DOI: 10.54254/2977-3903/2024.19435

A MDA-based multi-modal framework for panoramic viewport prediction

Jinghao Lyu

School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, China

791603612@qq.com

Abstract. Panoramic viewport prediction is crucial in 360-degree video streaming, aiming to forecast users' future viewing regions for efficient bandwidth management. To achieve accurate panoramic viewport prediction, existing frameworks have explored the utilization of multi-modal inputs, combining trajectory, visual, and audio data. However, they uniformly process different modalities through standardized pipelines and use concatenation-based feature fusion regardless of modality characteristics. With the unmodified application of computationally intensive Transformer architectures, the uniform design exacerbates computational overhead. Besides that, the concatenation-based feature fusion lacks the ability to model global dependencies and explicit interactions between different modalities, which limits the prediction accuracy. To overcome these issues, we introduce a lightweight Modality Diversity-Aware (MDA) framework including two primary components: a lightweight feature refinement module and a cross-modal attention module. The feature refinement module uses compact latent tokens to sequentially process audio-visual data, thereby filtering out irrelevant background signals and reducing model parameters. Following this, our cross-modal attention module effectively fuses trajectory features with the refined audio-visual features by allocating attention weights on the effective features, improving the prediction accuracy. Experimental results on a standard 360-degree video benchmark demonstrate that our MDA framework achieves higher prediction accuracy than current multi-modal frameworks, while requiring up to 50% fewer parameters.

Keywords: viewport prediction, deep learning, multi-modal fusion, panoramic video

1. Introduction

The advancement of panoramic video technology has revolutionized immersive experiences in applications such as virtual reality and live streaming, offering users a 360-degree spherical view of dynamic scenes. However, the data volume of panoramic video is significantly higher than that of traditional video formats, which has led to the urgent need to optimize transmission [1] and reduce bandwidth consumption. To address this issue, viewport prediction, which anticipates users' future viewing regions within 360-degree content, has emerged as a critical technique for optimizing resource allocation. Accurate prediction enables prioritized streaming of high-resolution content within the predicted viewport, minimizing bandwidth waste [2].

Early solutions predominantly relied on user trajectory data—historical head rotation angles and gaze positions—to model viewing patterns, using techniques like linear regression [3], recurrent neural networks (RNNs) [4], or long short-term memory (LSTM) networks [5]. While such single-modal approaches achieved reasonable results, their performance significantly degraded in scenarios involving dynamic scene-driven signals such as moving objects or salient sounds. Recent methods incorporate visual, audio, and trajectory information to improve prediction accuracy. Figure 1(a) shows the framework of existing methods [11, 28]. The data of audio, visual, and trajectory modality are encoded by the respective encoders E_a, E_v, and E_t. Then the encoded features are treated uniformly, going through the computationally intensive Transformer networks of multiple layers and get concatenated during the feature fusion module. These existing solutions treat each modality uniformly, causing two major challenges. First, uniformly processing all inputs often inflates model size and computational cost, hindering real-time performance. They lack an effective mechanism to filter out irrelevant information such as static backgrounds in visual or ambient noise in audio, resulting in less accurate predictions and added computational overhead. Secondly, the feature fusion based on concatenation fails to capture global dependencies and explicit interactions across different modalities, which restricts the overall prediction accuracy.

