

Article

STGCN360: A QoE Assessment Method with Spatio-Temporal Graph Convolutional Networks for Tile-Based 360° Video Streaming

Shijia Liu ¹ , Yong Wang ¹ , Danqing Wang ², Xuan Lei ³, Junqi Chen ^{1,*} and Yuming Liu ^{1,*}

¹ School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; lsjliu123@163.com (S.L.); ywang@guet.edu.cn (Y.W.)

² College of Digital Media, Geely University of China, Chengdu 610000, China; wangdanqing@guc.edu.cn

³ Guangxi Key Laboratory of Digital Infrastructure, Guangxi Zhuang Autonomous Region Information Center, Nanning 530000, China; leixuan@gxi.gov.cn

* Correspondence: cookiechen@mails.guet.edu.cn (J.C.); ymlu@guet.edu.cn (Y.L.)

Abstract

With the proliferation of 5G, wireless networks, and other infrastructure, 360° video streaming has experienced rapid development. Efficient scheduling of 360° video streams relies on accurate feedback of user-side Quality of Experience (QoE), necessitating the construction of more precise QoE assessment methods. However, the existing tile-based QoE assessment methods for 360° video streaming have several limitations. First, full-reference video quality assessment methods require additional overhead to transmit the source video as a reference. Second, many learning-based methods are too complex. This high complexity results in heavy computational costs during training, reducing their practicality in scenarios requiring frequent model adaptation. Third, some methods rely only on simple indicators like bitrate and stall duration or spatial features in isolation. They ignore the spatio-temporal coupling inherent in 360° videos, which reduces the QoE assessment accuracy. To sum up, there is a lack of a lightweight QoE assessment method that can effectively integrate multidimensional influencing factors like the spatio-temporal features of 360° video and network state. A matching QoE assessment dataset is also missing. Therefore, focusing on tile-based 360° video streaming, this paper proposes a QoE assessment method named STGCN360. This method comprehensively considers multidimensional influencing factors, including network state and the spatio-temporal features of the video stream. To reduce complexity, it limits spatio-temporal graph modeling to the key tiles within the user's viewport, avoiding the need to process all tiles. Then, a spatio-temporal graph convolutional network (STGCN) is employed to train the QoE assessment model. Furthermore, we integrate multi-source heterogeneous datasets through feature engineering, enabling the simultaneous representation of both video quality and multidimensional factors to support the training of STGCN360. The experimental results indicate that, compared to the existing methods, STGCN360 enables more accurate QoE assessment for 360° video streaming, improving accuracy by approximately 30.79% to 32.07%. Simultaneously, the training time cost is significantly reduced, with training efficiency improved by approximately 3.7 to 5.1 times.

Keywords: tile-based 360° video streaming; QoE assessment; spatio-temporal graph convolutional network; QoE assessment datasets



Academic Editor: Jérôme Gensel

Received: 13 January 2026

Revised: 4 February 2026

Accepted: 6 February 2026

Published: 9 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Future wireless networks are evolving to support diverse scenarios, finding applications in various critical fields [1–4]. In recent years, 5G/6G technologies [5] and wireless network multimedia technologies have garnered increasing attention. Relevant surveys indicate that, by the end of 2022, the number of users owning VR devices in the United States reached 55 million, comparable to the number of Netflix subscribers in 2018 [6]. Against this backdrop, major video service providers worldwide have been offering 360° video streaming services on various platforms. This trend further propels the development of 360° video streaming technology. Additionally, the Visual Networking Index (VNI) white paper released by Cisco points out that, by the end of 2022, video traffic accounted for 82% of the total global Internet traffic [7].

360° videos are captured by multiple cameras and, after stitching and calibration, projected onto a spherical model for playback, offering users a 360° free-viewing experience. The fovea centralis of the human eye possesses extremely high-resolution capability. To achieve a fully immersive experience, 360° video requires exceptionally high resolution. This, in turn, imposes more stringent demands on video data storage, transmission, and codec capabilities. Compared to conventional video streaming, 360° video streaming demands up to 80 times the network bandwidth and is highly sensitive to latency, becoming intolerable when exceeding 10 milliseconds [8].

The emergence of tile-based transmission mechanisms for 360° video streaming in recent years [9,10] has partially alleviated bandwidth pressure, allowing the client to select higher bitrates for the user's Field of View (FoV) region of interest while choosing lower bitrates for other unobserved regions to reduce bandwidth resource consumption. Researching efficient bitrate allocation strategies for 360° video streaming can further enhance bandwidth utilization, providing users with a more satisfactory QoE. Furthermore, QoE assessment plays a guiding role in the scheduling of 360° video streams, which is crucial for service providers. The accuracy of QoE assessment methods directly impacts the performance of bitrate allocation strategies in 360° video streaming [11].

The International Telecommunication Union Telecommunication Standardization Sector (ITU-T) defines QoE as “the overall quality of an application or a service as perceived subjectively by the end-user” [12]. This definition establishes QoE as a user-centric metric, which must consider not only the influence of objective factors but also emphasizes the role of subjective factors, undoubtedly posing significant challenges for QoE assessment research. Numerous studies have been conducted on QoE assessment for conventional video streaming. With the advancement of learning-based methods, recent research has introduced more complex models into QoE assessment [13–16]. These learning-based approaches facilitate the use of diverse input features, such as video quality assessment (VQA) objective metrics and impairment factors, to accurately evaluate their impact on user-perceived QoE. The overall design of the current learning-based QoE assessment methods and the core innovations of STGCN360 are illustrated in Figure 1.

Due to the greater complexity of 360° video in terms of encoding, projection, and transmission mechanisms, the factors influencing QoE have significantly increased. Consequently, QoE assessment methods for conventional video are difficult to apply directly to 360° video [17]. To address the aforementioned challenges, the current research on QoE assessment methods for 360° video streaming primarily unfolds along two directions. On one hand, studies focus on improving 360° video quality assessment models to better adapt to the spatial characteristics of panoramic content [18–21]. On the other hand, research explores the combined impact of other features and video quality on QoE, starting from the 360° video streaming transmission system [22,23].

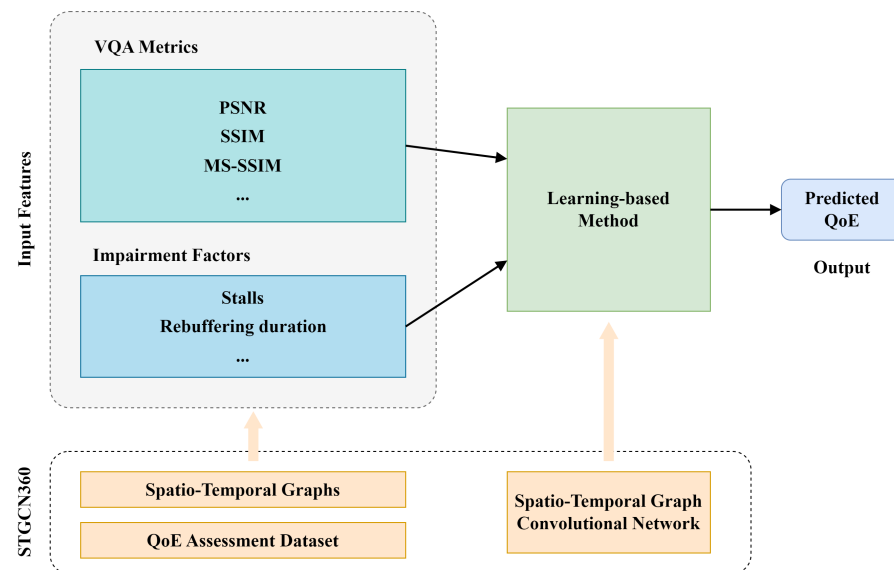


Figure 1. Overall design of learning-based QoE assessment methods.

However, the existing research still exhibits certain shortcomings. Firstly, although some learning-based methods have improved prediction accuracy, they incur prohibitively high training time overhead due to excessive model complexity, limiting their practicality. Secondly, many studies still adopt the traditional approach of uniformly assessing the entire panoramic frame, neglecting the core characteristic that users can only observe a portion of the viewport, resulting in unnecessary computational overhead. Finally, several QoE assessment methods predominantly focus on a single dimension or employ overly simplistic linear models, failing to effectively capture and model the complex dynamic coupling relationships of spatio-temporal features in 360° video streams. The lack of a QoE assessment model that is capable of simultaneously considering multidimensional influencing factors such as video quality and network state, along with a corresponding benchmark dataset, constrains model accuracy.

Focusing on the challenges identified, we focus on tile-based 360° video streaming and propose a novel QoE assessment method named STGCN360, which is based on the STGCN. This method analyzes the user's viewport trajectory during 360° video viewing, treating the tiles within the video as nodes in a graph structure. It constructs a spatial adjacency matrix based on the spatial relationship between the viewport and adjacent tiles. Building upon this, a spatio-temporal graph model that reflects the dynamic changes in the user's area of interest is established by incorporating temporal continuity. This graph structure can explicitly capture the distribution characteristics of user visual attention and its evolution over time, effectively reducing computational complexity. It provides structured spatio-temporal feature representations for QoE assessment, aiming to establish mapping relationships between various QoE-influencing factors and the MOS. Finally, we integrated a public video quality dataset with a self-constructed QoE dataset. Through feature engineering, we built a comprehensive QoE assessment dataset that is suitable for training STGCN360. Comparative experiments were conducted on this dataset to validate the performance of our proposed method.

In summary, the main contributions are summarized as follows:

- We propose a QoE assessment method for 360° video streaming based on the STGCN, named STGCN360, to establish mapping relationships between various QoE-influencing factors and the MOS. Specifically, the STGCN is employed to extract spatio-temporal features from video quality and other QoE-influencing factors, capturing dynamic changes in the temporal dimension and structural associations

in the spatial dimension. By fusing spatial features within time series, the model achieves a unified representation of multidimensional QoE-influencing factors. Finally, a feature aggregation layer outputs the comprehensive predicted MOS result reflecting subjective perception.

- Regarding tile-based 360° video streaming, we propose an explicit spatio-temporal graph modeling method focused on the viewport. This method abstracts tiles within the viewport as graph nodes, defines the spatial adjacency relationships between tiles as graph edges, and embeds multidimensional QoE-influencing factors as node embeddings. This approach functions similarly to the self-attention mechanism, but it guides the model to concentrate on key nodes and their neighboring information that shares similarities in visual content and perceptual importance through explicit spatio-temporal graph structure modeling. Thereby, it avoids redundant computations on all tiles, effectively reducing both model complexity and computational overhead during training.
- We constructed a QoE assessment dataset for STGCN360 via feature engineering. It integrates multi-source heterogeneous data and is capable of simultaneously considering video quality and multidimensional QoE-influencing factors. Based on this dataset, we conducted comparative experiments. The results demonstrate that STGCN360 delivers superior prediction accuracy while incurring significantly lower training time overhead.

The rest of this paper is organized as follows. Section 2 reviews the related work and current research landscape. Section 3 presents the problem formulation and details the proposed spatio-temporal graph modeling method. Subsequently, it provides a comprehensive description of the overall design of STGCN360 and the construction strategy for the QoE assessment dataset. Section 4 evaluates the performance of STGCN360 using the constructed dataset. Finally, Section 5 concludes the paper and discusses potential directions for future research.

2. Related Work

In this section, we first review the QoE assessment methods for video streaming, ranging from conventional to 360° video streaming. Then, we focus on the key metrics of video quality assessment and outline the mainstream research approaches. Finally, we survey existing QoE assessment datasets for 360° video streaming and analyze their limitations, thereby revealing the current research gaps and highlighting the novelty of our work.

2.1. QoE Assessment Methods for Video Streaming

Early research on conventional video streaming QoE assessment often adopted a relatively narrow perspective, with some studies primarily focusing on establishing mapping relationships between QoS metrics and QoE [24,25]. The complexity of video services has increased significantly with the advancement of adaptive streaming technologies. As a result, single-dimensional evaluation is insufficient to fully reflect the actual user experience. The impact of network transmission fluctuations and video content distortion on QoE is not a simple linear relationship; rather, it involves complex non-linear interactions and temporal dependencies. Consequently, some research has shifted toward jointly modeling integrated network QoS metrics and video quality assessment metrics. In addition to traditional heuristic algorithms [26–28], an increasing number of studies have begun to employ machine learning algorithms to address the complex correlation between multidimensional features and QoE [13,29].

To address the difficulties in feature extraction caused by the increasing variety and quantity of heterogeneous QoE features, learning-based QoE assessment models have gradually emerged as a new research direction [14]. Zhang et al. [15] proposed an end-to-end framework named DeepQoE for QoE assessment, which preliminarily explored the application of deep learning in video QoE prediction and evaluated its prediction accuracy using two datasets. Eswara et al. [16] designed an LSTM-QoE model for QoE assessment, which employs a target network composed of cascaded LSTM blocks. This model primarily focuses on the impact of varying network states on the client side to capture the non-linear and complex temporal dependencies involved in time-varying QoE. In [30], a model-based transfer learning method was proposed for the BVQA task to learn VQA frame-level features. This approach utilizes ResNet-50 for quality-aware pre-training on multiple IQA datasets and employs GRU to learn temporal information, thereby integrating spatio-temporal features with motion features. The authors of [31] introduced a lightweight CNN-based QoE model called LiteDC, which integrates temporally expanded convolutional networks with a customized pruning strategy for multi-device video streaming. By expanding the receptive field of kernels through dilated convolution operations while maintaining a streamlined model architecture, the network can effectively capture long-range dependencies without a corresponding increase in computational complexity.

Compared to conventional video streaming, QoE assessment for 360° video streaming faces unprecedented complexity and challenges. This is primarily due to its unique influencing factors, such as dynamically changing user viewports, spatio-temporal variations in video quality, complex network states, and potential motion sickness [32]. For instance, at the video content level, traditional assessment metrics struggle to accurately measure the impact of spatio-temporal relationships and spherical geometric characteristics of 360° video frames on quality perception. At the transmission level, the tile-based transmission method adopted to accommodate dynamic viewports leads to more frequent and spatially uneven occurrences of stalling and quality switching, making the impact of time-varying network states more complex. At the user level, subjective factors such as personal interests, gender, and cultural differences significantly influence viewport movement patterns, while the diverse range of user terminal devices directly determine the upper limits of rendering capability and immersion. These influencing factors interact with each other. Consequently, QoE modeling for 360° video streaming becomes a highly complex and dynamic system problem. The core challenge in this field lies in establishing accurate mapping relationships between these complex multidimensional QoE-influencing factors and the MOS. The current research in this area is still in its infancy, with several exploratory efforts as follows:

Jiang et al. [18] proposed a tile-based 360° video quality assessment method. It calculates the quality score for each tile using VMAF, generates a weight map based on head and eye movement data, and finally obtains the overall 360° video quality score by weighting the tile scores with this map. This method comprehensively considers the impact of encoding parameters, video content, and viewport. Yang et al. [19] proposed a binocular 360° video quality assessment model that combines spherical CNN and non-local attention modules. The model inputs difference maps of multiple sampled pairs of left and right panoramic video frames. Its core consists of two sets of spherical convolution operations and non-local attention modules, followed by a spherical convolution operation and a fully connected network to output the quality score. Xu et al. [33] proposed a viewport-based CNN quality assessment method named V-CNN. This method predicts and extracts viewports using CNN and Soft-NMS algorithms, employs Mini-DenseNet to predict viewport saliency and its corresponding weight, and finally fuses the quality scores of multiple viewports to obtain the overall 360° video quality score. The authors

of [22] proposed an ML-based QoE assessment method for 360° video streaming. Its core involves designing a non-invasive and lightweight heuristic method to extract key QoS features from encrypted video streams and then using machine learning to build a mapping relationship between network QoS metrics and end-user QoE scores. Zhou et al. [20] proposed a VQA method named Adaptive Score Alignment Learning (ASAL) for VR-VQA. By extending adaptive memory replay, the method constructs a novel Continual Learning (CL) framework. This framework utilizes keyframe extraction and feature adaptation techniques for 360° videos, aiming to address the non-stationary variations faced by VR devices under constraints of limited computing and storage resources, thereby enabling continuous and stable assessment of 360° video quality. However, the aforementioned methods still exhibit significant limitations. Some approaches fail to fully account for the core characteristic of 360° videos where users only watch a portion of the viewport, instead performing uniform assessment on the entire panoramic frame. This not only introduces substantial unnecessary computational overhead but also raises questions about the rationale behind the assessment logic.

In summary, the aforementioned research has sufficiently demonstrated the multifaceted nature of factors influencing video streaming QoE. Corresponding assessment models and supporting datasets have been established for different combinations of these influencing factors. Yet, these works also highlight a persistent core challenge in the field: how to systematically integrate highly heterogeneous spatio-temporal features and design QoE assessment methods that simultaneously balance high accuracy with computational efficiency. This challenge is particularly pronounced in the context of 360° video streaming with its increased feature dimensionality. The shortcomings of the existing methods in terms of both assessment accuracy and computational efficiency provide a clear direction for improvement in this paper.

2.2. Video Quality Assessment

In QoE assessment research, VQA constitutes a crucial component. Depending on whether human involvement is required in the evaluation process, the existing studies therein can be broadly classified into two categories: subjective assessment and objective assessment.

2.2.1. Subjective Assessment

Subjective assessment methods require participants to rate video quality based on pre-defined test conditions, aiming to study the degree to which human perception is affected by quality impairments. The methodology originates from psychophysics [34]. To ensure the reproducibility and verifiability of rating results, the Video Quality Expert Group (VQEG) [35] and ITU-T [12] have proposed detailed guidelines for subjective assessment methodologies. The ITU has also provided comprehensive standards for MOS scoring. However, subjective assessment suffers from inherent limitations, such as high cost, time consumption, limited participant scale, and few controllable variables, making it difficult to apply directly in practical systems. Therefore, its primary value lies in constructing high-quality benchmark datasets to guide the training and performance validation of QoE assessment models.

2.2.2. Objective Assessment

Objective assessment methods aim to establish mapping relationships between multi-dimensional influencing factors, such as video quality metrics or QoS parameters, and QoE, thereby enabling automated QoE evaluation. Among them, VQA can be categorized into three types based on the need for an original reference video: full reference (FR), reduced reference (RR), and no reference (NR) [36]. Full-reference methods require a complete and

accurate original signal source as a reference and are only suitable for controlled experimental environments. The common metrics include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and Visual Multimethod Assessment Fusion (VMAF) [37]. Reduced-reference methods utilize partial information from the source video, such as statistical features for comparing the original and distorted images, which can contribute to QoE assessment. No-reference methods do not rely on source video information. Therefore, NR-based QoE assessment models are more suitable for HTTP-based adaptive streaming systems as the client cannot obtain reference information from the source video.

Owing to the outstanding performance of Convolutional Neural Networks (CNNs) in various visual processing tasks, they have been widely applied in no-reference video quality assessment. Shen et al. [38] proposed a deep learning-based approach characterized by spatio-temporal feature fusion and hierarchical information integration. This method employs both 2D and 3D convolutional layers to progressively extract spatio-temporal features from raw videos, designs a hierarchical branch network to fuse multi-frame features, and ultimately integrates them into an end-to-end architecture through quality regression. Wu et al. [39] applied shallow neural networks and ResNet-50 to a semantic information-oriented no-reference video quality assessment model to extract both low-level and high-level features of videos. Zhang et al. [40] proposed a weakly supervised learning framework and a resampling strategy, training a CNN to construct the relationship between the 3D Discrete Cosine Transform distortion of video patches and their corresponding labels. Furthermore, in the M-3R framework designed by [41], an LSTM network was used to compute perceptual quality, capable of accurately modeling the temporal dynamics of streaming video quality under the 3R setting. This 3R setting combines the perceptual contributions of many FR, RR, and NR VQA metrics. The authors of [42] designed a framework named DeSVQ, which combines CNN and LSTM networks. This method adopts a two-stage feature processing approach: the first stage processes high-level spatio-temporal features with a CNN, and the second stage processes VQA metric features with an LSTM, finally mapping the video features to a continuous quality score through a linear layer. However, the aforementioned research primarily focuses on the visual domain of conventional video streaming. These models are often relatively complex and lack consideration for the dynamic nature of network conditions and the complex time-varying characteristics of video content. Consequently, they are more suited for scenarios where playback conditions are relatively controllable.

However, the panoramic and immersive nature of 360° videos introduces novel challenges for their VQA. This is because user-perceived quality is significantly influenced by viewing behavior (e.g., a dynamic viewport) and the projection methods employed. These projection methods fail to provide uniform pixel sampling density across the sphere. Therefore, effective quality metrics for 360° video must account for the remapping inherent in projection formats and investigate dedicated evaluation metrics. Currently, quality assessment for 360° video has emerged as a critical and growing research direction. Studies such as [21,43–45] aim to establish quality assessment frameworks that are suitable for 360° video to quantify the impact of operations like compression and encoding on video quality. This constitutes a prerequisite for building accurate QoE models for 360° video streaming.

2.3. QoE Assessment Datasets for 360° Video Streaming

In the field of conventional video streaming, there exists a substantial number of publicly available QoE assessment datasets. However, research on QoE for 360° video streaming is still in its developmental stage, and publicly accessible datasets that are suitable for comprehensive evaluation remain relatively limited. Most studies rely on

proprietary QoE datasets constructed independently, based on which the performance of corresponding QoE assessment models is validated.

For instance, the dataset constructed by Schatz et al. [46] includes two source 360° videos and 20 test 360° videos. By permuting three variables, tile-based streaming methods, head movement speed, and different quantization parameters (QPs), this dataset generated 10 different hypothetical reference circuits and employed post-stimulus scoring to rate video quality and acceptability. The dataset built by Anwar et al. [47] considers variations in 360° video resolution, QPs, initial delay, and interruptions. It provides a comprehensive evaluation of QoE across four key dimensions: immersiveness, acceptability, realism judgment, and attentional engagement. Furthermore, the dataset constructed by Fei et al. [48] utilized a network emulator to simulate wireless network environments under different network states, such as bandwidth, packet loss rate, and latency. The participants were asked to provide their ratings across four dimensions: quality score, immersion score, non-rotation sensation score, and overall score. The dataset by Zou et al. [49] consists of 10 high-quality source 360° videos and 100 processed video samples. All the source 360° videos were divided into tiles and encoded with different QPs. This study investigated the impact of dynamic quality boundary artifacts on user-perceived quality in viewport-adaptive 360° video streaming through subjective experiments and established an assessment model. Additionally, Meng et al. [50] constructed a viewport-based 360° video quality assessment dataset containing 18 original 360° videos and 774 processed video samples. The original 360° videos were clustered into two groups based on different content features for processing. Liu et al. [23] established a large-scale 360° video streaming QoE dataset named VRQoE-JUFE. This dataset includes 1440 adaptive 360° video streams with diverse content, collecting viewing behavior data and MOS ratings from 180 participants. Furthermore, the performance of state-of-the-art QoE models was evaluated on this dataset. STAV360 [51] established a video quality assessment dataset specifically for tile-based 360° video streaming. It includes six source 360° videos, which were tiled and encoded in 12 different patterns, systematically simulating quality variations occurring in real video streaming transmissions. This provides a data foundation for research on VQA methods for tile-based 360° video streaming. However, the aforementioned datasets either focus primarily on video quality distortions or include only partial network states and interaction logs. There is a lack of a publicly available benchmark dataset that systematically and synchronously encompasses multidimensional QoE-influencing factors, such as video quality and network state [23]. Simultaneously, the datasets that are currently available for tile-based 360° video streaming are relatively limited, and some valuable datasets have not been made public. This, to some extent, poses challenges for the reproducibility of research and fair comparisons between different methods.

In summary, the practicality of the current learning-based QoE assessment methods remains limited. This is primarily due to increased algorithmic overhead resulting from higher power consumption and greater computational resource demands when implementing complex QoE models. Furthermore, directly applying traditional methods that evaluate the entire video frame is both theoretically unsound for 360° video and incurs additional computational costs. Additionally, these studies lack a more comprehensive QoE model that is capable of accurately determining the impact of each individual influencing factor within adaptive streaming scenarios. Finally, the datasets employed in the aforementioned research mostly consider only a few isolated influencing factors, lacking a dedicated QoE assessment dataset designed to fuse multidimensional features.

