Swiping, Fast and Slow: Assessing the QoE of Short-Form Videos via Crowdsourcing

Filip Simonovski*, Samuel Hufen*, Lisa Karl†, Alperen Sayin*,
Nikolas Wehner†, Tobias Hoßfeld†, Michael Seufert*
*University of Augsburg, Chair of Networked Systems and Communication Networks, Augsburg, Germany
{filip.simonovski | samuel.hufen | michael.seufert}@uni-a.de, alperen.sayin@ajs.bz
†University of Würzburg, Chair of Communication Networks, Würzburg, Germany
lisakarl1@web.de, {nikolas.wehner | tobias.hossfeld}@uni-wuerzburg.de

***Abstract*—Short-form video (SFV) services, such as TikTok and Instagram Reels, have rapidly gained widespread popularity, accumulating billions of users. However, evaluating Quality of Experience (QoE) for these services poses challenges as they are typically consumed in mobile and interactive settings. In this paper, we introduce a novel QoE evaluation framework for SFV, which allows for a controlled presentation of stimuli and a reliable collection of valid QoE ratings in an unsupervised setting, while maintaining the authenticity of the mobile, interactive SFV experience. We use our framework to conduct two QoE studies on the impact of waiting times, i.e., initial delay and stalling, on the QoE of SFV via crowdsourcing. Our findings reveal that initial delay results in a three times higher probability that users swipe to the next video within the first ten seconds compared to stalling. In contrast, the Mean Opinion Score (MOS) of stalling is up to 0.4 lower than initial delay for the same waiting time conditions. These insights provide valuable guidelines for optimizing SFV content delivery to enhance user satisfaction, but also highlight the need for novel QoE models, which can describe not only perceived QoE but also resulting user engagement and behavior.**

***Index Terms*—Quality of Experience, Short-form Videos, Short Video Services, Waiting Times, Initial Delay, Stalling.**

I. Introduction

Short-form video (SFV) services, such as TikTok, Douyin, and Instagram Reels, have rapidly gained widespread popularity and accumulate billions of users [1]. Their rapid success is based on a novel platform design, which emphasizes easy content generation and instantaneous access. By facilitating the production and consumption of short video clips, which deliver quick bursts of information or amusement, these platforms cater to modern users' demand for immediate gratification and an easily digestible media experience in small snacks. This clearly distinguishes them from classical long-form video (LFV) platforms like Netflix, Amazon Prime, or YouTube.

SFVs typically last only a few seconds to approximately one minute and are watched almost exclusively on smartphones using the provided applications of SFV service platforms. Unlike with LFV platforms, users typically consume videos that are algorithmically recommended by the SFV platform based on individual user preferences, social interactions, and global content trends. The main user interaction is the swipe gesture to advance to the next recommended video.

As SFVs are often consumed in mobile networks with fluctuating conditions, SFV services leverage HTTP Adaptive Streaming (HAS) and adaptive bit rate (ABR) for dynamic quality. Compared to LFVs, SFVs typically only use a reduced number of encoding variants [2] and might possess distinct Quality of Experience (QoE) objectives. While avoiding stalling, i.e., playback interruptions, and delivering high visual quality are important, initial delay could be a much more relevant QoE factor compared to traditional HAS [3]–[5]. This is because users likely expect immediate content access due to short video duration and swiping interactivity.

Supporting this assumption, we can indeed observe that SFV services implement sophisticated pre-loading strategies in their applications to minimize waiting times, such as initial delay or stalling. Strategies vary by service, e.g., services might choose to download a few videos entirely, or might choose to download only the first few seconds of many videos [2]. This is a strong indicator that SFV services consider waiting times as highly important QoE factors. However, no explicit QoE model for SFV services has been derived so far.

Conducting QoE studies for SFV introduces several challenges. First, it is mandatory to replicate a realistic look and feel, capturing the high interactivity and dynamic user engagement in SFV platforms. Additionally, the study has to be designed for smartphones. Variations in screen size, resolution, brightness settings, and other customizable features can substantially affect the QoE, complicating the standardized presentation of test conditions. Last, crowdsourcing QoE studies are preferred to recruit a larger and more heterogeneous group of participants compared to lab studies. However, crowdsourced QoE studies typically involve passive testing in a browser without any interactive engagement. Thus, realistic SFV QoE studies on smartphones must ensure natural interactivity, like swiping, to accurately mimic actual user behavior.

In this work, we aim to close the gap and tackle the challenge of assessing the QoE of SFV via crowdsourcing:

- RQ1: How to achieve both a controlled presentation of stimuli and a reliable collection of valid QoE ratings in an unsupervised setting, while maintaining the authenticity of the mobile, interactive SFV service experience?
- RQ2: What is the impact of waiting times, i.e., initial delay and stalling, on the QoE of SFV?