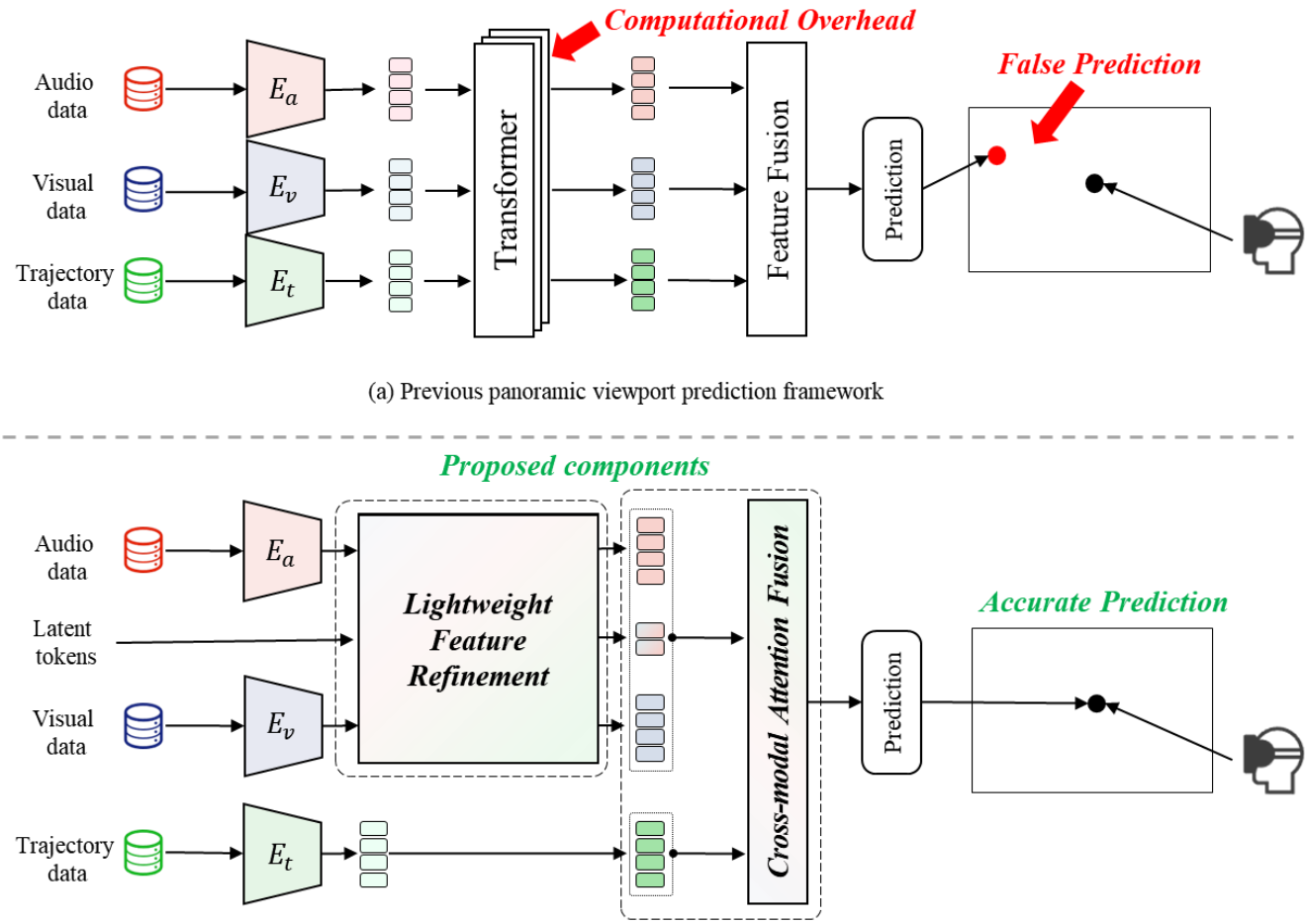

Figure 1. Comparison of existing frameworks and our framework. (a) shows the framework of existing methods [11, 28], where a uniform Transformer-based feature extraction and basic feature fusion lead to high complexity and unreliable predictions. (b) presents our Modality Diversity-Aware (MDA) framework, featuring a lightweight feature refinement module to remove irrelevant audio-visual cues and a cross-modal attention fusion module for selective integration with trajectory feature, thereby achieving more accurate prediction at reduced computational cost.

In this paper, we propose a Modality Diversity-Aware (MDA) framework that significantly reduces computational overhead while improving prediction accuracy. As illustrated in Figure 1(b), we first utilized the similar encoders E_a, E_v, and E_t to encode the audio, visual, and trajectory input data. After that, a lightweight feature refinement module processes audio-visual data via compact latent tokens, effectively filtering out irrelevant elements. This design substantially lowers parameters without undermining representational quality. Subsequently, a cross-modal attention fusion module fuses the refined audio-visual features with trajectory features by allocating attention weights to the most pertinent features, which improves the panoramic viewport prediction accuracy. Extensive experiments demonstrate that the MDA framework not only surpasses state-of-the-art multi-modal baselines in prediction accuracy but also achieves up to 50% fewer parameters, making it well-suited for resource-constrained deployment.

Our main contributions are summarized as follows:

- We propose a Modality Diversity-Aware (MDA) framework for efficient and accurate panoramic viewport prediction, effectively balancing computational complexity with prediction performance.
- We develop two novel components, a lightweight feature refinement module and a cross-modal attention module, which jointly filter out non-essential audio-visual information and dynamically integrate trajectory cues, leading to accurate viewport prediction and up to 50% fewer parameters compared to existing baselines.

