

BAT-VP: Behavior-Aware Temporal–Visual Modeling for Long-Term Viewport Prediction in 360° Videos

Peiyang Chen
pchen15@student.gsu.edu
Georgia State University
Department of Computer Science
Atlanta, Georgia, USA

Huadi Zhu
hzhu17@gsu.edu
Georgia State University
Department of Computer Science
Atlanta, Georgia, USA

Wei Li
wli28@gsu.edu
Georgia State University
Department of Computer Science
Atlanta, Georgia, USA

Dong Yang
dyang26@student.gsu.edu
Georgia State University
Department of Computer Science
Atlanta, Georgia, USA

Yue Wang
ywang182@gsu.edu
Georgia State University
Department of Computer Science
Atlanta, Georgia, USA

Yingshu Li[✉]
yili@gsu.edu
Georgia State University
Department of Computer Science
Atlanta, Georgia, USA

Abstract

Long-term viewport prediction is a key technique for improving the users' quality of experience of 360° video. However, existing methods suffer from two major limitations. First, they overlook the mismatch between saliency-based visual priors and user attention mechanisms, thus deriving saliency maps that align poorly with real user viewport movements. Meanwhile, they lack effective modeling of the correlation between head and eye behaviors, and fail to produce accurate viewport predictions. To address these issues, this paper proposes BAT-VP, a novel behavior-modulated long-term viewport prediction system for 360° videos. First, we propose a visual graph encoding module to extract prototypes of users' historical visual preferences. Then, we introduce a behavior-modulated visual attention module that dynamically generates user-level saliency maps. Meanwhile, we design a head & eye soft alignment module to model the collaborative behavior between head and eye-movements. Finally, these multimodal outputs are passed through a novel Temporal & Visual Fusion module that enables adaptive collaborative prediction across time horizons. Through comprehensive evaluation on two datasets, BAT-VP outperforms five representative baselines, improving overall and long-term IoU by at least 4.9% and 7.6%, respectively, while maintaining low computational overhead.

CCS Concepts

• **Information systems** → **Multimedia information systems**;
• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Machine learning approaches*.

Keywords

360° video, long-term viewport prediction, head & eye coordination, behavior-modulated saliency, multimodal fusion

ACM Reference Format:

Peiyang Chen, Huadi Zhu, Wei Li, Dong Yang, Yue Wang, and Yingshu Li. 2026. BAT-VP: Behavior-Aware Temporal–Visual Modeling for Long-Term Viewport Prediction in 360° Videos. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3835420>

1 INTRODUCTION

Virtual reality technology, as an important bridge connecting humans and virtual worlds, has in recent years been driving the evolution of video consumption from traditional flat viewing toward 360° panoramic immersive experiences, owing to its strong sense of immersion, high interactivity, and powerful scene reproduction capability. It has consequently become an important development direction for next-generation streaming systems [37, 43, 50]. However, compared with conventional 2D video, 360° panoramic video usually requires several times higher data rates to render and transmit the entire scene content [17, 43], while users in practice only attend to a local viewport at any given moment. This not only causes substantial bandwidth and computational waste, but may also lead to video stalling, degrade user experience, and even induce motion sickness or cybersickness [5, 11]. Therefore, viewport prediction for 360° videos has become a key technique for improving transmission efficiency and perceptual quality, and its accuracy directly determines the effectiveness of system resource allocation and the quality of user experience [17, 43, 45].

To improve viewport prediction performance, prior methods initially relied on purely temporal modeling, leveraging users' historical head trajectories or viewport sequences for short-term prediction [8, 34, 36]. However, in long-term prediction, user viewports are increasingly guided by video content and deviate from short-term inertial trajectories. Based on this observation, subsequent studies began to incorporate visual priors such as saliency maps, object semantics, and visual features to improve long-term prediction performance [17, 22, 26, 47]. However, directly adopting off-the-shelf saliency maps remains insufficient, since such priors reflect general visual importance rather than user-specific attention [30, 51]. Users in the same scene may attend to different regions, even with low saliency, due to heterogeneous historical



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
<https://doi.org/10.1145/3767308.3835420>

viewing trajectories, current positions, and motion states [30, 51]. Therefore, the model must not only perceive where is salient, but also determine its relevance to the current user, thus transforming video-centric saliency priors into dynamic, personalized visual guidance [17, 26]. However, establishing such user-centric visual guidance imposes an essential challenge, as user preferences are implicitly embedded in a long historical viewport sequence and mixed with short-term noise, redundant content, and temporally imbalanced information. Directly modeling all historical visual features is computationally expensive and may introduce redundancy.

Beyond visual priors, behavioral signals also provide important cues for future viewport changes. Prior studies have shown that the two signals are correlated but non-equivalent [14, 33], with eye-movements often revealing attention shifts and exploratory behavior earlier than head-movements [33, 40]. However, their correlation and contributions are variable over time. Existing methods typically encode the two modalities separately and fuse them directly, failing to model this dynamic relationship or exploit early eye-movement cues, thereby limiting prediction performance [33, 40, 47].

Motivated by the above issues, this paper first analyzes through preliminary studies the characteristics of saliency variation under behavioral constraints and the relationship between head and eye behaviors. Then, this paper proposes BAT-VP, a novel behavior-modulated long-term viewport prediction system for 360° videos. BAT-VP consists of multiple design modules to jointly improve the long-term viewport prediction performance. First, it introduces a novel Visual Graph Encoder (VGE) that innovatively models users' historical viewport contents as a graph to extract their historical visual preferences. Next, it proposes a Behavior-Modulated Visual Attention module (BMVA) to produce a user-level saliency map that matches the current user state by jointly considering historical viewing behavior, motion trends, and spatial constraints. Furthermore, to provide more sensitive behavioral intention cues, we design a head & eye soft alignment module (SAM) that adaptively aligns and interacts head & eye behavioral representations. Finally, a Temporal & Visual Fusion strategy is proposed to adaptively fuse multimodal features and improve prediction stability. The main contributions of this paper are summarized as follows:

- We study and verify the effectiveness of historical behavior modulation in improving the individualized visual saliency priors. We further quantitatively reveal the head & eye coupling relationship during 360° video viewing and the predictive role of eye-movements for future viewport changes.
- Based on our observations, we propose BAT-VP, a novel user behavior-modulated long-term viewport prediction system, and empower it with multiple novel modules such as VGE, BMVA, SAM, and a Temporal & Visual Fusion strategy, to enhance short-term and long-term prediction performance.
- We conduct a comprehensive evaluation of the proposed BAT-VP system on two representative datasets. The results show that BAT-VP outperforms mainstream solutions in short-term, long-term, and overall viewport prediction while maintaining low computational overhead.

2 RELATED WORK

2.1 Viewport Prediction in 360° Videos

Viewport prediction has been extensively studied as a key technique for improving quality of experience in 360° video [17, 43]. Early studies mainly relied on users' short-term motion inertia such as historical head-movements or viewport trajectories for short-term viewport prediction. These methods can be broadly categorized into trajectory-based prediction [3, 29], clustering-based prediction [8, 25], and sequential prediction [34, 42]. However, as the prediction horizon enlarges, users gradually deviate from inertial motion and are increasingly attracted by video content, resulting in large long-term prediction errors [8, 17, 42].

