DiffEye: Diffusion-Based Continuous Eye-Tracking Data Generation Conditioned on Natural Images

Ozgur Kara*, Harris Nisar*, James M. Rehg

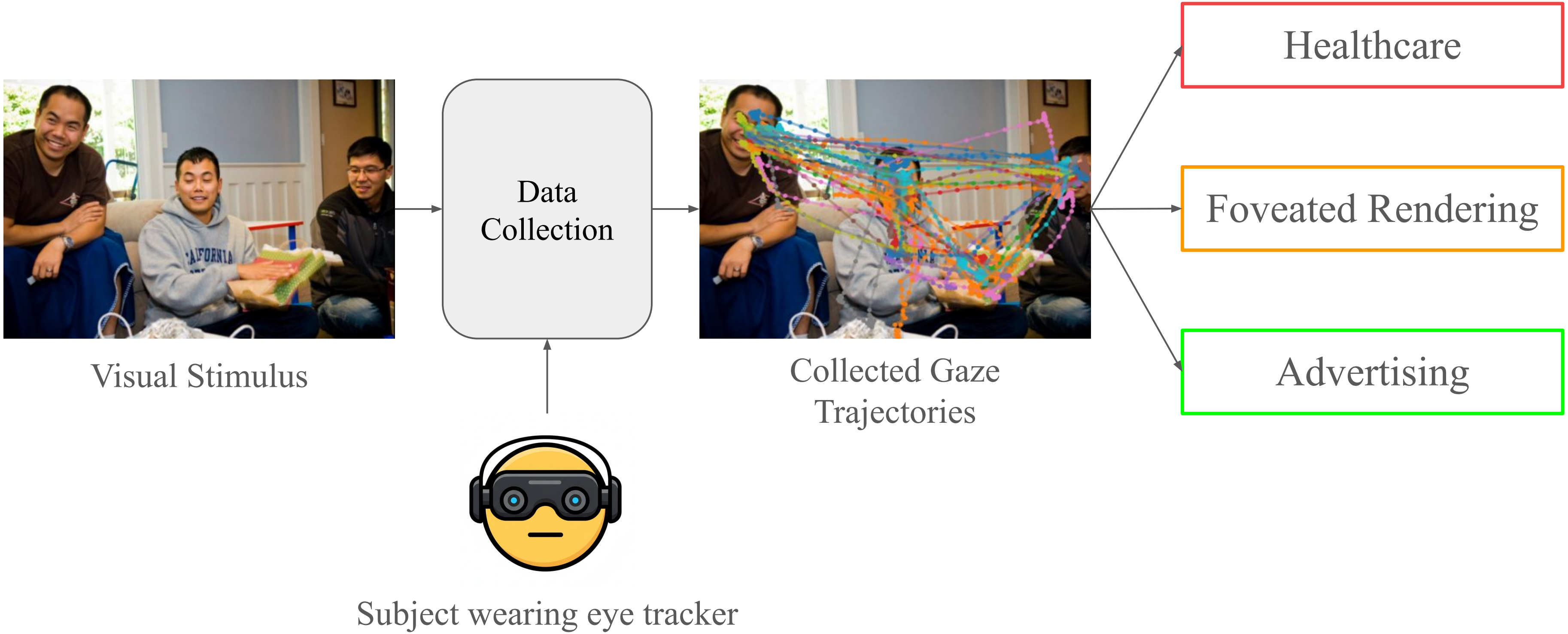
University of Illinois Urbana-Champaign

* Equal Contribution

NeurIPS 2025

https://diff-eye.github.io/

Introduction



How is Visual Attention Represented?

