05}$	$0.64_{\pm 0.05}$	$0.54_{\pm 0.13}$	$0.53_{\pm 0.10}$
	Pensieve [30]	$0.70_{\pm 0.05}$	$0.63_{\pm 0.05}$	$0.54_{\pm 0.13}$	$0.53_{\pm 0.12}$
	Puffer [43]	$0.61_{\pm 0.06}$	$0.54_{\pm 0.03}$	$0.29_{\pm 0.15}$	$0.08_{\pm 0.15}$
Regression	BSQI [11]	$0.51_{\pm 0.01}$	$0.50_{\pm 0.02}$	$0.02_{\pm 0.04}$	$0.06_{\pm 0.04}$
	Comyco-Lin [18]	$0.63_{\pm 0.06}$	$0.56_{\pm 0.04}$	$0.35_{\pm 0.16}$	$0.09_{\pm 0.15}$
$\mathcal{L}^o_\phi$(Equation (6))	Jade-Lin [17]	$0.71_{\pm 0.05}$	$0.68_{\pm 0.06}$	$0.56_{\pm 0.13}$	$0.58_{\pm 0.11}$
Personalized QoE (With Ablations)	Models	Evaluation Metrics (mean±std) IR_o	IR_c	$SRCC$	$PLCC$
Regression	Linear	$0.50_{\pm 0.11}$	$0.58_{\pm 0.06}$	$0.01_{\pm 0.31}$	$0.00_{\pm 0.32}$
	MLP	$0.50_{\pm 0.02}$	$0.50_{\pm 0.01}$	$0.00_{\pm 0.05}$	$0.01_{\pm 0.05}$
	MonMLP	$0.50_{\pm 0.02}$	$0.50_{\pm 0.01}$	$0.01_{\pm 0.05}$	$0.01_{\pm 0.06}$
$\mathcal{L}^o_\phi$(Equation (6))	Linear	$0.72_{\pm 0.05}$	$0.69_{\pm 0.06}$	$0.59_{\pm 0.13}$	$0.60_{\pm 0.11}$
	MLP	$0.84_{\pm 0.03}$	$0.81_{\pm 0.05}$	$0.83_{\pm 0.05}$	$0.79_{\pm 0.05}$
	MonMLP	$0.81_{\pm 0.05}$	$0.78_{\pm 0.04}$	$0.77_{\pm 0.09}$	$0.76_{\pm 0.05}$
$\mathcal{L}_\phi$(Equation (9))	Linear	$0.72_{\pm 0.05}$	$0.70_{\pm 0.06}$	$0.59_{\pm 0.13}$	$0.61_{\pm 0.11}$
	MLP	$\mathbf{0.89_{\pm 0.04}}$	$\mathbf{0.93_{\pm 0.03}}$	$0.88_{\pm 0.07}$	$\mathbf{0.88_{\pm 0.03}}$
	MonMLP (Ours)	$\mathbf{0.89_{\pm 0.03}}$	$0.88_{\pm 0.05}$	$\mathbf{0.90_{\pm 0.06}}$	$0.85_{\pm 0.05}$

Next, in Section 6.2.1, we evaluate our progressive policy learning framework using both the personalized QoE metric from Section 6.1 and traditional metrics (e.g., rebuffering time, video quality). All experiments use three random seeds per trial, with results averaged for robust statistics. See Appendix B [4] for hyperparameters.

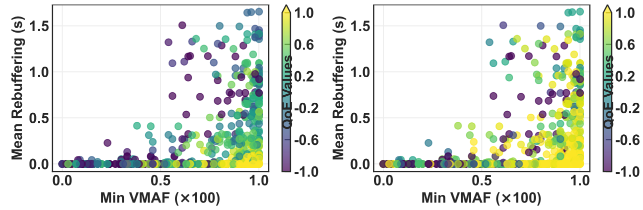
6.1 Evaluations of QoE Modeling in Q+

For each evaluator's personalized QoE, their subjective scores are randomly split with 80% for training and 20% for evaluation.

Evaluation metrics. For each testing session of an evaluator, we evaluate QoE predictions against subjective scores using the ordinal identity rate IR_o and cardinal identity rate IR_c, measuring the ratio of identical ordinal preferences (Equation (1)) and cardinal preference (Equation (2)), respectively, among all constructed pairwise comparisons. We also calculate the *SRCC* for rank correlation and the *PLCC* for numerical correlation. The reported results in the following are averaged over all sessions and evaluators.