To address this, subsequent studies incorporated visual content into viewport prediction. Feng *et al.* [16] first proposed hybrid motion tracking that jointly models user behavior and video content. Later work further improved long-term prediction by addressing projection distortion [22, 39], mobile efficiency [46], and tiling design [1]. Building on these advances, recent methods explore multi-scale trajectory modeling, content-aware fusion, multimodal Transformers, and spatiotemporal alignment [12, 17, 18, 32, 47, 49].

Meanwhile, correlations between head and eye-movements have been investigated and modeled for tasks such as eye-movement prediction and behavioral pattern analysis [20, 33, 40], supported by several public datasets [4, 14, 41]. On this basis, recent studies have integrated eye-movements to assist viewport prediction [7, 47].

Despite these advances, existing methods treat eye-movements merely as an additional modality, overlooking head & eye coordination and historical visual preferences. To address this gap, this paper first performs a behavioral study, then enhances viewport prediction by harnessing these two features.

2.2 Saliency-Guided Viewport Prediction

Video saliency models estimate visually attractive regions that approximate human attention and are widely used in immersive streaming, tile selection, and compression optimization [2, 13, 24, 27]. To further enhance long-term viewport prediction, researchers have incorporated salient visual features from 360° videos [17, 22, 26, 47] via various saliency prediction methods [44, 48, 51].

However, these methods simply take the saliency features predicted by the above models as visual priors; such saliency, reflecting population-level visual attraction, does not consider the personalized attention priors of the current user. To bridge this gap, Qiao *et al.* [30] proposed a novel saliency model, which adapts saliency prediction to users' viewport heterogeneity. Yet, its output remains a group-level saliency map based on visual features without considering user-level features. In this work, we propose to dynamically generate saliency maps for the current user at the current moment based on users' historical behaviors.

3 PRELIMINARY STUDY

3.1 Behavior-Aware Saliency Prior Analysis

Traditional saliency models primarily reflects population-level content attractiveness; in practice, user attention is jointly constrained by previously viewed regions, current motion trends, and spatial

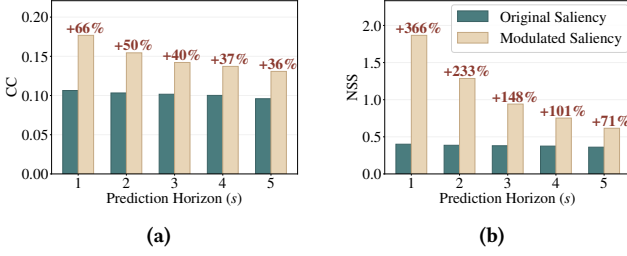


Figure 1: Quantification of Behavior-Aware modulation.



Figure 2: Visualization of Behavior-Aware modulation.

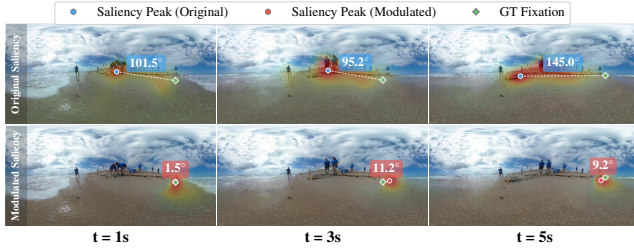


Figure 3: Comparison of saliency peaks with ground-truths.

reachability [35, 51]. This suggests necessity of constructing user-level saliency maps for viewport prediction.

To verify this, we first take the raw saliency maps as the basis and introduce three Behavior-Aware modulation mechanisms, namely inhibition of return, directional enhancement, and spatial decay, which respectively suppress previously viewed regions, enhance candidate regions aligned with the current motion tendency, and attenuate distant irrelevant areas. The constructed user-level saliency maps are then evaluated by two metrics: Normalized Scanpath Saliency (NSS), which measures saliency responses at ground-truth fixation locations, and Correlation Coefficient (CC), *i.e.*, the overall spatial similarity to the ground-truth. As shown in Fig. 1, after behavior modulation, the saliency maps achieve consistent improvements in NSS and CC across all prediction horizons.

Fig. 2 visualizes that traditional saliency maps often contain multiple hotspots and fail to accurately correspond to the user’s ground-truth current and future attention (top left). In contrast, the user-level saliency distribution becomes more concentrated and

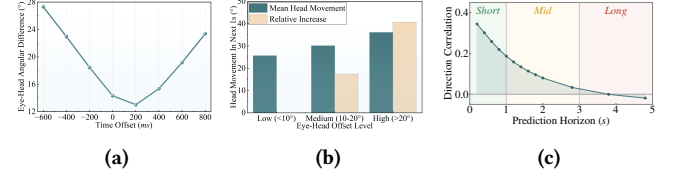


Figure 4: Statistical analysis of head & eye behavioral patterns. (a) Head & eye temporal shift. (b) Head & eye deviation. (c) Head & eye correlation.

better aligned with the user’s ground-truth attention (top right). This can be explained by the suppression in previously viewed hotspots (bottom left), the enhancement in regions along the future motion tendency (bottom middle), and the decay of peripheral irrelevant regions (bottom right). Fig. 3 further demonstrates lower error of the behavior-modulated saliency map compared to the original one across both short and long prediction horizons.

These findings indicate that user attention is dynamically shaped by behavioral history and motion state, suggesting the need to move from video-centric toward user-centric saliency. Therefore, we incorporate behavioral constraints into visual saliency modeling and dynamically construct personalized visual priors (§4.2-4.3).

3.2 Head & Eye Behavioral Pattern Analysis

Beyond visual priors, the coordination between head and eye-movements also serves as an important behavioral cue affecting viewport transitions. To validate this, we further conduct a statistical analysis of head & eye relationships over large-scale viewing sequences.

We make several observations from our experiment results. First, as shown in Fig. 4a, eye-movements generally exhibit a dynamic temporal shift from head-movements. Second, Fig. 4b illustrates that a higher head & eye deviation typically suggests a larger subsequent viewport change. Third, Fig. 4c indicates strong correlation between the eye-offset direction and future head motion direction over short term ranges, which gradually weakens as the prediction horizon increases.

These observations suggest that eye-movements contain important cues that can predict users’ viewport dynamics; however, such cues exhibit dynamic temporal shifts. In correspondence, we design a head & eye soft alignment mechanism to adaptively extract and exploit collaborative information between the two signals (§4.4).

4 SYSTEM DESIGN

4.1 System Overview

To address the aforementioned challenges in long-term viewport prediction, we propose **Behavior-Aware Temporal-Visual Viewport Prediction (BAT-VP)**. Fig. 5 illustrates the overall framework. It consists of four components: the Head & Eye Soft Alignment Module (SAM), Visual Graph Encoder (VGE), Behavior-Modulated Visual Attention (BMVA), and Temporal & Visual Fusion.

When a user starts watching a video, the system collects historical head and eye trajectories. To address the dynamic temporal offsets between eye and head-movements, SAM processes these

signals within a local temporal window and produces a unified motion representation for near-future prediction. Subsequently, VGE extracts local visual features from historical viewport contents and compresses them into a time-invariant user preference prototype, which is broadcast across all future prediction steps. BMVA combines future saliency information with head-motion-derived behavioral priors to generate user-level saliency representations, thereby complementing SAM with visual guidance as motion inertia weakens over longer horizons. Finally, the outputs of SAM, VGE, and BMVA are mapped into a shared feature space and assigned horizon-specific weights by the Temporal & Visual Fusion module, which gradually shifts prediction reliance from motion cues toward preference and visual cues to produce the final viewport prediction.

