Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Master's Thesis

Analysing Human vs. Neural Attention in VQA

Yingpeng Ma

Course of Study:	INFOTECH (Information Technology)
Examiner:	Prof. Dr. Andreas Bulling
Supervisor:	Yao Wang, M.Sc. Susanne Hindennach, M.Sc.
Commenced:	August 01, 2023
Completed:	February 01, 2024

Abstract

Visual Question Answering (VQA) has drawn substantial interest in both academic and industrial research fields in recent years. Driven by Vision Transformers (ViT) and the vision-text co-attention mechanism, these models have shown notable performance improvement. Yet, the black-box nature of neural attention impedes people from understanding its functionality and establishing their trustworthiness. Drawing inspiration from various scholars and their contributions, this thesis demystifies these mechanisms. We aim to 1) extract the neural attention weights of VQA models, 2) remap the weights to machine attention maps, 3) compare machine attention with human gazing heatmaps, and 4) compute the related metrics to provide deeper insights into the attention patterns.

First, the attempts to reproduce the MCAN model implementation and machine attention extraction on the VQA-MHUG dataset are performed on the MULAN framework. Through a comparison with official implementations, the accuracy and correctness of the re-implementation have been verified. Then, utilizing the toolkit of the MULAN framework, the $1D$ attention weights are remapped to $2D$ neural attention maps. Next, these attention maps are compared to human-gazing heatmaps of VQA-MHUG using explainable AI (XAI) metrics.

Following the above pipeline, another experiment on the AiR-D dataset is conducted and reports the Area Under ROC Curve (AUC), Spearman's rank correlation coefficient (ρ), and Jensen-Shannon Divergence (jsd) metrics to compare the neural attention with the human gazing heatmaps.

Finally, the discussion of the differences between the official and re-produced implementations is presented alongside insights on the interpretability of neural attention in VQA models.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature

Contents

List of Figures

List of Tables

1 Introduction

In recent years, the field of Question Answering (QA) has significantly evolved, driven by the advances in transformer technologies, Dosovitskiy et al., 2021, Vaswani et al., 2017, and attention mechanisms, Hochreiter et al., 1997, Pennington et al., 2014. Chart Question Answering (CQA) and Visual Question Answering (VQA), as distinct subfields, have become increasingly over-performance. These areas blend Natural Language Processing (NLP) and Computer Vision (CV) to address queries based on charts, graphs, and images, exemplifying the convergence of human-centered computing, visualization, and computational methodologies. With text-based questions and visual data inputs such as charts or images, these models implement a generative learning-to-answer mechanism. This integration of textual and visual data presents unique AI challenges, particularly in devising models that can effectively interpret and reason about visual content.

When humans answer questions about images, they naturally focus on the most pertinent parts of those images. This tendency can be reflected in human gazing maps, where the areas of highest focus often correspond to the most relevant aspects of the image in relation to the question. This selective attention in visual perception highlights how people prioritize certain elements in an image for understanding and responding, underscoring its importance in various applications like visual analytics. We present more detailed research and arguments on this topic in Chapter 5.

However, exploring how neural networks direct attention to images, particularly in the context of question-answering models, remains relatively under-researched. Concurrently, with the advancements in model performance, there's a growing emphasis on the interpretability and explainability of these systems. As VQA and CQA models increasingly contribute to decision-making in critical sectors like healthcare, finance, and automation, their transparency is both a technical requirement and an ethical and practical necessity. Traditional methods, such as GradCam, Selvaraju et al., 2019, are proving insufficient to meet these complex needs, indicating a demand for more advanced solutions in model interpretability.

Attention heatmaps in neural networks effectively reveal how the machine focuses its attention, offering a visual representation of the areas within an image that the model deems most significant for making a decision or prediction. These maps are a crucial

tool in understanding and interpreting the behavior of neural networks, as they provide insight into what the model 'sees' and 'considers important' in a given input. By analyzing these attention maps, we gain a better understanding of the model's decision-making process, identify potential biases, and improve the model's accuracy and reliability.

The Deep Modular Co-Attention Networks (MCAN) model, Yu et al., 2019b, a notable development in Visual Question Answering (VQA) research, leverages co-attention mechanisms for both text and images to generate answers, achieving state-of-the-art performance. It processes the inputs by first extracting features from images via a Faster R-CNN model, Ren et al., 2016, and tokenizing textual queries via 300-D GloVe word embeddings, Pennington et al., 2014. These extracted image features and tokenized questions are then fed into Modular Co-Attention (MCA) layers. This process facilitates an effective integration of visual and textual information. Following multimodal fusion and output classifier, the model can generate the final answer based on the cohesively processed visual and textual data.

The Multimodal Integration of Human-Like Attention (MULAN) framework is the first method to increase the training features for multimodal VQA models, Sood et al., 2021a. With the integration of human-like attention on image and text, MULAN allows the neural network to be supervised by human-like attention patterns, leading the model to concurrently process and analyze both visual and textual data, and closely mimicking human cognitive attention mechanisms. MULAN framework uniquely provides efficient tools for extracting attention maps from both images and texts, greatly facilitating research and development processes. Although initially trained on the VQAv2 dataset, Goyal et al., 2017, MULAN's versatile design is compatible with various other VQA models and datasets, showcasing its potential for widespread application and adaptability in diverse research contexts.

In this thesis, our initial approach involved following the MULAN framework, Sood et al., 2021a, to train MCAN models, Yu et al., 2019b, on the VQAv2 dataset, Goyal et al., 2017. Next, the MULAN framework provides tools to generate and output machine attention weights. We then engaged to remap the weight to machine attention heatmaps, so that they can be easily compared to human gazing heatmaps, which are provided by the VQA-MHUG dataset, Sood et al., 2021b. Consequently, we calculate the statistical metrics similar to the MCAN model and VQA-MHUG to validate our re-implementation. Following the aforementioned, we outlined and proposed a systematic pipeline for the analysis of human and neural attention. To demonstrate the effectiveness of this pipeline, we implemented it on the GQA dataset, Hudson et al., 2019, facilitating comparative analysis of attention heatmaps with the AiR-D dataset, Chen et al., 2022.

The task of analyzing human and neural attention is complex and multi-faceted. Fortunately, Explainable AI (XAI) methods provide essential tools for justifying, enhancing,

and unveiling the transparency of neural network models, Krajna et al., 2022. Techniques such as attention flow and attention roll-out, Abnar et al., 2020, offer post hoc methods for examining self-attention in Transformer models. Additionally, metrics like AUC, JSD, and the Spearman's rank correlation ρ offer a more numerical and statistical evaluation, enabling a thorough assessment of the model performance and analyzing results.

In summary, the main contributions of this thesis are the following five-folds:

1. Successfully reproduced the MULAN framework, Sood et al., 2021a, for the MCAN model, Yu et al., 2019b, on the VQAv2 dataset, Goyal et al., 2017, and validated the compatibility and consistency of the models.
2. Summarized and proposed a pipeline for analyzing human and neural attention mechanisms.
3. Effectively trained and validated the MCAN model on the GQA dataset, Hudson et al., 2019, extracted neural attention maps, and conducted a comparative analysis with AiR-D human gazing heatmaps, Chen et al., 2022.
4. Performed a detailed comparison between human and neural attention heatmaps, yielding comprehensive results and insights.
5. Identified and rectified various errors and bugs in the MULAN implementation, enhancing its functionality and reliability.

2 Related Works

In this chapter, we begin by introducing the intricacies of neural attention mechanisms in question-answering models, highlighting how they discern and prioritize relevant information in response to queries. Subsequently, we explore a range of models and datasets pertinent to Chart Question Answering (CQA) and Visual Question Answering (VQA) tasks, providing an overview of their architectures and performances. Lastly, we turn our attention to Explainable AI (XAI) methods, which are instrumental in demystifying machine attention processes, offering a clearer understanding of how AI models arrive at their conclusions and enhancing the transparency of the advanced neural network models.

2.1 Neural Attention Mechanism

Attention mechanisms in neural networks, inspired by the human cognitive process of selective attention, are first introduced in the context of Neural Machine Translation (NMT) by Bahdanau et al., 2016. This research aims to address the challenge of encoding long variable-length inputs into a fixed-length context vector while preserving information from earlier tokens. The method tries to generate output a^{pred} by projecting each source embedding e_i^s and target embedding e_j^t using the “alignment score” α.

$$a^{pred} = \sum_{i=1}^{N} \alpha e_i^s \tag{2.1}$$

$$\alpha = softmax(scoring(e_j^t, e_i^s)) \tag{2.2}$$

Where N is the source embedding length and the $scoring$ function can be variant depending on the model task: cosine similarity (Graves et al., 2014), additive/concat attention (Bahdanau et al., 2016), general/bilinear attention (Xu et al., 2016), scaled dot-product attention (Vaswani et al., 2017), sparse softmax (Peters et al., 2019), etc. Nowadays, the "alignment score" α is often called "attention weight".

In the work of Bahdanau et al., 2016, the source and target embeddings are the hidden states of recurrent networks, but more and more tasks and neural network models introduce the task-specific attention mechanism and co-attention mechanism, like Chart Question Answering (CQA) Masry et al., 2022 and Visual Question Answering (VQA) Yu et al., 2019b.

Vaswani et al., 2017 introduced a groundbreaking development in neural networks with their proposal of the Transformer architecture. This innovation enhanced the performance and design of natural language processing (NLP) tasks, offering a more extensive range of capabilities. Concurrently, there is a notable rise in the prominence of attention-based neural networks. These networks, leveraging the Transformer's efficient handling of sequences, revolutionize various aspects of NLP, from understanding context to generating more coherent and contextually relevant text, thereby marking a significant milestone in the evolution of machine learning and AI technologies.

Vision Transformer (ViT) architecture, Dosovitskiy et al., 2021, introduces the Transformer in NLP tasks to computer vision (CV) tasks. This method, deviating from conventional CNN-based strategies, processes images as a sequence of patches and applies a self-attention mechanism to capture global dependencies within the image.

To apply a standard Transformer to computer vision tasks, images, as $3D$ dimension input $image_{3D} \in \mathbb{R}^{H \times W \times C}$, are first flattened to $2D$ patches $image_{2D} \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where H and W stands for the height and width of the image, C for image color channels, $N = \frac{HW}{P^2}$ for number of patches, and P^2 for patch size. N is also the input sequence length of the Transformer, which is typically 512 or 1024. Since Transformer layers maintain a constant latent vector size D, all image patches are mapped to this D dimensional space, ensuring compatibility with the Transformer architecture.

Dosovitskiy et al., 2021 proposes the Multihead Self-Attention (MSA), which is the core unit of the ViT implementation, as an alternative qkv self-attention. Similar to Vaswani et al., 2017, for each element in an input sequence $z \in \mathbb{R}^{N \times D}$, a weighted sum over all values v is first calculated. Then the attention weights A_{ij} are based on the pairwise similarity between two elements of the sequence and their respective query q_i and key k_j representations. MSA extends the qkv self-attention with h self-attention operations, called "heads", in parallel, and projects their concatenated outputs:

$$A = softmax(qk^\top / \sqrt{D_h}), \qquad A \in \mathbb{R}^{N \times N} \tag{2.3}$$

$$SA(z) = Av \tag{2.4}$$

$$MSA(z) = [SA_1(z); SA_2(z); \ldots; SA_h(z)]U_{msa}, \tag{2.5}$$

Supplementary position information is also required to embed the flattened image patches to make sure that each patch can match the correct position. Similar to BERT, Devlin et al., 2019, the first embedding patch is pended as class tokens $z_0^0 = x_{class}$. The output of this token z_L^0 serves as the image representation y.

2.2 Visual Question Answering

In the task of Visual Question Answering (VQA), machines are endowed with the capability to interpret and respond to queries about images by synergizing visual and textual data. This process essentially equips machines with the skills to perceive and cognize visual content. The efficacy of VQA hinges on three pivotal components. Initially, image processing empowers the machine to scrutinize the image, distilling essential visual features. Subsequent to this is question comprehension, wherein the system parses and grasps the semantics and intent of the query posed. The culmination of this process is the answer generation phase, where the machine amalgamates the insights gleaned from both image analysis and question interpretation to formulate coherent and accurate answers.

