# Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention

Sounak Mondal[1], Zhibo Yang[1,2], Seoyoung Ahn[1], Dimitris Samaras[1], Gregory Zelinsky[2], Minh Hoai[1,3]
[1]Stony Brook University [2]Waymo LLC [3]VinAI Research

SCAN ME!
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA
TUE-AM-137

## Motivation

**Training Dataset:** search target in *N* categories

find **"fork"** find **"cup"**

Conventionally

find **"fork"** (**in** N categories)

**But what about**

find **"laptop"** (**outside** N categories)

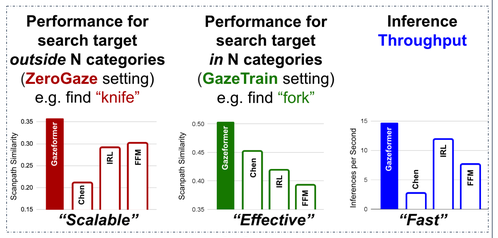
- Gaze prediction models for visual search require, *for each target category* – (1) **human gaze data**, (2) **detectors to encode the target**
- Hard to scale when **gaze/detection annotation is unavailable**
- Models must be ***scalable, accurate, fast*** to be used in HCI applications
  - Collecting annotation for all possible targets is impractical

## Solution

**Training Dataset:** search target in *N* categories

find **"tv"** find **"bowl"**

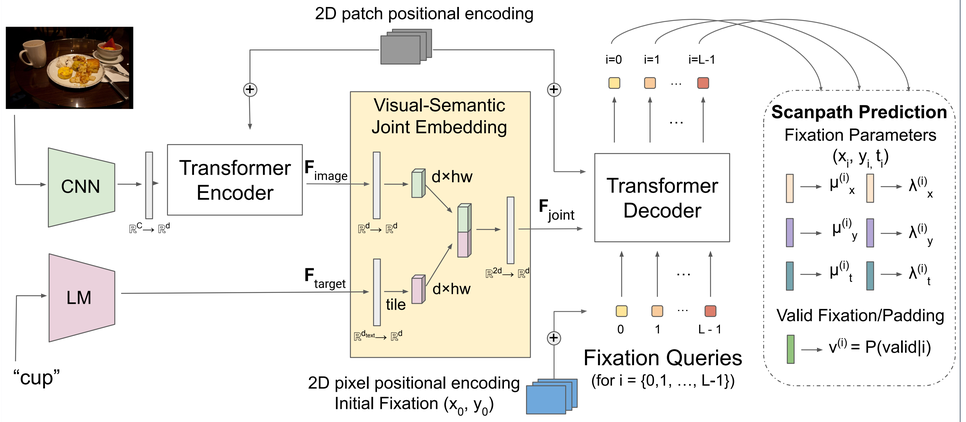
**Training**




**Inference**

- We propose a novel ***ZeroGaze*** task
  - Tests ***scalability*** of gaze prediction models for visual search
  - Extends **zero-shot learning** to gaze prediction
- We propose a novel ***Gazeformer*** *model to solve ZeroGaze*
  - Improves **scalability** – extends to unknown targets
  - Improves **effectiveness** – more accurate gaze prediction
  - Improves **efficiency** – faster than previous methods

## Gazeformer




- *Gazeformer is a **multimodal transformer encoder-decoder** based architecture*
- *Gazeformer* uses a **Language Model (LM)** to encode target name
  - ***Scalable*** – can hypothetically encode any target
  - Target semantics might help **extend** to unknown categories
- *Gazeformer* adopts a **transformer encoder-decoder architecture**
  - Learns **interactions** between image and target semantics
  - Models **spatio-temporal context** required for scanpath generation
  - Efficiently generates **entire sequence** of fixations in parallel
- Gazeformer ***regresses*** fixation parameters using Gaussian distributions
  - Previous methods predicted fixation probabilities over discrete image patches

$$x_i = \mu_{x_i} + \epsilon_{x_i} \cdot \exp(0.5\lambda_{x_i}), \quad y_i = \mu_{y_i} + \epsilon_{y_i} \cdot \exp(0.5\lambda_{y_i}),$$
$$t_i = \mu_{t_i} + \epsilon_{t_i} \cdot \exp(0.5\lambda_{t_i}), \quad \epsilon_{x_i}, \epsilon_{y_i}, \epsilon_{t_i} \in \mathcal{N}(0,1).$$