Baselines and ablations. As outlined in Table 1, we establish two sets of experiments[2] to validate our QoE modeling method comprehensively: (i) To compare with other general QoE, we select representative baselines, including those intuitively configured [30, 41, 43, 44], and trained on population-level subjective scores [11, 18] or ordinal comparisons [17]. (ii) Given the lack of prior work specifically focused on learning personalized QoE, we validate our personalized QoE modeling through ablation studies. Specifically, we ablate our training objective $\mathcal{L}_\phi$ (Equation (9)) by replacing it with two alternatives: traditional regression with MSE minimization [18] and learning from ordinal comparisons [17] with Equation (6). Additionally, we ablate our MonMLP architecture by substituting it with a linear model [17, 18] and a standard MLP [17].

[2]See implementation details about baselines and ablations in Appendix C [4].



(a) MonMLP: $\min(\boldsymbol{v})$ **vs.** $\text{mean}(\boldsymbol{u})$ **(b) MLP:** $\min(\boldsymbol{v})$ **vs.** $\text{mean}(\boldsymbol{u})$

Figure 6: QoE predictions from a MLP or MonMLP trained on an evaluator's subjective scores with $\mathcal{L}_\phi$ (Equation (9)).

Performance analysis. As shown in Table 1, our personalized QoE modeling significantly outperforms others, with findings summarized below: (i) Personalized QoE modeling with cross-session feedback requires learning from preferences with $\mathcal{L}^o_\phi$ or our $\mathcal{L}_\phi$, leading to performance exceeding SOTA general QoE. Regression fails due to numerical distortion. (ii) According to ablations on training objective, introducing cardinal loss $\mathcal{L}^c_\phi$ for QoE training (i.e., $\mathcal{L}^o_\phi$ vs. $\mathcal{L}_\phi$) enhances IR_c while also improving IR_o, *SRCC*, and *PLCC*. (iii) According to ablations about model architecture, MonMLP and MLP outperform the Linear model, with MonMLP generating more reasonable predictions that adhere to monotonicity constraints, as demonstrated in Figure 6[3]. For example, in Figure 6b, MLP predicts unreasonably high values (yellow points) for video experiences with large rebuffering (upper right zone) or poor video quality (lower left zone), while MonMLP produces smooth, monotonic predictions, demonstrating enhanced generalization in Figure 6a, further benefits RL policy training as shown in Section 6.2.4.

6.2 Evaluations of Q+ Policy Learning

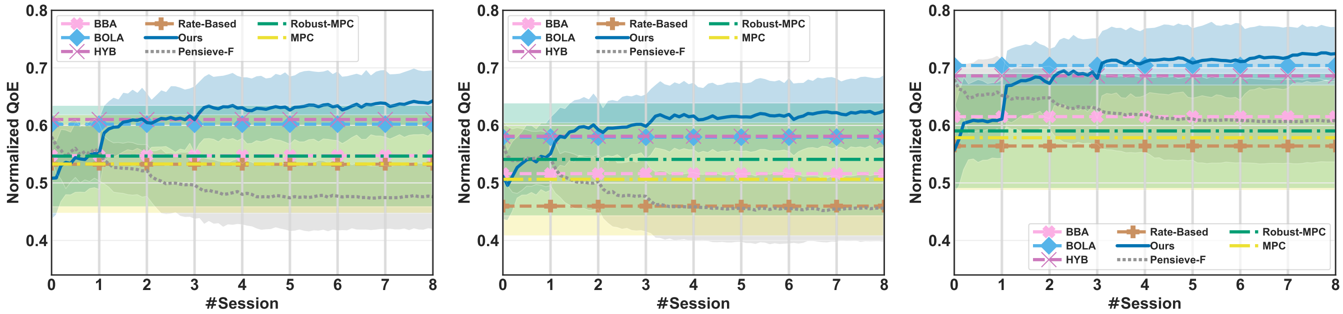
6.2.1 Experimental Setups.