The crux of emulating human-like cognition in Visual Question Answering (VQA) models lies in the adept construction of multimodal representations. These intricate representations serve as the linchpin for amalgamating the visual and textual realms, granting machines the prowess to harmonize imagery with language. By establishing a cohesive understanding of the image-question duo, these representations pave the way for responses that are not only accurate but also imbued with significance. In the intricate tapestry of VQA, multimodal representations are indispensable for they encapsulate the depth and subtleties inherent in both visual cues and linguistic elements. This dual comprehension empowers the system to perceive the finer details of the image while discerning the question with precision. Ultimately, multimodal representations are paramount in engendering responses that are both user-friendly and contextually astute.

Since the year of 2015, there has been an exponential surge in research aimed at enhancing the performance architecture performance of visual question answering (VQA) systems, underscoring the critical role of multimodal frameworks. The literature on VQA was sparse until 2015, with minimal publications each year. This data, curated from Google Scholar using the query "visual question answering" , reveals a stark inflection point post-2015. The annual output of research papers has escalated precipitously, rising from a modest count of 76 to an impressive tally nearing $6,000$ papers per year. This dramatic uptick reflects the burgeoning interest and significant advancements within the field of VQA.

In the following segments of this section, we provide a concise overview of various multi-modal VQA methodologies, alongside an introduction to some of the most widely utilized datasets in the field of Visual Question Answering.

2.2.1 Visual Question Answering Models

Researchers frequently reference Agrawal et al., 2016 for pioneering the concept of VQA, which integrates computer vision and natural language processing to address open-ended questions about images. This research melds multi-modal AI and spans various disciplines, focusing on goal-driven tasks like aiding visually impaired individuals or security monitoring. Unlike structured text, open-ended text lacks specific formatting, and invites unrestricted answers, contrasting with multiple-choice formats. The study also contrasts image captioning with image question answering. Stacked Attention Networks (SANs) Yang et al., 2016 uses question semantics to locate relevant image regions. Its multi-layered approach iteratively refines focus on key image areas, cumulatively leading to the answer in a layer-by-layer manner. Multi-Modal Factorized Bilinear Pooling (MFB) Yu et al., 2017 combines detailed image and question features using MFB pooling and a co-attention mechanism, creating a unified model that enhances VQA effectiveness. Feature Embedding for Visual Question Answering Lu et al., 2017, combines the visual attention of free-form regions with detection-based methods, integrating features from image regions, detection boxes, and question representations. This approach, using multi-modal multiplicative feature embedding, simultaneously focuses on relevant image areas and detection boxes, enhancing question-answering accuracy. Learning Cross-Modality Encoder Representations from Transformers (LXMERT) Tan et al., 2019 introduces a Transformer-based model to analyze vision-language interactions. It features three encoders for image relationships, language, and cross-modality. Pre-trained on extensive image-sentence pairs, it excels in linking visual and linguistic content. The study Jiang et al., 2020 revisited grid-based convolution features in VQA, discovering they perform comparably or better than bounding box and region-based features, with improved running times. Grid features also showed superior generalization in tasks like image captioning and enabled strong end-to-end VQA performance without pre-training region annotations. Deep Modular Co-Attention Networks for VQA Yu et al., 2019b presents a deep co-attention model, MCAN, which adeptly correlates keywords in questions with key objects in images. The model is built upon a series of Modular Co-Attention (MCA) layers, each designed to simultaneously manage self-attention for questions and images and guide the attention of images based on the questions. This is achieved through a composite of two fundamental attention units within each layer, allowing the model to intricately align and interpret visual and textual information. The layered structure of MCAN facilitates progressively refined understanding, making it a robust framework for complex VQA tasks.

In the paper **Multimodal Human-like Attention Network (MULAN)**, Sood et al., 2021a, the authors developed a novel approach for training VQA models by leveraging human-like attention on both image and text as a guiding signal for neural attention. This method demonstrated notable advancements in VQA tasks. It involves integrating attention predictions derived from cutting-edge text and image saliency models into the self-attention layers of contemporary transformer-based VQA models. MULAN stands out for its efficiency, requiring fewer trainable parameters and less runtime, while still outperforming previous models in terms of accuracy.

2.2.2 Visual Question Answering Datasets

DAtaset for QUestion Answering on Real-world image (DAQUAR) Malinowski et al., 2015, is foundational in VQA research. Based on NYU-Depth V2, Nathan Silberman et al., 2012, it comprises real-world images with $6,794$ training and $5,674$ test question-answer pairs, including both synthetic and human-generated content, offering a broad spectrum for VQA study. Visual Genome (VG) Krishna et al., 2016 was the largest dataset at its release, consisting of over $100,000$ images from MS-COCO, Lin et al., 2015, and YFCC100M, Thomee et al., 2016. VG's unique feature is its structured questioning format, starting each question with a 'W' word (What, Where, When, etc.), totaling around 1.7 million question-answer pairs. This design provides a substantial resource for visual understanding and question answering studies. VQAv2 Goyal et al., 2017, is developed to address biases and imbalance in answer distribution found in the earlier VQAv1 dataset (Agrawal et al., 2016). Recognizing that biases in language and inherent structures in the real world can often overshadow visual cues in learning, leading to an overestimated performance of models, VQAv2 aims to mitigate this issue. It achieves a more balanced dataset by pairing each question with two similar images that yield different answers. This approach effectively doubles the number of image-question pairs compared to its predecessor. The expanded VQAv2 dataset comprises $82,783$ training images, $40,504$ validation images, and $81,434$ test images, along with $443,757$ training questions, $214,354$ validation questions, and $447,793$ test questions, creating a more robust and challenging dataset for VQA research[1]. GQA dataset Hudson et al., 2019 stands as a significant resource for VQA tasks. The dataset's creation involved an innovative question engine that utilizes the scene graph structures from Visual Genome, Krishna et al., 2016, to generate a wide array of reasoning questions. Each question is accompanied by a functional program that defines its semantics. This approach allows for precise control over the answer distribution and effectively mitigates biases. The GQA dataset comprises a total of $113,000$ images and an impressive 22

[1] https://visualqa.org/

million questions. Additionally, the dataset includes pre-extracted image features for both grid and region-based analyses, further enriching its utility for VQA research[2]. The datasets highlighted above primarily concentrate on Visual Question Answering (VQA) tasks, offering tuples of (I, Q, A), where I represents the image, Q denotes the question, and A is the ground-truth answer. In this thesis, we focused on training models using the **VQAv2** and **GQA** datasets. This approach was instrumental in validating the transferability of the MULAN model and the effectiveness of our implementation. However, to delve deeper into machine attention mechanisms and to compare human and neural attention patterns, the incorporation of human eye-tracking data is equally crucial. This additional data provides valuable insights into how human attention navigates and processes visual information, thereby offering a more comprehensive understanding of attention dynamics in VQA systems.

2.2.3 Eye-Tracking Dataset

Compositional Language and Elementary Visual Reasoning diagnostics dataset (CLEVR), Johnson et al., 2016, aims to evaluate the reasoning abilities of VQA models. Noting biases in existing datasets, CLEVR offers 100k rendered images and around one million questions, with 853k unique, to isolate and analyze VQA reasoning capabilities.

Attention in Reasoning (AiR) framework, Chen et al., 2022, is one of the initial attempts to scrutinize the progression and rationality of attention in VQA tasks. AiR employs neural attention to enhance understanding and improve task outcomes. It introduces an evaluation metric (AiR-E) grounded in a series of atomic reasoning operations, facilitating a quantitative attention assessment that incorporates the reasoning process. The AiR-D dataset, part of this framework, includes human eye-tracking data and answer correctness, focusing on the GQA dataset's validation split with $1,454$ image-question pairs[3]. AiR-D's unique inclusion of human gazing maps serves as one of the ground truths in this thesis.

VQA Multimodal Human Gaze dataset (VQA-MHUG), Sood et al., 2021b, is a novel dataset comprising human gaze data on images and questions collected during VQA tasks using a high-speed eye tracker. It aims to compare human and neural attention strategies, revealing that a stronger correlation with human textual attention is indicative of better VQA performance. The VQA-MHUG dataset, based on the VQAv2 dataset's validation split, consists of $3,990$ image-question pairs, each answered by 3 participants, and includes human gazing heatmaps[4]. The study trained and tested five advanced VQA

[2]https://cs.stanford.edu/people/dorarad/gqa/index.html
[3]https://github.com/szzexpoi/AiR
[4]https://perceptualui.org/publications/sood21_conll/

models on this dataset to examine neural attention patterns. Our research replicates the findings as an initial step in our engineering tasks, validating our methodologies against this novel dataset.

2.3 Chart Question Answering

Chart Question Answering (CQA) is a specialized subset of Visual Question Answering (VQA) that concentrates on interpreting and responding to queries about charts, graphs, and various data visualizations. In CQA, machine learning models are specifically trained to identify chart-related elements and analyze data points, relationships, and trends depicted in these visual representations. The primary aim of a CQA system is to automatically provide answers to natural language questions about charts, thereby enhancing the process of visual data analysis.

The advent of robust VQA models has catalyzed focused research in the area of CQA. As researchers began tailoring VQA inputs to accommodate chart images, the unique capabilities and potential applications of CQA started to emerge more prominently. Given the absence of a universally recognized CQA dataset, our discussion will include an overview of some notable CQA models, as well as the datasets that have been developed and validated in conjunction with these models. This exploration offers insights into the evolution and current state of CQA tasks within the broader context of visual question answering.

Data Visualization Question Answering (DVQA), Kafle et al., 2018, addresses the challenge of interpreting bar charts in VQA. The DVQA dataset shows the difficulties advanced VQA algorithms face with bar charts. To tackle this, two models were developed: **Multi-Output Model (MOM)**, which combines Chart-OCR for data extraction and an LSTM for questions, and **SAN with DYnamic Encoding Mode (SANDY)**, an enhanced version of SAN that dynamically processes chart-specific question words. These models mark progress in automated bar chart analysis.

FigureQA, Kahou et al., 2018, offers an extensive visual reasoning corpus with over a million question-answer pairs based on $100,000+$ chart images, including various scientific-style figures. The dataset uses 15 templates to explore plot relationships and characteristics. The **FigureNet** model, developed to assess FigureQA, combines color probabilities, LSTM-generated question vectors, and one-hot vectors for color descriptors in a novel approach to graphical data interpretation.

ChartQA, Masry et al., 2022, stands as a significant benchmark in the research field of CQA tasks. The authors observed that people typically pose a range of complex reasoning questions about charts, often involving intricate logical and arithmetic operations tied

to the chart's visual features. To mirror this real-world complexity, they introduced a substantial benchmark encompassing 9.6K human-generated questions and an additional 23.1K questions derived from human-written chart summaries.

Addressing the intricate challenges of visual and logical reasoning in chart analysis, ChartQA introduces two transformer-based models: **VisionTAPAS** and **VL-T5**. These models innovatively integrate visual features with the chart's data table, presenting a cohesive approach to answering questions. This unified methodology enables the models to effectively process and interpret both the visual and data-driven aspects of charts, showcasing a significant advancement in the field of CQA.

To conclude this section, we consolidate the key information about the discussed VQA and CQA models in **??**. This table presents the backbone architectures and the fundamental approaches of each model, offering a comprehensive overview and facilitating easy comparison between the various methodologies employed in the field of Visual and Chart Question Answering.

2.4 Explainable AI

Explainable AI (XAI) methods offer valuable tools to justify, improve, and explore the transparency of neural network models Krajna et al., 2022. In our thesis, XAI is particularly valuable for gaining deeper insights into how neural network models function. One aspect of XAI is feature-based or global model explainability, which focuses on understanding the contribution of input features to the model's output. The **Shapley** value, Rozemberczki et al., 2022, offers both local and global interpretations, although it is computationally intensive. As an alternative, the **kernel SHAP** method, Chau et al., 2022, provides a more efficient kernel-based approximation of the Shapley value.

Another notable method is **Local Interpretable Model-Agnostic Explanations (LIME)**, which perturbs input data samples to observe changes in predictions, offering insights into the model's decision-making process, as detailed in Ribeiro et al., 2016.

