Exploring Entropy-Based Solutions for Trajectory Prediction in Virtual Reality

Varun Pradhan
CWI, TU Delft
Amsterdam, the Netherlands
Delft, the Netherlands
varunpradhan117@gmail.com

Silvia Rossi
Centrum Wiskunde & Informatica (CWI)
Amsterdam, the Netherlands
s.rossi@cwi.nl

Pablo Cesar
CWI, TU Delft
Amsterdam, the Netherlands
Delft, the Netherlands
p.s.cesar@cwi.nl

Abstract

This work explores the potential of entropy-based metrics to enhance the prediction of user navigation in Virtual Reality (VR). Specifically, we consider three entropy-based metrics: entropy of trajectories, which measures the overall variability and predictability of user behaviour; instantaneous entropy, which provides real-time assessments of movement predictability; and entropy of saliency maps, which offers insights into content-driven attention patterns. Through an exploratory behavioural analysis, we show that users with low entropy exhibit consistent and predictable navigation patterns, while high-entropy users pose greater challenges for prediction models. Building on these findings, we introduce three novel entropy-based solutions for VR trajectory prediction: a position-only baseline augmented with entropy information, an LSTM-based architecture with an entropy-based adaptive attention layer (E-AALSTM), and a multi-head attention-based architecture with adaptive attention (AMH). The proposed models performs as good as state-of-the-art methods, while improving stability and robustness in specific scenarios. This work highlights the importance of having an holistic metric to characterise the user behaviour in VR, and thus enhance trajectory prediction frameworks.

CCS Concepts

• **Human-centered computing** → **Virtual reality**; *User models*; • **Computing methodologies** → **Neural networks**; Artificial intelligence; • **Information systems** → *Data streaming*.

Keywords

Virtual Reality, Trajectory Prediction, User Behaviour Analysis, Predictive Model, Deep Learning, Entropy-based Adaptive Model

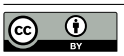
ACM Reference Format:
Varun Pradhan, Silvia Rossi, and Pablo Cesar. 2025. Exploring Entropy-Based Solutions for Trajectory Prediction in Virtual Reality. In *The 17th International Workshop on IMmersive Mixed and Virtual Environment Systems (MMVE '25), March 31-April 4, 2025, Stellenbosch, South Africa.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3712677.3720460

1 Introduction

Virtual Reality (VR) has significant growth in recent years, driven by advancements in computing technologies and its ability to create new and highly interactive digital experiences. Unlike traditional media, VR immerses individuals in a fully interactive virtual world, typically represented by a 360° or spherical video. This novel interactivity enhances users sense of presence and immersion [18, 24], making VR a valuable tool in diverse domains, including entertainment healthcare, engineering and education [14]. However, technical constraints, such as bandwidth limitations, computational demands, and latency pose significant challenges to delivering high-quality and immersive VR experiences. A promising solution is predicting users navigation trajectories (i.e., user behaviour) within the virtual space, enabling systems to mitigate latency, optimize resource allocation, and ensure seamless interactions [6, 9].

Trajectory prediction in VR has evolved significantly, from early regression models [5, 8, 21] to more advanced deep learning solutions. While sophisticated architecture, such as clustering [2, 17, 19], LSTM-based [1, 23], transformer-based methods [4, 12] and meta-learning approaches [13, 16], captures complex patterns and long-term dependencies, they overlook the role of behavioural insights derived from comprehensive data exploration. In fact, a growing attention has been recently put on the analysis and understanding of navigation trajectories in VR, highlighting key behavioural factors that play a crucial role in navigation patters. Content attributes can significantly affect fixation patterns and exploratory behaviours [11, 20, 25, 29]. For example, high spatial and low temporal complexity often result in more focused user attention, while content with no clear focus encourage slow exploratory head movements [30]. Individual preferences play also a role, with some users having a tendency to be more static or scattered independently by the content [26, 27]. A key insight from these behavioural analysis is the role of *entropy-based metrics* in quantifying and characterizing the predictability of human trajectories, resulting to be appealing for improving the accuracy of VR navigation movement predictions.

In this context, three entropy-based metrics emerge as key tools to analyse and model user navigation patterns. First, the *entropy of trajectory* quantifies the overall variability in user movement, distinguish between highly predictable trajectories and those characterized by randomness or unpredictability [26, 28]. Second, *instantaneous entropy* measures the evolving predictability of user movements in real-time, offering a finer temporal granularity [3]. Lastly, *entropy of saliency maps* evaluates the distribution of salient regions within a scene, capturing the relation between content-driven attention and navigation behaviour. These metrics, summarized in Table 1, form the basis of our approach aimed at exploring the potential of entropy metrics to capture and incorporate unpredictability into user navigation trajectories prediction.