To answer our research questions, we design a framework for SFV QoE studies and conduct two crowdsourcing QoE studies on the impact of waiting times. We analyze the interactions of QoE study participants with our framework, and to the best of our knowledge, we are the first to quantify the impact of initial delay and stalling on the QoE of SFV.

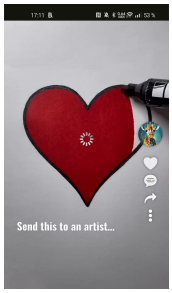
This work is structured as follows. Section II outlines related works on SFV and QoE. Section III describes the implemented crowdsourcing framework, the conducted QoE study and the implemented data filtering. Section IV presents the results on the QoE impact of waiting times and discusses open challenges with crowdsourced QoE studies for SFVs. Finally, Section V summarizes the findings and concludes.

II. Related Work

Quality of Experience: QoE is a widely accepted concept to evaluate the subjectively perceived quality of end users of Internet applications [6]. QoE research has focused on finding key relationships between Quality of Service (QoS), i.e., objectively measurable technical parameters of a system, and QoE as well as deriving QoE models for popular application classes. For example, in adaptive video streaming, high QoE can generally be reached by minimizing waiting times, i.e., initial delay and stalling (playback interruptions, rebuffering), while maximizing the visual quality of the streamed content in terms of high resolution, high frame rate, and low visible compression artifacts [3]. A large number of QoE models were proposed in literature, e.g., [3], [7], and recently, also machine learning (ML) [8] and explainable artificial intelligence [9], [10] were employed for QoE modeling. Standardized QoE models are P.1203 [4], [11], [12] and P.1204 [13].

Technical Aspects of Short-form Video Services: SFV services play videos in quick succession following the swiping interaction of the user. To ensure instant playback after a swipe, SFVs heavily rely on pre-loading algorithms to always have new content available for playback when the user swipes. Pre-loading strategies are either following a conventional [14]–[16] or a ML-based approach [17]–[20]. Due to the unique user interaction with SFV services, primarily the low-barrier swipe navigation, pre-loading algorithms often load more data than needed. Several works tackle the challenge of minimizing the used bandwidth [16] or data wastage [19], [20] while maintaining the QoE. Furthermore, in-network bandwidth capping of SFV services could substantially reduce download volume by 15%–45% and bandwidth wastage by 18%–52%, while at the same time bandwidth utilization fairness improved [21].Finally, a method for reordering SFV playlists to reduce server-side startup delays has been explored [22], but such an approach can interfere with SFV services recommendation algorithms [23].

QoE-related Studies for Short-form Video Services: Several measurements were conducted to understand the characteristics of SFV services [2], [24], [25]. A popular SFV service showed a median content length of 22s, with only 30.92% of videos watched completely [25]. Incompletely watched videos were skipped early, with a median playtime of only 12s.



(a) Loading Indicator

(b) Swipe Instruction

Fig. 1: Screenshots of the SFV QoE Study

Additionally, users appeared highly sensitive to stalling, as a single stalling event reduced the median playback percentage by more than 45% [25]. Although we can assume that this negative impact of stalling on user engagement carries over to QoE, as user engagement typically has a high correlation to QoE, cf. [26]–[28], there are no dedicated QoE studies or models for SFVs yet which provide definitive answers. Nevertheless, findings on stalling, alongside numerous pre-loading algorithms aimed at reducing initial delay, suggest that these two QoE factors, which also have a high impact on the QoE of traditional LFVs content, might still be relevant for SFVs [25]. Thus, in this work, to the best of our knowledge, we will conduct the first dedicated QoE study for SFVs. We will investigate if the impact of initial delay and stalling is still the same as for LFV or if the different nature of SFV necessitates novel SFV-tailored QoE models.

III. Methodology

A. Generation of Short-form Video Test Stimuli

SFV dataset: To create a realistic SFV dataset, we collected videos from YouTube using the library `pytubefix` [29]. As YouTube queries do not have the ability to distinguish between SFVs and LFVs, we define a set of criteria to select suitable videos. The selected videos have the following characteristics: (i) portrait format, (ii) video resolution of at least 1080×1920, representing 1080p (HD) [30], (iii) video duration of 25s $\pm$5s, and (iv) medium luminance. We classify the luminance as medium if the weighted average luminance (cf. [31]) is between 0.3 and 0.7. The limitation to medium luminance improves robustness with respect to automatic brightness adjustment of modern smartphones, which we cannot control during a crowdsourcing study. To appeal to different preferences, we select 33 different videos for our dataset, such that users can likely find videos they like when swiping through the videos.

Playlists: The selected videos are divided into one training playlist containing 3 videos and three stimuli playlists (P1, P2, P3), each consisting of 10 videos presented in a fixed order to ensure consistency across participants. The videos in each playlist are chosen so that they have different content, which ensures not only that participants can experience different contents and swipe to content they like, but also that content influence can be averaged out.

