Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An NLP-Based Multimodal Framework for Streaming QoE Assessment

Dongyan Zhang

zhangdy@ustb.edu.cn

University of Science and Technology Beijing

Yuzhi Yan

University of Science and Technology Beijing

Han Luo

Beijing PowerInfo CO.

Weihua Li

Beijing PowerInfo CO.

Research Article

Keywords: Video Quality Assessment, Quality of Experience, Adaptive Streaming

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8976064/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An NLP-Based Multimodal Framework for Streaming QoE Assessment

Dongyan Zhang[1*], Yuzhi Yan[2], Han Luo[3], Weihua Li[4]

[1,2*]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China.
[3,4]LTD, Beijing PowerInfo CO., Street, Beijing, 100083, China.

*Corresponding author(s). E-mail(s): zhangdy@ustb.edu.cn;
Contributing authors: m202320980@xs.ustb.edu.cn;

Abstract

With the large-scale deployment of adaptive video streaming services, Quality of Experience (QoE) assessment has become critical for optimizing bitrate adaptation and streaming strategies. Nevertheless, QoE modeling remains challenging due to the complex interplay between video content and Quality of Service (QoS). Most existing streaming QoE models encode QoS information using simple numerical features, which are limited in capturing contextual relationships among QoS attributes. Consequently, it is difficult to model long-term temporal dependencies and semantic correlations across video chunks. To address these limitations, this paper proposes a chunk-aware multimodal framework for QoE assessment in adaptive streaming. The proposed method jointly models network-related QoS dynamics and frame-level perceptual quality variations. Specifically, QoS attributes are converted into structured textual descriptions, which are then encoded by a language model to capture semantic correlations and temporal evolution across video chunks. In parallel, frame-level visual features are extracted within each chunk to retain local perceptual distortions caused by bitrate adaptation and compression. An attention-guided fusion mechanism is adopted to integrate the two modalities and generate the final QoE prediction. Experimental results demonstrate that the proposed framework achieves stable and competitive performance compared with representative state-of-the-art QoE assessment methods.

Keywords: Video Quality Assessment, Quality of Experience,Adaptive Streaming

1 Introduction

Video traffic currently accounts for a dominant portion of global Internet traffic. With audiences demanding higher visual quality and the rapid development of technologies, Quality of Experience (QoE) has become a critical indicator for evaluating the performance of adaptive video streaming systems. However, QoE assessment remains challenging. This is mainly due to the diversity of video content, the dynamic and unpredictable encoding , transmission processes, and the subjectivity of human visual perception. Moreover, service providers are often unable to access real-time streaming content at the client side. This further increases the difficulty of QoE evaluation.

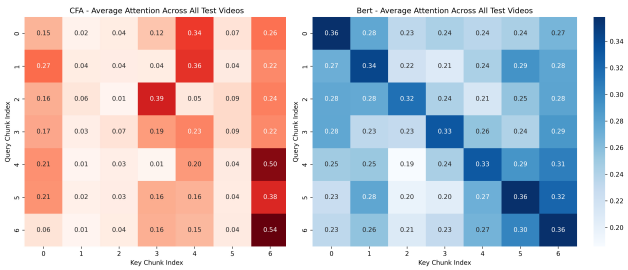
In adaptive streaming scenarios, no-reference QoE assessment methods are closely related to Quality of Service (QoS), as numerous studies have demonstrated a strong causal relationship between network and user-perceived video quality. Previous works have shown that network impairments, bitrate switching events, and variations in spatial or temporal resolution can significantly influence perceived quality and user experience [1, 2]. Although these approaches effectively characterize the impact of network transmission on user experience, they overlook the perceptual effects of human vision across different video contents and viewing conditions.

To address these limitations, several hybrid QoE models have been proposed. Duanmu [3] introduced a unified QoE prediction model that jointly considers video presentation quality, initial

buffering, and stalling events to provide a comprehensive evaluation of user experience. Similarly, Bampis [4] developed Video ATLAS, a learning-based framework that integrates multiple QoE-related features, including video quality metrics, rebuffering-aware indicators, and memory-driven effects, enabling more reliable QoE prediction. While these methods improve overall prediction accuracy, they don't capture fine-grained frame-level variations or model the continuous temporal evolution of QoE. As a result, their ability to represent transient distortions and dynamic playback effects in adaptive streaming is limited.

More recently, continuous-time and segment-level QoE prediction methods have attracted increasing attention. Li [5] proposed a weakly supervised domain adaptation framework that jointly predicts continuous-time and retrospective QoE. Jia [6] developed a unified evaluator that extracts both QoS information and deep video content features, adaptively aggregating them to produce continuous and overall QoE scores. Although these approaches enhance temporal modeling capability, QoS features and visual content are still often processed independently across video segments, which limits the ability to capture both short-term variations and long-range dependencies in streaming scenarios.