Additionally, example-based or local techniques delve into specific explanations for each decision a model makes. One such technique, **Anchors**, has shown promise in identifying the sufficient conditions under which a model makes high-precision predictions. It supports model-agnostic approaches across different data types, including text, images, and tabular data, making it a versatile tool in the XAI toolkit.

In this thesis, XAI holds promise as we aim to gain deeper insights into the neural network models. Feature-based / Global model explainability describes how input features contribute to the model output. **Shapley** value provides local and global

interpretation, but the computational expansiveness is its drawback, Rozemberczki et al., 2022. The **kernel SHAP** method can be a faster kernel approximation of it, Chau et al., 2022. **Local Interpretable Model-Agnostic Explanations (LIME)** perturbs the input of data samples and comprehends how the predictions change, which is also a promising method to explain the model processing flow, Ribeiro et al., 2016.

Besides, example-based / local techniques focus more on detailed explanations for every decision made. **Anchors** show its great potential to capture sufficient conditions of features at which the model gives a high-precision prediction and supports model-agnostic approaches for classification models of text, image, or tabular data.

3 Experiment on VQA Models

The **Visual Question Answering (VQA)** task is a compelling area in AI that focuses on automatically generating answers from input images in response to associated questions. It requires a sophisticated interplay of image understanding and language processing. Recent advancements in this field, particularly noted in Sood et al., 2021b, involve the extraction of attention maps from neural networks and their comparison with human attention patterns. These studies have been instrumental in highlighting the profound impact that both neural and human attention mechanisms have on the performance of VQA models. Inspired by these findings, our research aims to delve into extracting similar attention maps from VQA models. We seek to dissect and comprehend the variances between neural attention mechanisms and human attention processes. This exploration is expected to shed light on how neural models process visual and textual information and how their approaches differ from human cognitive processes, thereby contributing valuable insights to the field of VQA.

3.1 Basic framework

This section is dedicated to examining the foundational framework for our study. We focus on two advanced VQA architectures: the **Modular Co-Attention Network (MCAN)** model and the **Multimodal Human-like Attention Network (MULAN)** framework. These architectures represent the forefront of VQA technology, combining sophisticated attention mechanisms and deep learning strategies to process and interpret complex visual and textual data. Our exploration into these models involves dissecting their structural components, understanding their operational dynamics, and assessing their potential for adaptation and application in various VQA scenarios. By analyzing these state-of-the-art models, we aim to uncover new pathways for enhancing the accuracy and efficiency of VQA systems, thereby contributing to the broader field of artificial intelligence and machine learning.

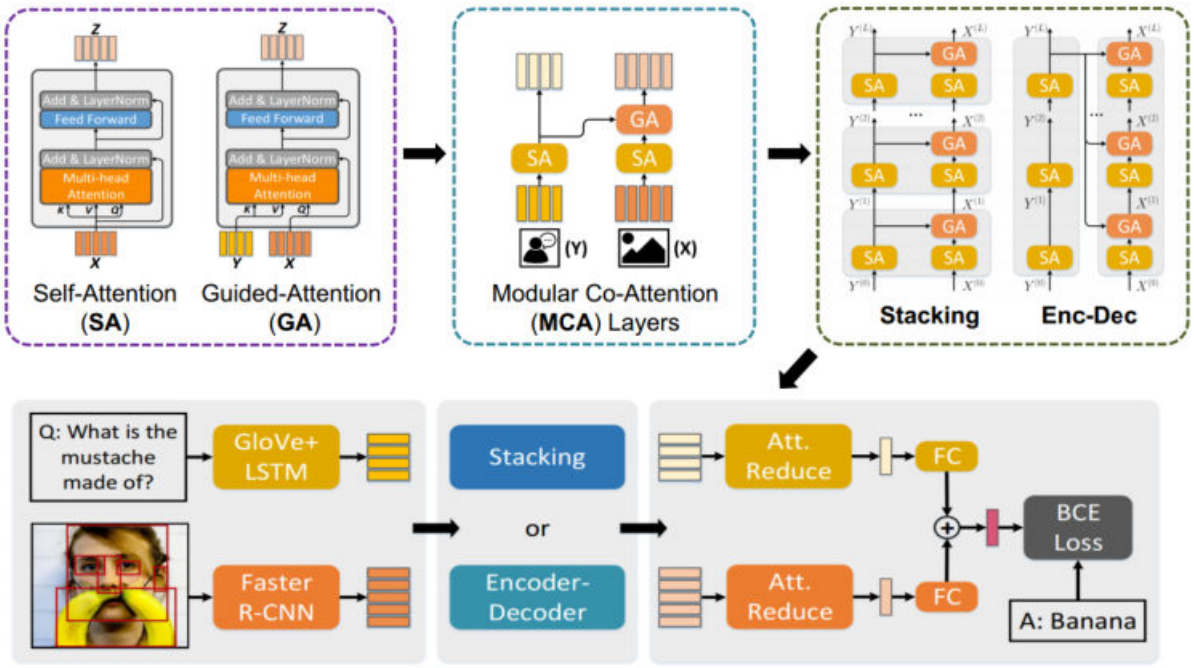

Figure 3.1: The MCAN architecture features the Modular Co-Attention (MCA) layer, based on Self-Attention (SA) and Guided-Attention (GA) units. Stacking these MCA layers in parallel forms the deep co-attention learning core. The graph is taken from the original Yu et al., 2019b.

3.1.1 MCAN model

The Modular Co-Attention (MCA) layer forms the foundational unit of the MCAN model, Yu et al., 2019b. This layer is ingeniously designed as a modular combination of two fundamental attention units: the self-attention (SA) unit and the guided-attention (GA) unit. This design is inspired by the scaled dot-product attention mechanism, as proposed in Vaswani et al., 2017.

In the scaled dot-product attention mechanism, the input comprises queries q and keys K, both of which share the same dimensionality, denoted as d. The attention weights A that act upon the values V are derived through a process where each query q is dot-multiplied with all the keys K. This dot product is then scaled down by the factor of $\sqrt{d}$ to normalize the results. Subsequently, a $softmax()$ function is applied to these scaled values, enabling the model to effectively determine the relative importance of each value V based on the corresponding attention weights (Equation (3.1)). This attention mechanism is crucial in the MCA layer, as it allows the model to focus on specific elements in the input data, enhancing its ability to interpret and respond to complex visual and textual queries.

$$A(q, K, V) = softmax(qK^\top/\sqrt{d})V \tag{3.1}$$

To enhance the representation capabilities of the attended features, Vaswani et al., 2017 introduced the concept of multi-head attention, a pivotal innovation in attention mechanisms. This approach involves parallelizing h independent 'heads', each representing a distinct scaled dot-product attention function.

In this multi-head attention framework, the attended output features f are computed by first processing the input through each of these individual attention heads (Equation (3.2)). Each head captures different aspects of the input data, allowing the model to attend to various features from multiple perspectives simultaneously. After the input passes through all the heads, the outputs of these heads are then concatenated and linearly transformed to produce the final attended feature set f. This process significantly enriches the model's ability to interpret complex data by enabling it to consider and integrate a diverse range of features and relationships within the input. As a result, multi-head attention provides a more comprehensive and nuanced understanding of the input data, contributing to the effectiveness of MCAN model, Yu et al., 2019b:

$$f = MA(q, K, V) = [head_1, head_2, \ldots, head_h]W^o \tag{3.2}$$

$$head_j = A(qW_j^Q, KW_j^K, VW_j^V) \tag{3.3}$$

where $W_j^Q, W_j^K, W_j^V \in \mathbb{R}^{d\times d_h}$ are the projection matrices for the j-th head to transform the queries, keys, and values respectively, and $W^o \in \mathbb{R}^{h*d_h\times d}$. $d_h = d/h$ is the output feature dimension from each head. The dimension $d_h = d/h$ represents the dimensionality of the output features from each individual head. This division ensures that the computational cost of multi-head attention is similar to that of single-head attention, while still benefiting from the diverse perspectives of multiple heads.

Given a set of question queries $Q = [q_1; q_2; ...; q_m] \in \mathbb{R}^{m\times d}$ as input, the multi-head attention (MA) weights can be calculated for each head. These weights are then used to generate a weighted sum of the values, which are processed through each head's attention mechanism. The outputs from all the heads are then concatenated and projected using W^o, resulting in the final output feature f.

MCAN leverages multi-head attention (MA) to develop two key attention units for handling multimodal VQA inputs: self-attention (SA) and guided-attention (GA), Yu et al., 2019b. SA processes a single set of features $X = [x_1; ...; x_m] \in \mathbb{R}^{m\times d_x}$, producing attended outputs by considering the interactions within X. GA, in contrast, operates on two sets of features, X and Y, generating attended features for X guided by the

context of Y. Then MCAN effectively combines these units to create Modular Co-Attention (MCA) layers. For example, the $SA(Y) - GA(X, Y)$ layer models intra-modal interactions among question word pairs in Y, while the $SA(Y) - SGA(X, Y)$ layer adds another dimension by considering image region pairs in X. This approach ensures comprehensive processing of multimodal data, crucial for accurately interpreting and answering VQA tasks, as shown in Figure 3.1.

In the MCAN model, images are not processed in their original pixel format. Instead, as described in Anderson et al., 2018, they are represented through a set of regional visual features extracted in a bottom-up approach. These features are derived from a pre-trained Faster R-CNN model, Ren et al., 2016, trained on the Visual Genome dataset, Krishna et al., 2016. The model applies a dynamic confidence threshold to the number of detected objects, ranging from $m \in [10, 100]$. Each object, denoted as $x_i \in \mathbb{R}^{d_x}$, is represented by mean-pooling the convolutional features from its detected region, resulting in an image feature matrix $X \in \mathbb{R}^{m \times d_x}$.

Questions are tokenized into words and embedded into a dimension of $n \in [1, 14]$. To accommodate the variable lengths of m and n, zero-padding is employed to standardize X and the question embeddings Y to their maximum dimensions ($m_{max} = 100$ and $n_{max} = 14$, respectively).

Therefore, MCAN processes the image features X and question features Y through deep co-attention learning, utilizing a series of L Modular Co-Attention (MCA) layers stacked sequentially (denoted as $MCA(1), MCA(2), \ldots, MCA(L)$), Yu et al., 2019b. Each layer $MCA(l)$ takes the output features $X(l-1)$ and $Y(l-1)$ from the previous layer as input, producing output features $X(l)$ and $Y(l)$. These outputs are then recursively fed into the subsequent $MCA(l+1)$ layer (Equation (3.4)). This cascading structure allows the MCAN to progressively refine its attention and understanding of the multimodal inputs.

$$[X(l), Y(l)] = MCA(l)([X(l-1), Y(l-1)]) \tag{3.4}$$

For the first Modular Co-Attention (MCA) layer in the MCAN model, denoted as $MCA(1)$, the initial input features are set as $X(0) = X$ and $Y(0) = Y$. In our specific implementation of the MCAN model, we employ $SA(Y) - SGA(X, Y)$ layers, structuring the network in an encoder-decoder format inspired by the Transformer architecture, as discussed in Vaswani et al., 2017.

Following the deep co-attention learning phase, the output image features $X(L) = [x_1(L); \ldots; x_m(L)] \in \mathbb{R}^{m \times d}$ and question features $Y(L) = [y_1(L); \ldots; y_n(L)] \in \mathbb{R}^{n \times d}$ are enriched with detailed information about the attention weights over both question

words and image regions, Yu et al., 2019b. To further process these features, a two-layer Multilayer Perceptron (MLP) is utilized in the MCAN model, consisting of a fully connected layer $FC(d)$, a ReLU activation, a dropout layer with a dropout rate of 0.1, and another fully connected layer $FC(1)$. This MLP is applied to either $Y(L)$ or $X(L)$ to derive the attended features $\widetilde{y}$ or $\widetilde{x}$. The attention weights $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_m] \in \mathbb{R}^m$ are learned during this process, allowing the calculation of the attended features as follows:

$$\alpha = softmax(MLP(X(L))) \tag{3.5}$$

$$\widetilde{x} = \sum_{i=1}^{m} \alpha_i x_i(L) \tag{3.6}$$

$$\widetilde{y} = \sum_{i=1}^{n} \alpha_i y_i(L) \tag{3.7}$$

These equations represent the computation of the final attended features, where $\widetilde{x}$ and $\widetilde{y}$ are weighted sums of the output features, modulated by the learned attention weights. This approach ensures that the most relevant features in the context of the given question and image are emphasized in the final model output.

