Redefining 360° Video Navigation: A Novel Benchmark for Enhanced Immersive and Automated Exploration

Xinyu Liu†, Mengke Song†, Ziming Wang, Chenglizhao Chen*, Aimin Hao, Shuai Li*, Hong Qin

Abstract—360° video navigation aims to model users' viewing behavior in immersive videos through continuous viewpoint selection. This task is challenging because human viewing is not driven by instantaneous responses to local saliency, but by continuous decisions that evolve with semantic events over time, involving attention shifts across events and long-term viewpoint changes. However, progress remains limited, mainly due to the lack of a well-established benchmark, resulting in limitations in data, annotation, and evaluation. Current datasets are limited in scale and diversity, and mostly consist of short or single-event videos, making cross-event viewing behavior underrepresented. Existing annotation methods rely on frame-level or short-term clustering, which emphasizes local consistency but fails to capture temporal continuity and the rationality of viewpoint selection during event transitions. Existing metrics focus on spatial overlap or path consistency, ignoring motion-direction agreement between predicted and ground-truth viewport trajectories. As a result, spatially accurate predictions may still follow incorrect motion directions, leading to unreliable evaluation of viewing behavior. To address these issues, we construct a new benchmark for 360° video navigation in multi-event scenarios, with a standardized data and annotation pipeline and a more discriminative evaluation protocol. We further develop an event-level model to capture cross-event viewing behavior and long-term viewpoint evolution. Extensive experiments show that the proposed benchmark provides more reliable evaluation, and the proposed model achieves superior performance to existing methods.

Index Terms—Virtual Reality, Automatic Viewport Selection, Global Semantic Awareness, 360° Video Navigation

1 Introduction

With the rapid advancement of virtual reality (VR) technologies and related devices, 360° videos have emerged as a promising medium for immersive experiences through omnidirectional content capture. However, existing viewing paradigms still suffer from fundamental limitations, preventing smooth and comfortable viewing in practice.

Currently, three primary approaches exist for viewing 360° videos. Head-mounted displays (HMDs) [1], [2] provide the most immersive experience, enabling viewpoint exploration via head or body movements, but require continuous interaction and can cause fatigue. Equirectangular projection (ERP) [3], [4] maps the full 360° content onto a 2D plane, offering a global view but introducing severe distortion near the poles, degrading visual quality. Normal field-of-view (NFOV) projection [5], [6], widely used on platforms such as YouTube and Facebook, allows viewpoint adjustment via dragging or swiping. Although it reduces distortion, it relies on frequent manual interaction, which may distract attention and cause important content to be missed.

- *Xinyu Liu, Mengke Song, Ziming Wang, and Chenglizhao Chen are with the College of Computer Science and Technology, and Qingdao Institute of Software, China University of Petroleum (East China), Qingdao, China. (E-mail: liuxy001005@163.com; songsook@163.com; wzmmzw159@163.com; cclz123@163.com).*
- *Aimin Hao and Shuai Li are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China. (E-mail: ham@buaa.edu.cn; lishuai@buaa.edu.cn).*
- *Hong Qin is with the Department of Computer Science, Stony Brook University, Stony Brook, NY, USA. (E-mail: qin@cs.stonybrook.edu).*
- *†Xinyu Liu and Mengke Song contributed equally to this work.*
- **Chenglizhao Chen and Shuai Li are the corresponding authors.*

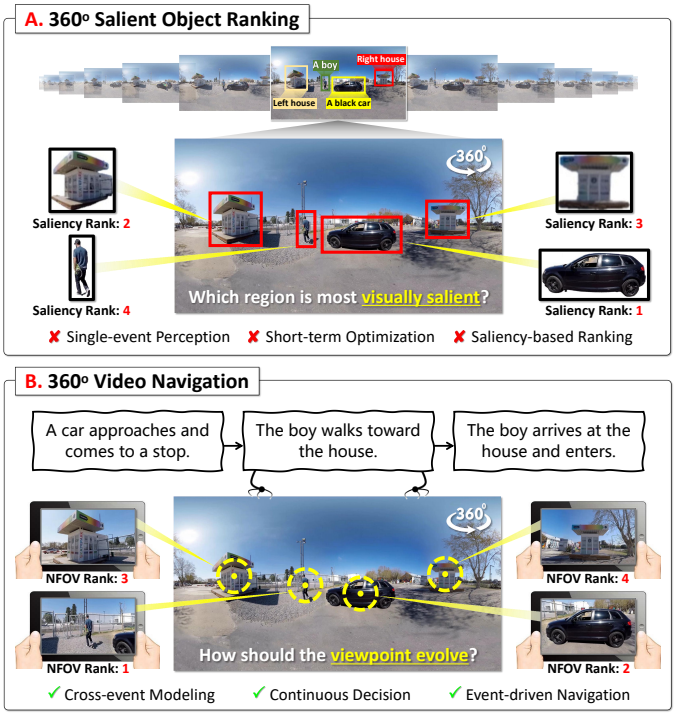


Fig. 1. Comparison between 360° salient object ranking and 360° video navigation. 360° salient object ranking **A** ranks objects by saliency to find those likely to draw attention. 360° video navigation **B** adjusts viewpoints to guide immersive exploration based on narrative flow.

To reduce the operational burden on users during video consumption, automated 360° video navigation has been proposed [7]–[9], aiming to select and recommend informative viewpoints in panoramic environments [10]–[13]. However,

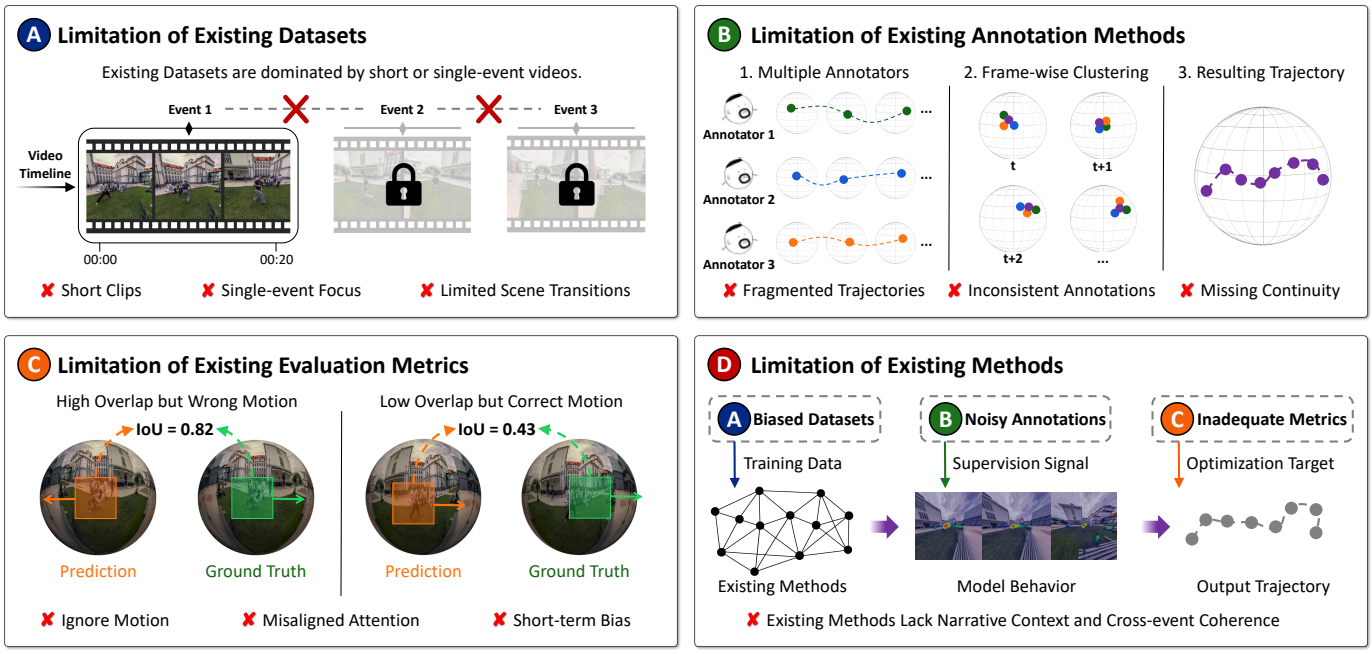


Fig. 2. Limitations of existing 360° video navigation. **A** Biased datasets lack cross-event coverage. **B** Noisy annotations yield fragmented trajectories. **C** Overlap metrics ignore motion dynamics. **D** These jointly bias models toward short-term decisions, hindering cross-event coherence.

from a holistic perspective, existing studies largely oversimplify this task as short-term viewpoint selection driven by local visual cues, i.e., 360° salient object ranking [14]–[17] (Fig. 1). This bias is not incidental, but arises from the combined effects of data, annotation, and evaluation [18]–[22], as illustrated in Fig. 2. At the data level, current datasets are dominated by single-event or short clips, lacking coverage of cross-event viewing behavior, preventing learning continuous decision patterns. At the annotation level, the absence of unified protocols and reliance on subjective labeling or local aggregation make it difficult to consistently capture complete viewing processes. At the evaluation level, existing metrics are primarily based on spatial overlap and fail to characterize dynamic viewing behavior, encouraging models to optimize local alignment. Together, these factors reinforce a local, short-term modeling paradigm and weaken cross-event continuous viewing behavior.

Therefore, without a unified framework across data, annotation, and evaluation to capture event-driven continuous viewing behavior, existing approaches remain confined to local, short-term decision paradigms and fail to represent cross-event dynamics in realistic scenarios. For example, in multi-event videos, models typically select locally salient regions within the field of view but fail to adjust viewpoints during event transitions, leading to missed key content or delayed attention shifts. As a result, even when predictions are spatially close to annotations, they may not reflect the underlying logic of continuous viewing, limiting performance gains and undermining reliability. This highlights the need for a unified modeling and evaluation framework, which remains a critical challenge in 360° video navigation.

To address these issues, we propose an event-level benchmark for 360° video navigation that jointly models data, annotation, and evaluation. Aligning these components with continuous viewing behavior establishes a consistent framework for supervision and evaluation.

For data and annotation, we adopt a standardized, perception-guided protocol representing viewing behavior at the event level, mitigating fragmentation of frame-level annotations. Based on this, we construct a multi-event dataset to capture cross-event viewing patterns. For evaluation, we propose a Motion-Adjusted Overlap (MAO) metric, which augments spatial overlap with motion consistency to assess spatial alignment and motion-direction agreement between predicted and ground-truth viewport trajectories.

Building upon this benchmark, we propose an event-level navigation model that fuses local visual information with global event semantics. By modeling cross-event dynamics and long-term viewpoint evolution, the model produces coherent and semantically consistent viewpoint transitions aligned with content progression.

Our framework lays a solid foundation for future research and advances more natural and seamless 360° video navigation, further enhancing immersive viewing experiences. Code and dataset are publicly available online.[1] The main contributions of this work are summarized as follows:

- We provide the first in-depth review of 360° video navigation, covering datasets, evaluation metrics, and models, and highlighting key challenges.
- We propose a standardized dataset annotation approach for 360° video navigation, introduce a dataset of 182 360° videos with diverse navigation patterns, and present a motion-consistency metric to assess navigation fluidity and alignment with video narratives.
- We develop a model that seamlessly integrates local saliency with global narrative context, thereby ensuring coherent and context-aware navigation paths.
- We present a comprehensive work encompassing a data annotation method, a dataset, an evaluation metric, and a network. The methodologies, discoveries, detailed analysis, and conclusions collectively provide valuable contributions to the research community.

[1] https://github.com/liuxy1005/PVN-Benchmark

TABLE 1
Comparison of existing 360° video navigation datasets. Attributes include collection equipment, ground-truth format, trajectory diversity, trajectory check, trajectory aesthetic evaluation, number of categories, videos, annotators, total duration, frames, and open-source availability. Compared to existing datasets, the proposed dataset (Ours) demonstrates clear advantages in **trajectory diversity, trajectory aesthetic evaluation, and the number and variety of video categories**. It provides more realistic and richer user interaction scenarios, making the navigation results more aligned with natural human behavior, thereby better supporting dynamic viewpoint adjustment and path optimization in 360° video navigation.

Datasets	Collection Equipment	Ground-truth	Trajectory Diversity	Trajectory Check	Aesthetic Evaluation	Categories	Videos	Annotators	Duration(h)	Frames	Open-source
Pano2Vid [2016]	Keyboard & Mouse	Viewport Coordinates	✓	✘	✘	4	86	3	7.3	-	✓
Sports360 [2017]	Keyboard & Mouse	Viewport Coordinates	✘	✓	✘	5	342	5	-	180K	✓
YouTube/Vimeo [2018]	-	-	✘	✘	✘	2	115	1	72	-	✘
Lee et al. [2018]	-	-	✘	✘	✘	5	285	1	92	-	✘
PVS-HM [2019]	Head-Mounted Display	Head Movement/Eye Fixation	✘	✘	✘	8	76	58	0.57	-	✓
Wu et al. [2017]	Head-Mounted Display	Head Movement	✘	✘	✘	5	18	48	1.5	-	✓
Nasrabadi et al. [2019]	Head-Mounted Display	Head Movement/Eye Fixation	✘	✘	✘	15	28	60	-	-	✓
Ours	Keyboard & Mouse	Viewport Coordinates	✓	✓	✓	8	182	12	1.63	176K	✓

2 Review of Existing Datasets, Models and Metrics

2.1 Existing Datasets

360° video navigation datasets are crucial for training and evaluating algorithms for navigating panoramic environments. These datasets typically include omnidirectional videos with continuously changing perspectives. However, due to the high cost of manual annotation, many rely on a single annotator, limiting data diversity and the ability to train robust models. Below, we summarize several widely used 360° video navigation datasets, as shown in Table 1.

Pano2Vid [8] is a widely used dataset consisting of 86 360° panoramic videos sourced from YouTube across four categories: soccer, mountain climbing, parades, and hiking. It also includes 343 hours of NFOV YouTube videos for training classifiers. Among the panoramic videos, 40 were annotated by three human editors, yielding 480 navigation trajectories with a total duration of approximately 717.2 minutes.

The Sports-360 dataset [7] comprises 342 360° videos from YouTube, covering five dynamic sports: basketball, parkour, BMX, skateboarding, and dance. These videos were selected based on the presence of identifiable human objects and dynamic activities, ensuring clear viewing angles for navigation. Five human annotators marked the most salient objects in each video, with annotations verified by multiple raters. The dataset includes annotated video segments with ground truth viewing angles for model training and testing.