To further investigate how existing methods model temporal dependencies in QoS information, we analyze attention mechanisms commonly used in prior work. Specifically, we visualize the attention distributions of the cross-feature attention (CFA) module [6] and a language-based attention model using heatmaps generated from the WaterlooSQoE-IV dataset [7], as illustrated in Fig. 1. The visualization reveals that CFA tends to focus on a limited number of key segments, resulting in sparse and localized attention patterns, whereas the language-based model distributes attention more broadly across segments, enabling the capture of dependencies at multiple temporal scales. This observation highlights the importance of modeling both short-term interactions and long-term temporal relationships in streaming QoE assessment.



(a) CFA Heatmap (b) BERT Heatmap

Fig. 1 Visualization of Attention Distribution

Motivated by these observations,our work makes the following contributions:

- We introduce a semantic-aware representation of QoS features that characterizes network and their temporal evolution, enabling effective modeling of both short-term variations and long-range dependencies across video chunks.
- We develop a chunk-aware multimodal QoE assessment framework that jointly models network-related QoS dynamics and frame-level perceptual quality variations, integrates them through an attention-driven interaction scheme to capture their combined influence on user experience.

2 Related Work

Quality of Experience (QoE) assessment for adaptive streaming has been widely studied in recent years. Existing approaches can be broadly categorized into full-reference (FR), reduced-reference (RR), and no-reference (NR) methods. FR methods rely on access to the original source video to compute quality metrics such as SSIM[8], or VMAF[9]. RR methods require only partial information from the reference video, such as STRRED[10]. In contrast, NR methods aim to predict QoE solely from the streaming video and network statistics available at the client side, without requiring any reference content. Given the unavailability of client-side information in streaming video, this paper focuses on the review and discussion of no-reference methods. Among NR approaches, existing NR methods can be broadly divided into three categories: QoS-based methods, video quality assessment (VQA)-based methods, and hybrid methods.

2.1 QoS-based Methods

QoS-based approaches estimate QoE primarily from network-level and system-level indicators, such as bitrate, rebuffering duration, startup delay, and bitrates switching frequency. Representative works include those proposed by Mok [11], Liu [1], Xue [2], and Yin [12]. These methods rely on handcrafted features and linear or shallow regression models to map QoS indicators to subjective QoE scores. However, by neglecting visual content characteristics, they are unable to capture perceptual quality variations across different videos and viewing conditions, which limits their generalization capability in adaptive streaming scenarios.

2.2 VQA-based Methods

VQA-based methods aim to predict video quality directly from visual signals using deep learning

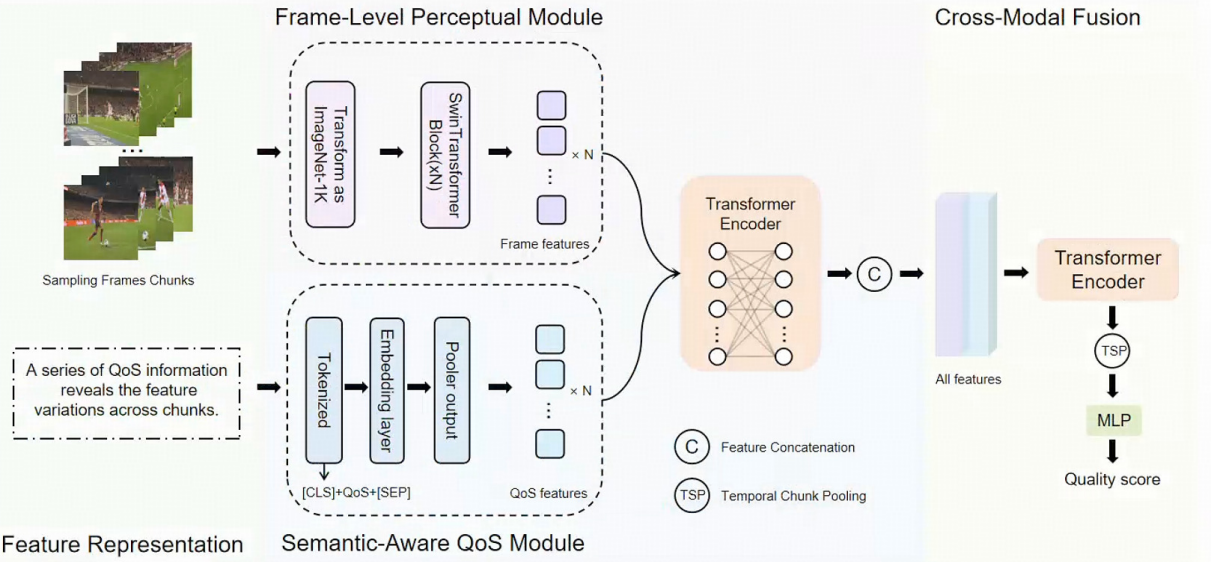


Fig. 2 Overview of the proposed multimodal framework for QoE assessment. QoS features are represented as structured textual descriptions. Frames sampled from each video chunk are processed by a Transformer-based video encoder to model frame-level perceptual features. A cross-attention fusion module aligns chunk-level QoS representations with frame-level visual features for QoE prediction.

