

Infinite Gaze Generation for Videos with Autoregressive Diffusion

Jenna Kang¹, Colin Groth¹, Tong Wu², Finley Torrens¹, Patsorn Sangkloy¹,
Gordon Wetzstein², and Qi Sun¹

¹ New York University, New York, NY, USA
jennakang@nyu.edu, c.groth@nyu.edu, ft2377@nyu.edu,
patsorn.sangkloy@gmail.com, qisun@nyu.edu

² Stanford University, Stanford, CA, USA
wutong16.thu@gmail.com, gordonwz@stanford.edu

Abstract. Predicting human gaze in video is fundamental to advancing scene understanding and multimodal interaction. While traditional saliency maps provide spatial probability distributions and scanpaths offer ordered fixations, both abstractions often collapse the fine-grained temporal dynamics of raw gaze. Furthermore, existing models are typically constrained to short-term windows ($\approx 3\text{--}5\text{s}$), failing to capture the long-range behavioral dependencies inherent in real-world content. We present a generative framework for infinite-horizon raw gaze prediction in videos of arbitrary length. By leveraging an autoregressive diffusion model, we synthesize gaze trajectories characterized by continuous spatial coordinates and high-resolution timestamps. Our model is conditioned on a saliency-aware visual latent space. Quantitative and qualitative evaluations demonstrate that our approach significantly outperforms existing approaches in long-range spatio-temporal accuracy and trajectory realism. Project website: <https://www.immersivecomputinglab.org/publication/infinite-gaze/>.

Keywords: gaze prediction, videos, saliency, scanpath, diffusion models, autoregression

1 Introduction

Modeling human visual attention through spatial saliency and temporal scanpaths is vital for scene understanding, enabling applications from foveated rendering [39, 60] to intention-aware robotics [29, 62]. However, the systematic collection of high-fidelity gaze data is hindered by expensive hardware requirements, cumbersome calibration, and privacy constraints. This data scarcity necessitates generative methods capable of synthesizing plausible, realistic gaze behavior. By leveraging a latent world model to understand evolving environmental dynamics, such approaches can move beyond static maps to produce continuous, infinite-horizon trajectories that accurately reflect human visual engagement with the world.

Existing gaze prediction models generally fall into two categories: static saliency maps, which provide time-independent spatial probability distributions [21, 51], and scanpaths, which discretize eye movements into sequences of fixations [25, 31]. While these abstractions are effective for images, they often collapse the high-frequency temporal dynamics of raw eye-tracking data and struggle to represent smooth pursuit in video. Furthermore, current models are typically constrained to short-term viewing windows ($\approx 3\text{--}5\text{s}$), failing to capture the long-range behavioral dependencies found in real-world, long-form content. These limitations prevent existing methods from synthesizing the continuous, infinite-horizon gaze trajectories required for deep scene understanding and realistic dynamic interaction.

Inspired by recent advances in video world models that generate content in a causal manner [12, 41, 53], we process video sequences frame-by-frame to synthesize gaze data autoregressively. Our approach predicts gaze for the current frames by conditioning on visual context, as well as the history of previously generated gaze coordinates. Notably, recognizing that visual saliency is a stronger prior than general visual features for inducing eye movements, we introduce the idea of applying saliency-aware latents to encode visual frame information. This causal formulation ensures temporal continuity and allows the model to act as a latent world model, capable of generating realistic, infinite-horizon gaze trajectories that remain grounded in the evolving scene dynamics.

Through extensive quantitative evaluations, qualitative analysis, and human subject studies, we demonstrate that our approach significantly advances the state-of-the-art in generating accurate and human-like spatio-temporal gaze trajectories for videos of arbitrary length. Our results confirm that the model effectively maintains long-range consistency while capturing the nuanced, high-frequency dynamics characteristic of real human vision. We envision this research establishing a new frontier in synthesizing natural human behaviors by leveraging the power of emerging autoregressive world models. To this aim, we will release our code and models upon acceptance.

2 Related Work

2.1 Scanpath Generation for Images

Visual attention acts as the selective mechanism that determines which parts of a scene humans prioritize [26]. Early computational models focused on predicting spatial saliency, estimating which regions of an image are most likely to attract human gaze [5–7, 23]. However, saliency maps ignore the temporal structure of eye movements and do not capture the order or timing of fixations, which are fundamental aspects of human visual behavior.

To address this limitation, researchers formulated the problem as scanpath prediction, modeling eye movements as sequences of discrete fixations [27, 31]. Early deep learning approaches used Recurrent Neural Networks (RNNs) to capture dependencies between successive fixations [2, 10, 49, 59]. More recently,

Transformer-based architectures have achieved state-of-the-art performance by better modeling long-range spatial and temporal dependencies in scanpath sequences [16, 37, 55, 56].

Despite these advances, representing gaze as discrete fixation sequences simplifies the rich dynamics present in raw eye-tracking data. To better capture fine-grained eye movement behavior, recent work has explored continuous gaze trajectory generation. For example, DiffEye [25] formulates gaze prediction as a diffusion process that generates fine-grained gaze trajectories conditioned on patch-level visual features, grounding the predicted gaze coordinates in the semantic structure of natural images.

2.2 Scanpath Modeling in 360-Degree Content

Modeling gaze behavior in immersive environments introduces additional challenges due to spherical geometry and the wider field of view. Several works extend scanpath prediction methods to 360-degree images, where gaze trajectories are represented as fixation sequences on a spherical panorama. For example, ScanDMM [48] is a probabilistic deep Markov model that generates realistic variable-length 3D scanpaths on panoramic images by modeling time-dependent attention dynamics and visual working memory. ScanGAN360 [32] uses a conditional generative adversarial network to synthesize human-like 3D scanpaths on 360-degree images using adversarial training with a DTW-based loss. Similarly, PathGAN [2] generates plausible fixation sequences by learning the joint distribution of fixation locations and their temporal order conditioned on an input image and its saliency map. More recently, diffusion-based approaches have also been applied to immersive content. DiffGaze [22] generates gaze trajectories on 360-degree images using a diffusion model conditioned on a global spherical feature representation obtained through spherical convolution.