MMVE '25, March 31-April 4, 2025, Stellenbosch, South Africa

ACM ISBN 979-8-4007-1468-9/2025/03
https://doi.org/10.1145/3712677.3720460

Metric (Notation)	Definition	Description
Entropy of trajectories (H_{traj})	$H_{traj}(\mathbf{P}) \approx (\frac{1}{T}\sum_{t=1}^{T}\lambda_t)^{-1}\log_2(T)$	Measures the predictability of an entire navigation trajectory.
Instantaneous entropy of trajectories (H^I_{traj})	$H^I_{traj}(\mathbf{P}_t) \approx (\frac{1}{t}\sum_{k=1}^{t}\lambda_k)^{-1}\log_2(t)$	Based the entropy of trajectories but offering temporal granularity.
Entropy of saliency maps (H_{sal})	$H_{sal}(\mathbf{S}) = -\sum_{i=1}^{N} s_i^2 log(s_i^2)$	Quantifies the alignment of head position across users in a frame.

Table 1: Key features of entropy-based metrics.

In this work, we begin with an exploratory analysis to investigate how content and individual differences influence navigation trajectories using entropy-based metrics. Our results confirm previous studies [27] showing that users with low entropy, who exhibit consistent and focused navigation patterns, are more predictable and bring lower prediction errors; conversely, users with high entropy, characterized by more erratic or exploratory behaviour, pose greater challenges for prediction models. Given these observations, we propose and evaluate three novel entropy-based solutions that dynamically allocate higher attention to more predictable and stable segments of users trajectory. We present a position-only baseline augmented with entropy information (*pos_augmented*), where entropy is directly appended as an input feature, together with two adaptive attention models, an LSTM-based architecture with an entropy-based adaptive attention layer (*E-AALSTM*), and a multi-head attention-based architecture (*AMH*). While the entropy-based models perform similarly to state-of-the-art methods, they enhance the stability and robustness of head pose prediction in specific scenarios. In summary, this work addresses the question: "how can behavioural metrics, such as entropy-based ones, improve the predictability and robustness of VR navigation trajectory models?". By emphasizing the importance of data exploration, this work highlights the role of entropy metrics as a bridge between algorithmic advancements and holistic user modelling in VR, with potential applications beyond trajectory prediction, including user-user-centric personalization and adaptive system design.

2 Methodology

We now introduce key elements of our study. Specifically, we define the entropy-based metrics (Table 1) used for our explorative analysis and in the proposed prediction mechanisms. We also outline the baseline models and navigation datasets employed in this study.
To ensure clarity and consistency, we now introduce the following notation. VR dataset collects navigation trajectories of a set of user who displayed a set of 360° videos. Viewers are provided by a VR device – typically a head-mounted display (HMD), that allows changing viewport according to their viewing direction. The sequence over time of user's viewing direction represents the navigation within the immersive content [24]. Thus, a VR trajectory of a given user can be denoted as $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T]$, where $\mathbf{p_t}$ denotes the head position at time t and can be represented in different formats, *e.g.*, Quaternion, Euler angles or Cartesian coordinates.

2.1 Entropy-Based Metrics

Entropy-based metrics quantify the variability of user navigation patterns, offering deeper insights into how users interact with immersive content [27]. Specifically, the *entropy of trajectories* H_{traj} measures the overall predictability of a user's trajectory by assessing the probability of repeated movement subsequences [31]. In our case, we discretize the spherical space of the 360° video, representing a trajectory $\mathbf{P}$ as a sequence of spatial blocks B. Thus, the trajectory at time t is defined as $P_t = [b_t, b_{t+1}, ..., b_{t-1+\lambda_t}]$, where λ_t is the length of the shortest subsequence starting at t that does not appear earlier in P. Since true probabilities cannot be obtained in real-world scenarios, we estimate the entropy of the trajectories using the Lempel-Ziv compression algorithm [34]:

$$H_{traj}(\mathbf{P}) \approx (\frac{1}{T}\sum_{t=1}^{T}\lambda_t)^{-1}\log_2(T) \tag{1}$$

where T is the total trajectory length. Higher values of H_{traj} indicate more unpredictable user navigation, while lower values reflect regular and repetitive patterns [27].
While H_{traj} provides a holistic measure over an entire trajectory, it does not capture temporal dynamics. To address this, we introduce the *instantaneous entropy* H^I_{traj}, which offers a real-time perspective by quantifying the predictability of user movements at any given moment [3]. At time t, it is derived from the cumulative entropy H_{traj} up to that point:

$$H^I_{traj}(\mathbf{P}_t) \approx (\frac{1}{t}\sum_{k=1}^{t}\lambda_k)^{-1}\log_2(t) \tag{2}$$

where λ_k is the shortest unseen subsequence in $\mathbf{P}_t$. This approach dynamically captures the evolving predictability of user behaviour in real-time, making it suitable for adaptive prediction algorithms. While H_{traj} and H^I_{traj} focus on trajectories of the same individual, the *entropy of saliency maps* H_{sal} quantifies the alignment of head position across users for a given frame [30]. More formally, given a single saliency map S of N pixels, H_{sal} is computed as the Shannon entropy of the frame:

$$H_{sal}(S) = -\sum_{i=1}^{N} s_i^2 log(s_i^2) \tag{3}$$

where s_i is a pixel bellowing to S. In this case, a low H_{sal} indicates that user focus is highly concentrated in specific regions, while a high H_{sal} suggests dispersed fixations across the scene, reflecting more varied attention patterns [30].