techniques. Recent models, such as FastVQA [13], DOVER [14], and MD-VQA [15], extract spatial or spatiotemporal features from video frames to estimate subjective quality. These approaches have demonstrated strong performance on conventional video quality assessment benchmarks. Nevertheless, they are not designed for adaptive streaming and overlook network-induced impairments, such as bitrate adaptation and rebuffering events. As a result, VQA-based methods often struggle to accurately reflect user experience in streaming scenarios where network dynamics play a critical role.

2.3 Hybrid Methods

Hybrid approaches combine QoS information with visual quality features to improve QoE prediction accuracy. Early hybrid models, such as those proposed by Duanmu [3] and Bampis [4], incorporate network-related indicators alongside video quality metrics and memory-related factors, achieving improved performance over single-modality methods. More recent works, including BSQI [16], Jia [17], and other deep learning-based hybrid frameworks, further explore temporal modeling and feature fusion strategies for streaming QoE assessment.

Despite these advances, most existing hybrid methods represent QoS features as numerical vectors or independent time-series signals and process them separately from visual content. This design makes it difficult to characterize QoS attributes, discrete event-driven behaviors (e.g., rebuffering), and long-range temporal dependencies across video segments. Moreover, effectively aligning network dynamics with fine-grained perceptual quality variations remains a challenging problem. These limitations motivate the development of multimodal QoE assessment frameworks that jointly model network-related QoS dynamics and perceptual video quality in a more coherent and integrated manner.

3 Method

This section presents the proposed multimodal framework for streaming QoE assessment. An overview of the proposed framework is illustrated in Fig. 2. We jointly model QoS dynamics and frame-level perceptual information at the chunk level. QoS attributes are first transformed into structured textual descriptions to encode their semantic relationships and temporal dependencies. In parallel, visual content is sampled and analyzed within each video chunk to capture perceptual distortions induced by bitrate adaptation and compression. The two modalities are then integrated through an attention-driven interaction mechanism to produce the final QoE prediction.

3.1 Feature Representation

In adaptive streaming, video playback is organized around chunks (or segments)[18], which serve as the fundamental units for bitrate adaptation and buffering decisions. User perception is therefore closely influenced by chunk-level variations in network conditions and visual quality. To capture such variations, we model each video as a sequence of chunks and extract QoS and visual features at the chunk level.

Table 1 Definition of QoS features adopted for chunk-level QoE modeling in adaptive streaming.

Feature Name	Unit	Description
Rebuffering_duration	s	Duration of playback interruption.
Bitrate	kbps	Target bitrate allocated to the video segment.
QP	–	Quantization parameter controlling the compression strength.
Framerate	fps	Frame display rate during playback.
Width	pixels	Horizontal resolution of the video frame.
Height	pixels	Vertical resolution of the video frame.
Video QoS Representation	–	[CLS] chunk_0: framerate is 25 fps , bitrate is 1000 kbps, qp is 27,rebuffering_duration is 0.333s,width * height is 1080x1920. chunk_1: framerate is 25 fps , bitrate is 300 kbps, qp is 21.3,rebuffering_duration is 1.6s,width * height is 1080x720.[SEP]

- **QoS Features**:We utilize commonly available QoS indicators, including bitrate, rebuffering duration, frame rate, quantization parameter (QP), and spatial resolution. These attributes characterize the streaming, such as transmission capacity, playback interruptions, encoding strength, and spatial–temporal presentation quality. Importantly, these QoS factors are in nature and exhibit non-linear interactions. For example, the perceptual effect of bitrate depends on both resolution and frame rate. This observation motivates a representation beyond simple numerical vectors. The detailed parameters are summarized in Table 1.
- **Video Features**: To capture perceptual quality variations, we sample frames uniformly within each video chunk rather than performing global random sampling. This chunk-aware sampling preserves local visual distortions that arise from adaptive bitrate and avoids diluting transient impairments across the entire video. The sampled frames are resized to 224×224 and fed into a Swin Transformer backbone to extract visual features. The resulting representations encode spatial characteristics at multiple scales and provide a compact description of frame-level perceptual quality for each chunk.

3.2 Semantic-Aware QoS Module

Fig. 3. Illustration of the semantic-aware QoS processing module in Method 3.2. Conventional QoS-based QoE models QoS attributes as independent numerical signals or time-series, which limits their ability to capture attribute interactions and long-range temporal dependencies. In contrast, we observe that QoS factors naturally convey semantic information (e.g., high bitrate and high resolution jointly imply better visual fidelity[19]) and that their impact on user experience accumulates over time due to memory and hysteresis effects.