In the MCAN model, the attended features $\widetilde{y}$ and $\widetilde{x}$ are stabilized and fused using a $LayerNorm()$ function into a unified feature vector z. This feature z is then projected into a vector $s \in \mathbb{R}^N$ and passed through a $sigmoid()$ function, where N is the number of most frequent answers in the training set. The model uses binary cross-entropy (BCE) as the loss function for training an N-way classifier based on z. The fusion of the features is mathematically represented as in Equation (3.8).

$$z = LayerNorm(W_x^T \widetilde{x} + W_y^T \widetilde{y}) \tag{3.8}$$

where W_x and $W_y \in \mathbb{R}^{d \times d_z}$ are linear projection matrices that help in transforming the attended features into a common dimensionality d_z.

3.1.2 MCAN with Grid image feature

The original MCAN model, as detailed in Yu et al., 2019b, relies on region image features extracted using the Faster R-CNN method, Ren et al., 2016. This approach utilizes a detector trained on a curated version of the Visual Genome dataset Krishna et al., 2016, encompassing numerous object categories and attributes annotated with bounding boxes. The extraction process involves a $14 \times 14 RoIPool$ layer, followed by an $AvgPool$

operation for each region, resulting in a set of region features, denoted as N, for each image.

However, a study by Jiang et al., 2020 revisited the use of both region and grid image features for VQA models, suggesting that grid features, extracted directly from the $C5$ layer of a variant of Faster R-CNN, Ren et al., 2016, do not underperform compared to region features. This finding indicates that grid features might be equally effective for VQA tasks.

Thus, in this thesis, we explore the use of both region and grid features within the MCAN model to assess their respective impacts. The grid features, as provided by Jiang et al., 2020, are directly extracted from the $C5$ layer and represented as $feat_{grid} \in \mathbb{R}^{H \times W}$, where H and W correspond to the height and width of the image. Simultaneously, the region features and bounding box information, as delineated in Yu et al., 2019b, are also incorporated. By employing both types of features, we aim to comprehensively evaluate their effectiveness and potential synergies in enhancing the performance of the MCAN model in VQA tasks.

3.1.3 MULAN framework

Sood et al., 2021a introduced the Multimodal Human-like Attention Network (MULAN), a groundbreaking approach for integrating human-like attention across both image and text modalities during the training of Visual Question Answering (VQA) models. MULAN's notable achievement lies in its ability to merge text and image attention within the neural self-attention layers of VQA models, thus enriching the multimodal learning process. Figure 3.2 shows the overview of the MULAN model.

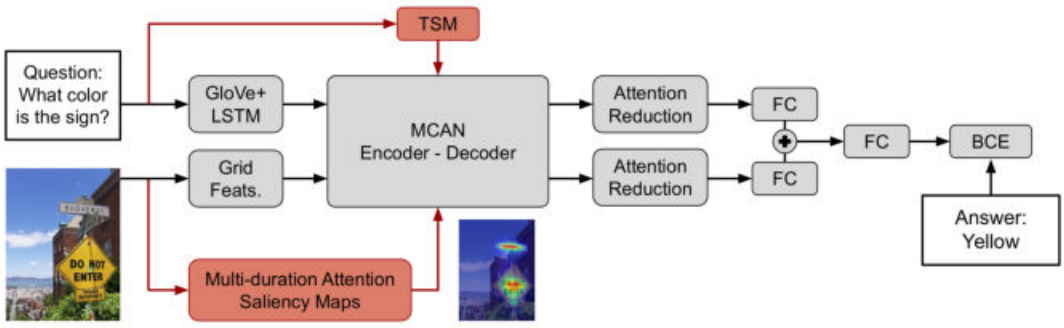

Figure 3.2: Overview of the Multimodal Human-like Attention Network (MULAN). This method is a multimodal integration of human-like attention on questions and images during the training of VQA models. The graph is taken from Sood et al., 2021a

The implementation of MULAN[1] is grounded in the OpenVQA platform[2], a versatile framework for VQA research. OpenVQA facilitates the implementation of various state-of-the-art VQA methodologies, such as BUTD, MFH, BAN, MCAN, and MMNasNet, across different benchmark datasets, including VQAv2, GQA, and CLEVR.

MULAN extends the capabilities of OpenVQA by allowing for the input of additional human-like attention data (MSA for image and TSM for text) and the extraction of neural attention information (with extra parameters during the validation step). This enhancement is particularly crucial for our thesis, as it enables more in-depth analysis and comparison of human and machine attention mechanisms in VQA tasks. By leveraging the extended functionality of MULAN, our implementation can delve deeper into understanding how neural networks mimic or differ from human cognitive processes in interpreting and responding to visual and textual stimuli.

3.2 Machine Attention on VQA-MHUG dataset

In our thesis, we have utilized the MCAN model, Yu et al., 2019b, within the framework of MULAN, Sood et al., 2021a. Our approach diverges slightly in that we did not employ additional human-like attention as a guidance signal in the MULAN framework. However, we place significant emphasis on the output of the neural attention mechanism, considering it a crucial aspect of our analysis.

This section focuses on reproducing the results and machine attention maps using region and grid features on the VQAv2 dataset, as outlined in Goyal et al., 2017. By conducting this reproduction, we aim to not only replicate the original findings but also to generate attention maps that are instrumental in our analysis. These results are then compared with those from the VQA-MHUG dataset, Sood et al., 2021b. The successful comparison and validation of our re-implementation against the VQA-MHUG results are pivotal. They not only confirm the correctness and reliability of our re-implementation but also establish a solid foundation for progressing to the subsequent phases of our research. This step is essential in ensuring that our approach and findings are grounded in robust and accurate methodologies, paving the way for further exploration and analysis.

As previously mentioned in Chapter 2, the VQA Multimodal Human Gaze dataset (VQA-MHUG), Sood et al., 2021b, is constructed using the validation split of the VQAv2 dataset, Goyal et al., 2017. Therefore, in the upcoming subsections of this thesis, our approach involves training and validating our model specifically on the VQAv2 dataset.

[1] https://git.hcics.simtech.uni-stuttgart.de/code/mulan
[2] https://github.com/MILVLG/openvqa

This strategy ensures that our model is aligned with the datasets and methodologies used in the VQA-MHUG study, facilitating a direct and relevant comparison of results and insights. By focusing on this dataset, we aim to comprehensively assess and validate the performance of our model in the context of established benchmarks.

3.2.1 Experiment on MCAN-Region model

The MULAN framework, Sood et al., 2021a, offers several variations of VQA models, including the **MCAN_small** and **MCAN_large** versions. The key difference between these two models lies in their configuration: MCAN_small is designed with a $Hidden\_Size = 512$, while MCAN_large features a larger $Hidden\_Size = 1024$. However, Yu et al., 2019b reports that the performance improvement of the MCAN_large model over the MCAN_small model is not substantial Thus, I have opted to use the **MCAN_small** model for all the implementations in this thesis. This choice is driven by considerations of efficiency and resource optimization, without significantly compromising the performance. By choosing a model that balances size and computational requirements against performance, I aim to achieve accurate and reliable results in the tasks, ensuring that the findings and analyses are both robust and practical.

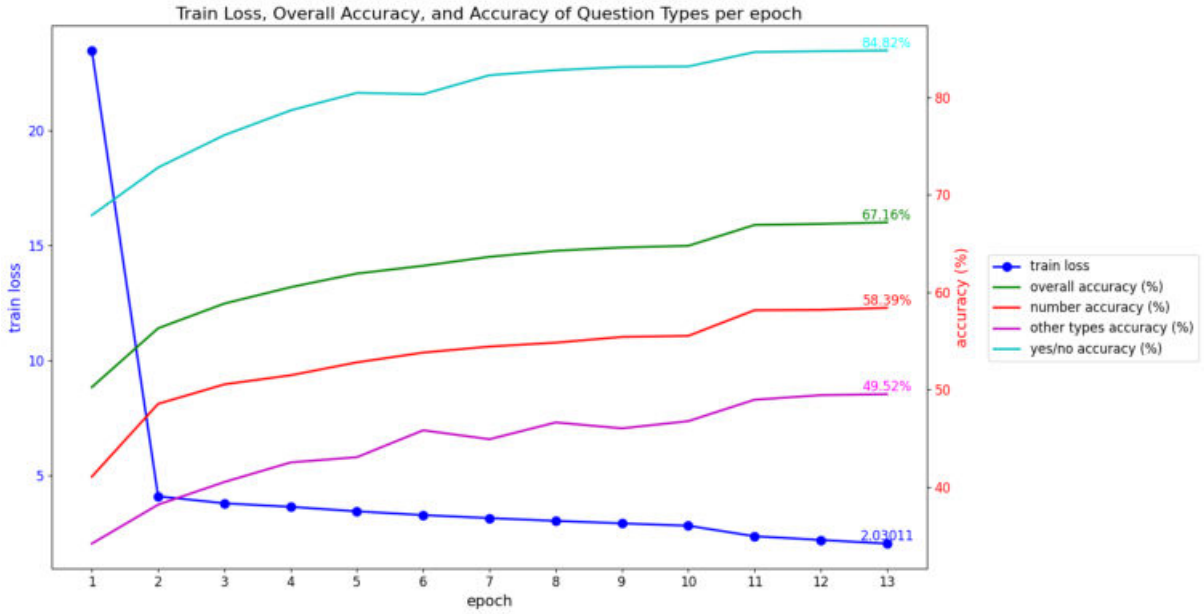

Figure 3.3: MCAN_small train loss and accuracy per epoch using VQAv2 Region features.

In the first experiment, the Region image features are obtained using the bottom-up-attention approach, Ren et al., 2016, where each image is represented by a variable number of features, ranging from 10 to 100, each with a dimensionality of 2048. These

Implemented by	**Overall** (%)	**Yes/No** (%)	**Number** (%)	**Other** (%)
Yu et al., 2019b	**67.17**	84.82	49.31	**58.48**
This thesis	67.16	84.82	**49.52**	58.39

Table 3.1: Per-question-type accuracy of region feature input for Yu et al., 2019b and our implementation.

features are conveniently stored in `.npz` files[3], which are readily compatible with the MCAN implementation, Yu et al., 2019b.

The training is conducted on a single Nvidia Tesla V100-SXM2 GPUs with 32GB RAM, taking approximately 7 hours over a span of 13 epochs. This process culminated in achieving an overall accuracy of 67.16%. The training loss and the detailed accuracy per epoch are illustrated in Figure 3.3. In the meantime, the official MCAN implementation reaches the overall accuracy of 67.17, Yu et al., 2019b, on the `train-to-val` split.

Continuing the experiment, I execute the validation mode (`--RUN = val`) of the MCAN model using the checkpoint from the best-performing epoch (i.e., epoch 13). During this process, I enable the parameter `--SAVE_ATTMAPS = True` to save the attention weights for each image-question pair evaluated during validation. In addition to the attention weights, the predicted answers for each image-question pair are also saved in one `.json` file. Each row of the saved attention weights text file corresponds to a predicted answer in the result file, maintaining a consistent order for easy correlation.

To create attention maps that align with the dimensions of the original images, it's crucial to have information about the image region features, specifically the bounding boxes. Fortunately, the region image features extracted from the Faster R-CNN model, Ren et al., 2016, include these necessary bounding box details[4]. By leveraging this information, the region attention maps based on Region image features are constructed. These maps visually represent how the model focuses on different regions of the image when answering a question. Figure 3.4 illustrates some examples of these region attention maps, showcasing the model's attention distribution across various image object regions.