Simulator. We conduct policy training experiments in Park [29], a trace-driven virtual player environment for realistic adaptive streaming simulations. Our video dataset includes 83 complete videos from YouTube [18], covering diverse genres such as gaming, cinema, music, sports, and television. Each video is encoded at nine target bitrates ({235, 375, 560, 750, 1050, 1750, 2350, 3000, 4300} kbps) and segmented into 4-second chunks, totaling 4,702 segments. Video lengths vary significantly, ranging from 8 to 231 chunks, ensuring diverse streaming scenarios. Our network traces include three trace collections covering varied network conditions: fixed broadband traces (FCC) and dynamic mobile traces (3G/HSDPA) from Pensieve [30], and heavy-tailed traces from Puffer [43]. Videos and network traces are randomly split with 80% for training and 20% for evaluation.

Information in a policy state s_t contains the VMAF of the last selected chunk, current buffer size, estimated bandwidth, delay time for receiving the last chunk, remaining chunks to complete the video, sizes of the next chunk across target bitrates, and the corresponding VMAFs.

To enable large-scale online RL training, the personalized QoEs trained with the whole SQoE-IV dataset using our method, which

[3]Supplementary figures for additional evaluators can be checked in Appendix H.1 [4].



(a) FCC traces. (b) 3G/HSDPA traces. (c) Puffer traces.

Figure 7: Comparisons Q+ with baselines. Lines and shadows represent average performance and standard deviation, respectively. Vertical lines mark interaction sessions where R_ϕ is updated with $\#F = 5$ new scores, triggering subsequent policy fine-tuning.

outperformed all other approaches as shown in Section 6.1, serve as the ground-truth QoEs for each evaluator e, denoted as R_e^*. Following previous works [23, 26, 28], we synthesize quality assessments by treating $R_e^*(\tau)$ as evaluator e's numerical rating for segment τ. Such simulation maintains human perceptual alignment while avoiding the prohibitive costs of continuous human annotation.

Performance evaluation. Following previous works [17], the performance of an ABR policy is measured by normalized R_e^* by linear mapped to [0,1], and we show the averaged performance over evaluators as the overall QoE performance. In addition, we also examine Q+ in conventional metrics, e.g., rebuffering time, video quality, and smoothness. Due to the computational cost for multi-user multi-session simulation, 8 evaluators (25% of the SQoE-IV corpus) are randomly selected, and we run 3 seeds for each user.

6.2.2 Compare with Baselines.

Baseline methods. Given the lack of prior ABR methods for the progressive learning scheme, we adapt representative baselines to fit our framework. These adaptations are summarized as follows[4] : (i) For learning-based methods, we extend Pensieve [30], a policy-gradient framework, to our progressive personalized QoE-driven setting, termed Pensieve-F. This extension uses the same experimental configurations as Q+ (see Section 6.2.1) to ensure a fair comparison. (ii) For Model Predictive Control (MPC) methods, we evaluate both the standard MPC and its conservative variant Robust-MPC [44]. Both methods use harmonic-mean bandwidth prediction and finite-horizon QoE optimization. They use the same last QoE model as Q+'s, which is trained with all personal feedback. (iii) For rule-based methods, we consider both buffer-based and rate-based methods. For buffer-based methods, we compare with BBA [20] that determines bitrate via linear interpolation based on buffer levels, and BOLA [41] that uses Lyapunov optimization to select bitrates constrained by buffer occupancy. For rate-based methods, we compare with Rate-Based [22] that selects the maximum bitrate below the predicted throughput estimated via harmonic mean. In addition, we consider HYB [6] that selects the maximum bitrate with chunk size below the estimated available file size based on buffer occupancy and predicted throughput.

[4]Further implementation details can be checked in Appendix D.1 [4].

