

Journal's Club: Halloween Edition

Anupam Pani



HKU Musketeers Foundation
Institute of Data Science
香港大學同心基金數據科學研究院



Follow Us:
linktr.ee/hkudatascience



Outline

Quick introduction about the work I will present

- Virtual staining of tissue images – related to histopathology
- Anticipating human actions using VLMs and eye gaze – for applications in assisted living environments for the patients who are senile or specially abled
- GazeSAM – can pathologists use eye gaze to help segment medical images
- **Objective** : Simple introduction about how we can apply some of the AI related projects in the medical world – and how we can take inspiration from the biological system to improve AI architectures

Virtual Staining of Tissue Images : Immunostaining

Using AI to help pathologists



HKU Musketeers Foundation
Institute of Data Science
香港大學同心基金數據科學研究院



Follow Us:
linktr.ee/hkudatascience



Introduction

Virtual Tissue Staining: A Novel Approach in Pathology

- Traditional staining methods are essential for disease diagnosis but are often time-consuming, costly, and can be environmentally harmful.
- Virtual tissue staining offers a faster, cost-effective alternative, supporting digital medicine and reducing hospital overcrowding.

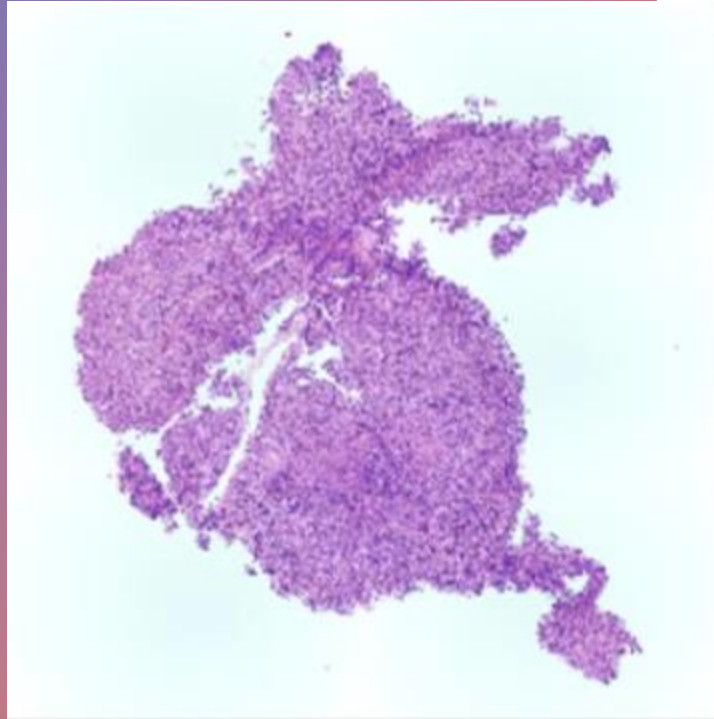
Opportunity and Need

"Why Virtual Staining?"

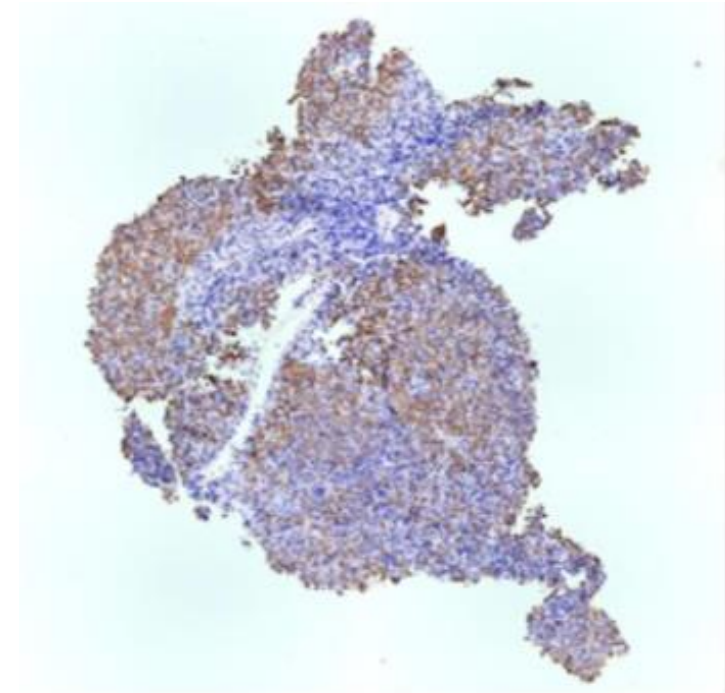
- Traditional methods involve chemical or antibody solutions and can contribute to workplace and environmental pollution.
- Virtual staining reduces pollutants, costs, and speeds up diagnostic processes, making it ideal for use in low-resourced hospitals and time-sensitive clinical scenarios

Example of Staining

Traditional Approach example



Brightfield H&E
images (hematoxylin and
eosin) unstained



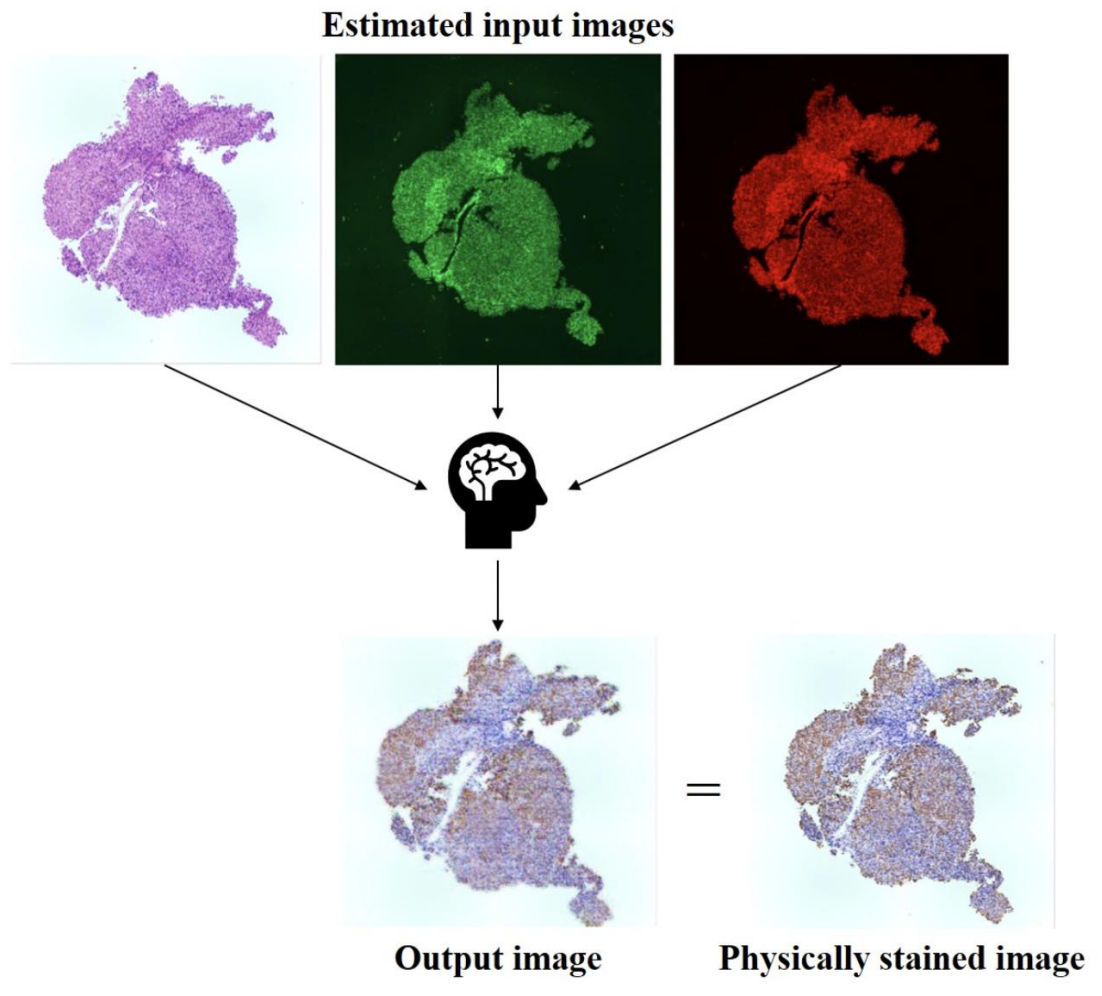
Stained output image

Method

Multi-Modal Image Synthesis for Virtual Staining

- Uses multiple input images (e.g., brightfield, GFP, RFP – green /red fluorescent proteins) to create a synthesized, virtually stained output image.

- Input images are often from unstained or lightly stained tissue sections, reducing the need for extensive chemical staining.