Conditions: We conduct two studies, one for initial delay and another for stalling. In each of the studies, the videos can contain four different waiting time conditions: 0s, 1s, 3s, or 5s. In the initial delay study, the waiting time is experienced on the first frame, before the playback starts. In the stalling study, the stalling event occurs at five seconds into the video. The stalling position at five seconds was chosen because it is far enough from the start of the video to clearly distinguish from the initial delay, while still being early enough so that participants often can see it before they swipe to the next video. The process of editing the videos to introduce the required waiting times is carried out using `ffmpeg` [32]. For this, we take the first frame of the video in the case of initial delay, or the frame at 5s for stalling, and replicate it multiple times depending on the condition value. Additionally, we also overlay a spinning loading icon, which is the typical for initial delay and stalling in video services. A screenshot of a video with a loading indicator can be seen in Figure 1a. The video editing process is performed for all selected SFVs.

Condition Block: A condition block consists of interacting with a stimuli playlist and answering a set of questions. The condition shown in the stimuli playlist is chosen randomly from the four options (0s, 1s, 3s, 5s) and this condition is included in all videos of the playlist to ensure that, regardless of the swiping behavior, users are exposed to it. Each playlist contains 10 videos and, to provide a realistic experience, users can freely swipe through the videos of a playlist. As we also must ensure a consistent stimulus presentation, we end a condition block after three videos have been attended to. We consider a video attended if it has been watched for more than 10 seconds, which includes perceiving the included condition in full. Thus, after the third video is watched for more than 10 seconds, the participant directly proceeds to the questionnaire. This way, all participants attended to the same number of videos before providing QoE ratings, avoiding that ratings are influenced by the number of attended videos. To further increase the realism, we also implemented overlay icons, which are typical for SFV platforms, i.e., profile picture, like icon, and share icon. Once a video finishes, a text notification is displayed at the center of the screen, prompting the user to swipe to continue, see Figure 1b. SFV platforms often loop (i.e., repeat) videos after playback ends. However, we chose not to include a looping mechanism in our QoE study, as repeated playback could introduce bias if some participants experience the initial delay or stalling condition more than once. During a condition block, participants can move to the next video anytime by swiping, even during a condition, i.e., while the loading icon is showing. However, to ensure that participants attend to three videos per playlist, we implement a mechanism that tracks the number of attended videos. If participants have not attended to the required number of videos, the mechanism restricts swiping for the first 10s of the final videos of the playlist, thus ensuring that users attend to the videos long enough to experience the condition. Specifically, if the participant has not attended to any video, swiping is blocked for the last three videos in the playlist to ensure that those last three videos are attended to. Similarly, if only two videos were attended to, swiping is restricted for the last video. This approach allows participants to interact with the swiping functionality, while also ensuring that they attend to the required number of videos in each condition block.

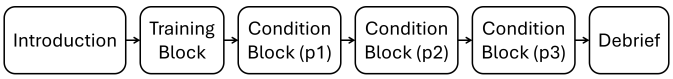


Fig. 2: Overview of the Study Design

B. Short-form Video Crowdsourcing QoE Study

The online framework used to conduct the study was created with jsPsych [33], a library for creating behavioral experiments in web browsers. Furthermore, the framework incorporates best practices for crowdsourcing studies, including pre-loading the video content to exclude network influences, monitoring the test execution, and conducting reliability checks [34]. Each study took about five minutes to complete, with participants recruited through advertisements for the QoE studies on the crowdsourcing platform Microworkers [35]. We designed our study framework such that the QoE study can only be accessed from smartphones. Thus, before loading the SFVs, a check is performed to determine whether the device is a smartphone based on the user agent string. In case of a non-mobile device, a QR code is generated, which, once scanned with a smartphone, redirects the participant to the study.

The study design is outlined in Figure 2. At the beginning, participants are introduced to the topic of the study, followed by a set of personal questions about their country and continent of residence, and their SFVs usage, i.e., they are asked to indicate the used SFV apps and their usage frequency. Afterwards, the rules of the study are explained, and participants familiarize themselves with the process by completing a short training block of three videos. After finishing the videos, participants are prompted to provide ratings for the video content, the video quality, and the overall quality of the video streaming on a 5-point Absolute Category Rating (ACR) [36] scale. To determine the impact of the condition, they also rate the level of annoyance caused by the waiting time, i.e., initial delay or stalling, on a Degradation Category Rating (DCR) [36] oriented 5-point scale (not at all/did not notice any, slightly annoying, moderately annoying, very annoying, extremely annoying). Finally, as a reliability check, participants are asked to select the content of the last watched video from a list. After training, participants repeat watching a stimulus playlist and answering these questions for the three condition blocks. After the condition blocks, participants are again asked for SFV application usage frequency as reliability check before the debrief block is presented, which includes a verification code to claim reimbursement for study participation.