2. Related work

The viewport prediction problem is often modeled as a time series prediction problem [6] because the users' viewport position is temporally correlated [7]. Some traditional temporal prediction methods are widely used in this scenario, including linear

regression and probabilistic statistics. However, these methods cannot maintain a high level of accuracy because they are difficult to learn the complex behavior patterns of users [8].

With the development of deep learning and reinforcement learning theories, researchers have proposed to apply learning-based methods to solve the challenge of low accuracy of viewport prediction. Bao et al. [14] proposed to use Long Short-Term Memory (LSTM) network model to improve the accuracy of prediction algorithms in long-term prediction. Xu et al. [15] established a deep reinforcement learning model for viewport prediction. Lee et al [16] introduced an attention module and combined the LSTM and Gated Recurrent Unit (GRU) to more accurately predict the users' viewport position in panoramic videos.

In addition to the viewport prediction method only based on history trajectory, there are video content-dependent methods that combine the users' history trajectory with the visual and audio content of the panoramic video for viewport prediction. Researchers integrated visual cues [12], as users naturally focus on dynamic objects [17] and high-contrast regions [18]. Some methods incorporated object detection models like YOLOv3 [19] to track moving elements, while others used convolutional neural networks like VGG-16 network [20, 21] and Inception-ResNet-V2 [13] to extract visual features. Recent studies explored spatial audio [22, 23], as directional sounds influence the movement of users' viewports. Zhang et al. [11] introduced an audio-assisted model mapping sound intensity to spatial locations. Wu et al. [28] leveraged spherical convolution networks to refine panoramic audio-visual processing. These two latest research utilized the Transformer-based pipeline for multi-modal viewport prediction, but their excessive computational overhead and prediction accuracy remain to be optimized. This is because of their uniform processing and concatenation-based feature fusion treatment of three modalities.

In summary, viewport prediction has evolved from trajectory-based models to multi-modal frameworks incorporating visual and audio elements. While accuracy has improved, challenges remain in balancing prediction effectiveness and computational efficiency. Our work proposes a modality-aware framework that accurately predicts the future viewport while maintaining efficiency for real-time applications.

3. Methodology

Our Modality Diversity-Aware (MDA) framework for panoramic viewport prediction comprises three primary phases: multi-modal feature encoder, lightweight feature refinement, and cross-modal attention fusion. As illustrated in Figure 2, each phase tackles a specific challenge associated with analyzing 360-degree data. First, trajectory, audio, and visual features are individually extracted via specialized encoders. Next, a lightweight feature refinement module leverages compact latent tokens to reduce parameter overhead while retaining the most salient cues of audio and visual modalities. Finally, the cross-modal attention module fuses these refined representations with trajectory data, enabling accurate and efficient viewport prediction.

3.1. Visual encoder module

Panoramic video frames exhibit distortions due to their spherical nature. To preserve spatial relationships, we apply a spherical convolution encoder network to the visual input X_v:

$$h_v = SphConv(X_v) \tag{1}$$

where h_v is the extracted visual feature representation, with three layers of spherical convolution and a convolution kernel of size 3x3. Each layer of spherical convolution is followed by the action of the Rectified Linear Unit activation layer and a maximum pooling layer of size 2x2 with a step size of 2. This step enhances the most salient regions of the panoramic frame while reducing irrelevant background information.

3.2. Audio encoder module

According to the research in [25], the directional perception of sound sources affect the viewers' viewport position in the panoramic video. Following [10], we first generate audio energy maps (AEM) to localize prominent sound sources. Then a convolutional encoder extracts the audio input X_a to obtain audio features h_a:

$$h_a = ConvEnc(X_a) \tag{2}$$

The encoder consists of two convolutional layers followed by batch normalization and ReLU activation. Each convolutional layer applies 2D convolutions with kernel sizes of 3x3 using a stride of 2. This process ensures that strong directional audio cues are retained, while irrelevant ambient cues are suppressed.