Gaze Trajectories

Saliency Map

Scanpaths

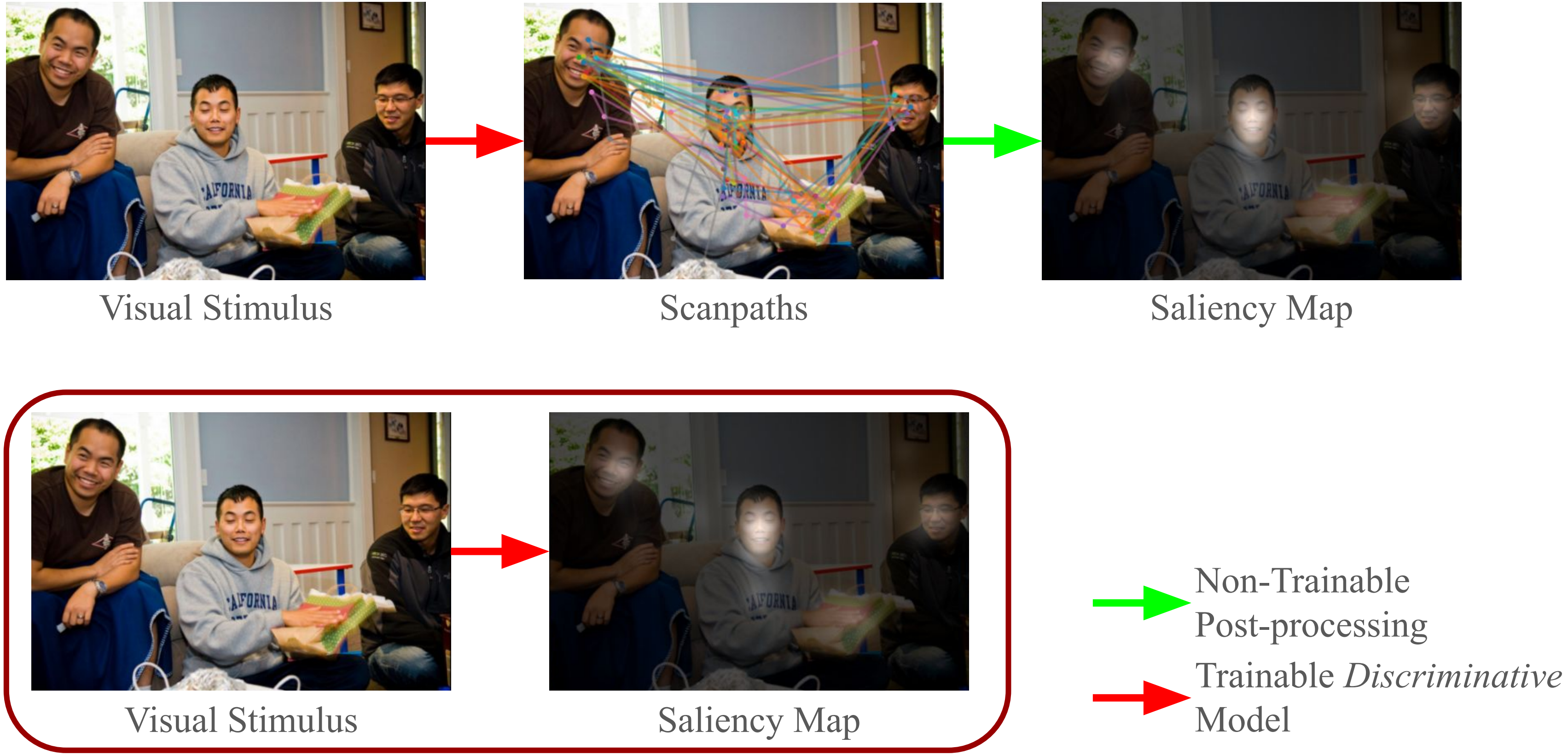
Non-Trainable Post-processing

Existing Approaches

Non-Trainable Post-processing

Trainable *Discriminative* Model

Existing Approaches



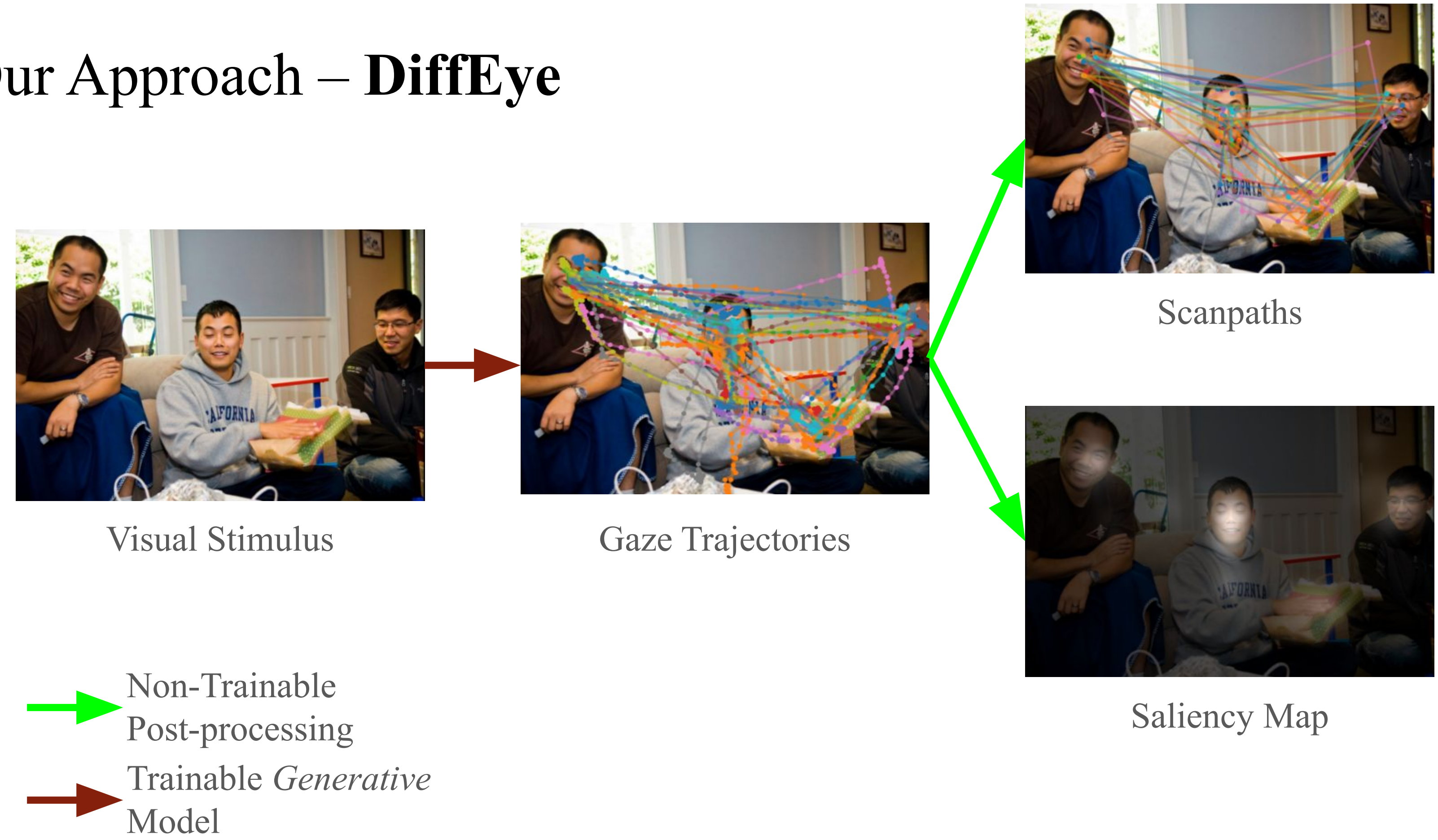
Limitations of Existing Approaches

- **Limitation 1 -** Scanpath & Saliency Only Training → *This process **discards the rich temporal information** contained in the raw eye-tracking trajectories*
  - ***Hypothesis:*** *Training on full, continuous trajectories will capture this lost information and lead to more accurate scanpath prediction.*