- Gazeformer learns ***scanpath termination***
  - Learns if a latent vector corresponds to a valid fixation or padding

## Experimental Results

| | SS↑ | | SemSS↑ | | FED↓ | | SemFED ↓ | | MM | CC | NSS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | w/o Dur | w/ Dur | w/o Dur | w/ Dur | w/o Dur | w/ Dur | w/o Dur | w/ Dur | ↑ | ↑ | ↑ |
| IRL | 0.290 | - | 0.314 | - | 4.606 | - | 4.377 | - | 0.774 | 0.241 | 4.018 |
| Chen *et al.* | 0.210 | 0.041 | 0.211 | 0.034 | 5.720 | 210.498 | 5.608 | 211.636 | 0.717 | 0.002 | 0.001 |
| FFM | 0.300 | - | 0.334 | - | 3.271 | - | 2.918 | - | 0.731 | 0.271 | **5.247** |
| Gazeformer-noDur | **0.359** | - | **0.391** | - | 2.788 | - | 2.474 | - | **0.822** | 0.316 | 4.671 |
| Gazeformer | 0.358 | **0.312** | **0.391** | **0.348** | **2.766** | **12.505** | **2.438** | **10.391** | 0.812 | **0.324** | 4.929 |

**"Scalable"** : ZeroGaze Setting

| | SS↑ | | SemSS↑ | | FED↓ | | SemFED ↓ | | MM | CC | NSS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | w/o Dur | w/ Dur | w/o Dur | w/ Dur | w/o Dur | w/ Dur | w/o Dur | w/ Dur | ↑ | ↑ | ↑ |
| Human | 0.490 | 0.409 | 0.548 | 0.456 | 2.531 | 11.526 | 1.637 | 8.086 | 0.857 | 0.472 | 8.129 |
| IRL | 0.418 | - | 0.499 | - | 2.722 | - | 2.182 | - | 0.833 | 0.434 | 6.895 |
| Chen *et al.* | 0.451 | 0.403 | 0.504 | 0.446 | 2.187 | 10.795 | 1.788 | 8.782 | 0.820 | 0.547 | 6.901 |
| FFM | 0.392 | - | 0.443 | - | 2.693 | - | 2.284 | - | 0.808 | 0.370 | 5.576 |
| Gazeformer-noDur | **0.504** | - | **0.534** | - | **2.061** | - | **1.742** | - | 0.849 | 0.559 | 8.356 |
| Gazeformer | **0.504** | **0.451** | 0.525 | **0.485** | 2.072 | **9.708** | 1.810 | **7.688** | **0.852** | **0.561** | **8.375** |

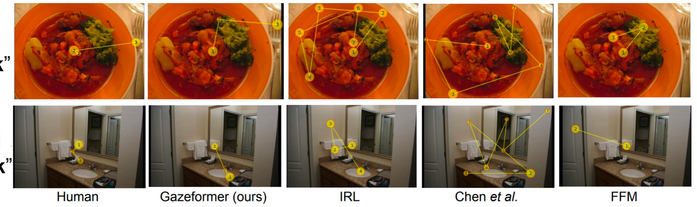
**"Effective"**: GazeTrain Setting

| | Time (in ms)↓ | Inferences/s ↑ | Speedup ↑ |
|---|---|---|---|
| Chen *et al.* | 386 | 2.59 | 1X |
| FFM | 133 | 7.52 | 2.9X |
| IRL | 85 | 11.77 | 4.5X |
| Gazeformer | **68** | **14.71** | **5.7X** |

**"Fast"**: Inference Time

find **"fork"**

find **"sink"**




***Qualitative Results*** *of Gazeformer and baselines for ZeroGaze Setting*


*Gazeformer also extends to **unknown and uncommon categories***

**Acknowledgements**: Supported by US NSF Awards IIS-1763981, IIS-2123920, NSDF DUE-2055406, the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe.