Learning Alignments of Human Gaze and Fine-grained Task Descriptions

TAKUMI NISHIYASU, The University of Tokyo, Japan
ZHIMING HU, The Hong Kong University of Science and Technology (Guangzhou), China
ANDREAS BULLING, University of Stuttgart, Germany
YOICHI SATO, The University of Tokyo, Japan

We propose GTANet – a novel approach to learning the alignments between human gaze scanpaths and *fine-grained* task descriptions in vision-language tasks. While the influence of tasks on gaze is well known, the relationship between gaze scanpaths and *fine-grained* task descriptions remains largely unexplored. GTANet addresses this gap by aligning encoded spatiotemporal gaze features with text descriptions. We utilize a patch-based gaze encoder to generate gaze features that reflect visual contexts, and a multimodal feature mixer to fuse the gaze features and the task descriptions, capturing cross-modal alignment. To validate our method, we introduce two novel tasks: gaze-to-question and question-to-gaze retrieval. Experiments on the AiR and MHUG datasets demonstrate that GTANet consistently outperforms baseline methods across all Recall@K metrics, achieving substantial improvements in both retrieval directions. These results confirm the strong link between human gaze and fine-grained task descriptions, thus validating the effectiveness of our approach.

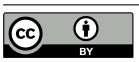
CCS Concepts: • **Computing methodologies → Computer vision**; **Artificial intelligence**; **Machine learning**; • **Human-centered computing → Human computer interaction (HCI)**; **Collaborative and social computing**.

ACM Reference Format:
Takumi Nishiyasu, Zhiming Hu, Andreas Bulling, and Yoichi Sato. 2026. Learning Alignments of Human Gaze and Fine-grained Task Descriptions. *Proc. ACM Comput. Graph. Interact. Tech.* 9, 2, Article pacmcgitv9n2-fp1031 (June 2026), 18 pages. https://doi.org/10.1145/3803535

1 Introduction

Modelling the relationship between human eye gaze and tasks that people perform while looking at images is a long-standing research challenge in computer vision and human-centred computing [Borji et al. 2012; Sood et al. 2023, 2021, 2020; Sugano and Bulling 2016; Yarbus 1967]. Many gaze-task alignment models have been explored [Borji et al. 2012; Coutrot et al. 2018; Hu et al. 2021a,b; Koulieris et al. 2016] and have mainly been used in two ways: (1) "task-to-gaze" – modelling task-related gaze patterns that can be leveraged for performance assessment or scanpath prediction, for example, as discussed in [Borji et al. 2012; Hu et al. 2021b; Koulieris et al. 2016], and (2) "gaze-to-task" – decoding gaze behaviour to infer task semantics, which can faciliate task recognition or user intent estimation as shown in [Karessli et al. 2017; Liao et al. 2019; Sattar et al. 2015]. Similarly, understanding how human gaze scanpaths reflect task descriptions can improve

Authors' Contact Information: Takumi Nishiyasu, The University of Tokyo, Tokyo, Japan, nisiyasu@iis.u-tokyo.ac.jp; Zhiming Hu, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, cranehzm@gmail.com; Andreas Bulling, University of Stuttgart, Stuttgart, Germany, andreas.bulling@vis.uni-stuttgart.de; Yoichi Sato, The University of Tokyo, Tokyo, Japan, ysato@iis.u-tokyo.ac.jp.



ACM 2577-6193/2026/6-ARTpacmcgitv9n2-fp1031
https://doi.org/10.1145/3803535

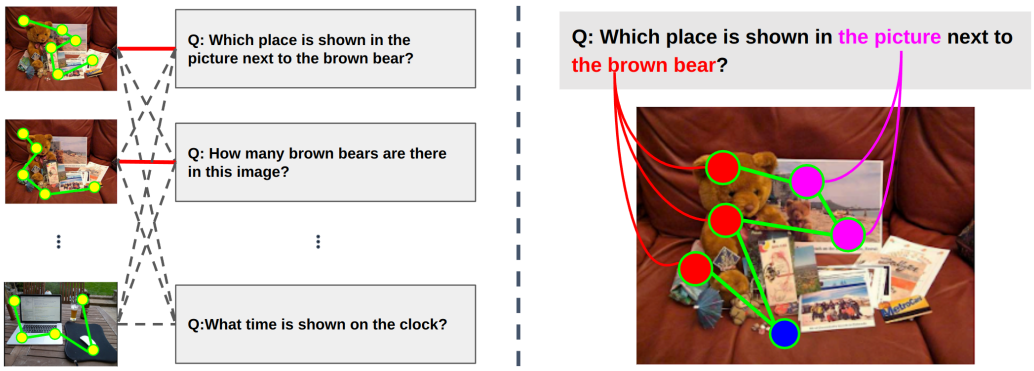


Fig. 1. Overview of Gaze-Task Alignment. The left side illustrates how human gaze scanpaths on images align with fine-grained task descriptions (questions) in Visual Question Answering (VQA). Strong alignment is indicated by solid red lines, while weaker or incorrect alignment is shown by dashed lines. The right side presents a qualitative example demonstrating the strong semantic and spatial grounding between the gaze scanpath and the specific keywords in the question (e.g., "brown bear," "picture next to"). GTANet is the first method designed to learn unified representations of human gaze and fine-grained task descriptions for improved cross-modal inference.

downstream applications such as automated usability analysis [Eger et al. 2007], cognitive workload estimation [Fridman et al. 2018], or assistive intelligent systems [Lang et al. 2019].

However, previous studies have primarily explored the relationship between gaze and *coarse* task descriptions, which are word-level descriptions or categorical labels that only classify tasks into broad categories, such as navigation [Hadnett-Hunter et al. 2019; Kothari et al. 2020; Liao et al. 2019] or visual search [Barz et al. 2020; Borji et al. 2015; Hu et al. 2021a; Nishiyasu and Sato 2024; Sattar et al. 2015; Stauden et al. 2018]. These coarse descriptions often only include high-level attributes, such as task type (e.g., freeview, visual search) or search target object categories (e.g., searching for "mouse" or "laptop"). While these coarse descriptions can generally describe gaze-task relationships, they fail to capture the detailed situational context that drives human attention. For instance, a coarse task description like "searching for laptop" cannot distinguish whether the user is simply looking for the object category or confirming a specific detail, such as whether "the laptop is next to the plate on the wooden desk."

In contrast, *fine-grained* task descriptions can provide detailed information required to perform a visual task. For example, in a Visual Question Answering (VQA) task, a coarse description might be a simple instruction like 'Answer the Question,' whereas a fine-grained description is the actual question itself, such as 'What color is the car on the right side of the image?'. Such specific descriptions necessitate highly targeted visual attention and influence gaze scanpaths significantly, enabling a more detailed understanding of gaze-task relationships. To better understand how gaze behaviour relates to such fine-grained task descriptions, it is crucial to examine their cross-modal alignment.

Understanding and leveraging this relationship opens up multiple applications. Firstly, it enables the inference of task intent from gaze behaviour, providing insights into how individuals process visual information in different contexts. Secondly, it facilitates task performance and engagement assessment by identifying task-relevant gaze patterns, which can be used to evaluate cognitive effort and user interaction quality [Fridman et al. 2018]. Additionally, leveraging gaze-task alignment can benefit assistive AI systems [Lang et al. 2019] by enabling adaptive interfaces that dynamically

respond to user attention and intent. Despite these potential applications, fine-grained cross-modal alignment remains largely unexplored, highlighting the need for further investigation.

