VLMS WITH GAZE-REGULARIZED ATTENTION FOR HUMAN ACTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Eye gaze, encompassing fixations and saccades, offers valuable insights into human intentions and future actions. This study presents a novel approach to enhancing Vision Language Models (VLMs) for human action prediction by integrating eye gaze data into egocentric video analysis. Existing methods for action prediction in egocentric videos often rely solely on visual data, potentially missing critical information provided by eye gaze. To address this limitation, we propose a unique gaze-augmented framework that integrates eye gaze directly into the VLM architecture and training process. By generating gaze heatmaps from eye gaze coordinates, our model dynamically focuses on regions highlighted by gaze patterns. Additionally, a gaze-regularization mechanism ensures the model maintains attention on gaze-allocated areas, thereby improving prediction accuracy and robustness. Our approach significantly enhances the model's ability to generate precise and detailed predictions of future actions. Compared to baseline models without leveraging gaze data, our method achieves a nearly 13% improvement in the semantic score of predictions. This substantial improvement underscores the effectiveness and novelty of integrating eye gaze with a gaze-regularized attention mechanism in VLMs for action prediction. Moreover, our work demonstrates that incorporating eye gaze through this gaze-augmented framework can significantly boost the predictive capabilities of VLMs, enhancing their potential in applications that require accurate human action prediction.

1 INTRODUCTION

VLMs are foundation models that combine computer vision and natural language processing to understand and generate both visual and textual information. Examples of such models include ViLBERT, LXMERT, and CLIP (Lu et al., 2019a; Tan & Bansal, 2019; Radford et al., 2021). These models can describe images, answer questions about visual content, and even generate images from textual descriptions. When oriented to predict future events and actions, VLMs have the potential to facilitate human-machine interaction in various applications like assistive robots (Li et al., 2024), accessibility for visually impaired individuals (Zhao et al., 2024b), and autonomous driving (Zhou et al., 2024), contributing to safer and more inclusive environments.

To achieve the full potential, we propose that the key is to endow VLMs with the capability for fine-grained action prediction, which provides more actionable information to facilitate human-machine interaction. While coarse predictions might identify general activities like "brewing coffee," fine-grained predictions specify detailed steps such as "reaching for the coffee capsule in the top-right cabinet, then filling the tank with the cup on the rack." These detailed predictions provide useful insights that allow machines to assist more effectively (Goyal & Durrett, 2021). However, achieving accurate fine-grained predictions requires understanding the human agent's short-term goals, which, we propose, can be inferred from eye gaze patterns. On one hand, eye gaze reveals which objects the human is focusing on or intends to interact with, making it a valuable component for enhancing action prediction (Frischen et al., 2007),(Tipper, 2010). Moreover, humans often scan their surroundings to collect information necessary to accomplish short-term goals before executing an action. Thus, incorporating eye gaze into VLMs for action prediction provides insights into a person's intentions, and shall lead to more accurate and reliable predictions of human activities.

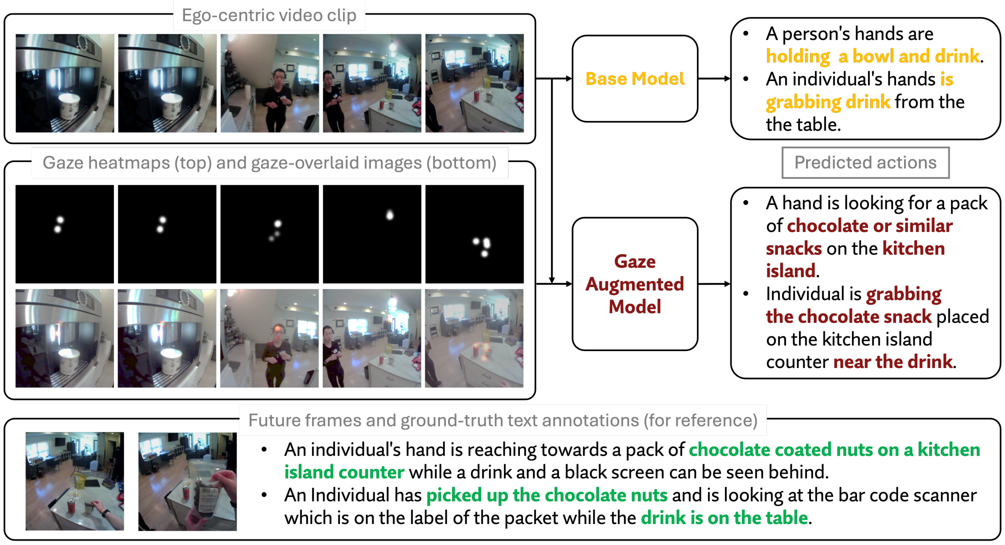

Ego-centric video clip
Gaze heatmaps (top) and gaze-overlaid images (bottom)
Base Model
Gaze Augmented Model
A person's hands are holding a bowl and drink.
An individual's hands is grabbing drink from the the table.
Predicted actions
A hand is looking for a pack of chocolate or similar snacks on the 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.

Figure 1: Illustration of human action prediction from egocentric video clips. The input consists of a sequence of image frames, with the output being predicted future actions in textual form. The base model predicts the presence of a bowl and selects the incorrect object, whereas the proposed model, which incorporates gaze data, correctly predicts the object that is about to be picked up. Predicted future actions (in text) are displayed on the right, while ground-truth annotations and immediate future frames are provided at the bottom for reference.

In this study, we propose a novel gaze-augmented framework that significantly enhances fine-grained action prediction by integrating eye gaze information into the VLM architecture and training process. Our approach involves generating gaze heatmaps from eye gaze coordinates, which dynamically guides the model's focus to regions of interest highlighted by gaze patterns. Specifically, the gaze-augmented framework employs a sophisticated attention mechanism where features derived from gaze-overlay images serve as queries in a Vision Transformer (ViT) (Dosovitskiy et al., 2021), while RGB image features act as keys and values. This method ensures that attention is calculated specifically to emphasize gaze-highlighted regions, which enhances the relevance of the extracted features. Furthermore, we introduce an explicit gaze-regularization mechanism that enforces the model to consistently allocate attention to gaze-concentrated areas, thereby promoting the predictive capability of the model.

Our experimental results show that augmenting the VLM architecture with gaze information greatly improves its performance, with nearly a 13% increase in the semantic score of predictions compared to baseline models without gaze data. This improvement highlights the effectiveness of using eye gaze with a regularized attention mechanism for detailed action prediction. By experimenting with both singular and aggregated gaze-augmented models, we validate the benefits of discounting noisy and irrelevant eye gaze points, which enhances robustness in highly dynamic environments. Further experiments with more detailed annotations show significant improvement in the performance of gaze-augmented models (by nearly 12%), whereas the base model's performance remains unchanged. This indicates that finer-grained annotations can help strengthen and unleash the proposed model's potential to leverage gaze data for better action prediction. In summary, our contributions are 1) A novel gaze-augmented VLM for fine-grained human action prediction; 2) A unique gaze-regularized attention mechanism that guarantees effective training of the gaze-augmented VLMs; 3) An efficient gaze sanity check to ensure the relevance of the gaze information; and 4) An extensive study that demonstrates the effectiveness of the proposed framework and validates the significance of gaze information for human action prediction.

2 Related Work

Action prediction and activity forecasting tasks Action prediction and activity forecasting tasks both aim to predict future human actions. While action prediction tasks focus on predicting immediate actions given an input image sequence or video clip, activity forecasting emphasizes predicting broader human activities over a longer time frame (Gao et al., 2017). However, both the tasks share the same goal of understanding human behaviour and making reasonable predictions. These tasks have been extensively studied, particularly through the challenges established by the Ego4D benchmark (Grauman et al., 2022). Previous approaches have leveraged Long Short-Term Memory (LSTM) networks and modality attention mechanisms (Furnari & Farinella, 2019) to predict future actions. Other studies have combined bottom-up and top-down approaches to infer latent goals and predict actions by modeling temporal dynamics (Zhao et al., 2024a). Mascaro et al. (2024) decomposed the action prediction problem into predicting low-level actions derived from high-level intentions, which showed an improvement in prediction performance. Similarly,Ashutosh et al. (2023) presented a hierarchical approach for action anticipation by modeling both short-term low-level actions ("what the person is doing right now") and high-level long-term intentions ("what the person wants to do") using contrastive learning. Cho et al. (2024) highlighted the significance of interactions between the next active object and human hands for short-term anticipation, predicting the next active object first to model future interactions. Transformers and their variants have also been employed to model interactions between objects and hands, enhancing the prediction of future actions (Roy et al., 2024). In our work, we aim to predict future human actions in the form of fine-grained and descriptive annotations by incorporating eye gaze as an additional modality and employing a gaze-regularized attention mechanism.

Attention and gaze-augmented models Attention-based models facilitate the discovery of important features and enhance performance on downstream tasks, and have been applied to various domains, including autonomous driving (Braunagel et al., 2017) , action prediction and human-computer interaction (Weber et al., 2020; Shafti et al., 2019; Aronson et al., 2018). For example - class activation maps have been used to leverage pooling layers in deep learning networks to generate class saliency maps (Sudhakaran & Lanz, 2018), allowing the model to focus on regions containing objects that correlate with the activity being considered for action anticipation and prediction. Guided-attention mechanisms have also been employed to model the next active object and predict future interactions (Thakur et al., 2023). The Spatiotemporal Attention Module (STAM) (Lu et al., 2019b) uses eye gaze information as supervision to train a network that predicts an attention map for activity recognition. Other studies similarly also use gaze as supervision to form attention maps for activity recognition tasks (Min & Corso, 2020; Awale & Sarikaya, 2022).

A lot of the previous work has utilized eye gaze data as ground truth to train attention prediction networks and to predict eye gaze. However, in our study, we use eye gaze data as a direct signal for human action prediction, building a gaze-regularized attention mechanism. With advancements in eye-tracking methods and devices, and the abundance of egocentric videos, eye gaze can be easily obtained and should be used as a direct signal. Our model conditions predictions on eye gaze data as an input signal to output fine-grained text annotations of human actions.

3 METHOD

We aim to develop a VLM that can accurately predict fine-grained human actions *in text* to facilitate human-machine interaction or collaboration in the physical world. In the following, we first formalize the problem and specify the training data, after which, we elaborate on the proposed mechanism that regularizes the attention in VLMs with gaze for enhanced predictions.

Problem setup Given egocentric video frames (observation) $\{\mathrm{I}_i\}_{i=1}^{\tau_o}$ over the past τ_o seconds (observation time), the *action prediction* VLM we aim to develop shall output text descriptions $\{\ell_i\}_{i=\tau_o+1}^{\tau_o+\tau_p}$ corresponding to future frames (what will happen) in the imminent τ_p seconds (prediction time). Let ϕ_{gaze} represent the gaze augmented VLM which aims to model the likelihood of the fine grained text descriptions (represented by ℓ_i) when eye gaze information is also provided:

$$\phi_{gaze}(\ell, \{\mathrm{I}\}, \{\mathrm{H}\}, \{\mathrm{G}\}) = \prod_{i=\tau_o+1}^{\tau_o+\tau_a} p(\ell_i \mid \ell_{<i}, \mathrm{I}_{\leq\tau_o}, \mathrm{H}_{\leq\tau_o}, \mathrm{G}_{\leq\tau_o}), \tag{1}$$

where $\{\mathrm{H}\}$ represents the gaze heatmaps and $\{\mathrm{G}\}$ represents the gaze overlaid images obtained by blending the gaze heatmaps and the video frames. It is shown by Lee et al. (2021) that human brains make predictions of the future over a hierarchy of timescales: ranging from fine-grained changes of words or images (in a duration of 1 to 4 seconds) to coarse-grained evolution of movie plots (up to 15 seconds). Since we focus on developing a model that can anticipate detailed human activities to facilitate human-machine interaction, we set the prediction time 'τ_p' to 2 seconds by default. Furthermore, Seidel et al. (2014) discovered that attention control mechanisms, in response to duress or stimuli, are activated for up to 2 seconds before the actual event, thus the observation time 'τ_o' is set to a value larger than 2 seconds. An ablation study of different observation/prediction horizons is reported in Sec. 8.3.4 in the Appendix. Next, we detail the training data and how the gaze information is utilized to promote human action prediction.

3.1 Dataset construction

We use the Ego4D dataset (Grauman et al., 2022), which includes video clips with eye-gaze data and textual action-object annotations. Gaze points are converted into images to highlight important visual regions. Videos are downsampled to one frame per second to reduce computational load.

For fine-grained action prediction, we enhance existing annotations with detailed textual descriptions using GPT-4V (OpenAI, 2023). This involves generating initial frame-by-frame descriptions and refining them by iteratively improving the prompt based on manual feedback. The final text annotations provide contextually coherent descriptions for each frame. During model training and testing, a sequence of images is input, and the subsequent text annotations serve as ground truth for action prediction.

Specifically, for every image frame I_i, we obtain a corresponding text annotation l_i. Figure 5 and Figure 6 in the Appendix provides a diagram illustrating this process, along with an example of the resulting annotations (Figure 7). For training and testing, the model is provided with a sequence of images $\{\mathrm{I}_i\}_{i=1}^{\tau_o}$, while the text annotations $\{l_i\}_{i=\tau_o+1}^{\tau_o+\tau_p}$ corresponding to the image frames $\{\mathrm{I}_i\}_{i=\tau_o+1}^{\tau_o+\tau_p}$ serve as the ground truth for predicting the anticipated actions. Here, τ_o represents the observation time, set to 5 seconds, and τ_p represents the prediction time, set to 2 seconds. Please refer to Sec. 8.1.1 for more details on the dataset construction pipeline. Next, we introduce the base model without eye gaze enhancement before delving into the proposed techniques that leverage eye gaze by regularizing the attention maps in the prediction.

3.2 Base model

Leveraging Flamingo's capability to model cross-modal interactions between images and language, we adapt the open-source Flamingo model (Awadalla et al., 2023) for our base model. We train the perceiver resampler in the vision block and the cross-attention layers in the language block, keeping other components frozen. The input image sequence $\{\mathrm{I}_i\}_{i=1}^{\tau_o}$ is processed by a pre-trained ViT to extract visual features, which are then fed into a perceiver resampler, learning a fixed-size representation of the visual data and leveraging time embeddings to capture temporal relationships. The output from the perceiver resampler connects to the language module to generate predicted text annotations. This base model, trained without gaze information, serves as a benchmark for evaluating our gaze-augmented models. Let ϕ represent the base model, then the problem can be formulated as follows:

$$\phi(\ell, \{\mathrm{I}\}) = \prod_{i=\tau_o+1}^{\tau_o+\tau_a} p(\ell_i \mid \ell_{<i}, \mathrm{I}_{i\leq\tau_o}) \,. \tag{2}$$

Furthermore, let $P(\ell)$ be the actual probability distribution of the ground-truth text, and $\phi(\ell, \{\mathrm{I}\})$ the model predicted distribution. The cross-entropy loss to be minimized for training can be written as:

$$\mathcal{L}_{\mathrm{CE}}(\{\mathrm{I}\}, \ell) = -\sum P(\ell) \log(\phi(\ell, \{\mathrm{I}\})) \,. \tag{3}$$

This setup establishes a clear baseline, allowing us to measure the impact of incorporating gaze information in subsequent models after we elaborate on the regularization technique for enhanced human action prediction in the following.

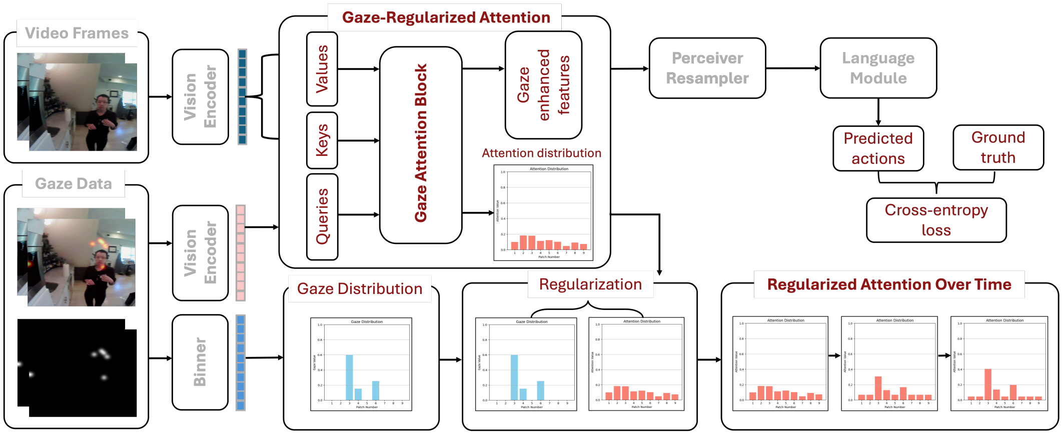


Figure 2: **Overview of the Architecture** A ViT encoder extracts features from video frames and gaze-overlaid images. A gaze-regularized attention block processes both the original and gaze-overlaid images, focusing on regions indicated by gaze patterns to produce gaze-enhanced features. These features are then passed to a Perceiver Resampler, which generates a fixed-size representation for the language module to predict future actions. A gaze regularizer aligns the model's attention with human gaze patterns by minimizing the Kullback–Leibler divergence between the attention distribution obtained from the attention block and the gaze distribution obtained from the binary heatmap images.

3.3 Gaze Augmented Model

Human gaze patterns reveal key insights into attention and focus, providing clues about intentions, thoughts, and potential actions (Tipper, 2010; Frischen et al., 2007). Leveraging this natural attention mechanism, we propose a gaze-regularized attention block to enhance visual features from the pre-trained ViT before passing them to the Perceiver Resampler. The input to the gaze-regularized block includes a sequence of RGB images from the video, $\{\mathrm{I}_i\}_{i=1}^{\tau_o}$, where i denotes time in seconds and τ_o is the observation time. Additionally, we incorporate corresponding binary heatmap images $\{\mathrm{H}_i\}_{i=1}^{\tau_o}$, highlighting regions of gaze. These heatmaps are generated from textual eye gaze data and are blacked out except for the gaze points, which are smoothed using a Gaussian filter to create a smooth heatmap. Moreover, these heatmaps generate a gaze distribution, indicating how gaze is distributed across image patches. Unlike the base model, the gaze-augmented model utilizes both eye gaze data and RGB images. The binary heatmap images and gaze-overlay images derived from the eye gaze data are shown in Figure 3. More information about the gaze-augmented model and it's components can be found in Sec.8.2 of the Appendix.

3.3.1 Gaze Distribution Computation

Each binary heatmap image is first divided into patches. Since gaze is usually concentrated on specific regions, not every patch will contain gaze information. Pixels occupied by gaze are indicated by non-zero pixel values, while pixels without gaze are blacked out, having values of zero. For a binary heatmap image denoted as H_t, let N represent the total number of patches. The proportion of gaze occupied within a patch $N_{i,j}$, where (i, j) represents the patch's position in the grid, can be calculated as follows:

$$N_{i,j} = \frac{\sum_{y=\frac{j(h)}{n_v}}^{\frac{(j+1)(h)}{n_v}} \sum_{x=\frac{i(w)}{n_h}}^{\frac{(i+1)(w)}{n_h}} p_{xy}}{\sum_{y=0}^{h} \sum_{x=0}^{w} p_{xy}} \quad \text{for } i \in \{0, n_h - 1\} \text{ and } j \in \{0, n_v - 1\}\,. \tag{4}$$

Here, h and w represent the height and width of the image, n_h and n_v denote the number of horizontal and vertical patches, respectively, and p_{xy} is a binary variable representing the pixel value at position (x, y), which can be either 0 or 1. The denominator corresponds to the total value of all pixels in the image. For each binary heatmap image H_t, we derive a gaze distribution $\hat{H}_t$, which is a vector of shape $(1, n_h \times n_v)$. The number of patches, N, in the gaze distribution is equal to the number of tokens produced per image by the ViT. This gaze distribution plays a crucial role in the gaze regularizer used in the attention block, acting as the target distribution that we want the attention distribution to mimic, which will be discussed in the following section.

3.3.2 Gaze-regularized attention block

To obtain features from the RGB images and gaze-overlaid images respectively, we utilize the ViT, and additional information can be found in Sec. 8.2.1 of the Appendix. The features derived from the gaze-overlaid images serve as queries (Q) in the gaze-regularized attention block. Simultaneously, the scene image features from the RGB images and are used as keys (K) and values (V). Attention is then calculated based on the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V = AV \,, \tag{5}$$

where A represents the attention weights. Simply using gaze features as queries does not guarantee that the attention will be concentrated on gaze-allocated regions. The attention scores will be distributed across the image. We aim to guide the model to prioritize these regions, directing more focus (attention) towards areas where gaze is present. To this end, we introduce a gaze regularizer designed to enhance the model's attention on gaze-allocated regions.

The gaze regularizer is implemented using Kullback-Leibler (KL) divergence, which is applied to the training of the gaze-augmented model. The aim of the regularizer is to ensure that the distribution of the attention weights is more closely aligned with the gaze distribution obtained from the binary heatmap image. We denote the attention weight distribution as A, while the gaze distribution is represented by H. The KL divergence for the two distributions can be obtained using the following equation:

$$D_{\text{KL}}(A \| H) = \sum_i A_i \log \frac{A_i}{H_i} \,. \tag{6}$$

The binary heatmap images are not used as queries because doing so could lead to important image information being overlooked. For instance, the heatmap images for two totally different images could be the same if one looks at the center of the image.

To avoid this, we use gaze-overlaid images as queries, which preserve the gaze information while also retaining the unique visual content of each image. It is crucial to preserve the visual content while aligning the attention distribution with the gaze distribution to ensure the model properly integrates both sources of information.

The overall objective function we minimize to train the gaze-augmented model is as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda * D_{\text{KL}}(A \| H) \,, \tag{7}$$

where λ is the coefficient of the regularization term in the proposed objective function, and the cross-entropy loss is similar to the loss from the base model. By utilizing this gaze-regularized attention mechanism within the model, we can orient the model's attention to regions highlighted by gaze, and thus improve the quality of the predicted future actions.

3.3.3 Singular and Aggregated gaze augmented models

In eye-tracking literature, eye movements are primarily categorized into fixations and saccades. During a fixation, the eyes pause and focus on a specific area, gathering detailed visual information. Most visual data is acquired during fixations, which typically last around 200 ms (Rayner, 2009). Although small involuntary eye movements may occur during fixations, they still concentrate on a specific region of interest. Saccades, on the other hand, are rapid eye movements that shift focus from one point of interest to another, causing significant changes in the visual scene. Saccades are crucial for visual search, and two consecutive saccades can cover considerable distances. Given the nature of saccades and fixations, as well as their average durations, detailed or relevant information is typically collected within a temporal window of $\delta = 200$ milliseconds.

Correspondingly, we provide both a singular gaze-augmented model and an aggregated gaze-augmented model. For each image frame I_t, we generate a heatmap image H_t and a gaze-overlaid image G_t using gaze coordinates available at time t. Gaze data from the Ego4D dataset (Grauman

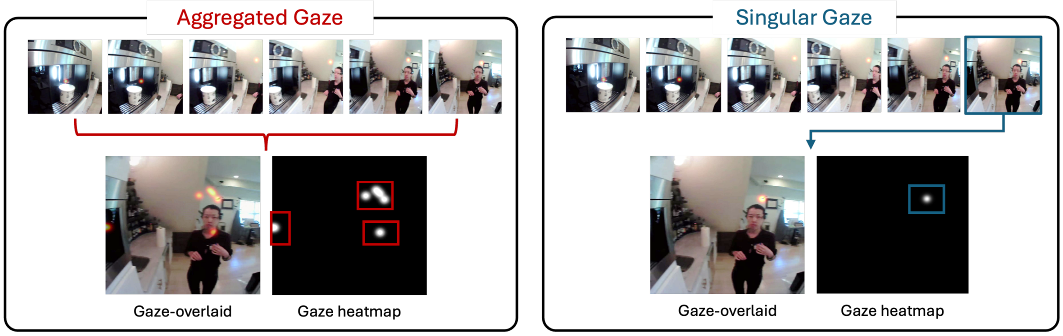


Figure 3: **Heatmap creation.** Illustration of gaze data collection for generating gray-scale heatmaps and gaze-overlaid images. On the left, the aggregated gaze model incorporates multiple gaze points collected over the interval $[t-\delta, t]$ to generate the heatmap. On the right, the singular gaze model uses a single gaze point collected at time t. Both with Gaussian smoothing.

et al., 2022) is sampled at 30 frames per second and stored in the format $g_t = (\text{timestamp}, x, y)$, where (x, y) is the coordinate of the gaze point. For the singular gaze-augmented model, the heatmap at time t, H_t, is:

$$\mathrm{H}_t = f(g_t)\,. \tag{8}$$

Here, the function f represents a mechanism to construct the heatmap, such as a Gaussian smoothing. In the singular gaze-augmented model, we use a single gaze point captured at the timestamp nearest to when the RGB image is obtained.

However, a singular gaze point can be noisy due to measurement errors or micro-saccades (Rolfs, 2009; Ratliff & Riggs, 1950; Collewijn & Kowler, 2008). An aggregated gaze pattern can account for such variations. To minimize the impact of noise and ensure a sufficient time frame for collecting detailed information, we propose aggregating points within a fixed time interval δ around the image frame I_t. The corresponding heatmap H_t at time t is then:

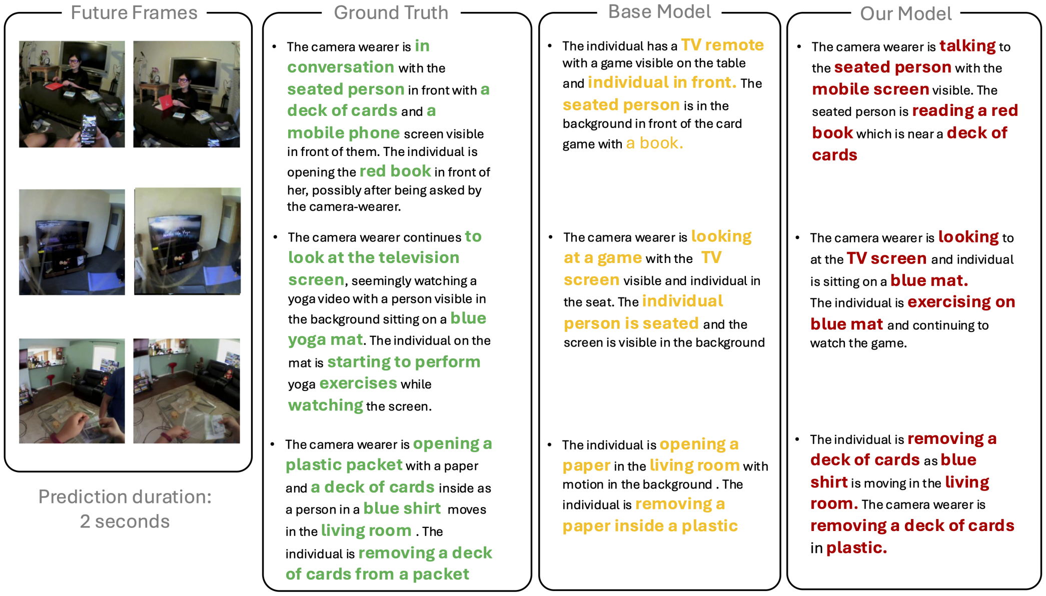
$$\mathrm{H}_t = \sum f(g_i), \quad \forall i \in [t-\delta, t]\,, \tag{9}$$

which is normalized by the length of the interval, and g_i for all $i \in [t-\delta, t]$ represents the set of gaze points lying within the vicinity of timestamp t. These gaze points are used to obtain both the binary heatmap image H_t and the gaze-overlaid image G_t. However, aggregation can sometimes include gaze points that may be occluded in the final frame. To address this, we implement an occlusion check using forward and backward optical flow consistency. The purpose of the occlusion check is to ensure that if significant occlusion occurs, the corresponding gaze points are excluded from the heatmap creation process. More information about the occlusion check can be found in Sec. 8.2.3. An example of using the occlusion check is shown in Figure 10.

To recap, for the aggregated gaze-augmented model, we collect gaze data points over a fixed interval, denoted as δ, to establish a more informative gaze pattern. This aggregation of gaze points, after an occlusion check, is then used to create a heatmap, which is subsequently overlaid on the RGB scene image to produce the gaze-overlaid image. Next, we demonstrate the effectiveness of the proposed gaze-regularization mechanisms for human action prediction.

4 Experiments

In this section, we begin by outlining the evaluation metrics used to assess model performance. Then, we compare the base models with the gaze-augmented models to emphasize the significance of incorporating gaze as a signal along with gaze-regularized attention. Following this, we explore multiple variations of the gaze-augmented models to identify the best-performing configuration and to justify the effectiveness of gaze in different settings discussed in the previous section. Finally, we present evidence supporting the use of finer-grained text annotations for gaze-augmented models to enhance overall performance. Additional information about the other experiments conducted as part of our study can be found in Sec.8.3 whereas the training and evaluation details for the models can be found in Sec.8.5.



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.
- 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.
- 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 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.
- 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
- 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 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
- 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.
- 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.

Figure 4: Action prediction results for the base model (without gaze) and our gaze-augmented model are presented for an observation horizon of 5 seconds. Past frames are omitted, but ground-truth annotations and future frames with a prediction duration of 2 seconds are provided as references. Keywords from each set of annotations are highlighted for easier reading.

4.1 Evaluation Metrics

For evaluation, we propose using a semantic transformer (Reimers & Gurevych, 2019) to provide a quantifiable score that compares the generated output with the ground-truth action text. This scoring system is designed to ensure that the model does not penalize variations of sentences that are semantically similar. Additionally, we aim for the model to penalize sentences where the word order is nonsensical or incomprehensible to humans. Alongside the semantic score, we also utilize the ROUGE-L score and METEOR score, which are widely employed to assess text similarity.

4.2 Comparison between Base Model and Gaze-Augmented Models

The objective of this experiment is to quantify the impact of eye gaze data on human action prediction by comparing model performance using measurable metrics. The base model receives only RGB image frames as input, while the gaze-augmented models incorporate multiple variations of gaze information. All gaze-augmented models share the same architecture but differ in the way gaze data is utilized. In the Singular Gaze-Augmented Model, for each image frame I_t at time t, the corresponding heatmap H_t is generated using a single gaze point g_t, obtained from the timestamp closest to t. In the Aggregated Gaze-Augmented Model, for each image frame I_t at time t, the corresponding heatmap H_t is generated using aggregated gaze points g_k, where $k \in [t - \delta, t]$ and δ is set to 200 milliseconds.

All models are trained under identical conditions. As shown in Table 1, the gaze-augmented models outperform the base model across all evaluation metrics. Using aggregated gaze points over a 200-millisecond interval shows a noticeable improvement in performance compared to using a single gaze point. This can be attributed to the larger focus area and the temporal gaze pattern enabled by the aggregation, allowing the model to capture a broader context for inferring the intent and better facilitate the anticipation of future actions. Please refer to Figure 4 for qualitative results.

4.3 Effect of the Gaze Regularizer

We assess the impact of the gaze regularizer using the aggregated gaze-augmented model to determine its importance. The final objective function includes the gaze regularization coefficient λ, and

Table 1: Evaluation of base model and gaze-augmented models with varying granularity in action annotations (i.e., GPT-4V and ShareCaptioner).

	Annotation	Semantic	Meteor	Rouge-L (↑)		
Model	source	Score(↑)	Score(↑)	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**

Table 2: Effect of regularization on aggregated gaze-augmented models.

Regularization	Semantic	Meteor	Rouge-L (↑)		
coefficient	Score(↑)	Score(↑)	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

we evaluate the model's performance on the test set using different values of λ. When $\lambda = 0$, the gaze regularizer is inactive and has no effect on the loss function, while larger values of λ increase its influence. As shown in Table 2, the performance of the aggregated gaze-augmented model without regularization is comparable to the base model. Optimal performance is achieved at moderate λ values, with a slight decline at higher regularization levels. The diminished performance without regularization underscores the significance of aligning attention with the gaze distribution through the gaze regularizer when the gaze attention block is included.

4.4 Gaze-regularized attention block

The gaze-regularized attention block is a critical addition to the baseline OpenFlamingo model. In our ablation studies, we examine the model's performance as the number of gaze-regularized attention blocks varies, aiming to identify the optimal number of blocks for further training. As shown in Table 3, the performance improves up to $n = 2$, after which it slightly declines with additional blocks. Further training confirms that using two gaze-regularized attention blocks yields the best results, as this configuration balances aligning the gaze and attention distributions while preventing over-alignment.

4.5 Effect of occlusion check on aggregated gaze-augmented models

In the aggregated gaze model, gaze points are collected within a specified time interval δ (200 ms). However, in dynamic environments, aggregating gaze points from the interval $[t-\delta, t]$ may introduce inaccuracies due to changes in the scene or camera movement. To mitigate this, we introduced an

Table 3: Effect of the number of gaze-regularized attention blocks on the performance of the aggregated gaze-augmented models.

Attention	Semantic	Meteor	Rouge-L (↑)		
blocks	Score(↑)	Score(↑)	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

Table 4: Comparison of the aggregated gaze-augmented models with and without occlusion check.

Aggregated gaze	Semantic	Meteor	Rouge-L (↑)		
model	Score(↑)	Score(↑)	Precision	Recall	F-score
w/o occlusion-check	0.7616	0.4718	0.5090	0.5508	0.5286
with occlusion-check	**0.7826**	**0.5033**	**0.5193**	**0.5644**	**0.5405**

occlusion check to ensure that only relevant gaze points are aggregated. More details about the occlusion check method can be found in Sec. 8.2.3 in the Appendix.

To evaluate the impact of this adjustment, we conducted experiments comparing models with and without the occlusion check. As shown in Table 4, the model incorporating the occlusion check slightly outperforms the one without it. The difference in the evaluation metrics can be attributed to the fact that only relevant and accurate gaze points are considered, which reduces noise and prevents the model from being confused by irrelevant data.

4.6 Impact of the annotation quality on model performance

To facilitate our experiments, we designed a prompt-based method to automatically generate text annotations for image frames using GPT-4V. Although these annotations were more detailed than the originals, we also generated additional annotations using ShareCaptioner (Chen et al., 2024), a tool specifically designed to produce comprehensive captions that capture changes between frames. We aim to compare the descriptive quality of these annotations and evaluate their impact on both the base and gaze-augmented models.

We find that annotations generated by ShareCaptioner are notably more detailed and fine-grained than those from our initial setup. While the base model's performance remained similar with both sets of annotations, the gaze-augmented model showed a nearly 12 percent improvement when using the more fine-grained annotations, as shown in Table 1. This improvement can likely be attributed to the model's ability to leverage areas of concentrated gaze when processing more descriptive captions, allowing for better prediction of detailed actions. From the experiment, we conclude that the importance of eye gaze increases even further when the granularity of the annotations is higher.

5 Conclusion

Our work demonstrates that incorporating gaze data significantly enhances the predictive abilities of VLMs, proving valuable for applications requiring precise human action prediction, such as assistive robotics and human-machine collaboration. By integrating eye gaze data, our approach improves the model's focus on critical image features, leading to enhanced prediction accuracy and reliability.

Our gaze-augmented model is trained with a gaze-regularized attention strategy, which simulates the natural attention mechanism driven by human gaze. This integration results in improved semantic and ROUGE scores, underscoring the effectiveness of gaze-regularized attention in boosting model performance. Through additional experiments, we highlight the crucial role of employing a gaze regularizer and gaze-augmented attention blocks, which provide insight into how to maximize the utilization of gaze information.

Looking ahead, we aim to further refine our model by enhancing the integration of gaze data and exploring other training techniques. We also plan to develop a dedicated large-scale dataset to support other researchers in running their models and advancing this methodology. To facilitate further research and development, we will make our code and dataset publicly available, allowing others to build upon our architecture and enhance its capabilities.

6 Ethics statement

We declare that our research does not present any potential ethical issues and is in compliance with the ICLR code of ethics. The study does not involve human subjects, sensitive data, or methodologies that could result in harmful outcomes or biases. All data this work uses is publicly available, and no privacy or security concerns are implicated.

7 Reproducibility statement

We have made significant efforts to ensure the reproducibility of our work. A code example is provided as supplementary material, demonstrating the core components of our approach. Upon acceptance, we will release all of the data and the complete training and testing code to facilitate the full reproducibility of our results.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

Reuben M. Aronson, Thiago Santini, Thomas C. Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. Eye-hand behavior in human-robot shared manipulation. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 4–13, 2018.

Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings, 2023. URL https://arxiv.org/abs/2301.02311.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023.

Abdishakour Awale and Duygu Sarikaya. Human gaze guided attention for surgical activity recognition, 2022.

Dutliff Boshoff, Raphael Elikplim Nkrow, Bruno Silva, and Gehard P. Hancke. Gait device pairing in unconstrained environments with signal morphing. *IEEE Transactions on Consumer Electronics*, 2024. URL https://api.semanticscholar.org/CorpusID:270081144.

Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intelligent Transportation Systems Magazine*, 9(4):10–22, 2017. doi: 10.1109/MITS.2017.2743165.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024.

Hyunjin Cho, Dong Un Kang, and Se Young Chun. Short-term object interaction anticipation with disentangled object detection @ ego4d short term object interaction anticipation challenge, 2024. URL https://arxiv.org/abs/2407.05713.

H. Collewijn and E. Kowler. The significance of microsaccades for vision and oculomotor control. *Journal of Vision*, 8(14):20–20, December 2008. ISSN 1534-7362. doi: 10.1167/8.14.20. URL http://dx.doi.org/10.1167/8.14.20.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4):694–724, July 2007. ISSN 0033-2909. doi: 10.1037/0033-2909.133.4.694. URL http://dx.doi.org/10.1037/0033-2909.133.4.694.

Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention, 2019.

Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation, 2017.

Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization, 2021. URL https://arxiv.org/abs/2104.04302.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022.

Michelle R. Greene, Benjamin J. Balas, Mark D. Lescroart, Paul R. MacNeilage, Jennifer A. Hart, Kamran Binaee, Peter A. Hausamann, Ronald Mezile, Bharath Shankar, Christian B. Sinnott, Kaylie Capurro, Savannah Halow, Hunter Howe, Mariam Josyula, Annie Li, Abraham Mieses, Amina Mohamed, Ilya Nudnou, Ezra Parkhill, Peter Riley, Brett Schmidt, Matthew W. Shinkle, Wentao Si, Brian Szekely, Joaquin M. Torres, and Eliana Weissmann. The visual experience dataset: Over 200 recorded hours of integrated eye movement, odometry, and egocentric video, 2024. URL https://arxiv.org/abs/2404.18934.

Junhwa Hur and Stefan Roth. Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Bruno Laeng, Ilona M. Bloem, Stefania D'Ascenzo, and Luca Tommasi. Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition*, 131(2):263–283, 2014. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2014.01.003. URL https://www.sciencedirect.com/science/article/pii/S0010027714000043.

Caroline S Lee, Mariam Aly, and Christopher Baldassano. Anticipation of temporally structured events in the brain. *elife*, 10:e64972, 2021.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators, 2024. URL https://arxiv.org/abs/2311.01378.

Yin Li, Miao Liu, and James M. Rehg. In the eye of the beholder: Gaze and actions in first person video, 2020. URL https://arxiv.org/abs/2006.00626.

Chen Lin and Xing Long. Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community. https://github.com/xiaoachen98/Open-LLaVA-NeXT, 2024.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019a. URL https://arxiv.org/abs/1908.02265.

Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4425–4434, 2019b. doi: 10.1109/ICCVW.2019.00543.

Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting, 2024.

Kyle Min and Jason Corso. Integrating human gaze into attention for egocentric activity recognition. 11 2020.

OpenAI. Gpt-4v(ision), 2023. URL https://cdn.openai.com/contributions/gpt-4v.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Floyd Ratliff and Lorrin A. Riggs. Involuntary motions of the eye during monocular fixation. *Journal of Experimental Psychology*, 40(6):687–701, 1950. ISSN 0022-1015. doi: 10.1037/h0057754. URL http://dx.doi.org/10.1037/h0057754.

Keith Rayner. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Martin Rolfs. Microsaccades: small steps on a long way. *Vision Res.*, 49(20):2415–2441, October 2009.

Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation, 2024. URL https://arxiv.org/abs/2211.14154.

Eva-Maria Seidel, Daniela M. Pfabigan, Andreas Hahn, Ronald Sladky, Arvina Grahl, Katharina Paul, Christoph Kraus, Martin Küblböck, Georg S. Kranz, Allan Hummer, Rupert Lanzenberger, Christian Windischberger, and Claus Lamm. Uncertainty during pain anticipation: The adaptive value of preparatory processes. *Human Brain Mapping*, 36(2):744–755, October 2014. ISSN 1097-0193. doi: 10.1002/hbm.22661. URL http://dx.doi.org/10.1002/hbm.22661.

Ali Shafti, Pavel Orlov, and A. Aldo Faisal. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 863–869, 2019. doi: 10.1109/ICRA.2019.8793804.

Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019. URL https://arxiv.org/abs/1908.07490.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. URL https://arxiv.org/abs/2003.12039.

Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Enhancing next active object-based egocentric action anticipation with guided attention, 2023. URL https://arxiv.org/abs/2305.12953.

Steven P. Tipper. Eps mid-career award 2009: From observation to action simulation: The role of attention, eye-gaze, emotion, and body state. *Quarterly Journal of Experimental Psychology*, 63 (11):2081–2105, November 2010. ISSN 1747-0226. doi: 10.1080/17470211003624002. URL `http://dx.doi.org/10.1080/17470211003624002`.

Daniel Weber, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling location proposals of unknown objects through gaze information for human-robot interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11086–11093, 2020. doi: 10.1109/IROS45743.2020.9340893.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations, 2024. URL `https://arxiv.org/abs/2401.18084`.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos?, 2024a. URL `https://arxiv.org/abs/2307.16368`.

Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. Vialm: A survey and benchmark of visually impaired assistance with large models, 2024b. URL `https://arxiv.org/abs/2402.01735`.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook, 2024. URL `https://arxiv.org/abs/2310.14414`.

8 Appendix

In this appendix, we present supplementary material related to our study. This includes detailed information on the dataset creation process and the design of the prompts used in our experiments. We also include results from several ablation studies conducted during our research. Finally, we provide essential details about the model training process to assist readers who may wish to reproduce our work from scratch.

8.1 Dataset

The Ego4D and Epic-Kitchens datasets consist of egocentric videos capturing camera-wearers performing daily activities in controlled settings (Grauman et al., 2022; Damen et al., 2022). Other relevant datasets include the EGTEA+ Gaze dataset, which offers 28 hours of content focused on cooking and kitchen activities (Li et al., 2020), and the more recent Visual Data Experience (VDE) dataset, which contains around 240 hours of clips documenting day-to-day activities coupled with gaze and head tracking (Greene et al., 2024).

These datasets are supplemented with text annotations in the form of (action, verb) tuples as well as coarse-grained narrations, which are used for action anticipation and activity forecasting challenges. For our study, we collected visual eye gaze data from the Ego4D dataset and modified the text annotations from coarse-grained descriptions to more descriptive fine-grained descriptions. In the future, we plan to incorporate the more recent VDE dataset into our study as well.

In this section, we provide details about the dataset creation process used for model training and testing, the prompt design, and alternative sources of annotations in the form of VLMs.

8.1.1 Dataset construction and prompt design

The Ego4D dataset comprises egocentric video clips along with supplementary data such as audio, text annotations, eye gaze data, and additional metadata (Grauman et al., 2022). For our project, we focused on the subset of video clips that include eye gaze data, containing approximately 33.3 hours of egocentric videos recorded from 80 participants. The eye gaze data is provided in numerical form, containing canonical timestamps and the pixel coordinates of gaze points. We transform gaze points to images to represent important visual regions, aligning with how humans perceive spatial information (Laeng et al., 2014). Due to the minute differences between consecutive images in the original videos, we perform downsampling to one image per second to reduce computational requirements while remaining effective.

Since our focus is fine-grained action prediction from ego-centric videos, we augment existing annotations with detailed textual descriptions to enhance human-machine interaction. We leverage GPT-4V to generate annotations for video frames by processing a sequence of images and prompting it to describe each frame. This method ensures contextual coherence across frames. After obtaining initial descriptions, we manually evaluate their quality and provide feedback to refine the output. This feedback is used to modify the prompt, which is then fed back into GPT-4V with the image frames. We conduct prompt selection and refinement using multiple sequences from different video clips to ensure generalizability. This iterative process continues until we establish an optimal prompt template that consistently yields accurate and contextually appropriate annotations, confirmed through human evaluation.

Specifically, we start by passing a small set of images to GPT-4V with a basic prompt like: “Describe what is happening in the image sequence and output the text descriptions.” The initial output was manually evaluated, and feedback was provided. This feedback, along with the original prompt, was passed to ChatGPT to refine the prompt for generating accurate and meaningful text annotations. The manual evaluation process ensured that the generated annotations met the following criteria:

1. A clear description of the objects being manipulated or focused on in the scene.
2. A detailed account of the actions being performed by both the camera wearer and other individuals present in the images.
3. Information about any trajectories or movements that take place within the scene.

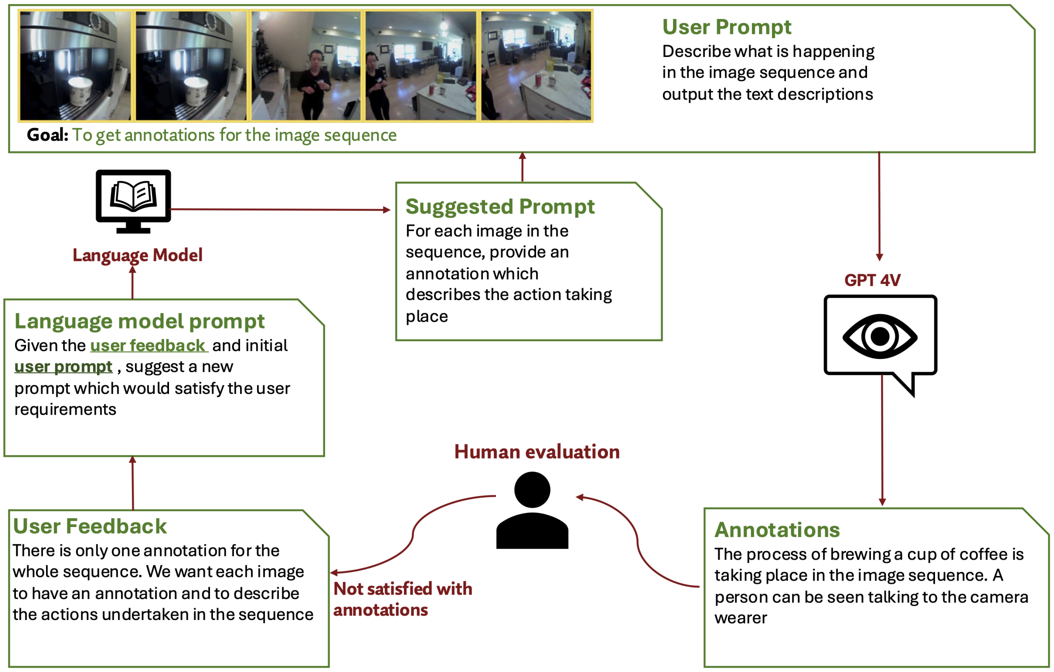

User Prompt
Describe what is happening in the image sequence and output the text descriptions
Goal: To get annotations for the image sequence
Language Model
Suggested Prompt
For each image in the sequence, provide an annotation which describes the action taking place
GPT 4V
Language model prompt
Given the user feedback and initial user prompt , suggest a new prompt which would satisfy the user requirements
Human evaluation
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
Not satisfied with annotations
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

Figure 5: **GPT-4V Iterative Prompting Workflow.** The process begins with a sequence of images and an initial prompt, which are input into GPT-4V to generate annotations for each image. The user then evaluates these annotations and provides feedback, which is incorporated into the prompt using a language model. This modified prompt is used to refine the annotations in a continuous cycle. The objective is to improve the quality and relevance of the output with each iteration until the user is satisfied. For GPT-4V, a set of 10 images is provided at once, ensuring that the annotations maintain contextual coherence.

4. Clear, fine-grained annotations that fulfill these criteria in a way that is easily understandable by both humans and machines, such as robots that may use these instructions for task execution.

After several iterations of refining the prompt and evaluating the results, we identified a suitable template for generating high-quality text annotations. This final template was then used to annotate the image sequences using GPT-4V. More details on the prompt design process can be found in Figure 5 and Figure 6.

8.1.2 Text annotations using an alternate VLM

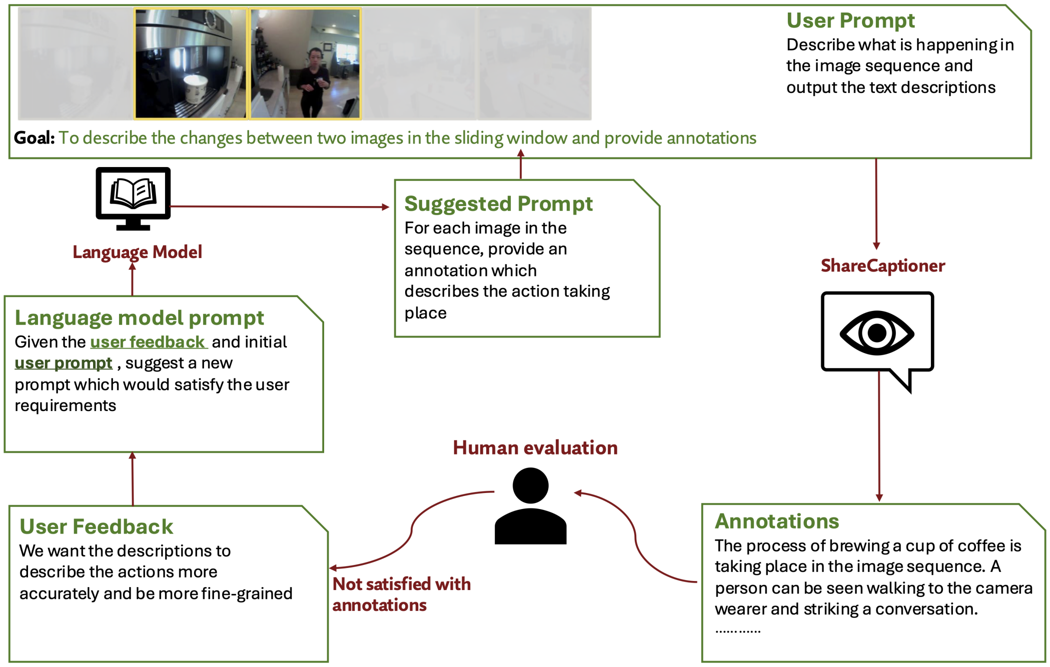
In the previous section, we obtained text annotations using GPT-4V. In addition, we conduct further experiments using an alternate VLM to generate text annotations for the same video clips.

Specifically, we employed ShareCaptioner (Chen et al., 2024) to annotate the entire dataset. Under similar training conditions, we compared the empirical performance of the models using annotations from GPT-4V and ShareCaptioner. The purpose of this comparison was to assess whether there is a significant performance variation – either a drop or improvement – when text annotations are sourced from a different VLM. Additionally, we aimed to confirm whether the trend of superior performance in gaze-incorporated models persists regardless of the annotation source. Finally, we sought to investigate if the granularity of the annotations impacts the model's performance.

As part of this investigation, we also experimented with an open-source implementation of LLaVA to generate finer-grained annotations Lin & Long (2024). However, upon manual evaluation, we found no significant difference between the annotations produced by ShareCaptioner and those generated by LLaVA. In order to have a dedicated comparison between fine-grained annotations and finer-grained annotations on the base model and the gaze-augmented model, we decided to proceed with GPT-4V annotations and ShareCaptioner annotations. However, a comparison of the semantic scores obtained by the base model and aggregated gaze-augmented models for all three annotation models can be found in Figure 7. We observed that the gaze-augmented models exhibited a 10-12 percent improvement in performance when utilizing the finer-grained annotations, indicating that more detailed descriptions can further enhance model accuracy in predicting future actions.

Examples of the annotations obtained for a sample image, as well as the semantic scores for the base models and the gaze-augmented models, can be seen in Figure 7. These examples highlight the difference in performance under varying quality of the fine-grained text annotations.

Note: It is important to emphasize that our goal is not to compare the performance of ShareCaptioner and GPT-4V, but rather to evaluate the quality of the annotations obtained from each system and the effect of the quality on the gains of employing gaze in the prediction. The differences in annotation granularity may be influenced by cost and time constraints, as the setups used to generate these annotations varied. Therefore, a direct comparison of the VLMs themselves would be inappropriate.



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

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

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

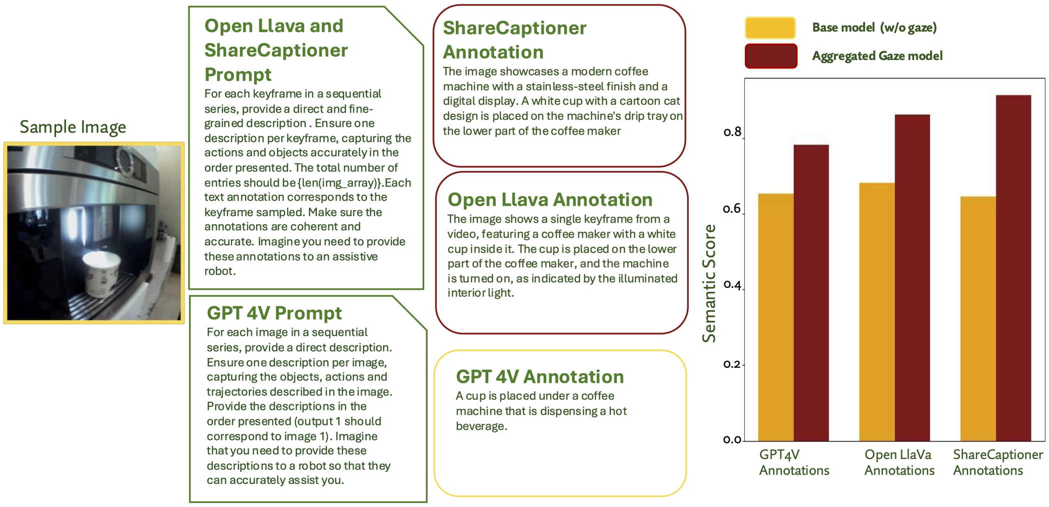
User Feedback: We want the descriptions to describe the actions more accurately and be more fine-grained

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. ....

Figure 6: **ShareCaptioner Iterative Prompting Workflow.** The process begins with a sequence of images accompanied by a starter prompt, which is processed by ShareCaptioner to generate initial annotations. The user then reviews these annotations and provides feedback. This feedback is integrated into the prompt using a language model. The updated prompt is subsequently used to enhance the annotations through an iterative cycle. This cycle continues until the output meets the user's satisfaction, focusing on improving both quality and relevance. Additionally, a sliding window mechanism is employed to traverse the image sequence, capturing and describing changes as they occur to ensure that all actions are accurately recorded.

8.2 Model overview and components

In our study, we employ the open-source version of the Flamingo model (Awadalla et al., 2023), built upon the foundational work of the original Flamingo developed by Alayrac et al. (2022). The Flamingo is a VLM designed to leverage interleaved text and image data. It features a pre-trained vision encoder to extract input features, a trainable perceiver resampler for creating fixed-size representations of the input features, and trainable cross-attention layers on the language side. The latent features learned by the perceiver resampler are integrated into the language module. These features, along with the processed text, are input to the cross-attention layers in the language module which combines the visual features with the language features. The model utilizes a cross-entropy



Open Llava and ShareCaptioner Prompt

For each keyframe in a sequential series, provide a direct and fine-grained description . Ensure one description per keyframe, capturing the actions and objects accurately in the order presented. The total number of entries should be {len(img_array)}.Each text annotation corresponds to the keyframe sampled. Make sure the annotations are coherent and accurate. Imagine you need to provide these annotations to an assistive robot.

GPT 4V Prompt

For each image in a sequential series, provide a direct description. Ensure one description per image, capturing the objects, actions and trajectories described in the image. Provide the descriptions in the order presented (output 1 should correspond to image 1). Imagine that you need to provide these descriptions to a robot so that they can accurately assist you.

ShareCaptioner Annotation

The image showcases a modern coffee machine with a stainless-steel finish and a digital display. A white cup with a cartoon cat design is placed on the machine's drip tray on the lower part of the coffee maker

Open Llava Annotation

The image shows a single keyframe from a video, featuring a coffee maker with a white cup inside it. The cup is placed on the lower part of the coffee maker, and the machine is turned on, as indicated by the illuminated interior light.

GPT 4V Annotation

A cup is placed under a coffee machine that is dispensing a hot beverage.

Figure 7: **Left**: Examples of annotations generated for a sample image by three different annotators: ShareCaptioner, Open Llava, and GPT-4V. The annotations from ShareCaptioner and Open Llava are more fine-grained compared to those produced by GPT-4V in our experimental setup. **Right**: A graph comparing the performance of the base model and gaze-augmented models, illustrating the impact of using text annotations of varying granularity.

loss mechanism, aiming to maximize the probability of predicting the correct text token based on preceding image and text.

In addition to traditional input modalities, such as RGB images, recent advancements have highlighted the potential of integrating various modalities beyond vision, including gaze, gait, and tactile sensors (Boshoff et al., 2024; Yang et al., 2024). Building on this premise, our approach incorporates eye gaze as an additional signal. Specifically, we introduce a gaze-regularized attention mechanism to make use of eye gaze as a signal. The base model without eye gaze data serves as a benchmark and relies exclusively on RGB images for input. In contrast, the gaze-augmented model integrates eye gaze data alongside the RGB images. The key distinction of the gaze-regularized attention models is their ability to condition the output using both RGB images and eye gaze data.

In the following section, we provide more details about some of the components in the model architecture, as well as provide information about the occlusion check which is used for gaze point correction during aggregation of gaze points.

8.2.1 Vision Transformer Encoder

In both the base and gaze-augmented models, we utilize a pre-trained Vision Transformer (ViT) as the image encoder. Images are processed as sequences, where each image is tokenized into patches that are flattened and then transformed into embeddings. To maintain spatial relationships between the patches, these embeddings are supplemented with positional embeddings, ensuring coherence across the entire image. An illustration of the ViT architecture is provided in Figure 8. In the figure, as an example, we show how an image is divided into nine tokens, each of which is assigned a positional embedding. These tokenized patches are then passed through the pre-trained transformer encoder, resulting in an embedding that represents the original sequence of image patches. The pre-trained vision encoder employed in our study is ViT-L-14, which results in the formation of 256 tokens for each image. In the example shown in Figure 8, the observation length (τ_o) is set to 5 seconds, meaning that the image sequences span a 5-second interval. The primary distinction between the original design and the approach used in our study is the omission of the final MLP head, as our focus is exclusively on extracting image features.

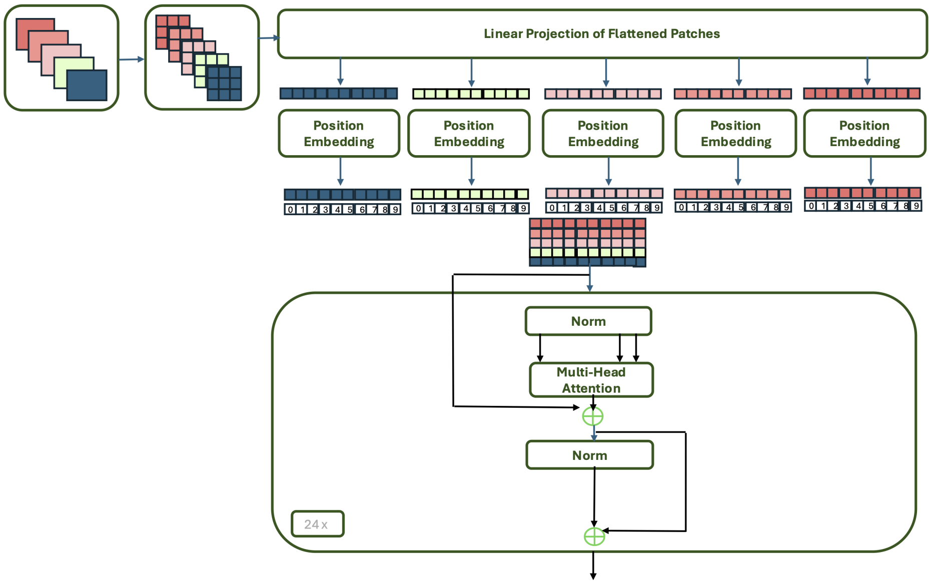


Figure 8: An example of how images are processed by the Vision Transformer in our model. In this case, each image is tokenized into 9 patches. Positional embeddings are then added to the patches, which are subsequently passed through an attention module. In this model, the final MLP head is omitted, as we are focused solely on extracting image features for further analysis.

8.2.2 Gaze-guided attention block

In our study, we propose the inclusion of a gaze-regularized attention mechanism, which features a gaze-guided attention block that operates in conjunction with the gaze regularizer. In the gaze-guided attention block, gaze-overlaid images are processed through the ViT to extract features that serve as queries for the attention module. These queries contain information about both the scene and the regions occupied by gaze, leading us to refer to this block as the gaze-guided attention block. Once the gaze-based features are obtained, they pass through a learnable linear layer to produce the queries involved in the attention mechanism. The visual features obtained from the RGB image frames, after passing through the ViT, are used as keys and values within the gaze-guided attention block. Similar to the gaze-based features , the features obtained from the RGB image frames pass through a linear layer to form the keys and values for the attention mechanism. The final output, consists of gaze-enhanced features that are informed by the gaze-based queries.

Additionally, the attention distribution is calculated, reflecting how attention is allocated across all tokenized patches. Initially, this attention distribution may not align with the gaze distribution. Therefore, through regularization, we aim to progressively align the attention distribution with the gaze distribution over time. A closer look at the gaze-guided attention block can be found in Figure 9.

8.2.3 Gaze aggregation with occlusion correction

The input image sequence can exhibit dynamic changes due to environmental movement or the movement of the camera wearer. The method for aggregating gaze points is suitable only when the frames within the δ interval for which aggregation is done, shows moderate movement. However, preventing movement in a dynamic environment is challenging. If we aggregate gaze points in the $[t-\delta, t]$ interval to construct the heatmap H_t and there is major occlusion between the earlier frames $[t-\delta, t)$ and the final frame at timestamp t, it becomes impractical to include gaze points from the occluded frames. In such cases, collecting gaze points from occluded frames may lead to inaccurate or misleading representations in the heatmap. To alleviate this issue, we perform an occlusion check between each frame in the $[t-\delta, t)$ interval and the final frame at time t to ensure appropriate gaze aggregation. In the case of significant occlusion or a drastic change in the scene,

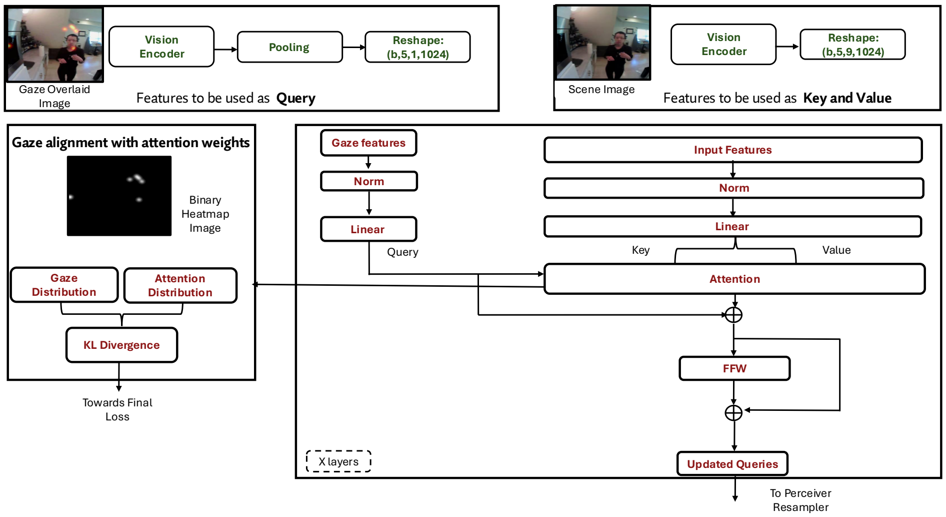


Figure 9: Closer look at the gaze-guided attention block in the pipeline. Gaze overlaid images are passed as queries, whereas the corresponding RGB images are passed as key and value. The binary heatmap image is used to obtain the gaze distribution, which is used in the regulariser by comparing it with the attention distribution obtained from the gaze-guided attention block. The regulariser attempts to align the attention distribution more towards the gaze distribution.

gaze points corresponding to the earlier frames should not be collected for the aggregation and subsequent formation of the heatmap H_t.

Using a method similar to the consistency check with optical flow presented by Hur & Roth (2017), we explicitly exclude gaze points that are occluded in the current frame. If a pixel is correctly translated and there is no major occlusion, then the difference between the forward optical flow displacement of this pixel (x, y) and the displacement of the translated pixel $(\hat{x}, \hat{y})$ with backward optical flow should be close to zero.

For an RGB image I_t at time t, we gather the image frames $\{\mathrm{I}_k\}$ for all $k \in [t-\delta, t)$. Let the forward optical flow between images $\{\mathrm{I}_k\}$ and $\{\mathrm{I}_t\}$ in the horizontal direction be denoted by $Fx_{k\to t}$ and the backward optical flow by $Fx_{t\to k}$. Similarly, $Fy_{k\to t}$ represents the forward optical flow in the vertical direction, and $Fy_{t\to k}$ represents the backward optical flow in the vertical direction. Let the coordinates of a designated pixel p_i be (x_i, y_i). The new coordinates of the translated pixel in the subsequent frame, using optical flow, are computed as follows:

$$\begin{aligned} \hat{x}_i &= x_i + Fx_{k\to t}(x_i)\,, \\ \hat{y}_i &= y_i + Fy_{k\to t}(y_i)\,. \end{aligned} \tag{10}$$

Next, we calculate the distance moved by this designated pixel in the horizontal and vertical directions according to the following equations:

$$\begin{aligned} d_{x_i} &= |Fx_{k\to t}(x_i)| - |Fx_{t\to k}(\hat{x}_i)|\,, \\ d_{y_i} &= |Fy_{k\to t}(y_i)| - |Fy_{t\to k}(\hat{y}_i)|\,. \end{aligned} \tag{11}$$

If the observed proportion of pixels η_{observed} exceeding the distance discrepancy is more than a predefined threshold η, we conclude that a major occlusion has occurred; otherwise, the occlusion is minor. The observed proportion of such pixels η_{observed} is calculated as:

$$\eta_{\text{observed}} = \frac{\sum_{i=1}^{h\times w} \mathbf{1}_{\text{condition}}\left(\sqrt{(d_{x_i})^2 + (d_{y_i})^2} > \epsilon\right)}{\sum_{i=1}^{h\times w} 1}\,, \tag{12}$$

where the denominator represents the total number of pixels in the image.

We disregard the gaze points for frames $\{\mathrm{I}_k\}$ where there is major occlusion with respect to the image frame I_t. If the occlusion is minor, the appropriate gaze points $\{g_i\}$ for all $i \in [t-\delta, t]$ are then translated into their new coordinates $\{\hat{g}_i\}$ and collected for the formation of heatmap H_t and subsequently, for the gaze-overlaid image G_t. The transformed gaze points are computed as follows:

$$\begin{aligned}(\hat{g}_i)_x &= (g_i)_x + Fx_{k\rightarrow t}((g_i)_x)\,, \\ (\hat{g}_i)_y &= (g_i)_y + Fy_{k\rightarrow t}((g_i)_y)\,.\end{aligned} \tag{13}$$

As mentioned above, the idea is that if there is a major occlusion, the difference between the distance traversed by a pixel during forward optical flow,and the distance traversed by the translated pixel during backward optical flow will be significantly greater than in cases where the occlusion is minor. Optical flow was calculated using the implemntation of the RAFT model developed by Teed & Deng (2020). Human feedback was utilized to distinguish between major and minor occlusions on sample image sequences, which informed the selection of hyperparameters ϵ and η. The hyperparameter ϵ is the threshold distance, which was set to 20, whereas η is the threshold proportion of pixels that have exceeded the occlusion limit, set to 0.60. An example of the result of our occlusion check can be found in Figure 10. It can be observed that if the occlusion check is not present, the aggregation points do not accurately reflect where the person was looking in the final frame on which gaze is overlaid. The occlusion check ensures that the correction is effective.

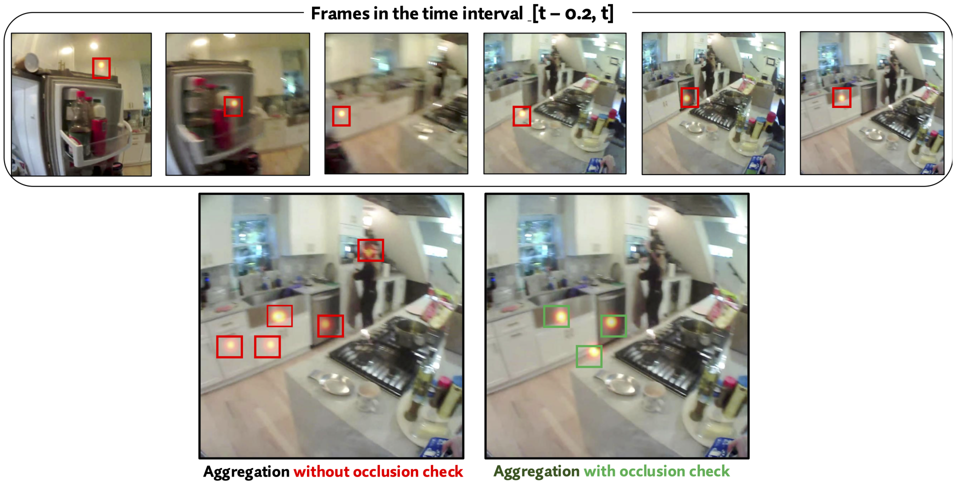

Frames in the time interval [t – 0.2, t]
Aggregation without occlusion check
Aggregation with occlusion check

Figure 10: An example of gaze aggregation on the final image frame is shown. **Bottom left**: Result without occlusion check. **Bottom right**: Result with occlusion check applied. The fridge is absent in the final frame, indicating significant occlusion. Therefore, the gaze points from the first two frames, which are associated with the fridge, should not be overlaid on the final frame as the object is no longer visible.

8.3 Other experiments

In this section, we provide some details about the ablation studies conducted during the course of our study. We first explore the impact of the size of the gaze points used in the formation of the heatmap on the model performance. In addition, we employ a self-attention block to the base model without gaze to assess the result comparison with the gaze-augmented model. We also utilize gaze in text form to investigate whether our need to use gaze based images is necessary or not, followed by an investigation of the impact of changing observational and prediction horizons on model performance.

8.3.1 Changes in the size of the overlays of the gaze points

In our exploration of gaze-augmented models, we investigated two primary approaches: the singular gaze model and the aggregated gaze model. The singular gaze model utilizes only one gaze point to

Table 5: Effect of the size of the gaze point overlays on model accuracy

Gaze	Semantic	Meteor	Rouge-L (↑)		
Model	Score(↑)	Score(↑)	Precision	Recall	F-score
Singular	0.7316	0.4501	0.4822	0.5309	0.5054
Singular (larger overlays)	0.7683	0.4922	0.5123	0.5508	0.5304
Aggregated	**0.7826**	0.5033	**0.5193**	**0.5644**	**0.5405**
Aggregated (larger overlays)	0.7816	**0.5060**	0.5128	0.5556	0.5330

Table 6: Comparison of base model with attention block against gaze augmented models

	Semantic	Meteor	Rouge-L (↑)		
Model	Score(↑)	Score(↑)	Precision	Recall	F-score
Base	0.6525	0.4075	0.4335	0.4301	0.4318
Base (w self-attention)	0.6701	0.4215	0.4292	0.4508	0.4393
Aggregated gaze	**0.7826**	**0.5033**	**0.5193**	**0.5644**	**0.5405**

create the binary heatmap and the gaze-overlaid image. To assess the impact of the gaze point size on model performance, we conducted experiments where we increased the size of the overlaid gaze points on the heatmap. Our findings indicated that performance improved with larger gaze point overlays in the singular gaze model.

In contrast, the aggregated gaze model, which combines multiple gaze points and includes an occlusion check, did not exhibit significant performance gains when the size of the overlaid points was increased. Table 5 highlights the results of this experiment. The results suggest that the aggregated gaze model's incorporation of multiple gaze points and an occlusion check is more effective than merely enlarging individual gaze points in the singular model.

8.3.2 Inclusion of self-attention block in base model without gaze

Building on the insights from the previous section, a natural question arises: what happens if we employ a self-attention block in the base model, providing the largest possible overlay—the entire image itself? In this experiment, we integrated a self-attention block into the base model without gaze. The features obtained from the ViT were passed into the self-attention block, where the queries, keys, and values were derived from these image features.

The objective of this modification was to investigate the effects of using a full image overlay as the query input to the attention block in the base model. The resultant features from the self-attention block were then forwarded to the perceiver resampler for further processing. Our results showed that while the performance of the base model improved with the inclusion of the self-attention block, it still remained below that of the gaze-augmented models, as shown in Table 6. This indicates that although leveraging the entire image enhances the base model's capabilities, it does not fully match the performance benefits achieved by incorporating eye gaze signal and using gaze-regularized attention mechanism.

8.3.3 Using gaze in text form

In our studies, we converted gaze data from coordinate text form into heatmaps, which were then overlaid on scene images to create a more visual representation. This transformation allows us to highlight important visual regions and aligns more closely with how humans perceive spatial information. (Laeng et al., 2014). To conduct a sanity check and assess whether using gaze data in visual form is more suitable than using gaze in text form , we trained a gaze-augmented model that utilizes gaze coordinates as text input. Let $\phi_{gaze,text}$ represent the gaze augmented VLM which aims to model the likelihood of the fine grained text descriptions (represented by ℓ_i) when eye gaze

Table 7: Effect of using gaze information in text form and comparison with other models

	Semantic	Meteor	Rouge-L (↑)		
Model	Score(↑)	Score(↑)	Precision	Recall	F-score
Base	0.6525	0.4075	0.4335	0.4301	0.4318
Aggregated gaze(in text form)	0.7021	0.4428	0.4642	0.4621	0.4630
Aggregated gaze	**0.7826**	**0.5033**	**0.5193**	**0.5644**	**0.5405**

information in the form of text coordinates $\{(x, y)\}$ is also provided:

$$\phi_{gaze,text}(\ell, \{\mathrm{I}\}, \{(x, y)\}) = \prod_{i=\tau_o+1}^{\tau_o+\tau_a} p(\ell_i \mid \ell_{<i}, \mathrm{I}_{\leq\tau_o}, \{(x, y)\}_{\leq\tau_o}) \tag{14}$$

Our results indicated that using gaze data in text form improved performance compared to the base model without gaze. However, it still fell short of the performance achieved by the aggregated gaze-augmented model (as shown in Table 7). This experiment demonstrates that incorporating gaze information as a signal is essential. However, utilizing gaze in conjunction with the gaze-regularized attention mechanism significantly enhances model performance

8.3.4 Changing prediction and observation horizons

To investigate the effect of sequence length on model performance, we conduct experiments by adjusting the prediction horizon (τ_p) and observation horizon (τ_o). First, we extend the prediction horizon from 2 to 5 seconds to evaluate both the base and gaze-augmented models (using the aggregated gaze model with occlusion check). As shown in Table 8, the gaze-augmented model consistently outperforms the base model, even with a longer prediction horizon reinforcing the intuition that gaze is important to predict intentions and future actions. We also reduce the observation horizon to 3 seconds while keeping the prediction horizon fixed at 2 seconds. The gaze-augmented model again outperforms the base model (Table 9). Interestingly, the shorter observation horizon shows slightly increased performance compared to the main results reported with a longer observation horizon. This could indicate a potential improvement of our approach, by accounting for the visual working memory humans utilize.

Table 8: On increasing the prediction horizon (from 2 seconds) to predict future actions up to 5 seconds, the gaze-augmented model is able to outperform the base model.

	Semantic	Meteor	Rouge-L (↑)		
Model	Score(↑)	Score(↑)	Precision	Recall	F-score
Base	0.6297	0.3972	0.4393	0.4065	0.4220
Aggregated Gaze	**0.7519**	**0.4891**	**0.5332**	**0.5410**	**0.5386**

Table 9: On decreasing the observation horizon (from 5 to 3 seconds) to predict future actions up to 2 seconds, the gaze-augmented model is again able to outperform the base model.

	Semantic	Meteor	Rouge-L (↑)		
Model	Score(↑)	Score(↑)	Precision	Recall	F-score
Base	0.6716	0.4284	0.4544	0.4627	0.4581
Aggregated Gaze	**0.7855**	**0.5132**	**0.5399**	**0.5523**	**0.5457**

8.4 Limitations

In this section, we highlight the key limitations of our work, which is as follows:

- **Dataset Creation**: The dataset was created using an iterative prompt mechanism to refine a suitable template, which was then tested on a smaller subset of the data. Template approval was carried out by two individuals, but we acknowledge that the suitability of annotations can be subjective, potentially leading to varying results across different individuals. Due to the impracticality of manually verifying all annotations, we randomly checked 200 annotations from various videos in the dataset to ensure suitability. While mistakes were found and rectified, we acknowledge that the annotations may still contain imperfections. Additionally, comparisons with different modalities (e.g., audio) were not feasible, as not all gaze-equipped videos in the dataset had accompanying modalities. As a result, we did not propose a new benchmark but focused on demonstrating the importance of including eye gaze as an additional modality in a VLM to enhance its predictive capabilities.
- **Inference Speed Analysis**: A detailed study comparing the real-time inference speed of the baseline model and the gaze-augmented model could further strengthen our findings. However, this analysis was not conducted due to the lack of appropriate equipment.
- **SOTA Comparison**: Since our dataset and problem definition differ from prior studies, a direct comparison with state-of-the-art (SOTA) methods was not performed. Nonetheless, we emphasize that the primary contribution of this work is not the introduction of a new dataset or benchmark but rather the demonstration of how incorporating eye gaze into VLMs can improve their predictive tasks compared to relying solely on traditional inputs.

8.5 Training and evaluation details

Both the base model and the gaze-augmented model were trained using two NVIDIA A800 40GB GPU cards. The base model required approximately 36-38 hours for training, while the gaze-augmented model took around 50 hours. The training utilized a batch size of 32 and a learning rate of $7\mathrm{e}-5$. The vision encoder is pre-trained by OpenAI. Additionally, the tokenizer for the text annotations used was the OPT (Open Pre-trained Transformer) language model developed by Meta. To accelerate the training process, we employed Fully Sharded Data Parallel (FSDP), a technique that efficiently distributes model parameters and gradients across multiple GPUs, reducing memory usage and improving training speed. Data loading was managed using the WebDataset loader, and the dataset was converted to .tar files to align with the format required for integration with both the WebDataset loader and FSDP.

In terms of model complexity, the base model has approximately 944 million parameters. The gaze-augmented model with 5 layers contains about 996 million parameters while the gaze-augmented model with 1 layer has around 955 million parameters. Lastly, the model with 2 layers, which is the best performing model, has approximately 966 million parameters.

For evaluation, both models were assessed using the semantic transformer (SBERT) developed by Reimers & Gurevych (2019), along with ROUGE-L and Meteor scores. The inference time on the test set for the base model was approximately one hour, while the gaze-augmented model had a slightly longer inference time but remained comparable to that of the base model.