

Forecasting 3D Scanpaths in Egocentric Video

Fiona Ryan^{1,2*} Ishwarya Ananthabhotla² Yijun Qian² Judy Hoffman^{1,3} James M. Rehg⁴
Vamsi Krishna Ithapu² Calvin Murdock²

¹Georgia Institute of Technology ²Meta Reality Labs Research ³UC Irvine ⁴UIUC

Abstract

Forecasting gaze behavior is an important task for understanding user intent and creating AR/VR systems that can anticipate where users will look and interact next. While prior works have addressed predicting scanpaths in static images, forecasting gaze in egocentric videos presents new challenges due to the dynamic nature of the scene and the camera wearer’s continuous movement through the 3D environment. To address these challenges, we formulate the novel task of egocentric scanpath prediction as forecasting a sequence of future fixations in 3D Cartesian coordinates relative to the last observed camera pose, producing a 3D scanpath that is grounded in the environment. We propose a transformer architecture that leverages egocentric video frames, head pose, and past 3D gaze observations to predict future 3D fixation sequences. We evaluate our method on the Aria Digital Twin dataset. Our findings establish a baseline for the novel task of 3D scanpath prediction and highlight important architectural elements for our task.

1. Introduction

Where a person looks is closely coupled with their intent, offering insight into what they are doing and what they may do next [24, 36, 43]. Forecasting visual attention, or predicting where someone will look next in a scene, is an important task with applications across behavior understanding, robotics, and augmented and virtuality reality (AR/VR) applications. In AR/VR, anticipating gaze targets enables systems to proactively render experiences that are driven by the user’s attention by identifying who or what they may attend to next [3]. Modeling attention in task-driven contexts can also inform robot learning and interaction [4, 5, 63] and algorithms that emulate human active perception [80, 81].

Most prior work on forecasting gaze has focused on the prediction of scanpaths in 2D static images [8, 16, 30, 37, 39, 44, 47, 60, 61, 67]. In this task, a temporally ordered

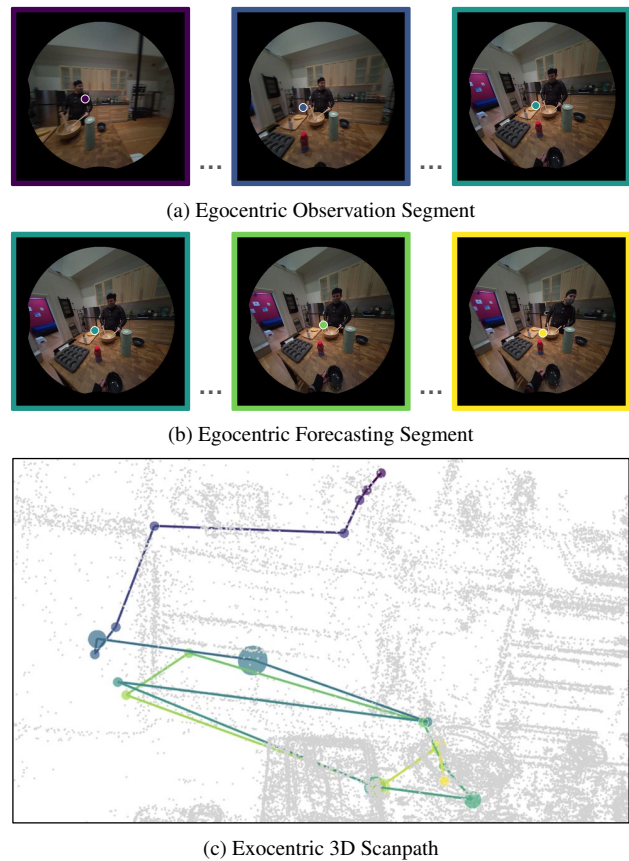


Figure 1. Forecasting gaze in egocentric videos gives insight into user attention and intent. Given a history of observed egocentric video frames and gaze targets (a), we aim to predict future targets of gaze (b). To account for the user’s motion through a dynamic 3D scene across frames, we formulate the problem as 3D scanpath prediction within a consistent exocentric frame of reference (c).

series of fixations is predicted for an image in pixel coordinates. Common datasets such as MIT1003 [37] and COCO-Search18 [15, 73] provide ground truth labels obtained from eye tracking, where gaze measurements are obtained from multiple subjects viewing the same images. Using images and videos displayed on a monitor as visual stimuli for gaze modeling has practical advantages, including enabling dis-

*Work done during internship at Meta.

tributions of responses to standardized visual stimuli, a simplified 2D output space, and eliminating the need to account for head movement during viewing.

While progress on image scanpath prediction represents an important step towards modeling visual attention, viewing an image is a highly constrained scenario compared to real-world gaze behavior. In this work, we formulate the novel task of *forecasting 3D scanpaths in egocentric video* (Fig. 1). We argue that predicting scanpaths in this embodied setting differs fundamentally from the 2D scanpath problem formulation. First, downstream applications in AR/VR rendering, robotics, and planning require gaze predictions that are consistent across multiple egocentric frames of reference, which mandates prediction in a fixed 3D coordinate system persistent across these frames. Second, our problem scenario is significantly more variable and technically complex than the 2D scanpath prediction scenario, as egocentric video data captured from a head-mounted device typically includes rapid head movements, wearer translation throughout the scene, dynamism in the environment, and task-oriented interactions between a user and their environment. Unlike in the image scanpath prediction case, as a person rotates their head and translates rapidly through a scene, a future gaze target at one instance in time may not be visible in subsequent video frames.

This is the first work to study the problem of forecasting 3D gaze scanpaths from egocentric videos. We introduce a novel problem formation that entails predicting gaze targets in a 3D coordinate system that is consistent across changing egocentric views and is grounded in the wearer’s 3D environment. To account for the dynamic nature of egocentric video and the continuous nature of gaze, we include past video frames, head poses, and gaze observations as inputs to our task. We develop a transformer-based architecture for our task and evaluate it against baselines and relevant prior work. Our contributions are as follows:

- We introduce and define the novel task of egocentric 3D scanpath prediction.
- We propose a transformer architecture to predict future 3D gaze points given egocentric video, observed head pose, and previous gaze points.
- We evaluate our model on the Aria Digital Twin dataset [55], establishing a baseline level of performance and identifying key architectural considerations.