Therefore, we propose STGCN360, a QoE assessment method designed for tile-based 360° video streaming. This method employs spatio-temporal graph structure modeling on tiles within the dynamic viewport, enabling explicit joint representation of temporal

dynamics and spatial correlations. Consequently, it can accurately extract features for QoE assessment, improving accuracy while also significantly enhancing computational efficiency. Furthermore, we introduce a preprocessing method that is capable of fusing multi-source independent datasets, constructing a comprehensive QoE assessment dataset that encompasses multidimensional influencing factors for training STGCN360.

3. STGCN360 Design

In this section, we present the detailed design of STGCN360, which includes: the problem formulation of the QoE assessment in tile-based 360° video streaming; the spatio-temporal graph modeling method for viewport tiles; the STGCN-based QoE assessment model; and the generation strategy for the QoE assessment dataset.

3.1. Problem Formulation

The QoE assessment problem in tile-based 360° video streaming can essentially be summarized as a process of representation learning from multidimensional complex features and establishing a mapping relationship between these features and MOS scores. The features influencing QoE originate from various dimensions and types, such as the transmission system, network state, and user behavior. These features exist in a sequential temporal order during video playback. Therefore, how to simultaneously model temporal and spatial features and perform joint learning becomes a key challenge in QoE assessment. In summary, we posit that an effective QoE assessment modeling method must possess the ability to concurrently process the spatio-temporal features of the video and demonstrate strong generalization capability and extensibility.

Therefore, we model the QoE assessment problem within a tile-based 360° video adaptive streaming system as follows: the video is divided into C consecutive chunks, each with an equal duration L . For any video chunk c ($c \in C$), the 360° video in ERP projection is further partitioned into $i \times j$ uniform tiles, denoted as V_{ij}^c . Each tile is encoded with three different quantization parameters Q_{ij}^c and organized via twelve encoding patterns to simulate the actual adaptive bitrate streaming process. During viewing, users can freely change their viewing perspective based on personal preference. We extract the corresponding quantization parameters Q_{mn}^c from the $m \times n$ tiles V_{mn}^c within the currently observed viewport. Concurrently, network states during the transmission of chunk c are collected, including throughput P^c , rebuffering time R^c , and buffer occupancy B^c . Based on the spatio-temporal characteristics of the viewport, the tile quantization parameters, and the network state features, a spatio-temporal graph is constructed. Finally, an extensible model is employed to predict the user's subjective quality perception MOS score after watching the complete video.

3.2. Modeling the Spatio-Temporal Graph of the Viewport

In current research on 360° video streaming QoE assessment, most methods incorporate the entire projected frame into the computational scope, overlooking the unique viewport-dependent nature of such videos. During actual viewing, users can only perceive the video content within their viewport, while the majority of areas outside the viewport remain invisible. Quality variations in these invisible regions have minimal impact on the user's actual QoE. Notably, tile-based 360° video streaming can leverage this characteristic by dynamically adjusting the preloaded quality of tiles in different regions to conserve transmission bandwidth. However, the existing methods still uniformly process and extract features from the entire frame. This approach not only introduces substantial computation unrelated to subjective perception but also results in significant wastage of computational resources. Particularly in high-resolution multi-bitrate adaptive 360° video streaming

scenarios, such computational redundancy grows exponentially with video complexity, severely limiting the practicality and scalability of QoE assessment models. Therefore, focusing on modeling the spatio-temporal graph exclusively for the user's actual viewport can effectively eliminate interference from irrelevant information, ensuring assessment accuracy while substantially reducing system training time overhead.

Therefore, our study focuses on explicitly addressing the QoE assessment for tiles within the viewport. We further formulate this problem as a regression task. First, a graph $G = (V, E, A)$ is constructed, where the graph nodes V are defined as the $m \times n$ tiles covered by the viewport, and the number of nodes is $N = |V|$. The graph edges E represent the spatio-temporal relationships between tiles. As illustrated in Figure 2b, each node is connected only to its immediate horizontal or vertical neighboring nodes, forming the adjacency structure of the graph. Accordingly, A denotes the adjacency matrix of this graph. Assuming nodes are only connected to their immediate orthogonal neighbors, if a direct edge exists from a node to one of its direct horizontal or vertical neighbors, the corresponding entry in the adjacency matrix is set to 1; otherwise, it is 0.

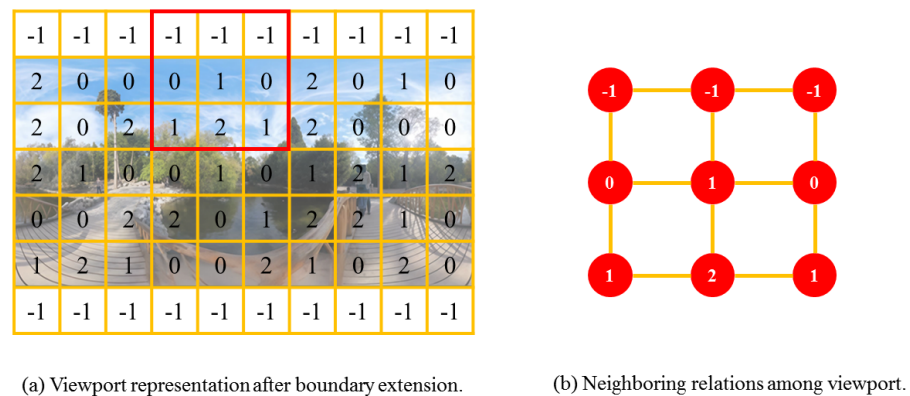


Figure 2. Illustration of the graph construction method based on the viewport at timestamp t .

Furthermore, we process the adjacency matrix using the normalized Laplacian matrix to obtain $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D is the degree matrix. Since node features change dynamically over time, the features of tile n at timestamp t are represented as a d -dimensional feature vector $X_n^t = [x_1, x_2, \dots, x_d]$, where each dimension incorporates multi-source information such as video quality, network state, and user behavior. Therefore, for the entire video playback duration T , the model input is the temporal feature sequence $(X^1, X^2, \dots, X^T)$, and the output is the predicted subjective MOS score of the video tile.

Considering the spatial particularity of 360° videos under ERP projection, when the user's viewport is at certain specific viewing angles, such as near the boundaries or polar regions, the actual number of covered tiles may decrease. This leads to dynamic changes in the number of graph nodes and their connectivity, making it unsuitable to model with a fixed static graph structure. Therefore, to ensure structural consistency and simplify computation, we apply a novel boundary extension strategy that expands the tile coverage at the polar regions of the frame.

The specific boundary extension strategy is as follows: when the viewport coverage is insufficient, virtual tiles are logically used to pad the upper and lower boundaries. As shown in Figure 2a, the feature values of these padded tiles are set to -1 . This setting helps to distinguish whether a tile belongs to the padded border regions or the actual video content. In computation, these tiles correspond to the low-quality level in QP encoding. In this paper, we map different QP levels to discrete numerical values for computational convenience; for example, low, mid, and high are encoded as 0, 1, and 2, respectively.

Using encoding pattern 12 shown in Figure 5 as an example, after boundary extension and flattening, the QP sequence corresponding to the tiles within the viewport in Figure 2a can be represented as $[-1, -1, -1, 0, 1, 0, 1, 2, 1]$. By applying this strategy, the number of nodes in the graph structure remains constant, thereby enabling the model to perform efficient spatio-temporal feature learning on a unified and static graph structure.

This paper introduces a spatio-temporal adjacency matrix to model the spatio-temporal dynamics of 360° videos. Inspired by the unified spatio-temporal graph convolution network [52], we integrate spatial adjacency relations and temporal state transitions into a unified spatio-temporal graph structure by combining spatial and temporal adjacency matrices. The schematic of the specific spatio-temporal graph construction method is illustrated in Figure 3. At each timestamp t , each graph node is connected only to its neighboring nodes along the horizontal and vertical axes of the graph structure. Multidimensional features of all nodes across all historical timestamps are extracted to regress and predict the subjective perception MOS score for the viewport where the blue node is located. Through this spatio-temporal graph architecture, the model can more accurately capture the complex spatio-temporal dependencies among neighboring nodes across different timestamps.

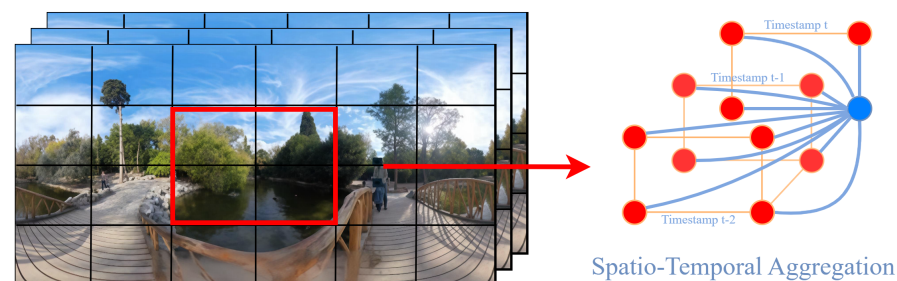


Figure 3. Illustration of the spatio-temporal graph construction method. The target node for prediction is marked in blue.

To correctly encode complex spatio-temporal dependencies across timestamps and achieve effective feature aggregation, we employ a spatio-temporal adjacency matrix A_{ST} constructed over T timestamps. Each element of this matrix is an $N \times N$ sub-matrix, with its specific structure given by Equation (1). This spatio-temporal adjacency matrix is designed as a lower triangular matrix to enforce a causal constraint: the prediction for the target node (the blue node in Figure 3) at timestamp t relies only on the features of itself and its neighboring nodes from timestamps 1 to $t - 1$. Consequently, all elements in the upper triangular part of the matrix are set to 0, indicating that no future features are aggregated, only historical ones. The diagonal elements are set to the spatial adjacency matrix A , ensuring that, within each timestamp, a node can aggregate features from its directly connected spatial neighbors. The lower triangular part (excluding the diagonal) is set to $\tilde{A} = A + I$, where I is the identity matrix. This configuration allows a node, at any given timestamp, to aggregate the combined spatio-temporal features of itself and its spatial neighbors from all previous timestamps.

$$A_{ST} = \begin{bmatrix} A & \mathbf{0} & \cdots & \mathbf{0} \\ \tilde{A} & A & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \tilde{A} & \cdots & \tilde{A} & A \end{bmatrix} \in \mathbb{R}^{NT \times NT} \quad (1)$$

This structured design yields a key advantage: by separating the diagonal component A from the lower triangular part $\tilde{A} = A + I$, the model can differentiate between the contributions of the target node itself and those of its neighboring nodes rather than

simply blending their features. This enhances the model's ability to distinguish and model heterogeneous information sources, thereby improving its overall expressive power. Conversely, using only $\tilde{A} = A + I$ throughout would homogenize the embeddings of the node and its neighbors, leading to a degradation in the model's expressive capability.

3.3. STGCN-Based QoE Assessment Model

The overall framework of the proposed STGCN360 is illustrated in Figure 4. It is designed as a lightweight model. The model takes a temporal multidimensional feature sequence as input, which primarily includes the tile quality within the viewport, denoted as Q_{mn}^c , and network state during video chunk transmission, namely throughput P^c , rebuffering time R^c , and buffer occupancy B^c . Given a more comprehensive multidimensional QoE-influencing factor dataset, the model can be extended to support inputs from additional feature dimensions. Subsequently, a viewport-based spatio-temporal graph is constructed to organize the inputs into a structured spatio-temporal feature sequence. This sequence is then fed into a multi-layer STGCN encoder for in-depth spatio-temporal feature fusion and extraction. The design of this layer is extensible, allowing for the flexible integration of various other lightweight large models to more effectively extract high-level semantic features from different types of data (e.g., non-spatio-temporal sequential data). Finally, a fully connected regression prediction layer aggregates the multidimensional features to output the user's subjective perception MOS score. This process accomplishes the mapping from raw multidimensional spatio-temporal features to the final subjective MOS rating.