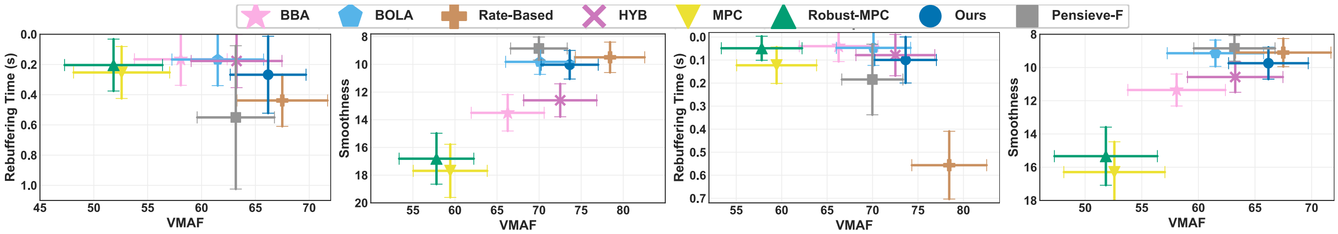
Performance analysis. For baselines involving personalized QoE-modeling (i.e., Q+, Pensieve-F, MPC, Robust-MPC), we evaluate them under varying *per-session feedback quantities* ($\#F$). Figure 7 compares their QoE performance under a moderate feedback setting (#F=5), where Q+ consistently surpasses all baselines after only 3 sessions with 15 feedbacks. Despite initial lag due to cold-start personalized QoE modeling, Q+ achieves sustained improvements through iterative feedback refinement. Although the initial general QoE-driven policy of Pensieve-F's outperforms Q+'s, Pensieve-F degrades progressively, while Q+'s value-based policy design enhances both the robustness against imperfect reward signals [15, 32, 40, 47] and training efficiency. As for MPC variants, they exhibit higher variance and lower average performance compared to Q+. Their vulnerability to imperfect QoE models highlights the limitations of model-based approaches within the progressive learning scheme. Additionally, Figure 8 compares the breakdown performance of the final policies shown in Figure 7[5], where Q+ consistently performs within the acceptable upper right regions.

6.2.3 Sensitivity Analysis for Feedback Quantity.

Settings. To study the impact of feedback quantity ($\#F$) on Q+'s adaptation, we conduct experiments with fixed $\#F$ values (3, 5, 7, 10) and a variable scenario where $\#F \sim Normal(\mu = 6, \sigma = 3)$ (samples below 3 are set to 3). The minimum $\#F = 3$ is chosen because $\mathcal{L}_\phi$ (Equation (9)) requires at least three scores to construct a cardinal comparison. The value $\#F = 10$ represents an ideal, highly engaged user scenario, while 5 and 7 reflect practical feedback levels. $Normal(\mu = 6, \sigma = 3)$ simulates spontaneous user feedback.

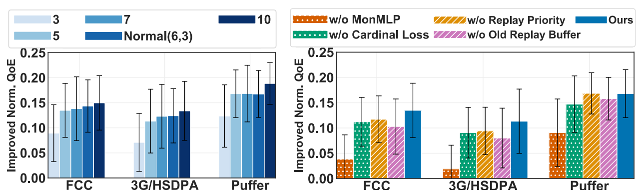
Performance analysis. As shown in Figure 9a, Q+ achieves better performance with increased user feedback, surpassing baseline methods. Even with the minimum $\#F = 3$, required to construct one cardinal pair per session, Q+ performs on par with the best-performing baseline, demonstrating its robustness under limited feedback. Furthermore, larger feedback quantities unlock more significant performance improvements, and the distributional $\#F \sim Normal(6, 3)$ setting also shows stable performance, indicating that occasional limited feedback does not hinder overall performance.

[5]Additional figures for 3G/HSDPA and other $\#F$ can be checked in Appendix G.1.2 [4]



(a) FCC: VMAF v.s. Rebuffering (b) FCC: VMAF v.s. Smoothness (c) Puffer: VMAF v.s. Rebuffering (d) Puffer: VMAF v.s. Smoothness

Figure 8: Breakdown performance for the last policies trained with $\#F = 5$. Smaller rebuffering and smoothness values, and larger VMAF values are desired.



(a) Sensitivity analysis for Q+ with different $\#F$. **(b) Ablation study for Q+.**

Figure 9: Relative normalized QoE improvement after the final training iteration compared to the initial policy. See complete training curves in Appendix G.2 and G.3 [4].

6.2.4 *Ablation Studies for* Q+.

Settings. We conduct four ablation experiments to evaluate Q+: two focus on personalized QoE modeling's impact on progressive policy learning, and two examine our policy training design. For personalized QoE modeling, we compare our method with a variant replacing MonMLP with a standard MLP (`w/o MonMLP`) and another removing the cardinal loss from $\mathcal{L}_\phi$ (Equation (9)) and use $\mathcal{L}_\phi^o$ (Equation (6)) instead (`w/o Cardinal Loss`). For policy training, we test cleaning the old replay buffer and recollecting new trajectories per training iteration (`w/o Old Replay Buffer`) and replacing priority replay with uniform replay (`w/o Prioritized Replay`). See more details for ablations' configurations in Appendix D.2 [4].