C. Filtering of Unreliable Study Participants

To begin with, we provide an overview of the study participants from the unfiltered dataset, consisting of 428 male, 425 female and 2 diverse participants. 53% of them are between 25 and 35 years old, 18% are between 20 and 24 years, and

TABLE I: Dataset Overview.

Condition	Condition Value [s]	Collected Ratings			After Generic Filters			After SFV-specific Filters		
		P1	P2	P3	P1	P2	P3	P1	P2	P3
Initial Delay	0	89	91	86	20	22	19	15	20	9
Initial Delay	1	123	127	130	24	33	30	20	26	16
Initial Delay	3	134	127	129	28	20	27	20	12	18
Initial Delay	5	123	124	124	34	31	30	25	24	16
Stalling	0	91	87	89	24	27	26	19	26	22
Stalling	1	90	103	104	24	32	22	17	27	16
Stalling	3	105	104	96	27	22	24	23	20	17
Stalling	5	100	92	97	22	16	25	17	11	8

less than 1% are under 20 years. This shows that, half of our participants are older than the predominant user group of SFV applications, which primarily consists of young adults (18-25 years) [37]. Nevertheless, 47% of participants reported using SFV applications several times per day and 37% using several times per hour, indicating that the collected data is from participants familiar with SFV applications.

To ensure data reliability, the dataset is post-processed by filtering out ratings of unreliable participants. For this, we follow general guidelines and best practices as established and implemented by the QoE research community, e.g., in [34]. Our filters include the removal of entries from participants who violated the study's rules, such as switching applications or restarting the study. Next, we check the consistency of ratings. For this, we compare responses to the twice presented consistency question on the SFV application usage frequency, for which we only allow a small deviation. We further compare responses to consecutive quality questions within the same condition block. Specifically, a participant is considered inconsistent if the level of annoyance caused by the waiting time is rated as "Extremely annoying" while simultaneously the overall experience is rated as "Excellent," and vice versa. Additionally, we filter out participants who take longer than 15s to answer a question, as we cannot assume active and attentive participation in the study in these cases. For the same reason, participants are removed who reported being annoyed by a waiting time while being exposed to a condition with no waiting time. Finally, we check the intra-rater reliability of the participants by analyzing their rating trends, similar to [38]. Since larger waiting times typically result in lower QoE ratings, we identify participants whose ratings do not follow this expected pattern. Specifically, as each condition block is assigned a random condition, three ratings are collected, one for each condition. Participants exhibiting inconsistent trends, such as constant or increasing QoE ratings despite longer waiting times, are excluded. After applying the generic filters on a per participant basis, we have 106 reliable participants for the initial delay study and 97 for the stalling study. Table I gives an overview of the ratings per study and per condition.

IV. Evaluation

A. Interactions of Study Participants

To ensure that users interacted naturally despite participating in a study, we analyze the behavior of the participants in the two SFV QoE studies, *Initial Delay* and *Stalling* using the data after applying the generic filters. Since swiping plays an important role in SFVs, in Figure 3 we analyze the cumulative distribution function (CDF) of swipe times per video, i.e., the time a user watched the video before swiping to the next one. The x-axis shows the swipe time relative to the start of a video. The *Initial Delay* study is depicted in blue and the *Stalling* study in orange. In the case of the *Initial Delay* study, there is a sharp increase of the blue CDF within the first 10s of the video content. At 10s we see a jump that is caused by users who are blocked as they have to watch the last videos of a playlist, and can swipe again only after 10s. After the 10s mark, the rate of increase slows down, but a sharp rise is observed again once the videos finish, around 25-35s. When we analyze the orange CDF for the *Stalling* study, we observe that it increases at a steady but lower rate compared to the *Initial Delay* study. However, a similar sharp increase is observed once the videos finish. Thus, we generally find that most participants either skipped videos early or watched them entirely.

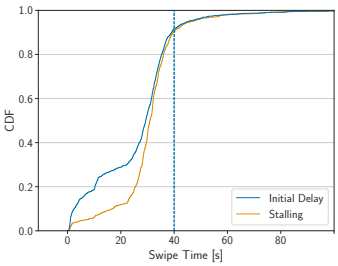


Fig. 3: Swipe Times

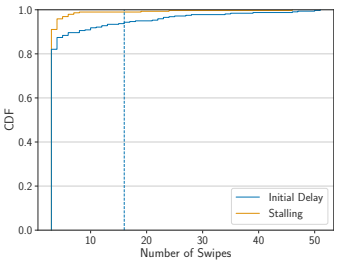


Fig. 4: Swipes per Block

A more detailed analysis of the conditions where stalling occurs shows that 6% of users swiped within the first 10s, while initial delay led to 18% swiping during the same period. This shows that initial delay was skipped more often than stalling, indicating a higher level of annoyance. One possible reason is that participants may find it easier to swipe during the initial delay, as the video has not yet started playing. In contrast, during stalling, they might be more inclined to wait since the video has already begun.