2. Related Work

Scanpath Prediction Several works have addressed the task of scanpath prediction, which aims to predict a viewer’s sequence of gaze fixations as they look at an image. This problem builds on a larger body of work addressing predicting visual saliency [9, 29, 33, 65], but differs in its goal to produce a temporally ordered gaze path of fixation locations with durations simulating a human actively perceiv-

ing the image. Early work on modeling dynamic visual behavior generates scanpaths via saliency maps using inhibition of return [30, 66] or information maximization [44, 60, 61, 67]. MIT1003 [37] provided the first large-scale dataset for this task, developed by collecting eye tracking data of humans viewing a set of images. MIT1003 and further image-based eye tracking databases [10, 70] have enabled the development of several architectures for scanpath prediction in free-viewing settings [8, 16, 38, 39, 47, 64, 69, 74]. Recent works have increasingly focused on goal-driven scanpaths, where the user is instructed to view the image with a goal such as searching for an object [15, 18, 20, 52, 73, 77], captioning the image [76], or visual question answering [12, 14]. Another line of research addresses personalization in scanpath prediction by predicting person-specific scanpaths given a support set of scanpaths from an individual viewer [13, 34, 71].

While modeling scanpaths in 2D images has facilitated progress on gaze behavior modeling, viewing a static 2D image is highly constrained compared to naturalistic gaze behaviors in the 3D world. A few architectures extend image scanpath prediction to 360° images to simulate viewing a wider scene, predicting gaze targets as direction on a unit sphere [35, 68]. A related task is viewport prediction in 360° videos [11], which predicts the viewer’s head direction while watching a 360° video with a VR headset. Recent work CT-ScanGaze [57] predicts scanpaths through 3D Computed Tomography volumes, however this constitutes a constrained, domain-specific 3D setting where the depth dimension is limited to image slices. Compared to these works, our formulation addresses predicting scanpaths from egocentric video in the dynamic 3D world, which necessitates predicting gaze targets as 3D world coordinates and accounting for viewer movement through the world.

Egocentric Gaze Modeling Prior work has addressed modeling gaze in egocentric video, but largely operates in the 2D pixel coordinates of individual video frames. Egocentric gaze estimation predicts 2D gaze locations within individual egocentric video frames [27, 40, 41, 45], with some approaches jointly modeling gaze and activity [28, 46, 51]. Zhang et al. [78] and subsequent works [42, 79] extend this formulation to predicting future gaze as 2D locations in image coordinate system of future frames. By predicting future gaze locations in 2D pixel coordinates of *unseen* future frames, this formulation requires models to implicitly predict future head motion in order to determine the frame of reference for future gaze predictions, and predictions cannot easily be interpreted between frames. In this work, we instead forecast gaze in a 3D coordinate system, such that gaze predictions are grounded in a coordinate system that is persistent and shared between individual video frames. Recent work EgoSpanLift [75] forecasts a temporally aggre-

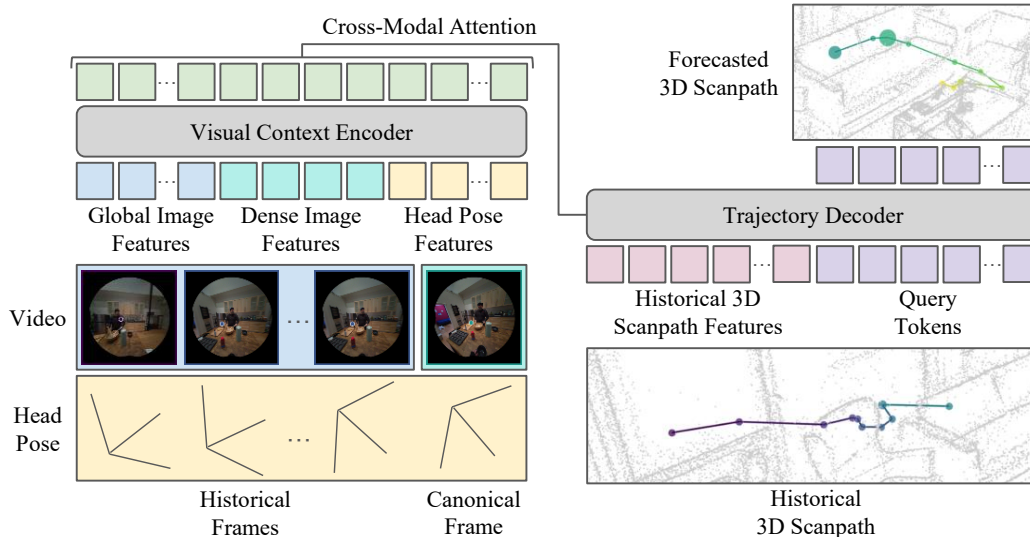


Figure 2. Overview of our proposed architecture. We encode global image and head pose features alongside dense patch features of the last observed video frame. Acting as a canonical reference frame, we project 3D scanpaths into this consistent world coordinate system. Then, with cross-modal attention between visual context and historical 3D scanpath features, we decode a forecasted 3D scanpath.

gated visual span region from egocentric point cloud observations. In contrast, we predict a dense, temporally ordered sequence of future gaze locations from egocentric video.

Egocentric Behavior Forecasting Prior works have tackled other behavior forecasting tasks in an egocentric setting, with some incorporating gaze as an additional predictive variable. GazeMotion [26] and VCR [31] exploit the tight coordination between human eye and body movements in predicting future body motions, jointly forecasting gaze direction with body pose. In contrast, we predict the future 3D gaze point, not gaze direction, and do not require body pose measurements to make effective gaze predictions. GIMO [72] contributes a dataset for human motion prediction conditioned on a gaze, where participants look at gaze targets before walking to them. While the gaze behavior in this dataset is not sufficiently diverse to support our gaze scanpath forecasting task, their experiments highlight the power of gaze predictions to drive behavior. Finally, FICTION [7] predicts future human-object interactions from an egocentric perspective in the 3D world. In particular, they use a coarsely voxelized 3D output space, which is too coarse for accurate prediction of gaze points. We take inspiration from these prior works by predicting behavior in a 3D frame of reference, but focus on the novel task of forecasting 3D gaze scanpaths.

3. Approach

3.1. Problem Formulation

We formulate egocentric 3D scanpath prediction as a forecasting task conditioned on historical gaze observations, head positions, and video frames. Previous gaze observa-

tions are given as a partial 3D scanpath composed of N_{obs} fixations, $\mathcal{S}_o = \{g_1, g_2, \dots, g_{N_o}\}$. Each gaze fixation $g \in \mathbb{R}^4$ consists of (x_i^W, m_i) , where $x_i^W \in \mathbb{R}^3$ is the 3D fixation location in Cartesian world coordinates and $m_i \in \mathbb{R}$ is the duration of the fixation in seconds. The input also includes the egocentric video stream \mathcal{V} corresponding to the observed segment \mathcal{S}_o and the corresponding set of egocentric camera poses \mathcal{P} where each pose p consists of rotation R^W and 3D translation t^W in the world coordinate system. The output is the future scanpath consisting of the next N_f fixations, $\mathcal{S}_f = \{g_{N_o+1}, g_{N_o+2}, \dots, g_{N_o+N_f}\}$.

We predict a fixed number of future fixations because time is a confounding factor for scanpaths. Our formulation focuses on the sequence of future gaze locations, treating the length of each fixation as a secondary, more ambiguous variable. We choose to use prior gaze observations as input to our task due to the continuous nature of egocentric gaze behavior. While image scanpaths are finite, comprising the full time that the viewer looks at a new image (typically 3 seconds), egocentric gaze behavior is continuous, without a defined starting or stopping point. By providing partial gaze history, we formulate our task similarly to other continuous forecasting tasks that condition future predictions on past observations such as 3D human motion estimation [32, 56, 58, 59]. In practice, the 3D gaze observations may be obtained from wearable eye tracking with gaze depth measured by vergence estimates [25], or intersection of gaze rays with 3D scene geometry provided by SLAM [53] or digital twin reconstructions[55].

3.2. Canonical Frame Definition

For each scanpath, we define a canonical 3D coordinate system \mathcal{C} centered at the camera position $p_{N_o}^W$ of the final frame of the observed sequence \mathcal{S}_o to represent the input and output space. We project all input fixations into this coordinate system to obtain the observed 3D scanpath in this canonical frame, \mathcal{S}_o^C , and we predict the future scanpath \mathcal{S}_o^C within the same coordinate system. By defining a canonical frame of reference, we model gaze in a coordinate system that is grounded in the 3D scene and persistent across the varying head poses within the scanpath. Furthermore, this 3D input and output space naturally handles cases when gaze points extend beyond the camera frame for a particular timestep.

3.3. Architecture

Our proposed architecture, shown in Fig. 2, consists of two main branches: multi-modal visual context encoding from visual frames and head positions, and cross-modal 3D scanpath decoding in the canonical frame of reference.

Visual Context Encoding Egocentric gaze forecasting requires temporal context beyond a single visual frame to account for the dynamic nature of the scene and the viewer’s egomotion. To capture temporal visual context, we encode visual features from multiple frames contextualized by the wearer’s head position. Given the egocentric video stream \mathcal{V} , we sample the sequence of frames $\{v_1, v_2, \dots, v_{N_o}\}$ such that v_i is the final video frame that temporally corresponds to fixation g_i in the observation scanpath \mathcal{S}_o . Following work on exocentric gaze estimation [62], we encode these frames using a frozen pretrained visual encoder, ψ , which is DINOv2-B [54] in our experiments. For the canonical video frame v_{N_o} , we leverage the full patch-wise feature map as the visual representation in order to capture dense semantic and spatial visual information for the reference frame in which the future scanpath will be predicted. For the preceding frames $v_{1:N_o-1}$, we use the global feature vector as the representation to capture broader visual motion in relation to the canonical frame. We learn two linear layers, E_{dense} and $E_{\text{global}} \in \mathbb{R}^{d_\psi \times d}$ to embed the dense features for the canonical frame and the global features for the past frames into d -dimensional features vectors, where d is the internal dimension of our transformer model.

Because these video frames represent different viewpoints due to egomotion, we leverage head pose (the position of the headworn RGB camera) to further contextualize the visual features. We project the camera position p_i^W corresponding to each input frame v_i into the canonical frame of reference, obtaining its relative pose $p_i^C \in \mathbb{R}^7$, which is composed the head orientation quaternion and 3D position. The poses are embedded via a linear layer $E_{\text{pose}} \in \mathbb{R}^{7 \times d}$, and concatenate them with the visual features to form the full visual context, C_{visual} . We add a fixed sinusoidal position embedding to each feature vector in C_{visual} to encode

relative time, such that head pose features and visual features for the same timestep have the same position embedding added to them. The concatenated visual and pose features are passed through 2 transformer layers with self attention to allow for interaction across the frames, producing the updated visual context C'_{visual} .

3D Scanpath Decoding To predict the scanpath, each observed fixation in \mathcal{S}_o is first embedded via a linear layer to obtain a list of observed trajectory features C_{traj} . We concatenate these features with a set of learnable trajectory query vectors, $Q = \{q_1, \dots, q_{N_f}\}$ where each $q_i \in \mathbb{R}^d$ will be transformed into a predicted future gaze point. We pass $[C_{\text{traj}}, Q]$ to a 2 layer transformer decoder to produce C'_{traj} and Q' . In each layer, the trajectory features interact across time via self attention, and cross-attend to the visual context C'_{visual} . To temporally align the observed trajectory features with the visual context, we add a fixed sinusoidal temporal position embedding to C_{traj} , such that the trajectory feature for each observed fixation receives the same position embedding as the corresponding pose and visual features. The future fixation predictions are decoded via a linear projection layer as $\mathcal{S}_f^{\text{pred}} = \text{Proj}(Q')$, where each $g_i^{\text{pred}} \in R^4$ in $\mathcal{S}_f^{\text{pred}}$ represents the predicted fixation at temporal position $N_o + i$ in the 3D canonical frame coordinate system with position $x_i^{C^{\text{pred}}}$ and duration m_i^{pred} .

Training We train our model using multitask loss to jointly supervise the prediction of fixation locations and durations. We use MSE loss for both, calculating the full loss as $\mathcal{L}(\mathcal{S}_f^{\text{pred}}, \mathcal{S}_f) = \lambda_1 \mathcal{L}_{\text{pos}}(X_f^{\text{pred}}, X_f) + \lambda_2 \mathcal{L}_{\text{dur}}(M_i^{\text{pred}}, M_i)$ where λ_1, λ_2 are weighting hyperparameters.

4. Experiments

4.1. Evaluation Protocol

Dataset We leverage the Aria Digital Twin (ADT) dataset [55] for training and evaluation, which is composed of egocentric data captured by Project Aria glasses [21], including 30Hz video, eye tracking measurements at 30Hz, and estimated camera pose via a SLAM system. Data is captured in a known environment with a 3D digital twin, facilitating obtaining ground truth 3D gaze points as the first intersection of the estimated gaze direction from the eye tracking direction with the 3D environment. We use 184 individual video sequences from Aria Digital Twin, and construct a split of 147 train, 18 val, and 19 test sequences. We process the 3D gaze points into fixation clusters by grouping together consecutive points with minimal distance between them. With an observed sequence length of $N_o = 10$ and predicted sequence length of $N_f = 10$, the test set consists of 646 non-overlapping segments, averaging 2.8s in length. Further details are provided in Supplement Sec. 8.

Table 1. Evaluation of our method and baselines on our new task of forecasting 3D scanpaths in egocentric video on the ADT dataset. We compare against heuristic baselines, an image scanpath prediction method modified to the 3D setting, and ablations of our architecture.

Method	DTW↓	EUC ↓	FRE↓	EYE ↓	TDE↓	Multimatch				
						Sh↓	Dir↓	Len↓	Pos↓	Dur↓
Dataset average	2.014	2.014	2.948	3.199	1.445	0.520	1.247	0.325	1.975	0.713
Center prior [37] at average depth	2.035	2.035	2.962	3.249	1.459	0.520	1.247	0.325	2.004	-
Average of observations	1.816	1.816	2.781	2.773	1.310	0.520	1.247	0.325	1.779	-
Last observed point	1.533	1.533	2.646	1.977	1.310	0.520	1.247	0.325	1.504	-
Linear extrapolation	4.642	4.660	8.133	5.506	1.504	0.984	1.568	0.603	4.309	-
TPP-Gaze [20] + GT depth	3.278	3.309	4.668	4.451	1.917	1.244	1.381	1.008	3.174	0.776
TPP-Gaze [20] (3D modified)	1.972	2.102	3.245	2.158	0.947	1.347	1.191	0.938	1.840	0.598
Ours - Trajectory only	1.450	1.456	2.280	1.901	0.930	0.507	1.254	0.406	1.419	0.654
Ours - Single image, no pose	1.410	1.421	2.212	1.836	0.860	0.509	1.185	0.377	1.367	0.671
Ours - Single image, pose	1.402	1.421	2.163	1.823	0.852	0.506	1.191	0.364	1.368	0.698
Ours - Video, no pose	1.395	1.412	2.215	1.814	0.853	0.503	1.188	0.373	1.344	0.740
Ours - Video, pose	1.377	1.382	2.173	1.800	0.859	0.507	1.194	0.384	1.350	0.630

Table 2. Euclidean error breakdown within the X-Y plane vs. depth for our full model compared to ablated versions that remove video frames and prior head poses.

Video	Pose	X-Y plane	Z
×	×	0.916 (+0.018)	0.927 (+0.037)
✓	×	0.907 (+0.009)	0.918 (+0.028)
✓	✓	0.898	0.890

Implementation Details We use $N_o = 10$ observed fixations and predict the next $N_f = 10$ fixations. Our input images are size 224×224 , producing a $16 \times 16 \times 768$ feature map from DINOv2 for the canonical frame. We use $d = 256$ as the latent dimension for our model. We train our model with the AdamW [48] optimizer using a fixed learning rate of $2e-4$ and weight decay of $1e-2$ for 3 epochs. A single epoch samples all possible starting points within the the training set, consisting of 87k overlapping trajectories.

Metrics We leverage commonly used distance-based measures [22] as well as the MultiMatch scanpath metrics [19] to measure similarity between predicted and ground truth scanpaths, adapting distance calculations to 3D for our setting. For distance-based measures, we consider metrics that align points in the predicted and ground truth scanpaths in different ways to measure distance: *Dynamic Time Warp* (DTW), *Euclidean Distance* (EUC), *Frechet Distance* (FRE) *Eyeanalysis* (EYE) [50], and *Time Delay Embedding* (TDE) [67]. DTW, EUC, EYE, and TDE are averaged over the sequence length, while FRE is reported for the full sequence. Distance metrics are reported in meters.

We also use the MultiMatch metric suite, which represents each scanpath as a set of gaze shifts between consecutive gaze points. Each gaze shift is represented as a 5-dimensional vector consisting of the characteristics of shape, length, direction, position, and duration. These vectors are used to calculate an optimal alignment between scanpaths, and distance measures are reported separately

across *Shape* (Sh), *Direction* (Dir), *Length* (Len), *Position* (Pos), and *Duration* (Dur). We choose to use distance measures over string edit-distance based scanpath metrics such as ScanMatch [17] and Levenshtein Distance due to the challenges in discretizing our output space to apply these measures. While a grid discretization is practical for a 2D image, our problem setting requires comparing scanpaths that cover a large 3D space, limiting the ability to meaningfully discretize the space into an alphabet for comparison.

4.2. Main Comparison

Baselines We adopt a recent state of the art image scanpath prediction model, TPP-Gaze [20], and a set of heuristic methods as our baselines. TPP-Gaze learns an image feature extractor and a temporal neural point process to predict a 2D scanpath from an image. We apply TPP-Gaze zero-shot by providing gaze observations in 2D pixel coordinates as prior context to the model and predicting the next N_f points and lifting these to 3D with ground truth depth. We also train their architecture on our dataset, both in 2D and 3D. For heuristic baselines, we consider the average fixation point across the training dataset, the center point of the canonical frame at average depth, average fixation point of the observation segment, and the last observed point. To calculate Multimatch metrics for static point baselines, we center a Gaussian distribution on the point with small variance (0.01) and sample points to simulate a scanpath with movement (we use a fixed seed, resulting in same Multimatch scores with exception of position). We also consider a linearly extrapolated scanpath based on the last 2 observed points. We additionally compare to ablations of our method: Trajectory only (no visual context or prior head poses), single image without head pose, single image with head pose, video without head pose, and our full method (multiple video frames and head poses).