Performance analysis. According to Figure 9b[6] with $\#F = 5$, omitting MonMLP (`w/o MonMLP`) causes the most significant decline, emphasizing the necessity of monotonicity constraints for effective personalized QoE modeling in the reward function. Among policy training ablations (`w/o Prioritized Replay` and `w/o Old Replay Buffer`), removing the old replay buffer (`w/o Old Replay Buffer`) has a more detrimental effect, highlighting the importance of retaining historical trajectories.

7 Related Work

Subjective QoE modeling. We review QoE modeling from the aspects of model training and model space. For training, methods learning from numerical scores via regression [11, 18, 33] or adversarial training [46] fail in cross-session settings due to numerical distortion. Recent work [17] uses pairwise comparisons for ordinal relationships but overlooks cardinal relations, limiting feedback utilization and risking misaligned predictions. In contrast, we learn reliable QoE from both ordinal and cardinal preferences with cross-session datasets. For model space, formula-based models [30, 44] provide interpretability yet misalign with real QoE, non-parametric approaches [11] are limited to numerical regression, and DNNs [17, 46] enhance complexity but struggle to balance expressivity and generalization. Our QoE modeling ensures both expressivity and generalization through monotonic neural networks.

Personalized QoE-driven ABR methods. Quality of Service (QoS)-driven ABR methods are often rule-based [20, 22, 41], focusing on system targets but underperform. Conventional QoE-driven methods optimize subjective perception but require costly retraining for QoE changes [17, 30, 43], while our method avoids this through progressive policy fine-tuning. Unlike policy-gradient DRL methods [17, 30], we use a value-based approach for sample efficiency. While some methods adapt to various QoE models without retraining, they are limited to 2-3 QoE parameters [18, 19, 34, 48] or are noise-sensitive [44]. In contrast, our method supports expressive QoE models that output scalar rewards and does not rely on precise environmental modeling.

8 Conclusion

In this work, we propose Q+, a novel personalized QoE-driven ABR framework that bridges the gap between user preferences and adaptive streaming through iterative human feedback. Our framework introduces a scalable, user-friendly progressive feedback collection scheme, a robust personalized QoE model that unifies ordinal and cardinal preferences via monotonic neural networks, and a calibrated RL algorithm that dynamically aligns bitrate polices with personalized QoE updates. Extensive experiments demonstrate that our framework outperforms SOTA baselines in both QoE accuracy and policy effectiveness, achieving significant improvements in user experience across diverse network scenarios.

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No. 2024YFC3017100) and the National Natural Science Foundation of China (Grant No. 62302292).

[6]Additional ablation results for $\#F \sim Normal(6,3)$ are provided in Appendix G.3 [4], where the findings for $\#F \sim Normal(6,3)$ align with those for $\#F = 5$.

References

[1] [n. d.]. Toward A Practical Perceptual Video Quality Metric | by Netflix Technology Blog | Netflix TechBlog. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652 VMAF-blog1.

[2] 2025. Amazon Prime Video. https://www.primevideo.com

[3] 2025. Netflix. https://www.netflix.com/

[4] 2025. Q+ Appendix. https://sites.ji.sjtu.edu.cn/yifei-zhu/wp-content/uploads/sites/19/2025/07/QPLUS_Appendix.pdf

[5] 2025. YouTube. https://www.youtube.com/

[6] Zahaib Akhtar, Yun Seong Nam, Ramesh Govindan, Sanjay Rao, Jessica Chen, Ethan Katz-Bassett, Bruno Ribeiro, Jibin Zhan, and Hui Zhang. 2018. Oboe: auto-tuning video ABR algorithms to network conditions. In *Proc. ACM SIGOCOMM* (Budapest, Hu..) *(SIGCOMM '18)*. 44–58.

[7] Avşar Asan, Werner Robitza, Is-haka Mkwawa, Lingfen Sun, Emmanuel Ifeachor, and Alexander Raake. 2017. Impact of video resolution changes on QoE for adaptive video streaming. In *ICME* (Hong Kong, China). IEEE, 499–504.

[8] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[9] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *NeurIPS* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[10] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. 2011. Understanding the impact of video quality on user engagement. In *SIGCOMM* (Toronto, Ontario, Canada) *(SIGCOMM '11)*. 362–373.

[11] Zhengfang Duanmu, Wentao Liu, Diqi Chen, Zhuoran Li, Zhou Wang, Yizhou Wang, and Wen Gao. 2023. A Bayesian Quality-of-Experience Model for Adaptive Streaming Videos. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 3s, Article 141 (feb 2023), 24 pages.