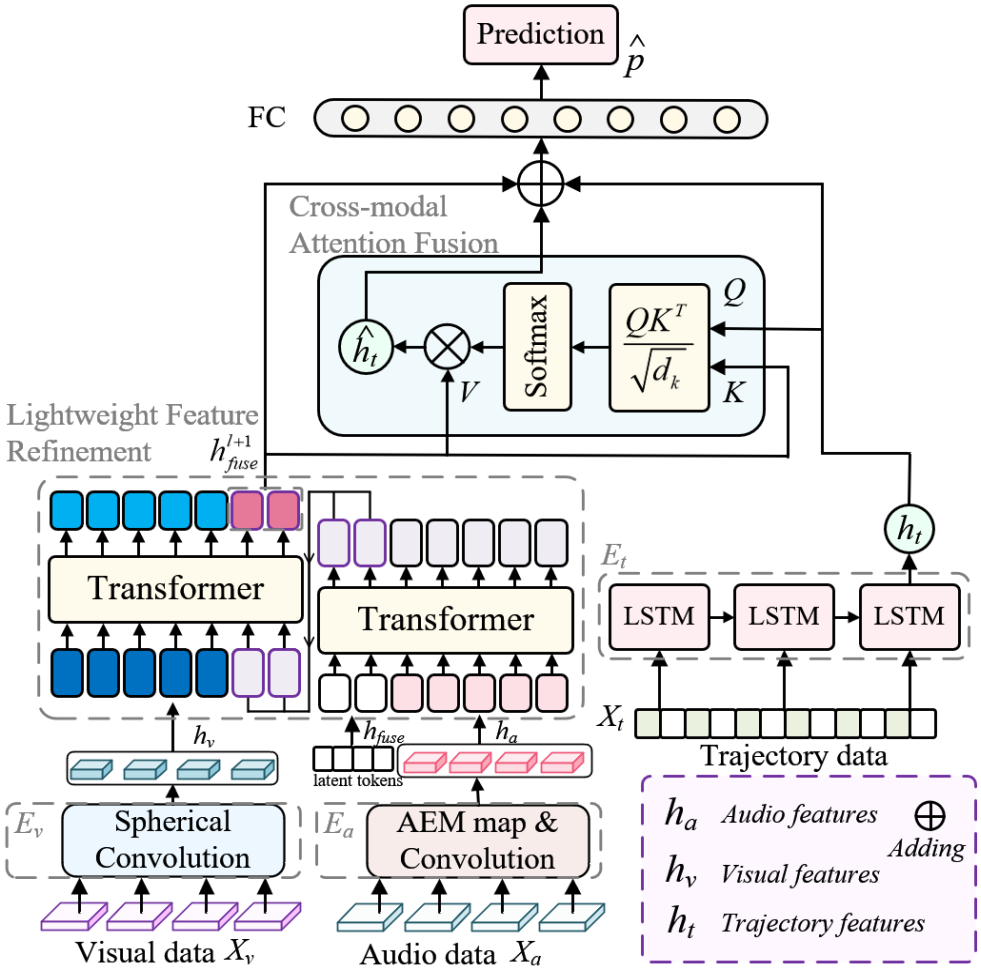

Figure 2. Overall structure of our proposed framework

3.3. Trajectory encoder module

The user's head movement is represented as a sequence of Euler angles capturing yaw, pitch, and roll across H timestamps. The trajectory sequence X_t is formulated as:

$$X_t = \{x_{t-H+1}, x_{t-H+2}, \dots, x_t\} \tag{3}$$

where x_i denotes the three-dimensional rotation angles at time step i. These trajectory sequential dependencies necessitate a model capable of effectively capturing both short-term fluctuations and long-term patterns in head movement. To achieve this, we employ a Long Short-Term Memory (LSTM) network to process this sequence to extract the trajectory features h_t:

$$h_t = LSTM(X_t) \tag{4}$$

This embedding encodes both short-term fluctuations and long-term motion trends.

3.4. Lightweight feature refinement module

Although h_v and h_a encode useful scene information, they also contain irrelevant details. To reduce computational cost while retaining important features, we introduce a compact latent token compression mechanism. This key procedure decreases pairwise attention's computational complexity by introducing the tiny latent fusion unit $h_{fuse}^l = [h_1^f, h_2^f, \dots, h_C^f]$ of length C to the visual and audio embedding units. Attention flows of audio and visual modalities are restricted within the latent fusion unit for information sharing. We calculate the fusion unit descriptions in the following format:

$$\left[h_a^{l+1}\right]\left[\widehat{h_{fuse}^{l+1}}\right] = Transformer\left(\left[h_a^l\right]\left[h_{fuse}^l\right]\right) \tag{5}$$

$$\left[h_v^{l+1}\right]\left[h_{fuse}^{l+1}\right] = Transformer\left(\left[h_v^l\right]\left[\widehat{h_{fuse}^{l+1}}\right]\right) \quad (6)$$

where l denotes the layer of the Transformer model [24]. To avoid irrelevant interactions across modalities, we refine them sequentially. The stream of attention between two modalities are restricted inside one Transformer following [26]. Since audio signals often serve as early indicators of user attention shifts, the audio features are processed before the visual features. This ensures that each modality is independently filtered before fusion, preventing unnecessary propagation of background information and irrelevant textures.