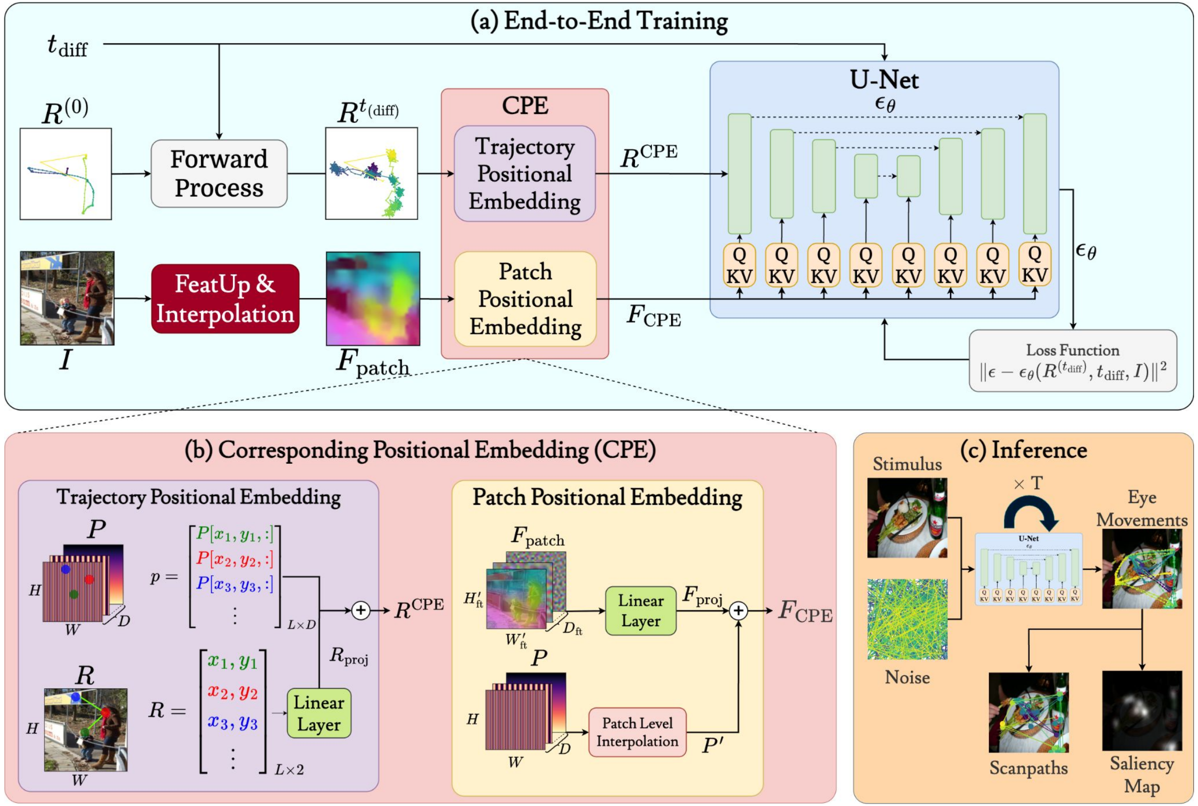
Limitations of Existing Approaches

- **Limitation 1 -** Scanpath & Saliency Only Training → *This process **discards the rich temporal information** contained in the raw eye-tracking trajectories*
  - ***Hypothesis:*** *Training on full, continuous trajectories will capture this lost information and lead to more accurate scanpath prediction.*
- **Limitation 2 -** Discriminative models → *This conflicts with **the inherent variability and stochastic nature** of human attention.*
  - ***Hypothesis:*** *Generative models are better suited to learn this rich, variable distribution compared to deterministic ones.*