Multimodal input challenges

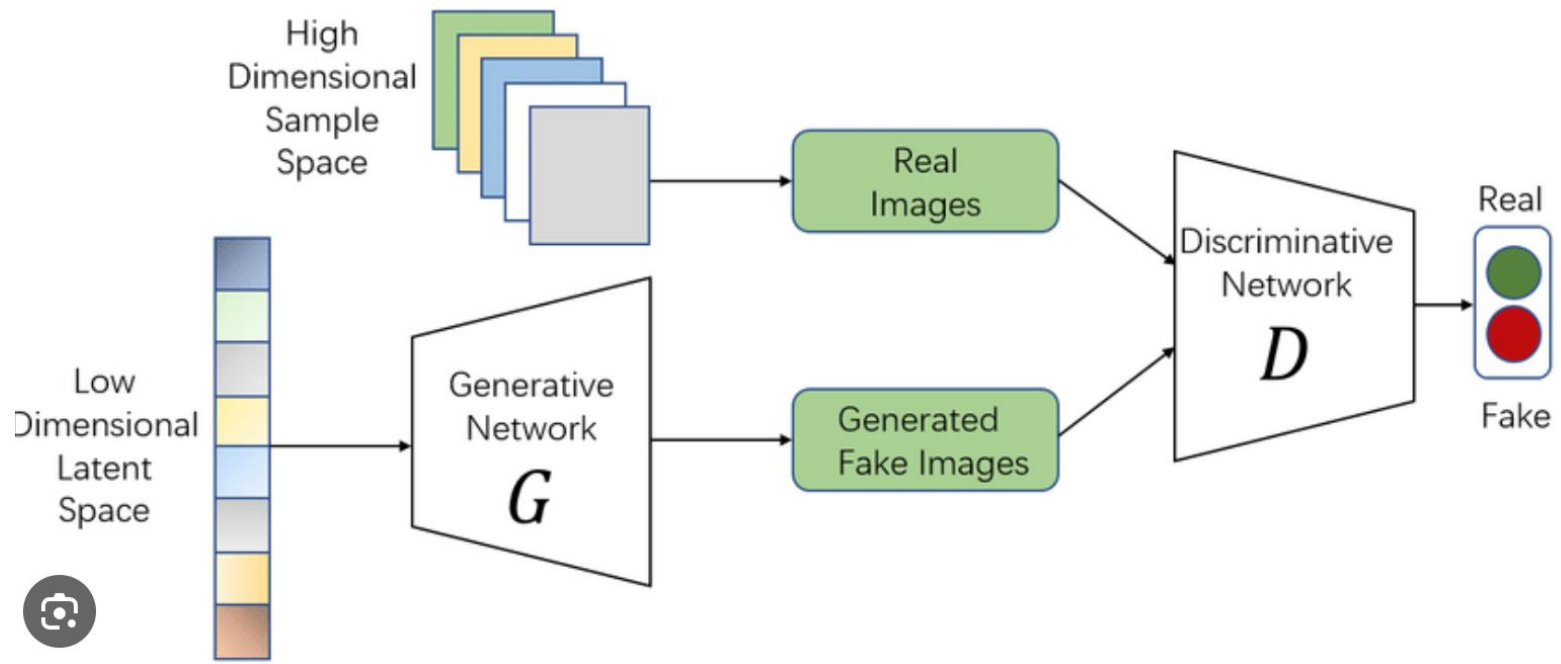
Addressing Image Blurring

- Challenge: Input and output images from different tissue sections can lead to blurring, making the output unsuitable for clinical use.
- Solution: Novel multi-modal image synthesis merges input images using a pixel-wise max procedure, extracting and combining the most important information
- Basically, selects the 'best' parts from each image to produce the final output which does not require

Use of GAN

Architecture Overview

- A Generative Adversarial Network (GAN) is trained to synthesize the final stained image by learning from multi-modal inputs.
- The GAN ensures high-quality output by reducing artifacts and maintaining image clarity, crucial for clinical diagnostics.
- It consists of two main parts:
 - Generator:** Tries to create images.
 - Discriminator:** Tries to figure out if the images are real or fake.
- They "compete" against each other, which helps the Generator get better at making realistic images.



GAN for staining

How is GAN used in our project

- The Generator creates a virtually stained image based on input images.
- The Discriminator checks if the generated image looks like a real, stained tissue image.
- Training Process:**
 - Over time, the Generator learns to make the virtual stain look just as real as the actual physical stain, without needing chemicals.
 - The Discriminator helps by pushing the Generator to improve its output until the images are almost indistinguishable.

Applications in pathology

How can this study/project help?

- Digital pathology: Supports faster diagnostics and reduces the need for extensive physical infrastructure.
- Veterinary pathology: Expected to become a standard method within 5–10 years.
- In-vitro diagnostics: With no commercial virtual staining software currently available, there is a large untapped market potential.

Conclusion

- Virtual tissue staining leverages advanced AI techniques, including GANs, to deliver fast, accurate, and cost-effective diagnostic solutions.
- This approach represents a significant step toward the future of digital medicine and diagnostic automation
- Patent granted May 2024 (US)

Gaze Regularized Attention for Human Action Prediction

Enhancing VLMs with eye gaze



HKU Musketeers Foundation
Institute of Data Science
香港大學同心基金數據科學研究院



Follow Us:
linktr.ee/hkudatascience



Introduction

Human action prediction

- Human action prediction is crucial in domains like **assistive robotics, autonomous driving, and accessibility**.
- Most existing models rely only on visual data from egocentric videos.
- **Problem:** Solely relying on visual data limits the model's ability to predict fine-grained actions accurately.
- **Proposal:** Eye gaze data provides critical insight into what a person is focusing on and can improve human action prediction accuracy.
- Can be used in assisted living environments and improve human-robot interaction (For example – a robotic nurse could be employed to look after the patients while the nurses are off duty or taking other responsibilities)

Motivation

Why incorporate eye gaze?

- **Human Attention Mechanism:** Eye gaze offers clues about intentions and upcoming actions by focusing on critical objects or regions.
- **Research Gap:** Previous approaches overlook the importance of integrating gaze data into models for action anticipation.
- **Key Insight:** Integrating gaze into models can improve predictions by guiding attention toward meaningful regions in the scene

Contributions

Key aspects of our study

- **Gaze-Augmented Framework:** Enhances Vision-Language Models (VLMs) by integrating gaze data directly into the architecture.
- **Gaze-Regularized Attention Mechanism:** Ensures the model focuses on gaze-highlighted areas.
- **Significant Performance Improvement:** Nearly 13% improvement in semantic score for action predictions when using gaze-regularized attention framework as compared to base model.
- **Extensive Experiments:** Evaluation on multiple gaze-augmented models and a comparison with baseline model coupled with ablation studies.

Background

Related work on action prediction

- Action prediction and activity forecasting have been well-studied in computer vision, with most models leveraging visual data only.
- **Ego4D Benchmark:** Previous work focuses on modeling human actions in egocentric videos but often lacks fine-grained predictions.
- **Attention Models:** Studies on attention-based models have improved feature selection, but gaze has typically been used as a **supervised signal** to predict attention, rather than a direct input.

Problem Statement

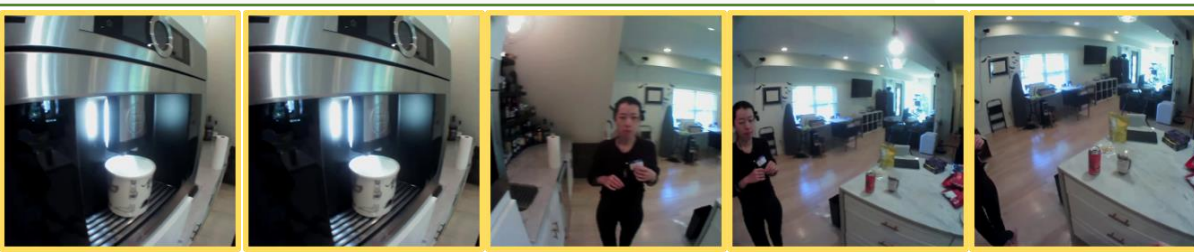
Define the problem setting

- Goal:** Predict fine-grained human actions from egocentric video frames, incorporating eye gaze to enhance the model's attention on relevant regions.
- Given past frames, the model must output descriptive text of future actions.
- Challenge:** How to effectively use gaze data as a guiding signal in Vision-Language Models.

Dataset and Annotations

Dataset and modifications made to the dataset

- **Dataset:** We use the Ego4D dataset, which contains egocentric videos with eye gaze data.
- Frames are collected every 1 second to reduce computational load.
- Annotations are enhanced using GPT-4V (and ShareCaptioner) and an **iterative prompt design** to provide detailed fine-grained annotations.
- ShareCaptioner provided more fine-grained captions as compared to GPT-4V



Goal: To get annotations for the image sequence

User Prompt

Describe what is happening in the image sequence and output the text descriptions



Language Model

Language model prompt

Given the **user feedback** and initial **user prompt**, suggest a new prompt which would satisfy the user requirements

User Feedback

There is only one annotation for the whole sequence. We want each image to have an annotation and to describe the actions undertaken in the sequence

Suggested Prompt

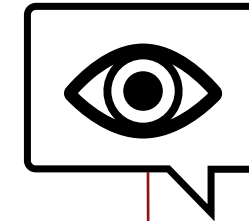
For each image in the sequence, provide an annotation which describes the action taking place

Human evaluation



Not satisfied with annotations

GPT 4V



Annotations

The process of brewing a cup of coffee is taking place in the image sequence. A person can be seen talking to the camera wearer



User Prompt

Describe what is happening in the image sequence and output the text descriptions

Goal: To describe the changes between two images in the sliding window and provide annotations

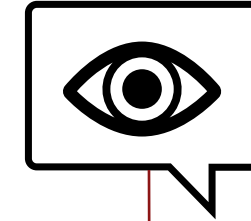


Language Model

Suggested Prompt

For each image in the sequence, provide an annotation which describes the action taking place

ShareCaptioner



Language model prompt

Given the user feedback and initial user prompt, suggest a new prompt which would satisfy the user requirements

Human evaluation



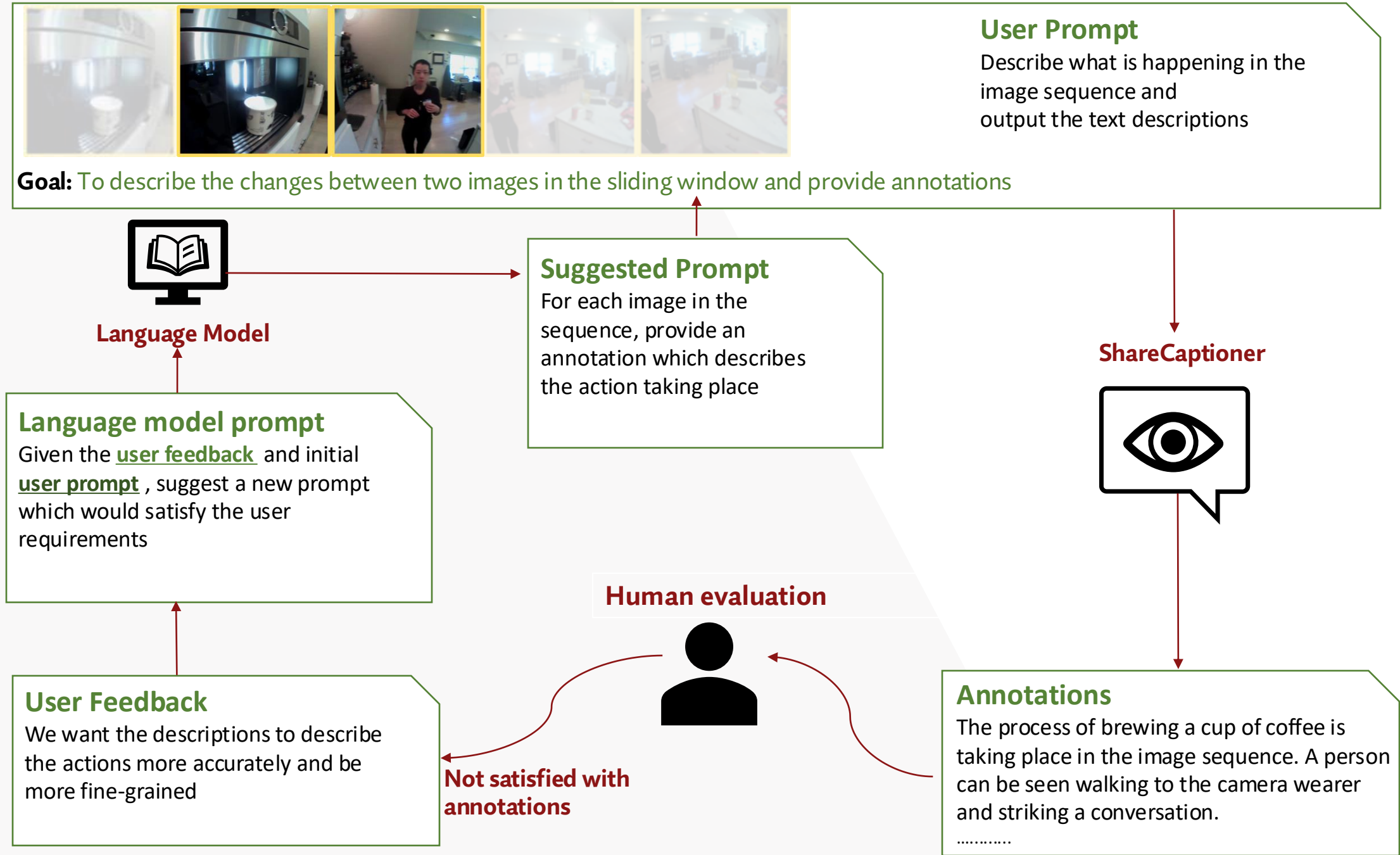
Annotations

The process of brewing a cup of coffee is taking place in the image sequence. A person can be seen walking to the camera wearer and striking a conversation.
.....

User Feedback

We want the descriptions to describe the actions more accurately and be more fine-grained

Not satisfied with annotations

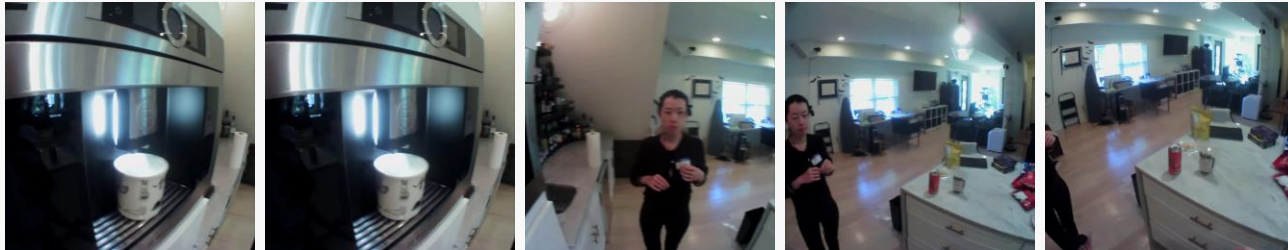


Models

Different models based on mechanism and input

- **Base Model:** relies only on RGB image frames obtained from the video clip , utilizes an open-flamingo model

Ego-centric video clip



Base Model

- A person's hands are **holding a bowl and drink**.
- An individual's hands **is grabbing drink** from the table.

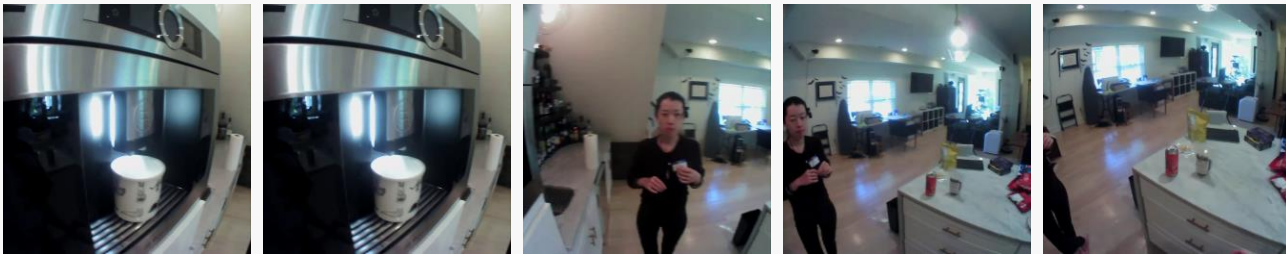
Predicted actions

Models

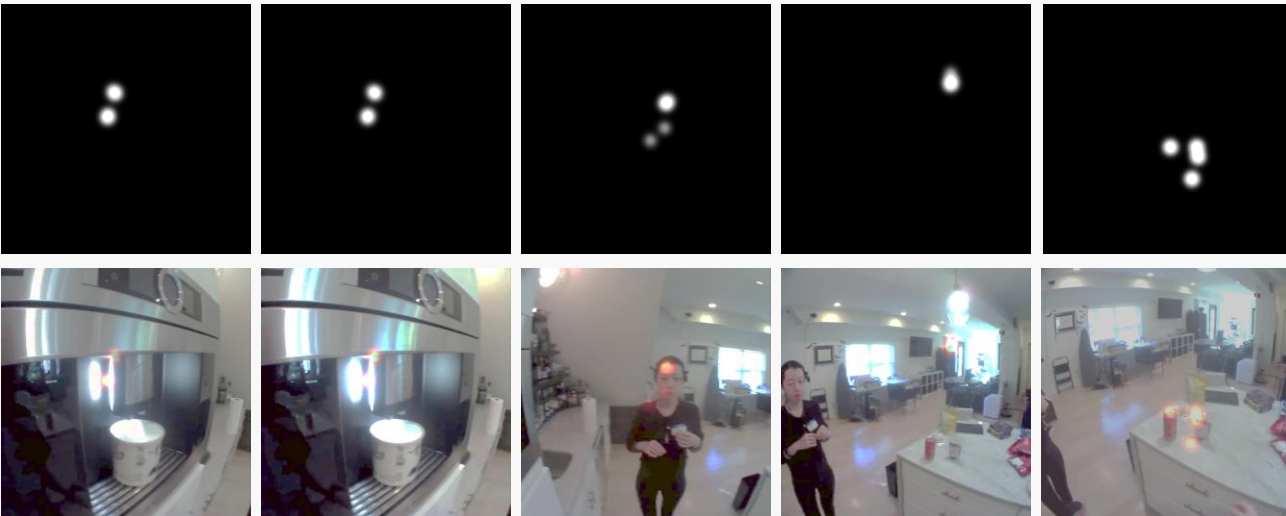
Different models based on mechanism and input