3.5. Cross-modal attention fusion module

The cross-modal attention procedure establishes directly paired attention between modalities by utilizing information from the source one to enhance the target one.

Since the lightweight feature refinement module has learned the audio-visual fused feature, we can now apply the cross-modal attention fusion to reinforce the trajectory feature h_t with the fused information h_{fuse}^{l+1}, which can be represented as follows:

$$\begin{aligned}\widehat{h_t} &= Cross - Attention\left(Q = h_t, K, V = h_{fuse}^{l+1}\right) \\ &= softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V\end{aligned} \quad (7)$$

where Q, K and V represent the query, key, and value matrices in the cross-attention mechanism and d_K represents the dimension of the key embedding. By computing attention between query and key sequences, cross-attention facilitates the capture of contextual relationships and outputs the enhanced trajectory feature embedding $\widehat{h_t}$. This ensures the improved comprehension of complex dependencies, contributing to accurate viewport prediction results.

3.6. Prediction output layer

Ultimately, we get the initial trajectory feature h_t, enhanced trajectory feature embedding $\widehat{h_t}$, and refined audio-visual embedding h_{fuse}^{l+1}. These result features from each modality are then added together to obtain $I = h_t + \widehat{h_t} + h_{fuse}^{l+1}$. The final viewport prediction is therefore obtained by a layer of fully connected network:

$$\hat{P} = W_{out}\, I + b_{out} \quad (8)$$

Where W_{out} is the weight vector, b_{out} is the bias.

4. Experiment

4.1. Dataset

To evaluate the effectiveness of our proposed framework, we utilize the Xu_CVPR_18 dataset [13], a widely recognized benchmark in multi-modal viewport prediction. This dataset consists of 208 high-quality 360-degree videos, each with an average duration of 36 seconds. A total of at least 31 participants watched and interacted with each video, ensuring diverse user behaviors. The dataset encompasses a variety of content categories, including film clips, sports events, outdoor adventures, and live concerts, providing a comprehensive and diverse set of test scenarios for evaluating framework performance. By leveraging this dataset, we ensure that our experimental results are well-founded and reflective of real-world viewport prediction tasks.

4.2. Implementation details

The dimension of the LSTM layer for encoding trajectory modality is set to 512, the length C of latent token is set to 4 and the attention heads are set to 4, with 2 Transformer layers in the lightweight feature refinement module. Motion features are extracted using the trajectory sequence of the previous five samples. We also use the spherical-convolution-based network to extract visual information from each video clip by extracting five frames at 0.2 second intervals. We utilize the Adam optimizer with a constant learning rate of 0.001 and mean square error (MSE) loss during the network training phase. The model is trained on an RTX 3090 GPU for a total of 60 epochs, with a batch size of 64.

Since the prediction of the FoV can be considered a classification problem, where pixels or tiles are classified in or out of future FoV, various metrics are considered for evaluation. We use the orthodromic distance [9] as the performance metrics following [27]. Given two points of three-dimensional coordinates in Euclidean space $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$, we

first transform them into points $P_1{}' = (\theta_1, \varphi_1)$ and $P_2{}' = (\theta_2, \varphi_2)$ on the surface of the unit sphere, where θ is the longitude and φ is the latitude of the point. The orthodromic distance OD, defined as the shortest path length between two points on a sphere, can subsequently be computed as:

$$OD(P_1, P_2) = arccos(cos(\varphi_1)\, cos(\varphi_2)\, cos(\theta_1 - \theta_2) + sin(\varphi_1)\, sin\,(\varphi_2)) \tag{9}$$

Another assessment metric for the viewport prediction model is the Intersection over Union (IoU). Once the panoramic frame is divided into tiles and the FoV is set, we can classify the tile whether within the viewport or not based on the angle difference between the viewpoint and the tile's central point. We let True Positive (TP) and True Total (TT) represent the intersection and the union of predicted and true viewport tiles respectively. The metric IoU can then be calculated as follows:

$$IoU = \frac{TP}{TT} \tag{10}$$

4.3. Comparison to state-of-the-art frameworks

We evaluate our framework in comparison to several competitive frameworks, which include Pos-only, TRACK [9], Wu_AAAI20 [28], and MFTR [11].

(1) Pos-Only (Single-modal, Trajectory-based): An LSTM-based encoder-decoder framework that solely utilizes user trajectory history for viewport prediction.