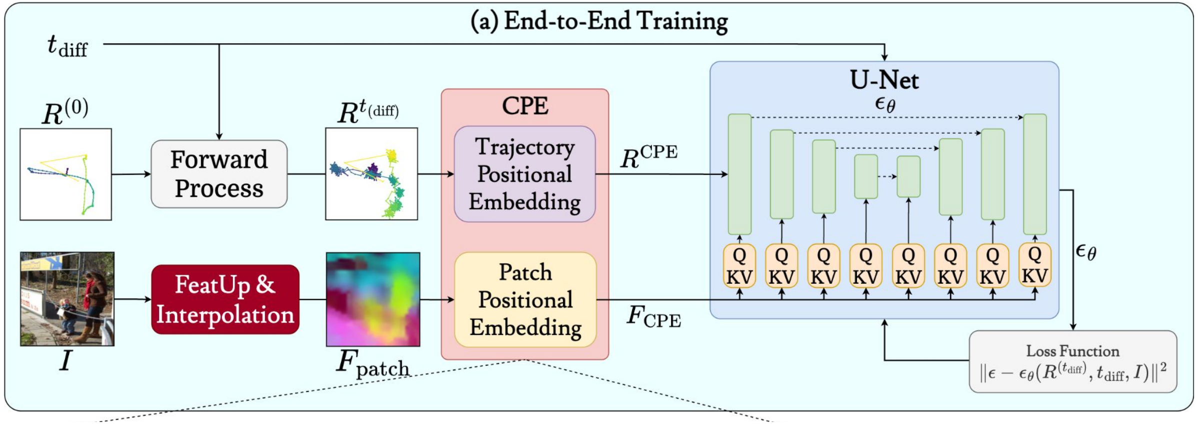
Our Approach – **DiffEye**



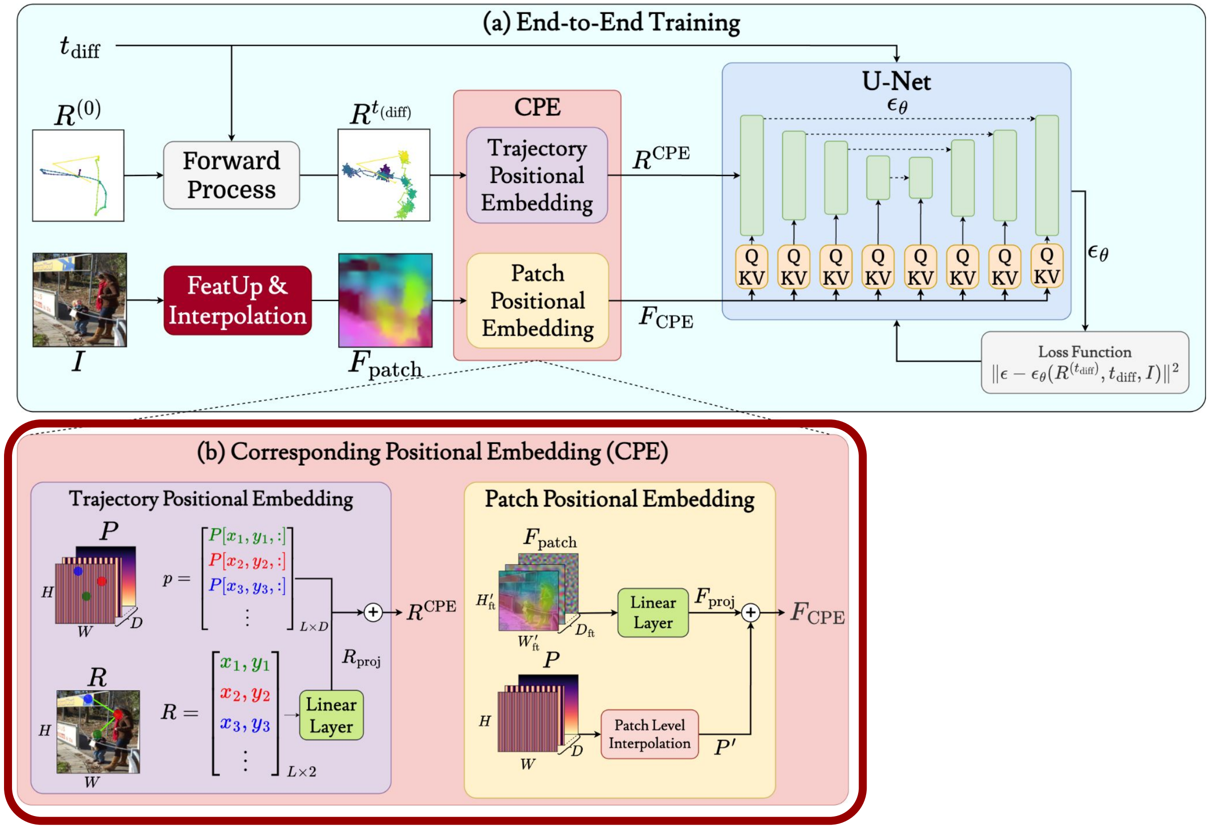
Our Approach - Overall Framework



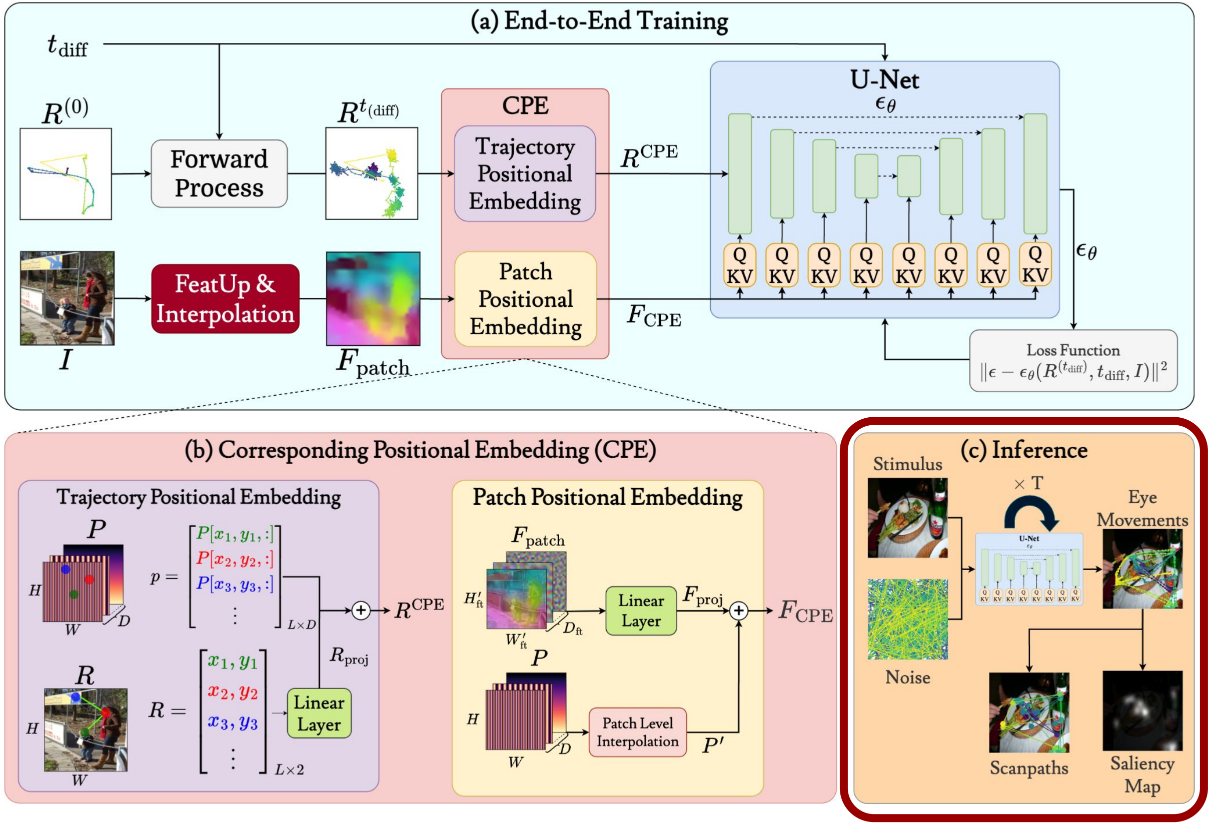
Our Approach - End-to-End Training



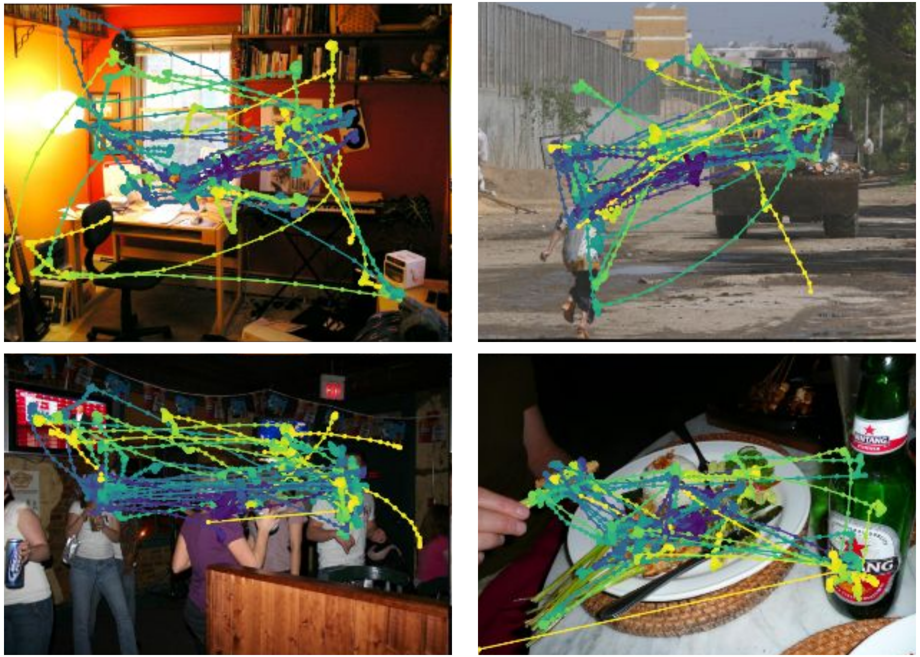
Our Approach - Corresponding Positional Embedding (CPE)



Our Approach - Inference



Datasets



MIT1003 Examples

- **We trained and evaluated DiffEye using the MIT1003 dataset (Judd et al.)**

Datasets

OSIE Examples

- We trained and evaluated DiffEye using the MIT1003 dataset (Judd et al.)
- **We also evaluated on the test split of the OSIE scanpath dataset (Xu et al.)**