In the approach to constructing attention maps, the region attention weights, derived from the region features, are mapped onto corresponding regions/bounding boxes in the image, reflecting a region-centric machine attention pattern. To enhance these maps and align them more closely with human-like attention patterns, I apply a Gaussian kernel to

[3]https://github.com/peteanderson80/bottom-up-attention
[4]The bounding box information is saved in a format of $[x\_min, y\_min, w, h]$

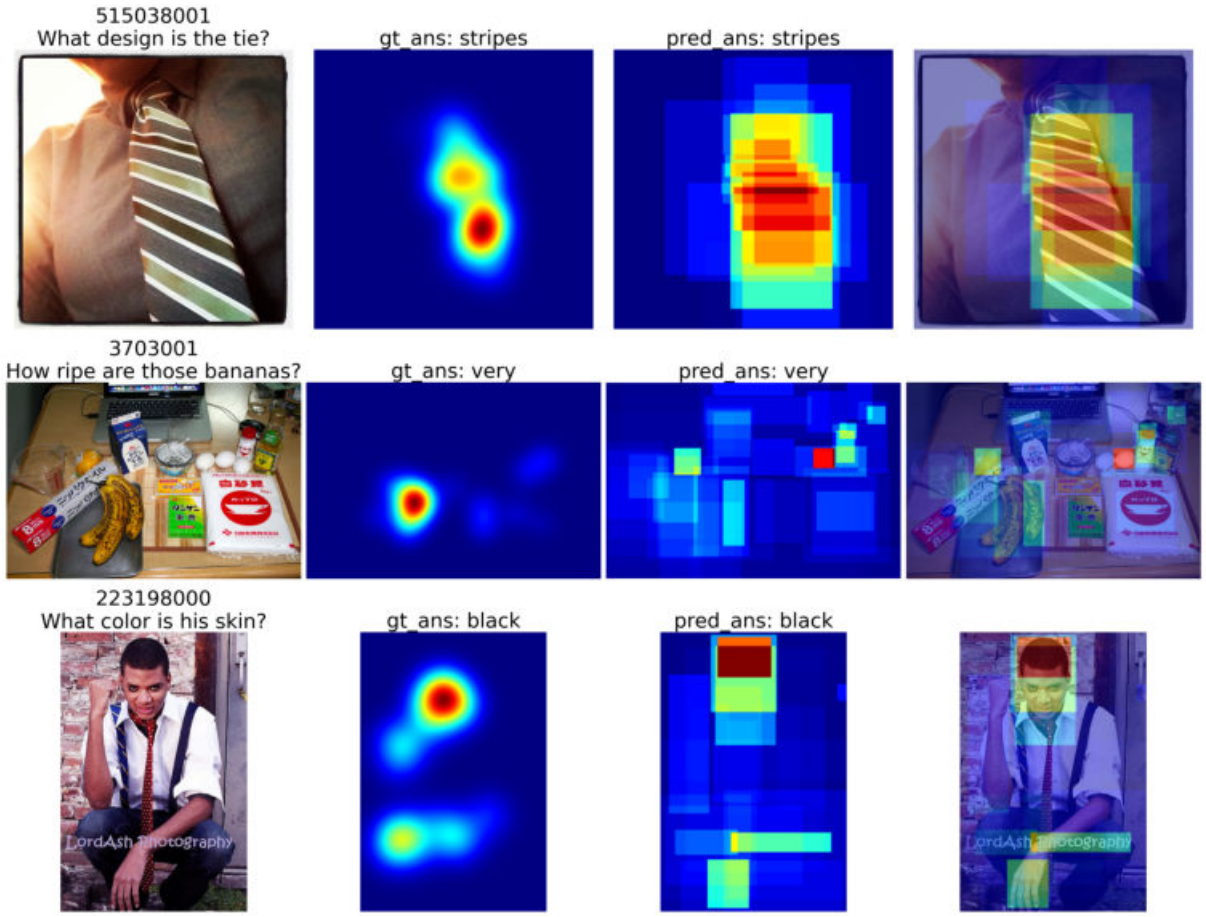

Figure 3.4: Some examples of human and machine Region attention on VQA-MHUG. From left to right: original image, VQA-MHUG human gazing heatmap, machine Region attention heatmap, machine Region attention masked image.

each region. This technique helps in smoothing the attention weights, providing a more natural and continuous distribution of attention across the image. The Gaussian kernel is defined by Equation (3.9), where α_g represents the Gaussianized attention weight, α is the original attention weight extracted from the model, x and y denote the center positions of the bounding box, and $(sig_x, sig_y) = (40, 40)$ specifies the variance of the Gaussian distribution, which determines the extent of the smoothing effect.

$$\alpha_g = \alpha + e^{\sqrt{x}/\sqrt{sig_x}+\sqrt{y}/\sqrt{sig_y}} \tag{3.9}$$

Applying the Gaussian kernel to attention maps is designed to mirror human visual attention, focusing on and distributing across an image. This method yields attention maps that are informative about the model's image processing and interpretable from a

human viewpoint, offering insights into the parallels and distinctions between machine and human visual attention mechanisms.

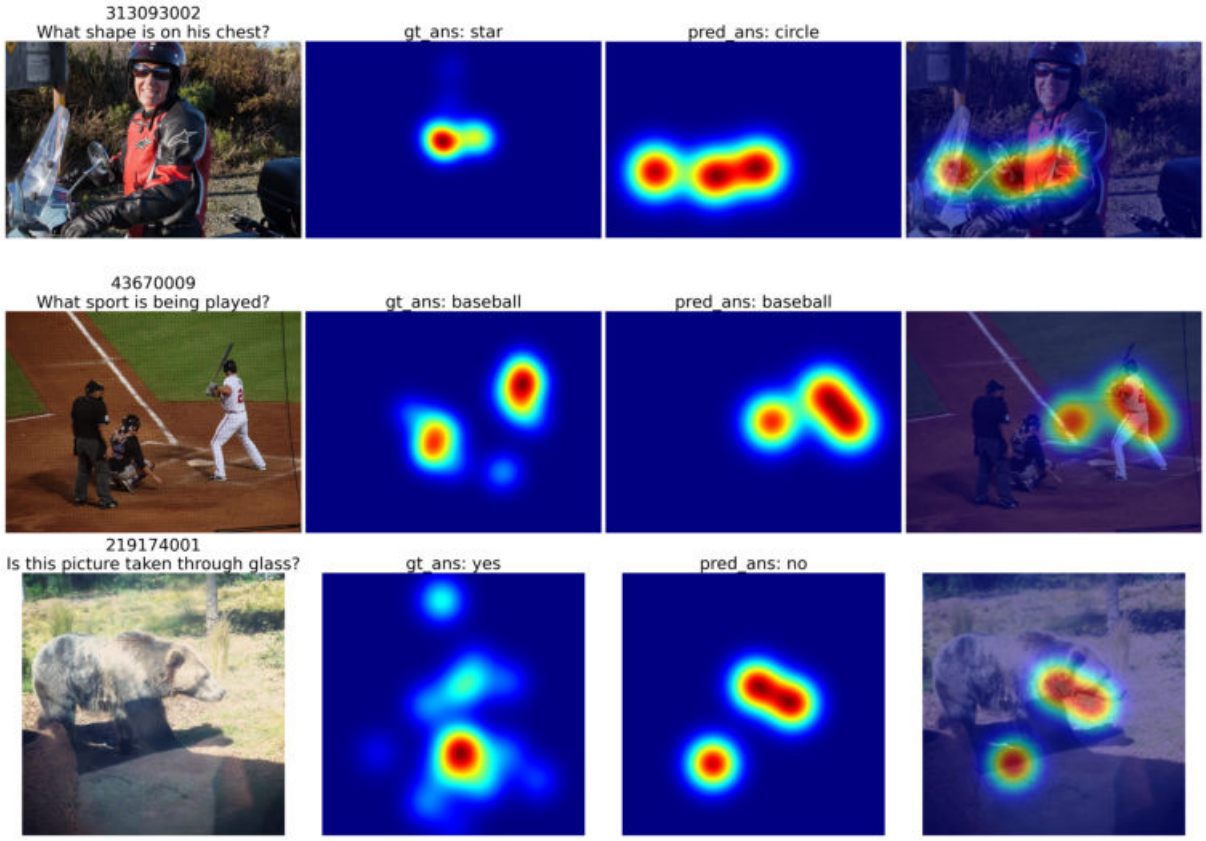

Figure 3.5: Some examples of human and Gaussianized Region attention on VQA-MHUG. From left to right: original image, VQA-MHUG human gazing heatmap, machine Region attention heatmap, Gaussianized machine Region attention masked image.

In refining my attention maps to more closely resemble human-like attention, we introduced a thresholding step in the Gaussianization process. By setting a threshold of 0.001, we filtered out attention weights below this value, effectively disregarding them as valid focus points. This step was crucial to avoid an excessive number of focus points in the attention map, ensuring that only the most relevant regions are emphasized. As a result, the final attention maps, showcasing the cleaned Gaussianized attention, provide a more focused and interpretable visualization of the model's attention distribution. These maps highlight key areas of interest in the image, as would be expected from human attention, offering a clearer understanding of the model's decision-making process. Figure 3.5 illustrates some examples of these attention maps from the VQA-mhug dataset, Sood et al., 2021b, demonstrating the effectiveness of this approach in generating human-like attention visualizations.

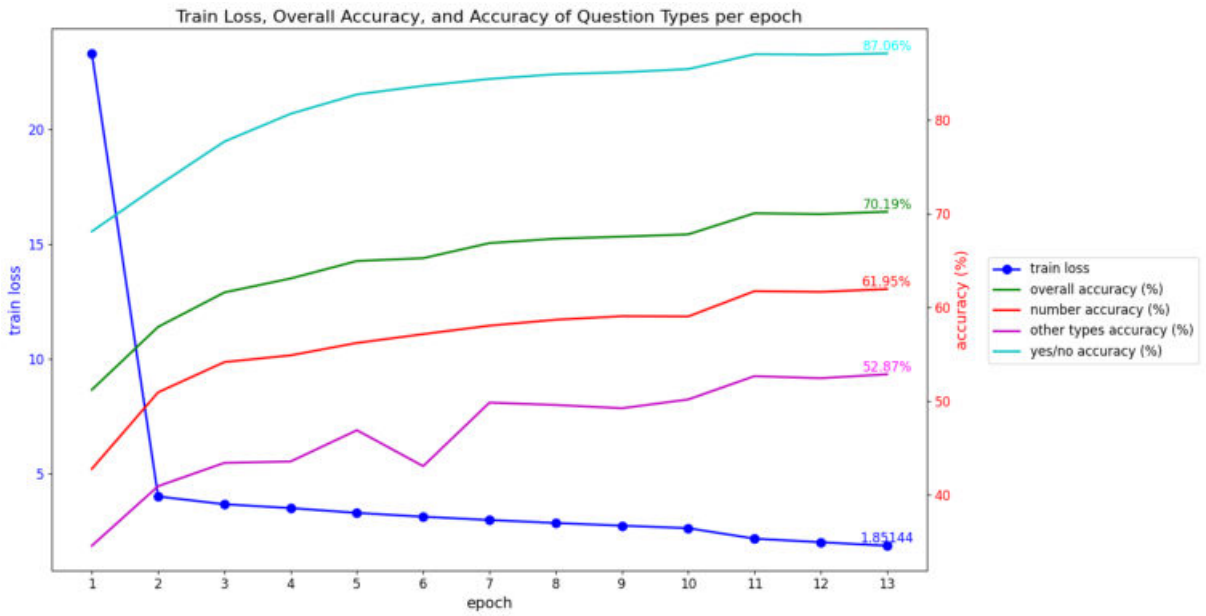

Figure 3.6: MCAN_small train loss and accuracy per epoch using VQAv2 Grid features.

3.2.2 Experiment on MCAN-Grid model

For the implementation of the MCAN model utilizing grid image features, we leverage the extracted grid features as provided in Sood et al., 2021b. Adhering to a similar training regimen as with the region features, we train the MCAN_small model on the VQAv2 dataset, Goyal et al., 2017, over 13 epochs. Upon completion of the training, we run the validation mode (`--RUN = val`) using the model checkpoint from the best performance epoch (epoch 13), with the parameter `--SAVE_ATTMAPS = True` enabled. This setting allows the model to save the attention weights generated during the validation process.