2.2 Baselines for Trajectory Prediction

In our study, we selected the following baselines based on their performance, uniqueness of approach, and availability of code:

- **Pos-only** [23] is a baseline LSTM based model that uses the seq2seq encoder-decoder architecture. As input, it only takes the user's head position over the input window;

- **TRACK** [23] is a variant of the seq2seq encoder-decoder model that processes content saliency maps with an RNN before merging it with head data. This approach ensures that the content information is attenuated in the short term while still considering its impact in later steps through the recurrent network's representation;
- **DVMS** [10] also uses a seq2seq decoder architecture, but it utilizes a latent variable, z, to generate multiple predictions. This introduces a degree of uncertainty in the head trajectories and allows to train the model based on the best prediction for each timestep. Specifically, we evaluate two versions: DVMS_2 and DVMS_5 trained to generate and 5 predictions per input, respectively, following [10];
- **VPT360** [4] utilizes a transformer-based architecture with multi-head attention, predicting the user's trajectory over the entire output horizon in a single step.

These baselines are included in the unified collection presented in [22]. However, the original framework performs the train-test split at video level by selecting a fraction of videos and training models on all user trajectories for the selected videos. This approach can lead to data contamination, as user trajectories from the same users might appear in both the training and test sets, albeit on different videos. This overlap could cause information leakage and affect performance metrics. To address this issue, in this work we randomly selected 60% of the videos and 60% of the users for the training set, leaving the remaining user-video pairs for testing. This approach ensures that the models are evaluated on previously unseen user-video pairs, allowing for a more accurate comparisons.

2.3 User Navigation Datasets

We consider 4 publicly available datasets which collect user navigation data during immersive VR experiences, specifically:

- **NOSSDAV17** [8] is composed of 10 videos of duration 60 seconds and bellowing to three types of video content: *Fast pace, computer generated*, *Fast pace, natural image*, and *Slow pace, natural image*. The trajectories of 30 participants have been collected and saliency maps are also available;
- **MMSys18** [7] consists of 57 navigation trajectories collected on 19 videos in indoor and outdoor settings in rural and urban environments. Each video is 20 seconds long and the dataset also contains saliency maps for each video;
- **PAMI18** [33] is composed of head-motion trajectories collected by 58 viewers in 76 different VR content. The videos range from 10 to 80 seconds long and include computer animation, acting, sports, driving and scenery content;
- **MM22** [11] consists of trajectories collected over a diverse collection based on video quality, amount of camera movement, and dispersion of regions of interest. Data is collected from 100 users in 27 videos, 20 seconds long.

3 Entropy-based Behavioural Analysis

As first step of our exploratory analysis, we investigate the relation between entropy of trajectories H_{traj} and saliency maps H_{sal} across the different datasets. As described in Section 2.1, the first quantifies the predictability of an entire user's trajectory, with higher values indicating more erratic and less structured movement

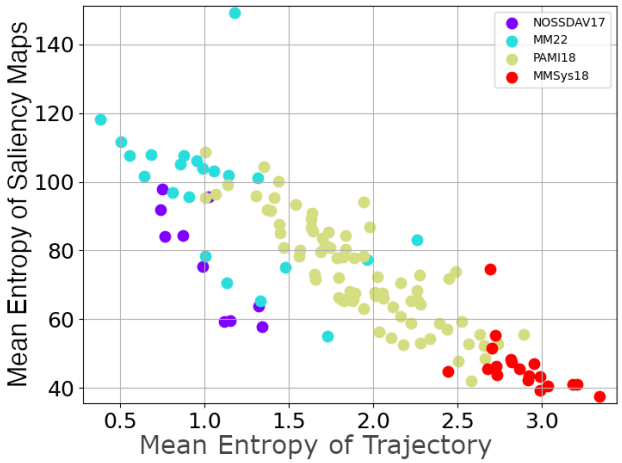
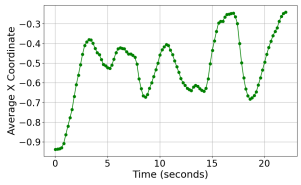


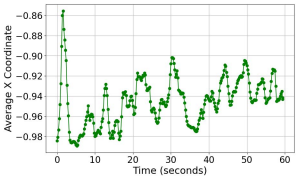
Figure 1: Mean H_{traj} vs Mean H_{sal} for videos across selected datasets (i.e., NOSSDAV17, MM22, PAMI18, MMSys18)

(a) *VRBasketball*



(b) *VRBasketball*: average X-coordinate

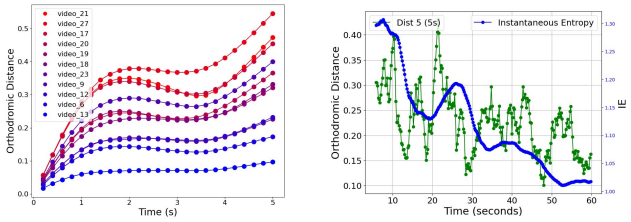
(c) *Mario*



(d) *Mario*: average X-coordinate

Figure 2: Representative scenes from *VRBasketball* (a) and *Mario* (c). Along with the average X coordinate of user trajectories over the course of *VRBasketball* (b) and *Mario* (d).

patterns [27]. In contrast, entropy of saliency maps captures the distribution of user alignment within a given frame, where lower values indicate a strong consensus [30]. Specifically, each dot in Figure 1 represents mean value of H_{traj} and H_{sal} in a video per each datasets described in Section 2.3. We can observe a consistent negative correlation between the entropy of trajectories and saliency maps across all four datasets. This suggests that as the average H_{sal} of a video decreases, user trajectories become more unpredictable, potentially due to heightened engagement levels where viewers exhibit faster movements that involve following a strong center of focus. There is also considerable variation in both H_{traj} and H_{sal} across the datasets, reflecting the diversity of user behaviours and video content. This negative correlation may seem unexpected, but as observed in [30], viewers tend to navigate scenes with very defined salient areas (low H_{sal}) more quickly, jumping form one focus to another. This rapid navigation results in highly random navigation patterns (high H_{traj}).

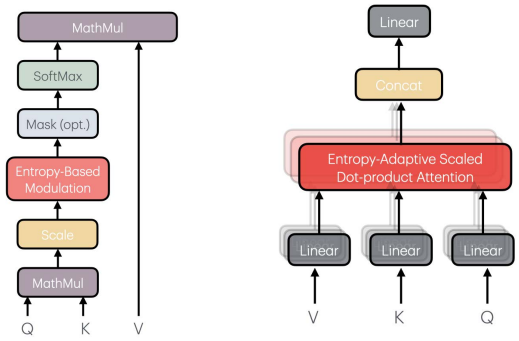


(a) *Orthodromic Distance Loss per video* (b) *Orthodromic distance vs IE*

Figure 3: (a) Performance of DVMS_5 on each video in MM22; (b) Orthodromic distance vs H^I_{traj} in video_23 (Chariot)

For more fine-gained exploration, we analyse two selected videos: *VRBasketball* and *Mario*, from PAMI18 [33] and MM22 [11] datasets, respectively. *VRBasketball*, characterised by a high mean H_{traj} (2.89) and a relatively low mean H_{sal} (55.56), is composed by two static basketball players passing a ball to each other, creating a dynamic center of focus (Figure 2 (a)). This dynamic object explains the relatively low value of H_{sal}. The high H_{traj} is, instead, due to the unpredictable and rapid movements of the basketball across the scene, which lead to more variable head movements as viewers track the ball. Since most of the movements occur on the horizontal plane, we plot the average X-coordinate of all the users' head positions in *VRBasketball* in Figure 2 (b). Users show rapid movements when the players pass the ball to each other (at timestamps 3 seconds, 6 seconds, and 16 seconds) or perform some tricks. For instance, at 10 seconds, one player performs dribbling tricks while moving with the ball. This suggests that a dynamic center of focus with unpredictable movements negatively influence the predictability of navigation trajectories. In contrast, *Mario*, with the lowest mean H_{traj} (0.37), and a high mean H_{sal} (118.18), has a static center of focus at the center of the frame, as shown in Figure 2 (c). The average X-coordinate over time, presented in Figure 2 (d), reveals a very stable trajectory, with the average X-coordinate being relatively unchanged with values between -0.98 and -0.86. This stability explain the low H_{traj}, as users exhibit slower and more consistent movements while viewing the video. This comparison highlights that videos with dynamic focal points, such as *VRBasketball*, lead to more unpredictable head movements as viewers actively track moving objects. In contrast, videos with static centers of attention, like *Mario*, result in slower and more predictable head movements due to the absence of significant visual stimuli.

To gain deeper insights into how entropy metrics reflect the predictability of users trajectories, and building on the work presented in [27], we compare the complexity of the navigation trajectories (measured by the entropy of trajectories) with their ability to be predicted. Specifically, we evaluate prediction performance in terms of orthodromic loss of the latest baseline in Section 2.2, DVMS_5, on each video in the MM22 dataset. Figure 3 (a) shows this loss with a 5 seconds prediction window, with videos ID ordered in the legend from the highest to the lowest mean H_{traj} (top to bottom). The results indicate that the model's accuracy tends to decrease for videos where users exhibit high entropy. Specifically, videos with the highest mean H_{traj} such as *video_21, video_27, video_17* also exhibit higher loss values. This aligns with findings in [27], confirming that trajectories with higher entropies are more challenging to predict due to the increased variability and unpredictability in user



(a) *Entropy-Adaptive Scaled dot-product attention* (b) *Multi-Head Entropy-Adaptive Attention*

Figure 4: Entropy-attention layers [32].

movements. For example, Figure 3 (b) shows the average prediction loss for DVMS_5 over time in the video *Chariot*, along with its H^I_{traj} highlighting a positive correlation between them.
In conclusion, our exploratory analysis highlights the diversity of 360° video content in the selected datasets, offering a broad spectrum of user behaviours and entropy values. This diversity is crucial for ensuring that trajectory prediction models generalize effectively across different content types. Additionally, our analysis of entropy correlations offers valuable insights: a negative correlation between H_{traj} and H_{sal} and a positive correlation between prediction error and H^I_{traj}. These findings suggest that integrating entropy metrics into prediction models could enhance their performance.