Experiment 1 - Scanpath Generation

- **Compared DiffEye to 5 scanpath models**

Experiment 1 - Scanpath Generation

- Compared DiffEye to 5 scanpath models
- **Generated 15 scanpaths per model and computed 2 scores (*Best* & *Mean*) for 4 standard metrics to compare trajectories**

Experiment 1 - Scanpath Generation

- Compared DiffEye to 5 scanpath models
- Generated 15 scanpaths per model and computed 2 scores (*Best* & *Mean*) for 4 standard metrics to compare trajectories
  - ***Best* was found by computing the metric between each pair of generated and ground truth scanpath, finding the smallest (or least distant) one per image and then averaging over all images.**

Experiment 1 - Scanpath Generation

- Compared DiffEye to 5 scanpath models
- Generated 15 scanpaths per model and computed 2 scores (*Best* & *Mean*) for 4 standard metrics to compare trajectories
    - *Best* was found by computing the metric between each pair of generated and ground truth scanpath, finding the smallest (or least distant) one per image and then averaging over all images.
    - ***Mean* score was found by computing the metric between each pair of generated and ground truth scanpath, finding the mean per image, and then averaging over all images**

Experiment 1 - Scanpath Generation

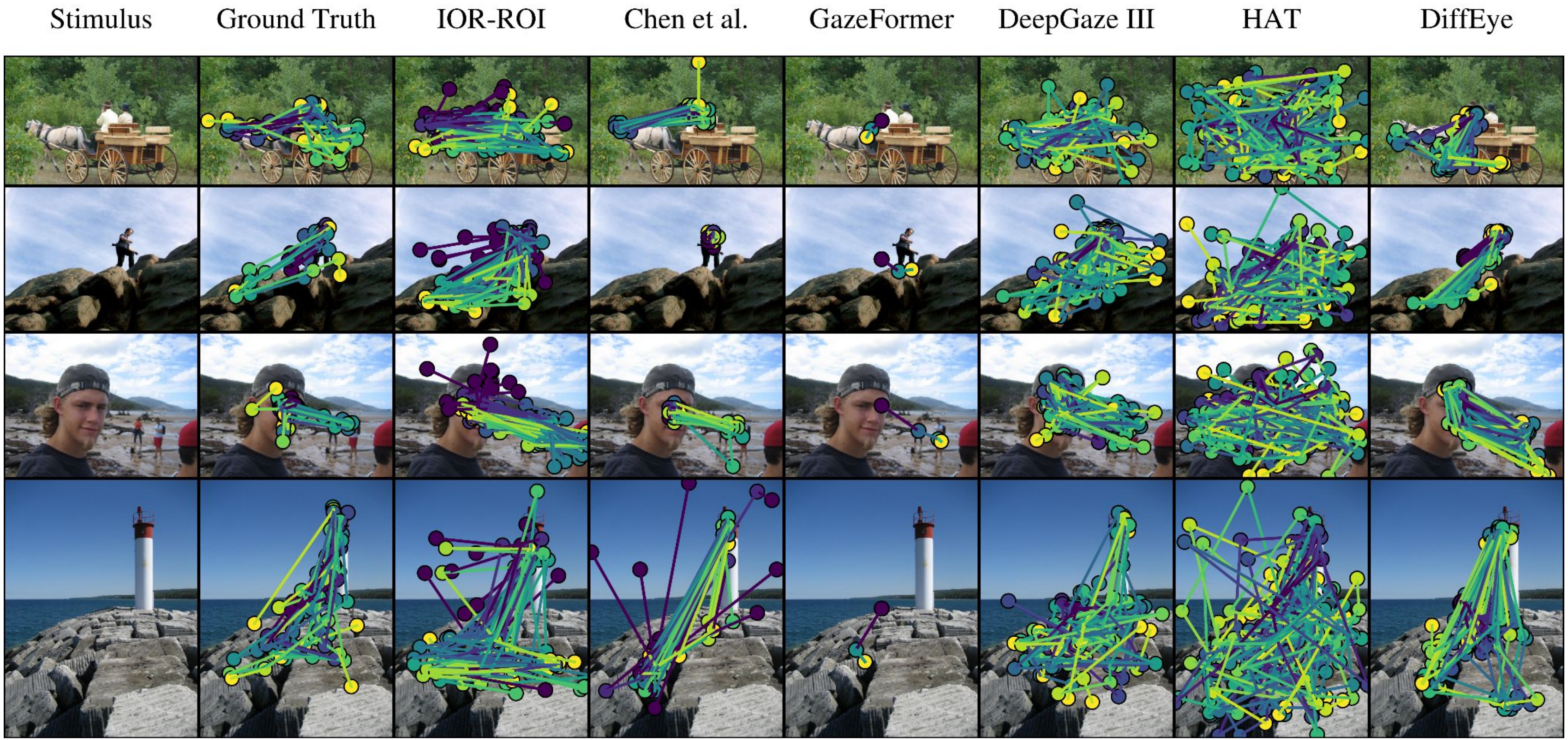
Test Dataset	Method	Levenshtein Distance ↓		Discrete Fréchet Distance ↓ ($\times 10^2$)		Dynamic Time Warping ↓ ($\times 10^3$)		Time Delay Embedding ↓	
		Mean	*Best*	*Mean*	*Best*	*Mean*	*Best*	*Mean*	*Best*
MIT1003	IOR-ROI	<u>13.574</u>	<u>11.092</u>	3.77	2.460	1.834	1.317	108.284	80.944
	DeepGaze III (Seen)	14.415	11.856	2.553	**2.160**	<u>1.757</u>	<u>1.141</u>	96.456	<u>65.408</u>
	Chen et al.	14.874	12.943	<u>3.704</u>	2.602	1.851	1.409	<u>92.100</u>	74.212
	GazeFormer	-	12.614	-	3.553	-	1.545	-	93.751
	HAT (seen)	18.440	14.645	4.293	2.940	2.680	1.862	131.516	97.232
	DiffEye	**13.009**	**9.709**	**3.529**	<u>2.449</u>	**1.573**	**1.067**	**88.661**	**53.486**
OSIE	IOR-ROI	<u>14.836</u>	<u>12.152</u>	3.357	<u>2.228</u>	<u>1.699</u>	1.167	92.960	70.624
	DeepGaze III	15.507	12.532	<u>3.206</u>	**2.077**	1.765	<u>1.166</u>	84.337	<u>57.786</u>
	Chen et al.	17.024	14.910	3.275	2.290	1.772	1.276	**78.286**	61.509
	GazeFormer	-	15.320	-	3.257	-	1.687	-	81.8789
	HAT	19.419	15.607	3.712	2.598	2.501	1.757	111.413	83.140
	DiffEye	**14.771**	**12.077**	**3.068**	2.238	**1.552**	**1.089**	<u>81.925</u>	**53.347**

Bold is best, <u>Underline</u> is second best

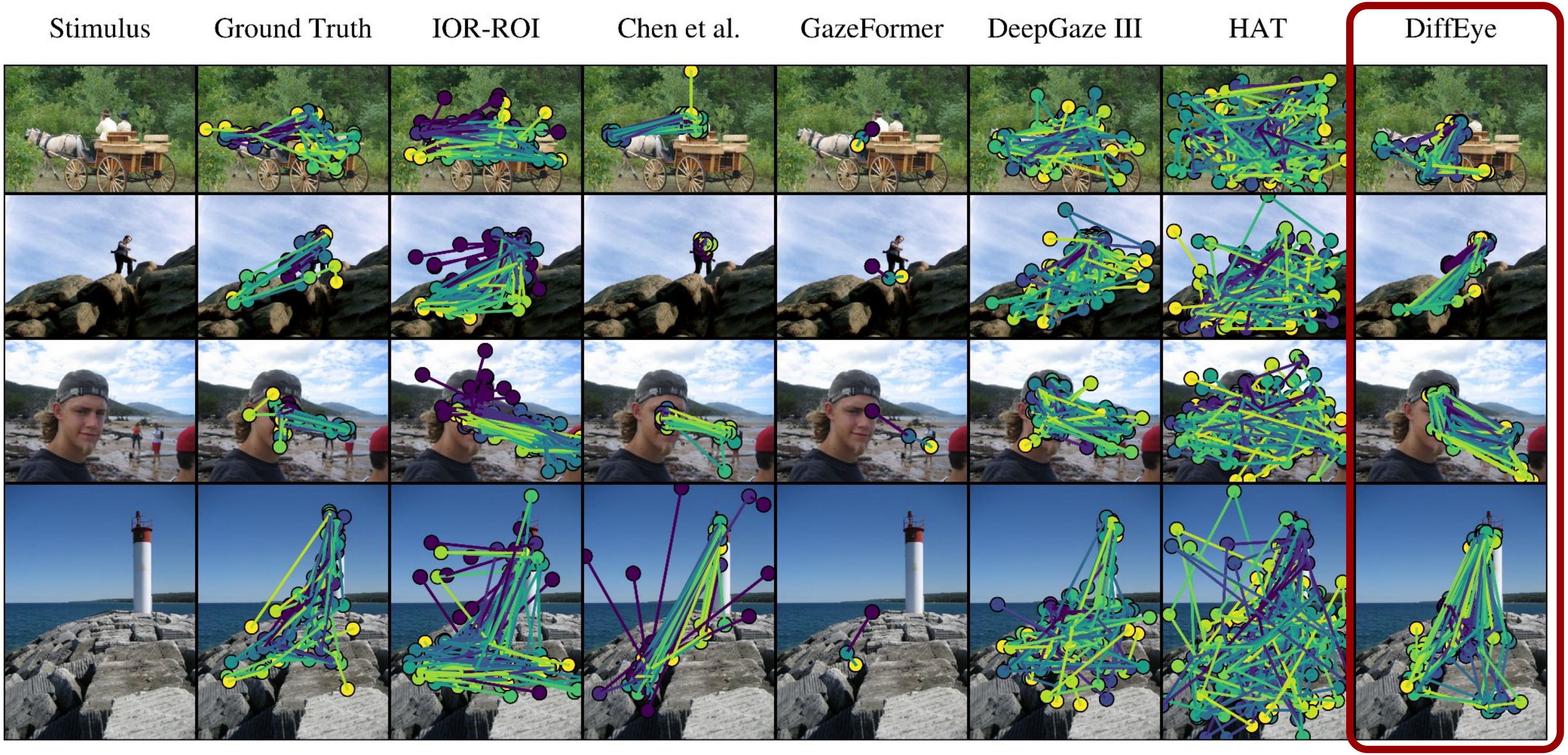
Experiment 1 - Scanpath Generation

	MIT1003		OSIE	
Model	SS (↑)	Sem SS (↑)	SS (↑)	Sem SS (↑)
DiffEye (ours)	**0.4782**	**0.6611**	0.4371	0.5837
HAT	0.4079	0.5794	0.4002	0.5791
GazeFormer	0.3531	0.4522	0.2713	0.3602
DeepGazeIII	0.4440	<u>0.6604</u>	**0.4623**	**0.6459**
ROI	<u>0.4506</u>	0.6603	<u>0.4404</u>	<u>0.6110</u>
Chen et al.	0.4237	0.6397	0.4333	0.5711