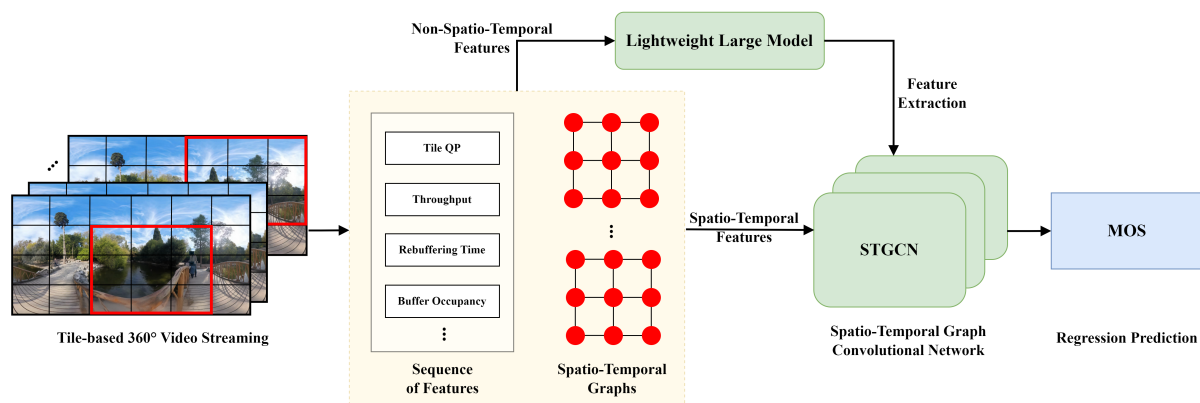


Figure 4. Overall framework of STGCN360.

To address the aforementioned challenges, our solution builds upon the unified spatio-temporal graph convolution network [52]. We adapt and enhance its network architecture for application in the novel domain of tile-based 360° video streaming QoE assessment.

The algorithm first consolidates input features $X^T \in \mathbb{R}^{N \times d}$ across T timestamps into a unified matrix $X_{self} \in \mathbb{R}^{NT \times d}$. Subsequently, spectral graph convolution is executed based on the spatio-temporal adjacency matrix A_{ST} and the feature matrix X_{self} , following the formulations in Equations (2) and (3). This convolution employs a learnable temporal weight parameter $W_{temp} \in \mathbb{R}^{NT \times d}$ to adaptively model the significance of node features across different timestamps from 1 to T . The diagonal degree matrix D_{ST} , required for normalization, is derived from A_{ST} , with each element defined as $D_{ST}[i, i] = \sum_{j=1}^{NT} A_{ST}[i, j]$.

The propagation of spatio-temporal features to the subsequent layer is governed by Equation (4). As depicted in Figure 5, the spatial adjacency matrix A is formed by connecting nodes both horizontally and vertically within a 3×3 viewport matrix. The spatio-temporal adjacency matrix A_{ST} is then constructed by combining A with its self-connected counterpart $\tilde{A}$. At convolutional layer l , the input features across timestamps are stacked

into X_{self}^l . Each element within X_{self}^l is scaled by the temporal weight W_{temp}^l , which encodes the relative importance of features at each timestamp. The graph convolution operation proceeds via a weighted fusion of the self-representation X_{self}^l and the spatio-temporally aggregated vector X_{ST}^l , where the weight matrix W_{final}^l learns to balance the contributions from the prior timestamp's self and aggregated features. The resulting sum is passed through a non-linear activation function $\sigma(\cdot)$ (typically ReLU) to produce the self-representation X_{self}^{l+1} for layer $l + 1$, which serves as the input to the next layer or as the final representation. The process concludes by applying two fully connected layers to predict the MOS score for each tile within the viewport, thereby completing the overall QoE prediction task.

$$\tilde{A}_{ST} = D_{ST}^{-\frac{1}{2}} A_{ST} D_{ST}^{-\frac{1}{2}} \quad (2)$$

$$X_{ST}^l = \tilde{A}_{ST} (X_{self}^l \cdot W_{temp}^l) \quad (3)$$

$$X_{self}^{l+1} = \sigma(W_{final}^l (X_{self}^l || X_{ST}^l)^T)^T \quad (4)$$

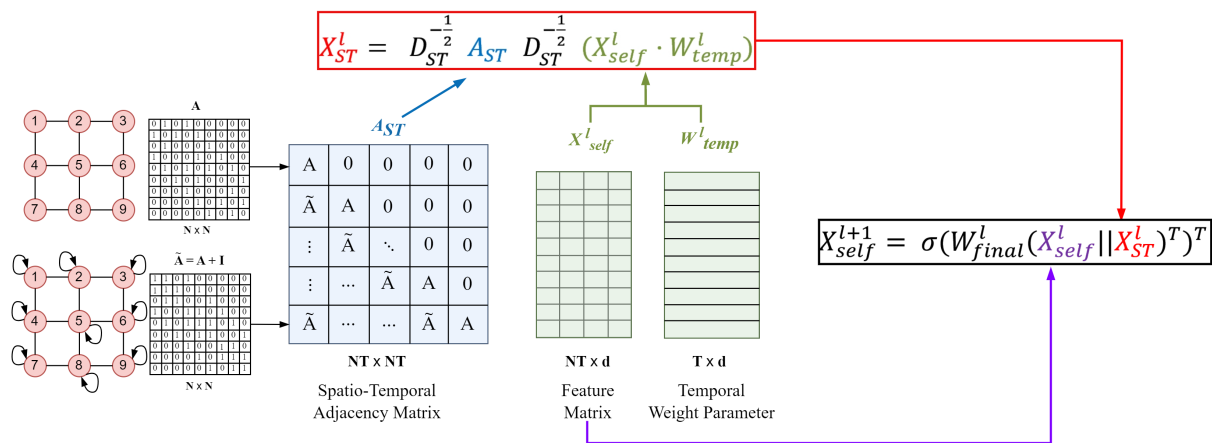


Figure 5. Architecture of a single STGCN layer.

3.4. Generation Strategy for the QoE Assessment Dataset

Current research primarily follows two branches: the first focuses on VQA within the field of computer vision, where recent studies predominantly employ neural networks to evaluate image quality; the second investigates multidimensional QoE-influencing factors such as device characteristics, network states, and user behavior. Both branches have corresponding publicly available datasets for validating the performance of QoE assessment models. However, for tile-based 360° video streaming, there is currently a lack of a publicly accessible QoE assessment dataset that simultaneously incorporates both VQA and multidimensional QoE factors, making it difficult to support the training of a comprehensive and accurate QoE assessment model. Therefore, we have integrated public tile-based 360° video VQA datasets with a self-constructed QoE dataset. Through feature engineering and alignment with operational logic, we have built a QoE assessment dataset suitable for training STGCN360. In future work, we plan to further develop a practical QoE assessment dataset that includes a more comprehensive set of QoE-influencing factors, providing a training foundation for QoE assessment models that better reflects real-world application environments. The detailed processing methodology for this QoE assessment dataset will be introduced next.

3.4.1. Viewport Trajectory and Video Quality

In tile-based 360° video streaming systems, to optimize bandwidth utilization, when a client requests the next video chunk, the system first obtains the viewport likely to be

watched by the user through viewport prediction. It then allocates appropriate video quality levels accordingly: high quality for encoding the key areas within the predicted viewport, while medium and low qualities are used for encoding the background areas outside the viewport. This strategy enables differentiated bandwidth allocation. During actual viewing, users freely watch the downloaded video content based on their personal behavioral preferences.

To construct data that aligns with the system described above, this paper utilizes the head movement trajectories provided by the STAV360 dataset [51] as the basis for authentic user viewing behavior, with the viewport being determined accordingly based on this trajectory. The dataset not only contains continuous head movement trajectories recorded in the form of unit quaternions but also encodes each tile using three QPs: high (QP = 22), mid (QP = 32), and low (QP = 42). Additionally, 12 different encoding patterns were designed for each video to simulate the impact of different scenarios on user-perceived quality. The visual representation of these patterns is shown in Figure 6. This provides data support for subsequently extracting video quality features corresponding to the viewport.

The specific data preprocessing pipeline is as follows: First, the continuous viewport spherical coordinate trajectory is decoded from the raw head movement trajectories in the form of unit quaternions and then mapped onto the 2D ERP projection plane to identify the corresponding specific tile. Subsequently, centered on this viewpoint tile, a 3×3 viewport covering its surrounding area is constructed, thereby transforming the original single-point trajectory data into a spatially more representative viewport trajectory sequence. Finally, by incorporating the aforementioned 12 tile encoding patterns, the QPs of the corresponding tiles are extracted from the viewport at each timestamp to serve as video quality features for subsequent processing.

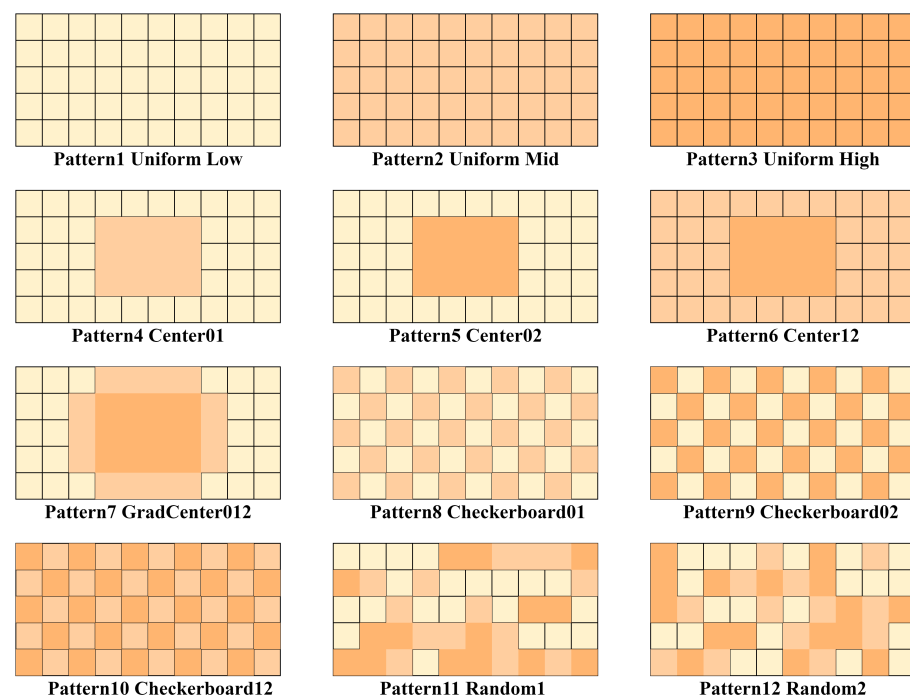


Figure 6. The 12 tile encoding patterns used in the dataset. The colors, ranging from light to dark, correspond to three different QPs.

3.4.2. Multidimensional QoE-Influencing Factors

Regarding the self-constructed dataset of QoE-influencing factors, we systematically collected fine-grained network state time-series data and user interaction logs during runtime based on the 360° video streaming system interaction environment built in [53].

After the raw data collection, rigorous data preprocessing was performed: first, multi-source data streams were aligned by timestamp; then, duplicate, missing, or anomalous entries were cleaned; finally, the cleaned data were integrated and restructured into a multidimensional QoE-influencing-factor dataset, which includes status features such as throughput, rebuffering time, and buffer occupancy, along with corresponding MOS scores. The core value of this dataset lies in the fact that it is not merely an isolated collection of network states. Rather, it fully encapsulates interactive data reflecting the causal chain from network states and system adaptive decisions to client playback states, thereby providing a more detailed representation of the complex factors affecting the final QoE. This dataset lays the foundation for subsequent logical alignment and integration with video quality features.

3.4.3. Feature Engineering for the QoE Assessment Dataset

Based on the processed viewport trajectory data and its corresponding 12 tile encoding patterns, we map QP values to bitrate levels low (QP = 42), corresponding to bitrates {1, 2.5} Mbps; mid (QP = 32) to {5, 8} Mbps; and high (QP = 22) to {16, 40} Mbps, and randomly match each viewport trajectory with a series of QoE-influencing factor data (including throughput, buffer occupancy, rebuffering time, etc.) from the self-constructed QoE dataset to simulate complex and dynamic network variations. Through temporal alignment, video quality features (from the public STAV360 dataset) and multidimensional QoE features (from the self-built QoE dataset) are fused to form multidimensional embeddings at each node of STGCN360, which are then fed into the model for training to obtain predicted MOS scores.

For constructing the training labels of the QoE assessment dataset, we did not directly adopt the single subjective MOS score from either dataset. Based on our fundamental assumption regarding the operational logic that video quality and multidimensional QoE-influencing factors are considered equally important, we fused the MOS scores from the two data sources using a weight of 0.5 for each. This fixed weighting reflects our prior knowledge of multi-factor influence and aims to guide the model to learn a QoE assessment logic that aligns with our defined operational rules rather than treating it as a hyperparameter to be optimized to fit any existing MOS score. Since our dataset is simulated, we have arbitrarily set the weights for the two data sources to be equal. The values can be adjusted according to specific requirements. In future work, we plan to collect datasets from real-world environments. The setting of this weight is not the primary focus of this paper. Furthermore, by varying the random seed, we can generate multiple distinct network-state matching combinations for the same viewport trajectory. This approach effectively expands the dataset scale, simulates complex and varying network states, and thereby enhances the diversity and robustness of model training.

4. Experimental Analysis

4.1. Experimental Setup

We implemented STGCN360 on a PC equipped with an 8-core Intel i7 processor and 16 GB of RAM. The key parameters of STGCN360 are configured as follows: the temporal length of the video spatio-temporal graph sequence is set to $T = 10$, corresponding to the total duration of the video; the viewport is set to 3×3 , resulting in a fixed number of nodes $N = 9$ in the spatio-temporal graph at any timestamp; the embedding dimension for each node is $d = 4$, corresponding to the 4 key input features. The number of graph convolutional network layers is set to 1. This hyperparameter was determined through extensive experimental tuning. Only with this configuration can the model learn effectively, while other layer settings lead to convergence difficulties. The model is trained using the Adam optimizer with an initial learning rate of 1×10^{-4} . These hyperparameters are set

according to [53]. To ensure a fair comparison, all baselines were evaluated under the same experimental setup. For MLP, LSTM, and GRU, the hidden layer dimension was set to 128, using the Adam optimizer with a learning rate of 1×10^{-3} , and each consisted of 2 neural network layers. For the Transformer model, the embedding dimension was set to 128, with 8 attention heads, a learning rate of 1×10^{-3} , and 2 encoder layers.

4.2. Dataset

We employed the dataset provided by STAV360 [51]. The 360° videos in this dataset were captured at 8K resolution (7680×4320) and a frame rate of 30 fps. The video content encompasses six distinct environmental types: FeedTheDucks, FootballFreestyling, LycabettusSunset, MuseumOfTheAncientAgora, PiraeusPort, and TempleOfHephaestus. Each video was tiled into 10×5 . Individual tiles were encoded at three quality levels defined by their QPs: high (QP = 22), medium (QP = 32), and low (QP = 42). Test sequences were then constructed using 12 different tiling patterns. A total of 27 participants (18 male, 9 female) were recruited for subjective testing. They provided MOS ratings on a scale of 1–5 for the aforementioned video sequences, and their head movement trajectories during viewing were recorded. This process resulted in the formation of this tile-based 360° video streaming VQA dataset. Since each user in the dataset includes head movement trajectories from viewing multiple videos, we divided the data on a per-user basis. Specifically, we randomly selected and retained trajectories from 20% of the users to form an independent test set. This test set remains completely unseen during training and validation and is used solely for final performance evaluation. The remaining 80% of user trajectories constitute the training set. Given the relatively limited scale of the generated dataset, we employed cross-validation during training and hyperparameter tuning to ensure the robustness of the model.

4.3. Baselines

MLP [54]: The multi-layer perceptron (MLP) is a classic fully connected feedforward neural network that served as the foundational model for many early QoE prediction studies. We employ MLP as a simple baseline to highlight the necessity of introducing spatio-temporal structure modeling.

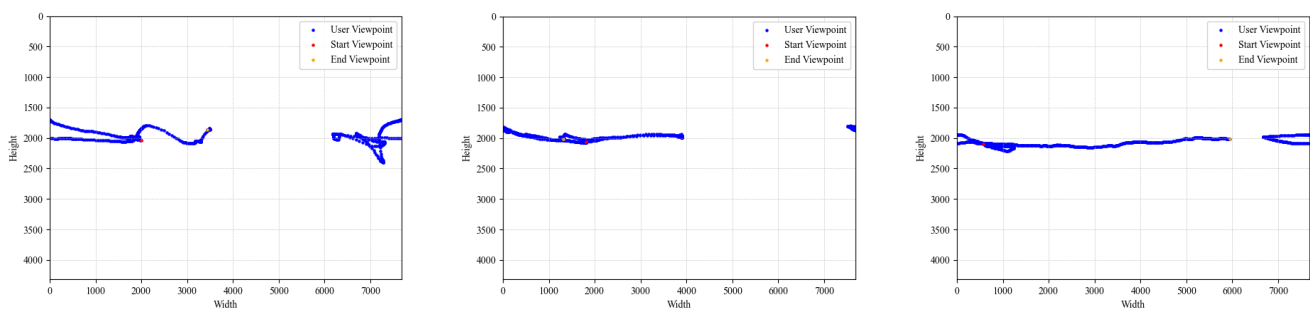
LSTM [16]: Long Short-Term Memory (LSTM) is a canonical model for processing sequential data, whose gating mechanism excels at capturing long-term dependencies. We extract the core prediction module from [16] and integrate it into the same experimental environment as STGCN360. This allows for a performance comparison to assess the impact of neglecting spatial features.

GRU [30]: The Gated Recurrent Unit (GRU) is a variant of LSTM that simplifies the model architecture by merging gating units, thereby achieving higher computational efficiency. Similarly, we extract the GRU module from [30] and enable it to learn dependencies within the input feature sequence under the same experimental conditions. This setup facilitates observation of the performance of different Recurrent Neural Network architectures on this specific QoE assessment task.

Transformer [55]: The Transformer, based on a self-attention mechanism, offers advantages in capturing long-range dependencies and enabling parallel computation. Ref. [55] utilized a Transformer for viewpoint prediction. We adapt its core Transformer module to function as a general-purpose spatio-temporal feature extractor. It learns global dependencies among features from the input sequence via self-attention to accomplish QoE assessment for 360° video streaming.

4.4. Analysis of the QoE Assessment Dataset

Figure 7a–c illustrate the head movement trajectories of three users watching the same 360° video content. In the figures, the blue points represent the user's viewpoint records, the red point indicates the viewpoint position at playback start, and the yellow point marks the position at playback end. Together, they form the user's head movement trajectory. Through cross-user comparison, we depict two key patterns in head movement trajectories. First, spatially, all users' attention is consistently drawn to similar high-interest content, demonstrating significant cross-user spatial aggregation and similarity. Second, temporally, viewpoint movement follows a typical pattern combining local smooth exploration with rapid jumps across different regions. Together, these patterns indicate that the visual attention features influencing QoE exhibit not only significant content-driven commonality spatially but also discernible dynamic patterns temporally.



(a) Head movement trajectory of User 1. (b) Head movement trajectory of User 2. (c) Head movement trajectory of User 3.

Figure 7. Head movement trajectories of different users watching the same video content.

Based on this finding, we argue that performing uniform QoE assessment on the entire video frame is neither efficient nor necessary. Instead, STGCN360 can dynamically identify and model the actual viewport being watched by the user and its spatio-temporal contextual features, thereby directly capturing this common attentional structure. Consequently, we propose constructing a spatio-temporal graph specifically for the viewport. In this graph, spatial adjacency effectively models dependencies for local smooth exploration, while temporal connections capture the dynamic shifts of the viewpoint between areas of interest. This provides an explicit, efficient, and structured method for extracting core QoE-influencing features, thereby reducing computational complexity.

We aim to construct a comprehensive QoE assessment dataset that integrates both video quality and multidimensional QoE factors by merging the STAV360 VQA dataset with our self-constructed QoE dataset. As shown in Figure 8, the MOS distribution of the generated dataset clusters towards the center, with fewer extreme scores at both ends, merging predominantly into the score of 2. This shift originates from the introduction of QoE-influencing factors. In real-world 360° video streaming scenarios, user ratings are a composite result of multiple factors such as video quality, stalling, and quality switching; instances of perfection in any single dimension are rare. Therefore, the new distribution more realistically simulates the distribution of user experience under complex conditions. Although the number of theoretically perfect (score 5) samples has decreased, the new dataset maintains sufficient sample density and distribution continuity in the main scoring range, fully capable of supporting STGCN360 in learning a robust mapping from multidimensional features to subjective MOS scores. The more concentrated and complex dataset forces the model to deeply explore the relationships between input features rather than relying on a simple linear mapping from quality to score. This explains, at the data level, why STGCN360 demonstrates superior generalization and robustness compared to

baselines in subsequent comparative experiments, especially when facing complex and dynamic scenarios.

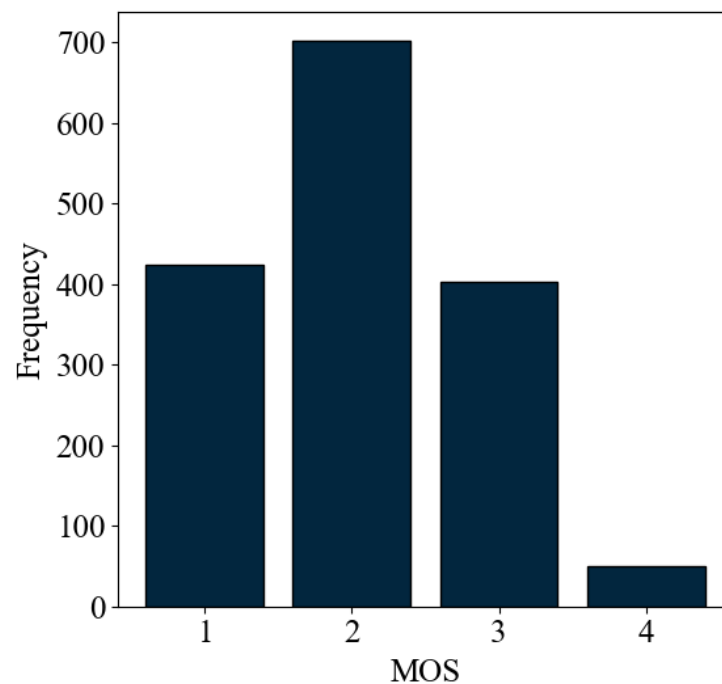


Figure 8. MOS distribution of the generated QoE assessment dataset.

4.5. Performance Analysis of STGCN360

Regarding the configuration of evaluation metrics, since the MOS scores provided in the STAV360 dataset are all discrete integers (ranging from 1 to 5), we applied post-processing to the regression predictions of STGCN360. The continuous predicted values were rounded to integers, and consistency across multiple node outputs was ensured through mode voting, resulting in integer MOS scores. These scores were then used to calculate accuracy against the dataset labels. This metric most directly reflects the alignment between model outputs and ground truth labels, enabling clear and intuitive performance comparison. Furthermore, since the dataset contains trajectories of each user watching multiple different 10 s videos, STGCN360 predicts the final MOS based on the spatio-temporal features within each individual 10 s video clip. To ensure that the model learns generalizable features rather than temporal noise, the dataset was randomly shuffled before training and testing, thereby disrupting the inherent sequential order of the trajectories under each user. Under this setting, traditional metrics such as PLCC and SROCC, which are designed to measure the linear correlation between predicted and subjective scores along the time dimension, are no longer suitable. Therefore, this study focuses on using accuracy to evaluate the model's prediction performance. Additionally, the MSE loss learning curves are provided as a supplementary analysis of the training and testing process, ensuring that the evaluation metrics align with the training setup.

Moreover, to ensure the statistical reliability and reproducibility of the experimental results, all reported numerical results in this study were obtained through the following procedure: for each experimental configuration, 5 independent trials were conducted by varying the random seed. The final result is the arithmetic mean of the outcomes from these 5 runs. The fluctuation range of the results is visualized using error bars representing plus or minus standard deviation.

We first conducted experiments on the full dataset, evaluating the proposed STGCN360 against various baselines from two key dimensions: prediction accuracy and training

time overhead. As shown in Figure 9, under the same dataset and experimental setup, STGCN360 achieved the highest accuracy of 0.701, significantly outperforming all compared baselines. Specifically, STGCN360 improved accuracy by approximately 31% to 32% relative to Transformer, GRU, LSTM, and MLP. This strongly validates the effectiveness of explicit spatio-temporal graph structure modeling for QoE assessment. By constructing a spatio-temporal graph, STGCN360 precisely captures the spatial adjacency relationships and temporal variation patterns of tiles within the viewport, reduces computational complexity, and thus effectively learns the complex spatio-temporal features of 360° videos.

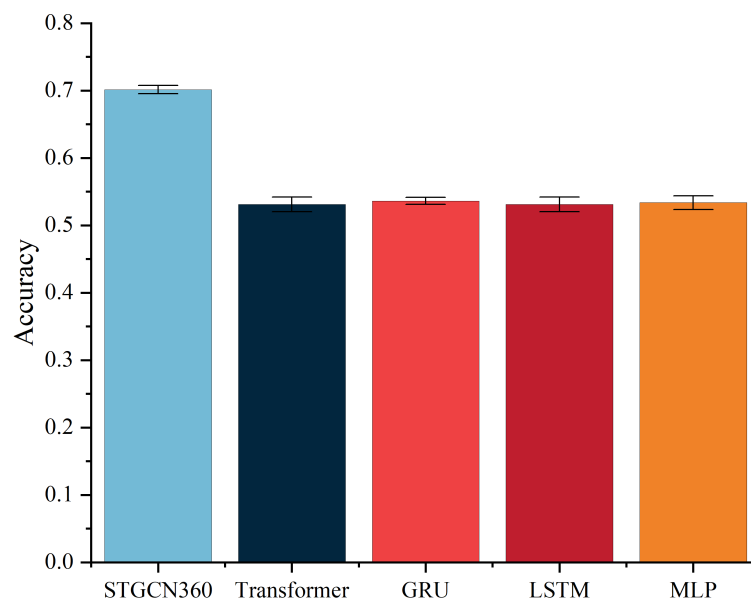


Figure 9. Comparison of accuracy among different QoE assessment methods.

In contrast, while Transformer possesses strong time-series modeling capabilities, its self-attention mechanism lacks explicit spatial topological priors for 360° video streaming, which may introduce irrelevant noise. Coupled with interference from the large-scale complex dataset, this prevents it from correctly learning the underlying mapping relationships. Meanwhile, GRU, LSTM, and MLP are constrained by their inherent algorithmic designs. Specifically, GRU and LSTM lack explicit modeling capability for the spatial relationships among tiles, while MLP struggles to effectively handle sequential time-series tasks. Consequently, their performance encounters a common ceiling as none can adequately capture the spatio-temporal correlation within the input features. This leads to similar accuracy performance among them, indicating a state that is nearly unpredictable.

In terms of training time overhead comparison, as shown in Figure 10, during the training process required to achieve their respective optimal performance, STGCN360 incurred the lowest training time overhead, taking only 320.2 s, which demonstrates exceptionally high training efficiency. The most time-consuming model was Transformer, requiring 1618 s. This is primarily due to the quadratic computational complexity introduced by its self-attention mechanism, leading to a significant drop in training efficiency when dealing with large-scale datasets. The training time overhead for GRU and LSTM was moderate, at 1299.8 s and 1402.4 s, respectively. Their inability to model spatial structure is a key factor. To process spatial features, the spatial dimensions must be flattened, which results in the loss of structural information and forces the models to implicitly learn spatial relationships from the data, thereby reducing learning efficiency. Notably, even the structurally simplest model, MLP (1185 s), exhibited low efficiency due to high-dimensional input and slow

convergence. Its training time overhead still exceeded that of STGCN360 by approximately 3.7 times, indicating poor practicality.

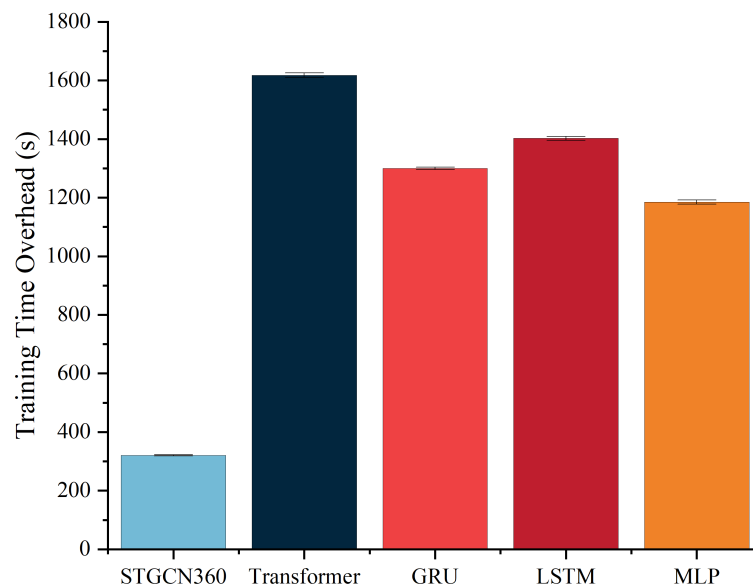


Figure 10. Training time overhead of different QoE assessment methods.

To further explore the optimization behavior and generalization capability of the models, the MSE loss learning curves for STGCN360 and Transformer on both the training and test sets are presented in Figure 11a and Figure 11b, respectively. Other baselines similarly exhibited difficulties in learning; therefore, we use Transformer as a representative model for analysis. The learning curves demonstrate that STGCN360 has learned correctly, displaying significantly superior and more stable convergence characteristics. Its loss in both training and testing phases decreased rapidly after 20 epochs and converged below 0.1. The two curves are closely aligned, indicating that the patterns learned by STGCN360 from the training data generalize well to the test data, with no obvious signs of overfitting. In contrast, Transformer's learning curve started from a higher initial value, decreased slowly with considerable fluctuation throughout the training process, and eventually converged around 0.45. This reflects Transformer's difficulty in efficiently learning stable spatio-temporal feature mappings from the data of this task, resulting in weaker generalization capability.

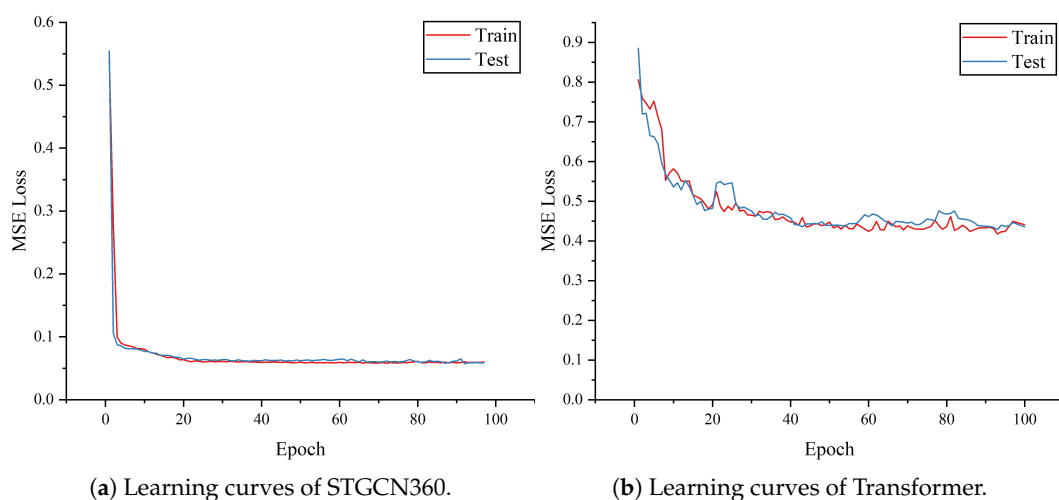


Figure 11. Learning curves of MSE loss for the QoE assessment method.

Through a combined analysis of the learning curves, accuracy, and training time overhead, a consistent and mutually reinforcing conclusion can be drawn: for the problem of 360° video streaming QoE assessment, which possesses a distinct spatio-temporal topological structure, with its specialized architecture utilizing explicit spatio-temporal graph modeling, STGCN360 demonstrates significantly superior performance in terms of model accuracy, training efficiency, and convergence stability compared to generic sequence prediction models that rely on implicit relationship learning.

However, as a lightweight model, STGCN360 has limited model capacity, and its learning ability is prone to reaching a bottleneck when handling complex large-scale data [56]. Therefore, to further explore the relationship between performance and data complexity, we constructed smaller-scale datasets for training and testing by reducing the number of user trajectories included in the original dataset. This allows us to analyze the impact of data scale and the complexity of spatio-temporal relationships on the model's generalization ability. These reduced datasets are referred to as downsampled datasets in the subsequent sections.

As shown in Figure 12, under the same downsampled dataset, the accuracy of all the methods improved compared to their performance on the full dataset. STGCN360 still significantly outperformed all the baselines, achieving the highest accuracy of 0.921, while the accuracy of Transformer, GRU, LSTM, and MLP increased to 0.709, 0.668, 0.642, and 0.650, respectively. These results indicate that reducing the complexity of the input data benefits the performance of all the methods, yet the final performance remains constrained by each model's inherent feature learning capability.

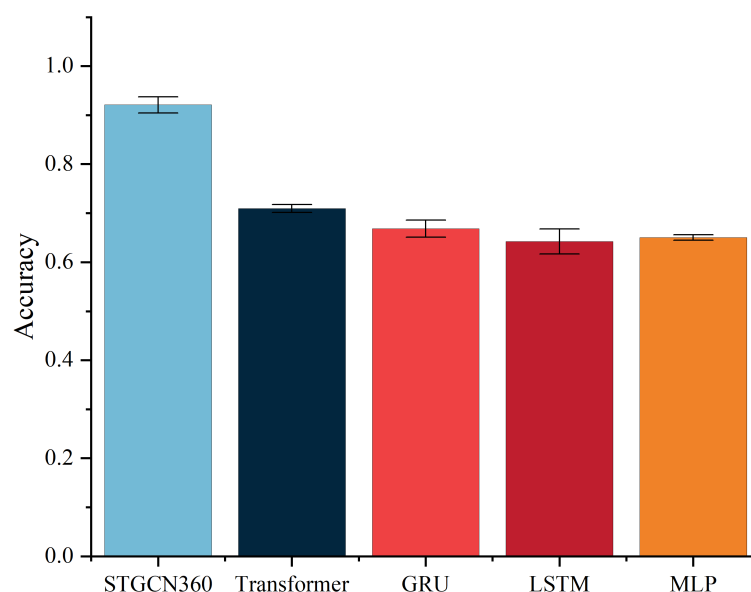


Figure 12. Accuracy of different QoE assessment methods on the downsampled dataset.

Transformer, leveraging its global self-attention mechanism, efficiently captures complex long-range dependencies across the video stream and enables feature interactions through parallel computation. Consequently, it demonstrates relatively superior generalization ability when data complexity is reduced. In contrast, GRU and LSTM, limited by the sequential nature of their recurrent structures and their bottleneck in modeling long-term dependencies, show performance that is more sensitive to changes in data scale. MLP, being entirely incapable of modeling sequential dynamics, performs poorly on both datasets. This proves that, for the QoE assessment task involving multidimensional spatio-temporal features, STGCN360, as a lightweight model with an architecture capable of explicitly and efficiently modeling global dependencies among the spatio-temporal fea-

tures of the viewport, reduces computational complexity during training and possesses a fundamental advantage.

As illustrated in Figure 13, within the downsampled dataset, the training time overhead of each method follows the same overall trend observed in the full dataset, further corroborating the inherent computational efficiency characteristics of the model architectures themselves. Among them, STGCN360 maintains a significant advantage with the lowest training time overhead of 113.2 s. Its efficiency primarily stems from the inductive bias introduced by the explicit spatio-temporal graph structure, which allows the model to avoid implicitly learning complex spatio-temporal relationships from the data, thereby substantially reducing the time cost during training. For clarity of comparison, we have tabulated the accuracy and training time overhead of different methods on both the full and downsampled datasets, as shown in Table 1.

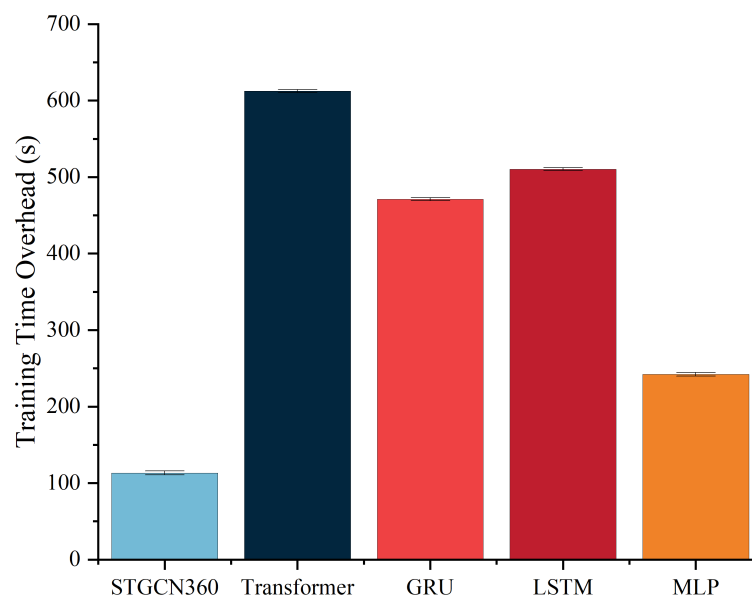


Figure 13. Training time overhead of different QoE assessment methods on the downsampled dataset.

Table 1. Accuracy and training time overhead of different methods on the full and downsampled datasets.

FULL DATASET					
Metrics	STGCN360	Transformer	GRU	LSTM	MLP
Accuracy	0.701	0.531	0.536	0.531	0.533
Training time overhead (s)	320.3	1618	1299.8	1402.4	1185
DOWNSAMPLED DATASET					
Metrics	STGCN360	Transformer	GRU	LSTM	MLP
Accuracy	0.921	0.709	0.668	0.642	0.650
Training time overhead (s)	113.2	612.4	471	510.4	242.2

In contrast, Transformer (612.4 s) still incurs the highest training overhead due to the considerable computational complexity of its self-attention mechanism. GRU (471 s) and LSTM (510.4 s), constrained by sequential computation and their gating mechanisms, demonstrate moderate efficiency. The training time overhead of MLP decreases notably in the downsampled dataset (to 242.2 s), which aligns with our earlier theoretical assumption: when input features are flattened, the model's input dimensionality increases sharply, making it difficult to learn effectively from originally highly complex data. However, once

data complexity is reduced, the patterns the model needs to fit become simpler, leading to a significant acceleration in its training speed.

Finally, the MSE loss learning curves in Figure 14 further reveal the differences in learning behavior between the two representative methods on the downsampled dataset. STGCN360 converges within the first 10 epochs, with its training and test learning curves declining rapidly and synchronously to a low level without showing signs of overfitting. This lightweight model demonstrates excellent learning efficiency and generalization stability even with limited samples. In contrast, Transformer's learning curve converges slowly in the later stages of testing and exhibits significant fluctuations. The gap between its training and test losses gradually widens, and, although it achieves a lower final loss, its robustness is suboptimal, exposing the limitations of generic sequence models when lacking explicit structural guidance. This result also indirectly indicates that the spatio-temporal graph structural prior knowledge leveraged by STGCN360 enables more efficient and stable learning from sparse data, further highlighting its dual advantages in both being lightweight and robust. This intuitively explains why STGCN360 leads in accuracy.

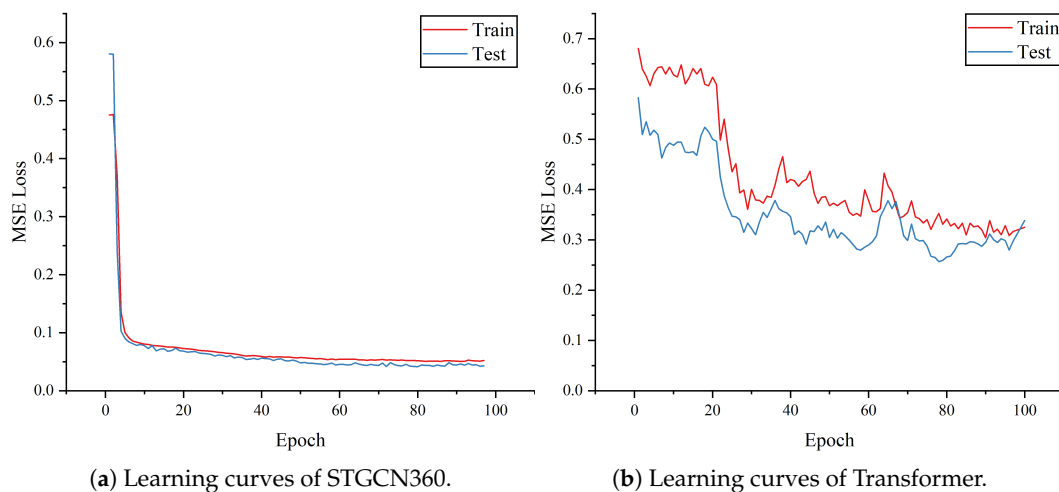


Figure 14. Learning curves of MSE loss for the QoE assessment method on the downsampled dataset.

5. Conclusions

Addressing the challenges of high computational complexity, substantial training time overhead, and the insufficient consideration of multidimensional influencing factors in the existing QoE models for tile-based 360° video streaming, this paper proposes an STGCN-based QoE assessment method named STGCN360. By constructing a spatio-temporal graph model for the viewport, the method effectively reduces extraneous computational costs. It leverages the STGCN to extract diverse spatio-temporal features during video playback, enabling more accurate QoE assessment. Furthermore, we constructed a comprehensive QoE assessment dataset by integrating multi-source heterogeneous datasets, providing robust data support for training STGCN360. The experimental results demonstrate that STGCN360 outperforms the compared methods in terms of both QoE assessment accuracy and training time overhead. Our future work will focus on the following two directions: first, collecting more extensive QoE-influencing factors and subjective rating data in real-world application environments to build a subjective QoE assessment dataset under practical scenarios, thereby providing a more substantial training foundation for STGCN360; second, exploring the integration of additional lightweight large models to expand the model capacity of STGCN360 through multidimensional feature extraction and fusion, further enhancing its QoE assessment accuracy on large-scale data.

Author Contributions: Methodology: S.L.; wrote the main manuscript text: S.L. and Y.L.; reviewed the manuscript: S.L., Y.W., Y.L., J.C., D.W. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangxi Key Research and Development Program (Grant No. AB25069178), the Open Project Program of Guangxi Key Laboratory of Digital Infrastructure (Grant No. GXDIOF2024008), the Project for Enhancing Young and Middle-Aged Teachers' Research Basis Ability in Colleges of Guangxi (No. 2025KY0250), the National Natural Science Foundation of China (Grant No. 62461057), and the National Natural Science Foundation of China (Grant No. 61861013).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He, Y.; Huang, F.; Wang, D.; Chen, B.; Zhang, R. Emergency Communications in Post-Disaster Scenarios: IoT-Enhanced Airship and Buffer Support. *IEEE Internet Things J.* **2025**, *12*, 11457–11468. <https://doi.org/10.1109/JIOT.2024.3520626>.
2. Wang, D.; Wang, Z.; Yang, W.; Zhao, H.; He, Y.; Li, L.; Wei, Z.; Zhou, F. Enhanced ISAC Framework for Moving Target Assisted by Beyond-Diagonal RIS: Accurate Localization and Efficient Communication. *IEEE Trans. Netw. Sci. Eng.* **2025**, *12*, 4299–4315. <https://doi.org/10.1109/TNSE.2025.3571278>.
3. He, Y.; Huang, F.; Wang, D.; Yang, L.; Zhang, R. Delay Minimization for NOMA-MEC Offloading in ABS-Aided Maritime Communication Networks. *IEEE Trans. Veh. Technol.* **2025**, *74*, 9577–9590. <https://doi.org/10.1109/TVT.2025.3539335>.
4. Liang, Y.; Liu, G.; Zhang, X.; Chen, Y.; Chen, J.; Xia, Y.; Liu, Y. Fully Anonymous Broadcast Signcryption for Secure Health Data Transmission in WBANs. *IEEE Trans. Mob. Comput.* **2026**, *25*, 1729–1743. <https://doi.org/10.1109/TMC.2025.3605205>.
5. Setayesh, M.; Wong, V.W.S. Viewport Prediction, Bitrate Selection, and Beamforming Design for THz-Enabled 360° Video Streaming. *IEEE Trans. Wirel. Commun.* **2025**, *24*, 1849–1865. <https://doi.org/10.1109/TWC.2024.3513221>.
6. Rodriguez, A.; Rodriguez, A. In Five Years, VR Could Be as Big in the US as Netflix. 2021. Quartz. Available online: <https://qz.com/1298512/vr-could-be-as-big-in-the-us-as-netflix-in-five-years-study-shows> (accessed on 12 January 2026).
7. CISCO. Cisco Visual Networking Index: Global Mobile Data Traffic FORECAST update. 2019. Available online: <http://media.mediapost.com/uploads/CiscoForecast.pdf> (accessed on 12 January 2026).
8. Chiariotti, F. A survey on 360-degree video: Coding, quality of experience and streaming. *Comput. Commun.* **2021**, *177*, 133–155. <https://doi.org/10.1016/j.comcom.2021.06.029>.
9. Zou, J.; Li, C.; Liu, C.; Yang, Q.; Xiong, H.; Steinbach, E. Probabilistic Tile Visibility-Based Server-Side Rate Adaptation for Adaptive 360-Degree Video Streaming. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 161–176. <https://doi.org/10.1109/JSTSP.2019.2956716>.
10. Li, Z.; Zhong, P.; Huang, J.; Gao, F.; Wang, J. Achieving QoE Fairness in Bitrate Allocation of 360° Video Streaming. *IEEE Trans. Multimed.* **2024**, *26*, 1169–1178. <https://doi.org/10.1109/TMM.2023.3277286>.
11. Ghosh, M.; Singhal, C. A review on machine learning based user-centric multimedia streaming techniques. *Comput. Commun.* **2025**, *231*, 108011. <https://doi.org/https://doi.org/10.1016/j.comcom.2024.108011>.
12. ITU, T.S.S.O. ITU-T P.1203: Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport. 2017. Available online: <https://www.itu.int/rec/T-REC-P.1203> (accessed on 12 January 2026).
13. Kanai, K.; Wei, B.; Cheng, Z.; Takeuchi, M.; Katto, J. Methods for Adaptive Video Streaming and Picture Quality Assessment to Improve QoS/QoE Performances. *IEICE Trans. Commun.* **2019**, *E102.B*, 1240–1247. <https://doi.org/10.1587/transcom.2018ANI0003>.
14. Yue, T.; Wang, H.; Cheng, S.; Shao, J. Deep learning based QoE evaluation for internet video. *Neurocomputing* **2020**, *386*, 179–190. <https://doi.org/https://doi.org/10.1016/j.neucom.2019.12.082>.
15. Zhang, H.; Hu, H.; Gao, G.; Wen, Y.; Guan, K. Deepqoe: A Unified Framework for Learning to Predict Video QoE. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2018; pp. 1–6. <https://doi.org/10.1109/ICME.2018.8486523>.

16. Eswara, N.; Ashique, S.; Panchbhai, A.; Chakraborty, S.; Sethuram, H.P.; Kuchi, K.; Kumar, A.; Channappayya, S.S. Streaming Video QoE Modeling and Prediction: A Long Short-Term Memory Approach. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 661–673. <https://doi.org/10.1109/TCSVT.2019.2895223>.
17. Kougioumtzidis, G.; Poulkov, V.; Zaharis, Z.D.; Lazaridis, P.I. A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction. *IEEE Access* **2022**, *10*, 19507–19538. <https://doi.org/10.1109/ACCESS.2022.3149592>.
18. Jiang, Z.; Xu, Y.; Sun, J.; Hwang, J.N.; Zhang, Y.; Appleby, S.C. Tile-Based Panoramic Video Quality Assessment. *IEEE Trans. Broadcast.* **2022**, *68*, 530–544. <https://doi.org/10.1109/TBC.2021.3136748>.
19. Yang, J.; Liu, T.; Jiang, B.; Lu, W.; Meng, Q. Panoramic Video Quality Assessment Based on Non-Local Spherical CNN. *IEEE Trans. Multimed.* **2021**, *23*, 797–809. <https://doi.org/10.1109/TMM.2020.2990075>.
20. Zhou, K.; Hao, Z.; Wang, L.; Liang, X. Adaptive Score Alignment Learning for Continual Perceptual Quality Assessment of 360-Degree Videos in Virtual Reality. *IEEE Trans. Vis. Comput. Graph.* **2025**, *31*, 2880–2890. <https://doi.org/10.1109/TVCG.2025.3549179>.
21. Gao, P.; Zhang, P.; Smolic, A. Quality Assessment for Omnidirectional Video: A Spatio-Temporal Distortion Modeling Approach. *IEEE Trans. Multimed.* **2022**, *24*, 1–16. <https://doi.org/10.1109/TMM.2020.3044458>.
22. Islam, M.T.; Rothenberg, C.E.; Gomes, P.H. Predicting XR Services QoE with ML: Insights from In-band Encrypted QoS Features in 360-VR. In *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*; IEEE: New York, NY, USA, 2023; pp. 80–88. <https://doi.org/10.1109/NetSoft57336.2023.10175481>.
23. Liu, X.; Yan, J.; Wan, Z.; Fang, Y.; Wang, Z. A Quality-of-Experience Database for Adaptive Omnidirectional Video Streaming. *IEEE J. Sel. Top. Signal Process.* **2023**, *17*, 949–963. <https://doi.org/10.1109/JSTSP.2023.3300529>.
24. Karim, S.; He, H.; Laghari, A.A.; Madiha, H. Quality of Service (QoS): Measurements of Video Streaming. *arXiv* **2020**, arXiv:2008.12017. <https://doi.org/10.48550/arXiv.2008.12017>.
25. Sawabe, A.; Iwai, T. A QoS Model to Identify Required QoS for Guaranteeing Quality of Internet Video Streaming Services. In *ICC 2021—IEEE International Conference on Communications*; IEEE: New York, NY, USA, 2021; pp. 1–6. <https://doi.org/10.1109/ICC42927.2021.9500711>.
26. Duanmu, Z.; Liu, W.; Chen, D.; Li, Z.; Wang, Z.; Wang, Y.; Gao, W. A Bayesian Quality-of-Experience Model for Adaptive Streaming Videos. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *18*, 1–24. <https://doi.org/10.1145/3491432>.
27. Taha, M.; Ali, A.; Lloret, J.; Gondim, P.R.L.; Canovas, A. An automated model for the assessment of QoE of adaptive video streaming over wireless networks. *Multimed. Tools Appl.* **2021**, *80*, 26833–26854. <https://doi.org/10.1007/s11042-021-10934-9>.
28. Eswara, N.; Reddy, D.S.V.; Chakraborty, S.; Sethuram, H.P.; Kuchi, K.; Kumar, A.; Channappayya, S.S. A linear regression framework for assessing time-varying subjective quality in HTTP streaming. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*; IEEE: New York, NY, USA, 2017; pp. 31–35. <https://doi.org/10.1109/GlobalSIP.2017.8308598>.
29. Jana, S.; Chan, A.; Pande, A.; Mohapatra, P. QoE prediction model for mobile video telephony. *Multimed. Tools Appl.* **2016**, *75*, 7957–7980. <https://doi.org/10.1007/s11042-015-2711-5>.
30. Li, B.; Zhang, W.; Tian, M.; Zhai, G.; Wang, X. Blindly Assess Quality of In-the-Wild Videos via Quality-Aware Pre-Training and Motion Perception. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5944–5958. <https://doi.org/10.1109/TCSVT.2022.3164467>.
31. Yang, H.; Lin, T.; Zhang, Y.; Xu, Y.; Chen, Z.; Yan, J. Enhancing QoE for Multi-Device Video Delivery: A Novel Dataset and Model Perspective. *IEEE Trans. Broadcast.* **2025**, *71*, 277–290. <https://doi.org/10.1109/TBC.2024.3443544>.
32. Zhao, T.; Liu, Q.; Chen, C.W. QoE in Video Transmission: A User Experience-Driven Strategy. *IEEE Commun. Surv. Tutorials* **2017**, *19*, 285–302. <https://doi.org/10.1109/COMST.2016.2619982>.
33. Xu, M.; Jiang, L.; Li, C.; Wang, Z.; Tao, X. Viewport-Based CNN: A Multi-Task Approach for Assessing 360° Video Quality. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2198–2215. <https://doi.org/10.1109/TPAMI.2020.3028509>.
34. Reichl, P.; Tuffin, B.; Schatz, R. Logarithmic laws in service quality perception: Where microeconomics meets psychophysics and quality of experience. *Telecommun. Syst.* **2013**, *52*, 587–600. <https://doi.org/10.1007/s11235-011-9503-7>.
35. Barkowsky, M.; Staelens, N.; Janowski, L. Open collaboration on hybrid video quality models - VQEG joint effort group hybrid. In *IEEE 15th International Workshop on Multimedia Signal Processing (MMSp)*; IEEE: New York, NY, USA, 2013; pp. 476–481. <https://doi.org/10.1109/MMSp.2013.6659335>.
36. Zhou, W.; Min, X.; Li, H.; Jiang, Q. A brief survey on adaptive video streaming quality assessment. *J. Vis. Commun. Image Represent.* **2022**, *86*, 103526. <https://doi.org/10.1016/j.jvcir.2022.103526>.
37. Hewage, C.T.E.R.; Ahmad, A.; Mallikarachchi, T.; Barman, N.; Martini, M.G. Measuring, Modeling and Integrating Time-Varying Video Quality in End-to-End Multimedia Service Delivery: A Review and Open Challenges. *IEEE Access* **2022**, *10*, 60267–60293. <https://doi.org/10.1109/ACCESS.2022.3180491>.
38. Shen, W.; Zhou, M.; Liao, X.; Jia, W.; Xiang, T.; Fang, B.; Shang, Z. An End-to-End No-Reference Video Quality Assessment Method With Hierarchical Spatiotemporal Feature Representation. *IEEE Trans. Broadcast.* **2022**, *68*, 651–660. <https://doi.org/10.1109/TBC.2022.3164332>.

39. Wu, W.; Li, Q.; Chen, Z.; Liu, S. Semantic Information Oriented No-Reference Video Quality Assessment. *IEEE Signal Process. Lett.* **2021**, *28*, 204–208. <https://doi.org/10.1109/LSP.2020.3048607>.
40. Zhang, Y.; Gao, X.; He, L.; Lu, W.; He, R. Blind Video Quality Assessment With Weakly Supervised Learning and Resampling Strategy. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2244–2255. <https://doi.org/10.1109/TCSVT.2018.2868063>.
41. Ghosh, M.; Singhal, C. M-3R: A Memory Based Approach for Streaming QoE Prediction under 3R settings. In *2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*; IEEE: New York, NY, USA, 2021; pp. 432–437. <https://doi.org/10.1109/ANTS52808.2021.9936944>.
42. Ghosh, M.; Singhal, D.C.; Wayal, R. DeSVQ: Deep Learning Based Streaming Video QoE Estimation. In *ICDCN'22: Proceedings of the 23rd International Conference on Distributed Computing and Networking*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 19–25. <https://doi.org/10.1145/3491003.3491023>.
43. Zakharchenko, V.; Alshina, E.; Singh, A. AhG8: Suggested testing procedure for 360-degree video. In *Proceedings of the JVET-D0027*; Joint Video Exploration Team of ITU-T: Geneva, Switzerland, 2016; Volume D0027, p. 1–1.
44. Sun, Y.; Lu, A.; Yu, L. Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video. *IEEE Signal Process. Lett.* **2017**, *24*, 1408–1412. <https://doi.org/10.1109/LSP.2017.2720693>.
45. Chen, S.; Zhang, Y.; Li, Y.; Chen, Z.; Wang, Z. Spherical Structural Similarity Index for Objective Omnidirectional Video Quality Assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2018; pp. 1–6. <https://doi.org/10.1109/ICME.2018.8486584>.
46. Schatz, R.; Zabrovskiy, A.; Timmerer, C. Tile-based Streaming of 8K Omnidirectional Video: Subjective and Objective QoE Evaluation. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*; IEEE: New York, NY, USA, 2019; pp. 1–6. <https://doi.org/10.1109/QoMEX.2019.8743230>.
47. Anwar, M.S.; Wang, J.; Ahmad, S.; Khan, W.; Ullah, A.; Shah, M.; Fei, Z. Impact of the Impairment in 360-Degree Videos on Users VR Involvement and Machine Learning-Based QoE Predictions. *IEEE Access* **2020**, *8*, 204585–204596. <https://doi.org/10.1109/ACCESS.2020.3037253>.
48. Fei, Z.; Wang, F.; Wang, J.; Xie, X. QoE Evaluation Methods for 360-Degree VR Video Transmission. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 78–88. <https://doi.org/10.1109/JSTSP.2019.2956631>.
49. Zou, W.; Zhang, W.; Yang, F. Modeling the Perceptual Quality for Viewport-Adaptive Omnidirectional Video Streaming Considering Dynamic Quality Boundary Artifact. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4241–4254. <https://doi.org/10.1109/TCSVT.2021.3050157>.
50. Meng, Y.; Ma, Z. Viewport-Based Omnidirectional Video Quality Assessment: Database, Modeling and Inference. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 120–134. <https://doi.org/10.1109/TCSVT.2021.3057368>.
51. Mahmoud, M.; Rizou, S.; Panayides, A.S.; Lazaridis, P.I.; Karagiannidis, G.K.; Kantartzis, N.V.; Zaharis, Z.D. STAV360: A Dataset for Subjective Tile-based Assessment of 360° Videos. In *12th International Conference on Information Technology (ICIT)*; IEEE: New York, NY, USA, 2025; pp. 67–72. <https://doi.org/10.1109/ICIT64950.2025.11049277>.
52. Roy, A.; Roy, K.K.; Ali, A.A.; Amin, M.A.; Rahman, A.K.M.M. Unified Spatio-Temporal Modeling for Traffic Forecasting using Graph Neural Network. In *2021 International Joint Conference on Neural Networks (IJCNN)*; IEEE: New York, NY, USA, 2021; pp. 1–8. <https://doi.org/10.1109/IJCNN52387.2021.9533319>.
53. Liu, S.; Wang, Y.; Li, S.; Liu, Y. MADRL-based bitrate allocation for QoE fairness in 360° video streaming with viewport prediction. *Multimed. Syst.* **2025**, *31*, 1–24. <https://doi.org/https://doi.org/10.1007/s00530-025-01920-4>.
54. Casas, P.; Wassermann, S. Improving QoE prediction in mobile video through machine learning. In *2017 8th International Conference on the Network of the Future (NOF)*; IEEE: New York, NY, USA, 2017; pp. 1–7. <https://doi.org/10.1109/NOF.2017.8251212>.
55. Wang, H.; Long, Z.; Dong, H.; El Saddik, A. MADRL-Based Rate Adaptation for 360° Video Streaming With Multiviewpoint Prediction. *IEEE Internet Things J.* **2024**, *11*, 26503–26517. <https://doi.org/10.1109/JIOT.2024.3398548>.
56. Liang, Y.; Sun, H.; Zhang, X.; Xie, Q.; Liu, G.; Liu, Z.; Tan, Z.; Liu, Y. Lightweight Multifactor Authentication and Key Agreement Scheme in Vehicle-to-Grid Networks. *IEEE Trans. Veh. Technol.* **2025**, 1–18. <https://doi.org/10.1109/TVT.2025.3621240>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.