4 Entropy-Based Solutions

Given the previous observations, we propose three entropy-based models that leverage entropy as a measure of user behaviour predictability to enhance head pose prediction in VR.

Entropy Enhanced Baseline (pos_augmented)
This model is a sequence-to-sequence (seq2seq) encoder-decoder model extends the position-only baseline [23] by appending the instantaneous entropy H^I_{traj} up to the current time as an additional input feature. By incorporating entropy, the model gains a deeper understanding of the predictability of the user's movements, enhancing its predictions. While it does not employ an explicit attention mechanism, the integration of entropy allows the model to adapt its predictions based on the variability of user behaviour.

Entropy-based Adaptive Attention LSTM (E-AALSTM)
We propose an LSTM-based adaptive attention architecture enhanced with entropy information, E-AALSTM, designed to address the challenges posed by variability in user navigation behaviour. Built on the standard seq2seq encoder-decoder with attention layer [32], E-AALSTM introduces a novel entropy-based adaptive attention mechanism which prioritises parts of the trajectory where predictability is higher, corresponding to lower values of instantaneous entropy H^I_{traj}. Specifically, we integrate a novel *entropy-based modulating factor* in the scaled dot-product attention layer [32] as shown in Figure 4 (a). This factor dynamically adjusts the

attention scores based on H^I_{traj} of the user's trajectory P at each time-step. More formally, the modulating factor $\mathbf{M}_t$ is defined as:

$$\mathbf{M}_t = \exp(-\mathbf{W_s} \cdot H^I_{traj}(P_t)) \tag{4}$$

where W_s are learnable weights that control the influence of the entropy on the attention mechanism, and P_t is the user trajectory up to time t. The modulating factor $\mathbf{M}_t$ is then applied to the traditional attention mechanism by scaling the attention scores. The modified attention weights AW_{adpt} are computed as:

$$AW_{adpt} = \mathbf{M}_t \cdot \text{softmax}(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k}) \tag{5}$$

where $\mathbf{Q}$, $\mathbf{K}$ and d_k are the query, key, and the dimensionality of the key vectors, respectively. Finally, the context vector, which captures the weighted representation of the input sequence adjusted by the entropy-based modulating factor, is then computed as:

$$c_t^{\text{adaptive}} = AW_{adpt} \cdot \mathbf{V} \tag{6}$$

where $\mathbf{V}$ represents the value matrix. Finally, the context vector is passed to linear layers to generate the final prediction.

Multi-head Adaptive Attention (AMH)

We propose a transformer-based adaptive attention model, AMH, inspired by the VPT360 architecture [4]. AMH leverages an *entropy-based modulation factor* defined in Equation 4 to adjust attention scores dynamically, as evaluated by multi-head entropy-adaptive attention module shown in Figure 4 (b). Unlike E-AALSTM, which iteratively adapts predictions based on entropy values, AMH uses entropy values from the input window to modulate attention scores for the entire sequence in parallel. This fixed, entropy-informed modulation makes AMH computationally more efficient while still incorporating the benefits of entropy-based adaptability. First, the input sequence $\mathbf{P}$ is transformed into embeddings, incorporating positional information as in [32]. These embeddings, $\mathbf{e_P}$, serve as the input to the multi-head entropy-adaptive attention mechanism. For each attention head $h \in \{1, ..., H\}$, the query, key, and value matrices are computed from the input embeddings as follows:

$$\mathbf{Q}_h = \mathbf{W}_h^q \mathbf{e_P} + \mathbf{b}_h^q, \quad \mathbf{K}_h = \mathbf{W}_h^k \mathbf{e_P} + \mathbf{b}_h^k, \quad \mathbf{V}_h = \mathbf{W}_h^v \mathbf{e_P} + \mathbf{b}_h^v \tag{7}$$

where $\mathbf{e}_P$ is the input embedding, and W_h^q, W_h^k and W_h^v are learnable projection matrices specific to head h. The entropy-based modulation factor $\mathbf{M}_t$ (as defined in Equation 4) is applied to adjust the attention scores for each head:

$$\mathbf{A}_h^{\text{adapt}} = \mathbf{M}_t \cdot \text{softmax}\left(\mathbf{Q}_h \mathbf{K}_h^\top / \sqrt{d_k}\right) \tag{8}$$

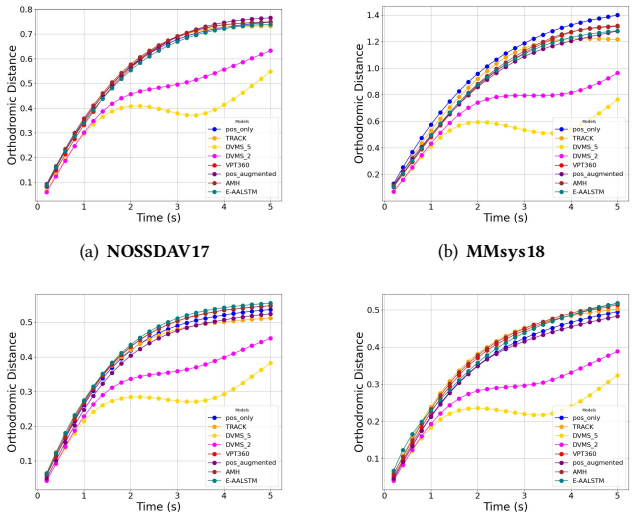
The modulated attention weights are then used to compute the context vector for each head:

$$\mathbf{C}_h = \mathbf{A}_h^{\text{adapt}} \cdot \mathbf{V}_h. \tag{9}$$

The outputs from all heads are concatenated and projected using a linear transformation to obtain the final attention output:

$$\mathbf{C}_{\text{AMH}} = \text{Concat}(\mathbf{C}_1, \ldots, \mathbf{C}_H)\mathbf{W}^O \tag{10}$$

where $\mathbf{W}^O$ is a learnable output projection matrix. Finally, $\mathbf{C}_{\text{AMH}}$, which represents the entropy-modulated multi-head attention context, is passed through additional feed-forward and linear layers to generate the final head trajectory predictions.



(a) **NOSSDAV17** (b) **MMsys18**

(c) **PAMI18** (d) **MM22**

Figure 5: Performance comparisons on selected datasets

5 Evaluation and Performance Analysis

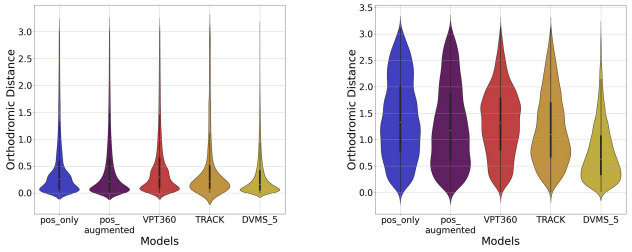
Experimental Setup

To evaluate the performance of our proposed models, we compare them with the baselines described in Section 2.2. To ensure that performance differences are solely due to architectural differences, we standardized hyperparameter configurations across all models. Specifically, the learning rate was set to 0.0001, and the AdamW optimizer [15] with a weight decay of 0.01 was used. Training was conducted for a maximum of 500 epochs with early stopping if the loss did not improve for 25 consecutive epochs, and a batch size of 128 was chosen to balance computational efficiency and training stability. In our proposed entropy modulation factor, weights were initialized randomly using a normal distribution, $\mathcal{N}(0, 0.1^2)$.

We evaluate our proposed models on the four datasets discussed in Section 2.3, and we follow the same train-test splitting approach utilized for our exploratory analysis. Specifically, we perform a stratified split based on the mean entropy of saliency maps and the mean actual entropy of user trajectories, ensuring both training and testing data represented the overall dataset. Videos and users were split into a 60-40 ratio, ensuring that the models are evaluated on data excluded from the training process for both categories. Performance is measured in terms of orthodromic distance between the predicted head position, $\hat{P}_t$, and the true position, P_t.

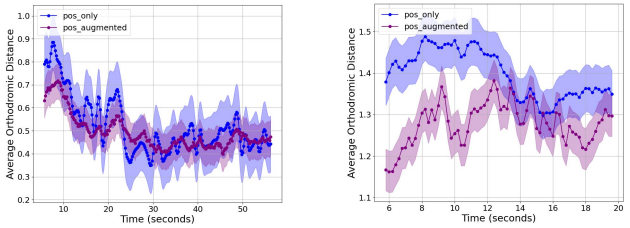
Results

Given the above settings, we compare the performance of our proposed models (i.e., pos_augmented, E-AALSTM and AMH) with the baseline (i.e., pos-only, TRACK, VPT360 and DVMS) described in Section 2.2. Figure 5 shows the average performance across videos in the selected datasets at each timestep up to the 5-second prediction horizon. Among the evaluated models, DVMS_5 achieves the lowest orthodromic distance, outperforming both the baselines and our proposed models. This shows the effectiveness of the multiple prediction training approach of DVMS, enabling it at better capturing head trajectory dynamics. Focusing on our proposed models, pos-augmented demonstrates promising results, slightly



(a) **MM22** (b) **MMsys18**

Figure 6: Violin plots of Orthodromic distance for selected models on (a) MM22, and (b) MMSys18 datasets



(a) **MM22** (b) **MMsys18**

Figure 7: Average prediction loss at each timestep for pos-only, and pos-augmented on (a) MM22, and (b) MMSys18.