4.2 Visual Graph Encoder

To address the difficulty of capturing structured user preferences from raw historical sequences, we propose the Visual Graph Encoder (VGE), which models historical viewport contents as a user's visual preference graph. For a history of length L , viewport regions are sampled along the user's head trajectory. We employ ResNet50 [19] as the backbone network to extract a sequence of local visual features $\mathbf{F}_{1:L} = \{\mathbf{f}_t\}_{t=1}^L$ where $\mathbf{f}_t$ denotes the visual representation of the viewport region at time step t . To capture the structural relationships among historical viewport series, we organize this feature sequence as a graph structure for modeling. Specifically, each historical feature is treated as a graph node, and edges are derived based on cosine similarities, a widely adopted in visual representation similarity modeling [38], between nodes: $S_{ij} = \frac{\mathbf{f}_i^T \mathbf{f}_j}{\|\mathbf{f}_i\| \|\mathbf{f}_j\|}$. To suppress noisy correlations introduced by long-range temporal connections, we impose a temporal neighborhood constraint such that candidate edges are only established within a local window δ , and a similarity threshold ε to eliminate low-similarity pairs, and construct the adjacency matrix:

$$A_{ij} = \begin{cases} 1, & S_{ij} > \varepsilon \text{ and } |i - j| \leq \delta, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

We add self-connections to keep nodes' own features: $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$.

Then, we employ PageRank [28] on the constructed graph to estimate the centrality scores of historical nodes [23], thereby measuring the relative importance of each viewport segment within the overall viewing trajectory. Since later observations are typically more informative for future prediction, we introduce a temporal weight vector $\boldsymbol{\tau}$ and combine it with the graph centrality scores $\mathbf{r}$ to obtain the final fusion weights $\mathbf{w}_{1:L} = \{\mathbf{w}_t\}$:

$$\mathbf{w} = \text{Softmax}(\beta \mathbf{r} + \log(\boldsymbol{\tau})), \quad (2)$$

where β is a scaling coefficient. The resulting user visual preference prototype is given by $\mathbf{X}^{VGE} = \sum_{t=1}^L \mathbf{w}_t \mathbf{f}_t$.

4.3 Behavior-Modulated Visual Attention

As aforementioned, existing saliency-based approaches rely on generic saliency models that fail to incorporate personalized attention priors. To bridge this gap, inspired by insights from prior studies [6, 9, 21, 35], we propose Behavior-Modulated Visual Attention (BMVA), which dynamically generates user-level saliency maps for future frames based on historical motion behaviors.

Let the future prediction horizon be P . For each future time step $t \in \{1, \dots, P\}$, a saliency map is precomputed as $\mathbf{S}_t \in \mathbb{R}^{H \times W}$. Given the historical head trajectory $\mathbf{H}_{1:L}$, BMVA injects three complementary behavior priors as detailed below.

Inhibition of Return (IoR). To characterize the behavioral phenomenon that previously explored regions are temporarily suppressed, we construct a soft cumulative exploration map on the spherical grid based on the historical trajectory and apply Gaussian diffusion to obtain a normalized exploration intensity map $\mathbf{E}$. Then, the IOR factor is defined as

$$\mathcal{M}^{\text{IoR}} = 1 - \lambda_{\text{IoR}} \mathbf{E}, \quad (3)$$

where λ_{IoR} is a learnable coefficient. This term redistributes attention weights from recently visited regions to areas that are more likely to attract attention under the current behavioral context. Meanwhile, it enhances the saliency of previously unobserved hotspots while preserving residual influence from historical hotspots.

Directional Enhancement. Viewport movements usually exhibit strong motion continuity. To capture the directional characteristics of future attention transitions, we further compute directional consistency between each spherical grid location and the current motion direction, and construct a directional enhancement term

$$\mathcal{M}^{\text{DE}} = 1 + \lambda_{\text{DE}} \cdot \max(0, \cos \theta), \quad (4)$$

where θ denotes the angle between the grid direction and the current motion direction, and λ_{DE} is a learnable coefficient. This term assigns higher gains to regions aligned with the current motion direction and potential turning trend, enabling the model to consider which regions are more likely to become the next viewing targets.

Spatial & Temporal Decay. In reality, users tend to attend more to salient regions closer to the current viewport to reduce efforts of head-movements. Meanwhile, the impact of historical behaviors on long-term attention distribution weakens as the prediction step increases. To reflect these phenomena, we first define the spatial decay d as the distance from the current viewport and a temporal decay σ_t that exponentially decreases over time. Then, we propose the following term

$$\mathcal{M}_t^{\text{STD}} = \exp\left(-\frac{d^2}{2\sigma_t^2}\right), \quad (5)$$

Behavior-modulated saliency. By combining the above behavioral priors, the behavior-modulated saliency map at future step t is formulated as

$$\tilde{\mathbf{S}}_t = \mathbf{S}_t \cdot \mathcal{M}_t^{\text{IoR}} \cdot \mathcal{M}_t^{\text{DE}} \cdot \mathcal{M}_t^{\text{STD}} \quad (6)$$

Feature encoding. Finally, $\tilde{\mathbf{S}}_t$ is normalized and mapped into a unified semantic feature space via a convolutional encoder $g_{\text{cnn}}(\cdot)$:

$$\mathbf{X}_t^{\text{BMVA}} = g_{\text{cnn}}(\text{Norm}(\tilde{\mathbf{S}}_t)) \in \mathbb{R}^D. \quad (7)$$

The resulting sequence $\mathbf{X}^{\text{BMVA}}$ serves as the behavior-modulated visual input to the subsequent multi-source fusion module.

4.4 Head & Eye Soft Alignment Module

As revealed in the §3.2, eye-movements precede head-movements with a dynamic temporal offset. To excavate their collaborative information without introducing noise or information redundancy,

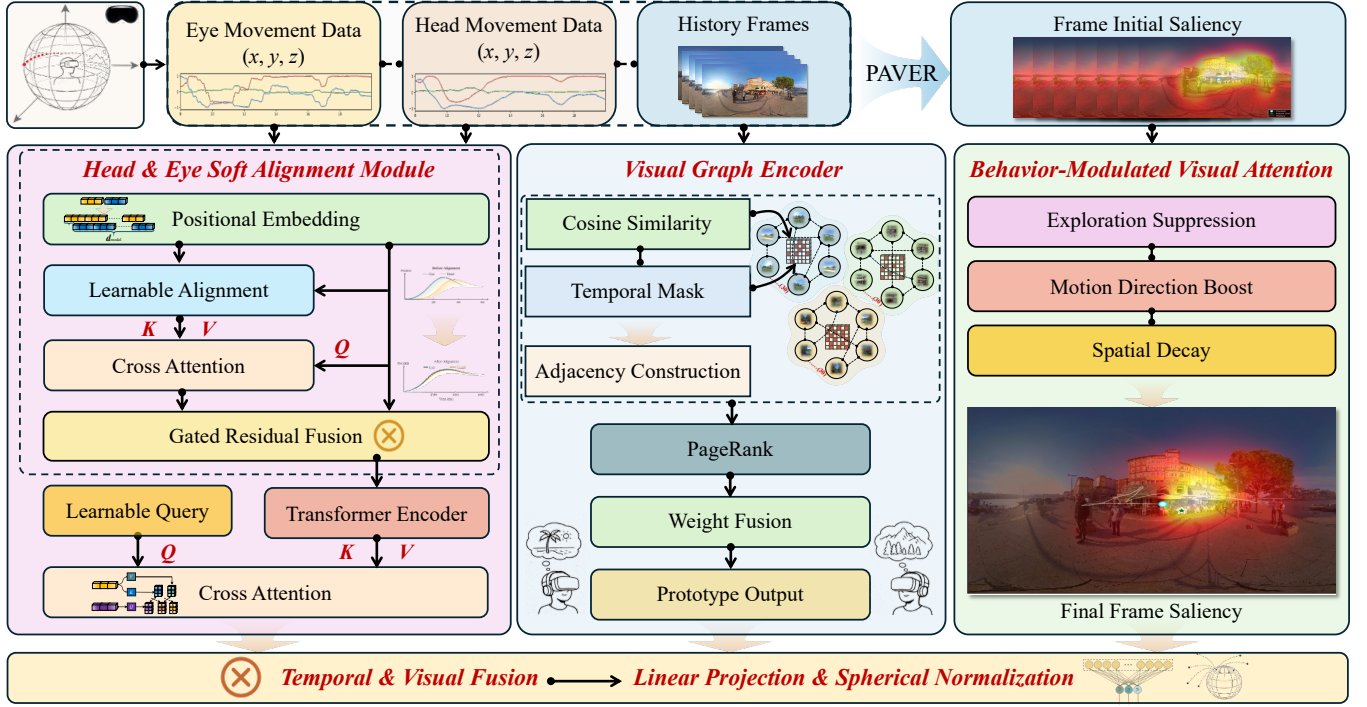


Figure 5: Overall architecture of BAT-VP.

we propose the Head & Eye Soft Alignment Module (SAM), which dynamically aligns head & eye movement representations within a local temporal window. Taking head motion as the dominant signal, SAM selectively injects informative eye-movements cues from the aligned eye-movements sequence, thereby forming a unified and temporally consistent motion representation.

Let the historical head and eye trajectories be denoted as $H_{1:L}$ and $E_{1:L}$, respectively, as three-dimensional unit vectors on the sphere. First, a linear projection followed by positional encoding is applied to map the two trajectories into a shared temporal feature space. Then, to characterize the dynamic temporal offset between head ($\mathbf{x}_t^h$) and eye signals ($\mathbf{x}_t^e$), we generate multiple eye-shift candidates within a limited local temporal window and compute alignment scores that optimize the offset value $a_{s,t} = f_{\text{align}}([\mathbf{x}_t^h; \mathbf{x}_{(s),t}^e])$. Then, the matching weight is defined as

$$w_{s,t} = \frac{\exp(a_{s,t})}{\sum_{s' \in S} \exp(a_{s',t})} \quad (8)$$

Accordingly, the aligned eye-movements representation is

$$\tilde{\mathbf{x}}_t^e = \sum_{s \in S} w_{s,t} \mathbf{x}_{(s),t}^e \quad (9)$$

Considering that head trajectories dominate the overall motion trend of the viewport, we employ head features as queries and the aligned eye features as keys and values, introducing a cross-attention mechanism for conditional interaction.

This representation is subsequently processed by a Transformer-based encoder-decoder for temporal modeling, which generates the final motion representation $\mathbf{X}^{\text{SAM}}$.

4.5 Temporal & Visual Fusion

Lastly, we unify representations from the three aforementioned modalities, i.e., $\mathbf{X}^{\text{VGE}}$, $\mathbf{X}^{\text{BMVA}}$, and $\mathbf{X}^{\text{SAM}}$, into the long-term viewport prediction. As they pose varying importance on the prediction horizon, we design a gating fusion module based on temporal information to adaptively adjust their contributions.

For the t -th future step, the per-modality gating weight is

$$g_t^k = \sigma\left(f_k([\mathbf{X}_t^k, \frac{t}{P}])\right), \quad k \in \{\text{SAM}, \text{VGE}, \text{BMVA}\} \quad (10)$$

where $\sigma(\cdot)$ is the Sigmoid activation function and $f(\cdot)$ is a light-weight feed-forward mapping. Then, the three gating coefficients are normalized as

$$\hat{g}_t^{(k)} = \frac{g_t^{(k)}}{g_t^{\text{SAM}} + g_t^{\text{VGE}} + g_t^{\text{BMVA}} + \varepsilon}, \quad k \in \{\text{SAM}, \text{VGE}, \text{BMVA}\} \quad (11)$$

where ε is a small constant for numerical stability.

Accordingly, the fused representation at step t is defined as

$$\mathbf{F}_t = \hat{g}_t^{\text{SAM}} \mathbf{X}_t^{\text{SAM}} + \hat{g}_t^{\text{VGE}} \mathbf{X}_t^{\text{VGE}} + \hat{g}_t^{\text{BMVA}} \mathbf{X}_t^{\text{BMVA}} \quad (12)$$

4.6 Loss Function Design

To ensure stable optimization of the proposed framework, we construct the objective function from the following perspectives:

Spherical Geometric Error. Since viewport directions lie on the unit sphere $\mathbb{S}^2$, we adopt the Orthodromic Distance (OD) as the fundamental geometric supervision $\mathcal{L}_1 = \frac{1}{P} \sum_{t=1}^P \arccos(\hat{\mathbf{y}}_t^\top \mathbf{y}_t)$

which penalizes the spherical angular deviation between the predicted direction $\hat{\mathbf{y}}_t$ and the ground-truth direction $\mathbf{y}_t$.

Motion Consistency Constraint. To preserve temporal continuity of the predicted trajectory, we further introduce first-order and second-order difference consistency losses, defined as $\mathcal{L}_2 = \|\Delta\hat{\mathbf{y}} - \Delta\mathbf{y}\|_2^2$ and $\mathcal{L}_3 = \|\Delta^2\hat{\mathbf{y}} - \Delta^2\mathbf{y}\|_2^2$, respectively.

Long-term Reinforcement. Considering that prediction errors tend to accumulate over longer horizons, we introduce an additional angular loss beyond t_0 , defined as $\mathcal{L}_4 = \frac{1}{P-t_0} \sum_{t=t_0}^P \arccos(\hat{\mathbf{y}}_t^\top \mathbf{y}_t)$

Temporal Gating Prior Regularization. To encourage temporal decay of the SAM's impact as the prediction horizon increases, we impose a soft prior on the temporal gating weights, defined as $\mathcal{L}_5 = \|\hat{g}_t^{\text{SAM}} - (1 - \alpha r_t)\|_2^2$, where $r_t = t/P$.

Overall Objective. The final loss function is defined as

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 + \lambda_4 \mathcal{L}_4 + \lambda_5 \mathcal{L}_5. \quad (13)$$

5 EVALUATION

5.1 Experimental Setup

5.1.1 Datasets. To evaluate the generalization ability and robustness of BAT-VP, we conduct experiments on two public datasets containing head & eye behavioral signals. Specifically, MMSys18 [14] is a widely used benchmark for long-term 360° viewport prediction, comprising 19 ERP videos of approximately 20 seconds viewed by 57 subjects under free-viewing conditions. Meanwhile, DSAV360 [4] is a larger-scale dataset containing 85 omnidirectional videos and 4,609 head & eye scanpaths from 87 subjects, with diverse content, complex scene dynamics, and substantial inter-subject variability. Moreover, to prevent user-level information leakage, we split users into training (60%), validation (10%), and test (30%) sets.

5.1.2 Compared Methods. We compare BAT-VP with five representative long-term viewport prediction baselines covering trajectory-only, content-aware, and multimodal fusion approaches. CVPR18-Improved [40] and TRACK [32] combine historical trajectories with visual saliency; EMD-ML [49] models multi-scale historical trajectories without visual content; MFTR [47] incorporates eye-movements through multimodal Transformer fusion; and STAR-VP [17] employs space-aligned and time-varying fusion.

To ensure fair and comparable evaluation, all experiments follow the unified evaluation framework introduced in [31]. All methods are reproduced or re-implemented with the same sampling frequency, history/future window setting, preprocessing pipeline, and evaluation metrics. Head and eye trajectories are uniformly resampled at 5 Hz and represented as 3D unit vectors on $\mathbb{S}^2$. Video saliency priors are generated offline using the official PAVER [44] implementation as a shared visual prior extraction pipeline.

5.1.3 Evaluation Metrics. We adopt two widely used metrics to evaluate prediction accuracy:

Orthodromic Distance (OD). The angular error at time step t between predicted viewport directions $\hat{\mathbf{y}}_t \in \mathbb{S}^2$ and ground-truth $\mathbf{y}_t \in \mathbb{S}^2$ is defined as $\text{OD}_t = \arccos(\text{clip}(\hat{\mathbf{y}}_t^\top \mathbf{y}_t, -1, 1))$.

Tile-based intersection-over-union (IoU). Let $\mathcal{T}(\cdot)$ be the set of tiles covered by the viewport. The IoU at time step t is defined as: $\text{IoU}_t = \frac{|\mathcal{T}(\hat{\mathbf{y}}_t) \cap \mathcal{T}(\mathbf{y}_t)|}{|\mathcal{T}(\hat{\mathbf{y}}_t) \cup \mathcal{T}(\mathbf{y}_t)|}$.

Lower OD and higher IoU suggest better performance. In addition to overall performance, we further divide the prediction horizon into short-term (0–2 s) and long-term (2–5 s) and analyze performance accordingly.

5.1.4 Implementation Details. We adopt a 6-second historical window to predict the subsequent 5 seconds, following typical settings in prior works [17, 32, 47]. Under 5 Hz sampling, the historical length is $L = 30$ and the prediction length is $P = 25$. The historical visual branch takes viewport-level local features as input, where each time step is represented by a 2048-dimensional feature vector. These features are extracted using a ResNet50 backbone [19] pretrained on ImageNet [15]; the final global average pooling and fully connected layers are removed, retaining convolutional feature maps only. The features are further projected onto a 9×16 spherical grid via adaptive pooling.

Video saliency priors are generated offline using the official PAVER implementation with publicly released pretrained weights. The original network structure is preserved. The predicted heatmaps are mapped to the same 9×16 grid and stored as compressed files for efficient loading during training. The saliency maps serve as fixed content priors and are not updated through backpropagation.

We adopt the following architectural hyperparameters: hidden dimension $D = 256$, # Transformer layers = 2, # attention heads = 2, and dropout rate = 0.25. Training is performed using the AdamW optimizer with an initial learning rate of 2×10^{-4} and weight decay of 0.05. A cosine annealing schedule is applied, with a minimum learning rate of 1×10^{-6} . The batch size is set to 32, and the model is trained for 30 epochs. Early stopping is adopted based on validation performance with a patience of 6 epochs. The loss weights are set as $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, $\lambda_3 = 0.05$, $\lambda_4 = 0.5$, and $\lambda_5 = 0.1$. The loss weights and module-specific hyperparameters were selected based on validation performance, while one-at-a-time perturbations around their default values caused only limited IoU variations, indicating low parameter sensitivity. All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 6000 Pro GPU.

5.2 Overall Performance

To intuitively illustrate the performance of BAT-VP, we visualize three representative sequences in Fig. 6. Specifically, the ground-truth and predicted viewport directions, represented as 3D unit vectors on the sphere, are projected onto the ERP frame. A fixed field-of-view (FoV) is used to indicate the viewport region, and the corresponding OD is reported.

It can be observed from Fig. 6 (left) that during short-term prediction ($t = 1\text{s}$), the predicted viewport highly overlaps with the ground-truth with a small OD. Although the error increases with the prediction horizon, the prediction remains spatially continuous and largely aligned with the true region of interest, demonstrating stable and geometrically plausible long-term prediction.

We also visualize the predicted trajectories vs ground-truth directions from three coordinates in Fig. 6 (right). We can observe that our predictions closely align with the ground-truth in the short-term interval and only slightly deviate in longer horizons. These results indicate that our system maintains high and robust performance across prediction horizons.

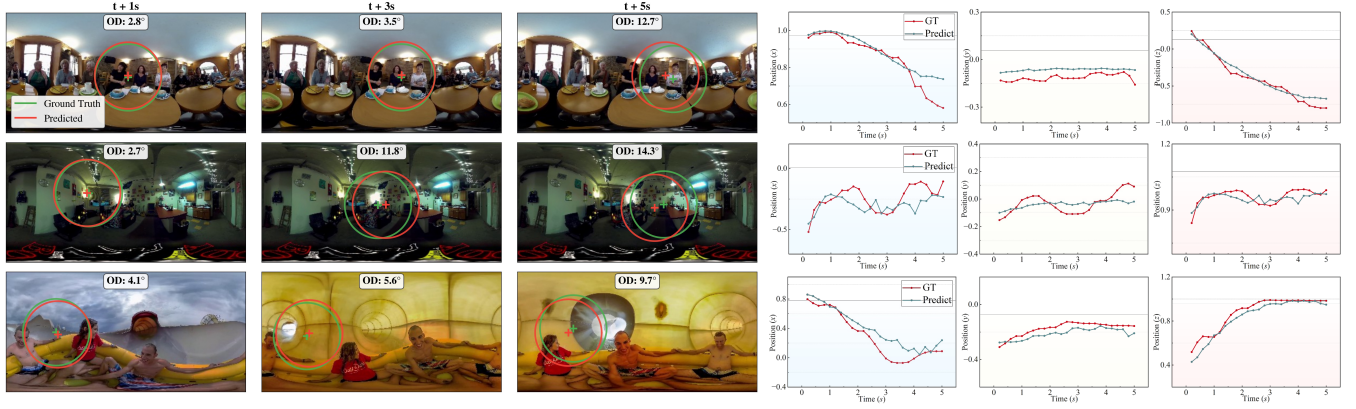


Figure 6: Qualitative visualization. Left: ERP overlays and the corresponding OD. Right: predicted and ground-truth trajectories.

Table 1: Performance comparison on the MMSys18 dataset. O = overall, S = short-term, L = long-term.

System	IoU-O	IoU-S	IoU-L	OD-O	OD-S	OD-L
CVPR18-Imp.	40.63	56.06	30.35	48.59	28.94	61.68
TRACK	42.76	59.11	31.86	47.74	27.47	61.25
EMD-ML	41.61	58.68	29.91	48.18	27.11	60.89
MFTR	41.57	56.61	31.55	51.99	31.34	65.76
STAR-VP	42.96	59.03	32.25	45.32	26.13	57.93
BAT-VP (Ours)	45.06	60.95	34.70	45.11	26.41	57.76

Table 2: Performance comparison on the DSAV360 dataset. O = overall, S = short-term, L = long-term.

System	IoU-O	IoU-S	IoU-L	OD-O	OD-S	OD-L
CVPR18-Imp.	48.87	63.80	38.91	41.93	24.44	53.59
TRACK	50.45	66.26	39.92	40.81	22.98	52.69
EMD-ML	48.58	66.90	37.36	41.72	22.89	53.61
MFTR	47.98	61.89	38.71	48.24	29.32	60.85
STAR-VP	47.55	59.44	39.62	40.93	26.20	50.75
BAT-VP (Ours)	54.17	68.13	44.85	38.03	21.84	48.83

5.3 Comparison Experiments

We compare the performance of BAT-VP against five baselines on two datasets mentioned in 5.1.2. First, we measure their prediction performance using IoU and OD over three temporal intervals: overall, short-term (0–2 s), and long-term (2–5 s). To assess statistical reliability, we further report 95% confidence intervals (95% CI) and p-values (p).

As shown in Tables 1–2, BAT-VP achieves the best overall performance on both datasets. On MMSys18, BAT-VP improves the overall IoU by 5.4%, 8.4%, and 8.3% over TRACK, MFTR, and EMD-ML, respectively. The advantage is more pronounced for long-term prediction, where BAT-VP reaches an IoU of 34.70, outperforming STAR-VP, the second best candidate, by 7.6%. This gain is statistically significant, with a 95% CI of 34.53–34.87 and $p = 1.75 \times 10^{-21}$.

Table 3: Overhead comparison. Params (M): model parameters; GPU-/CPU-Lat (ms): GPU/CPU latency; Mem (MB): memory at batch size 32; FLOPs (G): floating-point operations per second.

System	Params	GPU-Lat	Mem	CPU-Lat	FLOPs
CVPR18-Imp.	22.95	63.59	618.9	140.10	1.07
TRACK	172.57	226.46	2046.7	481.20	0.85
EMD-ML	3.10	4.47	53.3	8.35	0.18
MFTR	13.81	55.63	2386.4	72.18	2.60
STAR-VP	12.30	8.49	3268.2	53.05	3.25
BAT-VP (Ours)	6.24	9.10	74.3	9.15	0.50

Meanwhile, BAT-VP achieves OD comparable to STAR-VP across all intervals, indicating that improved viewport overlap does not increase geometric error.

On DSAV360, where user behaviors and video content are more diverse, the performance gaps are further enlarged. BAT-VP achieves an overall IoU of 54.17, outperforming TRACK by 7.4%. Its long-term IoU reaches 44.85, improving upon TRACK by 12.4%. Moreover, BAT-VP obtains the best overall and long-term OD of 38.03° (95% CI, 37.23°–38.83°) and 48.83° (95% CI, 47.91°–49.95°), respectively. Moreover, the improvements over STAR-VP are statistically significant, with $p = 7.65 \times 10^{-10}$ and $p = 2.31 \times 10^{-19}$, respectively. Overall, BAT-VP shows slower performance degradation over longer horizons while preserving short-term accuracy.

This improvement can be attributed to the complementary components of BAT-VP, which jointly improve user-adaptive visual guidance and behavioral motion modeling for long-term prediction. Compared with these baselines, BAT-VP better exploits head & eye behavioral cues through SAM, provides more user-adaptive visual guidance, and achieves stable user preference modeling.

Finally, we compare system overhead and deployability. As reported in Table 3, BAT-VP requires only 6.24M parameters, 74.3 MB batch memory, 9.15 ms CPU latency, and 0.50G FLOPs, while retaining a GPU latency of 9.10 ms for real-time inference. Although EMD-ML achieves the lowest overhead using only historical trajectories, its accuracy is substantially lower than BAT-VP across

all intervals. By contrast, TRACK and MFTR incur higher latency due to heavier temporal modeling and multimodal Transformer fusion, respectively. Although STAR-VP achieves slightly lower GPU latency, its visual branch requires substantially more parameters, memory, and FLOPs.

This efficiency advantage is primarily attributed to two design choices: (i) a lightweight temporal modeling backbone that avoids parameter growth typical of heavier architectures, and (ii) behavior-modulated saliency cues that remain spatially concentrated, enabling feature fusion at a compact 9×16 resolution. As a result, BAT-VP does not require high-resolution saliency inputs, which reduces computational overhead and simplifies visual integration.

5.4 Ablation Study

5.4.1 Module Ablation. To study the contribution of each key component, we conduct a module ablation study by removing each module from BAT-VP. The results are shown in Fig. 7.

By removing BMVA, we observe a performance drop of 4.7% in overall IoU and 6.8% in long-term IoU, accompanied by a noticeable increase in OD. Replacing the BMVA-modulated saliency map with a raw saliency map also leads to significant performance degradation. This indicates that generic visual saliency is insufficient for long-term prediction, as it fails to capture personalized viewing behaviors. In contrast, BMVA provides Behavior-Aware visual cues that better align with user attention, thereby improving prediction stability.

When removing the SAM module, the IoU decrease by 0.6–0.71, with correspondingly increasing OD. This is because SAM explicitly reduces the mismatch between eye and head-movements, allowing eye-movements to provide more reliable cues for future viewpoint changes and thus reducing error accumulation over time.

Finally, removing the VGE module also leads to performance degradation. This suggests that long-term prediction benefits from graph-based modeling of historical visual features to capture user-specific viewing preferences. The resulting personalized visual prior helps reduce prediction drift and improve long-term consistency.

5.4.2 Gate Analysis. We further analyze how the Temporal & Visual fusion module behaves across horizons. We analyze the evolution of the three fusion gate weights in Fig. 8. The three curves represent the mean normalized weights of the branches, and the shaded regions denote their standard deviations. It can be observed that SAM dominates during short-term prediction and gradually decreases as time progresses. In contrast, the weights of the VGE and BMVA steadily increase, indicating that the model progressively shifts from short-term motion reliance to stronger dependence on user preference and behavior-modulated visual cues in long-term prediction. Overall, this result provides intuitive evidence for the time-adaptive property of the proposed fusion mechanism.

5.4.3 Saliency Robustness. To evaluate dependence on the PAVER backbone, we further test BAT-VP under saliency perturbations and backbone replacement. Under a one-cell spatial shift of PAVER-generated saliency maps, long-term IoU remains 34.46 on MMSys18 and 44.01 on DSAV360, close to the clean results of 34.70 and 44.85. Replacing PAVER with CubePad [10] also maintains overall IoU at 44.86 and 53.51 on MMSys18 and DSAV360, indicating BAT-VP's robustness across the initial saliency backbones.

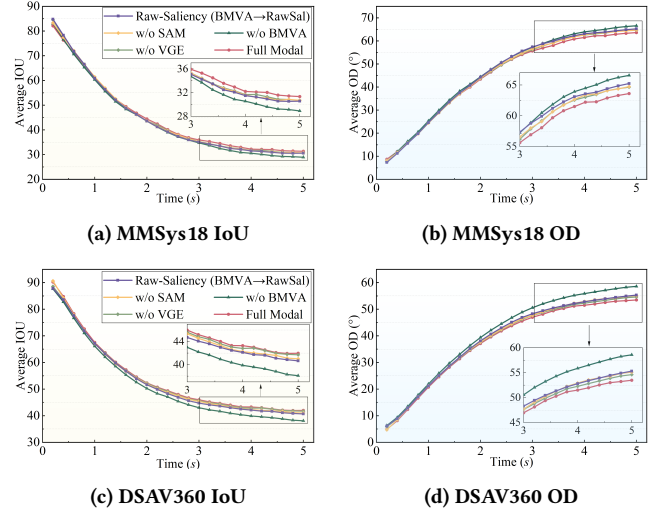


Figure 7: Performance under different ablation settings.