The MCAN model with employing grid features demonstrated performance with an overall accuracy of 70.19% on the `train-val` split. The achieved accuracy of 70.19% with the MCAN model using grid features is commendable, especially when contrasted with the 72.59% accuracy reported by Jiang et al., 2020 on the VQAv2 `test-dev` split. Considering that the `test-dev` split accuracy is typically about 2% higher than that of the `val` split, Jiang et al., 2020, our results are in line with expected performance benchmarks. The training loss reduction to 1.85143 by the final epoch underscores the model's efficiency and robustness in handling grid features.

Figure 3.6 provides a more granular view of the model's performance, detailing the accuracy metrics and illustrating the trend of training loss reduction across epochs. This graphical representation offers a clear insight into the model's learning trajectory and

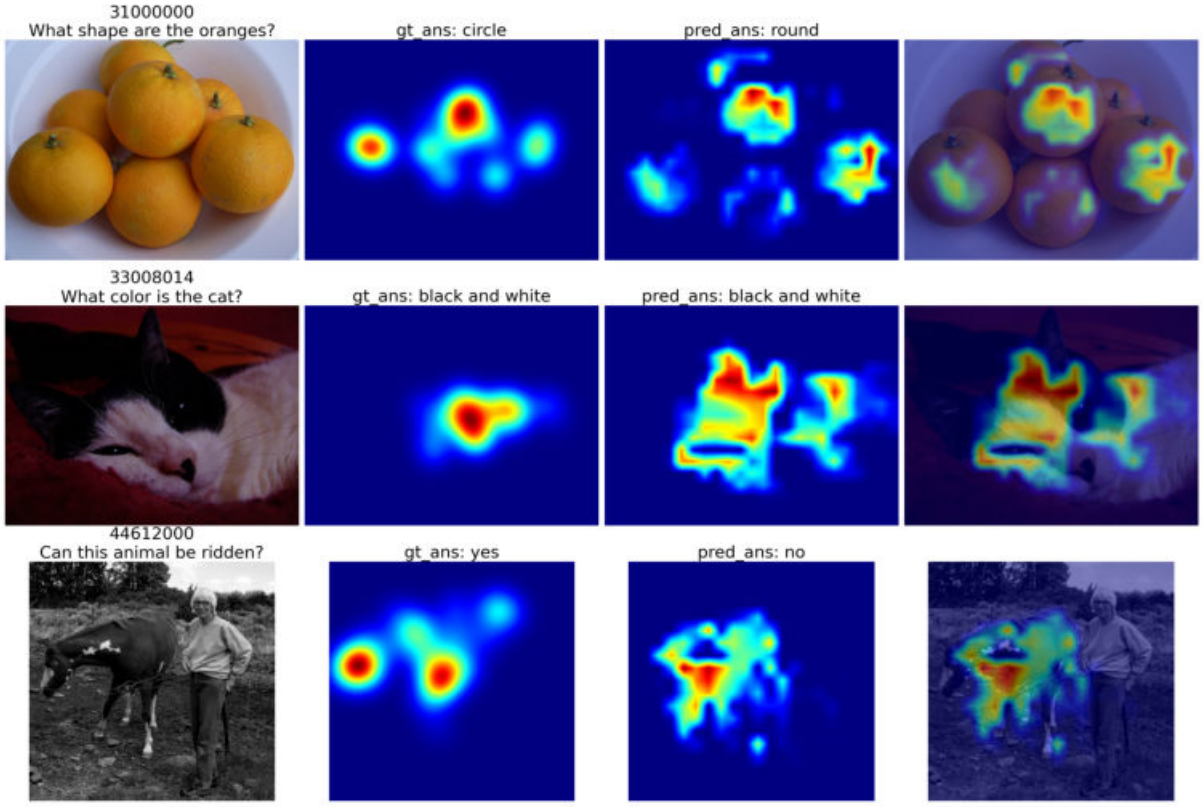

Figure 3.7: Some examples of human and machine Grid attention on VQA-MHUG. From left to right: original image, VQA-MHUG human gazing heatmap, machine Grid attention heatmap, machine Grid attention masked image.

effectiveness, further validating the use of grid features in the MCAN model for Visual Question Answering.

The neural attention weights from the MCAN model with grid features are saved in a plain text file, while the predicted answers for each image-question pair are stored in a `.json` file. The order of the attention weights matches the predicted answers, facilitating easy pairing of attention weights with (I, Q, A_{pred}) tuples. Unlike region features, the grid features' approach means the attention maps are inherently more human-like and don't require Gaussianization. Figure 3.7 showcases some of these attention map results.

The steps of training, validating, and constructing attention maps affirm the validity of our implementation against the official Jiang et al., 2020 models. The detailed comparison and analysis of human and neural attention, using the VQA-MHUG dataset Sood et al., 2021b and the AiR-D dataset Chen et al., 2022, will be discussed in Chapter 4. The methodology will mirror the approach detailed in this section, starting with reproducing VQA-MHUG results before extending the analysis to the AiR-D dataset.

3.3 Machine Attention on AiR-D dataset

In our research, we expanded the approaches mentioned above to the GQA dataset, Hudson et al., 2019, recognizing its foundational role in the AiR-D dataset's, Chen et al., 2022, construction. Our approach was comprehensive, involving not only the separate utilization of region and grid features as inputs to the MCAN model, Yu et al., 2019b, but also an innovative experiment combining both types of features within a single model. This allowed us to examine the potential synergies between different feature representations.

From Section 3.2, we can see that the model accuracy and train loss show not significant decrease during the training process. Thus, given the intensive computational demands of training and the constraints imposed by the thesis timeline, we pragmatically limited our training efforts on the GQA dataset, Hudson et al., 2019, to 3 epochs, carefully balancing the need for demonstrable model performance against the practicalities of time and resource availability.

3.3.1 Experiment on MCAN-Region model

In our study, we utilized the MCAN model, Yu et al., 2019b, as has been established within the MULAN, Sood et al., 2021a, and OpenVQA[5] frameworks, specifically tailored for the GQA dataset, Hudson et al., 2019. This approach allowed us to maintain a consistent methodology, facilitating a direct comparison with results obtained from the previous research. By keeping the model configuration unchanged and focusing on training and testing within the GQA dataset's context, we aimed to extend our analysis and validate the model's effectiveness across different VQA benchmarks.

The GQA dataset, Hudson et al., 2019, offers object-based/region features extracted via Faster-RCNN, Ren et al., 2016, with each image characterized by up to 100 objects/features bounded by pixel dimensions. It includes additional metadata like the number of regions per image (up to 1024), along with the original image's width and height. This comprehensive data supports model training and attention map construction in our study.

The GQA dataset's object-based/region features, provided in `.h5` format[6], necessitated a conversion for compatibility with the MULAN framework and MCAN model. OpenVQA offers a conversion script to transform these `.h5` files into `.npz` files, with each file

[5] https://openvqa.readthedocs.io/en/latest/
[6] https://docs.h5py.org/en/stable/index.html

representing a single image's features. Additionally, OpenVQA shares checkpoints and performance metrics of pre-trained models, including an MCAN model using region features, which achieved an accuracy of 53.41%. This pre-processing step and available resources facilitate the use of GQA dataset features in our experiments.

In our study, we independently trained the MCAN model using region features formatted as `.npz` files within the MULAN framework. By epoch 3, we observed a promising accuracy of 58.59% and a training loss of 1.27436. Given the satisfactory alignment of this performance with the benchmarks set by OpenVQA's implementation, and considering the sufficiently low training loss, we decided to conclude the training phase at this juncture. Subsequently, we utilized the trained model checkpoint to generate attention weights during the validation phase, employing the MULAN framework's functionality to save these weights as plain text files. We then undertook the process of remapping the $1D$ attention weights to $2D$ for both region-specific and Gaussian smoothed attention maps, applying the same Gaussian kernel used in previous analyses. The outcomes of this process, including illustrative examples of the attention maps, are showcased in Figure 3.8 and Figure 3.9, providing visual insights into the model's attention mechanisms.

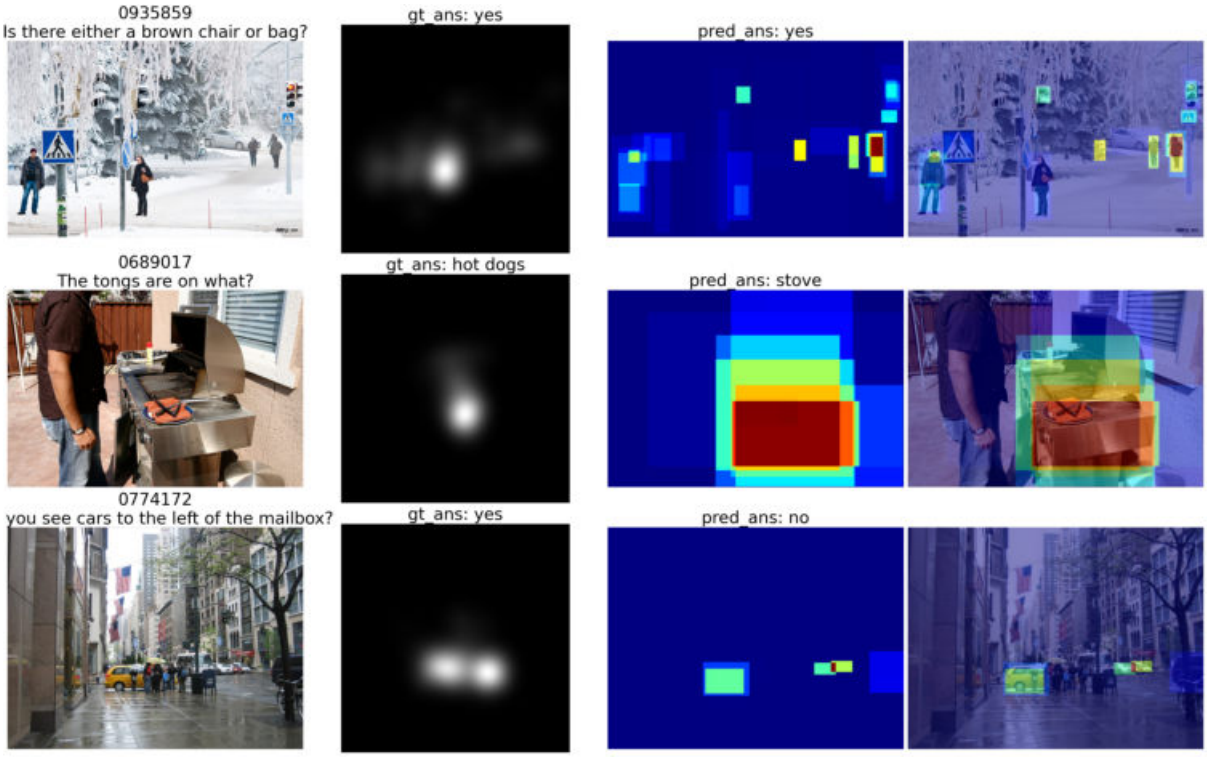

Figure 3.8: Some examples of human and machine Region attention on AiR-D.
From left to right: original image, AiR-D human gazing heatmap, machine region attention heatmap, machine region attention masked image.