(2) TRACK (Multi-modal, Trajectory-Visual-based): A framework employing three independent LSTM networks to process trajectory sequences, visual embeddings, and their fusion, capturing spatial-temporal dependencies.

(3) Wu_AAAI20 (Multi-modal, Trajectory-Visual-Audio-based): A Transformer-based framework incorporating spherical CNNs for 360-degree feature extraction and a preference-aware viewport prediction mechanism.

(4) MFTR (Multi-modal, Trajectory-Visual-Audio-based): A multi-modal fusion Transformer-based framework that models long-range dependencies across trajectory, visual, and audio modalities using Transformer encoders.

For a fair comparison, all frameworks are evaluated under identical experimental settings, including dataset splits, training configurations, and hyperparameters.

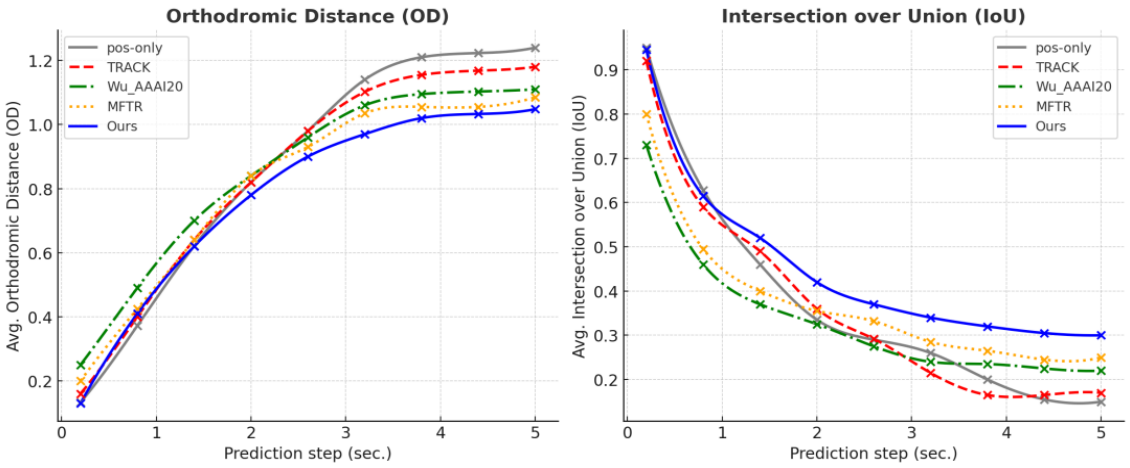

Figure 3. Prediction performance comparison with competitive frameworks across prediction steps

The quantitative evaluation results are presented in Figure 3, measuring viewport prediction accuracy based on two key metrics: orthodromic distance (OD) and intersection over union (IoU). The horizontal axis of Figure 3 represents the predicted future time window (0 to 5 seconds), while the vertical axis indicates the average orthodromic distance between predicted and actual viewport locations. A lower orthodromic distance signifies better alignment with user gaze behavior, whereas a higher distance suggests decreased prediction accuracy.

When analyzing short-term prediction performance (time window $\leqslant$1s), our proposed framework achieves the best results, with an orthodromic distance reduction of 0.12 compared to Wu_AAAI20. This improvement is attributed to our modality-aware fusion strategy, which effectively prioritizes trajectory signals while selectively integrating visual and audio cues. Meanwhile, Pos-Only and TRACK exhibit nearly identical performance, reinforcing the limitations of trajectory-only models in capturing contextual influences. As the prediction horizon extends to 5 seconds, all models experience a gradual decline in accuracy. However, our framework maintains a consistent lead. Notably, the single-modal framework (Pos-Only) exhibits the most

pronounced degradation, as it lacks external scene-awareness and relies purely on past movement patterns. MFTR, the most complex multi-modal model, performs comparably to our framework at longer horizons, but it incurs significantly higher computational costs. Our framework also demonstrates superior IoU scores, particularly in long-term predictions. This highlights the benefit of considering modality-specific information density variations—trajectory data dominates stable viewing phases, while visual and audio cues contribute more dynamically during abrupt scene changes. By addressing these modality interactions, our framework effectively balances precision and efficiency.

Beyond predictive accuracy, model efficiency is crucial for real-time applications. Figure 4 presents a comparative analysis of parameter counts across multi-modal fusion models. Our proposed framework significantly reduces parameter complexity compared to MFTR, achieving a more compact architecture while preserving high accuracy. This efficiency gain is primarily due to our modality diversity-aware fusion strategy, which eliminates excessive computations while maintaining essential cross-modal interactions. Compared to Wu_AAAI20 and TRACK, which rely on multiple stacked recurrent networks for trajectory modeling, our framework benefits from a lightweight Transformer-based refinement module, reducing computational overhead without sacrificing performance.

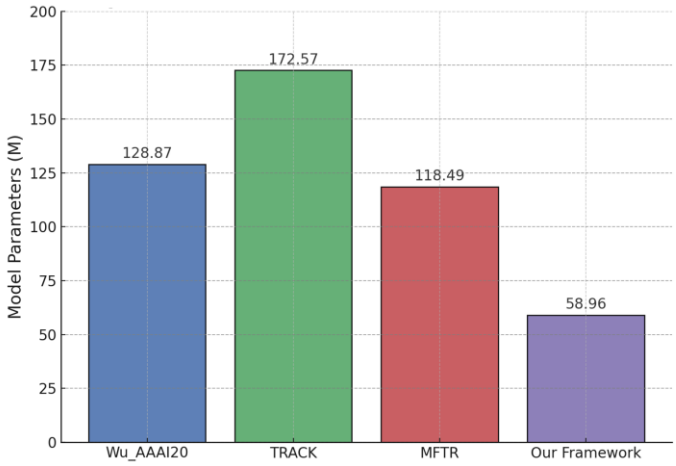

Figure 4. Model parameters comparison with competitive frameworks

5. Conclusion

In this paper, we introduced the Modality Diversity-Aware (MDA) framework, a lightweight yet effective solution for panoramic viewport prediction. Our framework addresses the limitations of existing multi-modal frameworks by reducing computational overhead while maintaining high prediction accuracy. Specifically, we designed two key components: a lightweight feature refinement module that leverages latent token aggregation to efficiently process audio-visual data, and a cross-modal attention fusion module that selectively integrates refined audio-visual cues with trajectory features. These innovations enable precise and computationally efficient viewport prediction, making our framework well-suited for real-time panoramic video streaming. Experimental evaluations on a benchmark 360-degree video dataset demonstrate that our MDA framework consistently outperforms state-of-the-art models in terms of viewport prediction accuracy, achieving lower orthodromic distances and higher IoU scores across different time horizons. Moreover, our framework reduces model parameter by up to 50% compared to existing multi-modal frameworks, ensuring a balance between efficiency and predictive performance. The results validate that our modality-aware fusion strategy effectively prioritizes trajectory signals while selectively integrating visual and audio cues, leading to accurate viewport predictions even in dynamic scenarios.

While our framework provides an accurate and efficient viewport prediction mechanism, future work could explore adaptive modality weighting to dynamically adjust the contribution of each modality based on scene transitions and user behavior. Further improvements could also be made by optimizing the framework for real-time edge computing scenarios and expanding the fusion process to incorporate physiological signals to better capture user intent. By enhancing the understanding of multi-modal interactions in immersive environments, our research paves the way for more intelligent and computationally efficient panoramic video streaming solutions.

References

[1] Hirway, A., Qiao, Y., & Murray, N. (2024). A Quality of Experience and Visual Attention Evaluation for 360 videos with non-spatial and spatial audio. In ACM Transactions on Multimedia Computing, *Communications and Applications*, vol. 20, no. 9, pp. 1-20.

[2] Wan, Z., Ma, M., & Liu, X. (2023). Collaborative Edge Caching for Panoramic Video Streaming. In 2023 IEEE International Performance, *Computing, and Communications Conference*, pp. 488-494.

[3] Qian, F., Ji, L., Han, B., & Gopalakrishnan, V. (2016). Optimizing 360 video delivery over cellular networks. In Proceedings of the 5th Workshop on All Things Cellular: Operations, *Applications and Challenges*, pp. 1-6.

[4] Yang, Q., Zou, J., Tang, K., Li, C., & Xiong, H. (2019). Single and sequential viewports prediction for 360-degree video streaming. In *2019 IEEE International Symposium on Circuits and Systems*, pp. 1-5.

[5] Jamali, M., Coulombe, S., Vakili, A., & Vazquez, C. (2020). LSTM-based viewpoint prediction for multi-quality tiled video coding in virtual reality streaming. In *2020 IEEE International Symposium on Circuits and Systems*, pp. 1-5.