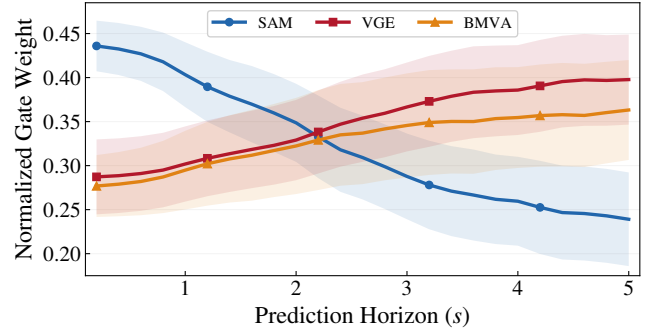


Figure 8: Visualization of gate fusion.

6 CONCLUSION

This paper proposes BAT-VP, a novel long-term viewport prediction system that jointly models visual preferences, personalized saliency, and head & eye behaviors. Specifically, VGE captures users' historical visual preferences through lightweight graph-based modeling, BMVA modulates generic saliency with historical behaviors to generate user-specific attention cues, and SAM softly aligns head and eye-movements to exploit their dynamic correlation. A time-aware multimodal gating strategy further balances temporal and visual cues across prediction horizons, improving long-term stability while preserving short-term accuracy. Evaluations on two public datasets show that BAT-VP consistently outperforms five mainstream baselines, achieving at least 4.9% and 7.6% gains in overall and long-term IoU, respectively. These gains are achieved with only 6.24 M parameters and 9.10 ms GPU latency, indicating practical deployability.

Acknowledgments

This work is partly supported by the National Science Foundation (NSF) under Grant Nos. 2343619, 2416872, 2315596, 2244219, 2146497, 2551417.

References

- [1] Jayasingam Adhuran and Maria G. Martini. 2024. Efficient viewport prediction and tiling schemes for 360 degree video streaming. In *Proceedings of the 15th ACM Multimedia Systems Conference*. 374–380. doi:10.1145/3625468.3653425
- [2] Duin Baek, Hangil Kang, and Jihoon Ryoo. 2020. SALI360: design and implementation of saliency based video compression for 360° video streaming. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 141–152. doi:10.1145/3339825.3391866
- [3] Yixuan Ban, Lan Xie, Zhimin Xu, Xinggong Zhang, Zongming Guo, and Yue Wang. 2018. CUB360: Exploiting Cross-Users Behaviors for Viewport Prediction in 360 Video Adaptive Streaming. In *2018 IEEE International Conference on Multimedia and Expo*. 1–6. doi:10.1109/ICME.2018.8486606
- [4] Edurne Bernal-Berdun, Daniel Martin, Sandra Malpica, Pedro J. Perez, Diego Gutierrez, Belen Masia, and Ana Serrano. 2023. D-SAV360: A Dataset of Gaze Scanpaths on 360° Ambisonic Videos. *IEEE Transactions on Visualization and Computer Graphics* 29, 11 (2023), 4350–4360. doi:10.1109/TVCG.2023.3320237
- [5] Nilotpal Biswas, Anamitra Mukherjee, and Samit Bhattacharya. 2024. “Are You Feeling Sick?” A Systematic Literature Review of Cybersickness in Virtual Reality. *Comput. Surveys* 56, 11, Article 284 (2024), 38 pages. doi:10.1145/3670008
- [6] Gaston Bujia, Melanie Sclar, Sebastian Vita, Guillermo Solovey, and Juan Esteban Kamienkowski. 2022. Modeling Human Visual Search in Natural Scenes: A Combined Bayesian Searcher and Saliency Map Approach. *Frontiers in Systems Neuroscience* 16 (2022), 882315. doi:10.3389/fnsys.2022.882315
- [7] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. 2021. Transformer-based Long-Term Viewport Prediction in 360° Video: Scanpath is All You Need. In *IEEE 23rd International Workshop on Multimedia Signal Processing*. 1–6. doi:10.1109/MMSP53017.2021.9733647
- [8] Jinyu Chen, Xianzhao Luo, Miao Hu, Di Wu, and Yipeng Zhou. 2021. Sparkle: User-Aware Viewport Prediction in 360-Degree Video Streaming. *IEEE Transactions on Multimedia* 23 (2021), 3853–3866. doi:10.1109/TMM.2020.3033127
- [9] Zhenzhong Chen and Wanjie Sun. 2018. Scanpath Prediction for Visual Attention using IOR-ROI LSTM. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 642–648. doi:10.24963/ijcai.2018/89
- [10] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. 2018. Cube Padding for Weakly-Supervised Saliency Prediction in 360° Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1420–1429. doi:10.1109/CVPR.2018.00154
- [11] Federico Chiariotti. 2021. A survey on 360-degree video: Coding, quality of experience and streaming. *Computer Communications* 177 (2021), 133–155. doi:10.1016/j.comcom.2021.06.029
- [12] Lovish Chopra, Sarthak Chakraborty, Abhijit Mondal, and Sandip Chakraborty. 2021. PARIMA: Viewport Adaptive 360-Degree Video Streaming. In *Proceedings of the Web Conference 2021*. 2379–2391. doi:10.1145/3442381.3450070
- [13] Xavier Corbillion, Gwendal Simon, Alisa Devlic, and Jacob Chakareski. 2017. Viewport-Adaptive Navigable 360-Degree Video Delivery. In *2017 IEEE International Conference on Communications*. 1–7. doi:10.1109/ICC.2017.7996611
- [14] Erwan J. David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A dataset of head and eye movements for 360° videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 432–437. doi:10.1145/3204949.3208139
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848
- [16] Xianglong Feng, Viswanathan Swaminathan, and Sheng Wei. 2019. Viewport Prediction for Live 360-Degree Mobile Video Streaming Using User-Content Hybrid Motion Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2, Article 43 (2019), 22 pages. doi:10.1145/3328914
- [17] Baoqi Gao, Daoxu Sheng, Lei Zhang, Qi Qi, Bo He, Zirui Zhuang, and Jingyu Wang. 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*. 5556–5565. doi:10.1145/3664647.3681268
- [18] Yichen Guo, Mai Xu, Lai Jiang, Xin Deng, Jing Zhou, Gaoming Chen, and Leonid Sigal. 2024. Proposal With Alignment: A Bi-Directional Transformer for 360° Video Viewport Proposal. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 11 (2024), 11423–11437. doi:10.1109/TCSVT.2024.3419910
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*. 770–778. doi:10.1109/CVPR.2016.90
- [20] Ming Jiang, Shi Chen, Jinhui Yang, and Qi Zhao. 2020. Fantastic Answers and Where to Find Them: Immersive Question-Directed Visual Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2980–2989. doi:10.1109/CVPR42600.2020.00305
- [21] Matthias Kümmeler, Matthias Bethge, and Thomas S. A. Wallis. 2022. DeepGaze III: Modeling Free-Viewing Human Scanpaths with Deep Learning. *Journal of Vision* 22, 5 (2022), 7. doi:10.1167/jov.22.5.7
- [22] Jie Li, Ling Han, Chong Zhang, Qiyue Li, and Zhi Liu. 2023. Spherical Convolution Empowered Viewport Prediction in 360 Video Multicast with Limited FoV Feedback. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19, 1, Article 3 (2023), 23 pages. doi:10.1145/3511603
- [23] Zhaowei Liu, Yihao Jiang, Rufe Gao, Jingli Liu, and Dong Yang. 2026. DA-DFGAS: Differentiable Federated Graph Neural Architecture Search with Distribution-Aware Attentive Aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 23963–23971. doi:10.1609/aaai.v40i28.39573
- [24] Yiyun Lu, Yifei Zhu, and Zhi Wang. 2022. Personalized 360-Degree Video Streaming: A Meta-Learning Approach. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3143–3151. doi:10.1145/3503161.3548047
- [25] Afshin Taghavi Nasrabadi, Aliehsan Samiei, and Ravi Prakash. 2020. Viewport prediction for 360° videos: a clustering approach. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*. 34–39. doi:10.1145/3386290.3396934
- [26] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. 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*. 1190–1198. doi:10.1145/3240508.3240669
- [27] Cagri Ozcinar, Julián Cabrera, and Aljosa Smolic. 2019. Visual Attention-Aware Omnidirectional Video Streaming Using Optimal Tiles for Virtual Reality. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2019), 217–230. doi:10.1109/JETCAS.2019.2895096
- [28] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report 1999-66. Stanford InfLab.
- [29] Stefano Petrangeli, Gwendal Simon, and Viswanathan Swaminathan. 2018. Trajectory-Based Viewport Prediction for 360-Degree Virtual Reality Videos. In *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality*. 157–160. doi:10.1109/AIVR.2018.00033
- [30] Minglang Qiao, Mai Xu, Zulin Wang, and Ali Borji. 2021. Viewport-Dependent Saliency Prediction in 360° Video. *IEEE Transactions on Multimedia* 23 (2021), 748–760. doi:10.1109/TMM.2020.2987682
- [31] Miguel Fabián Romero Rondón, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2020. A Unified Evaluation Framework for Head Motion Prediction Methods in 360° Videos. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 279–284. doi:10.1145/3339825.3394934
- [32] Miguel Fabián Romero Rondón, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2022. TRACK: A New Method From a Re-Examination of Deep Architectures for Head Motion Prediction in 360° Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2022), 5681–5699. doi:10.1109/TPAMI.2021.3070520
- [33] Ludvig Sidenmark and Hans Gellersen. 2019. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality. *ACM Transactions on Computer-Human Interaction* 27, 1, Article 4 (2019), 40 pages. doi:10.1145/3361218
- [34] Jashanjoy Singh Sidhu and Abdelhak Bentaleb. 2024. LCR360: Efficient Head Movement Prediction and Viewport Sharing in 360° Video Streaming. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 2084–2090. doi:10.1145/3636534.3696210
- [35] Wanjie Sun, Zhenzhong Chen, and Feng Wu. 2021. Visual Scanpath Prediction Using IOR-ROI Recurrent Mixture Density Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 6 (2021), 2101–2118. doi:10.1109/TPAMI.2019.2956930
- [36] Sam Van Damme, Maria Torres Vega, and Filip De Turck. 2022. Machine Learning Based Content-Agnostic Viewport Prediction for 360-Degree Video. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 2, Article 50 (2022), 24 pages. doi:10.1145/3474833
- [37] Haopeng Wang, Haiwei Dong, and Abdulmotaleb El Saddik. 2025. Immersive Multimedia Communication: State-of-the-Art on Extended Reality Streaming. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 7, Article 190 (2025), 33 pages. doi:10.1145/3721292
- [38] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5265–5274. doi:10.1109/CVPR.2018.00552
- [39] Chenglei Wu, Ruixiao Zhang, Zhi Wang, and Lifeng Sun. 2020. A Spherical Convolution Approach for Learning Long Term Viewport Prediction in 360 Immersive Video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 14003–14040. doi:10.1609/aaai.v34i01.7377
- [40] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. 2018. Gaze Prediction in Dynamic 360° Immersive Videos. In

- 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5333–5342. doi:10.1109/CVPR.2018.00559
- [41] Yutong Xu, Junhao Du, Jiahe Wang, Yuwei Ning, Sihan Zhou, and Yang Cao. 2024. Panonut360: A Head and Eye Tracking Dataset for Panoramic Video. In *Proceedings of the 15th ACM Multimedia Systems Conference*. 319–325. doi:10.1145/3625468.3652176
- [42] Qin Yang, Junni Zou, Kexin Tang, Chenglin Li, and Hongkai Xiong. 2019. Single and Sequential Viewports Prediction for 360-Degree Video Streaming. In *2019 IEEE International Symposium on Circuits and Systems*. 1–5. doi:10.1109/ISCAS.2019.8702654
- [43] Abid Yaqoob, Ting Bi, and Gabriel-Miro Muntean. 2020. A Survey on Adaptive 360° Video Streaming: Solutions, Challenges and Opportunities. *IEEE Communications Surveys & Tutorials* 22, 4 (2020), 2801–2838. doi:10.1109/COMST.2020.3006999
- [44] Heeseung Yun, Sehun Lee, and Gunhee Kim. 2022. Panoramic vision transformer for saliency detection in 360° videos. In *European Conference on Computer Vision*. 422–439. doi:10.1007/978-3-031-19833-5_25
- [45] Lei Zhang, Peng Chen, Cong Zhang, Cheng Pan, Tao Long, Weizhen Xu, Laizhong Cui, and Jiangchuan Liu. 2025. Optimizing Mobile-Friendly Viewport Prediction for Live 360-Degree Video Streaming. *IEEE Transactions on Mobile Computing* 24, 10 (2025), 10441–10455. doi:10.1109/TMC.2025.3571186
- [46] Lei Zhang, Weizhen Xu, Donghuan Lu, Laizhong Cui, and Jiangchuan Liu. 2022. MFVP: Mobile-Friendly Viewport Prediction for Live 360-Degree Video Streaming. In *2022 IEEE International Conference on Multimedia and Expo*. 1–6. doi:10.1109/ICME52920.2022.9859789
- [47] Zhihao Zhang, Yiwei Chen, Weizhan Zhang, Caixia Yan, Qinghua Zheng, Qi Wang, and Wangdu Chen. 2023. Tile Classification Based Viewport Prediction with Multi-modal Fusion Transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3560–3568. doi:10.1145/3581783.3613809
- [48] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. 2018. Saliency Detection in 360° Videos. In *Proceedings of the European Conference on Computer Vision*. 504–520. doi:10.1007/978-3-030-01234-2_30
- [49] Huiyi Zhou, Feng Zhao, and Chunhai Li. 2025. Multi-scale Historical Trajectory Decomposition for Viewport Prediction in 360-degree Videos. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 12, Article 352 (2025), 22 pages. doi:10.1145/3760533
- [50] Huadi Zhu, Chaowei Wang, Venkateshwar Reddy Darmanola, Hongbo Guo, Wenqiang Jin, and Ming Li. 2024. Bere: A Novel Video Recommender System for Virtual Reality Using Human Behavioral Signals. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 770–784. doi:10.1145/3636534.3690660
- [51] Yucheng Zhu, Guangtao Zhai, Yiwei Yang, Huiyu Duan, Xiongkuo Min, and Xiaokang Yang. 2022. Viewing Behavior Supported Visual Saliency Predictor for 360 Degree Videos. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2022), 4188–4201. doi:10.1109/TCSVT.2021.3126590