In the case of 360-degree videos, many works focus on predicting viewer head motion or viewport trajectories rather than true eye movements. These methods treat gaze as spatiotemporal points on the spherical panorama corresponding to the user’s viewport or head orientation. For example, Xu et al. [54] predict future gaze directions in 360-degree video sequences, while other works predict saliency or viewing patterns to optimize adaptive video streaming [35, 36]. Transformer-based models such as VPT360 [8] forecast future viewport trajectories using past viewing history, while diffusion-based approaches such as ScanTD [52] generate plausible future scanpaths by denoising sequences of spherical fixation points. Other methods formulate panoramic video scanpath prediction as an information-theoretic problem [30] or incorporate multimodal signals such as motion [63].

However, these 360-degree video approaches all rely on viewport or head-orientation signals rather than high-frequency eye-tracking trajectories. As a result, they are not directly comparable to datasets that record fine-grained 2D eye movements in conventional RGB videos.

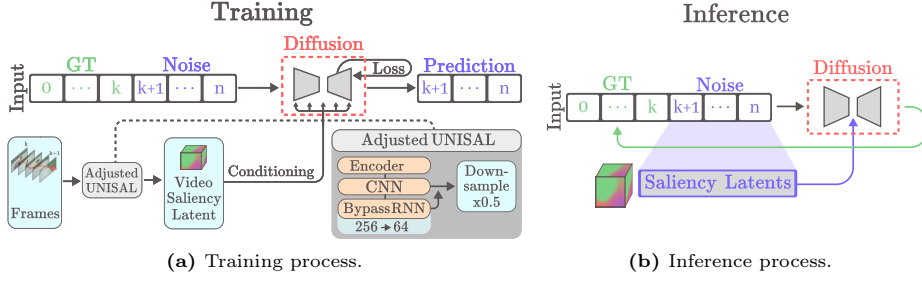


Fig. 1: Our training and inference pipelines. The input of each iteration consists of the n -dimensional input vector with gaze history (GT) and noise for future prediction, and video frames corresponding to the prediction window. During inference, the gaze history of this input vector is filled with former predictions in an autoregressive manner. Video latents are generated frame-by-frame with a modified version of UNISAL. The blue boxes indicate our modifications, while the orange boxes indicate the original architecture. The low-dimensional video latents are processed by the U-Net with cross-attention at each block.

2.3 Autoregressive Diffusion and World Models

Full-sequence diffusion models have demonstrated remarkable success in generating complex and high-dimensional data such as images [17, 40, 44, 47], videos [18, 38, 45], and continuous planning [13, 20]. However, they are typically constrained to fixed-length windows and struggle with variable-length generation. To unify the sequential flexibility of autoregressive next-token prediction with continuous diffusion, Diffusion Forcing [9] introduces a training paradigm that trains causal models to denoise future tokens while conditioning on a clean or partially noised past, stabilizing continuous long-horizon rollouts. This autoregressive diffusion paradigm has been particularly impactful in the development of generative world models, which require continuous, long-term predictions of physical environments and spatial trajectories. However, while these emerging models focus on generating long and variable-length videos [58], consistent environments [53], or autonomous agent actions [61], our work redirects this autoregressive diffusion paradigm toward modeling human eye movement reactions to the dynamic world.

3 Method

3.1 Problem Formulation

Let a video stimulus be represented as a sequence of RGB frames $V = \{I_t\}_{t=1}^T$, $I_t \in \mathbb{R}^{H \times W \times 3}$, where T denotes the number of frames, our goal is to perform gaze prediction $R \in \mathbb{R}^{T \times 2}$. The gaze prediction at each timestep is a two-dimensional coordinate x, y , representing the spatial location in the frame. Following DiffEye [25], we model gaze prediction as learning a conditional generative model $p_\theta(R |$

V) using a denoising diffusion probabilistic model (DDPM), which approximates the conditional distribution of gaze trajectories.

We further extend the gaze generator to operate in an autoregressive manner by introducing a historical gaze cache, with a subsequence of previously observed gaze coordinates. The historical sequence is concatenated before the predicted gaze trajectory to encourage temporal continuity and smooth transitions.

3.2 Model Architecture

As shown in Figure 1, we follow DiffEye [25] and adopt the same U-Net-based diffusion backbone as the noise prediction network. The model consists of hierarchical 1D convolutional blocks with multi-scale self-attention, and is conditioned on the diffusion timestep via sinusoidal embeddings. To extend the framework from static images to potentially infinite video sequences, we introduce two key modifications. First, we reformulate gaze generation in an **autoregressive manner**, allowing the model to iteratively predict future gaze trajectories conditioned on previously generated gaze coordinates. Second, due to the significantly larger information content in videos compared to images, we design an effective **stimulus conditioning** mechanism to make use of saliency latents and compress and aggregate temporal visual information before injecting it into the diffusion backbone.

Autoregressive Generation We adopt an autoregressive formulation similar to prior video diffusion models. Given a video sequence of length n , the first k gaze coordinates serve as conditional prefix, while the remaining $n - k$ coordinates are generated. During training, the windows of length k are randomly sampled from the ground-truth segments. The model conditions on the historical gaze prefix $R_{1:k}$ and predicts the future gaze sequence $R_{k+1:n}$. The diffusion loss is computed only on the predicted segment. At inference time, the model iteratively uses the most recent k generated gaze coordinates as conditioning input to predict subsequent gaze points, thereby enabling long-horizon autoregressive gaze generation. Following prior measurements of visual working-memory loss over 1–6 s [43], we chose $k = 90$ -frames (3 s) history as a midpoint balancing temporal context, noise, and compute cost.

Stimuli Conditioning To handle continuous video streams, an effective conditioning mechanism is crucial. Compared to static images, videos contain significantly larger information volume, making direct RGB conditioning inefficient and computationally expensive. We therefore emphasize information compression as a key design principle.