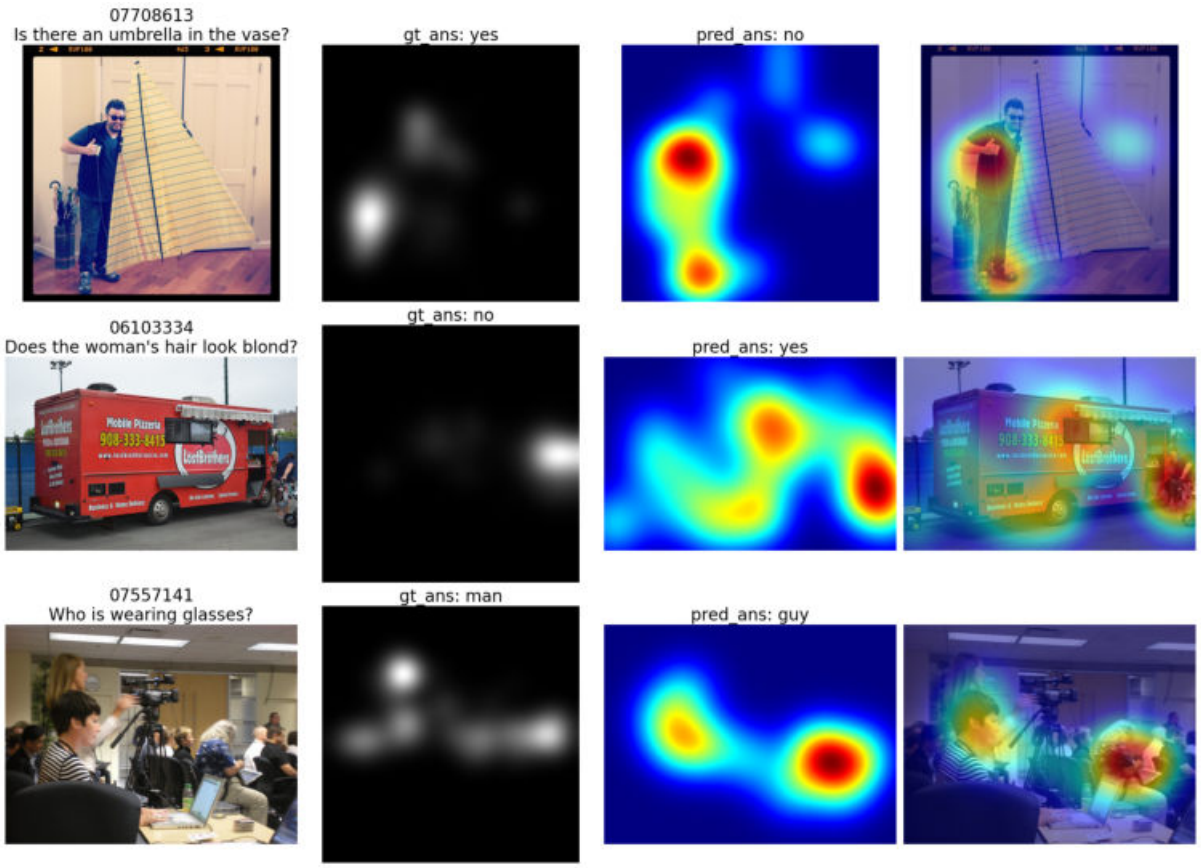

Figure 3.9: Some examples of human and Gaussianized machine Region attention on VQA-MHUG.
From left to right: original image, AiR-D human gazing heatmap, machine region attention heatmap, Gaussianized machine region attention masked image.

3.3.2 Experiment on MCAN-Grid model

The spatial/grid features on the GQA dataset, Hudson et al., 2019, are also provided simultaneously within the dataset. And same as the region features, the grid features are also in `.h5` format and structured into a $(7, 7)$ grid for each image. We applied the same script to transform `.h5` feature files into multiple `.npz` files, each one corresponding to one image. The training process is also stopped at epoch 3. Here we reached the train loss of 1.1639 and an overall accuracy of 55.10%, whereas the same implementation of OpenVQA reports the overall accuracy of 54.28%, Yu et al., 2019a. Table 3.2 shows more detailed accuracy of the training information.

Using the trained checkpoint of epoch 3 and enabling `--SAVE_ATTMAPS = True`, we saved $1D$ attention weights in `.txt` format, subsequently remapping them to $2D$ to match

the original image sizes. Figure 3.10 presents a comparison between these machine-generated attention maps and human attention data from the AiR-D dataset, Chen et al., 2022.

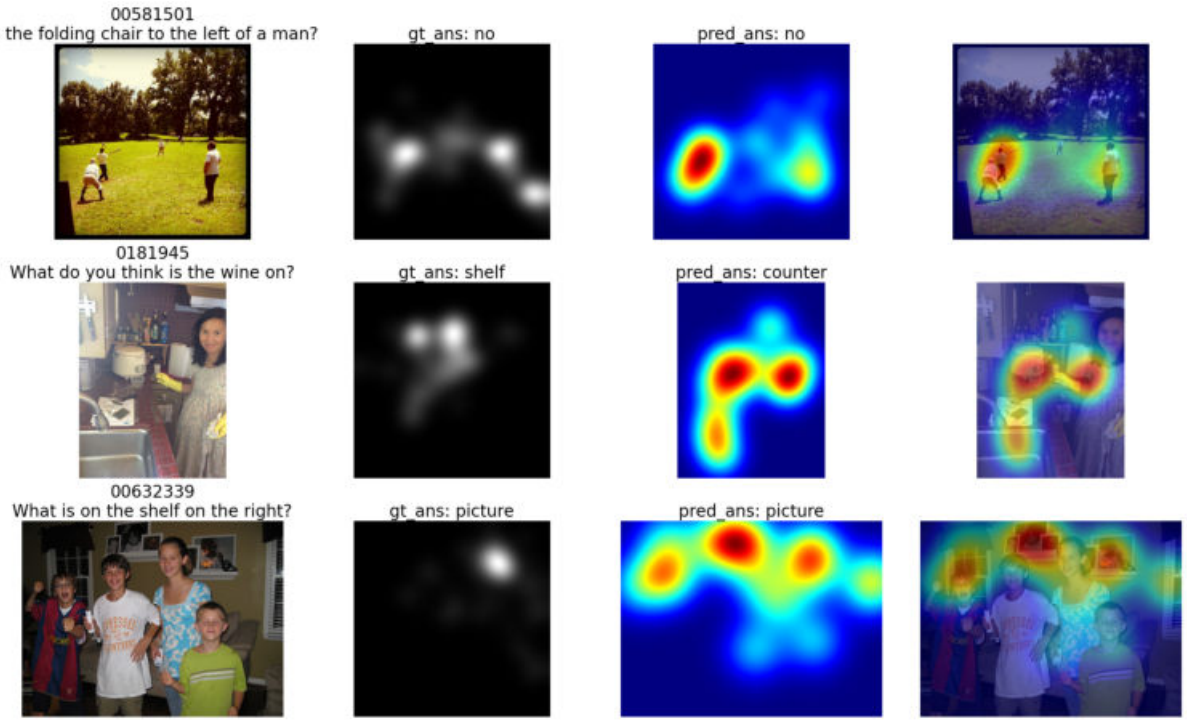

Figure 3.10: Some examples of human and machine Grid attention on AiR-D.
From left to right: original image, AiR-D human gazing heatmap, machine grid attention heatmap, machine grid attention masked image.

3.3.3 Experiment on MCAN Region+Grid model

Drawing on the insights from Lu et al., 2017 and Yu et al., 2019a, we explore a hybrid approach by training a model that utilizes both region and grid features on the GQA dataset Hudson et al., 2019. This approach aims to capitalize on the complementary strengths of both feature types. After training for 3 epochs, we achieve an overall accuracy of 59.62% and a training loss of 1.2609, slightly outperforming the OpenVQA benchmarks. Consequently, we decide to halt further training and proceed to the `--RUN = val` model, ensuring the `--SAVE_ATTMAPS = True` parameter is set to capture the attention weights for subsequent analysis.

Since the region feature map is adopted in this experiment, the initial attention maps are region-like, similar to the ones using region features as input. We thus apply the

Gaussianized kernel to smooth the attention. And then after converting the $1D$ attention weights to $2D$ to match the original image dimensions, we can directly compare machine and human attention areas. Figure 3.11 illustrates several examples.

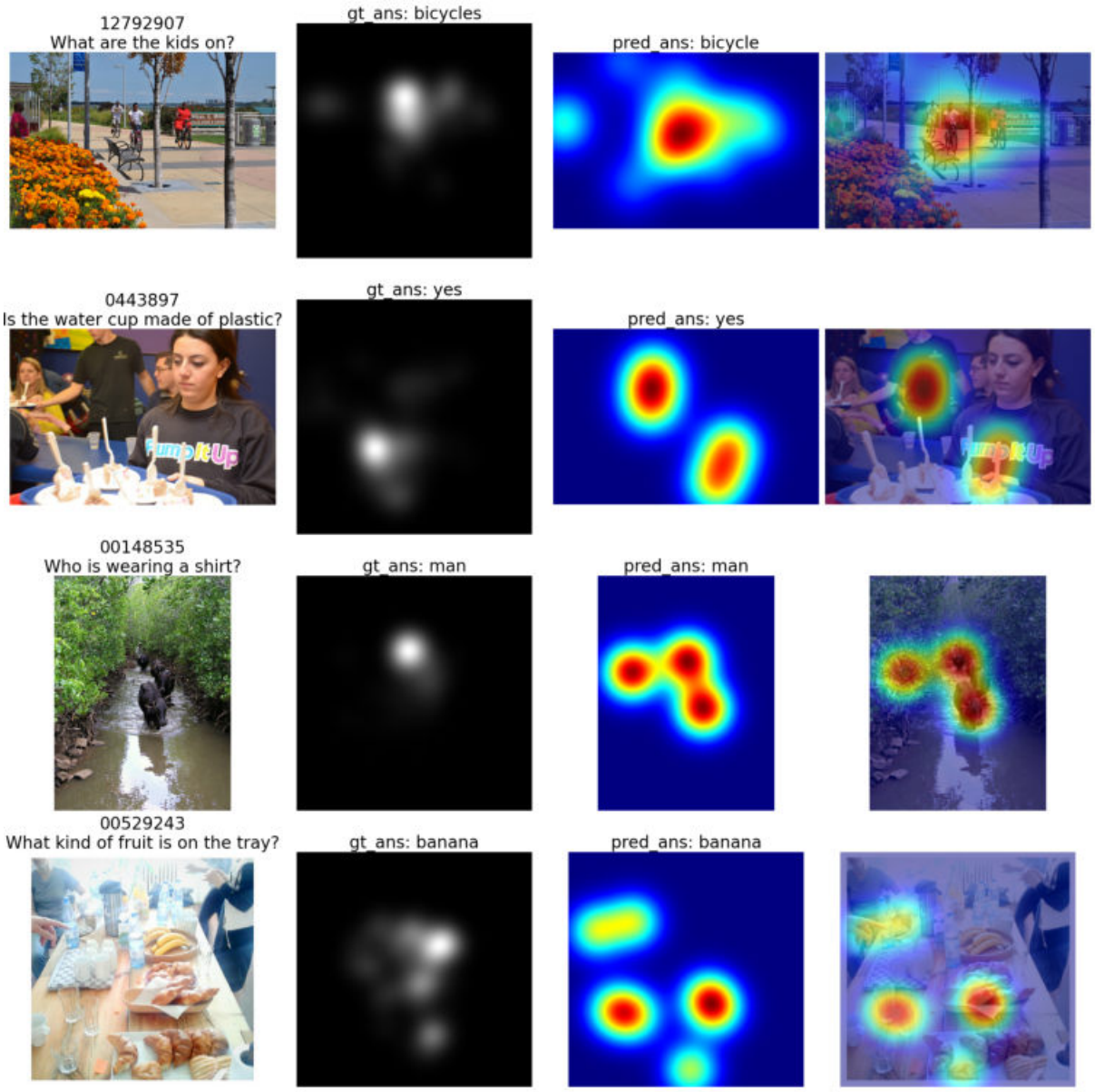

Figure 3.11: Some examples of human and machine Region + Grid attention on AiR-D. From left to right: original image, AiR-D human gazing heatmap, machine attention heatmap, machine attention masked image.

Imple.	Feat.	**Overall** (%) ↑	**Binary** (%) ↑	**Open** (%) ↑	**Validity** (%) ↑
OpenVQA	R	58.20	75.87	**42.66**	**97.01**
	G	54.28	71.68	38.97	96.79
	R + G	**58.38**	**76.49**	42.45	96.98
This thesis	R	58.59	72.83	45.24	94.77
	G	55.70	69.01	43.22	**94.95**
	R + G	**59.62**	**74.75**	**45.44**	94.89

Table 3.2: Accuracy of MCAN_small model on GQA. The upper part is the published accuracy by OpenVQA, Yu et al., 2019a; the lower part is my implementation.