- **Gaze-regularized model:** input is RGB image frames along with gaze data from gaze heatmaps and gaze overlaid images and uses a gaze-regularized attention mechanism

Ego-centric video clip



Gaze heatmaps (top) and gaze-overlaid images (bottom)

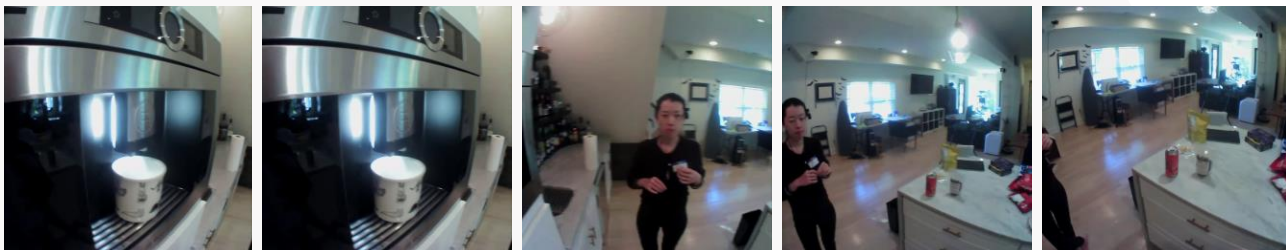


Predicted actions

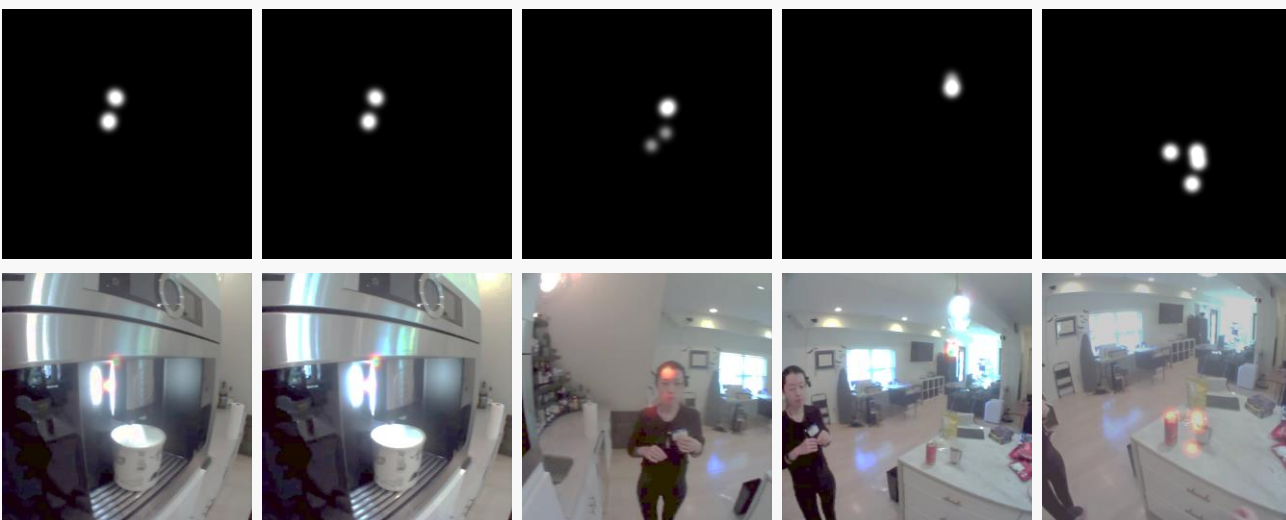
**Gaze
Augmented
Model**

- A hand is looking for a pack of **chocolate or similar snacks** on the **kitchen island**.
- Individual is **grabbing the chocolate snack** placed on the kitchen island counter **near the drink**.

Ego-centric video clip



Gaze heatmaps (top) and gaze-overlaid images (bottom)



Base Model

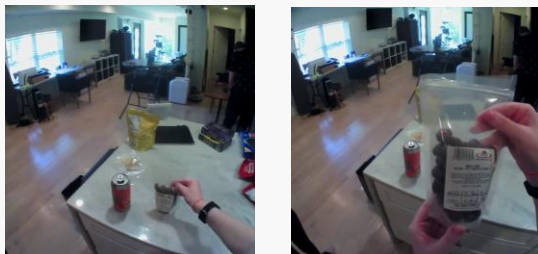
- A person's hands are **holding a bowl and drink**.
- An individual's hands **is grabbing drink** from the the table.

Predicted actions

Gaze Augmented Model

- A hand is looking for a pack of **chocolate or similar snacks** on the **kitchen island**.
- Individual is **grabbing the chocolate snack** placed on the kitchen island counter **near the drink**.

Future frames and ground-truth text annotations (for reference)



- An individual's hand is reaching towards a pack of **chocolate coated nuts on a kitchen island counter** while a drink and a black screen can be seen behind.
- An Individual has **picked up the chocolate nuts** and is looking at the bar code scanner which is on the label of the packet while the **drink is on the table**.

Proposed Method

Overview of the gaze-regularized attention framework

- **Gaze Heatmaps:** We use gaze heatmaps generated from gaze points to direct attention to important areas.
- **Gaze-Regularized Attention Mechanism:** Ensures model attention distribution aligns with human gaze distribution using Kullback-Leibler divergence. Gaze based queries are used in the attention block.
- **Prediction Pipeline:**
 - Extract features from RGB frames and gaze-overlaid frames.
 - Apply attention mechanism that highlights gaze-allocated regions.
 - Align model attention with human gaze attention.
 - Predict fine-grained human actions in textual form.

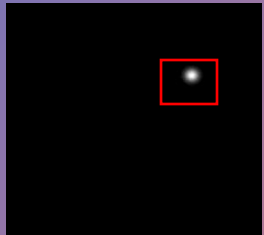
Architecture Overview

Framework for the gaze-regularized attention model

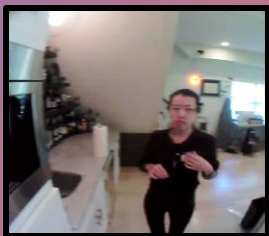
- **Input:** RGB image frames, gaze overlaid images and gaze-heatmap images
- **ViT Encoder:** Extracts features from RGB image frames and gaze-overlaid images.
- **Gaze-Regularized Attention block:** Processes both image features and gaze-overlaid features, focuses on regions indicated by gaze patterns to produce gaze-enhanced features.
- **Perceiver Resampler:** Generates a fixed-size representation from these features for the language module.
- **Language Module:** Predicts future actions based on the processed features.
- **Gaze Regularizer:** Aligns model attention with human gaze attention by minimizing Kullback–Leibler divergence between the attention distribution and gaze distribution.