[12] Zhengfang Duanmu, Wentao Liu, Zhuoran Li, Diqi Chen, Zhou Wang, Yizhou Wang, and Wen Gao. 2020. Assessing the Quality-of-Experience of Adaptive Bitrate Video Streaming. arXiv:2008.08804 [eess.IV]

[13] Xuening Feng, Zhaohui JIANG, Timo Kaufmann, Eyke Hüllermeier, Paul Weng, and Yifei Zhu. 2025. Comparing Comparisons: Informative and Easy Human Feedback with Distinguishability Queries. In *ICML*.

[14] Xuening Feng, Zhaohui Jiang, Timo Kaufmann, Puchen Xu, Eyke Hüllermeier, Paul Weng, and Yifei Zhu. 2025. DUO: Diverse, Uncertain, On-Policy Query Generation and Selection for Reinforcement Learning from Human Feedback. In *AAAI*.

[15] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling Laws for Reward Model Overoptimization. In *ICML* (Honolulu, Hawaii, USA). 10835–10866.

[16] Xiao Hu, Jianxiong Li, Xianyuan Zhan, Qing-Shan Jia, and Ya-Qin Zhang. 2023. Query-Policy Misalignment in Preference-Based Reinforcement Learning. *ArXiv* abs/2305.17400 (2023).

[17] Tianchi Huang, Rui-Xiao Zhang, Chenglei Wu, and Lifeng Sun. 2023. Optimizing Adaptive Video Streaming with Human Feedback. In *MM* (Ottawa ON, Canada) *(MM '23)*. 1707–1718.

[18] Tianchi Huang, Chao Zhou, Xin Yao, Rui-Xiao Zhang, Chenglei Wu, Bing Yu, and Lifeng Sun. 2020. Quality-Aware Neural Adaptive Video Streaming With Lifelong Imitation Learning. *IEEE Journal on Selected Areas in Communications* 38, 10 (2020), 2324–2342.

[19] Tianchi Huang, Chao Zhou, Rui-Xiao Zhang, Chenglei Wu, and Lifeng Sun. 2022. Learning Tailored Adaptive Bitrate Algorithms to Heterogeneous Network Conditions: A Domain-Specific Priors and Meta-Reinforcement Learning Approach. *IEEE Journal on Selected Areas in Communications* 40, 8 (2022), 2485–2503.

[20] Te-Yuan Huang, Ramesh Johari, Nick McKeown, Matthew Trunnell, and Mark Watson. 2014. A buffer-based approach to rate adaptation: evidence from a large video streaming service. In *SIGCOMM* (Chicago, Illinois, USA) *(SIGCOMM '14)*. 187–198.

[21] Liangyu Huo, Zulin Wang, Mai Xu, Yong Li, Zhiguo Ding, and Hao Wang. 2020. A Meta-Learning Framework for Learning Multi-User Preferences in QoE Optimization of DASH. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 9 (2020), 3210–3225.

[22] Junchen Jiang, Vyas Sekar, and Hui Zhang. 2012. Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE. In *CoNEXT* (Nice, France) *(CoNEXT '12)*. 97–108.

[23] Zhaohui JIANG, Xuening Feng, Paul Weng, Yifei Zhu, Yan Song, Tianze Zhou, Yujing Hu, Tangjie Lv, and Changjie Fan. 2025. Reinforcement Learning from Imperfect Corrective Actions and Proxy Rewards. In *ICLR* (Singapore).

[24] S. Shunmuga Krishnan and Ramesh K. Sitaraman. 2013. Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-Experimental Designs. *IEEE/ACM Transactions on Networking* 21, 6 (2013), 2001–2014.

[25] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. 2021. B-Pref: Benchmarking Preference-Based Reinforcement Learning. In *NeurIPS*. Neural Information Processing Systems Foundation, Curran Associates Inc.

[26] Kimin Lee, Laura M Smith, and Pieter Abbeel. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. In *ICML (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 6152–6163.

[27] Yiyun Lu, Yifei Zhu, and Zhi Wang. 2022. Personalized 360-degree video streaming: A meta-learning approach. In *ACM Multimedia*. 3143–3151.

[28] Jianlan Luo, Perry Dong, Yuexiang Zhai, Yi Ma, and Sergey Levine. 2024. RLIF: Interactive Imitation Learning as Reinforcement Learning. In *ICLR* (Vienna, Austria).