Table 3. Comparison in 2D pixel coordinates on the ADT dataset. We project 3D scanpaths into 2D pixel coordinates in the canonical (last observed) frame for comparison. We compare the pretrained TPP-Gaze applied zero-shot, TPP-Gaze and our model trained on our dataset projected into 2D, and predictions from 3D models projected into 2D. Distance metrics are reported in normalized image length.

Training	Method	DTW↓	EUC↓	FRE↓	EYE↓	TDE↓	Multimatch				
							Sh↓	Dir↓	Len↓	Pos↓	Dur↓
2D (zero-shot)	TPP-Gaze	0.151	0.165	0.266	0.160	0.071	0.163	1.315	0.138	0.139	0.606
2D (finetuned)	TPP-Gaze	0.121	0.125	0.209	0.133	0.058	0.068	1.143	0.053	0.112	0.583
	Ours	0.115	0.119	0.176	0.165	0.079	0.024	1.051	0.017	0.110	0.662
3D (finetuned)	TPP-Gaze	0.128	0.136	0.214	0.142	0.061	0.086	1.210	0.065	0.120	0.599
	Ours	0.088	0.091	0.151	0.105	0.051	0.023	1.029	0.016	0.086	0.624

Table 4. Gaze trajectory context length ablations.

Num. context	DTW↓	EUC↓	FRE↓	EYE↓	TDE↓
0	1.720	1.666	2.524	2.561	1.166
1	1.450	1.461	2.248	1.923	0.880
3	1.425	1.430	2.217	1.877	0.882
5	1.417	1.429	2.217	1.871	0.881
10	1.377	1.382	2.173	1.800	0.859

Table 5. Attention structure in trajectory decoder ablations.

Context Length	DTW↓	EUC↓	FRE↓	EYE↓	TDE↓
Causal	1.419	1.427	2.234	1.857	0.895
Partial causal	1.403	1.421	2.225	1.801	0.854
Bidirectional	1.377	1.382	2.173	1.800	0.859

Comparison on 3D Scanpath Prediction To demonstrate the challenges of our novel problem formulation, we compare our proposed baseline architecture against a variety of baselines in Tab. 1. Heuristic baseline comparisons are particularly revealing because, while simple static point baselines fail in image scanpath prediction where extensive motion around the full image is typical, the last observed point baseline provides a strong baseline in our setting due to freedom of head motion. This aligns with the egocentric 2D gaze estimation literature, where the center prior is an important and strong baseline due to gaze often being centered in the field of view [40]. The last observed point serves as a similar baseline in our case, since it is the gaze point corresponding to the canonical frame of reference.

Because there are no prior methods designed specifically for our task, we adapt the TPP-Gaze [20] architecture from 2D image scanpath prediction to 3D scanpath prediction and train it on our dataset for comparison. We adapt TPP-Gaze to predict 3D scanpaths by changing the output space to Cartesian world coordinates and altering the position Gaussian Mixture Model to include a z dimension. As with our model, we provide N_o fixation observations as input in addition to the canonical video frame. Our model outperforms this TPP-Gaze baseline, highlighting the need for different modeling techniques to model 3D scanpaths in egocentric videos compared to in static images.

Ablations of our method show that including visual information, temporal context via past video frames, and observed head pose improve performance. However, the relatively small overall improvements suggest that fully leveraging dynamic visual and pose information may require further exploration and larger-scale data to learn from. While head pose does not make a significant difference in all metrics, we show in Tab. 2 that including it most significantly reduces Euclidean error along the depth direction, illustrating the importance of accounting for head motion for producing geometrically grounded 3D predictions.

Comparison in 2D To enable more natural comparisons with standard 2D formulations of scanpath prediction, we project ground truth 3D scanpaths into pixel coordinates within the canonical frame (Tab. 3). This allows us to directly compare performance against the pretrained TPP-Gaze model in the zero-shot setting. We also compare to training the TPP-Gaze model and our model on our dataset projected into 2D. We exclude examples with scanpaths that move outside the field-of-view of the canonical frame during training and evaluation in 2D. Thus, the evaluations in Tab. 3 are compared on a restricted subset (523 sequences) of the full test dataset from Tab. 1. Finally, we compare to the predictions of our 3D model and 3D-adapted TPP-Gaze baseline projected into the 2D canonical frame. We observe that TPP-Gaze does not perform well zero-shot, reflecting the significant differences between the image scanpath domain and egocentric domain. Training TPP-Gaze on our data in 2D improves performance, but adapting their architecture to predict in 3D reduces 2D performance. Our architecture outperforms TPP-Gaze when trained on 2D, and training our architecture on 3D further improves predictions when projected into 2D. This result shows that by designing our architecture explicitly for the 3D setting, our model is able to leverage 3D information to produce predictions that are also more accurate in 2D.

4.3. Analysis

Qualitative Results We present a qualitative comparison in Fig. 3, showing scanpath predictions projected into 2D for visualization. We compare our full method with TPP-Gaze, adapted for 3D and trained on our dataset, and the