RGB scene image



Binary Heatmap Image

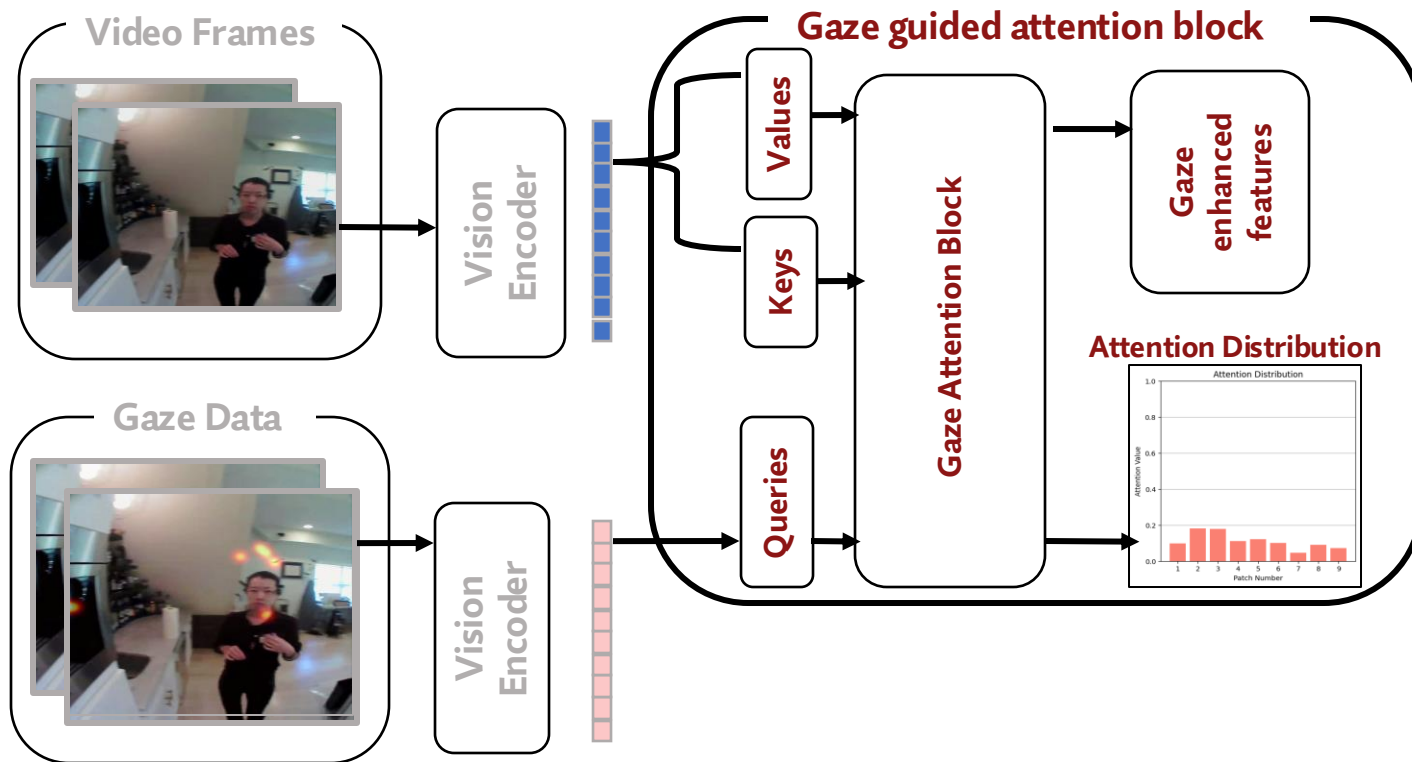


Gaze Overlaid Image

Gaze guided attention block

Use of gaze-based queries to obtain gaze enhanced features

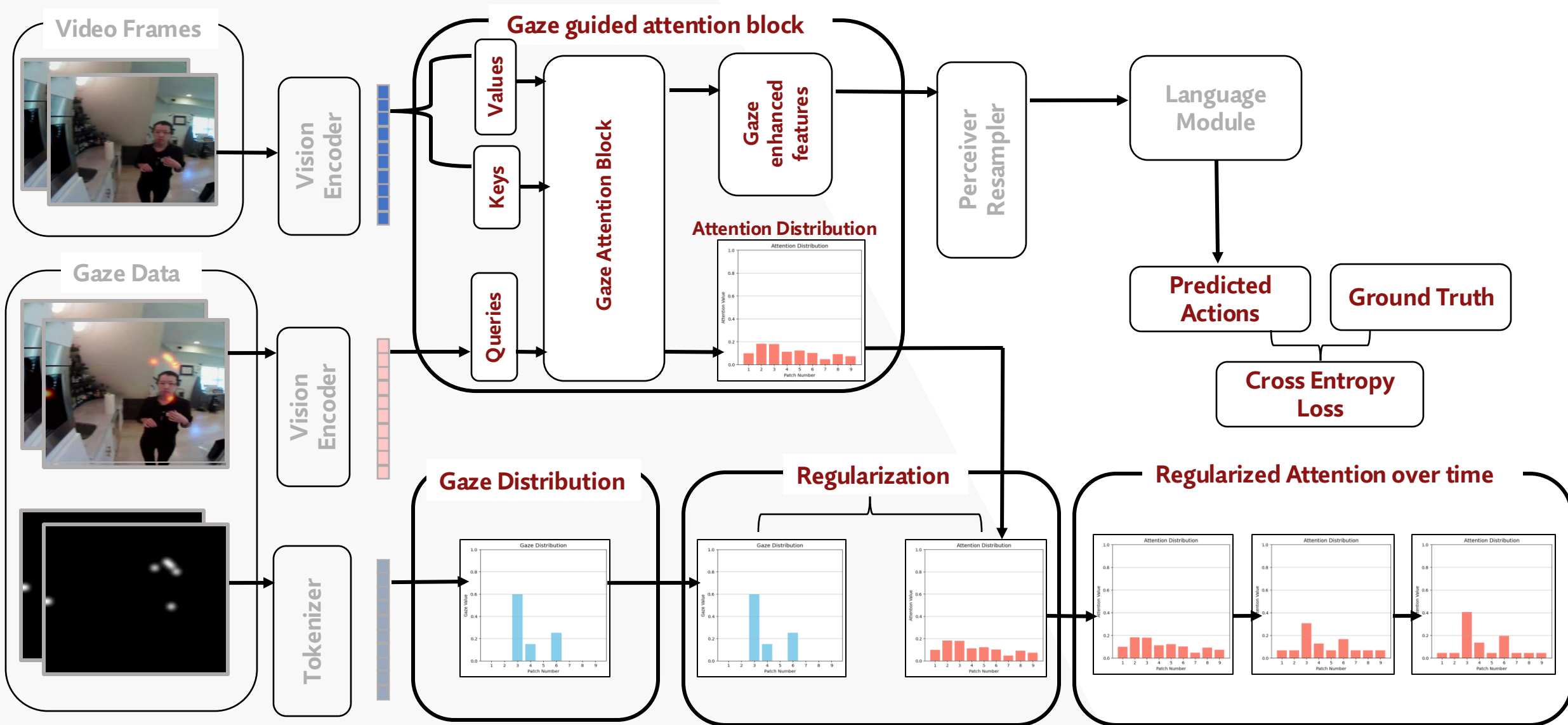
- **Queries, Keys and Values:** Gaze-overlaid features are used as queries, Image features are used as Keys and Values



Gaze Regularization Mechanism

Aligning attention with human gaze

- The gaze regularizer minimizes the difference between the model's attention distribution and the gaze distribution.
- The gaze distribution is obtained from gaze heatmap images. Gaze heatmap images are black and white images, where white regions are the gaze occupied regions.
- Helps ensure the model prioritizes gaze-highlighted areas due to 'soft-alignment' nature of attention mechanism.
- Over time, the model attention gets more and more aligned with the human gaze distribution.



Gaze equipped Models

Variations in the gaze equipped models

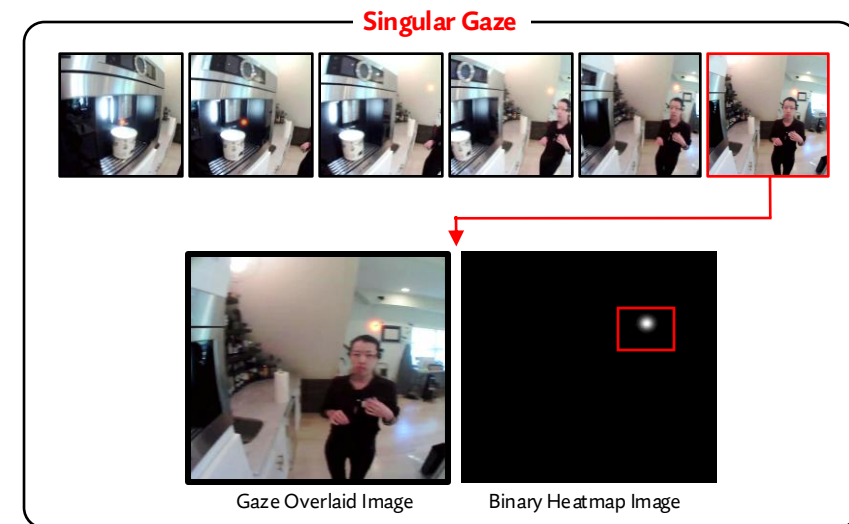
- Motivation:** A single gaze point can be noisy or can be a result of a tremor/ micro-saccade. Furthermore, information is usually collected during a fixation which lasts for a longer timeframe.
- Proposal:** Collect an aggregated gaze pattern over a time interval (200 ms) to account for noisy variations and to ensure a sufficient time frame for collecting detailed information

Gaze equipped Models

Variations in the gaze equipped models

- Motivation:** A single gaze point can be noisy or can be a result of a tremor/ micro-saccade. Furthermore, information is usually collected during a fixation which lasts for a longer timeframe.

- Proposal:** Collect an aggregated gaze pattern over a time interval (200 ms) to account for noisy variations and to ensure a sufficient time frame for collecting detailed information

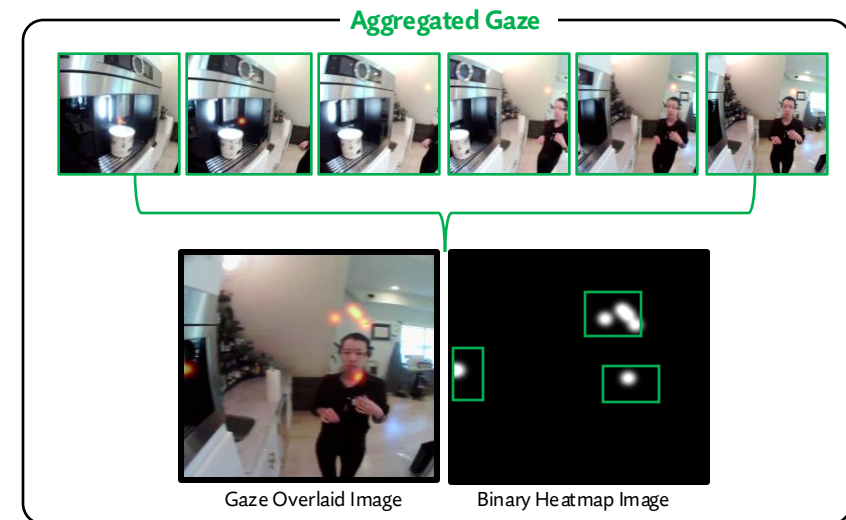


Gaze equipped Models

Variations in the gaze equipped models

•**Motivation:** A single gaze point can be noisy or can be a result of a tremor/ micro-saccade. Furthermore, information is usually collected during a fixation which lasts for a longer timeframe.

•**Proposal:** Collect an aggregated gaze pattern over a time interval (200 ms) to account for noisy variations and to ensure a sufficient time frame for collecting detailed information

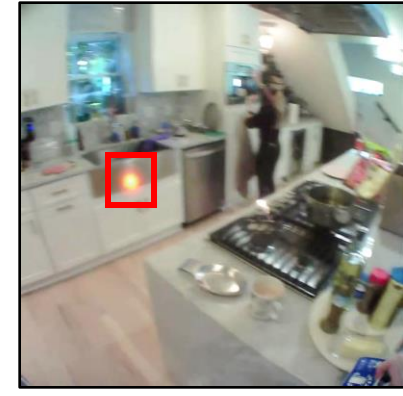
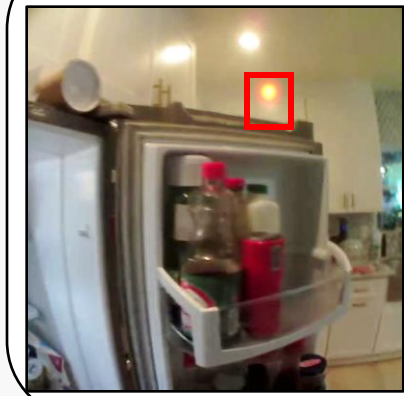


Occlusion due to aggregation of frames

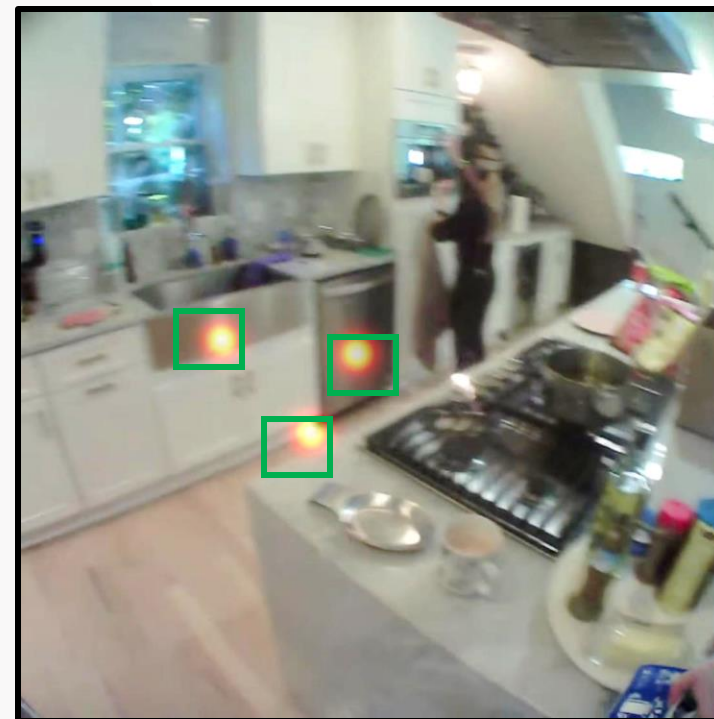
To ensure correct aggregation of gaze points

- **Consistency check** : Utilize forward and backward optical flow to check if there is major occlusion between previous frames in the time interval and frame utilized by the model.
- If there is major occlusion, do not collect the gaze point for heatmap formation.
- If occlusion is minor, collect the gaze point for subsequent heatmap formation.

Frames in the time interval $[t - 0.2, t]$



Aggregation **without occlusion check**



Aggregation **with occlusion check**

Experiments

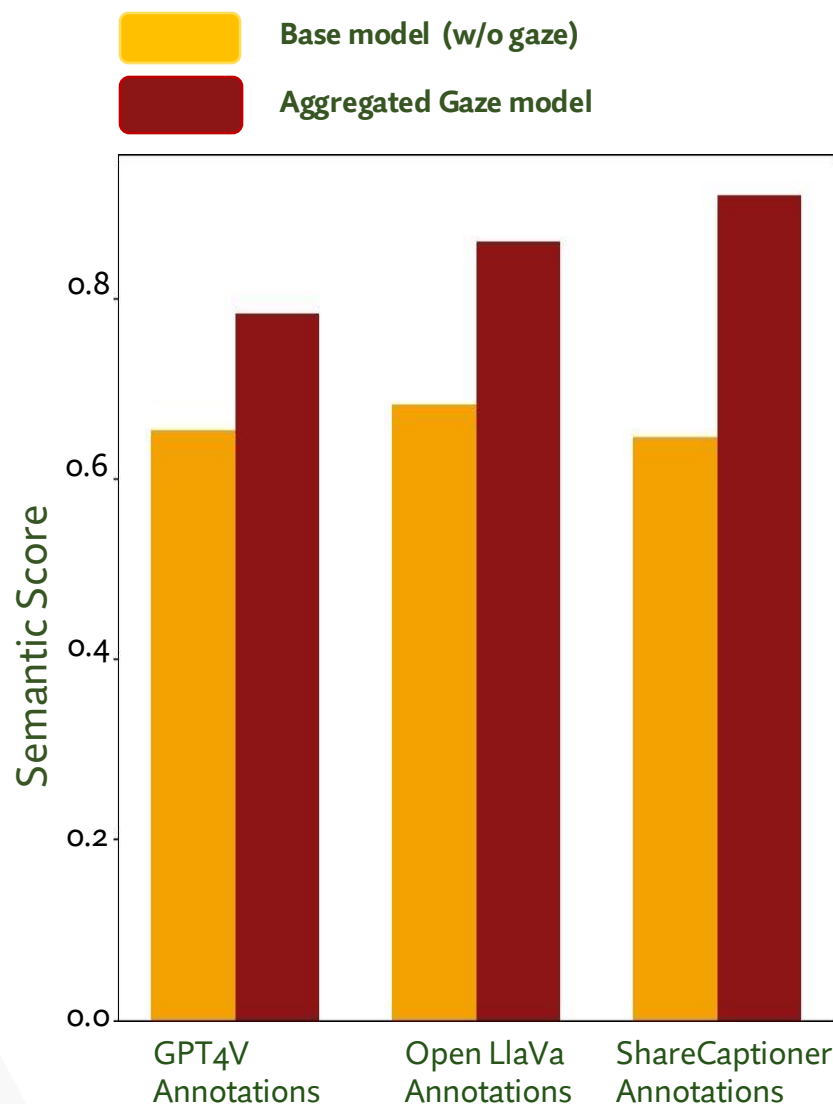
Questions we want to tackle in our study

- Is inclusion of gaze necessary?
- Does annotation quality impact performance?
- What is the effect of regularization on gaze-augmented models?
- How many gaze-guided attention blocks are needed?

Base Model vs. Gaze-Augmented Models

To highlight the importance of eye-gaze as a signal

| Model | Annotation source | Semantic Score(↑) | Meteor Score(↑) | Rouge-L (↑) | | |
|------------------------|-------------------|-------------------|-----------------|---------------|---------------|---------------|
| | | | | Precision | Recall | F-score |
| Base (no gaze) | GPT-4V | 0.6525 | 0.4075 | 0.4335 | 0.4301 | 0.4318 |
| Singular Gaze | GPT-4V | 0.7316 | 0.4501 | 0.4822 | 0.5309 | 0.5054 |
| Aggregated-Gaze | GPT-4V | 0.7826 | 0.5033 | 0.5193 | 0.5644 | 0.5405 |
| Base (no gaze) | ShareCaptioner | 0.6437 | 0.5060 | 0.5730 | 0.5566 | 0.5646 |
| Singular Gaze | ShareCaptioner | 0.8212 | 0.6514 | 0.6906 | 0.6914 | 0.6905 |
| Aggregated-Gaze | ShareCaptioner | 0.9125 | 0.7114 | 0.7617 | 0.7717 | 0.7666 |



Effect of Regularization

Performance when gaze-regularization coefficient is changed

- When no regularization is used in the gaze-regularization attention mechanism, there is a massive drop in performance
- Performance is the highest when a moderate coefficient is utilized
- There is a slight drop in performance when the coefficient is large

| Regularization coefficient | Semantic Score(↑) | Meteor Score(↑) | Rouge-L (↑) | | |
|-------------------------------|----------------------|--------------------|---------------|---------------|---------------|
| | | | Precision | Recall | F-score |
| 0 | 0.6317 | 0.4094 | 0.3872 | 0.3622 | 0.3738 |
| 100 | 0.7826 | 0.5033 | 0.5193 | 0.5644 | 0.5405 |
| 1000 | 0.7798 | 0.4963 | 0.5127 | 0.5558 | 0.5330 |

Number of attention blocks

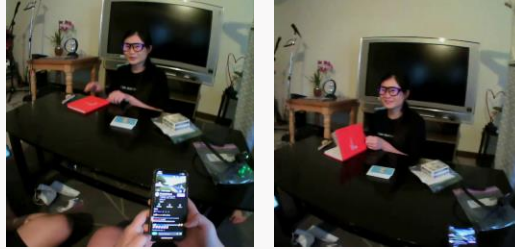
Performance with varying number of attention blocks

- Performance is the highest when 2 gaze-guided attention blocks are utilized (and when regularization coefficient is set to 100)
- This highlights the importance of combining the gaze-guided attention block, along with the the gaze-regularizer and utilizing them together.

| Attention blocks | Semantic Score(↑) | Meteor Score(↑) | Rouge-L (↑) | | |
|------------------|-------------------|-----------------|---------------|---------------|---------------|
| | | | Precision | Recall | F-score |
| 1 | 0.7434 | 0.4945 | 0.5065 | 0.5630 | 0.5328 |
| 2 | 0.7826 | 0.5033 | 0.5193 | 0.5644 | 0.5405 |
| 5 | 0.7765 | 0.5013 | 0.5098 | 0.5528 | 0.5301 |

Comparison of output predictions

Future Frames



Prediction duration: 2 seconds

Ground Truth

- The camera wearer is **in conversation** with the **seated person** in front with **a deck of cards** and **a mobile phone** screen visible in front of them. The individual is opening the **red book** in front of her, possibly after being asked by the camera-wearer.