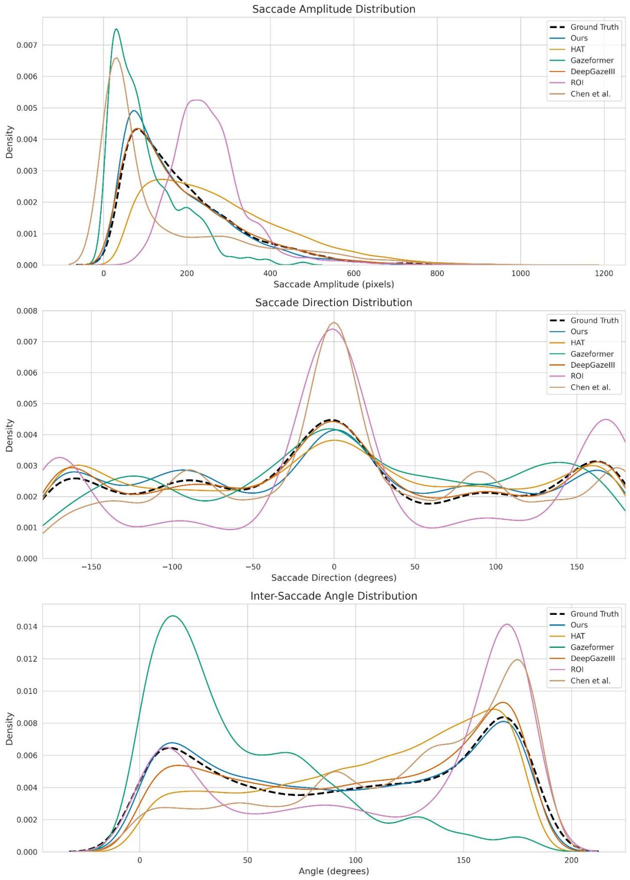
Experiment 1 - Scanpath Generation



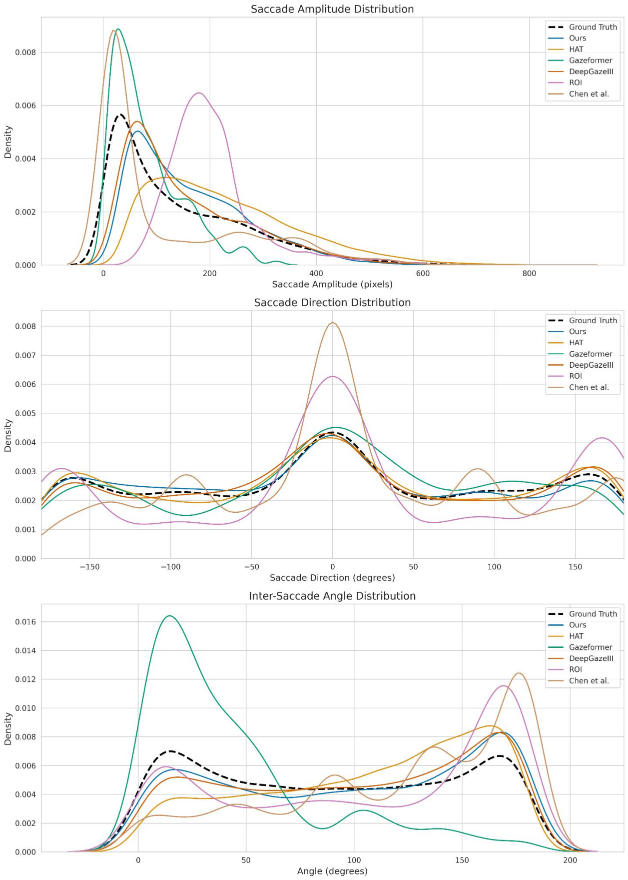
Experiment 1 - Scanpath Generation



Experiment 1 - Scanpath Generation



MIT1003



OSIE

Experiment 2 - Ablation

Bold is best

Task	Configuration	Levenshtein Distance ↓		Discrete Fréchet Distance ↓ ($\times10^2$)		Dynamic Time Warping ↓ ($\times10^3$)		Time Delay Embedding ↓	
		Mean	*Best*	*Mean*	*Best*	*Mean*	*Best*	*Mean*	*Best*
Scanpath Generation	Full Model: DiffEye	**0.130**	**0.097**	**3.529**	**2.449**	**0.157**	**0.107**	**88.661**	**53.486**
	Ablation 1: w/o FeatUp	0.133	0.100	3.546	2.423	0.163	0.110	91.007	60.103
	Ablation 2: w/o CPE	0.141	0.107	3.545	2.604	0.180	0.128	100.792	69.827
	Ablation 3: w/o U-Net Cross-Attention	0.143	0.107	3.701	2.557	0.189	0.130	107.962	68.353
	Ablation 4: w/o Patch Level Features	0.153	0.116	3.761	2.692	0.209	0.147	116.226	77.997
Eye Movement Trajectory Generation	Full Model: DiffEye	**10.083**	**8.289**	**3.601**	**2.460**	**11.834**	**8.212**	**35.228**	**20.968**
	Ablation 1: w/o FeatUp	10.265	8.736	3.844	2.623	12.513	8.645	41.224	26.453
	Ablation 2: w/o CPE	10.773	9.200	3.621	2.599	13.430	10.068	44.403	28.904
	Ablation 3: w/o U-Net Cross-Attention	10.971	9.394	3.828	2.587	14.716	10.992	56.739	38.264
	Ablation 4: w/o Patch Level Features	11.791	9.947	4.088	2.761	18.007	13.312	77.042	47.354

Experiment 2 - Ablation

Bold is best

Task	Configuration	Levenshtein Distance ↓		Discrete Fréchet Distance ↓ ($\times 10^2$)		Dynamic Time Warping ↓ ($\times 10^3$)		Time Delay Embedding ↓	
		Mean	*Best*	*Mean*	*Best*	*Mean*	*Best*	*Mean*	*Best*
Scanpath Generation	Full Model: DiffEye	**0.130**	**0.097**	**3.529**	**2.449**	**0.157**	**0.107**	**88.661**	**53.486**
	Ablation 1: w/o FeatUp	0.133	0.100	3.546	2.423	0.163	0.110	91.007	60.103
	Ablation 2: w/o CPE	0.141	0.107	3.545	2.604	0.180	0.128	100.792	69.827
	Ablation 3: w/o U-Net Cross-Attention	0.143	0.107	3.701	2.557	0.189	0.130	107.962	68.353
	Ablation 4: w/o Patch Level Features	0.153	0.116	3.761	2.692	0.209	0.147	116.226	77.997
Eye Movement Trajectory Generation	Full Model: DiffEye	**10.083**	**8.289**	**3.601**	**2.460**	**11.834**	**8.212**	**35.228**	**20.968**
	Ablation 1: w/o FeatUp	10.265	8.736	3.844	2.623	12.513	8.645	41.224	26.453
	Ablation 2: w/o CPE	10.773	9.200	3.621	2.599	13.430	10.068	44.403	28.904
	Ablation 3: w/o U-Net Cross-Attention	10.971	9.394	3.828	2.587	14.716	10.992	56.739	38.264
	Ablation 4: w/o Patch Level Features	11.791	9.947	4.088	2.761	18.007	13.312	77.042	47.354