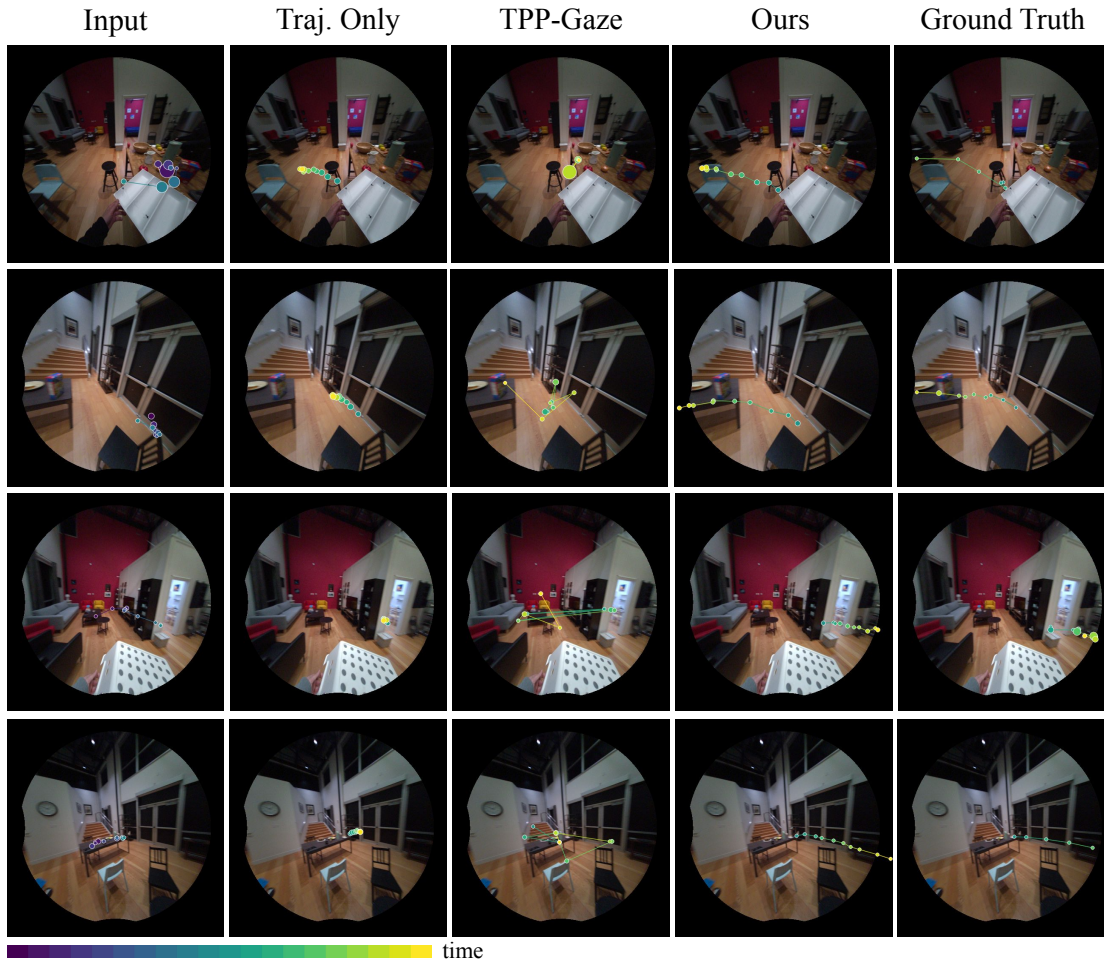


Figure 3. Qualitative comparison. We show the input observed scanpath, and the predicted future scanpaths from our trajectory-only model, TPP-Gaze adapted to 3D, and our model compared to the ground truth. For visualization purposes, we project the 3D scanpaths into the canonical video frame. Fixation duration sizes are indicated by radius. (*Best viewed zoomed in*).

trajectory-only variant of our method to illustrate the role of visual information in our architecture. Compared to the trajectory-only variant of our model, which lacks any visual context, our model produces predictions that are more aligned with the visual content of the scene, showing movement between salient objects. In comparison to TPP-Gaze, our model produces scanpaths that better align with the shape of gaze trajectories present in egocentric video. The scanpaths predicted by TPP-Gaze exhibit more movement around the image in different directions, which is more likely to occur in the image viewing setting for which TPP-Gaze’s neural point process architecture was designed. In contrast, our gaze model’s gaze trajectories better reflect the smooth and continuous nature of gaze in natural scenes and represent natural continuations of the input scanpaths.

We qualitatively explore the role of temporal context in our model in Fig. 4, which compares our full model leveraging video and prior head poses, to our single image model that uses only the canonical video frame and no prior head

poses. In cases with dynamic hand and object motion (row 1) and significant head movement (row 2), providing past video and head pose to our model produces predictions that are better aligned with the temporal dynamics of the scene.

We also show representative failure modes of our model in Fig. 5 to highlight the challenges of our new task and limitations of our model. A common failure mode is predicting that gaze remains in a small region (row 1), or moves minimal distances compared to the ground truth (row 2). These errors reflect the inherent biases present in egocentric gaze data: compared image free-viewing, continuous and task-driven gaze behavior has less movement between consecutive gaze points, resulting in a strong bias towards minimal motion. Row 3 highlights another error mode, in which the model predicts motion, but in a different direction than the ground truth. Like the ground truth, the predicted scanpath suggests that model is forecasting that the viewer will pick up and move the bowl with their outstretched hand. While this trajectory may be plausible, it obtains large error in

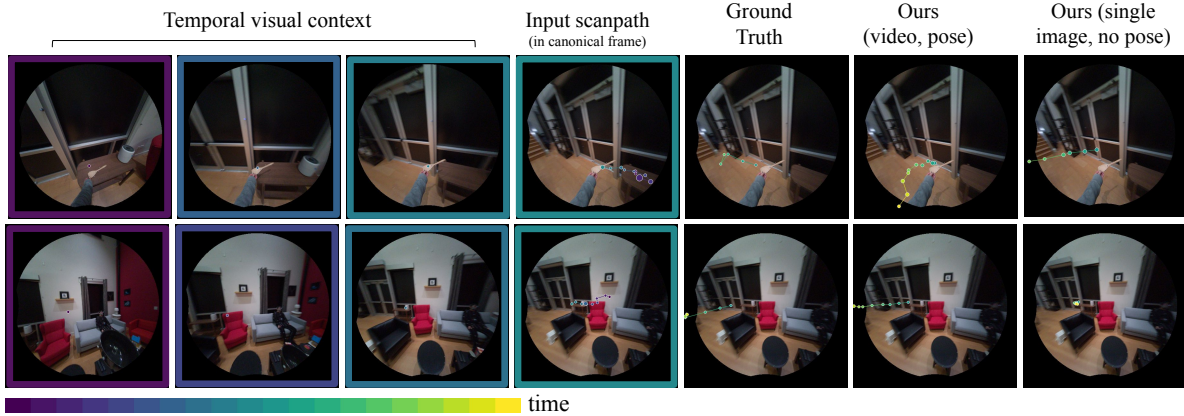


Figure 4. Role of temporal visual context. We show qualitative examples with hand/object motion (row 1) and significant head motion (row 2), where including prior visual context via multiple video frames and observed head poses improves forecasted scanpaths.

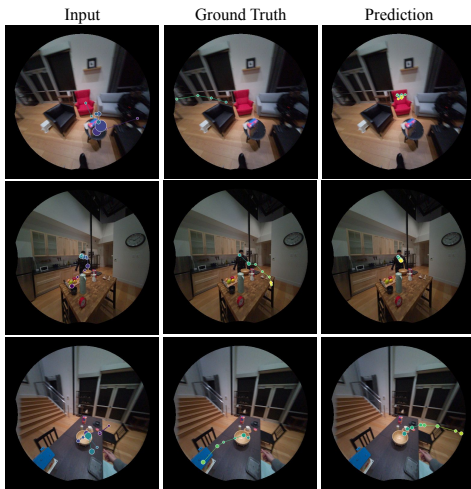


Figure 5. Representative failure modes. Our model experiences failures by predicting limited movement between gaze points (rows 1-2) and predicting motion in a different direction than the ground truth (row 3).

evaluation due to deviation from the ground truth in position and direction. This case reveals a larger challenge in evaluating egocentric gaze forecasting compared to image scanpath prediction: while image scanpath datasets present the same stimulus to multiple viewers to obtain multiple plausible scanpaths to evaluate against, by nature, our ground truth is unimodal. This finding suggests that establishing effective evaluation criteria for forecasting gaze in embodied settings is an important direction for future research.

Context Length We investigate the role of prior scanpath context in Tab. 4. Without any prior fixations provided as context, there is a significant increase in error, reflecting the importance of context in conditioning future gaze prediction. Even providing a single past fixation as prior context largely reduces this error, as it constrains the likely locations for the start of the forecasted scanpath. Providing additional

context results in further improvements, giving the model additional information about the shape, direction, and velocity of the partial scanpath up until the forecasting window to inform the continuation of the scanpath.

Attention Structure We also explore the attention structure within the trajectory decoder in our architecture (Sec. 3.3) in Tab. 5. While our model uses bidirectional attention across the embeddings for the observed scanpath and query embeddings for the future embeddings, several existing models for image scanpath prediction are causal, autoregressively predicting each point in the scanpath based only on prior predictions [16, 20, 39]. We compare our bidirectional decoder to a fully causal decoder and a partially causal decoder, which has bidirectional attention across the embeddings for the observed segment but causal attention across the embeddings for forecasted segment. Bidirectional attention performs best, suggesting benefit both in holistically modeling the observed segment with both forward and backward attention across timesteps and jointly predicting the full future scanpath.

5. Conclusion

Through this first formulation of egocentric 3D scanpath prediction, we demonstrate the potential of modeling gaze behavior in dynamic scenes enabled by wearable technologies like AR/VR. Contrasting with traditional approaches to single-image 2D scanpath prediction, we elucidate the challenges inherent to this setting wherein perception entails active interaction and movement within the environment. As a first step towards this task, we propose a demonstrative architecture that predicts 3D trajectories in a canonical Cartesian frame of reference conditioned on visual and behavioral temporal context provided by video and head poses, providing a strong initial baseline for further investigation. We hope our work enables future research towards modeling visual attention in embodied settings.

References

- [1] http://github.com/adswa/multimatch_gaze. 2
- [2] <http://github.com/rAmln/saliency>. 2
- [3] Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye tracking in virtual reality: a broad review of applications and challenges. *Virtual Reality*, 27(2):1481–1505, 2023. 1
- [4] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017. 1
- [5] Henny Admoni and Siddhartha S Srinivasa. Predicting user intent through eye gaze for shared autonomy. In *AAAI fall symposia*, pages 298–303, 2016. 1
- [6] Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47(4):1377–1392, 2015. 2
- [7] Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. Fiction: 4d future interaction prediction from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17613–17625, 2025. 3
- [8] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O’Connor. PathGAN: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1, 2
- [9] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012. 2
- [10] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 2
- [11] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need. In *MMSP*, pages 1–6, 2021. 2
- [12] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10876–10885, 2021. 2
- [13] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024. 2
- [14] Xianyu Chen, Ming Jiang, and Qi Zhao. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *European Conference on Computer Vision*, pages 314–333. Springer, 2024. 2
- [15] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021. 1, 2
- [16] Zhenzhong Chen, Wanjie Sun, et al. Scanpath prediction for visual attention using ior-roI lstm. In *IJCAI*, page 5, 2018. 1, 2, 8
- [17] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3):692–700, 2010. 5
- [18] Ryan Anthony Jalova De Belen, Tomasz Bednarz, and Arcot Sowmya. Scanpathnet: A recurrent mixture density network for scanpath prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5010–5020, 2022. 2
- [19] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012. 5, 2
- [20] Alessandro D’Amelio, Giuseppe Cartella, Vittorio Cuculo, Manuele Lucchi, Marcella Cornia, Rita Cucchiara, and Giuseppe Boccignone. TPP-Gaze: Modelling gaze dynamics in space and time with neural temporal point processes. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8786–8795. IEEE, 2025. 2, 5, 6, 8, 1
- [21] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talatof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project Aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 4, 2
- [22] Ramin Fahimi and Neil DB Bruce. On metrics for measuring scanpath similarity. *Behavior Research Methods*, 53(2):609–628, 2021. 5, 2
- [23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2
- [24] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005. 1
- [25] David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8(3):33–33, 2008. 3
- [26] Zhiming Hu, Syn Schmitt, Daniel Häufle, and Andreas Bulling. Gazemotion: Gaze-guided human motion forecasting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13017–13022. IEEE, 2024. 3
- [27] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 2
- [28] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 2

- [29] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 2
- [30] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 2002. 1, 2
- [31] Wenqi Jia, Bolin Lai, Miao Liu, Danfei Xu, and James M Rehg. Learning predictive visuomotor coordination. *arXiv preprint arXiv:2503.23300*, 2025. 3
- [32] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 3
- [33] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. 2
- [34] Yue Jiang, Zixin Guo, Hamed Rezazadegan Tavakoli, Luis A Leiva, and Antti Oulasvirta. EyeFormer: predicting personalized scanpaths with transformer-guided reinforcement learning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15, 2024. 2
- [35] Chuhan Jiao, Yao Wang, Guanhua Zhang, Mihai Băce, Zhiming Hu, and Andreas Bulling. Diffgaze: A diffusion model for continuous gaze sequence generation on 360 {deg} images. *arXiv preprint arXiv:2403.17477*, 2024. 2
- [36] Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. Eye–hand coordination in object manipulation. *Journal of neuroscience*, 21(17):6917–6932, 2001. 1
- [37] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 1, 2, 5
- [38] Ozgur Kara, Harris Nisar, and James M Rehg. Diffeeye: Diffusion-based continuous eye-tracking data generation conditioned on natural images. *arXiv preprint arXiv:2509.16767*, 2025. 2
- [39] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022. 1, 2, 8
- [40] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *arXiv preprint arXiv:2208.04464*, 2022. 2, 6
- [41] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132(3):854–871, 2024. 2
- [42] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pages 192–210. Springer, 2024. 2
- [43] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001. 1
- [44] Tai Sing Lee and Stella Yu. An information-theoretic framework for understanding saccadic eye movements. *Advances in neural information processing systems*, 12, 1999. 1, 2
- [45] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 2
- [46] Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 6731–6747, 2021. 2
- [47] Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, and Stephen Lin. Semantically-based human scanpath estimation with hmms. In *Proceedings of the IEEE international conference on computer vision*, pages 3232–3239, 2013. 1, 2
- [48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [49] Yusuf Mansour, Ajoy Savio Fernandes, Kiran Somasundaram, Tarek Hefny, Mahsa Shakeri, Oleg Komogortsev, Abhishek Sharma, and Michael J Proulx. Enabling eye tracking for crowd-sourced data collection with project aria. *IEEE Access*, 2025. 2
- [50] Sebastiaan Mathôt, Filipe Cristino, Iain D Gilchrist, and Jan Theeuwes. A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research*, 5(1):1–15, 2012. 5
- [51] Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8310–8320, 2025. 2
- [52] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1441–1450, 2023. 2
- [53] Raul Mur-Artal, JMM Montiel, and JD Tardos. ORB-SLAM: a versatile and accurate monocular slam system. In *IEEE Transactions on Robotics*, pages 1147–1163. IEEE, 2015. 3
- [54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4
- [55] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 2, 3, 4

- [56] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 3
- [57] Trong Thang Pham, Akash Awasthi, Saba Khan, Esteban Duran Marti, Tien-Phat Nguyen, Khoa Vo, Minh Tran, Son Nguyen, Cuong Tran, Yuki Ikebe, et al. Ct-scangaze: A dataset and baselines for 3d volumetric scanpath modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21732–21743, 2025. 2
- [58] Yijun Qian, Jack Urbanek, Alexander G Hauptmann, and Jungdam Won. Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2306–2316, 2023. 3
- [59] Yijun Qian, Jack Urbanek, Alexander Hauptmann, and Jungdam Won. Text motion translator: A bi-directional model for enhanced 3d human motion generation from open-vocabulary descriptions. In *European Conference on Computer Vision*, pages 398–414. Springer, 2024. 3
- [60] Laura Renninger, James Coughlan, Preeti Verghese, and Jitendra Malik. An information maximization model of eye movements. *Advances in neural information processing systems*, 17, 2004. 1, 2
- [61] Laura Walker Renninger, Preeti Verghese, and James Coughlan. Where to look next? eye movements reduce local uncertainty. *Journal of vision*, 7(3):6–6, 2007. 1, 2
- [62] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-llc: Gaze target estimation via large-scale learned encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28874–28884, 2025. 4
- [63] Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Understanding teacher gaze patterns for robot learning. In *Conference on Robot Learning*, pages 1247–1258. PMLR, 2020. 1
- [64] Xuan Shao, Ye Luo, Dandan Zhu, Shuqin Li, Laurent Itti, and Jianwei Lu. Scanpath prediction based on high-level features and memory bias. In *International conference on neural information processing*, pages 3–13. Springer, 2017. 2
- [65] Inam Ullah, Muwei Jian, Sumaira Hussain, Jie Guo, Hui Yu, Xing Wang, and Yilong Yin. A brief survey of visual saliency detection. *Multimedia Tools and Applications*, 79(45):34605–34645, 2020. 2
- [66] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural networks*, 19(9):1395–1407, 2006. 2
- [67] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *CVPR 2011*, pages 441–448. IEEE, 2011. 1, 2, 5
- [68] Yujia Wang, Fang-Lue Zhang, and Neil A Dodgson. Scantd: 360° scanpath prediction based on time-series diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7764–7773, 2024. 2
- [69] Chen Xia, Junwei Han, Fei Qi, and Guangming Shi. Predicting human saccadic scanpaths based on iterative representation learning. *IEEE Transactions on Image Processing*, 28(7):3502–3515, 2019. 2
- [70] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 2
- [71] Ruoyu Xue, Jingyi Xu, Sounak Mondal, Hieu Le, Greg Zelinsky, Minh Hoai, and Dimitris Samaras. Few-shot personalized scanpath prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13497–13507, 2025. 2
- [72] Yanchao Yang. Gimo: Gaze-informed human motion prediction in context. In *17th European Conference on Computer Vision, ECCV 2022 (23/10/2022-27/10/2022, Tel Aviv, Israel)*, 2022. 3
- [73] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2020. 1, 2
- [74] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693, 2024. 2
- [75] Heeseung Yun, Joonil Na, Jaeyeon Kim, Calvin Murdock, and Gunhee Kim. Gaze beyond the frame: Forecasting egocentric 3d visual span. *arXiv preprint arXiv:2511.18470*, 2025. 2
- [76] Dario Zanca, Andrea Zugarini, Simon Dietz, Thomas R Altstidl, Mark A Turban Ndjeuha, Moumita Chakraborty, Naga Venkata Sai Jitin Jami, Leo Schwinn, and Bjoern M Eskofier. Contrastive language-image pretrained models are zero-shot human scanpath predictors. *IEEE Transactions on Artificial Intelligence*, 2025. 2
- [77] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2
- [78] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4372–4381, 2017. 2
- [79] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Anticipating where people will look using adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1783–1796, 2018. 2
- [80] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJ-CAI: Proceedings of the Conference*, page 4951, 2020. 1

- [81] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl Muller, Jake Whritner, Luxin Zhang, Mary Hayhoe, and Dana Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6811–6820, 2020. 1