3.2.1 Structured Textual Representation of QoS

To model these semantic relationships, we convert raw numerical QoS metrics into structured textual descriptions. Each video chunk is described using a fixed linguistic template that encodes both attribute values and their contextual meaning. For example, a chunk can be represented as: "[CLS] chunk 0: framerate is 25 fps, bitrate is 1000 kbps, qp is 27, rebuffering duration is 0.333 s, width × height is 1080 × 1920 [SEP]".This representation transforms QoS attributes into a unified token space, enabling the model to learn attribute interactions through language semantics rather than relying on handcrafted feature combinations.

3.2.2 Intra-Chunk Semantic Modeling

Denoting n as the number of chunks and l as the length of the tokenized sequence, the i-th video's QoS representation can be expressed as:

$$\boldsymbol{T}_i = [s_{i,0}, s_{i,1}, \ldots, s_{i,n}], \quad \boldsymbol{T}_i \in \mathbb{R}^{n \times l} \tag{1}$$

where $s_{i,j}$ denotes the textual description of the j-th chunk. Let d denotes the dimensionality of the embedding produced by the BERT encoder. After tokenization and padding, these sequences are processed by BERT to produce chunk-level embeddings:

$$\boldsymbol{Q}_i = \mathrm{BERT}(\boldsymbol{T}_i), \quad \boldsymbol{Q}_i \in \mathbb{R}^{n \times d} \tag{2}$$

3.2.3 Inter-Chunk Temporal Modeling

While BERT aggregates QoS information within individual chunks, QoE perception also depends on temporal dependencies across chunks. Therefore, to model streaming effects over time, the chunk-level embeddings $\boldsymbol{Q}_i$ are further processed by a Transformer-based QoS encoder.The resulting representations are $\hat{\boldsymbol{Q}}_i$.

$$\hat{\boldsymbol{Q}}_i = \mathrm{QoSEncoder}(\boldsymbol{Q}_i), \quad \hat{\boldsymbol{Q}}_i \in \mathbb{R}^{n \times d} \tag{3}$$

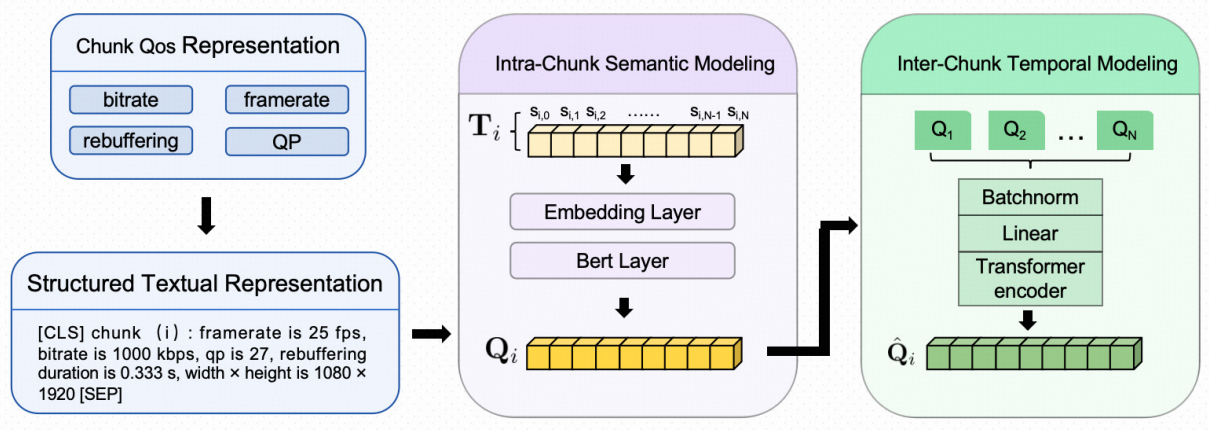


Fig. 3 Overview of the proposed semantic-aware QoS modeling framework for streaming QoE assessment.

3.3 Frame-Level Perceptual Module

We differ from previous VQA methods that perform global random sampling. Instead, we sample frames within each chunk. This strategy preserves fine-grained variations in streaming content. The extracted frame-level features can encode global video characteristics more effectively. We use $\boldsymbol{v_{i,n,m}}$ denotes the m-th frame in the n-th chunk of the i-th video.The number of sampled frames can be expressed as:

$$\boldsymbol{F}_i = [v_{i,0,0}, v_{i,0,1} \dots v_{i,n,m}] \tag{4}$$

where $F_i \in \mathbb{R}^{n\times m\times c\times h\times w}$ denotes the sampled frame features of the i-th video, c representing the number of channels, h and w representing the height and width of each frame, respectively. After simple sampling, the temporal information of the original video is partially lost. As a result, it becomes difficult to capture accurate contextual dependencies. Therefore, a frame-level Encoder (VideoEncoder($\cdot$)) is introduced to encode the sampled features. Using a self-attention mechanism, it captures the spatiotemporal relationships among the sampled frames, which can be expressed as follows:

$$\hat{\boldsymbol{F}}_i = \text{VideoEncoder}(\text{Swin}(\boldsymbol{F}_i)) \tag{5}$$

where $\hat{\boldsymbol{F}}_i \in \mathbb{R}^{n\times b}$ denotes the frame-level features containing spatiotemporal information, and b denotes the dimensionality of each feature vector.

3.4 Cross-Modal Fusion

Based on the extracted features, we obtain the QoS-related features $\hat{\boldsymbol{Q}}_i$ and the video-related features $\hat{\boldsymbol{F}}_{\boldsymbol{i}}$, each independently encoded to capture contextual relationships. This module aims to fuse the two features and model the correlation between semantics and video quality. For instance, high frame rates and resolutions correspond to clear and smooth videos. This can be formally expressed as follows:

$$\boldsymbol{H}_i = \text{CrossEncoder}\big([\hat{\boldsymbol{Q}}_i; \hat{\boldsymbol{F}}_{\boldsymbol{i}}]\big) \tag{6}$$

where $[\hat{\boldsymbol{Q}}_i; \hat{\boldsymbol{F}}_{\boldsymbol{i}}] \in \mathbb{R}^{n\times(b+l)}$ denotes the concatenation of the encoded QoS and video features for the i-th video, and $\boldsymbol{H_i}$ represents the fused multimodal representation capturing cross-modal dependencies.Finally, The fused features are then fed into a multi-layer perceptron (MLP) for regression, producing the predicted MOS value $\hat{\boldsymbol{y}}_i$ as:

$$\hat{\boldsymbol{y}}_i = \text{MLP}(\boldsymbol{H}_i) \tag{7}$$

3.5 Loss function

To simultaneously optimize numerical accuracy and ranking consistency, we propose a composite loss combining Smooth L1 regression with a linear correlation-inspired regularization. Let $\mathbf{P} \in \mathbb{R}^N$ denotes the predicted scores and $\mathbf{G} \in \mathbb{R}^N$ denotes the ground-truth MOS for N samples. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}}(\mathbf{P}, \mathbf{G}) + \beta\, \mathcal{L}_{\text{plcc}}(\mathbf{P}, \mathbf{G}), \tag{8}$$

where β balances regression accuracy and linear consistency. The regression term uses Smooth L1 loss to reduce the impact of outliers while maintaining stable convergence:

$$\mathcal{L}_{\text{reg}}(\mathbf{P}, \mathbf{G}) = \frac{1}{N}\sum_{i=1}^{N} \text{SmoothL1}(P_i, G_i). \tag{9}$$

The correlation-inspired term encourages the predictions to preserve the relative ordering of MOS values by incorporating a global correlation

factor ρ:

$$\rho = \frac{1}{N}\sum_{i=1}^{N} P_i G_i. \tag{10}$$

$$\mathcal{L}_{\text{plcc}} = \frac{1}{8N}\Big(\sum_{i=1}^{N}(P_i - G_i)^2 + \sum_{i=1}^{N}(\rho P_i - G_i)^2\Big) \tag{11}$$

After multiple ablation studies and batch training, we set β to 0.5.

4 Exeperiments

This section first introduces the datasets used for evaluation and the adopted performance metrics, followed by the experimental setup and result analysis.

4.1 Evaluation Metrics and Dataset Information

The proposed method is evaluated on three publicly adaptive streaming QoE datasets: LIVE-NFLX-II[23], WaterlooSQoE-III[24] and WaterlooSQoE-IV[7]. These datasets include user subjective Mean Opinion Scores (MOS), encoder information, as well as details on video streaming stalling and buffering.The specific information of the datasets used in our experiment is summarized in Table 2 .

We adopt the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Order Correlation Coefficient (SRCC) to evaluate linear consistency and rank-order agreement between predicted scores and subjective MOS values.

4.2 Experiment Setup

We split the datasets using an 8:2 ratio, as the original datasets do not provide predefined training and testing sets. To prevent potential data leakage, videos derived from the same source content are strictly assigned to either the training or testing set, but not both.After dataset splitting, MOS values are normalized using statistics computed from the training set only. Specifically, the normalized MOS value for the i-th sample is computed as follows:

$$y_i^{\text{norm}} = \frac{y_i - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{12}$$

where y_i denotes the original MOS score, y_i^{norm} represents the normalized score, and μ_{train} and σ_{train} are the mean and standard deviation of the MOS values in the training set. This normalization ensures that both the training and testing sets follow the same scale, which helps stabilize the regression training.

Swin Transformer is adopted as the visual backbone, with four frames uniformly sampled from each video chunk for frame-level feature extraction. During training, we pad the chunks to ensure consistent batch processing since chunk sizes vary across datasets. A mask is applied to ignore the padded parts and prevent them from introducing noise during training.

4.3 Experiment Results and Discussion

We compare our method against several state-of-the-art QoE prediction models, including QoS-based, VQA-based, and hybrid methods.Table 3 presents the detailed comparison between our method and the existing models.

The results indicate that models based solely on video quality, such as Fast-VQA[13] and SimpleVQA [20], perform poorly on streaming datasets. This can be attributed to the relatively limited diversity of source content in streaming datasets. Traditional QoS-based models, such as Liu2012[1] and Xue2014[2], also exhibit limited performance, as they primarily depend on bitrate, rebuffering, and quality-switching indicators while giving less consideration to perceptual processes of the Human Visual System (HVS). As a result, their predictions may deviate from actual user perception. Hybrid models, such as BSQI [16] and DSA-QoE [6], achieve improved performance by integrating both QoS and visual quality features. However, their feature fusion strategies may fail to fully capture all QoS features.

In contrast, our proposed model demonstrates superior performance across all three datasets. On LIVE-NFLX-II and Waterloo-IV, it achieves the highest SRCC and PLCC. On Waterloo-III, it remains highly competitive, achieving results comparable to or better than the best existing models. These gains mainly stem from the proposed NLP-based semantic modeling of QoS features, which enables tighter coupling among attributes and more effective integration of local and long-range temporal context.

Figure 4 illustrates the scatter plots of predicted MOS versus ground-truth MOS on three datasets. Each point represents a test video segment, and the red dashed line denotes the ideal prediction, where the estimated score exactly matches the subjective MOS. From these visualizations, a clear linear trend can be observed across all datasets, indicating strong consistency between the predicted scores and human subjective judgments.

4.4 Ablation Study

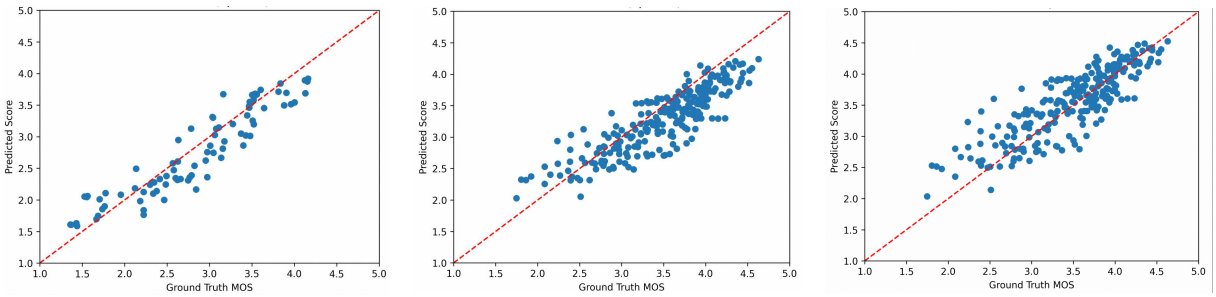
To evaluate the effectiveness of the proposed composite loss function, we conducted an ablation

Table 2 Summary of Streaming QoE Datasets Used in Experiments

Dataset	Content	Videos	ABR Algorithms	MOS Range
LIVE-NFLX-II	15	420	4	$[-2, 2]$
WaterlooSQoE-III	20	450	6	$[0, 100]$
WaterlooSQoE-IV	5	1350	5	$[0, 100]$

Table 3 Performance Comparison Across Different Streaming QoE Perdition Model. **Bold** values indicates the best performance, and underlined values indicate the second-best.

QoE Model	Method Category	LIVE-NFLX-II		Waterloo-III		Waterloo-IV	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Liu2012 [1]	QoS	0.6962	0.6899	0.5141	0.4793	0.4593	0.2732
Xue2014 [2]	QoS	0.5134	0.5439	0.5764	0.5388	0.4874	0.4641
SimpleVQA [20]	VQA	0.6193	0.6247	0.5783	0.5934	0.5834	0.6047
FAST-VQA [13]	VQA	0.6030	0.5936	0.7018	0.7237	0.6718	0.7014
BSQI [16]	Hybrid	0.8930	0.9050	0.7760	0.7940	0.6990	0.7200
ASPECT [21]	Hybrid	0.7739	0.7898	0.8630	0.8578	0.7839	0.7947
Jia2024 [17]	Hybrid	0.9220	0.9150	0.8700	0.8510	0.7950	0.8010
DSA-QoE [6]	Hybrid	0.9250	0.9340	0.8130	0.7968	0.7680	0.7520
Huang2025 [22]	Hybrid	0.8700	0.8760	**0.8780**	0.8930	0.8010	0.8160
Ours	Hybrid	**0.9274**	**0.9385**	0.8692	**0.8994**	**0.8172**	**0.8265**



(a) LIVE-NFLX-II (b) Waterloo-III (c) Waterloo-IV

Fig. 4 Scatter plots of predicted MOS versus ground-truth MOS on three representative streaming QoE datasets.

study comparing traditional loss functions (L1 and MSE) with our method under different values of the weighting parameter β.