Experiment 2 - Ablation

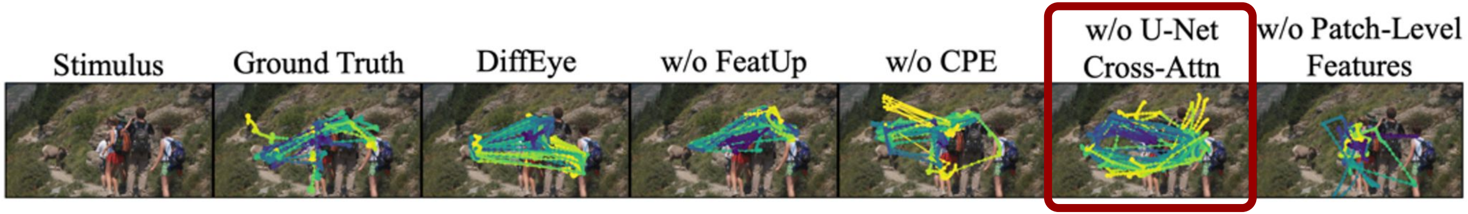
Bold is best

Task	Configuration	Levenshtein Distance ↓		Discrete Fréchet Distance ↓ ($\times 10^2$)		Dynamic Time Warping ↓ ($\times 10^3$)		Time Delay Embedding ↓	
		Mean	*Best*	*Mean*	*Best*	*Mean*	*Best*	*Mean*	*Best*
Scanpath Generation	Full Model: DiffEye	**0.130**	**0.097**	**3.529**	**2.449**	**0.157**	**0.107**	**88.661**	**53.486**
	Ablation 1: w/o FeatUp	0.133	0.100	3.546	2.423	0.163	0.110	91.007	60.103
	Ablation 2: w/o CPE	0.141	0.107	3.545	2.604	0.180	0.128	100.792	69.827
	Ablation 3: w/o U-Net Cross-Attention	0.143	0.107	3.701	2.557	0.189	0.130	107.962	68.353
	Ablation 4: w/o Patch Level Features	0.153	0.116	3.761	2.692	0.209	0.147	116.226	77.997
Eye Movement Trajectory Generation	Full Model: DiffEye	**10.083**	**8.289**	**3.601**	**2.460**	**11.834**	**8.212**	**35.228**	**20.968**
	Ablation 1: w/o FeatUp	10.265	8.736	3.844	2.623	12.513	8.645	41.224	26.453
	Ablation 2: w/o CPE	10.773	9.200	3.621	2.599	13.430	10.068	44.403	28.904
	Ablation 3: w/o U-Net Cross-Attention	10.971	9.394	3.828	2.587	14.716	10.992	56.739	38.264
	Ablation 4: w/o Patch Level Features	11.791	9.947	4.088	2.761	18.007	13.312	77.042	47.354

Experiment 2 - Ablation

Bold is best

Task	Configuration	Levenshtein Distance ↓		Discrete Fréchet Distance ↓ ($\times 10^2$)		Dynamic Time Warping ↓ ($\times 10^3$)		Time Delay Embedding ↓	
		Mean	*Best*	*Mean*	*Best*	*Mean*	*Best*	*Mean*	*Best*
Scanpath Generation	Full Model: DiffEye	**0.130**	**0.097**	**3.529**	**2.449**	**0.157**	**0.107**	**88.661**	**53.486**
	Ablation 1: w/o FeatUp	0.133	0.100	3.546	2.423	0.163	0.110	91.007	60.103
	Ablation 2: w/o CPE	0.141	0.107	3.545	2.604	0.180	0.128	100.792	69.827
	Ablation 3: w/o U-Net Cross-Attention	0.143	0.107	3.701	2.557	0.189	0.130	107.962	68.353
	Ablation 4: w/o Patch Level Features	0.153	0.116	3.761	2.692	0.209	0.147	116.226	77.997
Eye Movement Trajectory Generation	Full Model: DiffEye	**10.083**	**8.289**	**3.601**	**2.460**	**11.834**	**8.212**	**35.228**	**20.968**
	Ablation 1: w/o FeatUp	10.265	8.736	3.844	2.623	12.513	8.645	41.224	26.453
	Ablation 2: w/o CPE	10.773	9.200	3.621	2.599	13.430	10.068	44.403	28.904
	Ablation 3: w/o U-Net Cross-Attention	10.971	9.394	3.828	2.587	14.716	10.992	56.739	38.264
	Ablation 4: w/o Patch Level Features	11.791	9.947	4.088	2.761	18.007	13.312	77.042	47.354



Experiment 2 - Ablation

Bold is best

Task	Configuration	Levenshtein Distance ↓		Discrete Fréchet Distance ↓ ($\times 10^2$)		Dynamic Time Warping ↓ ($\times 10^3$)		Time Delay Embedding ↓	
		Mean	*Best*	*Mean*	*Best*	*Mean*	*Best*	*Mean*	*Best*
Scanpath Generation	Full Model: DiffEye	**0.130**	**0.097**	**3.529**	**2.449**	**0.157**	**0.107**	**88.661**	**53.486**
	Ablation 1: w/o FeatUp	0.133	0.100	3.546	2.423	0.163	0.110	91.007	60.103
	Ablation 2: w/o CPE	0.141	0.107	3.545	2.604	0.180	0.128	100.792	69.827
	Ablation 3: w/o U-Net Cross-Attention	0.143	0.107	3.701	2.557	0.189	0.130	107.962	68.353
	Ablation 4: w/o Patch Level Features	0.153	0.116	3.761	2.692	0.209	0.147	116.226	77.997
Eye Movement Trajectory Generation	Full Model: DiffEye	**10.083**	**8.289**	**3.601**	**2.460**	**11.834**	**8.212**	**35.228**	**20.968**
	Ablation 1: w/o FeatUp	10.265	8.736	3.844	2.623	12.513	8.645	41.224	26.453
	Ablation 2: w/o CPE	10.773	9.200	3.621	2.599	13.430	10.068	44.403	28.904
	Ablation 3: w/o U-Net Cross-Attention	10.971	9.394	3.828	2.587	14.716	10.992	56.739	38.264
	Ablation 4: w/o Patch Level Features	11.791	9.947	4.088	2.761	18.007	13.312	77.042	47.354

Conclusion

- **We introduced DiffEye, a diffusion-based model that generates diverse eye-movement trajectories for natural images.**

Conclusion

- We introduced DiffEye, a diffusion-based model that generates diverse eye-movement trajectories directly from raw eye-tracking data.
- **Our Corresponding Positional Embedding (CPE) helps the model align gaze positions with image semantics through shared spatial embeddings.**

Conclusion

- We introduced DiffEye, a diffusion-based model that generates diverse eye-movement trajectories directly from raw eye-tracking data.
- Our Corresponding Positional Embedding (CPE) helps the model align gaze positions with image semantics through shared spatial embeddings.
- **Our evaluation shows that DiffEye generates realistic eye movement trajectories.**

Conclusion

- We introduced DiffEye, a diffusion-based model that generates diverse eye-movement trajectories directly from raw eye-tracking data.
- Our Corresponding Positional Embedding (CPE) helps the model align gaze positions with image semantics through shared spatial embeddings.
- Our evaluation shows that DiffEye generates realistic eye movement trajectories.
- **DiffEye shows promise for modeling population-specific gaze patterns, with potential applications in developmental research.**