Instead of conditioning on raw RGB frames, we leverage video saliency maps as stimulus representation. Saliency provides a compact, single-channel description of the probability that an average viewer attends to a given spatial region. This representation is both semantically meaningful and content-independent, since saliency values directly correspond to gaze likelihood rather than raw pixel appearance. We adopt UNISAL [14], a well-established video saliency model,

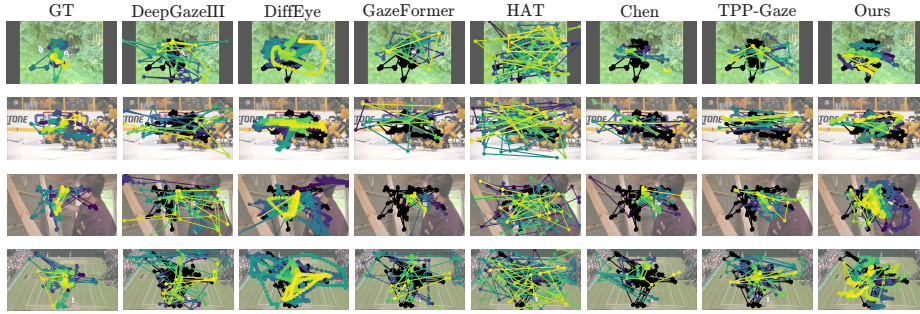


Fig. 2: Comparison of the trajectories of video gaze predictions with different methods. The image shows a representative frame selected from the video, but the path is generated for the entire duration of the video. Generated trajectories (color map indicates timestamps) are overlaid on the ground truth human gaze (black trajectories, which are also shown in the left column with the color map).

as the backbone for saliency extraction. However, existing saliency models are designed for prediction accuracy rather than compact representation. To better align with our conditioning objective, we introduce a compressed bottleneck within the encoder–decoder architecture, as shown in Figure 1. Specifically, the RNN channels are decreased from 256 to 64, which still provides enough temporal information yet saves significant size. The latent after the encoder is further compressed in its spatial dimensions through 2D average pooling. The compressed saliency latent is used as conditioning input to the diffusion backbone. During training and inference, saliency features are extracted per frame and concatenated temporally, while conditioning is applied at a reduced temporal frequency (every 5 frames) to further limit computational overhead. This design enables efficient long-horizon gaze generation without sacrificing performance.

4 Experiments

4.1 Experimental Setup

Implementation Details The model was trained for 70 epochs using the Adam optimizer with a constant learning rate of 1×10^{-4} . We employed a DDPM linear noise schedule, spanning from 1×10^{-4} to 2×10^{-2} . For sampling, we adopted Denoising Diffusion Implicit Models (DDIM) [46] to achieve high-quality generation in significantly fewer steps without compromising performance. Accordingly, the number of diffusion steps was set to 1000 during training and reduced to 50 during sampling for improved efficiency.

Datasets We evaluate the proposed method on two public video datasets that provide diverse visual content and human gaze annotations. We use *DIEM* as the in-domain benchmark [33]. The dataset contains 84 videos with a resolution of 1280×720 and durations between 30 to 209 seconds. Each video provides gaze

data from 67 observers on average, with a total viewing time of 553,116 seconds (16,593,480 frames). In comparison, the only other gaze prediction method that outputs raw gaze trajectories (DiffEye [25]) is trained on 45,135 seconds of viewing time (MIT1003 Dataset [24]). We adopt an 80/20 train-test split. We select DIEM for training due to its comparatively high spatial resolution, long temporal extent, and dense gaze annotations.

We further evaluate generalization on *DHF1K* [50]. The dataset comprises 1,000 videos with durations between 17 and 42 seconds and a spatial resolution of 640×360 . Each video includes gaze data from 17 observers. The content spans a wide range of scene motion and camera motion patterns. For evaluation, we select 16 videos, with four videos from each motion category defined by the authors: slow camera motion, fast camera motion, slow scene motion, and fast scene motion. We manually verify that the randomly selected subset covers diverse content and motion characteristics.

Evaluation Metrics We evaluate the results using four standard metrics from path assessment of video data [30]. (1) *Levenshtein Distance* measures the sequence similarity based on the number of insertion, deletion, and substitution operations required to match a GT [42]; (2) *Discrete Fréchet Distance* measures the similarity between two curves by evaluating the minimal maximum pointwise distance under a monotonic traversal of both sequences [1]; (3) *Dynamic Time Warping* (DTW) computes an optimal alignment between two sequences that may differ in speed or length by minimizing cumulative alignment cost [4]; (4) *Maximum Temporal Correlation* quantifies the highest cross-correlation between two temporal signals over possible time shifts [30].

Our evaluation protocol follows established procedures in prior work on image-based scanpath prediction [22, 52]. For each video, we generate 10 gaze trajectories per model. Given the multiple ground-truth (GT) gaze paths per video (see Dataset section for numbers), we compute the evaluation metrics by performing pairwise comparisons between each GT path and all 10 predicted trajectories. Metric scores are first computed per GT path. We report two variants. The best score selects the maximum (or minimum, depending on the metric direction) value among the 10 pairwise comparisons. The mean score corresponds to the average over all 10 comparisons. The best score reflects the peak performance of a method under multiple samples. The mean score quantifies the consistency of the predicted trajectories. Finally, for each metric, we average both the best and mean scores across all GT gaze paths and all videos in the dataset.

Comparison Methods Since our work introduces the first video-based gaze prediction model, no direct baselines exist. We adapt existing state-of-the-art image-based scanpath and gaze prediction models to the video domain for comparison. The baselines include the transformer-based architectures Human Attention Transformer (*HAT*) [56] and *GazeFormer* [34], *TPP-Gaze* which is based on neural temporal point processes [11], the convolutional network *DeepGaze III* [28], the reinforcement learning model of Chen et al. (originally designed for

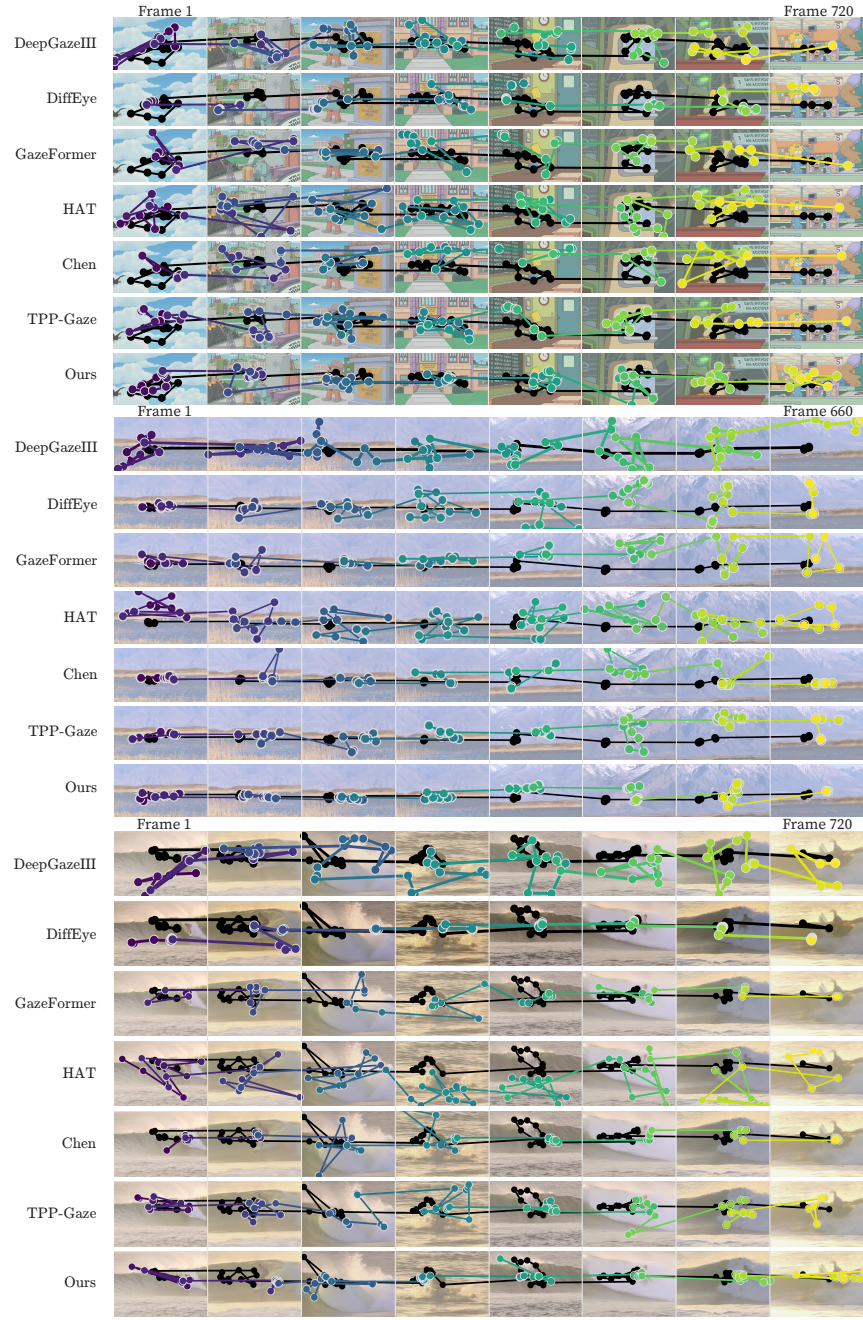


Fig. 3: Temporal progression of video gaze predictions with different methods. Generated trajectories are overlaid on the ground-truth human gaze (black). Note the closer proximity between our method and the ground truth, indicating greater similarity compared with alternative approaches.

Table 1: Scanpath evaluation on DIEM and DHF1K test sets. Lower is better for all metrics except Maximum Temporal Correlation. Metrics are scaled as indicated in the column headers. **Bold** values indicate the best scores, while underlined values denote the second-best scores.

Dataset	Method	Levenshtein ($\times 10^3$) ↓		Disc. Fréchet ($\times 10^2$) ↓		DTW ($\times 10^4$) ↓		Max. Temp. Corr. ↑	
		Mean	Best	Mean	Best	Mean	Best	Mean	Best
DIEM	DeepGaze III [28]	4.36	4.26	4.78	4.12	20.07	18.02	0.149	0.230
	DiffEye [25]	<u>3.78</u>	<u>3.71</u>	<u>3.82</u>	<u>3.29</u>	<u>12.47</u>	<u>11.66</u>	<u>0.176</u>	<u>0.254</u>
	GazeFormer [34]	4.26	4.12	4.17	3.42	15.64	13.71	0.188	0.272
	HAT [56]	4.45	4.32	5.16	4.28	23.44	20.42	0.131	0.223
	Chen et al. [10]	4.23	4.09	4.36	3.35	15.97	13.51	0.175	0.270
	TPP-Gaze [11]	4.26	4.14	3.60	3.01	15.17	13.36	0.131	0.261
	Ours	3.40	3.31	4.14	3.35	12.33	11.44	0.235	0.314
DHF1K	DeepGaze III [28]	3.61	3.48	3.12	2.67	11.12	9.41	0.136	0.317
	DiffEye [25]	<u>1.76</u>	<u>1.71</u>	2.34	1.95	<u>4.24</u>	<u>3.79</u>	<u>0.232</u>	<u>0.352</u>
	GazeFormer [34]	<u>1.76</u>	<u>1.70</u>	2.49	1.93	4.43	3.69	0.258	0.392
	HAT [56]	1.84	1.80	3.12	2.67	6.08	5.43	0.179	0.289
	Chen et al. [10]	1.77	1.69	2.54	<u>1.84</u>	4.70	3.70	<u>0.265</u>	0.432
	TPP-Gaze [11]	1.78	<u>1.70</u>	2.34	1.68	4.30	<u>3.47</u>	0.225	<u>0.435</u>
	Ours	1.57	1.50	2.50	1.95	4.01	3.40	0.290	0.444

visual question answering) [10], and the diffusion-based method *DiffEye* [25]. Among these, DiffEye is the only method that predicts dense, frame-to-frame gaze motion. To ensure fair comparison, we post-process the sample density of all scanpath prediction models to match the number of samples from our method (30 samples per second) by duplicating fixation positions proportionally. Since DiffEye produces raw gaze at a higher sampling rate than our method, we evenly downsample its gaze sequence. All comparison method predictions are generated with their default 3-second length. However, videos are typically much longer. To achieve autoregressive behavior for longer sequences, we process the video in 3-second chunks and concatenate outputs. For methods allowing us to control the gaze starting position (HAT and DeepGaze III), we use the last predicted gaze of the previous chunk as the initial position for the next chunk. For each 3-second segment, the first frame is provided as input, since the baseline models are single-image based. For multi-task models, we ensure that scanpaths are generated for the free-viewing task specifically. For additional fairness, we also evaluate all methods within their native 3-second window without autoregressive adjustment. Additionally, we retrain DiffEye using the same saliency latents and no autoregression to demonstrate that saliency conditioning alone is insufficient for long-horizon video gaze generation.

4.2 Quantitative Comparisons

Long Window Generation Quantitative results for long window video gaze prediction are shown in Table 1. Our method achieves the best performance in almost all evaluated metrics across both datasets. In particular, we obtain the strongest results in Levenshtein distance, DTW, and Maximum Temporal Correlation. The low Levenshtein distance indicates that the predicted gaze path closely matches the ordering and structure of human gaze transitions. The superior DTW performance demonstrates accurate temporal alignment, even under local variations in duration or sampling density. The high Maximum Temporal

Table 2: Additional results for 3-second predictions that do not require autoregressive adjustments for the comparison methods. Lower is better for all metrics except Maximum Temporal Correlation. Metrics are scaled as indicated in the column headers. **Bold** values indicate the best scores, while underlined values denote the second-best scores.

Dataset	Method	Levenshtein ($\times 10^3$) ↓		Disc. Fréchet ($\times 10^2$) ↓		DTW ($\times 10^4$) ↓		Max. Temp. Corr. ↑	
		Mean	Best	Mean	Best	Mean	Best	Mean	Best
DIEM	DeepGaze III	<u>0.459</u>	0.402	5.985	3.439	3.463	1.579	0.000	0.000
	DiffEye	0.499	0.459	9.857	8.358	6.233	5.112	0.021	0.325
	DiffEye retrained w/ saliency	0.464	0.426	4.402	3.269	2.224	1.586	<u>0.037</u>	0.348
	GazeFormer	0.466	0.430	4.538	3.363	2.364	1.616	0.012	0.322
	HAT	0.478	0.450	5.195	3.468	2.713	1.974	0.009	0.324
	Chen et al.	0.463	0.424	4.327	3.077	2.204	1.491	0.005	<u>0.384</u>
	TPP-Gaze	0.461	0.424	<u>3.947</u>	<u>2.741</u>	<u>2.066</u>	<u>1.318</u>	0.029	0.409
	Ours	0.447	<u>0.420</u>	3.312	2.450	1.599	1.205	0.156	0.374
DHF1K	DeepGaze III	0.441	0.401	2.410	1.585	1.446	0.941	0.017	0.462
	DiffEye	<u>0.426</u>	0.406	<u>2.031</u>	1.735	<u>1.139</u>	0.949	0.039	0.319
	DiffEye retrained w/ saliency	0.439	0.413	2.457	1.793	1.441	1.096	0.065	0.354
	GazeFormer	<u>0.430</u>	0.406	2.236	1.712	1.264	0.985	0.047	0.301
	HAT	0.444	0.422	3.001	2.332	1.501	1.226	0.028	0.302
	Chen et al.	<u>0.432</u>	0.400	2.341	1.562	1.370	0.897	<u>0.075</u>	0.402
	TPP-Gaze	0.434	<u>0.398</u>	2.193	1.357	1.204	<u>0.770</u>	0.028	<u>0.442</u>
	Ours	0.422	0.395	1.935	<u>1.363</u>	1.006	0.745	0.109	0.393

Correlation further confirms that the predicted gaze trajectories follow similar temporal dynamics as human observers, capturing consistent motion patterns over time. These temporal motion patterns are particularly important for videos as content changes constantly and may only be visible for a short period of time. Together, these results indicate that the model captures both the sequential structure and the temporal dynamics of gaze behavior.

The results of the best scores show that individual samples generated by our model can closely approximate specific human gaze paths. In contrast, the consistently strong mean scores indicate that the model reliably aligns with the overall distribution and dynamics of human fixations, rather than producing occasional high-quality outliers.

Improvements in Discrete Fréchet distance are less pronounced. Discrete Fréchet emphasizes the maximum deviation along the trajectory and is dominated by local outliers. Minor timing shifts or short exploratory saccades, which are common in videos, can substantially increase the distance, even when the overall trajectory structure and temporal progression match well. As a result, Discrete Fréchet is less sensitive to global sequence similarity and temporal alignment, which are central to video gaze prediction.

Three Second Window Generation Quantitative results show a similar trend to the full-video evaluation. Additionally, the modular baseline “DiffEye retrained w/ saliency” remains substantially below our method, demonstrating that the performance gain comes from the combination of saliency-aware conditioning and autoregressive diffusion rather than from saliency conditioning alone.

4.3 Qualitative Comparisons

Visual Inspection Figures 2–3 show how the gaze predictions progress with the methods as the video advances. We further refer to the supplementary videos for

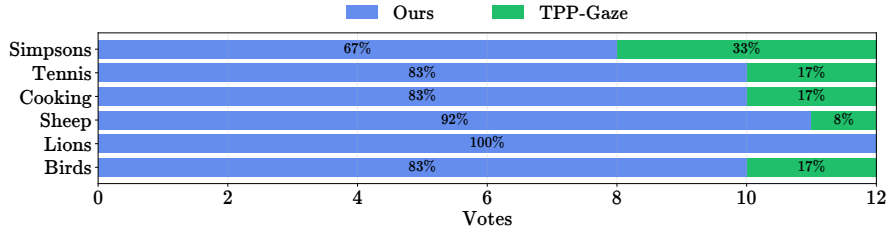


Fig. 4: *User study results.* Each row corresponds to one video. Our model (blue) receives consistently more votes than TPP-Gaze (green).

a comprehensive assessment of temporal dynamics and motion characteristics. The qualitative results support the quantitative findings. Our model produces gaze trajectories that are temporally smooth and spatially stable. The predicted paths avoid abrupt jumps and exhibit coherent motion over time. In addition, the gaze behavior responds consistently to scene content and camera motion, adapting to moving objects and global motion patterns. In contrast, the comparison methods often generate temporally fragmented or unstable trajectories that do not react well to content change, which explains their lower scores in sequence- and time-aware metrics.

User Study In the quantitative evaluation, our method achieved the best performance on all metrics except Discrete Fréchet Distance, where TPP-Gaze obtained the lowest score. To assess whether this outcome reflects limitations of our model or potential biases from the Discrete Fréchet metric for video-based gaze prediction, we conducted a qualitative user study comparing our model against TPP-Gaze. We randomly selected six videos from the benchmark, three from each dataset. For randomly chosen ground-truth trajectories, we identified the closest predictions from both methods based on the Discrete Fréchet score. We rendered the ground truth and both model predictions as moving dot overlays on the videos (see supplementary material). The ground truth trajectory was displayed in a fixed color across all videos, while the two predicted trajectories were assigned random colors. Twelve participants took part in the study. For each video, participants were required to watch the full sequence and could replay it before making a decision. They were asked to select the trajectory that appeared closest to the ground truth over the entire video.

The results are shown in Figure 4. Across all trials, our method was identified as more similar to the ground truth trajectories in 85% of the cases. Our predictions received consistently better ratings than TPP-Gaze in every evaluated instance. These findings indicate that the inferior ranking under Discrete Fréchet Distance does not necessarily correspond to poorer perceptual quality. Instead, the metric appears less suitable for evaluating gaze prediction in videos than for static scanpath comparison tasks.

Table 3: Computational efficiency on DIEM validation batches using an NVIDIA RTX A6000. **Bold** values indicate the better scores.

Metric	Ours	Raw Pixels	Improvement
Parameters (Mil)	8.27	32.33	$3.91\times$ fewer
Checkpoint size (MB)	37.89	348.54	$9.20\times$ smaller
Data load time (s)	0.0085	0.6353	$74.85\times$ faster
Forward FLOPs (G)	1.18	71.65	$60.95\times$ fewer
Throughput (samples/s)	49.19	17.61	$2.79\times$ higher
Peak memory (MB)	49.50	974.90	$19.69\times$ lower

Table 4: *Ablation study results.* Regular training and if not denoted otherwise is run with 90 history samples and 45 prediction samples on the DIEM dataset. Lower scores are better for all metrics except Maximum Temporal Correlation. Metrics are scaled as indicated in the column headers. **Bold** values indicate the best scores, while underlined values denote the second-best scores.

Method	Levenshtein ($\times 10^3$) $\downarrow$		Disc. Fréchet ($\times 10^2$) $\downarrow$		DTW ($\times 10^4$) $\downarrow$		Max. Temp. Corr. $\uparrow$	
	Mean	Best	Mean	Best	Mean	Best	Mean	Best
Ours regular	3.40	<u>3.31</u>	4.14	3.35	12.33	<u>11.44</u>	0.235	0.314
10-Prediction	3.97	3.10	3.67	2.90	17.31	6.88	0.048	0.126
90-Prediction	3.77	3.63	4.38	3.58	14.65	13.84	<u>0.230</u>	<u>0.309</u>
No initial history	<u>3.41</u>	3.31	<u>4.13</u>	<u>3.34</u>	<u>12.39</u>	11.57	0.235	0.314
MAGVIT2 Conditioning	3.82	3.76	7.97	6.96	54.92	48.15	0.054	0.082

4.4 Computational Efficiency

Our saliency latent compression architecture was empirically chosen to preserve gaze-relevant information while reducing conditioning dimensionality and improving convergence. Tab. 3 compares the computational efficiency of our saliency-compressed conditioning against naive raw-pixel conditioning. Our compressed representation reduces the parameter count from 32.33M to 8.27M, yielding a $3.91\times$ smaller model, while decreasing checkpoint size by $9.20\times$. It also substantially improves runtime efficiency: data loading is $74.85\times$ faster, forward computation is reduced from 71.65G to 1.18G FLOPs, and peak memory drops from 974.90MB to 49.50MB. As a result, throughput increases from 17.61 to 49.19 samples/s, corresponding to a $2.79\times$ speedup.

4.5 Ablation Studies

Conditioning We compare video saliency latents with a visual tokenizer-based alternative using an open-source implementation of MAGVIT2 [15]. To mirror our saliency pipeline, we converted each video frame into tokens using the pre-trained encoder. The resulting token embeddings were then applied as per-frame conditioning signals. The goal was to compress image frames into a general-purpose latent representation that could serve as conditioning without relying on saliency-specific structure. Unlike spatial saliency latents, this representation

provides a discrete global summary of each frame. However, in practice, models trained with these tokenized conditioning signals failed to converge, while saliency latent conditioning trained stably. This convergence problem is also highlighted in Table 4 with significantly degraded results on video gaze prediction.

Prediction Length We analyze the influence of the prediction horizon in the autoregressive diffusion model. The history length is fixed to 90 frames. The prediction length is varied during both training and inference. Table 4 summarizes the results (*Prediction* rows). Short predictions (10 frames) reduce the mean Discrete Fréchet distance and achieve the best scores for several best scores. This improvement indicates that short-horizon prediction can simplify the temporal modeling problem and reduces trajectory drift. However, the mean DTW error increases substantially, which suggests reduced robustness when evaluated over full sequences. Extending the prediction length to 90 frames degrades most metrics compared to the regular setting. The increase in Levenshtein and DTW distances indicates accumulated autoregressive error and higher temporal uncertainty over long horizons. Overall, we find that a 45-frames prediction provides the best trade-off between short-term accuracy and long-term stability.

Minimal Temporal Initialization To evaluate the model’s dependency on historical motion priors, we conduct an ablation study using a minimal initialization strategy (Table 4, *No initial history* row). The primary experiment initially conditions the autoregressive model on a 90-frame ground-truth history. We run a round of predictions on DIEM that replaces the initial buffer with a static cold start approach by sampling only a single ground-truth coordinate and replicating it to fill the full history window. This setup effectively removes all velocity and directional cues, forcing the model to initiate predictions from a zero-motion state. Even under this constrained scenario, the results remain robust: both Levenshtein distance and Maximum Temporal Correlation show identical results to the baseline, suggesting that the model could recover the necessary temporal dynamics from the latent features. Only DTW decreases slightly, potentially due to an initial misalignment before the history is built through autoregression. However, the overall consistency of the results demonstrates stable and plausible gaze trajectory generation even without motion priors.