outperforming E-AALSTM and AMH across most datasets except for NOSSDAV17 (Figure 5 (a)). It achieves competitive short-term predictions, outperforming TRACK up to 3 seconds on PAMI18 in Figure 5 (c) and both VPT360 and TRACK on the MM22 dataset in Figure 5 (d). We also note that all models tend to perform worse on MMSys18 dataset, as the losses exceed 0.6 shortly after the first second, likely due to its shorter videos. Indeed, users tend to have higher trajectory entropy due to their rapidly exploring movements at the beginning of the immersive experience. These findings confirm that higher trajectory entropy introduce greater unpredictability, posing challenges for accurate prediction. Given these observation, we select the pos-augmented model as the most robust among our proposed solutions. To evaluate model stability, we computed violin plots of the errors across the prediction steps (5 seconds into the future), shown in Figure 6 in two selected datasets, MM22 and MMsys18. DVMS_2 was excluded, as its distribution is similar to DVMS_5 with a slightly higher mean error. As expected, DVMS_5 exhibited the most stable distributions, consistently skewed towards lower error values. In the MM22 dataset (Figure 6 (a)), both VPT360 and pos-only exhibit a prominent main cluster of errors centred around 0.08, with a secondary band visible above this range. In contrast, pos-augmented and TRACK lack this secondary band, as their distributions taper off more smoothly from their respective modes at 0.06 for pos-augmented and 0.13 for TRACK. While TRACK lacks the secondary band, its main error cluster is situated higher, reflecting a broader spread of errors compared to the other models. Similarly, for the MMSys18 dataset (Figure 6 (b)), we observe that the loss distribution for VPT360 is skewed towards higher losses, with its mode at 1.59, indicating a larger proportion of higher error values. In contrast, the TRACK model's distribution is skewed towards lower losses, but its central tendency occurs at higher error levels, with a mode of 1.17, compared to the pos-augmented model, which has a mode of 0.55. This suggests that, while the TRACK model has a tendency towards lower losses than VPT360, its errors still tend to be more concentrated in the higher error range compared to the pos-augmented model, which demonstrates the most favourable distribution with its mode of 0.34. Finally, we compare the position-only and the pos-augmented models by analysing the average orthodromic distance between the predicted head pose and the true head pose. Specifically, Figure 7 shows the averaged prediction error, for each time step across all test videos in the MM22, and MMSys18 datasets. We also include the standard deviation at each time-step to study the stability of the models. In MM22 (Figure 7 (a)), the performance of the two models are quite similar, even if pos-augmented model showed a slightly better average error (0.49 vs. 0.53). More interestingly, for MMSys18 (Figure 7 (b)), the pos-augmented model outperforms the baseline, with greater accuracy, particularly in the first half of videos where users exhibit more exploratory behaviour. This highlights the value of instantaneous entropy early in sessions for handling more unpredictable behaviour. We also calculate the coefficient of variation (CV) for MM22, which normalizes variance relative to the mean. The pos-augmented model achieved a CV of 15.4%, compared to 23.4% for the baseline, demonstrating more stable and reliable predictions. This consistency is likely due to the incorporation of entropy-based features, enabling the model to better handle variability in user behaviour.

To conclude, these findings highlight the potential of integrating trajectory entropy into head pose prediction models to enhance their performance, especially in scenarios characterized by highly unpredictable user behaviours. Among the proposed approaches, the pos-augmented emerges showing meaningful improvements, demonstrating enhanced stability and accuracy under high-entropy conditions. However, it is important to acknowledge the current limitations regarding real-time applicability. At present, for sequences of 60 seconds, the entropy calculation introduces an additional 5–6 seconds per batch, making real-time implementation infeasible. Future efforts should focus on reducing computational overhead to enable deployment on real-time VR systems.

6 Conclusion

This work explored the role of trajectory entropy in improving trajectory prediction models for VR navigation. Through extensive evaluation across multiple datasets, we demonstrated that integrating entropy-based features enhances prediction performance, particularly under conditions of high user behaviour variability. Among the proposed solutions, the pos-augmented model exhibited the best balance of accuracy and stability, performing similarly to baseline approaches and achieving robust results across diverse scenarios. While these results highlight the potential of entropy-based features, real-time applicability remains a challenge. Currently, entropy computation introduces a processing overhead, requiring further optimization. Some failure cases were observed in highly dynamic settings, suggesting the need for adaptive mechanisms to handle extreme unpredictability. Addressing these aspects in future work could further improve the feasibility and robustness of entropy-based trajectory prediction opening avenues for personalized VR experiences.

Acknowledgments

This work has been in part supported through the European Commission Horizon Europe program under grant 101070109 TRANSMIXR (https:// transmixr.eu/).

References

[1] Avi M. Aizenman, George A. Koulieris, Agostino Gibaldi, Vibhor Sehgal, Dennis M. Levi, and Martin S. Banks. 2023. The statistics of eye movements and binocular disparities during VR gaming: Implications for headset design. *ACM transactions on graphics* (2023).

[2] Yixuan Ban, Lan Xie, Zhimin Xu, Xinggong Zhang, Zongming Guo, and Yue Wang. 2018. Cub360: Exploiting cross-users behaviors for viewport prediction in 360 video adaptive streaming. In *International Conference on Multimedia and Expo*. IEEE, 1–6.

[3] Paul Baumann and Silvia Santini. 2013. On the use of instantaneous entropy to measure the momentary predictability of human mobility. In *Workshop on Signal Processing Advances in Wireless Communications*. IEEE, 535–539.

[4] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. 2021. Transformer-based Long-Term Viewport Prediction in 360° Video: Scanpath is All You Need.. In *International Workshop on Multimedia Signal Processing*. IEEE, 1–6.

[5] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. 2022. Privacy-Preserving Viewport Prediction using Federated Learning for 360° Live Video Streaming. In *International Workshop on Multimedia Signal Processing*. IEEE.

[6] Federico Chiariotti. 2021. A survey on 360-degree video: Coding, quality of experience and streaming. *Computer Communications* 177 (2021), 133–155.