Base Model

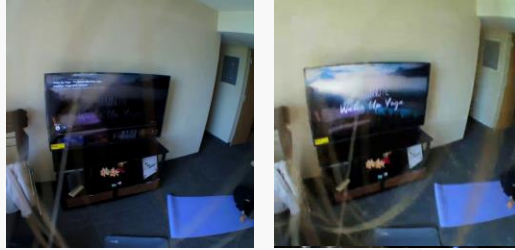
- The individual has a **TV remote** with a game visible on the table and **individual in front**. The **seated person** is in the background in front of the card game with **a book**.

Our Model

- The camera wearer is **talking** to the **seated person** with the **mobile screen** visible. The seated person is **reading a red book** which is near a **deck of cards**

Comparison of output predictions

Future Frames



Prediction duration: 2 seconds

Ground Truth

- The camera wearer continues **to look at the television screen**, seemingly watching a yoga video with a person visible in the background sitting on a **blue yoga mat**. The individual on the mat is **starting to perform** yoga **exercises** while **watching** the screen.

Base Model

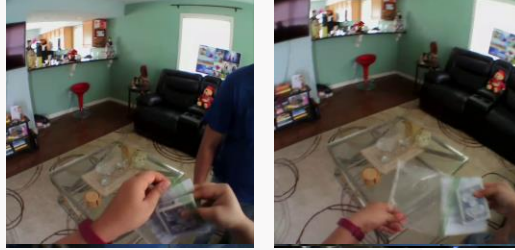
- The camera wearer is **looking at a game** with the **TV screen** visible and individual in the seat. The **individual person is seated** and the screen is visible in the background

Our Model

- The camera wearer is **looking** to at the **TV screen** and individual is sitting on a **blue mat**. The individual is **exercising on blue mat** and continuing to watch the game.

Comparison of output predictions

Future Frames



Prediction duration: 2 seconds

Ground Truth

- The camera wearer is **opening a plastic packet** with a paper and **a deck of cards** inside as a person in a **blue shirt** moves in the **living room**. The individual is **removing a deck of cards from a packet**

Base Model

- The individual is **opening a paper** in the **living room** with motion in the background. The individual is **removing a paper inside a plastic**

Our Model

- The individual is **removing a deck of cards** as **blue shirt** is moving in the **living room**. The camera wearer is **removing a deck of cards in plastic**.

Conclusion

Summarizing what we have uncovered during our study

- Gaze-regularized attention improves fine-grained human action prediction if the gaze guided attention block is utilized in synchronization with the gaze regularizer.
- Aggregation of gaze points provides better performance as opposed to when a single gaze point is used
- Improvement is seen in output predictions for gaze augmented models when finer-grained annotations are used for training.
- Quality of output predictions decreases for models without gaze when finer-grained annotations are used.
- Gaze data can significantly enhance model focus on critical areas.

Future Work

Where do we go from our current work

- Several more datasets have appeared containing egocentric videos and eye gaze data. Preparing a large-scale dataset can further improve our work.
- Annotations were obtained using VLMs. With further advancements in VLMs, we will obtain higher quality annotations.
- Exploration using other modalities – humans use auditory signals as well as tactile signals. Is there a way to integrate those signals to enhance VLMs?
- Eye gaze is closely related to the human visual system. Can we take inspiration from the biological mechanism and further improve the architecture?

GazeSAM

Using eye gaze to help segment
medical images



HKU Musketeers Foundation
Institute of Data Science
香港大學同心基金數據科學研究院



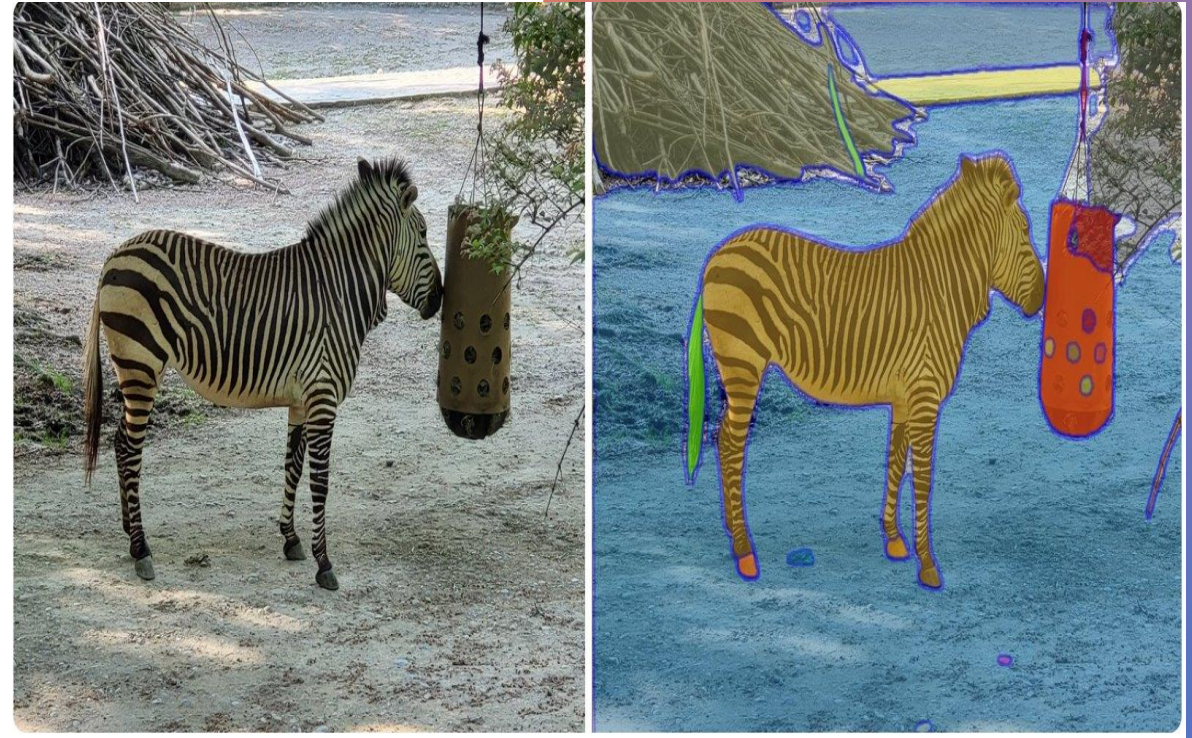
Follow Us:
linktr.ee/hkudatascience



Introduction

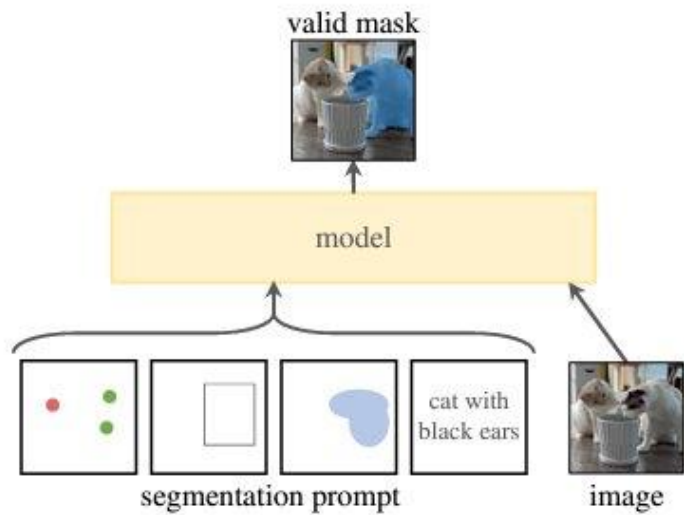
What is SAM?