5 Limitations and Discussions

Long-Range Dependencies. Modeling long-range behavioral consistency remains a challenge for gaze generation. The proposed autoregressive diffusion model captures high-frequency temporal dynamics of gaze trajectories. However, the model relies on a sliding window of recent predictions with a fixed length. Human gaze behavior can depend on information beyond such short-term history. Long-term memory and evolving scene understanding can influence viewing behavior in dynamic scenes [26]. This dependency is particularly relevant in videos

with static cameras and persistent background structure [57]. Future work could adopt mechanisms from video world models that maintain information over long temporal horizons [53].

Evaluation metrics. Our method achieved the highest performance in most of the quantitative analyses and human subject studies. One exception is the Discrete Fréchet Distance. This metric is typically applied to low-dimensional scanpath data and emphasizes the maximum pointwise deviation along a trajectory. Discrete Fréchet Distance is highly sensitive to local outliers. Running the metric with a dense gaze path can lead to short exploratory saccades, minor temporal shifts, or brief deviations from the ground-truth path dominating the score. It is not designed to evaluate the alignment of global temporal structure and motion trends, which is better expressed by Levenshtein distance or DTW. The user study results further indicate that lower Discrete Fréchet scores do not necessarily correspond to better perceptual similarity. More generally, all current trajectory metrics rely on pairwise comparisons to a single ground-truth path and do not account for the inherent stochasticity and multi-modality of human gaze. Human observers exhibit substantial inter-subject variability, and multiple qualitatively different trajectories can be equally plausible. Future work should therefore design metrics that are specifically built to quantify the perceptual validity of generated video gaze trajectories.

Head movements. In our training datasets, the gaze trajectories were acquired with fixed head poses. While head movements may not commonly be a factor in desktop viewing settings, they may be common in physical world and wide-field displays [3, 19]. Therefore, extending the current gaze trajectory towards higher-dimensional models that jointly generate head orientations and their relative gaze may shed light on application scenarios such as imitation learning and virtual/augmented reality (VR/AR).

6 Conclusion

In this paper, we introduce emerging autoregressive diffusion models to generate infinite and realistic human gaze movements in videos of arbitrary length. We believe this work establishes a new frontier, offering a path toward generative world models that not only simulate physical environments but also understand the context-aware behavioral responses of the humans observing them.

References

1. Aronov, B., Har-Peled, S., Knauer, C., Wang, Y., Wenk, C.: Fréchet distance for curves, revisited. In: European symposium on algorithms. pp. 52–63. Springer (2006)

2. Assens, M., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: Pathgan: Visual scanpath prediction with generative adversarial networks. In: ECCV Workshop on Egocentric Perception, Interaction and Computing (EPIC) (2018), <https://arxiv.org/abs/1809.00567>
3. Bahill, A.T., Adler, D., Stark, L.: Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science* **14**(6), 468–469 (1975)
4. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd international conference on knowledge discovery and data mining. pp. 359–370 (1994)
5. Borji, A.: Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(2), 679–700 (2021)
6. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence* **35**(1), 185–207 (2013)
7. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A.: Where should saliency models look next? In: European Conference on Computer Vision. pp. 809–824. Springer (2016)
8. Chao, F.Y., Ozcinar, C., Smolic, A.: Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need (10 2021). <https://doi.org/10.1109/MMSP53017.2021.9733647>
9. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems* **37**, 24081–24125 (2025)
10. Chen, X., Jiang, M., Zhao, Q.: Predicting human scanpaths in visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10876–10885 (2021)
11. D’Amelio, A., Cartella, G., Cuculo, V., Lucchi, M., Cornia, M., Cucchiara, R., Boccignone, G.: Tpp-gaze: modelling gaze dynamics in space and time with neural temporal point processes. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 8775–8784 (2025)
12. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al.: Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys* **58**(3), 1–38 (2025)
13. Dong, Z., Hao, J., Li, P., Ni, F., Wang, Y., Yuan, Y., Zheng, Y.: Diffuserlite: Towards real-time diffusion planning. In: *Advances in Neural Information Processing Systems*. vol. 37, pp. 122556–122585 (2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/dd6a47bc0aad6f34aa5e77706d90cdc4-Abstract-Conference.html
14. Droste, R., Jiao, J., Noble, J.A.: Unified image and video saliency modeling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) *Computer Vision – ECCV 2020*. pp. 419–435. Springer International Publishing, Cham (2020)
15. Fang, X., Yao, Y., Jiang, X., Li, X., Yu, N., Wang, Y.: O2-magvit2 (2024), <https://github.com/cofe-ai/O2-MAGVIT2>
16. Fang, Y., Yu, J., Zhang, H., van der Lans, R., Shi, B.: Oat: Object-level attention transformer for gaze scanpath prediction. In: European Conference on Computer Vision. pp. 366–382. Springer (2024)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239* (2020)
18. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. *arXiv:2204.03458* (2022)