The YouTube/Vimeo dataset [23] is designed for highlight detection in 360° videos. It contains 115 360° videos, primarily from wedding and music video genres. Videos are subdivided into multiple segments, with manual annotations marking important moments or "highlights" for processing. This dataset is valuable for researchers working on highlight detection and content analysis in immersive environments.

Lee et al. [24] collected a dataset of 285 360° videos, covering scenarios such as weddings, parades, boating, scuba diving, and hot air ballooning. The total duration is approximately 92 hours. Each video is accompanied by human-generated summaries, marking key events crucial to the narrative. This dataset is used for evaluating temporal summarization models and story-based video analysis.

PVS-HM database [25] is designed to study head movement and attention distribution in 360° videos. It contains 76 video sequences along with head movement (HM) data from 58 participants, who wore eye-tracking devices while watching the videos. The dataset records the trajectory of head movements and gaze points, providing valuable insights into user attention patterns in immersive environments.

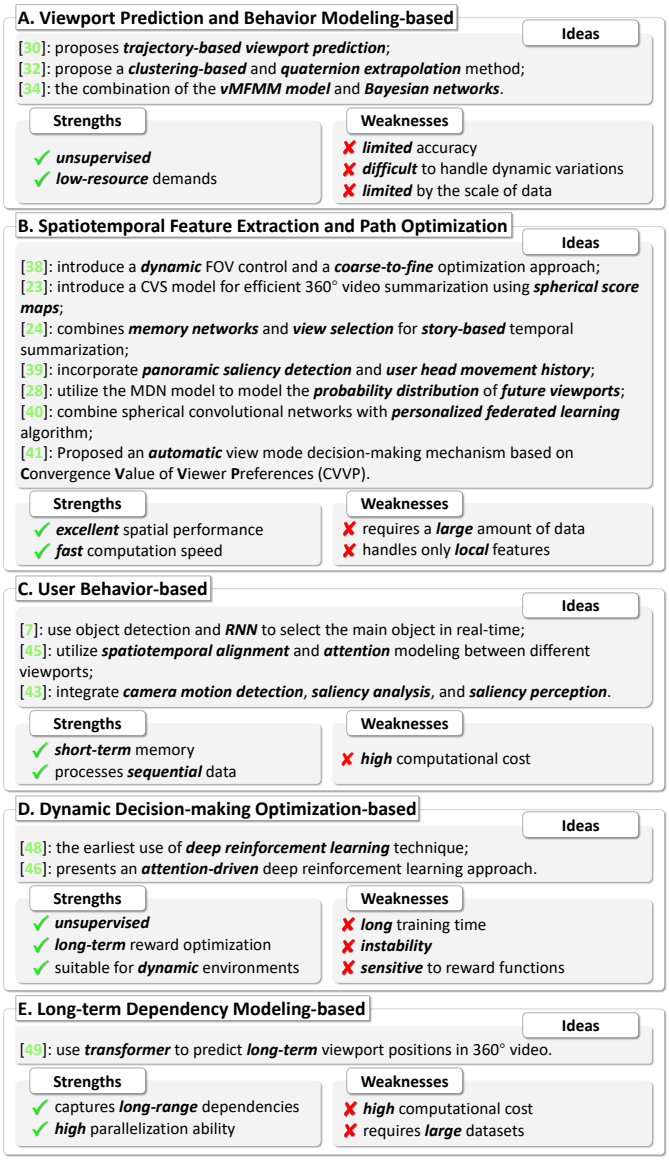


Fig. 3. The overall summary of existing automatic methods.

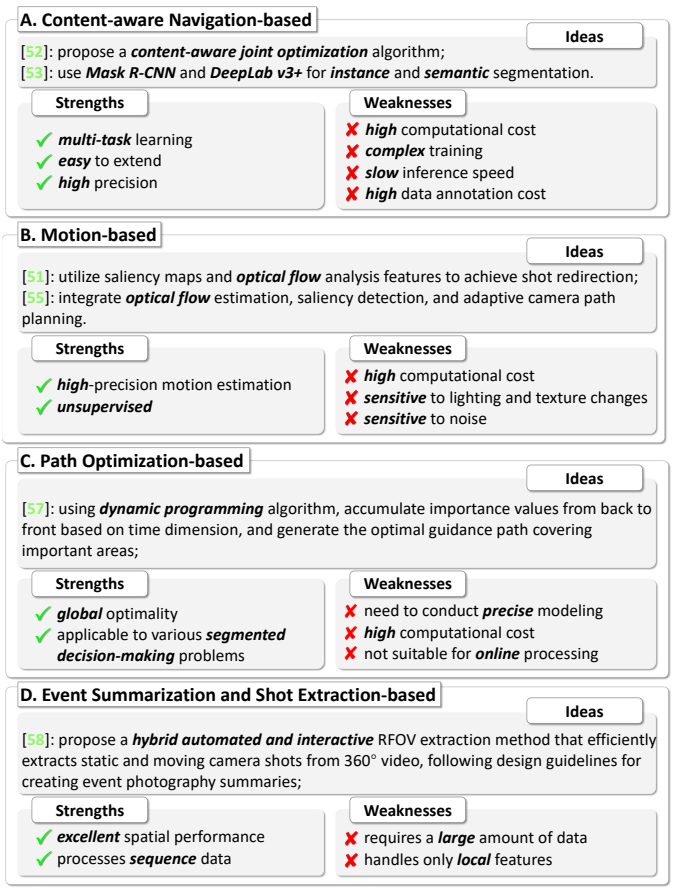


Fig. 4. The overall summary of existing interactive methods.

The dataset by Wu et al. [26] explores user behavior in spherical videos using HTC Vive headsets. It includes 18 videos spanning categories such as basketball and skiing, with participants' head movements and other data recorded during playback. The dataset provides metadata, including timestamps, head orientations, and HMD positions, helping researchers analyze user behavior in virtual environments.

Nasrabadi et al. [27] developed a dataset that combines gaze tracking data with responses from questionnaires filled out by participants watching 360° videos. The dataset includes gaze heatmaps and video samples, enabling in-depth analysis of viewer behavior and experiences. It also serves as a foundation for developing gaze prediction algorithms.

The PanoUCF101 dataset [28] adapts the UCF101 action recognition dataset [29] to 360° videos. It consists of 101 action categories, with each video clip projected into a 360° viewport. The dataset is divided into training, validation, and test sets, providing a standardized benchmark for 360° video navigation and action recognition tasks.

Summary. While these datasets have advanced 360° video navigation research, they have notable limitations. Most rely on a small number of annotators, which limits diversity and generalizability. Annotation methods are often subjective and inconsistent, making cross-study comparisons difficult. Additionally, many datasets are narrow in scope and fail to capture the full range of real-world user behaviors. These challenges emphasize the need for larger, more diverse, and standardized datasets to develop robust 360° video navigation systems. A summary of widely used 360° video navigation datasets is presented in Table 1.

2.2 Existing Models

The use of 360° video navigation systems has led to the development of two key approaches: automatic and interactive methods. These approaches address the challenges of the "Pano2Vid" problem, which involves converting panoramic footage to conventional video formats. This section explores both methodologies, analyzing their strengths, limitations, and practical applications. It aims to provide a comprehensive overview of current "Pano2Vid" solutions and offer insights into future developments in the field.

2.2.1 Automatic

Automated 360° video navigation methods use intelligent algorithms to optimize video perspectives for a smoother, more personalized viewing experience. These methods include viewport prediction, user behavior modeling, spatiotemporal feature extraction, path optimization, and personalized viewpoint adjustments based on historical data and user interactions. They can be categorized into five approaches.

Firstly, viewport prediction and behavior modeling methods [30]–[36] analyze users' past viewing trajectories to predict future positions, ensuring key content remains in the viewport. By learning from head movements and historical patterns, these algorithms preload relevant areas to prevent delays (Fig. 3-A). Recent work has also explored audio-visual saliency prediction in 360° videos to model viewer attention over omnidirectional content [37], providing complementary perceptual cues for identifying attention-worthy regions. Secondly, spatiotemporal feature extraction and path optimization methods [23], [24], [28], [38]–[41] (Fig. 3-B) analyze 360° videos to create optimal viewing paths based on dynamic objects, movement trajectories, and attention shifts. These methods adapt in real time to ensure important content is always visible, preventing users from missing significant moments. Thirdly, user behavior-based viewpoint prediction methods [7], [30], [42], [43] (Fig. 3-C) adjust perspectives in real time based on users' viewing habits and historical data. By modeling viewpoint-switching patterns, these methods automatically offer the most relevant views, enhancing the viewing experience and eliminating manual adjustments [44], [45]. Fourthly, dynamic decision-making optimization methods [25], [46], [47] (Fig. 3-D), using reinforcement learning [48], refine the viewing path based on user feedback. These methods continuously adapt the viewpoint, ensuring the user always sees the most relevant content in real time. Finally, long-term dependency modeling methods [49], [50] (Fig. 3-E) use self-attention mechanisms to capture long-term patterns in viewpoint changes, ensuring consistent and coherent navigation during extended sessions.

Summary. Although automated 360° video navigation methods offer advantages in simplifying user interaction and enhancing the viewing experience, they still have some drawbacks. Firstly, these methods demand high computational resources, especially for real-time prediction and dynamic optimization, which can lead to latency and increased hardware requirements. Secondly, they struggle with tracking accuracy when handling fast-moving or complex scenes, particularly in dynamic content. Further, personalized recommendations are limited by the diversity and quality of the data, making it difficult to meet the needs of different users. Finally, over-

TABLE 2
Comparison of evaluation metrics for 360° video navigation, highlighting their abbreviations, descriptions, formulas, advantages, and limitations. These metrics assess viewport accuracy, trajectory dynamics, saliency-map consistency, and motion direction. Each metric captures different aspects of predicted and ground-truth navigation paths. The proposed Motion-Adjusted Overlap (MAO) combines **spatial overlap and motion consistency**, enabling a more comprehensive evaluation of spatial and temporal trajectory alignment.

Metrics	Abbr.	Description	Formula	Advantages	Disadvantages
Mean Overlap	MO	Measures overlap between predicted and actual viewports.	$\mathrm{MO} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IoU}\,(l_t, l_t^{gt})$	Directly measures viewport accuracy; size-independent.	Ignores viewport direction and temporal continuity.
Mean Velocity Difference	MVD	Measures the curvature and smoothness of the predicted viewing-angle trajectory.	$\mathrm{MVD} = \frac{1}{T-2}\sum_{t=3}^{T} \parallel v_t - v_{t-1} \parallel_2$	Captures trajectory curvature and transition smoothness.	Ignores spatial accuracy and ground-truth alignment.
Mean Cosine Similarity	MCS	Assesses trajectory accuracy via cosine similarity of key viewpoints.	$\mathrm{MCS} = \frac{1}{N}\sum_{i=1}^{N} \frac{v_i \cdot u_i}{\parallel v_i \parallel \parallel u_i \parallel}$	Direct similarity measurement.	Ignores the temporal aspect.
Normalized Scanpath Saliency	NSS	Measures the average normalized saliency value at human fixation locations.	$\mathrm{NSS}(P, Q) = \frac{1}{N}\sum_{i} \frac{P_i - u_p}{\sigma_p} \times Q_i$	Standardized and directly evaluates fixation alignment.	Requires fixation annotations and ignores temporal order.
Pearson's Correlation Coefficient	CC	Evaluates correlation between predicted significance maps and user fixation heatmaps.	$\mathrm{CC}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)}$	Sensitive to symmetry and direction; evaluates global consistency.	Lacks local detail sensitivity.
Kullback-Leibler divergence	KL	Measures the divergence between predicted and ground-truth saliency distributions.	$\mathrm{KL}(P \parallel Q) = \sum_{i} P(i) \log\left(\frac{P(i)}{Q(i)}\right)$	Quantifies distributional mismatch between saliency maps.	Asymmetric and sensitive to zero probabilities.
Great Circle Distance	GCD	Measures the angular distance between two viewport centers on a sphere.	$\mathrm{GCD} = \arccos(\sin(\phi_1) \cdot \sin(\phi_2) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \cos(\lambda_1 - \lambda_2))$	Provides accurate spherical measurements; handles edge cases on the sphere.	Ignores time and only measures distance, not directional differences.
Motion Consistency	MC	Measures consistency between predicted and actual trajectory motion directions at adjacent points.	$\mathrm{MC}_{t\to t+1} = \mathrm{Cos}(\mathbf{m}_{t\to t+1}, \hat{\mathbf{m}}_{t\to t+1})$	Considers motion direction and avoids false high scores for incorrect directions.	Complex calculation; needs combination with overlap metrics.
Motion-Adjusted Overlap	MAO	Combines viewport overlap with motion consistency to evaluate navigation quality in dynamic 360° videos.	$\mathrm{MAO} = \frac{1}{T-1}\sum_{t=1}^{T-1} S^{\mathrm{overlap}} \cdot \mathrm{norm}\,(\mathrm{MC}_{t\to t+1})$	Evaluates both spatial and motion alignment.	Computationally more complex and sensitive to motion tracking errors.

reliance on historical data may result in a lack of flexibility, failing to adapt to real-time shifts in user interests.

2.2.2 *Interactive*

Interactive 360° video navigation [51] allows users to engage actively with content by adjusting their viewpoint, field of view (FoV), and perspective during playback. Through interactions such as mouse dragging, touchscreen swiping, or specialized controllers, users can explore immersive video environments and tailor the experience to their preferences. Some systems even enable users to influence video content directly. This method can be divided into four categories.

Firstly, the content-aware navigation [52]–[54] relies on image segmentation to identify important areas in the video, such as people or objects. It dynamically adjusts the view in response to user inputs, ensuring key content remains prominent (Fig.4-A). This approach enhances personalization and optimizes viewing paths but demands significant computational resources for accurate segmentation and path planning. Secondly, motion-based adjustment [51], [55], using optical flow [56], analyzes motion between consecutive frames to dynamically shift viewpoints. By tracking motion changes, such as object displacement or human movement, it ensures users stay focused on dynamic elements (Fig. 4-B). Effective for fast-changing scenes like sports, it efficiently handles motion but can struggle with low-resolution, complex backgrounds, or lighting changes. Thirdly, path optimization for seamless navigation [57] calculates optimal viewing paths to smoothly transition between viewpoints while focusing on important content (Fig.4-C). Using global path planning and visual saliency, this method enhances the navigation experience but requires intensive computational resources, making real-time implementation challenging. Lastly, event summarization and shot extraction [58] identifies key events or segments within videos using computer vision, summarizing them into short, navigable shots (Fig.4-D). This approach is ideal for lengthy or content-rich videos, reducing viewing time and simplifying navigation. However, it may struggle with capturing full context or handling complex dynamic scenes.

Summary. The four methods of interactive 360° video navigation each offer unique advantages but share common challenges. They often require high computational resources, which can lead to delays or lags, affecting the user experience. Additionally, content-aware and motion-adjustment methods may struggle with complex scenes, low-quality videos, or environmental factors like lighting changes, causing inaccuracies in viewpoint adjustments. Path optimization and event-extraction methods, while providing smooth navigation and simplified content, may miss important details or context, potentially impacting the overall viewing experience.

2.3 Existing Evaluation Metrics

Evaluation metrics are crucial for assessing algorithm performance, especially in tasks like 360° video navigation, where user experience and content effectiveness are paramount. This section introduces commonly used metrics for 360° video navigation, as summarized in Table 2.

The Mean Overlap (MO) is a key metric for evaluating predicted viewport accuracy in 360° video navigation [8]. It measures overlap between predicted and ground truth viewports for each frame, with the final score as the average. Higher MO values (closer to 1) indicate better accuracy, providing a useful measure of prediction quality.

The Mean Velocity Difference (MVD) evaluates the smoothness of a predicted trajectory by measuring changes between consecutive viewing-angle velocities [7]. A lower MVD indicates smoother viewpoint transitions, whereas a higher value reflects sharper trajectory curvature.

The Mean Cosine Similarity (MCS) measures similarity between the model's predicted trajectory and the manually annotated trajectory. It computes cosine similarity between key segments of both trajectories, with higher values indicating better alignment. MCS is valuable for evaluating how closely the model's path follows the human-annotated path.

The Normalized Scanpath Saliency (NSS) [59] evaluates a standardized saliency map at fixation locations. It averages the normalized saliency values at fixated pixels, with higher scores indicating better fixation alignment.

Pearson's Correlation Coefficient (CC) [60] quantifies the linear relationship between the predicted saliency map and the ground truth user fixation heatmap. A high CC suggests predicted and actual gaze patterns are closely related, which is useful for evaluating saliency prediction models.

Kullback-Leibler (KL) divergence measures the discrepancy between predicted and ground-truth saliency distributions. A lower KL divergence indicates closer agreement.

Great Circle Distance (GCD) [61] measures the shortest angular distance between predicted and ground-truth viewport centers on the unit sphere. A smaller GCD indicates more accurate viewport localization.

Summary. Previous studies have shown that different evaluation metrics may emphasize different aspects of prediction quality and consequently lead to different assessments of model performance [62]. While these metrics are widely used, they have limitations in the context of 360° video navigation. Most focus on accuracy, such as overlap or similarity, but fail to capture user experience or the dynamic nature of navigation. For instance, MO and MVD respectively characterize spatial overlap and trajectory smoothness, but neither fully captures semantic coherence, motion alignment, or immersion. Additionally, saliency-based metrics (e.g., NSS, CC, and KL) rely on gaze tracking data, which can be noisy or incomplete. The dependence on a single ground truth (often human-annotated) can also introduce bias, as different annotators may interpret paths or saliency differently. These limitations highlight the need for more comprehensive evaluation frameworks that better capture the complexities of 360° video navigation and user interaction.

3 Main Course

3.1 Event-level Annotation Framework for 360° Video Navigation

3.1.1 Why Do We Need This New Annotation Approach?

Existing annotation approaches fail to capture the event-driven and continuous nature of human viewing behavior in 360° videos. Most existing datasets and protocols are built upon frame-level or short-term observations, assuming that viewing decisions are dominated by local visual cues. However, this assumption contradicts immersive viewing, where attention evolves across semantic events and involves long-term viewpoint transitions. From a modeling perspective, current annotation strategies suffer from three key limitations. First, annotations collected via head-mounted displays (HMDs) are constrained by limited field of view, restricting global perception and leading to incomplete observation of event transitions. Second, the absence of standardized protocols introduces variability across datasets, resulting in inconsistent and incomparable annotations. Third, frame-wise aggregation produces fragmented trajectories that lack narrative coherence and fail to reflect the underlying decision process of continuous viewing behavior.

These issues reveal a fundamental misalignment between existing annotation schemes and event-driven viewing dynamics. To address this gap, we propose a new annotation framework that explicitly models event structure and enforces coherent trajectory representation. As illustrated in Fig. 5, the framework consists of two stages: event-aware perception-guided decoupling and consensus-driven trajectory integration.

3.1.2 Perception-Guided Decoupling

We reformulate 360° video annotation as a behavior reconstruction problem, where continuous viewing decisions are decomposed into event-level semantics and local perceptual evidence. The perception-guided decoupling stage (Fig. 5-A) operationalizes this formulation by separating what to attend from how attention evolves. Instead of directly annotating frame-wise trajectories, which entangle short-term visual responses with long-term decision dynamics, this design improves controllability, consistency, and interpretability.

Specifically, the pipeline first extracts keyframes via low-level visual difference detection, and then constructs event-level units through semantic grouping, ensuring both sensitivity to dynamic changes and preservation of narrative structure. Based on this event-aware representation, annotators perform region-of-interest (RoI) selection within each event, followed by a quality evaluation procedure to ensure annotation reliability. By integrating perception-driven cues with structured constraints, this stage establishes a standardized and robust annotation space for subsequent coherent and consistent trajectory modeling.

Annotation Preparation. To balance immersion and annotation reliability, we adopt a desktop-based visualization system rather than relying solely on HMD devices, thereby avoiding the limited field-of-view and physiological discomfort inherent to immersive setups. Annotators interact with ERP-formatted 360° videos using a customized PyQt5 interface [63], which supports synchronized audio-visual perception for improved identification of salient regions. Prior to annotation, all annotators are required to review the full video to ensure a consistent understanding of the underlying content and event structure.

Event-level Keyframe Construction. As illustrated in Fig. 5-C, we first construct an event-level keyframe representation to bridge low-level temporal dynamics and high-level semantic structure. Specifically, keyframes are extracted based on pixel-level differences between adjacent frames [64]–[66], ensuring that local temporal variations are preserved. However, such low-level representations alone fail to capture the semantic transitions that govern human attention.

To address this limitation, we introduce an event-level grouping mechanism. Semantic features are extracted using a pre-trained vision-language model, and temporally adjacent keyframes with high similarity are grouped into the same event unit. This design establishes a dual structure: smooth, coherent perception within events, and flexible viewpoint transitions across events, thereby aligning the annotation structure with the natural evolution of viewing behavior.

Local Annotation. Given the event-level structure, annotators perform region-of-interest (RoI) selection on each keyframe, as illustrated in Fig. 5-D. Instead of making isolated frame-level decisions, annotations are guided within an event-aware context, reducing ambiguity in viewpoint selection. The viewport is defined with a fixed configuration (79.2° horizontal FoV and 16:9 aspect ratio), consistent with mainstream 360° platforms, ensuring comparability across

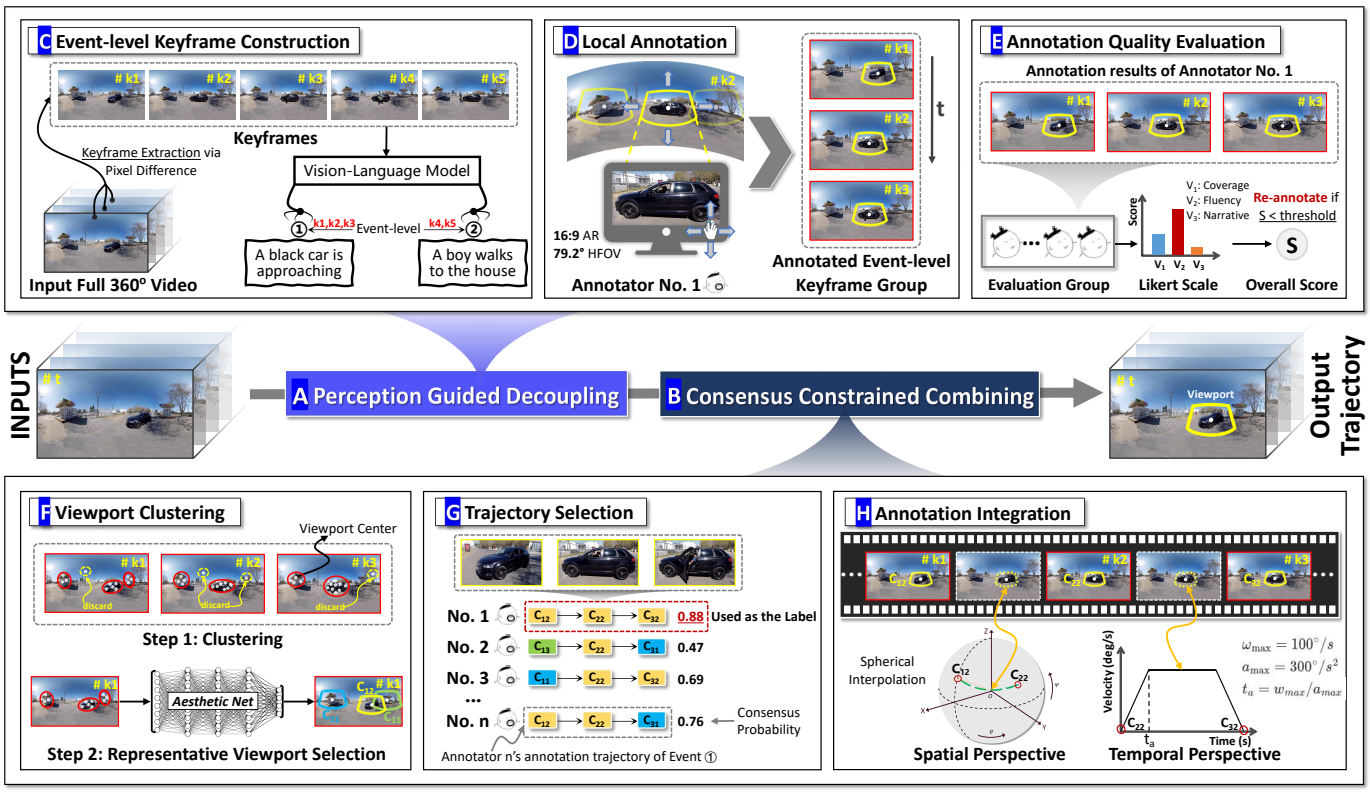


Fig. 5. Pipeline of the proposed 360° video navigation dataset annotation. The framework comprises Perception Guided Decoupling (**A**) and Consensus Constrained Combining (**B**). Stage **A** extracts keyframes, groups them into event-level units, and performs RoI annotation with expert validation. Stage **B** merges local annotations into trajectories via clustering, aesthetic selection, and probabilistic consensus. Finally, spatial spherical interpolation and temporal trapezoidal profiling ensure trajectory continuity, naturalness, and reproducibility.

samples. Annotators can freely adjust viewing direction and iteratively refine their selections, grounding annotations in both local perception and event-level semantics.

Annotation Quality Evaluation. To ensure annotation reliability, we adopt a multi-criteria evaluation protocol (Fig. 5-E). An expert panel evaluates annotations using a five-point Likert scale [67], [68] along three dimensions: (i) *coverage*, measuring the completeness of captured content; (ii) *fluency*, assessing the smoothness of viewpoint transitions; and (iii) *narrative*, reflecting alignment with visual semantics and storytelling. The final score S is computed as the average of these criteria. Annotations with S below a predefined threshold are discarded and re-annotated, ensuring consistency and high-quality supervision. The choice of S_t is empirically validated in Sec. 4.2.5.

This design restores the event-driven structure of human viewing behavior lost in conventional frame-level annotation, effectively bridging annotation design and continuous behavior modeling.

3.1.3 Consensus Constrained Combining

Reconstructing a behaviorally valid viewing trajectory is essential for faithfully modeling continuous human decision processes. To this end, the consensus-constrained combining stage (Fig. 5-B) transforms dispersed local annotations into a coherent trajectory grounded in real observations. Existing frame-by-frame aggregation or clustering strategies often produce fragmented results and may yield fictitious paths that do not correspond to actual viewing behavior, thereby lacking narrative coherence and interpretability. To address this issue, we introduce consensus constraints to guide trajectory construction, ensuring that the final result reflects collective viewing preferences while remaining consistent with human observations.

Viewport Clustering. Annotations from different annotators are inherently dispersed, and directly using them introduces noise. To address this, we cluster all annotations such that each cluster corresponds to a set of highly consistent viewport selections (Fig. 5-F).

Notably, cluster centers are not directly adopted, as they do not correspond to actual annotations and may lead to perceptually implausible results. Instead, within each cluster, candidate viewports are evaluated using an aesthetic model (e.g., [69]), and the real annotation with the highest score is selected as the representative viewport, denoted as C_{ij} (i: keyframe index, j: cluster index). This ensures that all representative viewports originate from real observations while maintaining perceptual quality, thereby avoiding artifacts introduced by virtual cluster centers.

The complete trajectory of each annotator can then be expressed as:

$$\mathcal{T} = \{C_{1j_1}, C_{2j_2}, \ldots, C_{ij_i}, \ldots, C_{nj_n}\}, \quad (1)$$

where $\mathcal{T}$ denotes the trajectory sequence. This representation enables comparison across annotators and provides a structured foundation for subsequent event-level modeling.

Trajectory Selection. 360° video navigation requires not only comprehensive coverage of critical information but also the avoidance of fictitious paths. Such paths arise when frame-by-frame aggregation produces sequences that do not

correspond to any real trajectory, thereby lacking behavioral validity and interpretability.

To address this issue, we propose an event-level trajectory selection strategy based on probabilistic consensus (Fig. 5-G). Within each event unit, the consensus probability of each viewport cluster is computed as the normalized proportion of annotations assigned to that cluster, capturing group-level preferences. Each annotator's trajectory is then mapped to a sequence of cluster probabilities, and a global trajectory score is accumulated across the event sequence using a log-sum formulation. The trajectory with the highest global score is selected as the final output, ensuring that the resulting path corresponds to a real observation sequence while best reflecting collective preferences.

Annotation Integration. To ensure smooth and behaviorally plausible transitions between keyframes, we design a trajectory refinement strategy from both spatial and temporal perspectives (Fig. 5-H).

At the spatial level, adjacent keyframes have fixed positions but uneven temporal intervals. As a result, simple linear interpolation often produces irregular viewpoint motion, with abrupt accelerations and decelerations that deviate from natural gaze patterns. To address this issue, we employ time-normalized spherical interpolation, ensuring smooth angular transitions and avoiding discontinuities near the poles or across longitudinal boundaries.

At the temporal level, to suppress abrupt speed variations, we adopt a trapezoidal angular velocity profile ("acceleration-uniform-deceleration") to regulate motion along the time axis. The acceleration stage is defined as:

$$\Theta(t) = \tfrac{1}{2} a_{\max} t^2, \quad 0 \le t \le t_a, \tag{2}$$

where $a_{\max}$ denotes the maximum angular acceleration and $t_a = \omega_{\max}/a_{\max}$ is the acceleration duration. Following physiological studies on eye and head movements [70], [71], we set $\omega_{\max} = 100°/s$ and $a_{\max} = 300°/s^2$, ensuring that the generated trajectories align with natural gaze dynamics.

After the acceleration phase, the trajectory proceeds at the maximum velocity $\omega_{\max}$, and then symmetrically decelerates to zero as it approaches the target keyframe. For short angular distances, the profile degenerates into a symmetric triangular velocity curve. When the segment duration is insufficient to support a complete trapezoid, feasibility is maintained by reducing the maximum angular velocity or redistributing the angular displacement across neighboring segments. This ensures that the final trajectory is not only smooth, but also consistent with the dynamics of human viewing behavior.

3.2 The Proposed New Dataset

3.2.1 *Why Should We Build This New Dataset?*

Existing 360° video navigation datasets are largely composed of short or single-event clips, which fail to capture cross-event viewing behavior and long-term viewpoint transitions. As a result, models trained on such data tend to learn short-term, locally driven strategies rather than continuous decision processes aligned with semantic event evolution.

Moreover, the lack of unified annotation and evaluation standards often leads to inconsistent results across datasets, hindering fair comparison and progress in the field. To address these issues, we construct a new multi-event 360° video navigation dataset under a standardized annotation framework. With multi-annotator annotations under consistent protocols, the dataset improves both diversity and reliability, providing a more suitable foundation for modeling and evaluating event-driven navigation. This dataset is specifically designed to align with the proposed annotation protocol and evaluation metric, forming a unified benchmark for modeling event-driven viewing behavior.

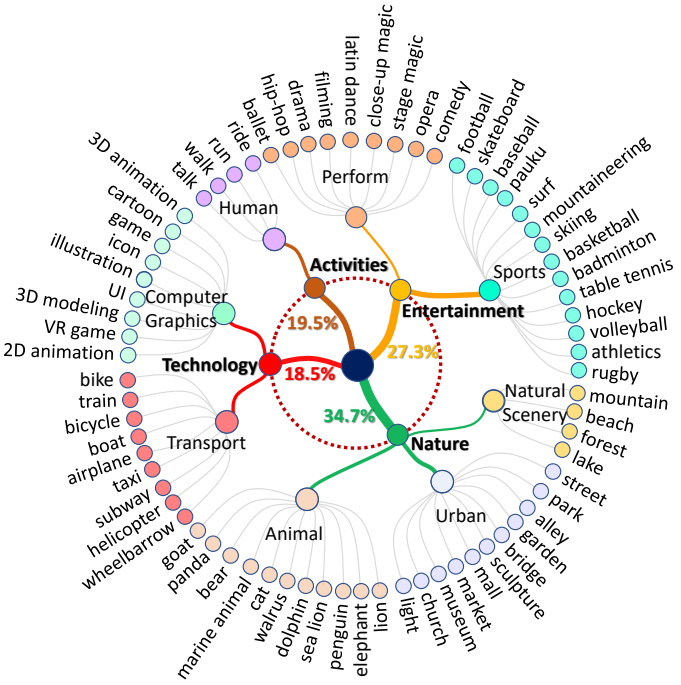


Fig. 6. Hierarchical category distribution of our dataset. The 182 videos are organized into four super-categories and eight major categories, with the outer ring showing fine-grained content labels. The percentages indicate the proportions of videos in the four super-categories.

3.2.2 360° *Video Clip Collection and Annotation*

To construct the dataset, we initially collected 300 360° video clips from YouTube and discarded 118 low-quality samples (e.g., low resolution or severe visual distortion), resulting in 182 high-quality videos. Unlike existing datasets dominated by short or single-event clips, the selected videos primarily contain multi-event sequences, enabling the modeling of cross-event viewing behavior. The dataset covers four super-categories and eight major categories, with representative fine-grained content labels illustrated in Fig. 6 and key statistics summarized in Table 1.

Using the proposed standardized annotation protocol, we recruited 20 annotators (13 females, 7 males, aged 18–28), among whom 12 performed annotation and 8 conducted quality evaluation. The HMD-free annotation process was carried out on PCs with a resolution of 2560×1440. Each annotator spent approximately 30 minutes per video and was allowed to take breaks to mitigate fatigue and ensure annotation quality.

3.2.3 *Advantage Discussions*

The proposed dataset addresses key limitations of existing 360° video navigation benchmarks, particularly limited cross-event coverage, inconsistent annotation protocols, and fragmented trajectory construction. Its main advantages are summarized as follows:

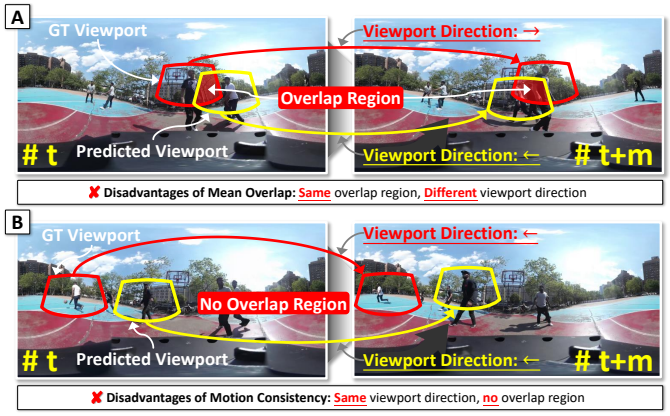


Fig. 7. Comparison of existing and proposed evaluation metrics for 360° video navigation tasks. Subfigure **A** shows the limitation of Mean Overlap, where the viewport regions overlap but directions differ, failing to capture motion consistency. Subfigure **B** illustrates the limitation of Motion Consistency alone, where viewports share direction but lack overlap, potentially missing key content. The proposed metric addresses these issues by combining viewport overlap with motion direction consistency, providing a more comprehensive evaluation of navigation quality.

- **Multi-event Coverage:** Unlike datasets dominated by short or single-event clips, ours consists of multi-event sequences, enabling modeling of cross-event viewing behavior and long-term viewpoint transitions.
- **Diverse and Reliable Annotations:** Each keyframe is annotated by 12 annotators and assessed by eight quality evaluators. Quality screening, viewport clustering, and consensus-based trajectory selection reduce bias and improve annotation agreement, as confirmed by component and consistency analyses.
- **Coherent Trajectory Construction:** The event-level strategy selects complete trajectories from real annotations rather than aggregating frames independently. Spherical interpolation and trapezoidal velocity profiling further smooth keyframe transitions, producing more behaviorally plausible navigation paths.
- **Efficient and Controllable Annotation:** The keyframe-based pipeline reduces frame-wise annotation by focusing on representative content changes. The desktop interface, quality screening, and re-annotation ensure controllable and reliable trajectory construction.

3.3 The Proposed New Evaluation Metric

3.3.1 Why Should We Propose This New Metric?

Current evaluations of 360° video navigation rely on metrics such as mean overlap, which measure alignment between predicted and ground-truth viewports based solely on spatial overlap. However, such formulations are insufficient for dynamic scenarios. Specifically, mean overlap neglects motion dynamics, which play a critical role in immersive viewing. In practice, user attention is strongly influenced by moving objects [72]–[74], requiring a navigation system to model not only where users look, but also how viewpoints evolve over time. As a result, predictions with high spatial overlap may still deviate from true viewing behavior when motion direction is inconsistent with the ground truth (Fig. 7-A), leading to unreliable evaluation.

More importantly, this spatial-only evaluation paradigm implicitly biases models toward short-term alignment, rather than event-level continuous decision making. To address these limitations, we propose a new evaluation metric, termed Motion-Adjusted Overlap (MAO), which jointly captures spatial alignment and motion consistency.

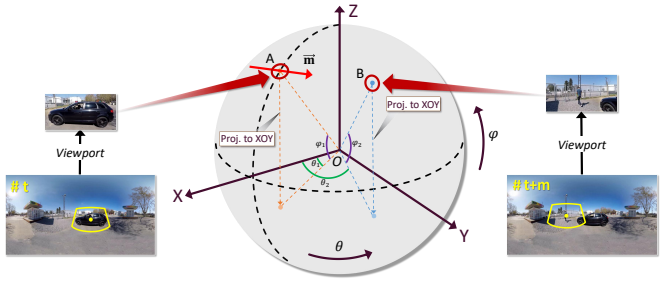


Fig. 8. Conceptual diagram of the proposed evaluation metric for 360° video navigation. Points A and B represent viewport centers at times t and $t+1$, respectively, with $\vec{m}$ indicating the viewport's motion direction at time t. Viewport projection on the spherical coordinate system shows how motion consistency is evaluated based on the alignment of predicted and ground-truth viewport motion paths.

3.3.2 Motion-Adjusted Overlap

Motion-Adjusted Overlap (MAO) is designed to address the limitations of spatial-only metrics in dynamic 360° video scenarios. In addition to measuring viewport overlap, MAO evaluates motion-direction consistency between predicted and ground-truth viewport trajectories, enabling a more accurate characterization of spatiotemporal alignment.

As shown in Fig. 8, we denote the viewport center at time t as $A(\theta_1, \phi_1)$ and at time $t+1$ as $B(\theta_2, \phi_2)$ in spherical coordinates. For computational convenience, these points are mapped to a 3D Cartesian space:

$$A(\theta_1,\phi_1) \to A(x_1,y_1,z_1) = \begin{cases} x_1 = r\cdot\cos\phi_1\cdot\cos\theta_1 \\ y_1 = r\cdot\cos\phi_1\cdot\sin\theta_1 \\ z_1 = r\cdot\sin\phi_1 \end{cases},$$
$$B(\theta_2,\phi_2) \to B(x_2,y_2,z_2) = \begin{cases} x_2 = r\cdot\cos\phi_2\cdot\cos\theta_2 \\ y_2 = r\cdot\cos\phi_2\cdot\sin\theta_2 \\ z_2 = r\cdot\sin\phi_2 \end{cases}, \tag{3}$$

where r is the unit sphere radius, while θ and ϕ denote longitude and latitude, respectively.

The ground-truth motion direction is obtained by projecting the chord $\vec{AB}$ onto the tangent plane at A:

$$\vec{m}_g = (I - \mathbf{a}\mathbf{a}^\top)(\mathbf{b}-\mathbf{a}), \quad \mathbf{a} = \frac{A}{\|A\|},\ \mathbf{b} = \frac{B}{\|B\|}, \tag{4}$$

where I is the identity matrix. The predicted motion vector $\vec{m}_p$ is computed analogously.

We define Motion Consistency (MC) as the cosine similarity between motion directions:

$$MC_{t\to t+1} = \frac{\vec{m}_g\cdot\vec{m}_p}{\|\vec{m}_g\|\,\|\vec{m}_p\|}, \quad t = 1,\ldots,T-1. \tag{5}$$

It is then normalized to $[0,1]$ as:

$$\tilde{MC}_{t\to t+1} = \frac{MC_{t\to t+1}+1}{2}. \tag{6}$$

For completeness, when both ground-truth and predicted viewports remain stationary, we define $\widetilde{MC}_{t\to t+1} = 1$; when only one remains stationary, we set it to 0. Otherwise, the score is computed using Eqs. 5 and 6.

While MC captures motion consistency, it does not guarantee spatial alignment with salient regions. To jointly



Fig. 9. Comparison of viewports generated by existing 360° video navigation methods and the proposed method. Existing methods (yellow) produce visually striking but disjointed viewports, lacking narrative coherence. In contrast, the proposed method (red) generates a smooth narrative thread linking key viewports, ensuring a coherent and engaging viewing experience.

model spatial and temporal agreement, we weight viewport overlap by motion consistency:

$$MAO = \frac{1}{T-1}\sum_{t=1}^{T-1} \overbrace{S_{t\to t+1}^{overlap} \cdot \tilde{MC}_{t\to t+1}}^{\Downarrow}, \quad \frac{1}{2}\left[\mathrm{IoU}\left(l_t^p, l_t^{gt}\right) + \mathrm{IoU}\left(l_{t+1}^p, l_{t+1}^{gt}\right)\right]. \tag{7}$$

where $S_{t\to t+1}^{\text{overlap}}$ denotes the average viewport IoU at the two endpoints of the transition, and l_t^p and l_t^{gt} denote the predicted and ground-truth viewports at time t, respectively. This formulation ensures that accurate predictions are both spatially aligned and motion-consistent, providing a more faithful evaluation of navigation quality.

3.4 The Proposed New Network

3.4.1 Why Should We Develop This New Network?

The goal of 360° video navigation is to guide viewers through immersive content by dynamically selecting informative viewpoints. However, existing methods [7], [8] are predominantly designed under a local and short-term paradigm, relying on saliency cues to highlight prominent regions. While effective in capturing striking elements, such designs largely overlook the underlying semantic structure of videos, leading to fragmented navigation that prioritizes secondary details over key narrative content (Fig. 9).

More importantly, these methods lack the ability to model cross-event dependencies and long-term viewpoint evolution. As a result, they often produce locally plausible yet globally inconsistent trajectories, failing to adapt to event transitions and continuous viewing behavior. Furthermore, treating saliency and semantic information as independent cues neglects their interaction, which may bias the model toward visually salient but semantically irrelevant regions.

These limitations are fundamentally inconsistent with the event-level annotation, multi-event data distribution, and motion-aware evaluation introduced in this work. To address this gap, we propose a new network that jointly models local saliency, temporal dynamics, and global event semantics. By integrating content-aware temporal segmentation, adaptive spatial modeling, and semantic consistency optimization, the proposed framework generates coherent and behaviorally aligned navigation trajectories.

3.4.2 Network Overview

As illustrated in Fig. 10, the proposed framework consists of three components: (A) Content-aware Temporal Segmentation, (B) Adaptive Spatial Segmentation, and (C) Semantic Consistency Optimization.

Given ERP video frames as input, the model first performs content-aware temporal segmentation to partition the video into semantically coherent segments by detecting event boundaries and content transitions, thereby capturing the underlying event structure (Sec. 3.4.3). For each temporal segment, adaptive spatial segmentation is then applied to identify and prioritize salient regions. By jointly leveraging object cues and saliency information, the model selects representative viewports and projects them into cube maps for improved spatial representation (Sec. 3.4.4). Finally, semantic consistency optimization enforces alignment between local viewpoint selection and global narrative context, ensuring that the generated trajectories are both temporally coherent and semantically consistent (Sec. 3.4.5).

3.4.3 Content-aware Temporal Segmentation

Technical Rationale. In 360° video navigation, human attention evolves across semantic events rather than isolated frames. However, most existing methods operate on frame-wise or uniformly segmented inputs, implicitly assuming short-term independence. This prevents models from capturing event transitions and long-term viewpoint evolution, resulting in fragmented navigation that fails to align with the underlying narrative.

To address this limitation, it is necessary to explicitly model event structure by segmenting videos according to content dynamics. We therefore propose a content-aware temporal segmentation strategy (Fig. 10-**A**), which partitions the video into semantically coherent segments based on content changes instead of uniform intervals. This design ensures that each segment corresponds to a meaningful event unit, enabling consistent viewpoint modeling within events while preserving flexibility across event transitions. Furthermore, adjacent segments with similar semantic content are merged to reduce redundancy and improve temporal coherence, providing a stable foundation for subsequent spatial modeling and trajectory generation.

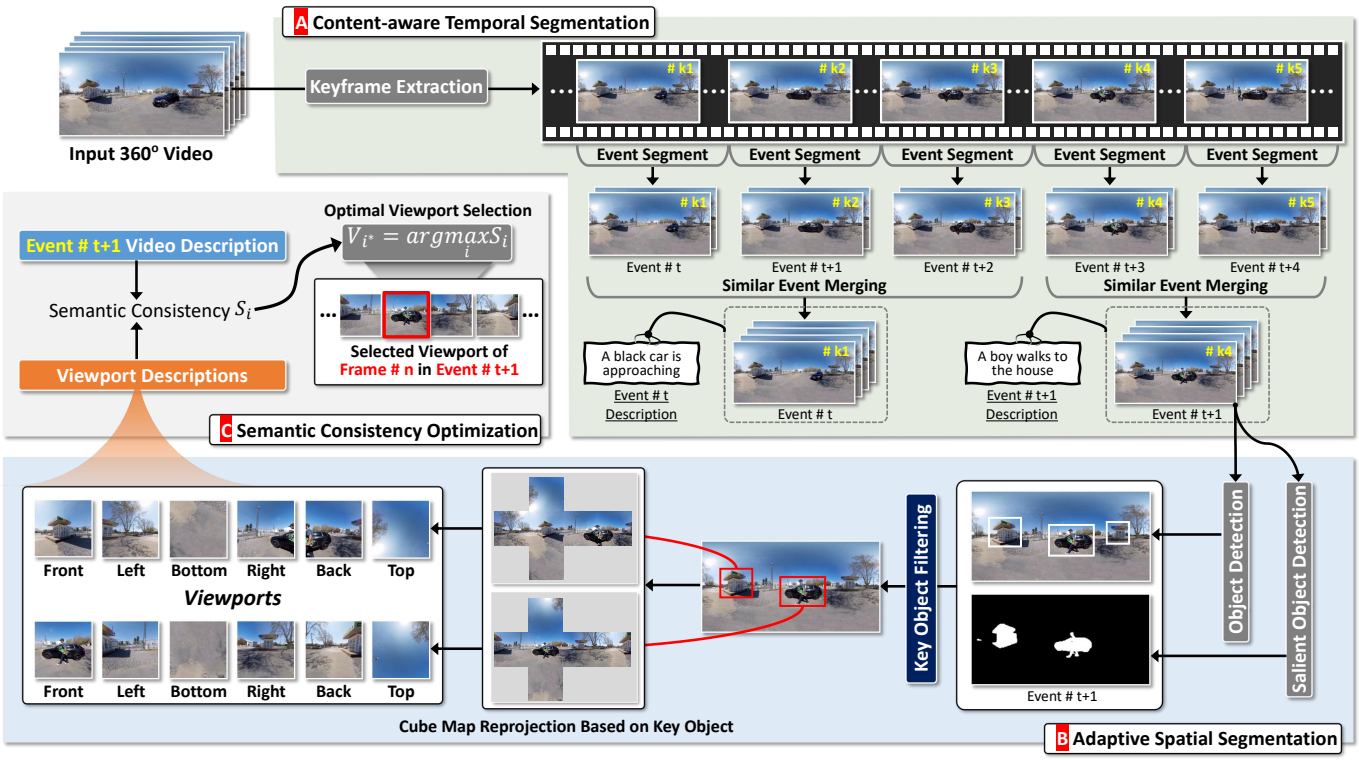


Fig. 10. Overview of the proposed 360° video navigation model, comprising three main components: **A** Content-aware Temporal Segmentation (Sec. 3.4.3) divides the video into event-based segments by detecting key transitions, capturing essential narrative moments. Building on these temporal segments, **B** Adaptive Spatial Segmentation (Sec. 3.4.4) highlights and centers visually relevant regions within each segment, using object and saliency detection to guide viewer attention. **C** Semantic Consistency Optimization (Sec. 3.4.5) aligns these viewports with the global narrative, ensuring thematic coherence. Together, these steps create a smooth navigation path that enhances engagement and narrative comprehension.

Keyframe Extraction. To detect major transitions, we compute pixel-wise differences between consecutive frames $\{F_1, F_2, \ldots, F_N\}$, where F_i denotes the i-th frame. The difference between F_i and F_{i-1} is defined as:

$$D_i = \sum_{x=1}^{H} \sum_{y=1}^{W} \sum_{c=1}^{C} |F_i(x,y,c) - F_{i-1}(x,y,c)|, \tag{8}$$

where H, W, C denote the frame height, width, and number of channels. This measure captures global content variation for identifying candidate keyframes.

To suppress noise caused by minor changes (e.g., illumination variations or small motions), we smooth the sequence $\{D_i\}$ using a normalized Hann window of length $L = 2h+1$, yielding $\{S_i\}$:

$$S_i = \sum_{j=-h}^{h} \frac{1 + \cos\left(\frac{\pi j}{h}\right)}{\sum_{\ell=-h}^{h} \left[1 + \cos\left(\frac{\pi \ell}{h}\right)\right]} \cdot D_{i+j}. \tag{9}$$

where h denotes half the window length. The normalization ensures that the window weights sum to one, suppressing high-frequency noise while preserving major transitions.

Keyframes are then detected as local maxima in $\{S_i\}$:

$$S_i > S_{i-1} \quad \text{and} \quad S_i > S_{i+1}. \tag{10}$$

These points correspond to significant content changes and serve as boundaries for temporal segmentation, forming the basis of event-level units.

Event Segmentation. The detected keyframes define temporal boundaries that partition the video into M segments $\{v_1, v_2, \ldots, v_M\}$, each corresponding to a semantically coherent event unit. This content-aware segmentation aligns with the natural evolution of video content, enabling consistent modeling of viewpoint behavior within each event.

However, direct segmentation based on keyframes may lead to over-fragmentation, producing redundant or weakly separated segments. To address this issue, we further refine the segmentation via semantic-based merging, grouping adjacent segments with similar content into unified event units. This step reduces redundancy and ensures more stable and coherent event-level representations.

Similar Event Merging. Although the initial segmentation captures fine-grained transitions, it may introduce over-fragmentation when adjacent segments are semantically similar. To obtain more stable event units, we further perform semantic-based merging.

Specifically, we extract semantic features for each segment using a vision-language model, resulting in a sequence $\{s_1, s_2, \ldots, s_M\}$. The semantic similarity between adjacent segments is computed via a pre-trained sentence-transformer model. If the similarity $\mathrm{Sim}(s_i, s_{i+1})$ exceeds a threshold ω, the corresponding segments are merged:

$$v_i, v_{i+1} \rightarrow v_i \cup v_{i+1}, \quad \text{if } \mathrm{Sim}(s_i, s_{i+1}) \geq \omega, \tag{11}$$

where $i = 1, \ldots, M-1$.

This merging process groups semantically consistent segments into unified event units, reducing redundancy and preventing over-segmentation. As a result, the temporal segments better reflect coherent event structures, providing a more reliable basis for subsequent viewpoint selection.

3.4.4 *Adaptive Spatial Segmentation*

Technical Rationale. While temporal segmentation captures the event structure of a video, it does not resolve spatial

ambiguity within each event. In 360° videos, salient content is distributed across panoramic space, and a single event may contain multiple competing regions of interest, causing viewpoint selection to be biased toward locally prominent yet semantically irrelevant areas without spatial modeling.

To address this issue, we introduce adaptive spatial segmentation, which identifies and prioritizes candidate regions within each event. By jointly modeling visual saliency and semantic relevance, this module enables more precise viewpoint selection, ensuring that the selected viewports are both informative and aligned with the event context, thereby maintaining consistency with both local visual cues and the global narrative structure.

Candidate Object Localization. For each temporal segment, we perform object localization on a representative frame using the pre-trained YOLO-World detector [75], identifying candidate regions with high semantic relevance. The detected objects are represented by their center coordinates, forming an initial set of candidate viewports.

To refine these candidates, we incorporate saliency filtering using the pre-trained CSMA-Net saliency model [76]. The resulting saliency maps emphasize prominent regions, allowing the model to prioritize candidates that are both semantically meaningful and perceptually salient.

Key Object Filtering. We combine object detection results with saliency information to select the top-k candidates (with $k = 3$). For each object j, a fused score is defined as:

$$\overbrace{F_j = \sigma \cdot S_j + (1-\sigma)\cdot C_j,}^{\Downarrow} \quad \overbrace{S_j = \frac{1}{|A_j|}\sum_{(x,y)\in A_j} S(x,y)}^{} \tag{12}$$

where A_j denotes the pixel set within the bounding box of object j, $S(x, y)$ is the saliency value at location (x, y), and C_j is the detection confidence.

This formulation jointly considers semantic confidence and visual saliency, enabling robust candidate selection. The top-k objects are retained as key candidates and re-centered within each frame to guide subsequent viewpoint generation.

Cube Map Reprojection based on Key Objects. For each selected object, its center coordinates are used as a reference to reproject the frame sequence, ensuring that the object remains within the primary field of view. This step preserves spatial integrity by avoiding the splitting of salient regions across projection boundaries.

After re-centering key objects, we apply Cube Map projection to transform the spherical 360° video into six perspective planes: $front$, $back$, $left$, $right$, top, and $bottom$. Each plane provides a localized view, capturing spatial information within the panoramic scene. This representation facilitates consistent viewpoint generation by maintaining the prominence of key objects across viewing directions.

3.4.5 Semantic Consistency Optimization

Technical Rationale. While temporal and spatial segmentation guide viewpoint selection at the event and local levels, they do not enforce global semantic coherence across the entire video. Without such a constraint, the model may focus on visually salient yet semantically inconsistent regions, leading to discontinuous or unstable viewing trajectories.

To address this issue, we introduce Semantic Consistency Optimization, which incorporates global semantic context to regularize viewpoint selection. By measuring the alignment between local viewport features and the overall video semantics, this module enforces consistency across events and ensures that the generated trajectories remain coherent over time. As a result, the navigation process better reflects both local relevance and global narrative structure.

Local and Global Semantic Description. Given $6 \times k \times M$ candidate sequences (where 6 denotes Cube Map faces, k the number of key objects, and M temporal segments), we extract semantic features for each sequence $V_{q,i}$ using Video-LLaVA [77], yielding local representations $E_{\text{local}}(V_{q,i})$.

To capture global context, we compute a video-level semantic representation E_{global} from the entire 360° video, which encodes the overall narrative and high-level semantics. This global feature serves as a reference for evaluating the relevance of each local candidate.

Semantic Consistency Calculation. For each temporal segment q, we define the semantic consistency score of candidate sequence $V_{q,i}$ as:

$$S_{q,i} = \text{Sim}\left(E_{\text{local}}(V_{q,i}), E_{\text{global}}\right), \quad i = 1, \ldots, 6k, \tag{13}$$

where $\text{Sim}(\cdot)$ denotes the semantic similarity function and $q = 1, \ldots, M$ indexes the temporal segments. This score measures the alignment between each locally selected candidate and the global video semantics.

Optimal Viewport Selection. For each temporal segment, the $6k$ candidates are ranked according to their semantic consistency scores. The optimal candidate index and corresponding viewport sequence are determined as:

$$\overbrace{V_q^* = V_{q,i_q^*}, \qquad q = 1, \ldots, M,}^{\Downarrow} \quad \overbrace{i_q^* = \arg\max_{i\in\{1,\ldots,6k\}} S_{q,i}\,.}^{} \tag{14}$$

This selection maintains global semantic alignment while preserving locally relevant content. To generate a temporally coherent navigation trajectory, we map the selected viewport centers onto the unit sphere and apply time-normalized spherical interpolation between adjacent keyframes. Their frame indices serve as temporal anchors, while a trapezoidal angular-velocity profile regulates transition speed. This design preserves spherical geometry, handles longitudinal wrap-around, and avoids abrupt viewpoint changes. The resulting formulation jointly optimizes spatial selection and temporal coherence under global semantic constraints.

Overall, the proposed network is tightly coupled with the event-level annotation, multi-event dataset, and motion-aware evaluation, forming a unified framework for modeling continuous viewing behavior.

4 Experiments

4.1 Experiments of Annotation Approach and Collected Dataset

4.1.1 Effectiveness of Our Annotation Approach

Ablation Study of Our Annotation Approach. The proposed annotation approach enhances diversity, consistency, and quality through multiple annotators and an evaluation procedure. To assess the contribution of each component, we conducted a component-wise user study involving 40

TABLE 3
Quantitative component studies for proposed annotation method (See Sec. 4.1.1).

	O	A	B	C	D	E	Q1	Q2	Q3
1	✓						5.41	5.28	5.37
2		✓					5.24	4.86	4.96
3		✓	✓				5.58	5.47	5.54
4		✓		✓			5.44	5.63	5.62
5		✓	✓	✓			5.86	5.84	5.83
6			✓	✓	✓		6.17	6.20	6.02
7			✓	✓	✓	✓	6.22	6.29	6.38

O Baseline (each video annotated by one person)
A Average Viewpoint Position of original viewport
B Annotation Quality Evaluation (Sec. 3.1.2-E)
C Viewport Clustering (Sec. 3.1.3-F)
D Trajectory Selection (Sec. 3.1.3-G)
E Annotation Integration (Sec. 3.1.3-H)
Q1 Q2 Q3 Three Questions Used to Evaluate Scores

participants, who were divided into four groups of ten to reduce fatigue. An annotation set comprised a subset of evaluation videos with the navigation trajectories generated by all seven configurations in Table 3. Each group reviewed two annotation sets, and each set was assigned to at least two groups, ensuring that it was evaluated by at least 20 participants. All participants passed a color-vision screening test before the study. For each video, participants watched the original panoramic content to understand its events and narrative and then evaluated the trajectories generated by the seven configurations. The questionnaire included three questions: Q_1^M: Does the video navigation focus on key visual targets? Q_2^M: Is the navigation aesthetically composed? Q_3^M: Do you like this navigation result? Each question was rated on a 7-point Likert scale. For each configuration, valid ratings were averaged across participants and evaluation videos to obtain the final scores reported in Table 3.

Table 3 presents the results. Simple coordinate averaging (row 2) degraded the scores relative to the baseline (row 1), indicating that averaging viewport coordinates is ineffective. Annotation Quality Evaluation (row 3) and Viewport Clustering (row 4) improved performance, while their combination (row 5) achieved the best results among the first five configurations, increasing Q_1^M, Q_2^M, and Q_3^M by 0.45, 0.56, and 0.46, respectively. Replacing frame-wise selection with event-level Trajectory Selection (row 6) further improved these scores by 0.31, 0.36, and 0.19. Finally, Annotation Integration (row 7) increased Q_3^M by an additional 0.36 through spherical interpolation and trapezoidal velocity profiling, demonstrating its effectiveness in producing smoother and more coherent navigation trajectories.

Overall, the proposed annotation approach produces more consistent, coherent, and perceptually preferred trajectories that better align with natural human viewing behavior.

Our Method Reduces Subjectivity Bias in Annotations. Our method effectively reduces subjectivity bias in annotations by alleviating inconsistencies caused by limited annotator diversity in existing approaches. To evaluate its effectiveness, we conduct an experiment where video segments in the test set are divided into five groups, each annotated by 12 annotators. Consistency is assessed using two metrics: "Trajectory Coverage" and "Trajectory Maximum Similarity".

For "Trajectory Coverage", each annotator's trajectory is treated as the baseline, and its coverage is computed with respect to others by measuring the proportion of its viewports appearing in their trajectories. This reflects the degree of agreement among annotators. The average coverage across segments yields the overall "Trajectory Coverage" score. For "Trajectory Maximum Similarity", we measure the overlap of viewports at keyframes between each trajectory and others. The highest overlap with the baseline trajectory is taken as its maximum similarity, and averaging these values across trajectories and segments yields the overall "Trajectory Maximum Similarity" score.

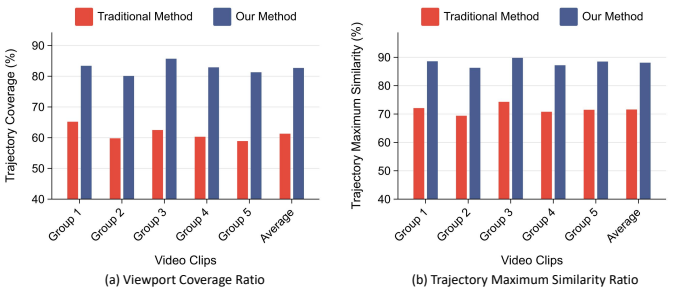


Fig. 11. Comparison of "Trajectory Coverage" and "Trajectory Maximum Similarity" shows that our method achieves higher scores on both metrics across all video groups than traditional methods, indicating reduced subjectivity bias and more consistent annotations.

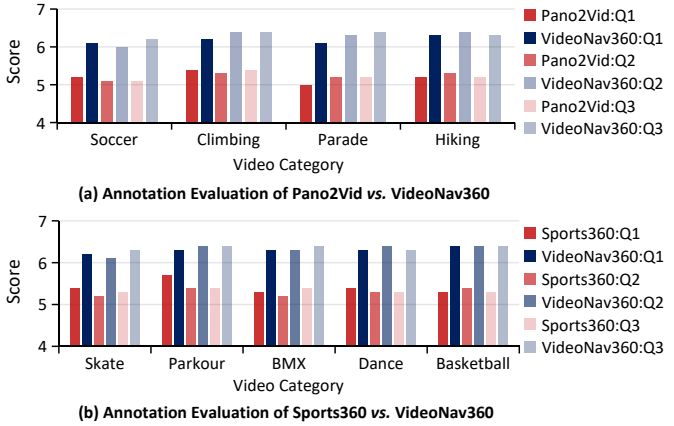


Fig. 12. User study to validate the rationality and effectiveness of the annotated navigation trajectories in our VideoNav360 dataset by comparing them with existing datasets Pano2Vid (a) and Sports360 (b).

The results in Fig. 11 show that our method consistently outperforms traditional approaches on both metrics. The baseline, which relies on independent annotations from 12 annotators and evaluations from 8 raters, does not incorporate the proposed trajectory selection and integration steps. Our method improves inter-annotator consistency, increases coverage of key regions, and reduces subjectivity bias, resulting in more reliable and consistent annotations for downstream navigation and visual saliency tasks.

4.1.2 *User Study on Annotated Navigation Paths*

We conducted a user study to compare the navigation paths annotated by our method with those from two existing methods: Sports360 and Pano2Vid. For this experiment, we randomly selected 40 clips from the Sports360 dataset and 11 from Pano2Vid, annotating the videos using our method while using the existing navigation paths from the other two datasets. Ten users were recruited to evaluate the annotated paths using a questionnaire from Sec. 4.1.1. Effective navigation paths, which enhance the viewing experience, generally received higher scores. The results in Fig. 12 show that our method achieves consistently higher average scores than Pano2Vid (a) and Sports360 (b), indicating that the generated navigation paths better align with user preferences and improve the viewing experience.

4.1.3 *Effectiveness of Our VideoNav360 Dataset*

This experiment aims to evaluate whether the proposed VideoNav360 dataset can improve the performance of exist-

TABLE 4
Cross-dataset quantitative evaluation between VideoNav360 and Sports360. See Sec. 4.1.3 for details.

Dataset	Train on VideoNav360 (Ours)						Train on Sports360					
	❶ Test on Ours			❷ Test on Sports360			❸ Test on Ours			❹ Test on Sports360		
Metric	MO↑	MVD↓	MAO↑	MO↑	MVD↓	MAO↑	MO↑	MVD↓	MAO↑	MO↑	MVD↓	MAO↑
Deep360	0.630	7.130	0.555	0.651	7.293	0.547	0.630	7.589	0.480	0.677	7.216	0.639
A-DRL	0.629	6.675	0.524	0.619	6.595	0.513	0.574	6.989	0.479	0.596	7.167	0.481
S2CNet	0.592	8.030	0.481	0.602	7.425	0.486	0.542	7.594	0.435	0.577	8.022	0.453
DCNet	0.458	6.121	0.349	0.521	5.400	0.518	0.487	5.870	0.498	0.443	3.973	0.336
LESOD	0.622	7.844	0.504	0.640	7.488	0.515	0.621	7.785	0.479	0.611	7.692	0.496

TABLE 5
Quantitative comparison of PanoNavNet with other SOTA methods on Pano2Vid, Sports360, and VideoNav360 datasets.

Dataset	Pano2Vid			Sports360			VideoNav360		
Methods	MO↑	MVD↓	MAO↑	MO↑	MVD↓	MAO↑	MO↑	MVD↓	MAO↑
Deep360$_{[2017]}$	0.604	14.730	0.544	0.670	6.892	0.623	0.635	7.384	0.571
A-DRL$_{[2020]}$	0.588	**14.102**	0.498	0.603	6.975	0.493	0.612	6.984	0.510
IRSR$_{[2021]}$	0.315	21.852	0.211	0.341	12.075	0.232	0.328	11.535	0.223
PSR$_{[2023]}$	0.397	16.878	0.297	0.402	7.194	0.302	0.415	7.386	0.315
ASSRE$_{[2024]}$	0.301	18.648	0.199	0.345	10.929	0.240	0.330	10.086	0.228
QAGNet$_{[2024]}$	0.330	14.280	0.226	0.350	9.684	0.263	0.356	9.123	0.263
S2CNet$_{[2024]}$	0.552	15.930	0.442	0.576	7.684	0.461	0.584	7.746	0.472
DCNet$_{[2025]}$	0.422	15.726	0.317	0.442	**3.822**	0.345	0.455	7.060	0.355
LESOD$_{[2026]}$	0.598	15.210	0.487	0.623	7.617	0.501	0.635	7.684	0.514
Ours	**0.715**	14.560	**0.636**	**0.710**	6.974	**0.653**	**0.736**	**6.743**	**0.671**

ing 360° video navigation models. The dataset is constructed based on our annotation pipeline, with the goal of providing higher-quality and more consistent navigation trajectories. Considering that existing 360° navigation models are relatively limited and some are not publicly available, we select Deep 360 Pilot [7] (Deep360), A-DRL [46], S2CNet [78], DCNet [22], and LESOD [79] as representative models for evaluation. Our hypothesis is that if the proposed dataset provides higher-quality navigation trajectories, models trained on it should achieve better performance.

To verify this hypothesis, we randomly select 50 videos for training and 30 videos for testing from each of VideoNav360 and Sports360. We repeat this random split five times and report the average results. Each model is trained on one dataset and evaluated on both datasets.

The results in Table 4 show that training on VideoNav360 improves cross-dataset performance, although gains vary across models and metrics. On the Sports360 test set, VideoNav360 training improves all three metrics for A-DRL, S2CNet, and LESOD, while improving MO and MAO for DCNet. On the VideoNav360 test set, it provides gains for A-DRL and improves most metrics for Deep360, S2CNet, and LESOD. These results suggest that VideoNav360 provides useful supervision for learning transferable event-level navigation patterns, although model-specific differences remain.

4.2 Experiments of PanoNavNet and Metric MAO

4.2.1 Implementation Details, Datasets and Metrics

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB memory and 16 GB RAM. Each ERP video frame was resized to 960×480. PanoNavNet does not require end-to-end training; instead, it integrates publicly available pre-trained models to improve deployment flexibility and avoid overfitting on limited navigation data. Specifically, we use YOLO-World [75] for object detection, CSMA-Net for 360° saliency estimation, and Video-LLaVA [77] for local and global semantic description. Cosine similarity is used for semantic consistency calculation. The number of retained key objects is set to $k = 3$, while the balancing factor, event-merging threshold, and annotation-quality threshold are set to $\sigma = 0.75$, $\omega = 0.73$, and $S_t = 4.0$, respectively. These settings are fixed across all experiments.

We evaluate the proposed method on the publicly available Pano2Vid and Sports360 datasets and our newly constructed VideoNav360 dataset. For Pano2Vid and Sports360, only videos that remained publicly accessible were included because several original download links were invalid. We adopt Mean Overlap (MO), Mean Velocity Difference (MVD), and the proposed Motion-Adjusted Overlap (MAO) for evaluation. MO measures spatial overlap between predicted and ground-truth viewports, while MVD evaluates trajectory smoothness, with lower values indicating smoother transitions. MAO jointly measures spatial alignment and motion-direction consistency, providing a complementary evaluation of dynamic navigation quality.

4.2.2 Evaluations with SOTA Methods

Since most existing 360° video navigation models are not publicly available, we adopt Deep 360 Pilot [7] (Deep360) and A-DRL as representative methods for comparison. In addition, we include an aesthetic cropping model, S2CNet [78], and six saliency ranking models, including IRSR [80], ASSRE [81], PSR [82], QAGNet [83], DCNet [22], and LESOD [79]. All models are retrained to ensure fair comparison. It is important to note that saliency ranking and cropping models differ from 360° video navigation models in their objectives. The former focus on identifying visually important or aesthetically appealing regions, while the latter aim to model dynamic viewport selection and continuous navigation over global content. To ensure comparability, we adapt these methods by selecting the most salient region as the initial viewport and using its centroid as the final coordinate, thereby establishing a consistent basis for comparison across methods.

Quantitative Evaluation. The results in Table 5 show that our method achieves the best MO and MAO on all three datasets while remaining competitive in MVD. The gains are particularly pronounced in MAO, demonstrating the ability to jointly model spatial alignment and motion consistency. Although several competing methods obtain lower MVD on Pano2Vid and Sports360, their substantially lower MO and MAO indicate that trajectory smoothness alone does not guarantee accurate or motion-aligned navigation. These results confirm that our approach effectively captures event-level temporal dynamics and continuous viewing behavior rather than relying solely on local saliency cues. Unlike saliency-ranking methods that prioritize visually prominent regions without considering temporal coherence, our model integrates local saliency with global narrative context.