[6] Liu, X., Yan, J., Huang, L., Fang, Y., Wan, Z., & Liu, Y. (2024). Perceptual Quality Assessment of Omnidirectional Images: A Benchmark and Computational Model. In *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 6, pp. 1-24.

[7] Li, J., Han, L., Zhang, C., Li, Q., & Liu, Z. (2023). Spherical convolution empowered viewport prediction in 360 video multicast with limited FoV feedback. In *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 1, pp. 1-23.

[8] Park, S., Bhattacharya, A., Yang, Z., Das, S. R., & Samaras, D. (2021). Mosaic: Advancing user quality of experience in 360-degree video streaming with machine learning. In *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 1000-1015.

[9] Rondón, M. F. R., Sassatelli, L., Aparicio-Pardo, R., & Precioso, F. (2021). TRACK: A New Method from a Re-Examination of Deep Architectures for Head Motion Prediction in 360 Videos. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5681-5699.

[10] Chao, F. Y., Ozcinar, C., Zhang, L., Hamidouche, W., Deforges, O., & Smolic, A. (2020). Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *2020 IEEE International Conference on Visual Communications and Image Processing*, pp. 355-358.

[11] Zhang, Z., Chen, Y., Zhang, W., Yan, C., Zheng, Q., Wang, Q., & Chen, W. (2023). Tile classification based viewport prediction with multi-modal fusion transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3560-3568.

[12] Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, & Zhang, C. (2024). Multimodal fusion on low-quality data: A comprehensive survey. arXiv preprint arXiv:2404.18947.

[13] Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., & Gao, S. (2018). Gaze prediction in dynamic 360 immersive videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5333-5342.

[14] Bao, Y., Wu, H., Zhang, T., Ramli, A. A., & Liu, X. (2016). Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. In *2016 IEEE International Conference on Big Data*, pp. 1161-1170.

[15] Xu, M., Song, Y., Wang, J., Qiao, M., Huo, L., & Wang, Z. (2018). Predicting head movement in panoramic video: A deep reinforcement learning approach. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2693-2708.

[16] Lee, D., Choi, M., & Lee, J. (2021). Prediction of head movement in 360-degree videos using attention model. *Sensors*, pp. 3678-3699.

[17] Tang, J., Huo, Y., Yang, S., & Jiang, J. (2020). A viewport prediction framework for panoramic videos. In *2020 International Joint Conference on Neural Networks*, vol. 21, no. 11, pp. 1-8.

[18] Feng, X., Swaminathan, V., & Wei, S. (2019). Viewport prediction for live 360-degree mobile video streaming using user-content hybrid motion tracking. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1-22.

[19] Chopra, L., Chakraborty, S., Mondal, A., & Chakraborty, S. (2021). Parima: Viewport adaptive 360-degree video streaming. In *Proceedings of the Web Conference 2021*, pp. 2379-2391.

[20] Fan, C. L., Lee, J., Lo, W. C., Huang, C. Y., Chen, K. T., & Hsu, C. H. (2017). Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*, pp. 67-72.

[21] Nguyen, A., Yan, Z., & Nahrstedt, K. (2018). Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1190-1198.

[22] Yang, Q., Li, Y., Li, C., Wang, H., Yan, S., Wei, L., & Frossard, P. (2023). SVGC-AVA: 360-degree video saliency prediction with spherical vector-based graph convolution and audio-visual attention. In *IEEE Transactions on Multimedia*, vol. 26, pp. 3061-3076.

[23] Corrêa De Almeida, G., Costa de Souza, V., Da Silveira Júnior, L. G., & Veronez, M. R. (2023). Spatial Audio in Virtual Reality: A systematic review. In *Proceedings of the 25th Symposium on Virtual and Augmented Reality*, pp. 264-268.

[24] Vaswani, A. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998-6008.

[25] Bernal-Berdun, E., Martin, D., Malpica, S., Perez, P. J., Gutierrez, D., Masia, B., & Serrano, A. (2023). D-SAV360: A Dataset of Gaze Scanpaths on 360 Ambisonic Videos. In *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 11, pp. 4350-4360.

[26] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, vol. 34, pp. 14200-14213.

[27] Gao, B., Sheng, D., Zhang, L., Qi, Q., He, B., Zhuang, Z., & Wang, J. (2024). STAR-VP: Improving Long-term Viewport Prediction in 360 Videos via Space-aligned and Time-varying Fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5556-5565.

[28] Wu, C., Zhang, R., Wang, Z., & Sun, L. (2020). A spherical convolution approach for learning long term viewport prediction in 360 immersive videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 14003-14040.