SAM – segment anything model

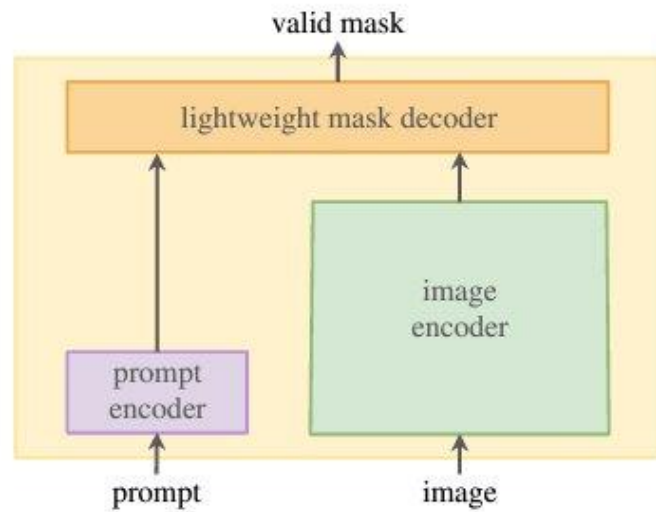


Input prompts for SAM

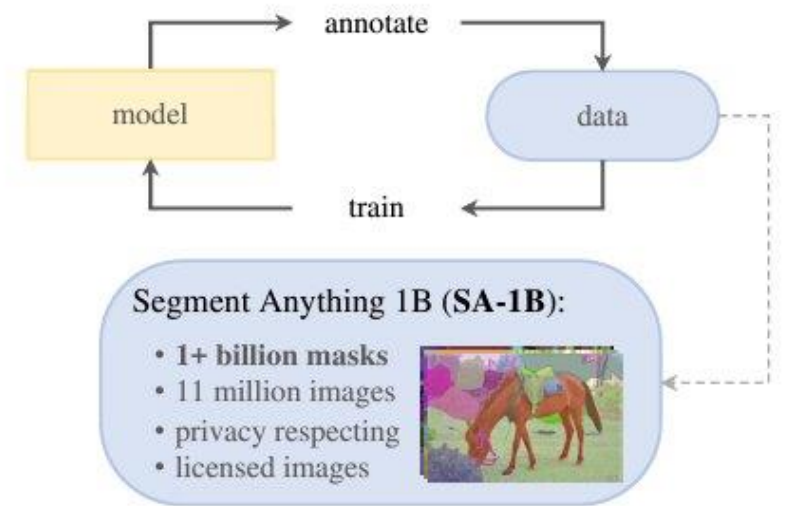
Normal approach



(a) **Task:** promptable segmentation



(b) **Model:** Segment Anything Model (SAM)



(c) **Data:** data engine (top) & dataset (bottom)

Segmentation and medicine

Why do we need segmentation in medical world?

- Image segmentation has many medical applications – disease diagnosis , treatment planning and surgical navigation.
- Segmentation helps with finding ROIs (region of interest)
- Can help radiologists – pathologists with diagnosis
- Manual segmentation is costly, time-consuming and can take days

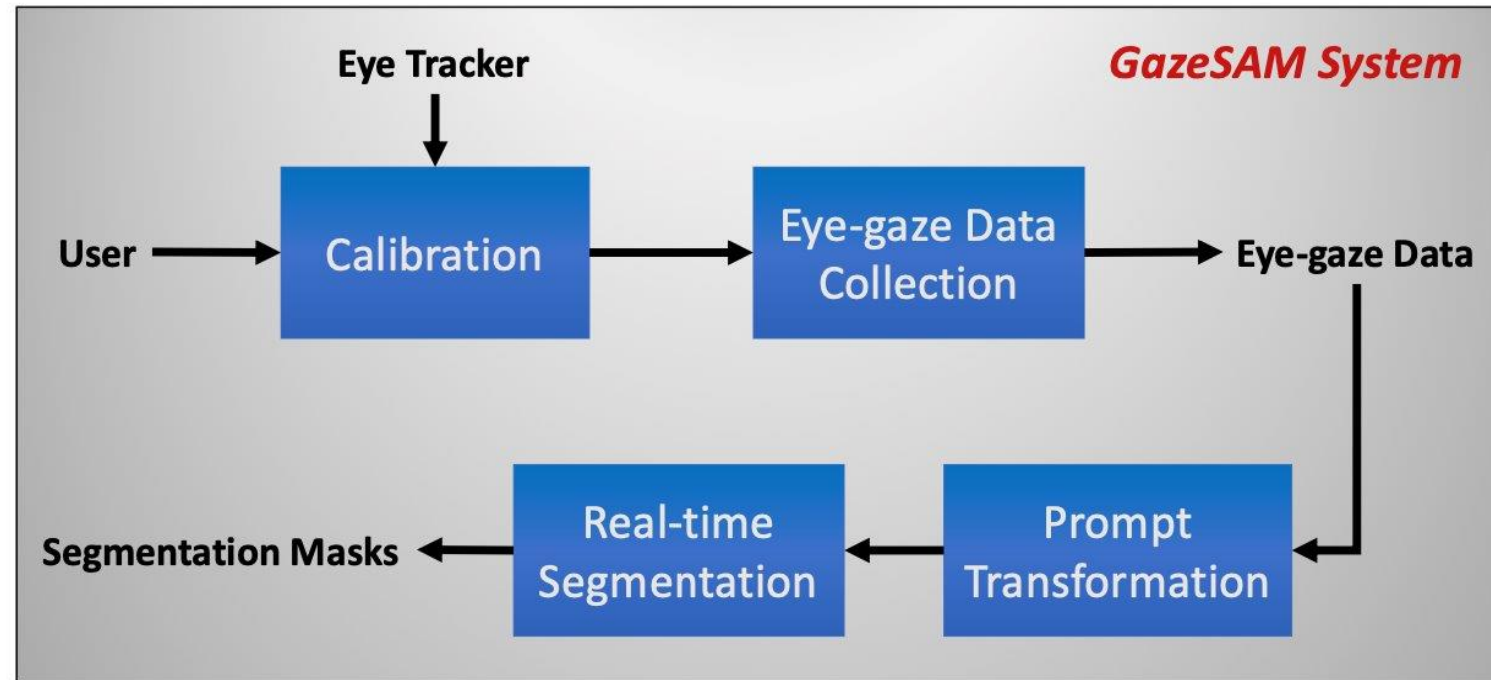
Motivation

Motivating factors behind GazeSAM

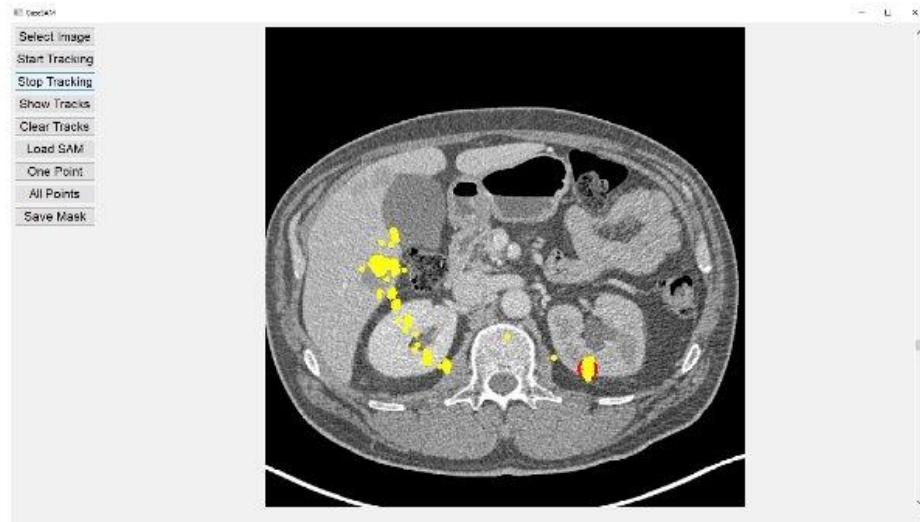
- SAM trained on large scale dataset (natural images)
- Ability to infer on medical images is limited
- Collaborative framework by using other modalities especially for medical images – has not been explored

Method

Overview of the system



GazeSAM interface



(a) Option 1 (All Points)



(b) Option 2 (One Point)

Contributions

- Using eye-tracking technology with SAM for automated medical image segmentation
- **Real-Time Segmentation** with SAM, generating segmentation masks.
- **User-Friendly & Fast:** Automatically updates segmentation based on eye movement.

- **Maintaining Focused Attention:**
- Directly using a noisier, Gaussian-like gaze distribution could dilute the model's focus. In contrast, gaze-regularized attention aligns attention distribution with gaze patterns without sacrificing focus, allowing the model to selectively attend to relevant regions with precision.
- This approach maintains the discriminative power of gaze data, avoiding the broad, potentially uninformative spread that noisy distributions may introduce.
- **Selective Responsiveness to Gaze Patterns:**
- By using gaze regularization, the model learns to attend to gaze-relevant regions as a preference rather than a strict rule. This prevents the model from over-relying on the gaze distribution, which can become problematic if the gaze data has significant noise or slight errors. Regularization encourages an alignment that is more robust and adaptable than directly incorporating noisy gaze.
- **Noise Reduction Without Attention Disruption:**
- Noisier distributions, even when Gaussian, can spread attention to irrelevant regions, causing the model to attend to areas that are not crucial for the prediction task. Regularization, however, reduces attention to noise in gaze data by setting a controlled alignment with gaze points, limiting the risk of capturing irrelevant regions influenced by noise.
- **Adaptability Across Tasks:**
- In tasks requiring both broad context and specific interactions, gaze-regularized attention can balance between gaze-prioritized regions and the overall image. Directly using a noisy distribution may cause misalignment in contexts where precise, focused attention is necessary. Regularization allows for an adaptable mechanism that can fit more diverse scenarios, especially where gaze data alone might miss relevant areas outside its direct focus.