BP-LLM: A Large Language Model-based Approach for Accurate and Adaptable Bandwidth Prediction

Zheng jiang[*] **Iat Long long**[*] **Xin Chen**[*]
Department of Computer Science and Technology, Tsinghua University
{jz24,rongyl24,chenxin24}@mails.tsinghua.edu.cn

Abstract

In the rapidly evolving streaming media landscape, accurate bandwidth prediction is crucial for optimizing user experience and resource utilization. This work introduces BP-LLM, a novel approach that leverages the capabilities of large language models (LLMs) to enhance bandwidth prediction. Traditional algorithms face significant limitations due to their reliance on historical data, inability to incorporate multimodal inputs, and challenges in generalization across diverse network conditions. BP-LLM addresses these challenges by employing the Transformer architecture to capture long-term dependencies in network traffic and integrating various input modalities—such as user location and communication latency—through text representations. Our method not only improves prediction accuracy but also demonstrates superior adaptability to new tasks and environments. Key contributions include the establishment of a comprehensive benchmark for evaluating bandwidth prediction algorithms across different application scenarios, the innovative application of LLMs to enrich feature representation and align textual and temporal data, and the demonstration of robust performance across various error metrics. The results indicate that BP-LLM outperforms state-of-the-art algorithms, providing reliable guidance for downstream tasks such as resource allocation and quality of service management. This advancement paves the way for more efficient network management strategies, enhancing the competitiveness of streaming media applications.

1 Introduction

In the streaming media industry, bandwidth prediction is essential for ensuring user experience and optimizing resource use. In low-latency live streaming, it estimates user network conditions in real-time, adjusts transmission strategies, and reduces stuttering and latency. For long videos, it helps adaptive algorithms select bitrates intelligently, adapting to network changes and balancing picture quality, smoothness, and buffering. In short videos, it determines video bitrate combinations for seamless and high-definition (HD) playback, enhancing user retention. Bandwidth prediction also affects the cost and efficiency of CDN distribution by optimizing transcoding bitrates and scheduling. Accurate bandwidth prediction drives decision-making algorithms, optimizes user experience and technical architecture, and is crucial for competitiveness in the streaming media industry.

Recently, the field of bandwidth prediction has witnessed a surge in innovative approaches, as researchers and practitioners have sought to harness the power of statistical methods, machine learning techniques, and time-series analysis to model and forecast network traffic patterns. These traditional algorithms, designed to predict future bandwidth usage, have predominantly relied on historical data, employing a range of sophisticated techniques to extract meaningful insights. Among these techniques, autoregressive models have been widely used for their ability to capture temporal

[*]These authors contributed equally.

Submitted to Tsinghua University Course: Advanced Machine Learning (AML 2024). Do not distribute.

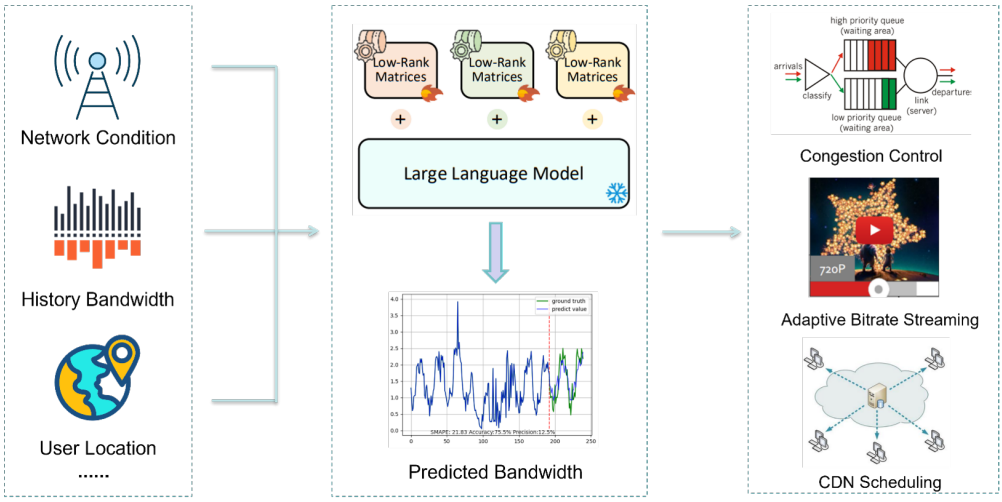

Figure 1: Background Elaboration: To further enhance the accuracy and adaptability of bandwidth prediction, more advanced and flexible prediction models leveraging large language models (LLMs) should be developed, ultimately better catering to the evolving demands of network applications.

dependencies in the data, while moving averages have proven effective in smoothing out short-term fluctuations to reveal underlying trends. More recently, supervised learning methods, such as support vector machines and neural networks, have gained prominence due to their capacity to learn complex patterns and make accurate predictions.

However, despite these advances, the complexity and dynamics of network conditions pose significant challenges to traditional bandwidth prediction algorithms, primarily in the following aspects:

1. **Long-Term Trend Analysis.** Due to the volatility and uncertainty of network traffic, traditional network architectures often fail to capture the long-term trends of bandwidth changes, resulting in limited accuracy of prediction results. This limitation is particularly pronounced during peak network traffic periods, making it difficult for prediction models to provide stable and reliable bandwidth predictions.
2. **Multi-Modal Input Handling.** Existing algorithms typically model bandwidth prediction as a single time-series task, neglecting domain-specific knowledge and failing to account for other modalities that influence bandwidth, such as network conditions, user location, and communication latency. This single-perspective modeling approach overlooks the complexity and variability of network environments, leading to prediction models that cannot comprehensively consider various factors affecting bandwidth. For instance, changes in user location may impact network signal strength and stability, while communication latency may be influenced by network congestion and routing selection. Consequently, prediction models relying solely on time-series data struggle to accurately reflect the impact of these factors on bandwidth.
3. **Generalization and Adaptability.** Bandwidth prediction plays a crucial guiding role in downstream tasks, but traditional algorithms' simplistic modeling limits their generalizability, making it difficult to adapt quickly to new tasks and scenarios. In practical applications, bandwidth prediction results are often used for optimizing network resource allocation and adaptive video streaming. However, due to the limited generalizability of traditional algorithms, they often struggle to maintain consistent prediction performance across different network environments and application scenarios. This restricts the value and effectiveness of bandwidth prediction in practical applications, particularly when facing emerging network technologies and application scenarios, where the limitations of traditional algorithms are more pronounced.

In summary, to overcome these challenges, we need to develop more advanced and flexible prediction models to improve the accuracy and adaptability of bandwidth prediction, thereby better meeting the demands of network applications.

In this work, we propose **BP-LLM**, a strategy leveraging large language models (LLMs) for feature extraction and long-term dependency modeling, significantly improving bandwidth prediction and related tasks (as shown in Figure 1). First, by incorporating the Transformer architecture of LLMs,

our model processes longer sequences and captures dependencies through self-attention, crucial for predicting volatile network traffic trends. Second, we use LLMs' pre-trained network knowledge to align input data (like network conditions, user location, and latency) to text, enhancing prediction accuracy. This cross-modal fusion harnesses LLMs' natural language processing capabilities to integrate multimodal inputs affecting bandwidth. Finally, pre-trained on extensive language and network knowledge, our model adapts and generalizes well to new tasks and scenarios. Fine-tuning on bandwidth prediction optimizes performance, ensuring consistent accuracy across various network environments and applications.

Our main contributions can be summarized as follows:

- We have developed a comprehensive benchmark for evaluating bandwidth prediction algorithms by employing a tail importance sampling method to collect bandwidth data of varying granularity from three distinct application scenarios: Video-on-Demand (VoD), Live Streaming, and Real-Time Communication (RTC). This unified benchmark not only enriches the diversity of the dataset but also ensures that the evaluation is conducted under a wide range of network conditions, thereby providing a more robust and realistic assessment of algorithm performance.
- We are the first to introduce Large Language Models (LLMs) into the domain of bandwidth prediction. **BP-LLM** not only enriches the feature representation of the input data but also facilitates the alignment between textual and temporal features, pointing towards the potential of multimodal foundation models that excel in both language understanding and network conditions analysis.
- Our proposed method is not only capable of handling data from different time windows but also provides stable and efficient prediction performance across various error measures. This versatility is achieved through the effective utilization of LLMs, which are adept at processing and understanding complex patterns in data. As a result, our method outperforms state-of-the-art algorithms in terms of comprehensive metrics on the bandwidth prediction task, demonstrating its superiority in both accuracy and reliability.
- By providing more reliable and effective guidance for downstream tasks, our approach drives the development and optimization of network applications. This is particularly significant in scenarios where accurate bandwidth prediction is crucial for resource allocation, congestion control, and quality of service (QoS) management. The success of our method in downstream tasks opens up new possibilities for the enhancement of existing algorithms, paving the way for more sophisticated and efficient network management strategies.

2 Related Work

Bandwidth Prediction. Network transmission time series prediction is essential for forecasting future network traffic and latency using historical data, aiming to optimize resource management. Recent advancements include linear model-based methods like TiDE, N-Hits, and Dlinear [5, 8, 29], which use various forms of linear regression. TiDE excels with simple datasets [8], N-Hits improves accuracy with cyclic data [5], and Dlinear adapts to rapid changes [29]. Transformer-based approaches, leveraging deep learning, include PatchTST, FEDformer, Pyraformer, Autoformer, and Informer [18, 20, 25, 32, 33]. These methods use intricate mechanisms like segmented data patches [20], combined attention mechanisms [33], and sparse attention for efficient long-term predictions and handling complex data patterns [25, 32]. Linear models are simple and computationally efficient, while Transformer-based methods excel with complex time series data. Future research may explore hybrid models that combine the strengths of both approaches for better accuracy and efficiency.

LLMs for Time Series Forecasting. Large language models (LLMs) like GPT-4 and Llama-3 [1, 21] have achieved significant success in natural language processing (NLP) and computer vision (CV) due to their powerful sequence modeling abilities [4]. Researchers are now exploring their application in time series forecasting for tasks such as bandwidth prediction [17, 30]. Time series analysis is vital in many domains, including climate modeling, bandwidth prediction, and financial analysis [10, 31]. While classical methods like frequency analysis have been used, deep learning techniques such as CNNs [27], LSTMs [13], and transformers [23] have proven highly effective.

However, LLMs are initially trained on discrete text data, which presents challenges for aligning different modalities in time series forecasting tasks. Addressing these challenges is critical to

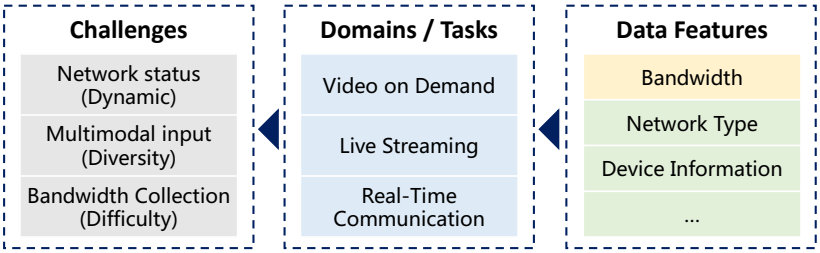

Figure 2: Benchmark Construction.

enhancing model performance in this context [6, 12]. Existing methods for cross-modality alignment include prompting [10, 26], quantization [3, 9], and aligning [11, 15, 34]. Through these methods, LLMs can significantly improve their ability to handle multi-modal data [22], thereby enhancing their performance in time series forecasting tasks.

3 Benchmark Construction

The construction of a suitable benchmark is indeed a critical prerequisite for training a bandwidth prediction algorithm that can effectively handle the complexities of real-world network conditions. Given the intricate nature of network dynamics and the diversity of downstream tasks, there currently exists no unified benchmark that comprehensively captures the full spectrum of online distributions. To address this gap and better fit the online data distribution, we have adopted a strategy based on tail importance sampling to prioritize data collection from scenarios with lower bandwidth, which are often more challenging and critical for accurate prediction.

Our approach involves categorizing data based on tail importance sampling, focusing on low-bandwidth scenarios that are typically underrepresented in conventional datasets. To cater to a wide range of business scenarios, we have collected data from three distinct application domains: Video-on-Demand (VoD), Live Streaming, and Real-Time Communication (RTC). These domains present unique challenges and requirements, necessitating a comprehensive dataset that can capture the nuances of each scenario. By doing so, we aim to ensure that the offline dataset mirrors the real-world online environments more accurately, especially in conditions that are prone to instability and variability.

As shown in Figure 2, we have successfully constructed a dataset that encompasses over 5 million records, meticulously gathered to cover a broad spectrum of business scenarios. This dataset not only includes bandwidth data but also encompasses a rich set of parameters such as network type, device information, and other relevant metadata. By including such a wide array of input features, we aim to provide a standardized input framework that can be readily applied across different applications and network environments.

4 Methodology

4.1 System Overview

Our model architecture, illustrated in Figure 3, is designed to address the complexities of bandwidth prediction by integrating historical bandwidth data and network conditions. Given a sequence of historical bandwidth $X_B \in \mathbb{R}^{N \times T}$, which consists of N different bandwidth data points across T time steps, along with the corresponding network conditions X_N presented in text modality, including details such as network type, device information, and app platform, our objective is to train a large language model $f(\cdot)$ to understand the input series and accurately forecast the readings at H future time steps. The forecast is denoted by $\hat{Y} \in \mathbb{R}^{N \times H}$, where each element $\hat{Y}_h$ represents the predicted values at the h-th future time step for all N variables. The overall goal is to minimize the errors between the ground truth values Y and the predictions $\hat{Y}$. This objective can be quantified by the following loss function:

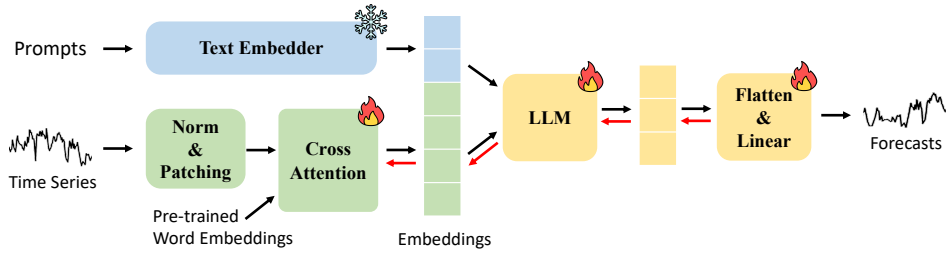

Figure 3: The System Architecture of BP-LLM. Prompts are encoded with a text embedder, and time series data undergoes normalization and cross-attention. The combined embeddings are processed by a large language model (LLM) to produce forecasts.

$$\mathcal{L} = \frac{1}{H} \sum_{h=1}^{H} \left\| \hat{Y}_h - Y_h \right\|_F^2$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm, which measures the Euclidean norm of the matrix after flattening it into a vector. The Frobenius norm is employed to calculate the distance between the predicted matrix $\hat{Y}_h$ and the ground truth matrix Y_h at each future time step h. By minimizing this loss function, our model is trained to learn the underlying patterns and dependencies in the historical data and network conditions, enabling it to make accurate predictions for future bandwidth readings.

Indeed, our method is meticulously designed with three core components: (1) Network Context-Informed Embedding Guidance, (2) Alignment between Text and Temporal Modalities, (3) Parameter-Efficient Fine-Tuning of Large Language Models, each addressing specific challenges and leveraging unique strengths to enhance the accuracy and adaptability of bandwidth prediction. This framework is a significant departure from traditional approaches, integrating advanced techniques to address the complexities of network dynamics and multi-modal data.

We highlight that by integrating these three components, our model can seamlessly integrate network context, handle multi-modal data, and optimize the performance of a large language model for bandwidth prediction. This framework represents a significant advancement in the field, offering a robust solution to the challenges of bandwidth prediction in complex and dynamic network environments. Our approach not only addresses the limitations of traditional approaches but also leverages the strengths of LLMs, aiming to achieve more accurate and adaptable bandwidth predictions.

4.2 Network Context-Informed Embedding Guidance

Recognizing the vital importance of domain-specific knowledge and the influence of factors beyond historical bandwidth data, we introduce a novel Network Context-Informed Embedding Guidance mechanism. We acknowledge that additional network status information, such as packet loss rates, delay variability, and network congestion levels, play a crucial auxiliary role in the bandwidth prediction task. To overcome the challenge of integrating these heterogeneous data types, which are often represented in different modalities, we harness the powerful understanding ability of LLMs in the text modality. By converting these network status data into text descriptions as inputs, we enable the model to effectively tap into their valuable information, thereby enhancing its predictive capabilities.

This innovative approach allows the model to leverage textual representations that capture the intricate nuances of network dynamics, including the complex interplay between network components and the subtle variations in user behavior. By incorporating this contextual information, the model transcends traditional data-driven approaches, becoming not only data-driven but also informed by the broader network environment. This context-aware approach leads to more accurate and informed predictions, ultimately enhancing the model's ability to optimize network resource allocation and improve user experience.

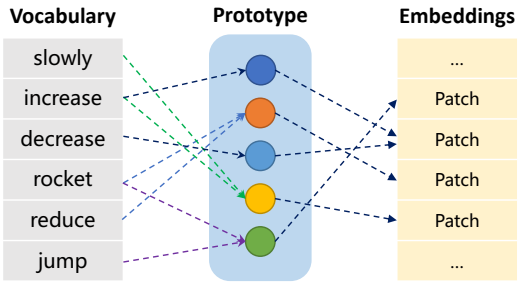

Figure 4: Preliminary feature extraction. Identifying prototypes from vocabulary to streamline cross-attention, enabling efficient and accurate integration of multi-modal data.

4.3 Alignment between Text and Temporal Modalities

Recognizing that time series data and textual information belong to different modalities, direct processing by Large Language Models (LLMs) is not straightforward. To bridge this gap and facilitate the integration of multi-modal data, we propose the utilization of a cross-attention mechanism as a pivotal component in our approach to bridge the gap between these modalities, thereby enabling the seamless integration of multi-modal data.

The cross-attention mechanism we employ is designed to facilitate the alignment of textual and temporal features, a critical step in amalgamating the contextual richness provided by the Network Context-Informed Embedding Guidance with the historical bandwidth data. This alignment is achieved by allowing the model to focus on relevant features across modalities, ensuring that the textual context is effectively integrated with the temporal dynamics of the network conditions.

As shown in Figure 4, we introduce a preliminary feature extraction step to optimize computational efficiency and mitigate the potential computational overhead associated with processing multi-modal data. This involves a meticulous analysis of the vocabulary to identify prototypes that encapsulate the essence of network conditions. These prototypes serve as a condensed representation of the key aspects of the network environment, thereby streamlining the cross-attention process. By leveraging these prototypes, our model is able to process and integrate multi-modal data with enhanced efficiency, without compromising on performance or accuracy.

Furthermore, this approach ensures that the model can effectively leverage the contextual information embedded in textual data to enrich its understanding of the time series data. This not only enhances the model's ability to make informed predictions but also opens up new avenues for exploring the complex interplay between textual and temporal data in network analysis. In essence, our methodology represents a significant advancement in the field of multi-modal data integration, offering a robust and efficient solution for processing and understanding complex data landscapes.

4.4 Parameter-Efficient Fine-Tuning of Large Language Models

In the culmination of our architectural framework, the training of the Large Language Model (LLM) module assumes a pivotal role, serving as the linchpin for our bandwidth prediction endeavors. We have judiciously chosen the Llama2-7B model as the foundational architecture, a decision underpinned by its established prowess in grappling with intricate patterns and its inherent scalability. This selection is not merely arbitrary; it is a testament to the model's capability to handle the multifaceted nature of our data, ensuring that our predictions are grounded in robust computational foundations.

To tailor this pre-trained LLM to our specific bandwidth prediction task, we harness the power of the LoRA (Low-Rank Adaptation) technique. LoRA's elegance lies in its ability to adapt the model to our task with a minimal alteration of the original model parameters. This approach is not only efficient but also strategic, as it allows us to preserve the rich knowledge encapsulated within the pre-training phase while simultaneously optimizing the model for our bandwidth prediction objectives. By doing so, we strike a delicate balance between leveraging the general knowledge acquired during pre-training and acquiring task-specific insights, a synergy that is crucial for enhancing the model's performance.

Table 1: Bandwidth Prediction Result. A lower value indicates better performance. The best results for each metric are noted in **bold** and the second best results are underlined.

Method Time	Metric	BP-LLM (Ours)	Reformer (2020)	Informer (2021)	TSMixer (2023)	MICN (2023)	DLinear (2023)	Autoformer (2021)
4-4	MSE	0.296	**0.287**	0.288	0.293	0.291	0.298	0.334
	MAE	0.407	**0.400**	**0.400**	0.413	0.408	0.412	0.420
	RMSE	0.545	**0.536**	0.537	0.541	0.539	0.546	0.578
16-16	MSE	0.319	**0.309**	0.310	0.311	0.313	0.319	0.348
	MAE	0.430	0.422	**0.421**	0.428	0.429	0.432	0.436
	RMSE	0.564	**0.556**	0.557	0.558	0.559	0.565	0.590
64-64	MSE	**0.296**	0.321	0.319	0.320	0.321	0.323	0.328
	MAE	**0.407**	0.435	0.433	0.439	0.437	0.437	0.441
	RMSE	**0.545**	0.567	0.565	0.566	0.567	0.568	0.572
Average	MSE	**0.304**	0.306	0.306	0.308	0.308	0.313	0.337
	MAE	**0.415**	0.419	0.418	0.427	0.424	0.427	0.433
	RMSE	**0.551**	0.553	0.553	0.555	0.555	0.560	0.580
Count		6	5	2	0	0	0	0

The fine-tuning process, facilitated by LoRA, is indispensable for our bandwidth prediction task. It empowers the model to learn patterns that are uniquely relevant to our specific domain, thereby enriching its predictive capabilities. This enhancement is not merely quantitative; it is qualitative, leading to predictions that are not only more accurate but also more adaptable to the dynamic nature of network conditions. In essence, this fine-tuning phase is the crucible in which the model's potential is fully realized, transforming it into a powerful tool for bandwidth prediction that is both precise and versatile.

5 Experiments

5.1 Bandwidth Prediction

Baselines. In this experiment, we compare the results of BP-LLM with the current state-of-the-art time series forecasting methods. The methods used as comparisons include Reformer [16], Informer [32], TSMixer [7], MICN [24], DLinear [29] and Autoformer [25].

Setup. The experiment's input consisted of three dimensions from the previous time step: sampling time, sampling duration, and bandwidth value, with the output being the predicted bandwidth value. We evaluated the models across three different time windows (4-4, 16-16, 64-64) and measured their performance using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Result. The results are shown in Table 1. Despite exhibiting slightly inferior performance in shorter time windows such as 4-4, where its Mean Squared Error (MSE) was marginally higher compared to models like Reformer and Informer, BP-LLM showcased its true prowess as the time window expanded. Notably, in the 64-64 time window, BP-LLM achieved the lowest MSE value of 0.296, underscoring its exceptional capability to maintain high-precision bandwidth prediction performance even with significantly larger data volumes. This trend is indicative of BP-LLM's strength in long-term modeling, where it effectively captures and predicts bandwidth trends over extended periods.

Moreover, BP-LLM's robust performance was not limited to MSE alone. It also demonstrated strong results in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) across different time windows, further solidifying its position as a superior model in the bandwidth prediction domain. These comprehensive results highlight BP-LLM's versatility and reliability, as it not only handles data from varying time windows adeptly but also delivers stable and efficient prediction performance across a spectrum of error measures.

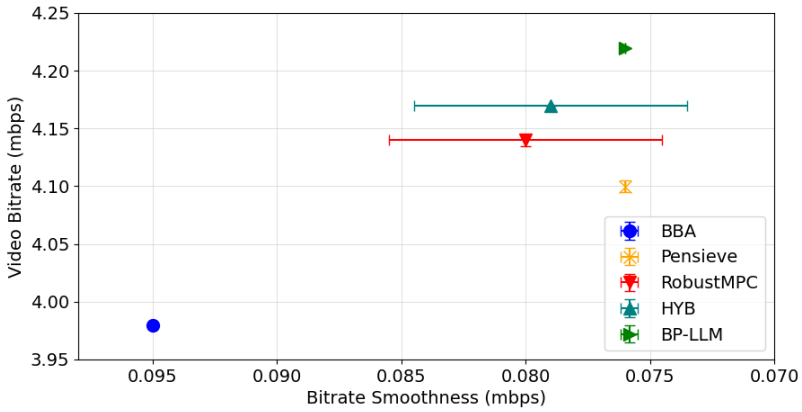

Figure 5: BP-LLM for Adaptive Bitrate Streaming (ABR).

In summary, while BP-LLM may face slight challenges in very short time windows, its overall performance, especially in longer time frames, is outstanding. It consistently outperforms or matches state-of-the-art models in critical metrics such as MSE and MAE, positioning it as a highly competitive solution for bandwidth prediction tasks. The model's ability to adapt and excel across different time windows and error measures makes it a reliable choice for optimizing network resource allocation and enhancing user experience in various network applications.

5.2 Downstream Evaluation

We conducted a comprehensive evaluation of our proposed method, BP-LLM, against a variety of traditional algorithms in the context of downstream tasks that are critical for optimizing network performance and enhancing user experience. Our evaluation focused on two key downstream tasks: Adaptive Bitrate Streaming (ABR) and Quality of Experience (QoE) improvement. ABR is a critical component in streaming media applications, where the goal is to dynamically adjust the video quality based on the available network bandwidth to ensure smooth playback and minimize buffering. QoE, on the other hand, encompasses a broader range of factors that contribute to the overall satisfaction of users, including video quality, buffering frequency, and playback continuity.

Baselines. The algorithms we compared against include BBA [14], Pensieve [19], RobustMPC [28], and HYB [2], each of which represents a different approach to addressing network challenges and improving service quality.

ABR Test. We assessed the ability of BP-LLM and the competing algorithms to adapt the bitrate of the streaming content in response to changes in network conditions. The results are depicted in Figure 5. Notably, BP-LLM demonstrates a bitrate smoothness of approximately 0.075 Mbps, with a corresponding video bitrate nearing 4.25 Mbps. This positioning suggests that our method effectively manages bandwidth fluctuations while maintaining a high level of video quality. In comparison, other algorithms exhibit higher bitrate smoothness but achieve lower video bitrates. This indicates that while they may handle bandwidth variability well, they compromise on video quality. Overall, BP-LLM's performance strikes a commendable balance between bitrate smoothness and video bitrate, thereby providing a more consistent and higher-quality streaming experience.

QoE Improvement Test. We evaluated the overall impact of each algorithm on user satisfaction. This included analyzing factors such as video quality, buffering duration, and the overall smoothness of the playback. Our findings revealed that BP-LLM significantly enhanced QoE compared to the traditional algorithms, with users reporting higher satisfaction levels due to fewer interruptions and higher video quality throughout the streaming session. As is shown in Figure 6, BP-LLM achieves the highest QoE score of 4.22, surpassing all other algorithms, with RobustMPC [28] and HYB [2] following at 4.14 and 4.17 respectively and BBA [14] performing the lowest. The CDF indicates that as the QoE score increases, the probability of achieving that score also rises, with BP-LLM demonstrating a steeper ascent, suggesting a higher concentration of users experiencing superior

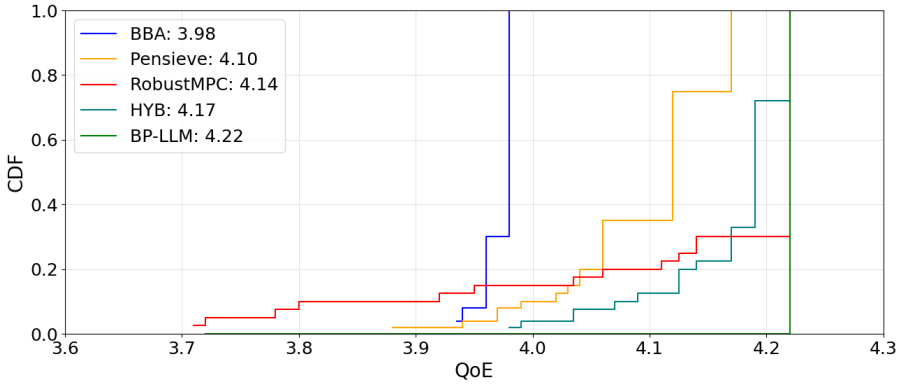

Figure 6: BP-LLM for QoE improvement.

quality. This analysis highlights BP-LLM's effectiveness in optimizing user experience, making it a promising choice for applications demanding high QoE.

In summary, our comparative analysis of BP-LLM against traditional algorithms in the context of ABR and QoE improvement tasks showcased the superior performance of our method. BP-LLM's ability to accurately predict bandwidth and adapt to changing network conditions resulted in a more stable and higher-quality streaming experience, as evidenced by the improved ABR performance and enhanced QoE metrics. This underscores the potential of BP-LLM to revolutionize network management strategies and significantly contribute to the optimization of streaming media applications.

6 Conclusion

In summary, this research has introduced a groundbreaking methodology for bandwidth prediction, coined BP-LLM, which leverages the sophisticated capabilities of large language models (LLMs) to significantly enhance prediction accuracy and adaptability. By confronting and overcoming the inherent limitations of traditional bandwidth prediction algorithms, BP-LLM has demonstrated its superiority in several critical aspects. It excels in capturing long-term dependencies in network traffic patterns, effectively handling multimodal inputs that include network status, user location, and communication latency, and demonstrates remarkable generalization capabilities across a wide spectrum of network conditions.

We establish a comprehensive benchmark that spans diverse application scenarios such as Video-on-Demand (VoD), Live Streaming, and Real-Time Communication (RTC). This benchmark enhances the evaluation process and ensures that BP-LLM is tested under realistic and varied network conditions, providing a robust foundation to assess its capabilities.

BP-LLM excels across error metrics such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE), outperforming state-of-the-art algorithms. This reliability supports downstream tasks like resource allocation, congestion control, and quality of service (QoS) management, enhancing network resource utilization and user experience in streaming applications.

Future research directions include integrating BP-LLM with advanced machine learning techniques like reinforcement learning or graph neural networks, and exploring its applicability to emerging technologies such as 5G. Continuous refinement and optimization will ensure BP-LLM's effectiveness across diverse network conditions and scenarios.

In conclusion, the development and evaluation of BP-LLM represent a significant milestone in the field of bandwidth prediction. By addressing critical challenges and demonstrating superior performance, this work not only advances the state of the art but also paves the way for more efficient and intelligent network management strategies. The potential impact of BP-LLM on enhancing the competitiveness of streaming media applications and optimizing network resource allocation is substantial, underscoring the importance of continued research and innovation in this domain.

References

[1] Meta AI. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[2] Zahaib Akhtar, Yun Seong Nam, Ramesh Govindan, Sanjay Rao, Jessica Chen, Ethan Katz-Bassett, Bruno Ribeiro, Jibin Zhan, and Hui Zhang. Oboe: Auto-tuning video abr algorithms to network conditions. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 44–58, 2018.

[3] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[5] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 6989–6997, 2023. URL https://arxiv.org/abs/2201.12886.

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020.

[7] Si-An Chen, Chun-Liang Li, Sercan O Arik, Nathanael Christian Yoder, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecast-ing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wbpxTuXgm0.

[8] Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan K Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=pCbC3aQB5W.

[9] Yiqun Duan, Charles Chau, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters, 2024. URL https://arxiv.org/abs/2310.07820.

[11] William Han, Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Douglas Weber, Bo Li, and Ding Zhao. An empirical exploration of cross-domain alignment between language and electroencephalogram. *arXiv preprint arXiv:2208.06348*, 2022.

[12] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), June 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9 (8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[14] Te-Yuan Huang, Ramesh Johari, Nick McKeown, Matthew Trunnell, and Mark Watson. A buffer-based approach to rate adaptation: Evidence from a large video streaming service. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 187–198, 2014.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[16] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

[17] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020. URL https://arxiv.org/abs/1912.09363.

[18] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X. Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0EXmFzUn5I.

[19] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. Neural adaptive video streaming with pensieve. In *Proceedings of the conference of the ACM special interest group on data communication*, pages 197–210, 2017.

[20] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

[21] OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[23] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[24] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *The eleventh international conference on learning representations*, 2023.

[25] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021. URL https://arxiv.org/abs/2106.13008.

[26] Hao Xue and Flora D Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[27] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.

[28] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. A control-theoretic approach for dynamic adaptive video streaming over http. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pages 325–338, 2015.

[29] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023. URL https://arxiv.org/abs/2205.13504.

[30] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*,

KDD '21, page 2114–2124, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467401. URL https://doi.org/10.1145/3447548.3467401.

[31] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. Large language models for time series: A survey, 2024. URL https://arxiv.org/abs/2402.01801.

[32] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[33] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, 2022. URL https://arxiv.org/abs/2201.12740.

[34] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.