Next, we look at the CDF of the number of swipe events per condition block shown in Figure 4. The figure is structured similarly to Figure 3, with the difference being that the number of swipes is shown on the x-axis. The results show that more than 80% of users in both studies swiped three times, indicating that most participants did what was needed to do to finish the study. This does not invalidate the provided QoE ratings since the waiting time conditions were still presented in a realistic SFV context. However, the swiping behavior deviates from our expectations and is more similar to traditional LFV streaming consumption than to natural SFV behavior. We will analyze and discuss this finding in more detail below.

We also see that both distributions of swipe times and swipe events contain outliers. In the case of swipe times, in Figure 3, we observe that some users did not swipe to another video for a long time after the video finished. This is surprising, as the video shows a swipe instruction, see Figure 1b, after it is

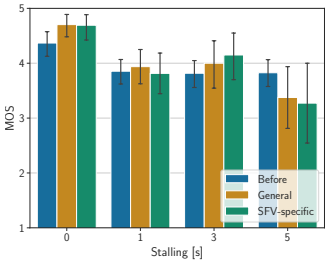


Fig. 5: Impact of Filters

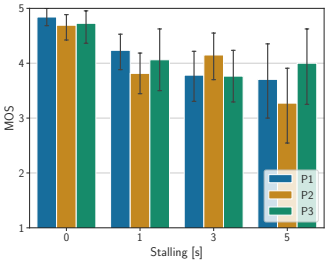


Fig. 6: Content Influence

finished. We can thus assume that these participants were not actively engaged in the QoE study. In Figure 4, we can see that other users continued swiping, generating a high number of swipes, without realizing that swipes are blocked for the last three videos in each content block. Again, this indicates that these participants were not actively watching the content. These issues, such as hyperactive users swiping too fast, and inactive users swiping too slowly, highlight the need for new filters specific to SFVs that account for users' swipe behavior to increase the reliability of the QoE results.

B. Impact of Filtering and Content

To address both hyperactive and inactive swiping behaviors, we implement and analyze the impact of a new group of SFV-specific filters. First, ratings after low engagement or inactivity in a condition block, are excluded from the dataset. We define inactivity as a swipe time occurring after more than 40s, accounting for a maximum content length of 30s, the worst-case scenario with an additional 5s of waiting time, and a reaction time of 5s. Second, we consider hyperactive swiping behavior as continued swiping despite being blocked by the mechanism, thereby generating many swiping events. Therefore, we exclude ratings from condition blocks with more than 16 swipes, which include 10 swipes for content changes and a buffer of two additional swipes for each of the last three videos, where blocking may occur. After applying the new filters, 221 reliable ratings remain from the *Initial Delay* study and 223 from the *Stalling* study, cf. Table I. Note that the SFV-specific filters are applied per condition block, which resulted in different numbers of valid ratings across playlists.

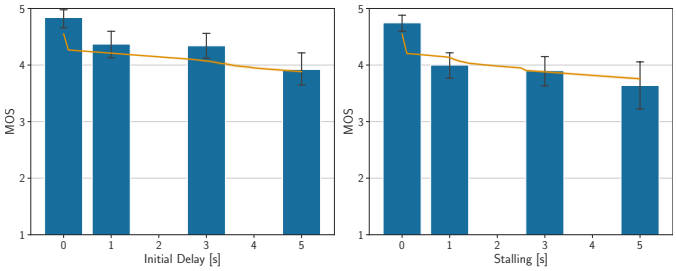
To analyze the effect of the filtering, Figure 5 illustrates its impact on the resulting QoE ratings. For this example, we use data from the *Stalling* study and the stimuli playlist P2, noting that similar results can be observed for the *Initial Delay* study and all other playlists. The figure shows the conditions on the x-axis and the MOS values on the y-axis. Since participants' QoE ratings were highly correlated across the different conditions, we focus on the MOS results from the question regarding the annoyance caused by the waiting time. Black whiskers on top of the bars represent the 95% confidence intervals. The different filters are presented using distinct colors: blue represents the results before applying any filters, orange represents the results after the general filters are applied, and green represents the final results after applying the SFV-specific filters. First, we observe that the confidence intervals overlap, indicating no statistically significant difference between the different filtering mechanisms. Second, we note that filtering does not result in a trend of increasing or decreasing the scores. Instead, by filtering out unreliable ratings, we reduce noise in the data without introducing bias. Note that this does not lead to lower confidence intervals, as this process simultaneously reduces the sample size. However, it ensures that our results include only ratings from condition blocks without extreme swiping behavior.

Figure 6 shows the impact of the different stimuli playlists on the QoE ratings for various conditions. It is structured similarly to Figure 5, with the key difference being that the different colors represent the different playlists shown to the participants: blue for P1, orange for P2, and green for P3. Our first observation is that the confidence intervals within each condition overlap, meaning we cannot conclude a statistically significant difference caused by the different content in each playlist. The MOS for each playlist within a condition also does not follow a clear trend, so we cannot identify any playlist that is consistently rated better or worse. This indicates that our content selection was successful in averaging out content influences. As the rating distributions are highly similar, in the following, we merge all ratings from the three different playlists to obtain more general QoE models.