[7] Erwan J. David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A dataset of head and eye movements for 360 videos. In *Proceedings of the Multimedia Systems Conference*. ACM.

[8] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the workshop on network and operating systems support for digital audio and video*. 67–72.

[9] Monalisa Ghosh and Chetna Singhal. 2024. A review on machine learning based user-centric multimedia streaming techniques. *Computer Communications* (2024), 108011.

[10] Quentin Guimard, Lucile Sassatelli, Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2022. Deep variational learning for multiple trajectory prediction of 360° head movements. In *Proceedings of the Multimedia Systems Conference*. ACM.

[11] Yili Jin, Junhua Liu, Fangxin Wang, and Shuguang Cui. 2022. Where Are You Looking? A Large-Scale Dataset of Head and Gaze Behavior for 360-Degree Videos and a Pilot Study. In *Proceedings of the International Conference on Multimedia*. ACM.

[12] ChengDong Lan, Xu Qiu, Chenqi Miao, and MengTing Zheng. 2024. A self-attention model for viewport prediction based on distance constraint. *The Visual Computer* 40, 9 (2024), 5997–6014.

[13] Junjie Li, Yumei Wang, and Yu Liu. 2023. Meta360: Exploring User-Specific and Robust Viewport Prediction in360-Degree Videos through Bi-Directional LSTM and Meta-Adaptation. In *International Symposium on Mixed and Augmented Reality*. IEEE, 652–661.

[14] Matthew J Liberatore and William P Wagner. 2021. Virtual, mixed, and augmented reality: a systematic review for immersive systems research. *Virtual Reality* 25, 3 (2021), 773–799.

[15] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:53592270

[16] Yiyun Lu, Yifei Zhu, and Zhi Wang. 2022. Personalized 360-degree video streaming: A meta-learning approach. In *Proceedings of the International Conference on Multimedia*. ACM, 3143–3151.

[17] Dario DR Morais, Lucas S Althoff, Ravi Prakash, Marcelo M Carvalho, and Mylène CQ Farias. 2021. A content-based viewport prediction model. *Electronic Imaging* 33 (2021), 1–8.

[18] Joschka Mütterlein. 2018. The Three Pillars of Virtual Reality? Investigating the Roles of Immersion, Presence, and Interactivity. In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2018.174

[19] Afshin Taghavi Nasrabadi, Aliehsan Samiei, and Ravi Prakash. 2020. Viewport prediction for 360 videos: a clustering approach. In *Proceedings of the Workshop on Network and Operating Systems Support for Digital Audio and Video*. ACM.

[20] Cagri Ozcinar and Aljosa Smolic. 2018. Visual attention in omnidirectional video for virtual reality applications. In *International conference on quality of multimedia experience*. IEEE, 1–6.

[21] Feng Qian, Lusheng Ji, Bo Han, and Vijay Gopalakrishnan. 2016. Optimizing 360 video delivery over cellular networks. In *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*. 1–6.

[22] Miguel Fabián Romero Rondón, Lucile Sassatelli, Ramon Aparicio-Pardo, and Frédéric Precioso. 2020. A unified evaluation framework for head motion prediction methods in 360° videos. In *Proceedings of the ACM on Multimedia Systems Conference*. ACM. https://doi.org/10.1145/3339825.3394934

[23] Miguel Fabian Romero Rondon, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2021. Track: A new method from a re-examination of deep architectures for head motion prediction in 360-degree videos. *Transactions on Pattern Analysis & Machine Intelligence* 01 (2021), 1–1.

[24] Silvia Rossi, Alan Guedes, and Laura Toni. 2022. Streaming and User Behaviour in Omnidirectional Videos. *Immersive Video Technologies* (2022), 49.

[25] Silvia Rossi, Cagri Ozcinar, Aljosa Smolic, and Laura Toni. 2020. Do Users Behave Similarly in VR? Investigation of the User Influence on the System Design. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2020).

[26] Silvia Rossi and Laura Toni. 2020. Understanding user navigation in immersive experience: an information-theoretic analysis. In *Proceedings of the International Workshop on Immersive Mixed and Virtual Environment Systems*. ACM.

[27] Silvia Rossi, Laura Toni, and Pablo Cesar. 2023. Correlation between Entropy and Prediction Error in VR Head Motion Trajectories. In *Proceedings of the 2nd International Workshop on Interactive eXtended Reality*.

[28] Silvia Rossi, Irene Viola, and Pablo Cesar. 2022. Behavioural Analysis in a 6-DoF VR System: Influence of Content, Quality and User Disposition. In *Proceedings of the 1st Workshop on Interactive eXtended Reality*.

[29] Ana Serrano, Vincent Sitzmann, Jaime Ruiz-Borau, Gordon Wetzstein, Diego Gutierrez, and Belen Masia. 2017. Movie editing and cognitive event segmentation in virtual reality video. *ACM Transactions on Graphics* 36, 4 (2017), 1–12.

[30] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How Do People Explore Virtual Environments?. In *IEEE Transactions on Visualization and Computer Graphics*.

[31] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[32] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[33] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. 2018. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence* 41, 11 (2018), 2693–2708.

[34] Jacob Ziv and Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory* 24, 5 (1978), 530–536.