To fill this gap, we propose GTANet– a novel method to learn the cross-modal alignments between human gaze scanpaths and *fine-grained* task descriptions. As illustrated in Figure 1 (left), the goal is to accurately distinguish strong alignment between gaze and task (solid red line) from weak or incorrect alignments (dashed lines). Our method models this distinction through a two-stage process built upon two core components: the Patch-based Gaze Encoder and the Multimodal Feature Mixer . The gaze scanpath is first processed by our novel Patch-based Gaze Encoder to extract spatially and temporally enriched features. This encoder achieves the goal by explicitly grounding the temporal gaze sequence in the visual context of the fixated image patches. This creates visual-context-aware gaze representations highly sensitive to *what* and *when* the user was looking, thereby ensuring the gaze features accurately reflect the visual evidence attended. Subsequently, these extracted gaze features are input into a Multimodal Feature Mixer along with the task text and image features. The Mixer utilizes a self-attention mechanism to explicitly learn the detailed dependencies and influences between the gaze, image, and task semantics. This fusion mechanism allows the model to identify the precise semantic correspondence between the observed attention (gaze) and the required semantic content (task), enhancing the model's ability to align gaze behavior precisely with task-relevant content for final compatibility scoring, as conceptually shown in Figure 1 (right).

To evaluate our approach, we propose two new retrieval tasks that enable a systematic and quantitative evaluation of this fine-grained gaze-task alignment: gaze-based question retrieval and question-based gaze retrieval. These tasks allow us to measure the model's ability to retrieve the correct task description given a gaze scanpath, and conversely, to retrieve the expected gaze scanpath for a given task description, serving as valuable benchmarks for future research. Experimental results on the AiR [Chen et al. 2020b] and MHUG [Sood et al. 2021] datasets show that GTANet significantly and consistently outperforms all baseline methods for gaze-based question retrieval across all Recall@K metrics. Specifically, on the challenging AiR dataset, our method achieves an improvement of over 56% in gaze-based question retrieval compared to the best performing baseline method. Similarly, on the MHUG dataset, GTANet achieves an improvement of over 38% in $R@1$ compared to the best baseline method. Furthermore, GTANet demonstrates highly competitive performance in question-based gaze retrieval, showing that modeling gaze-task alignment significantly enhances cross-modal inference by strengthening the gaze-task correspondence.

Our main contributions are summarised as follows:

(1) We propose GTANet, a novel method for learning the alignments between gaze scanpaths and fine-grained task descriptions. We demonstrate that our Patch-based Gaze Encoder and Multimodal Feature Mixer effectively enhance the alignment between gaze scanpaths and task descriptions.
(2) We introduce a new evaluation framework for assessing gaze-task alignments, incorporating two retrieval-based tasks: gaze-based question retrieval and question-based gaze retrieval. By measuring retrieval performance using ranking-based metrics, we quantitatively assess how well gaze scanpaths align with task descriptions and evaluate the model's ability to distinguish meaningful gaze-task relationships.

2 Related Work

2.1 Visual Attention and Task Prediction

Predicting visual attention has been an active area of research, focusing on modelling the sequence of eye movements (scanpaths) that individuals follow when viewing visual stimuli. Scanpath

prediction has been explored in both free-viewing settings [Assens et al. 2018; Assens Reina et al. 2017; de Belen et al. 2022; Sui et al. 2023] and task-driven scenarios [Chen et al. 2021, 2024; Mondal et al. 2023; Qiu et al. 2023]. In particular, several models have been proposed for predicting human attention in visual search and question-answering tasks, leveraging gaze behaviour to understand attention allocation [Chen et al. 2021, 2024; Ilaslan et al. 2023; Mondal et al. 2023; Qiu et al. 2023; Sood et al. 2023; Wang et al. 2024b].

Beyond predicting scanpaths, gaze has been widely used for task and activity recognition. Yarbus [Yarbus 1967] demonstrated that eye movements reveal an observer's task, a concept that has been further explored in task classification based on gaze features [Borji and Itti 2014; Haji-Abolhassani and Clark 2014]. Activity recognition from gaze has also been studied using first-person videos and wearable eye trackers [Bulling et al. 2010], with more recent work focusing on task inference in VR environments [Hu et al. 2021a]. Additionally, predicting target categories from gaze patterns has gained attention, with researchers aiming to infer target objects in visual search and natural scenes [Barz et al. 2020; Borji et al. 2015; Nishiyasu and Sato 2024; Sattar et al. 2015; Stauden et al. 2018].

Previous studies have framed task prediction as category-level inference using coarse task descriptions, such as activity categories or object attributes. These coarse descriptions often only include high-level attributes, such as task type (e.g., freeview, visual search) or search target object categories (e.g., searching for "mouse" or "laptop"). However, these descriptions fail to capture the key factors that shape gaze behaviour. Our approach moves beyond these predefined categories by introducing *fine-grained* task descriptions in natural language. This allows us to learn the precise cross-modal alignments between the dynamic human gaze scanpaths and the detailed task semantics, providing a significantly more comprehensive understanding of gaze-task relationships.

2.2 Multimodal Feature Alignment

Feature alignment across different modalities has been widely studied, particularly in the context of image-text learning modalities [Cao et al. 2022; Chen et al. 2020a; Frome et al. 2013; Karpathy and Li 2015; Li et al. 2019] and audio-visual learning [Arandjelovic and Zisserman 2017, 2018]. Foundational works in image-text alignment have mapped visual and textual features into a shared space using techniques such as hard negative mining [Faghri et al. 2017] and cross-attention [Chen et al. 2020a; Lee et al. 2018]. More recent approaches use large-scale pre-training to achieve state-of-the-art performance in tasks such as image captioning and VQA [Chen et al. 2020c; Li et al. 2023, 2022, 2021]. In audio-visual learning, researchers exploit the natural synchronization between vision and sound [Arandjelovic and Zisserman 2017, 2018], using ambient sounds for scene analysis [Owens and Efros 2018] and self-supervised methods to enhance alignment [Morgado et al. 2020].

Inspired by these prior works, we develop a multimodal alignment framework for gaze-task alignment. This framework is designed to effectively capture the correspondence between fine-grained task descriptions and gaze scanpaths, allowing us to quantify their semantic alignment. To achieve this, we use contrastive learning to align paired gaze-task representations while separating non-matching pairs.

2.3 Gaze-Assisted Learning in VLMs

Gaze information has been used as an auxiliary modality to enhance model performance in various vision-language tasks. In the context of image captioning, for example, studies have demonstrated that incorporating human gaze can enhance attention mechanisms, refine attention maps, and generate more accurate captions [Alahmadi and Hahn 2022; Chen and Zhao 2018; Cornia et al. 2018; Das et al. 2017; He et al. 2019; Sugano and Bulling 2016; Takmaz et al. 2020; Tavakoli et al. 2017]. Gaze information has also been used in medical image analysis to improve visual-textual

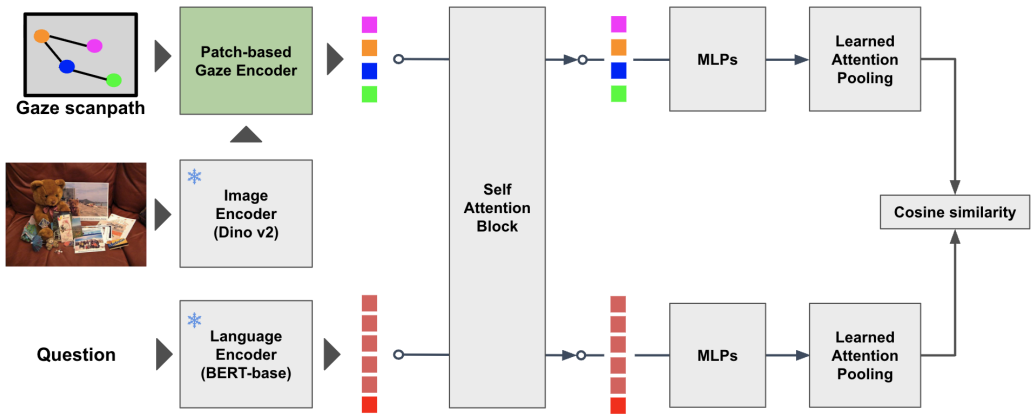


Fig. 2. GTANet architecture overview. Our method takes raw gaze scanpath data, a question, and an image as input. The image and question are processed by pretrained image and text encoders, respectively. A novel patch-based gaze encoder (detailed in Figure 3) is employed to extract gaze features. These features then undergo cross-modal interaction within the Self Attention Block to learn fine-grained correspondences between gaze and task (question) representations. The resulting features are further refined by MLPs and aggregated via Learned Attention Pooling. Finally, the alignment score is computed using cosine similarity.

alignment. Studies have shown that incorporating gaze enhances radiology report generation [Kumar and Marttinen 2024; Ma et al. 2024; Peng et al. 2024; Sultana et al. 2024] . Other research has explored integrating gaze into multimodal graph neural networks (GNNs), demonstrating improved performance in gaze-assisted Chest X-ray classification [Wang et al. 2024a] .

Unlike previous studies, which used gaze as an additional signal to improve vision-language tasks, our work focuses on using gaze as the main source for retrieving fine-grained task descriptions. Furthermore, we examine the bidirectional relationship between gaze and task descriptions within vision-language models, a topic that has not received much attention in previous research.

3 Learning Gaze-Task Alignment

3.1 GTANet

Overview. We formulate the problem as learning the alignment between gaze scanpaths and fine-grained task descriptions, given an image. The input consists of a gaze scanpath $G = \{(x_i, y_i, d_i)\}_{i=1}^{n_g}$, where n_g is the number of fixation points and (x_i, y_i, d_i) represents the fixation coordinates and duration, a task description T, which is a sequence of n_t tokens, and an image $I \in \mathbb{R}^{H \times W \times C}$. The output is an alignment score $S(G, T, I)$, which quantifies the alignment between G and T. The core objective is to learn this alignment score $S(G, T, I)$, which captures the semantic alignment—the degree to which the visual information attended by the gaze sequence corresponds to the content expressed by the task description. This score is subsequently used to rank gaze-task pairs based on their correspondence. The model achieves this by projecting gaze and task descriptions alongside image representations, where the image content serves as auxiliary context for gaze encoding. The model is trained to assign higher scores to correct gaze-task alignment by optimizing a contrastive loss while penalizing mismatched pairs. During inference, the computed score $S(G, T, I)$ is used to rank candidate gaze-task pairs based on their correspondence.

Figure 2 shows an overview of our method, GTANet, which implements this alignment learning. The process involves two main stages: feature encoding and cross-modal mixing. First, during

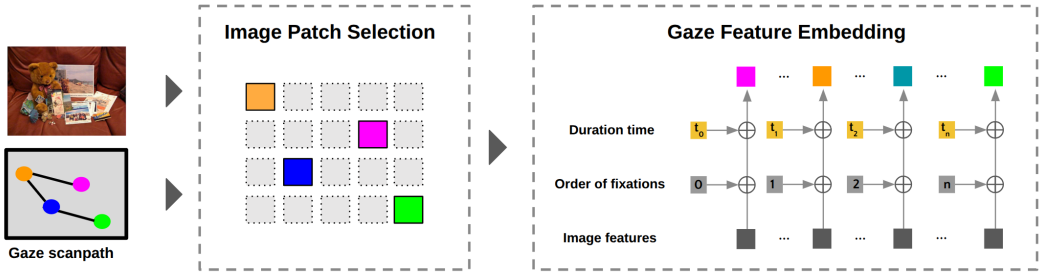


Fig. 3. Architecture of our Patch-based Gaze Encoder. Given an input image and a gaze scanpath, the encoder first selects image patch features that correspond to each fixation location on the image. Simultaneously, gaze-specific information, such as fixation order and duration, is encoded and combined with these selected image patch features, forming enriched gaze representations. These combined features serve as the input for subsequent gaze-task modeling, effectively integrating spatial visual context with temporal gaze dynamics.

the feature encoding stage, the image, task description, and gaze scanpath are processed by their respective encoders. An image feature extractor processes the image, and a language model encodes the task description into text features, Z_T. Crucially, the gaze scanpath is embedded using a Patch-based Gaze Encoder. This encoder captures the spatiotemporal properties of fixations by projecting them onto corresponding image patches, thereby generating visual-context-aware gaze features, Z_G. Second, to model the cross-modal interaction, the encoded gaze features Z_G and task features Z_T are input into a Multimodal Feature Mixer. This mixer uses a self-attention mechanism to fuse the modalities, allowing each feature token to attend to all others and capturing the underlying correspondence. Finally, the refined representations from the mixer are processed through separate Multilayer Perceptrons (MLPs) and a learned attention pooling mechanism to produce compact, fixed-length embeddings. The final alignment score is then computed by measuring the cosine similarity between the final gaze and task feature embeddings, which is optimized using a contrastive loss such as InfoNCE [Oord et al. 2018].

Image and Text Encoder. To establish robust feature representations across all modalities, we utilize state-of-the-art pre-trained models. We use BERT-base [Devlin et al. 2019] to encode text-based task descriptions T, which are questions in our formulation as shown in Figure 2. We extract a fixed number of token embeddings from the last hidden layer. This approach ensures a consistent, high-dimensional semantic representation of the task description regardless of the input length. We use DINOv2 (ViT-B/14) [Oquab et al. 2023] for encoding the image I. Crucially, we obtain features from the final transformer layer by using the patch tokens instead of the global [CLS] token. This is a deliberate design choice: by using the patch tokens, we preserve fine-grained spatial information across the image grid. This patch-level representation is essential as it serves as the input dictionary for the Image Patch Selection mechanism within our Patch-based Gaze Encoder. This allows the gaze fixations to be directly mapped to the corresponding visual features they attended.

Patch-based Gaze Encoder. We propose a novel Patch-based Gaze Encoder that effectively transforms raw gaze scanpaths into structured, context-aware embeddings. While our approach draws inspiration from methods that encode gaze through image regions [Nishiyasu and Sato 2024; Yang et al. 2024], our encoder presents a novel architecture that comprehensively integrates image patch features attended by the gaze with specific gaze dynamics (fixation duration and order), as

illustrated in Figure 3. This encoder is crucial for creating a compact gaze representation Z_G suitable for cross-modal alignment.

Encoding Process: For an input gaze scanpath $G = \{(x_i, y_i, d_i)\}_{i=1}^{n_g}$, where (x_i, y_i) are the coordinates and d_i is the duration of the i-th fixation, the encoding proceeds in two main steps:

(1) **Image Patch Selection (IPS)**: Each fixation point (x_i, y_i) is mapped to its corresponding spatial location within the grid of image patches, Z_I. Specifically, we utilize a 14×14 grid of patch embeddings extracted from a 224×224 image via ViT-B/14. Although spatially coarser than foveal acuity, these embeddings capture global context through self-attention, ensuring sufficient semantic detail despite the lower resolution. This step selects the visually attended patch feature P_i for the i-th fixation.

(2) **Gaze Feature Embedding (GFE)**: The selected image feature P_i is integrated with two essential gaze-specific features as shown in the right panel of Figure 3: Gaze Duration Embedding (GDE) and Gaze Sequential Embedding (GSE).

Inspired by absolute positional encoding methods [Vaswani et al. 2017], GDE and GSE are generated using sine and cosine functions to encode the fixation duration d_i and the sequential index i (order of fixation), respectively. We incorporate these embeddings directly into the selected image patch feature P_i. The final integrated embedding E_i for the i-th fixation is computed as:

$$E_i = P_i + \text{GDE}(d_i) + \text{GSE}(i) \tag{1}$$

The fixation-level embeddings $\{E_i\}_{i=1}^{n_g}$ collectively form the Gaze Feature Embedding Z_G. This sequence is then directly passed to the Multimodal Feature Mixer for cross-modal interaction with the text features. This architecture ensures that the gaze representation is contextually rich, reflecting not only *where* the human looked, but also the *duration* and the *order* of attention, grounded within the visual scene.

Multimodal Feature Mixer. To capture the underlying semantic alignment between the gaze scanpath and the task description, we use a Self-Attention Block to model the cross-modal interaction between the encoded features of the two modalities. Given the sequence of gaze features $Z_G \in \mathbb{R}^{n_g \times D}$ and task features $Z_T \in \mathbb{R}^{n_t \times D}$ where n_g and n_t are the sequence lengths and D is the feature dimension, we first concatenate these sequences along the token dimension to obtain $Z_{GT} = [Z_G; Z_T] \in \mathbb{R}^{(n_g+n_t)\times D}$. The concatenated sequence is then passed through a Self-Attention Block as shown in Figure 2 to generate contextually enriched feature sequences, $\hat{Z}_{GT}$. The Self-Attention Block follows a standard Transformer encoder architecture. It consists of a Multi-Head Self-Attention (MHSA) module followed by a Feed-Forward Network (FFN). Layer normalization (LayerNorm) is applied before each module, and residual connections are employed after each block. Specifically, the input Z_{GT} serves as the Query, Key, and Value ($Q = K = V = Z_{GT}$) for the MHSA mechanism. This comprehensive structure allows each token within the combined sequence to attend to all other tokens, explicitly modeling the mutual dependencies and semantic correspondence between the two modalities:

$$\hat{Z}_{GT} = \text{SelfAttentionBlock}(Z_{GT}) \tag{2}$$

After processing by the mixer, the enriched features $\hat{Z}_{GT}$ are split back into their respective sequences: $\hat{Z}_G \in \mathbb{R}^{n_g \times D}$ and $\hat{Z}_T \in \mathbb{R}^{n_t \times D}$.

Alignment Scoring. To obtain compact, fixed-length embeddings for compatibility scoring, the sequences $\hat{Z}_G$ and $\hat{Z}_T$ are passed through separate two-layer Multilayer Perceptrons (MLPs) and then subjected to a learned pooling mechanism. Specifically, we employ a *Learned Attention Pooling* mechanism. Unlike mean pooling, this layer dynamically aggregates features based on their

semantic importance. Formally, for an input sequence of feature vectors $H = \{h_1, \dots, h_L\} \in \mathbb{R}^{L \times D}$ (representing either the gaze or text sequence), we introduce a learnable weight vector $\mathbf{w} \in \mathbb{R}^D$. The attention score α_i for the i-th token and the final pooled representation z are computed as:

$$\alpha_i = \frac{\exp(\mathbf{w}^\top h_i)}{\sum_{j=1}^{L} \exp(\mathbf{w}^\top h_j)}, \quad z = \sum_{i=1}^{L} \alpha_i h_i \tag{3}$$

This mechanism produces the final fixed-length embeddings z_G and z_T. Finally, the gaze-task alignment score $S(G, T, I)$ is computed as the cosine similarity between these normalized embeddings: $S(G, T, I) = \mathit{cosine\_similarity}(z_G, z_T)$ This alignment score is used for ranking in retrieval-based tasks.

3.2 Alignment Objective

The model is trained using the InfoNCE loss [Oord et al. 2018], which is designed to maximize the probability that a positive (correct) pair is correctly identified among a specific set of explicitly generated negative pairs. This loss enforces the post-mixing embeddings z_G and z_T to lie close in the latent space if they correspond to a matched gaze-task pair, and far apart otherwise. The loss function $\mathcal{L}$ is computed based on the negative log-likelihood of the probability that the correct text T_i matches the gaze G_i. We utilize the correct score $S_i = S(G_i, T_i, I)$, a set of N_q explicit negative text scores $S_{i,j}^{T'} = S(G_i, T'_j, I)$, and a set of N_g explicit negative gaze scores $S_{i,k}^{G'} = S(G'_k, T_i, I)$. The probability of the matching text T_i for a given gaze G_i is defined using the softmax function over the positive score and all explicit negative scores:

$$P(T_i|G_i) = \frac{\exp(S_i/\tau)}{\exp(S_i/\tau) + \sum_{j=1}^{N_q} \exp(S_{i,j}^{T'}/\tau) + \sum_{k=1}^{N_g} \exp(S_{i,k}^{G'}/\tau)} \tag{4}$$

where τ is the temperature hyperparameter. The total InfoNCE loss $\mathcal{L}$ is the negative log-likelihood averaged over the batch for both the gaze → text ($G \to T$) and text → gaze ($T \to G$) directions:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^{B} [\log P(T_i|G_i) + \log P(G_i|T_i)] \tag{5}$$

This formulation directly reflects the implementation where only the explicit negatives are used in the InfoNCE denominator for each anchor, ensuring the model learns robust alignments by distinguishing the true match from a diverse, explicitly curated set of incorrect pairs.

4 Experiments

To evaluate the proposed model, we performed experiments on **Gaze-to-Question Retrieval** and **Question-to-Gaze Retrieval** tasks on two publicly available datasets [Chen et al. 2020b; Sood et al. 2021].

Evaluation Tasks. We define two core retrieval tasks to evaluate the learned Gaze-Task Alignment. Both tasks require the model to retrieve the correct counterpart, which is either the question or the gaze scanpath, from a pool of negative candidates based on the computed alignment score $S(G, T, I)$.

(1) **Gaze-to-Question Retrieval ($G \to T$)**: Given a ground truth gaze scanpath G and an image I as the query, the model ranks a set of candidate questions $\{T, T'_1, T'_2, \dots\}$ based on the alignment score $S(G, T', I)$. This task evaluates the model's ability to infer the high-level task intention from the observed attention behavior.

(2) **Question-to-Gaze Retrieval ($T \rightarrow G$)**: Given a task description T and an image I as the query, the model ranks a set of candidate gaze scanpaths $\{G, G'_1, G'_2, \dots\}$ based on the alignment score $S(G', T, I)$. This task evaluates the model's ability to predict plausible human attention patterns for a given task.

Evaluation Metrics. For evaluation metrics, we measure performance using R@K, a standard metric in information retrieval [Karpathy and Li 2015]. R@K is defined as the fraction of queries for which the correct counterpart is successfully retrieved among the top K ranked candidates. The total number of candidate items for each query is 30, consisting of the correct item plus $N_q = 29$ or $N_g = 29$ negative examples. Following common practice in vision-language retrieval, we report R@1, R@5, and R@10.

4.1 Dataset Preparation

Datasets Overview. We use the AiR [Chen et al. 2020b] and MHUG [Sood et al. 2021] datasets, which contain human gaze scanpaths recorded while answering vision-language tasks. Each sample in these datasets consists of an image, a corresponding question, and the recorded eye movements of participants as they viewed the image and processed the question. The questions follow a visual question answering format, requiring reasoning about image content in relation to text. The AiR dataset contains approximately 13,000 gaze samples collected from 20 participants, while the MHUG dataset contains approximately 12,000 gaze samples collected from 49 participants. We follow the standard data splits defined by the original papers. Crucially, we ensure that there is no overlap of images between the training, validation, and test sets.

Generating Negative Question Examples. To enhance the model's robustness and its ability to distinguish between correct and incorrect task intents, we introduce hard negative samples by generating a specialized set of dummy questions. These dummy questions serve as distractors, helping the model learn to differentiate between relevant and irrelevant queries. We generated 10 new VQA pairs for each ground truth (GT) sample using the Gemini 2.0 Flash. For this process, the Gemini 2.0 Flash was supplied with the original image where a mask was applied to the regions corresponding to the ground truth gaze fixation areas. This masking strategy ensures that the generated questions could not be answered by observing the objects or regions that the participant had fixated upon for the original GT question. This highlights the importance of learning the specific alignment between the GT gaze scanpath and the original question, as the generated negative questions are visually plausible and solvable even without the GT gaze information, making them difficult to distinguish from the original question based on image content alone. The specific prompts and constraints used for generation can be found in the appendix **??**.

In addition to these 10 MLLM-generated questions, we augment the negative set by randomly selecting 19 ground truth questions from other samples within the same batch (in-batch negatives). This results in a total of $N_q = 29$ negative questions for each anchor pair.

Generating Negative Gaze Scanpath Examples. Negative gaze scanpaths are generated by randomly selecting gaze data from other samples within the same batch (in-batch negatives). Each ground truth scanpath is paired with a set of $N_g = 29$ randomly chosen negative scanpaths from different instances, ensuring diversity. No constraints on task or image similarity are applied. This setup provides a balanced set of negative examples for the contrastive objective.

4.2 Baselines

As existing studies have not explicitly addressed the alignment between fine-grained task descriptions and gaze scanpaths, we adopt a hybrid gaze encoding technique from prior works

on gaze-based task recognition [Hu et al. 2021a] as one baseline, and introduce three structural cross-modal baselines for rigorous comparison. First, we evaluate the hybrid sequential model: **1DCNN+BiGRU** [Hu et al. 2021a], which first uses 1DCNN for local pattern extraction before employing BiGRU for temporal dependencies. The comparison between this hybrid baseline and our method allows us to validate the effectiveness of integrating visual context directly into the gaze feature extraction process. In addition to this sequential model, we introduce three dedicated cross-modal baselines whose architectures process Gaze and Task features independently, and perform alignment directly via final similarity comparison, without sequence-level interaction. The Task Description in all three is represented solely by the **CLS** token $\mathbf{z}_T$ from the BERT encoder. The first of these is **GST** [Nishiyasu and Sato 2024], which utilizes the Gaze Search Transformer (GST) structure for the gaze branch and employs the original GST input reliant on panoptic segmentation results for feature aggregation. To ensure a fairer comparison using the same visual features as GTANet, the second model is **GST (Image)**, a variant that replaces the panoptic segmentation input of GST with the output of our shared Image Encoder. Finally, to assess the unique contribution of gaze information, we introduce the **Image-Task Baseline**, which excludes the gaze scanpath entirely. In this case, the Image Encoder's patch features are directly aggregated via the Self-Attention Block and pooling into a single vector $\mathbf{z}_I$, which is then aligned with the Task $\mathbf{z}_T$. Comparing the GST-based baselines with GTANet allows us to isolate the benefit of our dedicated Mixer for cross-modal token-level interaction, while comparing the Image-Task Baseline quantifies the exact performance gain attributable to the integration of human gaze attention.

4.3 Implementation Details

The input image resolution is set to $(C, H, W) = (3, 224, 224)$, and image features are extracted as 14×14 patch embeddings. Text features, with a maximum sequence length of $n_t = 20$, and gaze scanpaths, with a maximum length of $n_g = 20$, are projected into a common feature dimension of $D = 128$. The final Gaze and Task representations are fixed to this dimension $D = 128$ for computing the alignment score via cosine similarity. To ensure robust evaluation, we applied dataset-specific splitting strategies. For the AiR dataset, we adopted the official training, validation, and testing splits provided by the authors. For the MHUG dataset, which lacks a standardized split, we partitioned the data into training, validation, and testing sets with an approximate ratio of 8:1:1. When constructing this split, we strictly enforced that no identical questions appear across the splits. Furthermore, we minimized the overlap of images between the training and testing sets to evaluate the model's robustness under diverse visual contexts. For training, we use the Adam optimizer with a learning rate of 2×10^{-4}, managed by a Cosine decay scheduler. The model is trained for 20 epochs using a mini-batch size of 32. To mitigate overfitting, a dropout rate of 0.20 is applied to the feature mixer blocks. Following standard contrastive learning practices, the temperature parameter τ is set to 0.07 for the InfoNCE loss. The model checkpoint with the highest validation score was selected for final testing. All experiments were performed on a single NVIDIA TITAN RTX. Training took approximately 10 hours, and inference requires 635 ms per sample with 1.25 M trainable parameters.

5 Results

5.1 Retrieval Performance Analysis

Gaze-to-Question Retrieval. Our proposed method, GTANet, significantly and consistently outperforms all baseline approaches in gaze-to-question retrieval across all Recall@K metrics (R@1, R@5, and R@10) on both the AiR (Table 1) and MHUG (Table 2) datasets. The GTANet model

Table 1. Quantitative results of Question retrieval and Gaze retrieval on AiR [Chen et al. 2020b]. dataset in terms of Recall@K. The GTANet model significantly improves Question Retrieval performance compared to previous methods.

Model	Question Retrieval			Gaze Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Image-Task Baseline	0.2297	0.6494	0.8380	N/A	N/A	N/A
1DCNN+BiGRU [Hu et al. 2021a]	0.2441	0.6859	0.8760	0.0669	0.2875	0.4966
GST [Nishiyasu and Sato 2024]	0.1042	0.4814	0.7856	0.0395	0.1863	0.3658
GST (Image) [Nishiyasu and Sato 2024]	0.2160	0.6487	0.8684	0.4798	0.8525	0.9437
GTANet (ours)	**0.3810**	**0.7878**	**0.9300**	**0.5095**	**0.8631**	**0.9506**

Table 2. Quantitative results of Question retrieval and Gaze retrieval on MHUG [Sood et al. 2021] dataset in terms of Recall@K. Our proposed method (GTANet) significantly improves Question Retrieval metrics on this dataset.

Model	Question Retrieval			Gaze Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Image-Task Baseline	0.2435	0.5889	0.7691	N/A	N/A	N/A
1DCNN+BiGRU [Hu et al. 2021a]	0.2401	0.6028	0.7927	0.0858	0.3541	0.5671
GST [Nishiyasu and Sato 2024]	0.0933	0.3731	0.6189	0.0340	0.1819	0.3558
GST (Image) [Nishiyasu and Sato 2024]	0.2481	0.6195	0.8071	0.4917	0.8198	0.9229
GTANet (ours)	**0.3431**	**0.7513**	**0.8998**	**0.5112**	**0.8348**	**0.9234**

achieves the highest Recall@K values, unequivocally demonstrating its effectiveness in capturing the fine-grained alignment between gaze scanpaths and task descriptions.

Specifically, on the AiR dataset, GTANet achieves an $R@1$ of 0.3810, representing a substantial improvement of over 69% compared to the best performing cross-modal baseline, the Image-Task Baseline ($R@1$: 0.2297), and a 56% improvement over GST (Image) ($R@1$: 0.2160). This dominance extends to broader retrieval, with $R@5$ reaching 0.7878 and $R@10$ reaching 0.9300. The poor performance of the original GST model ($R@1$: 0.1042) highlights the limitations of relying on coarse, pre-processed features like panoptic segmentation for fine-grained alignment. Similarly, on the MHUG dataset, our model achieves an $R@1$ of 0.3431, outperforming the best performing baseline, GST (Image) ($R@1$: 0.2481), by over 38%.

The substantial gap between GST (Image) which uses the same image features as GTANet and GTANet highlights the necessity of the Multimodal Feature Mixer for capturing fine-grained, token-level dependencies. Furthermore, the low performance of the Image-Task Baseline clearly quantifies the essential contribution of the gaze scanpath itself for task understanding, which far surpasses methods relying solely on sequential processing (1DCNN+BiGRU).

Question-to-Gaze Retrieval. In the question-to-gaze retrieval task, the performance trends show a different structure, emphasizing the difficulty of accurately predicting raw gaze patterns.

On the AiR dataset, GTANet achieves the highest metrics across all K values, with $R@1$ reaching 0.5095 and $R@10$ reaching 0.9506. The GST (Image) baseline shows highly competitive performance, with $R@1$ of 0.4798, demonstrating that the combination of our Image Encoder features and the GST aggregation mechanism is highly effective for synthesizing a high-quality gaze representation

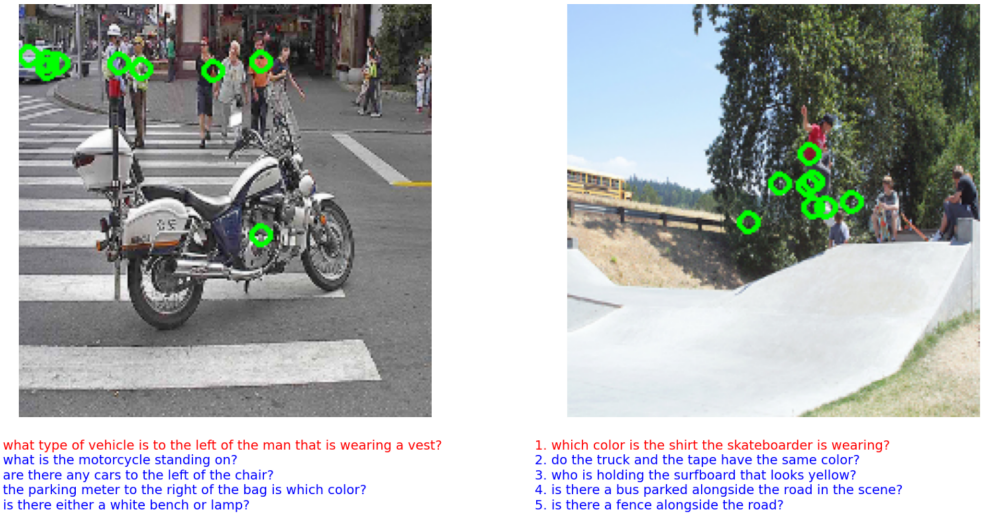


Fig. 4. Visualisation of gaze-to-question retrieval results. Each example includes an input image with gaze fixations (green dots) and the retrieved top5 questions. The ground-truth (GT) question is shown in red, and dummy questions are highlighted in blue. This visualisation demonstrates how the model retrieves semantically and spatially relevant questions based on gaze patterns (Best viewed in colour).

Fig. 5. Visualisation of question-to-gaze retrieval results. Each example consists of an input image with different gaze scanpaths. The two images in each pair represent the top2 retrieved gaze scanpaths based on the model's predictions. The red border indicates cases where the retrieved gaze matches the ground-truth (GT), while the blue border represents non-GT retrieved gaze scanpaths. This visualisation illustrates how the model retrieves gaze scanpaths that align with the input question and scene context (Best viewed in colour).

$\mathbf{z}_G$. Similarly, on the MHUG dataset, GTANet leads at $R@1$ (0.5112) and $R@5$ (0.8348), closely followed by GST (Image) ($R@1$: 0.4917).

The strong performance of GST (Image) suggests that for the one-way retrieval of a gaze scanpath, a high-quality, aggregated feature vector of the gaze is highly effective. However, GTANet's superior, balanced performance across both retrieval directions, especially the significantly dominant Gaze-to-Question task, emphasizes the importance of our Multimodal Feature Mixer. GTANet ensures robust, reciprocal alignment by integrating gaze, image, and task representations through structured attention, which is crucial for comprehensive inference and achieving state-of-the-art results.

Table 3. Effects of gaze embedding in patch-based gaze encoder on AiR dataset regarding Recall@K. IPS = Image Patch Selection, GFE = Gaze Feature Embedding.

Gaze Encoder	Question Retrieval			Gaze Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
No Gaze Embeddings	0.2707	0.7004	0.8958	N/A	N/A	N/A
IPS	0.3354	0.7719	**0.9384**	0.4935	0.8608	**0.9544**
Ours (IPS + GFE)	**0.3810**	**0.7878**	0.9300	**0.5095**	**0.8631**	0.9506

5.2 Visualisation and Analysis

Examples of Gaze-to-Question Retrieval. To further illustrate the effectiveness of our approach, we present qualitative examples of gaze-to-question retrieval. Figure 4 showcases cases where our model retrieves the top five most relevant questions based on gaze scanpath. In each example, the green circles indicate gaze fixations, and the retrieved questions are listed below the image. Among these, the ground-truth (GT) question is highlighted in red, while other retrieved questions are shown in blue. The results demonstrate strong semantic and spatial alignment between the retrieved questions and the given gaze scanpath. The model effectively captures the relationship between human attention and task-relevant content in the scene, retrieving questions corresponding to objects and regions. The presence of the GT question among the top-ranked retrievals further validates the effectiveness of our cross-modal feature alignment in associating gaze scanpaths with relevant task descriptions.

Examples of Question-to-Gaze Retrieval. Similarly, we analyse question-to-gaze retrieval results to evaluate how well our model retrieves gaze scanpaths based on the question. Figure 5 presents examples where the retrieved gaze scanpaths correctly highlight relevant image regions, demonstrating the model's ability to infer visual attention from textual queries. The model retrieves the top two gaze scanpaths for a given question in each example. The red-bordered image represents the GT gaze scanpath, while the blue-bordered image shows another retrieved gaze scanpath from the top-ranked candidates. Green circles indicate gaze fixation points. The results show that the retrieved gaze scanpaths often align well with the key visual elements related to the question, illustrating the effectiveness of our cross-modal alignment. These findings highlight the model's ability to capture human visual attention dynamics and retrieve meaningful gaze scanpaths for a given question. The strong relationship between the retrieved gaze and the question-relevant image regions further supports the validity of our approach.

5.3 Ablation Study

To evaluate the impact of different components in our model, we conduct an ablation study by selectively removing or modifying key features, specifically focusing on the gaze encoding method and the multimodal feature mixer.

Effect of Patch-Based Gaze Encoder. Table 3 presents the impact of the gaze embedding module in the patch-based gaze encoder. We compare the performance of using only Image Patch Selection (IPS) and our full encoder (IPS + GFE) against a baseline without gaze features.

The results demonstrate the necessity of explicit gaze encoding for effective Question Retrieval. Using only Image Patch Selection (IPS) significantly improves Question Retrieval (R@1: 0.3354) compared to the baseline without gaze embeddings (R@1: 0.2707). This suggests that focusing feature extraction only on fixated image regions is highly effective for linking gaze to task semantics.

Table 4. Effects of Multimodal Feature Mixer on the AiR dataset regarding Recall@K. SA = Self-Attention.

Fusion Method	Question Retrieval			Gaze Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
MLP + Regression	0.3072	0.7285	0.9125	0.4989	**0.8700**	0.9483
SA + Regression	0.3156	0.7559	**0.9468**	0.4989	0.8456	0.9399
Ours (SA + Similarity)	**0.3810**	**0.7878**	0.9300	**0.5095**	0.8631	**0.9506**

However, PS alone results in a slight decrease in Gaze Retrieval performance compared to the final model (R@1 drops from 0.5095 to 0.4935), indicating that raw patch features are less distinct for identifying the correct gaze path. Crucially, integrating our proposed Gaze Embedding (IPS + GFE) addresses this by incorporating localized gaze statistics (e.g., fixation order and duration). This addition dramatically boosts Question Retrieval performance, achieving the highest R@1 of 0.3810, while maintaining a competitive Gaze Retrieval score (R@1: 0.5095). This validates our gaze encoder design: the IPS component grounds the task in relevant visual regions, and the GFE component enriches the gaze representation with temporal and statistical depth, proving essential for accurate alignment with task descriptions.

Effect of Multimodal Feature Mixer. We evaluate the effectiveness of the Multimodal Feature Mixer using a self-attention (SA) mechanism combined with a cosine similarity score predictor, contrasting it against alternatives that use simpler fusion and different score prediction heads.

The results in Table 4 reveal a strong dependence of Question Retrieval performance on both the feature fusion mechanism and the choice of the score function. Our full model (Ours, utilizing Self-Attention (SA) and Cosine Similarity) achieves a superior Question Retrieval R@1 of 0.3810. Models employing a Regression score head, regardless of whether they use a simple MLP (0.3072) or an SA mixer (0.3156), show significantly lower Question Retrieval performance. Specifically, the difference between SA + Regression and Our model (SA + Similarity) is over 6.5% in R@1, confirming that the cosine similarity loss function is highly effective for establishing the alignment boundary in the contrastive learning setting. While the use of Self-Attention (SA) for feature mixing provides a small gain over MLP fusion (0.3156 vs. 0.3072) when regression is used, the largest benefit comes from combining SA with the similarity-based objective. In contrast, Gaze Retrieval performance is relatively robust across the tested fusion methods, suggesting that the gaze embedding itself is sufficiently distinct, and the fusion mechanism primarily serves to refine the task representation for better correspondence with the gaze modality.

5.4 Limitations

This study has several limitations. First, regarding generalizability, while this work successfully utilizes the VQA setting, extending the learned gaze-task alignment to broader eye-tracking paradigms remains an area for future investigation. Specifically, applying our framework to dynamic scenes or implicit visual search behaviors involves challenges beyond static image-based tasks. Additionally, current datasets linking human gaze scanpaths to fine-grained task descriptions are limited in scale and diversity, potentially constraining the model's adaptability to real-world scenarios. Second, from a behavioral perspective, our work primarily focuses on the *computational* alignment of gaze and language. While the model effectively retrieves task intents based on scanpaths, it does not explicitly explicate the underlying cognitive mechanisms—*why* specific gaze patterns

emerge for specific tasks. Bridging the gap between deep learning representations and behavioral interpretability remains a vital direction for future research.

5.5 Privacy and Ethics Statement

Our study demonstrates that high-level user intent and specific task goals can be readily inferred by aligning minimal gaze scanpath data with task descriptions. This capability highlights a severe privacy risk inherent in human attention data, underscoring the urgent need for robust safeguards. We urge the community to implement strong privacy-preserving techniques to protect users' cognitive intentions and ensure the responsible development of adaptive human-computer interaction (HCI) systems.

6 Conclusion

In this paper, we explored the alignment between human gaze scanpaths and fine-grained task descriptions in vision-language tasks. We introduced GTANet, a novel method designed to measure the alignment between these two modalities. To provide a concrete evaluation of the learned gaze-task alignment, we proposed two novel retrieval tasks: gaze-based question retrieval and question-based gaze retrieval. Extensive experimental results on the AiR and MHUG datasets consistently demonstrated that GTANet significantly outperforms baseline methods, strongly underscoring the importance and effectiveness of explicitly modeling the relationship between human visual behavior and task intent. Our findings highlight the potential of our cross-modal attention-based framework for accurately quantifying gaze-task alignment.

Acknowledgement

This work was supported by JST Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE), Grant Number JPMJAP2303, and SPRING Grant Number JPMJSP2108.

References

Rehab Alahmadi and James Hahn. 2022. Improve Image Captioning by Estimating the Gazing Patterns From the Caption. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1025–1034.

Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 609–617.

Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *Proc. European Conference on Computer Vision (ECCV)*. 435–451.

Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2018. PathGAN: Visual scanpath prediction with generative adversarial networks. In *Proc. European Conference on Computer Vision Workshops (ECCVW)*.

Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2017. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2331–2338.

Michael Barz, Sven Stauden, and Daniel Sonntag. 2020. Visual search target inference in natural interaction settings with machine learning. In *Proc. ACM Symposium on Eye Tracking Research and Applications (ETRA)*. 1–8.

Ali Borji and Laurent Itti. 2014. Defending Yarbus: Eye movements reveal observers' task. *Journal of vision* 14, 3 (2014), 29–29.

Ali Borji, Andreas Lennartz, and Marc Pomplun. 2015. What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. *Neurocomputing* 149 (2015), 788–799.

Ali Borji, Dicky N Sihite, and Laurent Itti. 2012. Probabilistic learning of task-specific visual attention. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 470–477.

Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. 2010. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33, 4 (2010), 741–753.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713* (2022).

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020a. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12655–12663.

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. 2020b. Air: Attention with reasoning capability. In *Proc. European Conference on Computer Vision (ECCV)*. 91–107.

Shi Chen and Qi Zhao. 2018. Boosted attention: Leveraging human attention for image captioning. In *Proc. European Conference on Computer Vision (ECCV)*. 68–84.

Xianyu Chen, Ming Jiang, and Qi Zhao. 2021. Predicting human scanpaths in visual question answering. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10876–10885.

Xianyu Chen, Ming Jiang, and Qi Zhao. 2024. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *Proc. European Conference on Computer Vision (ECCV)*. 314–333.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020c. Uniter: Universal image-text representation learning. In *Proc. European Conference on Computer Vision (ECCV)*. 104–120.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2 (2018), 1–21.

Antoine Coutrot, Janet H Hsiao, and Antoni B Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods* 50, 1 (2018), 362–379.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding (CVIU)* 163 (2017), 90–100.

Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. 2022. Scanpathnet: A recurrent mixture density network for scanpath prediction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 5010–5020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conference of the North American chapter of the association for computational linguistics (NAACL)*. 4171–4186.

Nicola Eger, Linden Ball, Robert Stevens, and Jon Dodd. 2007. Cueing retrospective verbal reports in usability testing through eye-movement replay. In *Proc. British Computer Society Conference on Human-Computer Interaction (BCS-HCI)*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

Lex Fridman, Bryan Reimer, Bruce Mehler, and William T Freeman. 2018. Cognitive load estimation in the wild. In *Proc. CHI Conference on Human Factors in Computing Systems (CHI)*. 1–9.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 26 (2013).

Jacob Hadnett-Hunter, George Nicolaou, Eamonn O'Neill, and Michael Proulx. 2019. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception (TAP)* 16, 3 (2019), 1–17.

Amin Haji-Abolhassani and James J Clark. 2014. An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision research* 103 (2014), 127–142.

Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. Human Attention in Image Captioning: Dataset and Analysis. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2021a. Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2021).

Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2021b. FixationNet: forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 27, 5 (2021), 2681–2690.

Muhammet Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Lim, and Mike Shou. 2023. GazeVQA: A Video Question Answering Dataset for Multiview Eye-Gaze Task-Oriented Collaborations. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 10462–10479.

Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. 2017. Gaze Embeddings for Zero-Shot Image Classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6412–6421.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3128–3137.

Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: a dataset for studying eye and head coordination in everyday activities. *Scientific Reports* 10, 1 (2020), 1–18.

George Alex Koulieris, George Drettakis, Douglas Cunningham, and Katerina Mania. 2016. Gaze prediction using machine learning for dynamic stereo manipulation in games. In *Proc. IEEE Virtual Reality (VR)*. 113–120.

Yogesh Kumar and Pekka Marttinen. 2024. Improving Medical Multi-modal Contrastive Learning with Expert Annotations. *arXiv preprint arXiv:2403.10153* (2024).

Yining Lang, Wei Liang, and Lap-Fai Yu. 2019. Virtual agent positioning driven by scene semantics in mixed reality. In *Proc. IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 767–775.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proc. European Conference on Computer Vision (ECCV)*. 201–216.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. International Conference on Machine Learning (ICML)*. 19730–19742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. International Conference on Machine Learning (ICML)*. 12888–12900.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), 9694–9705.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*. 4654–4662.

Hua Liao, Weihua Dong, Haosheng Huang, Georg Gartner, and Huiping Liu. 2019. Inferring user tasks in pedestrian navigation from eye movement data in real-world environments. *International Journal of Geographical Information Science* 33, 4 (2019), 739–763.

Chong Ma, Hanqi Jiang, Wenting Chen, Yiwei Li, Zihao Wu, Xiaowei Yu, Zhengliang Liu, Lei Guo, Dajiang Zhu, Tuo Zhang, et al. 2024. Eye-gaze guided multi-modal alignment for medical representation learning. *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), 6126–6153.

Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. 2023. Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1441–1450.

Pedro Morgado, Yi Li, and Nuno Nvasconcelos. 2020. Learning representations from audio-visual spatial alignment. *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 4733–4744.

Takumi Nishiyasu and Yoichi Sato. 2024. Gaze Scanpath Transformer: Predicting Visual Search Target by Spatiotemporal Semantic Modeling of Gaze Scanpath. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 625–635.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proc. European Conference on Computer Vision (ECCV)*. 631–648.

Peixi Peng, Wanshu Fan, Yue Shen, Wenfei Liu, Xin Yang, Qiang Zhang, Xiaopeng Wei, and Dongsheng Zhou. 2024. Eye gaze guided cross-modal alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics* (2024).

Mengyu Qiu, Quan Rong, Dong Liang, and Huawei Tu. 2023. Visual ScanPath Transformer: Guiding Computers to See the World. In *Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 223–232.

Hosnieh Sattar, Sabine Muller, Mario Fritz, and Andreas Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 981–990.

Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2023. Multimodal integration of human-like attention in visual question answering. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2648–2658.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A gaze dataset to study multimodal neural attention in VQA. In *Proc. ACL SIGNLL Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 27–43.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*. 1–15.

Sven Stauden, Michael Barz, and Daniel Sonntag. 2018. Visual search target inference using bag of deep visual words. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. 297–304.

Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203* (2016).

Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. 2023. ScanDMM: A Deep Markov Model of Scanpath Prediction for 360deg Images. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6989–6999.

Jamalia Sultana, Ruwen Qin, and Zhaozheng Yin. 2024. Seeing Through Expert's Eyes: Leveraging Radiologist Eye Gaze and Speech Report with Graph Neural Networks for Chest X-ray Image Classification. In *Proc. Asian Conference on Computer Vision (ACCV)*. 2579–2595.

Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4664–4677.

Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2487–2496.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. 2024a. GazeGNN: A Gaze-Guided Graph Neural Network for Chest X-Ray Classification. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2194–2203.

Yao Wang, Weitian Wang, Abdullah Abdelhafez, Mayar Elfares, Zhiming Hu, Mihai Bâce, and Andreas Bulling. 2024b. SalChartQA: Question-driven saliency on information visualisations. In *Proc. CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14.

Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. 2024. Unifying top-down and bottom-up scanpath prediction using transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1683–1693.

Alfred L Yarbus. 1967. *Eye movements and vision.* Springer.