C. Impact of Waiting Times on QoE

The results of the influence of initial delay and stalling on the QoE of SFV are presented in Figure 7, with Figure 7a displaying the results from the *Initial Delay* study and Figure 7b showing the results from the *Stalling* study. In the bar plots, the x-axis represents the condition values, while the y-axis displays the calculated MOS values. Similar to Figure 5, we present the MOS results based on the explicit question about waiting time annoyance, as other quality ratings showed strong correlations. Black whiskers on top of the bars represent the 95% confidence intervals. Based on the observed interaction behavior, we compare our SFV findings with the standardized LFV QoE assessment model P.1203 to evaluate how well they align with validated results for LFV. The QoE estimations of the P.1203 model are represented by a yellow line. For this, we provided all 30 condition videos as input to the available `itu-p1203` code [39], setting "mobile" as the device type and changing the initial delay and stalling parameters from 0s to 5s in intervals of 0.1s. We show the average O46 scores for each waiting time in the plots, leaving out the very small 95% confidence intervals for readability.

When analyzing the results for initial delay, we observe that the 0s condition without any waiting time has the highest MOS, at 4.8. The scores for the 1s and 3s conditions are lower, with values of approximately 4.4 and 4.3, respectively, followed by the 5s condition, which has the lowest score of 3.9. We can also see that the confidence intervals for the 0s condition do not overlap with those of the other conditions. In contrast, the intervals for all conditions involving initial delays (1s, 3s, 5s) overlap, indicating that no statistically significant differences can be concluded among them. When comparing our results to those from the P.1203 model, we observe that the



(a) Initial Delay (b) Stalling

Fig. 7: Impact of Waiting Times on QoE

trend is consistent, with only a minor deviation noted for the 0s condition. This could be due to the fact that crowdsourcing participants might tend to rate the conditions more favorably to satisfy the employer, cf. [40].

Next, we analyze the influence of stalling on the QoE of SFVs. In the baseline 0s condition, the MOS score is 4.7, which drops to 4.0 for the 1s condition, 3.9 for the 3s condition, and ultimately 3.6 for the 5s stalling condition. We again see that the confidence intervals for the 0s condition do not overlap with those of the stalling conditions, indicating a statistically significant difference. The values from the P.1203 model also align with the trend, with the only exception being again the 0s condition. Finally, when comparing the results of stalling to those from initial delay, we observe that the MOS of stalling is up to 0.4 lower for the same waiting time condition.

D. Discussion

In this research, we designed a framework to conduct QoE studies for SFVs on mobile phones. Crowdsourcing studies typically apply generalized filters, such as excluding participants who do not follow the rules or checking for inconsistent responses to ensure data reliability. However, compared to LFVs, SFVs introduce additional interactivity, namely swiping, which allows users to quickly start the next videos. Despite creating an SFV-like study with a realistic look and feel, our crowdsourcing participants generally did not show natural swiping behavior, leaving RQ1 partially open. This could be due to the fact that many of them are used to conducting such studies on computers and not on mobile phones. Switching to another device should not pose a problem per se. However, the focus on adhering to and finishing the task, in our case watching three videos, might lead to the observed behavior where the majority of participants waited for the first three videos to finish. Another explanation could be that this behavior may be influenced by the age group of our participants, as most are older than 25 years. Based on their past experiences with LFVs, their natural swiping behavior might more resemble traditional LFV streaming consumption, in which content is switched less frequently. In the future, age restrictions could be enforced on study participants and the training phase could be improved to provide early feedback if too slow or too fast swiping behavior is observed. This could incentivize participants to behave more naturally.

Finally, we observe some ambivalence in our findings with respect to RQ2. If we consider the results from Figure 3, we have observed that initial delay leads to more swipes, but on the other hand, initial delay conditions are rated better than the corresponding stalling conditions, cf. Figure 7. This makes it difficult to conclude which waiting time has a more severe impact on the QoE. We deliberately refrain from doing so, and instead, want to raise awareness that the focus on high interactivity in SFV poses a novel QoE modeling challenge. Classical QoE mapping functions, which take objective characteristics of the content and map to a QoE score, are no longer sufficient as they cannot solve the ambivalence observed in our study. This means, while the P.1203 model offered good MOS estimations in our studies, it does not account for users swiping away videos when annoyed rather than watching them fully. Thus, in future work, we need to develop models that can not only map to QoE scores but also describe resulting user behavior or engagement. Moreover, we might need to better research fundamental relationships between observed user interactions and self-reported subjective QoE ratings. Ideally, both indicators of QoE degradation can be combined towards a more holistic QoE assessment for interactive applications.

V. Conclusion

In this work, we developed a realistic and interactive framework to conduct QoE studies for SFVs on mobile devices. We conducted two crowdsourcing QoE studies on the impact of waiting times, i.e., initial delay and stalling, on the QoE of SFV. Our framework allowed us to analyze swiping behavior of the participants, which revealed that initial delay triggered swiping three times more frequently compared to stalling, indicating that this waiting time was perceived as more annoying. We also observed some extremely unnatural swiping behavior, namely, participants that either swiped too fast or too slow, which indicated that they did not pay sufficient attention to the study. Thus, we were required to implement SFV-specific filters to ensure the collection of reliable QoE results.

Our QoE results showed that both initial delay and stalling negatively impacted the QoE of SFV. In particular, the presence of a waiting time led to a significantly lower QoE compared to SFVs without waiting times. Comparing both types of waiting time, the MOS results suggested that initial delay had a less negative impact on the QoE than stalling, which contradicted the findings from the analysis of the swiping behavior. Thus, we refrain from providing a definitive answer to which degradation is perceived as more annoying as well as a final QoE model for SFV. Instead, we highlight that our work presents open challenges for future work. These include not only to research additional QoE parameters of SFV, such as video resolution or frame rate. More importantly, in future work, we need to research how we can motivate more natural user behavior in interactive QoE studies and how we can consider interactive user behavior as an additional indicator of QoE degradation for QoE modeling.

Acknowledgement

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant SE 3163/3-1, project number: 500105691. The authors alone are responsible for the content.

REFERENCES

[1] M. Iqbal, "TikTok Revenue and Usage Statistics (2024) - Business of Apps," 2024, accessed: 2025-06-15. [Online]. Available: https://www.businessofapps.com/data/tik-tok-statistics/

[2] S. Zhu, T. Karagioules, E. Halepovic, A. Mohammed, and A. D. Striegel, "Swipe along: A Measurement Study of Short Video Services," in *ACM MMSys*, 2022, p. 123–135.

[3] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 469–492, 2014.

[4] International Telecommunication Union (ITU), "ITU-T Recommendation P.1203: Parametric Bitstream-based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services over Reliable Transport," 2016.

[5] M. Seufert, N. Wehner, and P. Casas, "Studying the Impact of HAS QoE Factors on the Standardized QoE Model P.1203," in *ICDCS*, 2018, pp. 1636–1641.

[6] K. Brunnström, K. De Moor, A. Dooms, S. Egger-Lampl, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, C. Larabi, B. Lawlor, P. Le Callet, S. Möller, F. Pereira, M. Pereira, A. Perkis, A. Pinheiro, U. Reiter, P. Reichl, R. Schatz, and A. Zgank, *Qualinet White Paper on Definitions of Quality of Experience*, 2013.

[7] N. Barman and M. G. Martini, "QoE Modeling for HTTP Adaptive Video Streaming–A Survey and Open Challenges," *IEEE Access*, pp. 30 831–30 859, 2019.

[8] G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis, and P. I. Lazaridis, "A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction," *IEEE Access*, pp. 19 507–19 538, 2022.

[9] N. Wehner, A. Seufert, T. Hoßfeld, and M. Seufert, "Explainable Data-Driven QoE Modelling with XAI," in *QoMEX*, 2023, pp. 7–12.

[10] ——, "A Tutorial on Data-Driven Quality of Experience Modeling With Explainable Artificial Intelligence," *IEEE Communications Surveys & Tutorials*, vol. PP, pp. 1–41, 2025.

[11] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A Bitstream-based, Scalable Video-Quality model for HTTP Adaptive Streaming: ITU-T P.1203.1," in *QoMEX*, 2017.

[12] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, "HTTP Adaptive Streaming QoE Estimation with ITU-T rec. P. 1203: Open Databases and Software," in *ACM MMSys*, 2018.

[13] A. Raake, S. Borer, S. M. Satti, J. Gustafsson, R. R. R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza, G. Heikkilä, S. Broom, C. Schmidmer, B. Feiten, U. Wüstenhagen, T. Wittmann, M. Obermann, and R. Bitto, "Multi-Model Standard for Bitstream-, Pixel-Based and Hybrid Video Quality Assessment of UHD/4K: ITU-T P.1204," *IEEE Access*, pp. 193 020–193 049, 2020.

[14] H. Zhang, Y. Ban, X. Zhang, Z. Guo, Z. Xu, S. Meng, J. Li, and Y. Wang, "APL: Adaptive Preloading of Short Video with Lyapunov Optimization," in *VCIP*, 2020, pp. 13–16.

[15] D. Nguyen, P. Nguyen, V. Long, T. T. Huong, and P. N. Nam, "Network-aware Prefetching Method for Short-form Video Streaming," in *MMSP*, 2022.

[16] C. Zhou, Y. Ban, Y. Zhao, L. Guo, and B. Yu, "PDAS: Probability-Driven Adaptive Streaming for Short Video," in *ACM MMSys*, 2022, p. 7021–7025.

[17] J. He, M. Hu, Y. Zhou, and D. Wu, "LiveClip: Towards Intelligent Mobile Short-Form Video Streaming with Deep Reinforcement Learning," in *NOSSDAV*, 2020, p. 54–59.

[18] G. Zhang, K. Liu, H. Hu, and J. Guo, "Short Video Streaming With Data Wastage Awareness," in *ICME*, 2021.

[19] N. T. Phong, T. T. Huong, P. N. Nam, T. C. Thang, and D. Nguyen, "Joint Preloading and Bitrate Adaptation for Short Video Streaming," *IEEE Access*, vol. 11, pp. 121 064–121 076, 2023.

[20] B. Hou, S. Yang, F. Li, L. Zhu, L. Jiao, X. Chen, and X. Fu, "Gamora: Learning-Based Buffer-Aware Preloading for Adaptive Short Video Streaming," *IEEE TPDS*, vol. 35, no. 11, pp. 2132–2146, 2024.

[21] N. Wehner, T. Karagioules, E. Halepovic, F. Simonovski, T. Hoßfeld, and M. Seufert, "To Cap or not to Cap: Bandwidth Capping Effects on Network Interactions and QoE of Competing Short Video Streams," in *MMSys*, 2025.

[22] Z. Gao, C. Li, Y. Zhao, B. Zhang, and C. Li, "Startup delay aware short video ordering: Problem, model, and a reinforcement learning based algorithm," *Peer-to-Peer Networking and Applications*, vol. 18, p. 74, 2025.

[23] Y. Wu, R. Fu, T. Xing, Z. Yu, and F. Yin, "A User Behavior-Aware Multi-Task Learning Model for Enhanced Short Video Recommendation," *Neurocomputing*, vol. 617, p. 129076, 2024.

[24] Z. Chen, Q. He, Z. Mao, H.-M. Chung, and S. Maharjan, "A Study on the Characteristics of Douyin Short Videos and Implications for Edge Caching," in *ACM TURC*, 2019.

[25] Y. Zhang, Y. Liu, L. Guo, and J. Y. B. Lee, "Measurement of a Large-Scale Short-Video Service Over Mobile and Wireless Networks," *IEEE TMC*, vol. 22, no. 6, pp. 3472–3488, 2023.

[26] M. Diallo, F. Fieau, and J.-B. Hennequin, "Impacts of Video Quality of Experience on User Engagement in a live event," 2014.

[27] C. Moldovan and F. Metzger, "Bridging the Gap between QoE and User Engagement in HTTP Video Streaming," in *ITC 28*, 2016, pp. 103–111.

[28] N. Wehner, M. Seufert, S. Egger-Lampl, B. Gardlo, P. Casas, and R. Schatz, "Scoring High: Analysis and Prediction of Viewer Behavior and Engagement in the Context of 2018 FIFA WC Live Streaming," in *ACM MM*, 10 2020, pp. 807–815.

[29] J. Bindez, "pytubefix," 2025, accessed: 2025-06-15. [Online]. Available: https://github.com/JuanBindez/pytubefix

[30] Google, "YouTube Help - Video resolution & aspect ratios," 2025, accessed: 2025-06-15. [Online]. Available: https://support.google.com/youtube/answer/6375112

[31] International Telecommunication Union (ITU), "ITU Parameters for HDTV (PDF)," 2015, item 3.2: ITU, p.3.

[32] FFmpeg-Team, "FFmpeg," 2025, accessed: 2025-06-15. [Online]. Available: https://www.ffmpeg.org/

[33] J. de Leeuw, "jsPsych," 2023, accessed: 2025-06-15. [Online]. Available: https://www.jspsych.org/latest/

[34] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force "Crowdsourcing"," *Qualinet White Paper*, 2014.

[35] Microworkers, "Microworkers," 2009, accessed: 2025-06-15. [Online]. Available: https://www.microworkers.com/

[36] International Telecommunication Union (ITU), "ITU-T Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications," 2008, iTU-T Recommendation P.910.

[37] K. Liu, "Research on the Core Competitiveness of Short Video Industry in the Context of Big Data—A Case Study of Tiktok of Bytedance Company," *American Journal of Industrial and Business Management*, vol. 12, pp. 699–730, 2022.

[38] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," in *2011 IEEE ISM*, 2011, pp. 494–499.

[39] International Telecommunication Union (ITU), "ITU-T Rec. P.1203 Standalone Implementation," https://github.com/itu-p1203/itu-p1203, 2021, accessed: 2025-06-15.

[40] M. Varela, T. Mäki, L. Skorin-Kapov, and T. Hoßfeld, "Increasing Payments in Crowdsourcing: Don't Look a Gift Horse in the Mouth!" in *PQS*, 2013, pp. 14–19.