Qualitative Evaluation. To further validate the effectiveness of our approach in modeling continuous viewing behavior, we present qualitative comparisons with representative SOTA methods in Fig. 13. In Scene A, the video depicts a cyclist in red passing by a group of people in a park at night. Several competing methods tend to focus on locally salient but less event-relevant regions, such as static pedestrians, background structures, or illumination changes, resulting in viewpoint deviations from the main moving subject. In

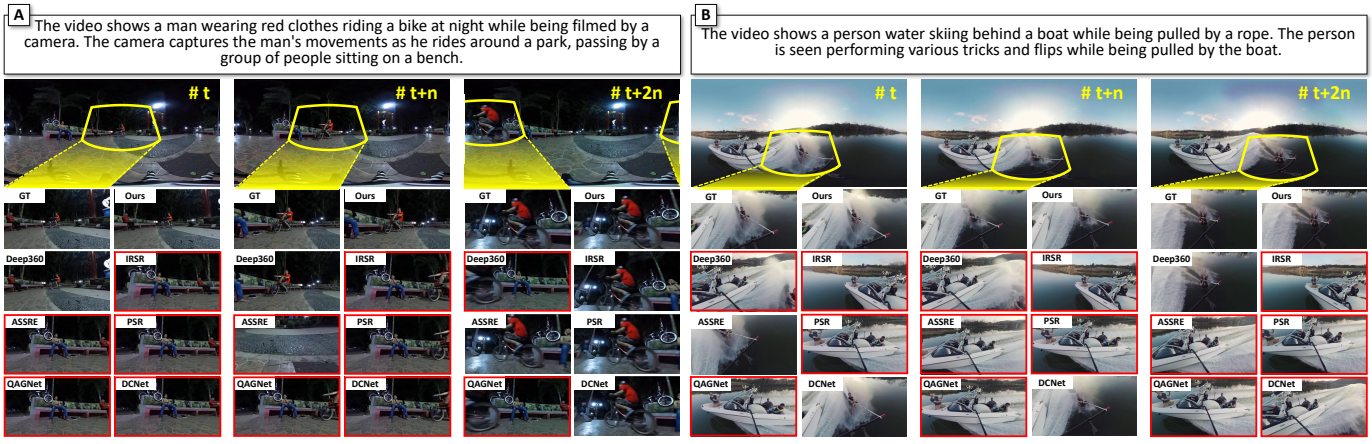


Fig. 13. Qualitative comparisons between PanoNavNet and several representative state-of-the-art methods on the VideoNav360 dataset. All trainable competing methods were retrained on the VideoNav360 training set and evaluated on the corresponding test set, whereas PanoNavNet employed fixed pre-trained components without end-to-end training. See Sec. 4.2.2 for details.

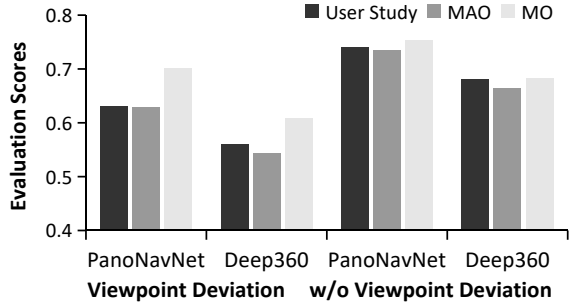


Fig. 14. User study of the proposed metric MAO (see Sec. 4.2.3) shows that in videos with viewport shifts, MAO exhibits trends more consistent with user ratings than MO. In contrast, for videos without viewport shifts, MAO, MO, and user ratings exhibit similar trends. These findings suggest that MAO is more consistent with the user's viewing experience.

contrast, our method consistently attends to the cyclist and the surrounding human interaction context, which better reflects the progression of the scene. This demonstrates the benefit of incorporating global semantic cues for viewpoint selection. In Scene B, the main event is a water skier being pulled by a boat and performing dynamic movements on the water. Existing methods are easily attracted by visually dominant regions such as the boat, water splashes, wake patterns, or strong reflections, which may cause unstable viewpoint shifts across time. Our method maintains attention on the skier and the action-relevant region, producing viewpoints that are more consistent with the ground truth. This highlights the ability of our model to suppress short-term visual distractions and preserve temporal coherence. Overall, these examples show that our method prioritizes event-relevant regions over transient or locally salient stimuli, resulting in more coherent and stable navigation trajectories.

4.2.3 *Effectiveness Evaluation of Metric MAO*

To evaluate the effectiveness of the proposed Motion-Adjusted Overlap (MAO) metric, we conduct a user study. From three datasets (ours, Sports360, and Pano2Vid), we select 20 viewpoint-shift videos per dataset, where the predicted viewport trajectory deviates from the ground-truth viewport motion direction, yielding 60 samples. Additionally, we select 10 non-viewpoint-shift videos per dataset, resulting in 30 samples.

We compare two 360° navigation methods: Deep 360 Pilot [7] (Deep360) and our PanoNavNet, under both conditions. Ten evaluators, who previously participated in the dataset study (Sec. 4.1.1), rate the predicted trajectories via a standardized questionnaire. For each trajectory, we record (1) subjective scores, (2) MAO, and (3) MO. All 1–7 ratings are normalized to [0,1] for consistency.

As shown in Fig. 14, in viewpoint-shift cases, both user scores and MAO are low, indicating motion misalignment, while MO remains high, exposing its limitation in dynamic scenarios. In non-viewpoint-shift cases, user scores, MAO, and MO exhibit similarly high trends, indicating that both metrics are reliable when motion is consistent. Overall, our method achieves higher user satisfaction and MAO, with improved MO. MAO exhibits trends more consistent with human ratings than MO in dynamic scenes, suggesting that jointly modeling spatial overlap and motion consistency provides a more representative measure of navigation quality. This validates MAO as a strong complement to MO, especially for motion-intensive 360° content.

4.2.4 *Effectiveness Evaluation on Different Components*

To assess the contribution of each component, we construct a baseline and progressively introduce modules. The baseline performs equidistant sampling of video clips (A), applies random Cube Map projection (E), and uses plain ranking (G) for navigation. As shown in Table 6, this baseline achieves an MO of 0.656, MVD of 7.153, and MAO of 0.572, and serves as the reference for comparisons. Processing time per 600 frames (FPS = 25) is also reported.

Effectiveness of Content-Aware Temporal Segmentation. Adding keyframe extraction (B) (row 2 vs. row 1) improves MO and MAO to 0.679 and 0.602, respectively, indicating that informative frames are more effectively captured. This improves alignment with key events, although MVD slightly increases due to stronger viewpoint transitions. Further incorporating Similar Event Merging (C) (row 3 vs. row 2) reduces MVD from 7.314 to 7.104 and processing time from 9293 to 5629 ms, corresponding to a 39.4% reduction. Meanwhile, MO and MAO decrease moderately from 0.679 and 0.602 to 0.658 and 0.576, respectively. These results show that semantic merging removes redundant segments and improves computational efficiency and trajectory smoothness, with a modest trade-off in navigation accuracy.

TABLE 6
Quantitative component studies for PanoNavNet (See Sec. 4.2.4). The unit of time is milliseconds (ms).

	A	B	C	D	E	F	G	H	MO	MVD	MAO	Time
1	✓				✓		✓		0.656	7.153	0.572	9285
2		✓			✓		✓		0.679	7.314	0.602	9293
3	✓		✓		✓		✓		0.658	7.104	0.576	5629
4		✓	✓		✓		✓		0.676	7.306	0.599	5637
5		✓	✓			✓	✓		0.709	6.940	0.634	5787
6		✓	✓	✓	✓		✓		0.673	7.241	0.596	2393
7		✓	✓	✓		✓	✓		0.714	6.821	0.645	2455
8		✓	✓	✓		✓		✓	0.736	6.743	0.671	2573

A Equidistant Sampling of Video Clips
B Keyframe Extraction (Eqs. 8-10 in Sec. 3.4.3)
C Similar Event Merging (Eq. 11 in Sec. 3.4.3)
D Key Object Filtering (Eq. 12 in Sec. 3.4.4)
E Random Projection of Cube Maps
F Cube Map Reprojection based on Key Object (Sec. 3.4.4)
G Plain Ranking
H Semantic Consistency Calculation (Eqs. 13-14 in Sec. 3.4.5)

TABLE 7
Ablation study on the threshold S_t in annotation quality evaluation (Sec. 3.1.2).

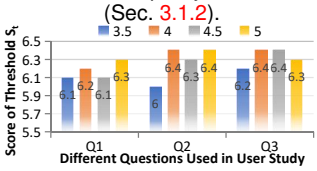


TABLE 8
Ablation study on the semantic similarity in similar event merging (Eq. 11).

ω	MO	MVD	MAO	Time
0.53	0.473	3.758	0.341	1682
0.63	0.592	4.660	0.503	2248
0.73	0.736	6.743	0.671	2573
0.83	0.734	6.774	0.668	3786

TABLE 9
Ablation study on the balancing factor for key object filtering (Eq. 12).

σ	MO	MVD	MAO
0.55	0.694	7.083	0.620
0.65	0.713	7.042	0.643
0.75	0.736	6.743	0.671
0.85	0.720	6.820	0.655

TABLE 10
Ablation study on different choices of video caption methods in semantic consistency calculation (Sec. 3.4.5).

Video Caption	MO	MVD	MAO
BMT	0.732	6.748	0.664
MDVC	0.731	6.739	0.674
PDVC	0.734	6.747	0.669
Video-LLaVA	0.736	6.743	0.671

Effectiveness of Adaptive Spatial Segmentation. We next examine spatial modules D–F. Key object filtering (D) (line 6 vs. line 4) reduces processing time from 5637 to 2393 ms while maintaining comparable performance, indicating that removing irrelevant regions improves efficiency. Introducing Cube Map reprojection based on key objects (F) (line 7 vs. line 6) leads to clear improvements, increasing MO to 0.714 and MAO to 0.645, while reducing MVD to 6.821. This demonstrates that centering key objects strengthens spatial coherence and aligns the viewport with user attention.

Effectiveness of Semantic Consistency Optimization. Semantic consistency calculation (H) (line 8 vs. line 7) further improves all metrics, achieving MO = 0.736, MVD = 6.743, and MAO = 0.671. By ranking candidates by semantic relevance, the model prioritizes event-consistent regions and avoids locally salient yet contextually irrelevant selections.

These results show that each component addresses a distinct aspect of the problem. Temporal segmentation captures event structure, spatial segmentation ensures focus on key regions, and semantic consistency enforces alignment with global context. Together, they enable stable, coherent, and event-driven navigation trajectories.

4.2.5 *Ablation Study*

Effect of the Threshold S_t in Annotation Quality Evaluation. We conduct an ablation study on the score threshold S_t, which determines whether annotations are re-collected (Sec. 3.1.2). Four thresholds (3.5, 4.0, 4.5, and 5.0) are evaluated to balance data quality and annotation cost. As shown in Table 7, setting $S_t = 4.0$ effectively filters out low-quality annotations, leading to improved trajectory coverage and similarity without introducing excessive re-annotation overhead. In contrast, higher thresholds yield marginal gains while increasing re-annotation overhead.



(a) A scene with no obvious focus (b) Multi-focus dynamic scene

Fig. 15. The limitations of the proposed annotation approach and PanoNavNet are discussed in Sec. 4.3 for more details.

These results indicate that $S_t = 4.0$ achieves the best trade-off between data quality and annotation efficiency, ensuring both consistency and stability.

Semantic Similarity in Similar Event Merging. We study the effect of the threshold ω in Similar Event Merging (Eq. 11), which determines whether adjacent segments are merged based on their semantic similarity. We evaluate four settings, $\omega \in \{0.53, 0.63, 0.73, 0.83\}$. As shown in Table 8, increasing ω from 0.53 to 0.73 substantially improves MO and MAO, indicating more accurate event separation and better navigation alignment. Meanwhile, MVD increases from 3.758 to 6.743, reflecting a trade-off between content alignment and trajectory smoothness as stricter merging preserves more event transitions. Further increasing ω to 0.83 slightly decreases MO and MAO, further increases MVD, and raises the processing time by 47.1%. These results indicate that $\omega = 0.73$ achieves the best trade-off among spatial alignment, motion consistency, trajectory smoothness, and computational efficiency.

Balancing Factor for Key Object Filtering. In Eq. 12, the parameter σ controls the trade-off between saliency and detection confidence when estimating object importance. A larger σ emphasizes salient regions, while a smaller value preserves detection confidence. As shown in Table 9, increasing σ from 0.55 to 0.75 improves performance, with MAO rising from 0.620 to 0.671, indicating more accurate identification of key regions. However, further increasing σ to 0.85 leads to a performance drop, suggesting that over-reliance on saliency suppresses detection confidence and causes less salient yet semantically important objects to be overlooked. These results indicate that $\sigma = 0.75$ achieves the best balance, effectively integrating visual saliency and semantic relevance for robust object selection.

Different Choices of Video Caption Methods. We evaluate the impact of different vision-language models used for semantic representation in Sec. 3.4.5. Specifically, we compare four video captioning models, including BMT [84], MDVC [85], PDVC [86], and Video-LLaVa [77], which differ in their feature extraction strategies. As shown in Table 10, performance remains stable across all choices, with only marginal variations in MO, MVD, and MAO. This indicates that the proposed Semantic Consistency Optimization is robust to the choice of captioning model and does not rely on a specific architecture. These results suggest that our method effectively leverages high-level semantic cues in a model-agnostic manner, enabling consistent alignment between viewer attention and event-level narrative structure.

4.3 Limitations

Despite the effectiveness of our annotation method in addressing subjectivity issues, there are two key limitations. First, our approach struggles in scenes lacking a clear focal

point, as shown in Fig. 15-a. In such scenes, where there is no obvious narrative focus, the system has difficulty prioritizing a viewport area, leading to dispersed attention and inconsistent viewer guidance. Second, in dynamic scenes with multiple focal points (Fig. 15-b), where several narrative areas coexist, our navigation path may frequently switch between focus points. This results in scattered viewport transitions and a less stable trajectory, which can impact the smoothness and coherence of the viewing experience. These examples demonstrate how attention dispersion across multiple areas can fragment navigation and reduce the overall viewing experience.

5 Conclusion

This paper addresses a fundamental challenge in 360° video navigation: the lack of a unified framework for modeling consistent, event-driven viewing behavior, which leads to subjective annotations and unreliable evaluation. To address this issue, we propose a comprehensive benchmark that jointly improves data construction, annotation, evaluation, and modeling. We introduce an event-level annotation framework that integrates user feedback with structured evaluation criteria, enabling the collection of high-quality and behaviorally consistent navigation trajectories. We further present the Motion-Adjusted Overlap (MAO) metric, which incorporates motion consistency to capture dynamic alignment between predicted and ground-truth trajectories, overcoming the limitations of purely spatial evaluation. We also develop a navigation model that integrates content-aware temporal segmentation, adaptive spatial modeling, and semantic consistency optimization. By jointly leveraging local saliency and global semantic context, the model produces coherent and narrative-aligned navigation paths. This work establishes a unified and principled foundation for 360° video navigation, advancing event-level modeling of continuous viewing behavior and enabling more reliable and semantically consistent immersive experiences.