[29] Hongzi Mao, Parimarjan Negi, Akshay Narayan, Hanrui Wang, Jiacheng Yang, Haonan Wang, Ryan Marcus, ravichandra addanki, Mehrdad Khani Shirkoohi, Songtao He, Vikram Nathan, Frank Cangialosi, Shaileshh Venkatakrishnan, Wei-Hung Weng, Song Han, Tim Kraska, and Dr.Mohammad Alizadeh. 2019. Park: An Open Platform for Learning-Augmented Computer Systems. In *NeurIPS*, Vol. 32. Curran Associates, Inc.

[30] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *SIGCOMM* (Los Angeles, CA, USA) *(SIGCOMM '17)*. 197–210. doi:10.1145/3098822.3098843

[31] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[32] Jane Pan, He He, Samuel R. Bowman, and Shi Feng. 2024. Spontaneous Reward Hacking in Iterative Self-Refinement. *CoRR* abs/2407.04549 (2024).

[33] Leonardo Peroni, Sergey Gorinsky, Farzad Tashtarian, and Christian Timmerer. 2023. Empowerment of Atypical Viewers via Low-Effort Personalized Modeling of Video Streaming Quality. *CoNEXT* 1, CoNEXT3, Article 17 (nov 2023), 27 pages.

[34] Chunyu Qiao, Jiliang Wang, and Yunhao Liu. 2021. Beyond QoE: Diversity Adaptation in Video Streaming at the Edge. *IEEE/ACM Transactions on Networking* 29, 1 (2021), 289–302.

[35] Werner Robitza, Marie-Neige Garcia, and Alexander Raake. 2017. A modular HTTP adaptive streaming QoE model — Candidate for ITU-T P.1203 ("P.NATS"). In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. Erfurt, Germany, 1–6.

[36] Davor Runje and Sharath M Shankaranarayana. 2023. Constrained Monotonic Neural Networks. In *ICML* (Honolulu, Hawaii, USA) *(Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 29338–29353.

[37] Sandvine. 2024. The Global Internet Phenomena Report January 2024. https://www.sandvine.com/blog/sandvines-2024-global-internet-phenomena-report-global-internet-usage-continues-to-grow

[38] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized Experience Replay. In *ICLR* (San Juan, Puerto Rico).

[39] Max Schwarzer, Johan Samir Obando Ceron, Aaron Courville, Marc G Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. 2023. Bigger, better, faster: Human-level atari with human-level efficiency. In *ICML*. PMLR, 30365–30380.

[40] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and Characterizing Reward Gaming. In *NeurIPS*, Vol. 35. Curran Associates, Inc., 9460–9471.

[41] Kevin Spiteri, Rahul Urgaonkar, and Ramesh K. Sitaraman. 2020. BOLA: Near-Optimal Bitrate Adaptation for Online Videos. *IEEE/ACM Transactions on Networking* 28, 4 (2020), 1698–1711.

[42] Thomas Stockhammer. 2011. Dynamic adaptive streaming over HTTP –: standards and design principles. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems* (San Jose, CA, USA) *(MMSys '11)*. 133–144.

[43] Francis Y. Yan, Hudson Ayers, Chenzhi Zhu, Sadjad Fouladi, James Hong, Keyi Zhang, Philip Levis, and Keith Winstein. 2020. Learning in situ: a randomized experiment in video streaming. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. USENIX Association, Santa Clara, CA, 495–511.

[44] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. 2015. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. (2015), 325–338. doi:10.1145/2785956.2787486

[45] Huaizheng Zhang, Linsen Dong, Guanyu Gao, Han Hu, Yonggang Wen, and Kyle Guan. 2020. DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction. *IEEE Transactions on Multimedia* 22, 12 (2020), 3210–3223.

[46] Huanhuan Zhang, Liu zhuo, Haotian Li, Anfu Zhou, Chuanming Wang, and Huadong Ma. 2024. AraLive: Automatic Reward Adaption for Learning-based Live Video Streaming. In *MM*. 11099–11108.

[47] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Haoran Huang, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Delve into PPO: Implementation Matters for Stable RLHF. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*. Curran Associates Inc.

[48] Xutong Zuo, Jiayu Yang, Mowei Wang, and Yong Cui. 2022. Adaptive Bitrate with User-level QoE Preference for Video Streaming. In *INFOCOM*. IEEE, London, United Kingdom, 1279–1288.