Table 4 summarizes the PLCC and SRCC on the WaterlooSQoE-III dataset.The results demonstrate that the proposed composite loss consistently outperforms standard regression losses. Incorporating the correlation-inspired term notably improves ranking consistency. Different values of the weighting parameter β show that a balanced setting achieves the best trade-off between regression accuracy and rank correlation. Excessive emphasis on either term can degrade the other, highlighting the importance of jointly optimizing numerical and ordinal aspects of QoE prediction.

To evaluate the effectiveness of the proposed framework, an ablation experiment is conducted on the WaterlooSQoE-III dataset. The experiment is designed to assess the impact of different module combinations within the model, and four experimental structures are considered as follows:

Table 4 Performance of Different Loss Functions on the WaterlooSQoE-III Database

Loss Type	β	PLCC	SRCC
L1	–	0.8792	0.8512
MSE	–	0.8623	0.8483
Ours	0.1	0.8824	0.8554
Ours	0.3	0.8902	0.8623
Ours	0.5	0.8994	0.8692
Ours	0.7	0.8932	0.8701
Ours	0.9	0.8838	0.8628

1. Training the model using only the Frame-Level Perceptual Module (FLPM).
2. Training the model using the Frame-Level Perceptual Module without the encoder component (FLPM-Encoder).
3. Training the model using only the Structured Textual Representation of QoS module (STROQ).
4. Training the model using the Structured Textual Representation of QoS module with ICTM (STROQ-ICTM).

5. Training the model with the joint use of the FLPM and STROQ modules.
6. Training the model with all modules of the proposed framework.

Table 5 presents the outcomes of our ablation study. The study considers four main components: FLPM, FLPM-Encoder, STROQ, STROQ-ICTM. The results indicate that incorporating each individual module leads to performance improvements over the baseline. Specifically, the addition of the NLP-based QoS module significantly enhances prediction accuracy, demonstrating the importance of semantic modeling of network conditions. Combining multiple modules consistently yields further gains, with the full model achieving the highest SRCC and PLCC. These findings suggest that each module provides complementary information and that their integration is essential for capturing both visual and temporal QoS dynamics in streaming , resulting in more accurate QoE predictions.

Table 5 Ablation Analysis of Module Contributions on the WaterlooSQoE-III Dataset

FLPM	FLPM Encoder	STROQ	STROQ ICTM	SRCC	PLCC
✓	–	–	–	0.7625	0.7578
–	✓	–	–	0.8043	0.7945
–	–	✓	–	0.7828	0.7994
–	–	✓	✓	0.8094	0.8182
✓	–	✓	–	0.8745	0.8526
✓	✓	✓	✓	0.8994	0.8692

5 Conclusions

This paper proposed a multimodal framework for streaming QoE assessment that integrates frame-level visual features with semantic-aware QoS representations. An NLP-based module is employed to capture both short-term variations and long-term temporal dependencies in QoS signals, enabling effective modeling of the semantic relationships between network conditions and perceptual quality. Experiments conducted on multiple public streaming QoE datasets demonstrate that the proposed method consistently achieves strong performance compared with state-of-the-art QoE prediction models. Ablation studies further verify that each module contributes complementary information, and that the jointly modeling of visual content and semantic QoS features is critical for accurate QoE prediction. Future work will focus on improving computational efficiency and enabling real-time QoE assessment in practical adaptive streaming systems.

Declarations

- Funding. No funding was received for conducting this study
- Ethics approval and consent to participate. Not applicable.
- Consent for publication. Not applicable.
- Data availability. All datasets used in this study are publicly available.
- Materials availability. Not applicable.
- Code availability. The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.
- Competing interests . The authors declare that they have no conflict of interest.
- Author contribution Conceptualization : Yuzhi Yan, Dongyan Zhang, Han Luo, Weihua Li. Investigation and data curation: Yuzhi Yan. Writing—original draft: Yuzhi Yan. Writing—review and editing: Dongyan Zhang.

References

[1] Liu, X., Dobrian, F., Milner, H., Jiang, J., Sekar, V., Stoica, I., Zhang, H.: A case for a coordinated internet video control plane. SIGCOMM Comput. Commun. Rev. **42**(4), 359–370 (2012) https://doi.org/10.1145/2377677.2377752

[2] Xue, J., Zhang, D.-Q., Yu, H., Chen, C.W.: Assessing quality of experience for adaptive http video streaming. In: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6 (2014). https://doi.org/10.1109/ICMEW.2014.6890604

[3] Duanmu, Z., Zeng, K., Ma, K., Rehman, A., Wang, Z.: A quality-of-experience index for streaming video. IEEE Journal of Selected Topics in Signal Processing **11**(1), 154–166 (2017) https://doi.org/10.1109/JSTSP.2016.2608329

[4] Ghadiyaram, D., Pan, J., Bovik, A.C.: Learning a continuous-time streaming video qoe model. IEEE Transactions on Image Processing **27**(5), 2257–2271 (2018) https://doi.org/10.1109/TIP.2018.2790347

[5] Li, L., Chen, P., Lin, W., Xu, M., Shi, G.: From whole video to frames: Weakly-supervised domain adaptive continuous-time qoe evaluation. IEEE Transactions on Image

Processing **31**, 4937–4951 (2022) https://doi.org/10.1109/TIP.2022.3190711

[6] Jia, Z., Min, X., Zhai, G.: Dsa-qoe: Quality of experience evaluation for streaming video based on dual-stage attention. In: 2024 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5 (2024). https://doi.org/10.1109/ISCAS58744.2024.10558538

[7] Duanmu, Z., Liu, W., Li, Z., Chen, D., Wang, Z., Wang, Y., Gao, W.: The Waterloo Streaming Quality-of-Experience Database-IV. https://doi.org/10.21227/j15a-8r35 . https://dx.doi.org/10.21227/j15a-8r35

[8] Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2, pp. 1398–14022 (2003). https://doi.org/10.1109/ACSSC.2003.1292216

[9] Li, Z., Aaron, A., Manohara, M.: Toward a practical perceptual video quality metric (2016)

[10] Soundararajan, R., Bovik, A.C.: Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Transactions on Circuits and Systems for Video Technology **23**(4), 684–694 (2013) https://doi.org/10.1109/TCSVT.2012.2214933

[11] Mok, R.K.P., Chan, E.W.W., Chang, R.K.C.: Measuring the quality of experience of http video streaming. In: 12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops, pp. 485–492 (2011). https://doi.org/10.1109/INM.2011.5990550

[12] Yin, X., Jindal, A., Sekar, V., Sinopoli, B.: A control-theoretic approach for dynamic adaptive video streaming over http. SIGCOMM Comput. Commun. Rev. **45**(4), 325–338 (2015) https://doi.org/10.1145/2829988.2787486

[13] Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., Lin, W.: Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, pp. 538–554. Springer, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20068-7_31 . https://doi.org/10.1007/978-3-031-20068-7_31

[14] Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20144–20154 (2023)

[15] Zhang, Z., Wu, W., Sun, W., Tu, D., Lu, W., Min, X., Chen, Y., Zhai, G.: Md-vqa: Multi-dimensional quality assessment for ugc live videos. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1746–1755 (2023). https://doi.org/10.1109/CVPR52729.2023.00174

[16] 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. Multimedia Comput. Commun. Appl. **18**(3s) (2023) https://doi.org/10.1145/3491432

[17] Jia, Z., Min, X., Sun, W., Zhai, G.: Continuous and overall quality of experience evaluation for streaming video based on rich features exploration and dual-stage attention. IEEE Transactions on Circuits and Systems for Video Technology **34**(11), 11709–11723 (2024) https://doi.org/10.1109/TCSVT.2024.3418941

[18] Timmerer, C., Griwodz, C.: Dynamic adaptive streaming over http: from content creation to consumption. In: Proceedings of the 20th ACM International Conference on Multimedia. MM '12, pp. 1533–1534. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2393347.2396553 . https://doi.org/10.1145/2393347.2396553

[19] Sullivan, G.J., Ohm, J.-R., Han, W.-J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on Circuits and Systems for Video Technology **22**(12), 1649–1668 (2012) https://doi.org/10.1109/TCSVT.2012.2221191

[20] Sun, W., Min, X., Lu, W., Zhai, G.: A deep learning based no-reference quality assessment model for ugc videos. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, pp. 856–865. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548329 . https://doi.org/10.1145/3503161.3548329

[21] Li, C., Lim, M., Bentaleb, A., Zimmermann,

R.: A real-time blind quality-of-experience assessment metric for http adaptive streaming. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 1661–1666 (2023). https://doi.org/10.1109/ICME55011.2023.00286

[22] Huang, J., Kong, G., Duan, X., Long, H.: Adaptive video streaming quality of experience assessment based on multi-layer feature perception. Signal, Image and Video Processing **19**(4), 335 (2025) https://doi.org/10.1007/s11760-025-03906-1

[23] Bampis, C.G., Li, Z., Katsavounidis, I., Huang, T.-Y., Ekanadham, C., Bovik, A.C.: Towards perceptually optimized adaptive video streaming-a realistic quality of experience database. IEEE Transactions on Image Processing **30**, 5182–5197 (2021) https://doi.org/10.1109/TIP.2021.3073294

[24] Duanmu, Z., Rehman, A., Wang, Z.: A quality-of-experience database for adaptive video streaming. IEEE Transactions on Broadcasting **64**(2), 474–487 (2018) https://doi.org/10.1109/TBC.2018.2822870