References

[1] W. Fang, L. Chen, T. Zhang, C. Chen, Z. Teng, and L. Wang, "Head-mounted display augmented reality in manufacturing: A systematic review," *RCIM*, vol. 83, 2023.

[2] A. D. Souchet, S. Philippe, D. Lourdeaux, and L. Leroy, "Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review," *IJHCI*, vol. 38, no. 9, pp. 801–824, 2022.

[3] H. Ai and L. Wang, "Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion," in *CVPR*, 2024, pp. 9926–9935.

[4] R. Cong, K. Huang, J. Lei, Y. Zhao, Q. Huang, and S. Kwong, "Multi-projection fusion and refinement network for salient object detection in 360º omnidirectional image," *IEEE TNNLS*, 2023.

[5] T. Hara, Y. Mukuta, and T. Harada, "Spherical image generation from a few normal-field-of-view images by considering scene symmetry," *IEEE TPAMI*, vol. 45, no. 5, pp. 6339–6353, 2022.

[6] Z. Lu, K. Hu, C. Wang, L. Bai, and Z. Wang, "Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation," in *AAAI*, vol. 38, no. 13, 2024, pp. 14 211–14 219.

[7] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos," in *CVPR*, 2017, pp. 1396–1405.

[8] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2vid: Automatic cinematography for watching 360 videos," in *ACCV*, 2016, pp. 154–171.

[9] Z. Zhao, M. Wang, X. Sui, P. Chen, X. He, and S. Wang, "Rate control for 360° versatile video coding based on visual gaze mechanism," *IEEE TMM*, 2026.

[10] X. Sui, Y. Fang, H. Zhu, S. Wang, and Z. Wang, "Scandmm: A deep markov model of scanpath prediction for 360deg images," in *CVPR*, 2023, pp. 6989–6999.

[11] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Saltinet: Scan-path prediction on 360 degree images using saliency volumes," in *CVPR*, 2017, pp. 2331–2338.

[12] X. Chen, M. Jiang, and Q. Zhao, "Beyond average: Individualized visual scanpath prediction," in *CVPR*, 2024, pp. 25 420–25 431.

[13] R. A. J. de Belen, T. Bednarz, and A. Sowmya, "Scanpathnet: A recurrent mixture density network for scanpath prediction," in *CVPR*, 2022, pp. 5010–5020.

[14] M. Kalash, M. A. Islam, and N. D. Bruce, "Relative saliency and ranking: Models, metrics, data and benchmarks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 204–219, 2019.

[15] R. Cong, Z. Chen, H. Fang, S. Kwong, and W. Zhang, "Breaking barriers, localizing saliency: A large-scale benchmark and baseline for condition-constrained salient object detection," *IEEE TPAMI*, vol. 48, no. 4, pp. 4167–4183, 2026.

[16] Z. Wu, A. Alhilal, Y. H. Tsui, W. J. Chai, M. Siekkinen, and P. Hui, "Saliency-guided foveated video encoding for low-latency and immersive cloud vr," *IEEE TVCG*, 2026.

[17] H. Liu, Y. Gao, S. Zhou, and L. Zhou, "More is not always better: Toward general cross-modal saliency prediction for immersive communications," *IEEE TMM*, 2026.

[18] M. Song, L. Li, D. Wu, W. Song, and C. Chen, "Rethinking object saliency ranking: A novel whole-flow processing paradigm," *IEEE TIP*, vol. 33, pp. 338–353, 2024.

[19] H. Fang, D. Zhang, Y. Zhang, M. Chen, J. Li, Y. Hu, D. Cai, and X. He, "Salient object ranking with position-preserved attention," in *CVPR*, 2021, pp. 16 331–16 341.

[20] Z. Wang, X. Yan, Y. Han, and M. Sun, "Ranking video salient object detection," in *ACM MM*, 2019, pp. 873–881.

[21] M. A. Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *CVPR*, 2018, pp. 7142–7150.

[22] J. Zhu, X. Qin, and A. Elsaddik, "Dc-net: Divide-and-conquer for salient object detection," *Pattern Recognition*, vol. 157, p. 110903, 2025.

[23] Y. Yu, S. Lee, J. Na, J. Kang, and G. Kim, "A deep ranking model for spatio-temporal highlight detection from a 360° video," in *AAAI*, vol. 32, 2018.

[24] S. Lee, J. Sung, Y. Yu, and G. Kim, "A memory network approach for story-based temporal summarization of 360 videos," in *CVPR*, 2018, pp. 1410–1419.

[25] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE TPAMI*, vol. 41, no. 11, pp. 2693–2708, 2018.

[26] C. Wu, Z. Tan, Z. Wang, and S. Yang, "A dataset for exploring user behaviors in vr spherical video streaming," in *ACM MSC*, 2017, pp. 193–198.

[27] A. T. Nasrabadi, A. Samiei, A. Mahzari, R. P. McMahan, R. Prakash, M. C. Farias, and M. M. Carvalho, "A taxonomy and dataset for 360 videos," in *ACM MSC*, 2019, pp. 273–278.

[28] C. Wu, R. Zhang, Z. Wang, and L. Sun, "A spherical convolution approach for learning long term viewport prediction in 360 immersive video," in *AAAI*, vol. 34, no. 1, 2020, pp. 14 003–14 040.

[29] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[30] S. Petrangeli, G. Simon, and V. Swaminathan, "Trajectory-based viewport prediction for 360-degree virtual reality videos," in *IEEE AIVR*, 2018, pp. 157–160.

[31] S. Atev, G. Miller, and N. P. Papanikolopoulos, "Clustering of vehicle trajectories," *IEEE TITS*, vol. 11, no. 3, pp. 647–657, 2010.

[32] A. T. Nasrabadi, A. Samiei, and R. Prakash, "Viewport prediction for 360 videos: a clustering approach," in *ACM WNOSSDAV*, 2020, pp. 34–39.

[33] Y. Lee and W. C. Kim, "Concise formulas for the surface area of the intersection of two hyperspherical caps," *KAIST Technical Report*, 2014.

[34] C. Wu, Z. Wang, and L. Sun, "A spherical mixture model approach for 360 video virtual cinematography," in *ICIP*, 2021, pp. 1179–1183.

[35] T. Pintaric, U. Neumann, and A. Rizzo, "Immersive panoramic video," in *ACM MM*, vol. 493, 2000.
[36] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *ICCV*, 2015, pp. 262–270.
[37] M. Cokelek, H. Ozsoy, N. Imamoglu, C. Ozcinar, I. Ayhan, E. Erdem, and A. Erdem, "Spherical vision transformers for audio-visual saliency prediction in 360° videos," *IEEE transactions on pattern analysis and machine intelligence*, 2025.
[38] Y.-C. Su and K. Grauman, "Making 360 video watchable in 2d: Learning videography for click free viewing," in *CVPR*, 2017, pp. 1368–1376.
[39] A. Nguyen, Z. Yan, and K. Nahrstedt, "Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction," in *ACM MM*, 2018, pp. 1190–1198.
[40] M. Setayesh and V. W. Wong, "A content-based viewport prediction framework for 360° video using personalized federated learning and fusion techniques," in *ICME*, 2023, pp. 654–659.
[41] Q. Zhou, M. Wu, Y. Zhang, M. Zink, R. Sitaraman, and K. Nahrstedt, "360tripleview: 360-degree video view management system driven by convergence value of viewing preferences," in *IEEE ISM*, 2023, pp. 28–35.
[42] M. Qiao, M. Xu, Z. Wang, and A. Borji, "Viewport-dependent saliency prediction in 360 video," *IEEE TMM*, vol. 23, pp. 748–760, 2020.
[43] I. Kontostathis, E. Apostolidis, and V. Mezaris, "An integrated system for spatio-temporal summarization of 360-degrees videos," in *ICMM*, 2024, pp. 202–215.
[44] Y. Dahou, M. Tliba, K. McGuinness, and N. O'Connor, "Atsal: an attention based architecture for saliency prediction in 360° videos," in *ICPR*, 2021, pp. 305–320.
[45] E. Bernal-Berdun, D. Martin, D. Gutierrez, and B. Masia, "Sst-sal: A spherical spatio-temporal approach for saliency prediction in 360° videos," *Computers & Graphics*, vol. 106, pp. 200–209, 2022.
[46] J. Wang, M. Xu, L. Jiang, and Y. Song, "Attention-based deep reinforcement learning for virtual cinematography of 360° videos," *IEEE TMM*, vol. 23, pp. 3227–3238, 2020.
[47] C. Liu *et al.*, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.
[48] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *ICML*, 2016, pp. 1928–1937.
[49] F.-Y. Chao, C. Ozcinar, and A. Smolic, "Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need." in *MMSP*, 2021, pp. 1–6.
[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS*, vol. 30, 2017.
[51] K. Kang and S. Cho, "Interactive and automatic navigation for 360 video playback," *ACM TOG*, vol. 38, no. 4, pp. 1–11, 2019.
[52] M. Wang, Y.-J. Li, W.-X. Zhang, C. Richardt, and S.-M. Hu, "Transitioning360: Content-aware nfov virtual camera paths for 360 video playback," in *ISMAR*, 2020, pp. 185–194.
[53] A. Rao, L. Xu, and D. Lin, "Shoot360: Normal view video creation from city panorama footage," in *ACM SIGGRAPH*, 2022, pp. 1–9.
[54] H. Sak, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.
[55] A. Pavel, B. Hartmann, and M. Agrawala, "Shot orientation controls for interactive cinematography with 360 video," in *ACM SUIST*, 2017, pp. 289–297.
[56] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *IJCV*, vol. 56, pp. 221–255, 2004.
[57] S. Cha, J. Lee, S. Jeong, Y. Kim, and J. Noh, "Enhanced interactive 360° viewing via automatic guidance," *ACM TOG*, vol. 39, no. 5, pp. 1–15, 2020.
[58] A. Truong, S. Chen, E. Yumer, D. Salesin, and W. Li, "Extracting regular fov shots from 360 event footage," in *CHI CHFCS*, 2018, pp. 1–11.
[59] J. Gutiérrez, E. David, Y. Rai, and P. Le Callet, "Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images," *SPIC*, vol. 69, pp. 35–42, 2018.
[60] P. Sedgwick, "Pearson's correlation coefficient," *Bmj*, vol. 345, 2012.
[61] E. Porcu, M. Bevilacqua, and M. G. Genton, "Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere," *JASA*, vol. 111, no. 514, pp. 888–898, 2016.
[62] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.
[63] P. M. Kulkarni, M. Venkatesh, J. Hemalatha, and C. Sunanda, "Empowering engineering training: A portable pyqt5-based gui for system validation with simulink dlls," in *CSITSS*, 2023, pp. 1–5.
[64] G. M. E. Elahi and Y.-H. Yang, "Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition," *Pattern Recognition*, vol. 122, 2022.
[65] Y. Zhang, J. Zhang, R. Liu, P. Zhu, and Y. Liu, "Key frame extraction based on quaternion fourier transform with multiple features fusion," *ESWA*, vol. 216, 2023.
[66] K. Tan, Y. Zhou, Q. Xia, R. Liu, and Y. Chen, "Large model based sequential keyframe extraction for video summarization," in *ICCMLDS*, 2024, pp. 1–5.
[67] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *BJAST*, vol. 7, no. 4, pp. 396–403, 2015.
[68] J. Dawes, "Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales," *IJMR*, vol. 50, no. 1, pp. 61–104, 2008.
[69] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *ICCV*, 2023, pp. 20 144–20 154.
[70] A. T. Bahill, M. R. Clark, and L. Stark, "The main sequence, a tool for studying human eye movements," *Mathematical biosciences*, vol. 24, no. 3-4, pp. 191–204, 1975.
[71] E. G. Freedman and D. L. Sparks, "Coordination of the eyes and head: movement kinematics," *Experimental brain research*, vol. 131, no. 1, pp. 22–32, 2000.
[72] H. Ding, C. Liu, S. He, X. Jiang, and C. C. Loy, "Mevis: A large-scale benchmark for video segmentation with motion expressions," in *CVPR*, 2023, pp. 2694–2703.
[73] C. Chen, G. Wang, C. Peng, Y. Fang, D. Zhang, and H. Qin, "Exploring rich and efficient spatial temporal interactions for real-time video salient object detection," *IEEE TIP*, vol. 30, pp. 3995–4007, 2021.
[74] Y. Kong, Y. Wang, A. Li, and Q. Huang, "Self-sufficient feature enhancing networks for video salient object detection," *IEEE TMM*, vol. 25, pp. 557–571, 2021.
[75] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *CVPR*, 2024, pp. 16 901–16 911.
[76] Y. Zhang, W. Hamidouche, and O. Deforges, "Channel-spatial mutual attention network for 360º salient object detection," in *ICPR*, 2022, pp. 3436–3442.
[77] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.
[78] Y. Su, Y. Cao, J. Deng, F. Rao, and Q. Wu, "Spatial-semantic collaborative cropping for user generated content," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4988–4997.
[79] M. Zhong, J. Sun, F. Wang, and F. Sun, "Lesod: Lightweight and efficient network for rgb-d salient object detection," *Pattern Recognition*, vol. 171, p. 112103, 2026.
[80] N. Liu, L. Li, W. Zhao, J. Han, and L. Shao, "Instance-level relative saliency ranking with graph reasoning," *IEEE TPAMI*, vol. 44, no. 11, pp. 8321–8337, 2021.
[81] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Inferring attention shifts for salient instance ranking," *IJCV*, vol. 132, p. 964–986, 2024.
[82] C. Sun, Y. Xu, J. Pei, H. Fang, and H. Tang, "Partitioned saliency ranking with dense pyramid transformers," in *ACM MM*, 2023, pp. 1874–1883.
[83] B. Deng, S. Song, A. P. French, D. Schluppeck, and M. P. Pound, "Advancing saliency ranking with human fixations: Dataset models and benchmarks," in *CVPR*, 2024, pp. 28 348–28 357.
[84] V. Iashin and E. Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," *arXiv preprint arXiv:2005.08271*, 2020.
[85] V. Iashin and E. Rahtu, "Multi-modal dense video captioning," in *CVPRW*, 2020, pp. 958–959.
[86] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *ICCV*, 2021, pp. 6847–6857.