19. Hu, Z., Li, S., Zhang, C., Yi, K., Wang, G., Manocha, D.: Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics* **26**(5), 1902–1911 (2020)
20. Janner, M., Du, Y., Tenenbaum, J.B., Levine, S.: Planning with diffusion for flexible behavior synthesis. In: *International Conference on Machine Learning* (2022)
21. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: Deepvs: A deep learning based video saliency prediction approach. In: *Proceedings of the european conference on computer vision (eccv)*. pp. 602–617 (2018)
22. Jiao, C., Wang, Y., Zhang, G., Bâce, M., Hu, Z., Bulling, A.: Diffgaze: A diffusion model for modelling fine-grained human gaze behaviour on 360° images. *ACM Transactions on Interactive Intelligent Systems* (2025)
23. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *IEEE International Conference on Computer Vision (ICCV)*. pp. 2106–2113 (2009). <https://doi.org/10.1109/ICCV.2009.5459462>
24. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *2009 IEEE 12th international conference on computer vision*. pp. 2106–2113. IEEE (2009)
25. Kara, O., Nisar, H., Rehg, J.M.: Diffeye: Diffusion-based continuous eye-tracking data generation conditioned on natural images. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025), <https://openreview.net/forum?id=P5yoTfwyD>
26. Kruijne, W., Meeter, M.: Implicit short-and long-term memory direct our gaze in visual search. *Attention, Perception, & Psychophysics* **78**(3), 761–773 (2016)
27. Kümmerer, M., Bethge, M., Wallis, T.S.: Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision* **22**(5), 7–7 (2022). <https://doi.org/10.1167/jov.22.5.7>, <https://doi.org/10.1167/jov.22.5.7>
28. Kümmerer, M., Bethge, M., Wallis, T.S.A.: Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision* **22**(5), 7–7 (04 2022). <https://doi.org/10.1167/jov.22.5.7>, <https://doi.org/10.1167/jov.22.5.7>
29. Li, M., Zhong, B., Lobaton, E., Huang, H.: Fusion of human gaze and machine vision for predicting intended locomotion mode. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **30**, 1103–1112 (2022)
30. Li, M., Fan, K., Ma, K.: Scanpath prediction in panoramic videos via expected code length minimization (2025), <https://arxiv.org/abs/2305.02536>
31. Martin, D., Serrano, A., Bergman, A.W., Wetzstein, G., Masia, B.: Scangan360: A generative model of realistic scanpaths for 360 images. *IEEE Transactions on Visualization and Computer Graphics* **28**(5), 2003–2013 (2022)
32. Martin, D., Serrano, A., Bergman, A.W., Wetzstein, G., Masia, B.: Scangan360: A generative model of realistic scanpaths for 360° images. *IEEE Transactions on Visualization and Computer Graphics* **28**(5), 2003–2013 (2022). <https://doi.org/10.1109/TVCG.2022.3150502>
33. Mital, P.K., Smith, T.J., Hill, R.L., Henderson, J.M.: Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation* **3**(1), 5–24 (2011)
34. Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., Hoai, M.: Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1441–1450 (June 2023)
35. Nguyen, A., Yan, Z.: Enhancing 360 video streaming through salient content in head-mounted displays. *Sensors* **23**(8), 4016 (2023)

36. Nguyen, A., Yan, Z., Nahrstedt, K.: Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1190–1198 (2018)
37. Nishiyasu, T., Sato, Y.: Gaze scanpath transformer: Predicting visual search target by spatiotemporal semantic modeling of gaze scanpath. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 625–635 (2024)
38. OpenAI: Video generation models as world simulators. <https://openai.com/index/video-generation-models-as-world-simulators/> (2024), accessed: 3/3/2026
39. Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., Lefohn, A.: Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions On Graphics (TOG)* **35**(6), 1–12 (2016)
40. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
41. Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., Huang, X.: Long-context state-space video world models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8733–8744 (2025)
42. Privitera, C.M., Stark, L.W.: Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence* **22**(9), 970–982 (2002)
43. Ricker, T.J., Cowan, N.: Loss of visual working memory within seconds: the combined use of refreshable and non-refreshable features. *Journal of experimental psychology: Learning, memory, and cognition* **36**(6), 1355 (2010)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
45. Singer, O., Polyak, G., Hayes, T., Ping, X., Chen, S.F., Ding, Y., Yuan, S., Adejuwon, E., Korthy, J., Tesfaye, H., et al.: Make-a-video: Text-to-video generation without text-video data (2022), <https://arxiv.org/abs/2209.14792>
46. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
47. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=PXTIG12RRHS>
48. Sui, X., Fang, Y., Zhu, H., Wang, S., Wang, Z.: Scandmm: A deep markov model of scanpath prediction for 360° images. In: Proceedings - 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 6989–6999. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, United States (2023). <https://doi.org/10.1109/CVPR52729.2023.00675>, <https://cvpr2023.thecvf.com/Conferences/2023>, <https://openaccess.thecvf.com/menu>, <https://ieeexplore.ieee.org/xpl/conhome/1000147/all-proceedings>, full text of this publication does not contain sufficient affiliation information. With consent from the author(s) concerned, the Research Unit(s) information for this record is based on the existing academic department affiliation of the author(s).; 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), CVPR2023 ; Conference date: 18-06-2023 Through 22-06-2023

49. Sun, W., Chen, Z., Wu, F.: Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(6), 2101–2118 (2019). <https://doi.org/10.1109/TPAMI.2019.2957262>
50. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: *Proceedings of the IEEE Conference on computer vision and pattern recognition*. pp. 4894–4903 (2018)
51. Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., Borji, A.: Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence* **43**(1), 220–237 (2019)
52. Wang, Y., Zhang, F.L., Dodgson, N.A.: Scantd: 360° scanpath prediction based on time-series diffusion. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. p. 7764–7773. MM '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3664647.3681315>, <https://doi.org/10.1145/3664647.3681315>
53. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284* (2025)
54. Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., Gao, S.: Gaze prediction in dynamic 360° immersive videos. pp. 5333–5342 (06 2018). <https://doi.org/10.1109/CVPR.2018.00559>
55. Xue, R., Xu, J., Mondal, S., Le, H., Zelinsky, G., Hoai, M., Samaras, D.: Few-shot personalized scanpath prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13497–13507 (2025)
56. Yang, Z., Mondal, S., Ahn, S., Xue, R., Zelinsky, G., Hoai, M., Samaras, D.: Unifying top-down and bottom-up scanpath prediction using transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1683–1693 (2024)
57. Yarbus, A.L.: *Eye movements and vision*. Springer (2013)
58. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: *CVPR* (2025)
59. Zelinsky, G., Yang, Z., Huang, L., Chen, Y., Ahn, S., Wei, Z., Adeli, H., Samaras, D., Hoai, M.: Benchmarking gaze prediction for categorical visual search. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–0 (2019)
60. Zeng, H., Liu, W., Xia, T., Chen, J., Li, Z., Zhang, S.Q.: Foveated instance segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 24496–24505 (2025)
61. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., Cao, X., Yin, W.: Epona: Autoregressive diffusion world model for autonomous driving. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2025)
62. Zhang, Q., Hu, Z., Song, Y., Pei, J., Liu, J.: The human gaze helps robots run bravely and efficiently in crowds. In: *2023 IEEE international conference on robotics and automation (ICRA)*. pp. 7540–7546. IEEE (2023)
63. Zhu, Y., Wang, Y., Zhang, W., Chen, J., Li, Y.: Scanpath prediction in panoramic videos through multimodal fusion. *Displays* **91**, 103199 (2026). <https://doi.org/https://doi.org/10.1016/j.displa.2025.103199>, <https://www.sciencedirect.com/science/article/pii/S0141938225002367>