4 Evaluation and Results

4.1 Evaluation Metrics

The primary goal of this thesis is to delve into attention mechanisms within VQA models, focusing on understanding and comparing machine and human attention, rather than surpassing state-of-the-art (SOTA) model performance. Accordingly, the evaluation lies in twofold: first, assessing model performance through conventional metrics, and second, examining the learned attention patterns to discern correlations between machine and human attention processes.

4.1.1 Model performance evaluation metrics

To assess VQA model performance, we employ standard metrics prevalent in deep learning and VQA research, ensuring a comprehensive evaluation of the model's effectiveness in interpreting and responding to visual questions.

Overall accuracy provides a high-level performance score for deep neural network models, calculated as detailed in Equation (4.1), using the validation set of the dataset for evaluation.

$$Acc(\text{pred\_ans}) = \min(\frac{\text{pred\_ans} = \text{gt\_ans}}{3}, 1) \tag{4.1}$$

Individual accuracy. For the VQAv2 dataset, Goyal et al., 2017, the evaluation script not only provides an overall accuracy score but also breaks down the performance into three specific question types: "number", "yes/no", and "other". This detailed scoring allows for a more granular assessment of the model's capabilities in handling different types of questions.

In the case of the GQA dataset, Hudson et al., 2019, the evaluation script offers an even more comprehensive analysis. It categorizes questions into three main types: "structural type", "semantic type", and "steps number". Each of these categories is further subdivided: "structural type" includes "choose", "compare", "logical", "query", and "verify"; "semantic

type" encompasses "attribute", "categorization", "global", "objective", and "relevance"; and "steps number" ranges from 1 to 9 steps. This extensive categorization provides more insights into the model's performance across a diverse array of question complexities and types.

Training loss The training loss is a metric used to assess how a deep learning model fits the training data. That is to say, it assesses the error of the model on the training set. Note that, the training set is a portion of a dataset used to initially train the model. Computationally, the training loss is calculated by taking the sum of errors for each example in the training set. It is also important to note that the training loss is measured after each batch. In our training, the Binary Cross-Entropy (BCE) loss function (as Equation (4.2)) is used and computed for the VQAv2 dataset, and Cross-Entropy (CE) (as Equation (4.3)) is used for the GQA dataset, Hudson et al., 2019.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \tag{4.2}$$

where N is the total number of observations, y_i is the actual label for observation i, which can be either 0 or 1, and $\hat{y}_i$ is the predicted probability that observation i is of class 1.

$$CE = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \tag{4.3}$$

where N is the total number of observations in the dataset, C is the number of classes, $y_{i,c}$ is a binary indicator of whether class c is the correct classification for observation i, and $\hat{y}_{i,c}$ is the predicted probability that observation i is of class c.

4.1.2 Attention maps evaluation metrics

Area Under the Curve (AUC) score is a widely used metric for evaluating the performance of the salience maps. The AUC score is computed from the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

To compute AUC, a saliency map is first treated as a binary classifier of fixations at various threshold values (THRESH). Then the TPR and FPR, representing True Positive Rate and False Positive Rate, are calculated:

$$TPR = \frac{TP}{TP + FN} \tag{4.4}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.5}$$

where TP, FN, FP, and TN stand for True Positive, False Negative, False Positive, and True Negative respectively. For each threshold, the ROC curve can be plotted using the FPR as the x-coordinate and the TPR as the y-coordinate.

In signal detection theory, the Receiver Operating Characteristic (ROC) measures the tradeoff between true and false positives at various discrimination thresholds. And the AUC score is the area under the ROC curve. This can be approximated by summing up the area of trapezoids formed between successive points on the curve.

Spearman's rank correlation (ρ) is used to retrieve the pairwise monotonic relationship between human and machine attention, which is computed as Equation (4.6).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.6}$$

where d_i is the difference between the two ranks of each observation and n is the number of observations.

In alignment with the approach outlined in Sood et al., 2021b, we process both machine-generated attention maps and human gazing heatmaps, resizing them to a uniform 14×14 size. This standardization facilitates a direct comparison, where each machine-generated heatmap is assessed against the corresponding human gazing heatmap, serving as the ground truth. This evaluation covers 3990 samples from the VQA-MHUG dataset, Sood et al., 2021b, and 1454 samples from the AiR-D dataset, Chen et al., 2022, ensuring a comprehensive analysis. Subsequently, we calculate the average of these individual metrics for each dataset, providing a holistic view of the overall performance across the entire dataset of both.

Jensen–Shannon Divergence (jsd) is a method of measuring the similarity between two probability distributions. Here, it provides information about the distances of neural and human attention heatmap distributions. jsd is based on the Kullback–Leibler divergence, and calculated as the Equation (4.7):

$$jsd(P, Q) = (D(P, M) + D(Q, M))/2 \tag{4.7}$$

$$D(P, Q) = \sum_i P_i \log(P_i / Q_i) \tag{4.8}$$

where $D(P, Q)$ is the Kullback–Leibler divergence between distribution P and Q and $M = (P + Q)/2$ is the mixture distribution of P and Q.

Similar to Spearman's rank correlation ρ, we first down-sample the attention maps and human gazing heatmaps to 14×14 and compute the jsd individually on each sample. For each dataset, we also average the individual jsd to get the overall jsd value on the entire dataset.

Attention visualization. In our qualitative analysis, we visualize attention maps to identify highly focused areas. Unlike the approach for MULAN, Sood et al., 2021a, we average raw attention from the model's last encoder and decoder layers across heads. For a better visualized comparison, we also include attention visualizations from the attention reduction module. Raw attention vectors are normalized akin to human-like attention, allowing direct neural attention heatmap visualization. After removing padding, image attention weight vectors are reshaped to the original image's aspect ratio for visualization.

Sood et al., 2021b identified a strong link between model performance enhancement and a higher correlation with human attention. Their analysis used raw neural attention, which lacks a direct input mapping. We aim to extend this by analyzing remapped attention maps, offering a more accurate comparison. Our approach will consider both Spearman's rank correlation (ρ) and Jensen–Shannon divergence (jsd) between model and human-like attention, building on the foundational insights provided by Sood et al., 2021b.

4.2 Results on VQA-MHUG dataset

In this section, we first present the overall results of the VQA-MHUG dataset from the implementation of both ours and Sood et al., 2021b reported. This shows an intuitive view of how our approach for attention map extraction performs. Table 4.1 shows the AUC score (auc), Spearman's rank correlation (ρ), and Jensen–Shannon divergence (jsd) results.

The AUC scores (auc) in our study, which were not reported by Sood et al., 2021b, revealed interesting insights. The Gaussianized region features show lower AUC scores compared to the intuitive region features, and the grid features score the lowest. This pattern suggests that the detailed information provided by region features, such as object positions and bounding boxes, enhances the model's ability to focus attention effectively, reflected in higher AUC scores. The grid features, lacking such detailed object-specific information, may not capture the nuanced aspects of visual attention as effectively, resulting in lower AUC scores. These findings underscore the importance of detailed

	Feature type	$auc \uparrow$	$\rho \uparrow$	$jsd \downarrow$
VQA-MHUG	Region	–	**0.602**	**0.467**
	Grid	–	0.509	0.537
ours	Region	**0.781**	**0.684**	**0.320**
	Guassianized Region	0.772	0.683	0.342
	Grid	0.712	0.648	0.346

Table 4.1: Overall metrics on VQA-MHUG dataset, Sood et al., 2021b; VQA model: MCAN_small, Yu et al., 2019b; total number of samples: 3990.

object information in achieving effective attention alignment. Nevertheless, these values of auc still proved a high relationship between the neural and human attention patterns. This is to say that the MCAN model, Yu et al., 2019b, owns a potentially good attending mechanism compared to humans.

Spearman's rank correlation (ρ) in our implementation, aligning with Sood et al., 2021b, indicates that intuitive region features without Gaussianization correlate better with human gazing patterns than other features. This correlation suggests a shared object-centric approach in image processing by both humans and AI models, where identifying and understanding object attributes and locations take precedence over detailed pixel analysis. This aligns with human cognitive processes where rapid visual assessments typically focus on object recognition rather than scrutinizing every pixel. This object-focused approach in neural networks highlights their potential in mimicking human-like perception, crucial in applications not only the VQA tasks. The values above 0.65 indicate that the neural attention patterns have a strong correlation with human gazing patterns, Akoglu, 2018, on the VQA-MHUG dataset.

The Jensen-Shannon Divergence (jsd) measures the similarity between two probability distributions. In our thesis, human gazing heatmaps serve as the ground truth, compared against the extracted machine attention maps. Our jsd values are lower than those reported by VQA-MHUG Sood et al., 2021b, for reasons not fully understood. However, the trend remains consistent, with lower jsd for region features suggesting neural attention aligns more closely with human patterns when better pre-training methods are employed. This insight could guide future efforts to enhance VQA model performance through more refined approaches.

4.3 Results on AiR-D dataset

In this section, we delve into an overview of our implementation outcomes on the AiR-D dataset, Chen et al., 2022. These numeric values aim to shed light on the efficacy of our methodology for extracting attention maps. The comparative analysis is succinctly encapsulated in Table 4.2, which enumerates the metrics of AUC score (auc), Spearman's rank correlation (ρ), and Jensen–Shannon divergence (jsd), offering a nuanced perspective on the performance and alignment of our approach with human attention patterns.

In our implementation, the AUC scores on the AiR-D dataset, Chen et al., 2022, revealed intriguing patterns. Particularly, the MCAN model, Yu et al., 2019b, with grid features outperformed the combined region and grid features, possibly due to the unique nature of the AiR-D dataset's human gazing heatmaps. These heatmaps, unlike those in the VQA-MHUG dataset, Sood et al., 2021b, are grayscale and lack a focus point threshold, contrasting with our RGB neural attention maps which use a 0.001 threshold. This disparity might explain the lower AUC scores for region features, as attention weights are more evenly distributed across object regions. Despite these findings, the overall strong AUC scores suggest that neural networks are adept at focusing on object regions in a manner akin to human attention, reinforcing the potential of these models in tasks requiring visual attention.

Spearman's rank correlation (ρ) on the AiR-D dataset, Chen et al., 2022, suggests a better correlation between neural attention using grid features and human gazing patterns compared to other feature types. With values exceeding 0.65, there's a strong indication of significant alignment between neural attention mechanisms and human visual behavior, supporting the findings in Akoglu, 2018 regarding the strength of such correlations.

The Jensen-Shannon Divergence (jsd) values on the AiR-D dataset, Chen et al., 2022, are notably higher than those observed on the VQA-MHUG dataset, Sood et al., 2021b, with values exceeding 0.5 suggesting a lesser degree of similarity between neural attention and human gazing patterns than anticipated. The specific cause of this discrepancy remains unclear to us yet.

Feature type	$auc \uparrow$	$\rho \uparrow$	$jsd \downarrow$
Region	0.710	0.639	0.558
Guassianized Region	0.789	**0.694**	0.512
Grid	**0.792**	0.682	**0.511**
Region + Grid	0.680	0.656	0.567

Table 4.2: Overall metrics on AiR-D dataset, Chen et al., 2022; VQA model: MCAN_small, Yu et al., 2019b; total number of samples: $1,454$. The best results are shown in bold.

5 Discussion

5.1 Human vs. neural attention in VQA

The comparison of metrics (Chapter 4) and visualized attention maps (Chapter 3) supports Sood et al., 2021b's assertion of a statistically significant correlation between human and machine attention in terms of Spearman's correlation and jsd. This suggests that while neural attention patterns bear similarities to human attention, they are still not identical.

The study of human attention spans multiple disciplines, including computer science, biology, psychology, and cognitive sciences. Despite advances, many aspects of human vision remain elusive, lacking a unified theoretical framework for human visual attention mechanisms. As noted by Yang, 2020, deepening the understanding of human attention through research and modeling is crucial for enhancing computer information processing, suggesting a need for interdisciplinary efforts to bridge gaps between human cognitive processes and computational models.

The exploration in this thesis of neural visual attention through Vision Transformer architecture and Human-Like attention mechanisms underscores the nuanced ways in which AI models can interpret visual information. By comparing these models with human attention patterns, we gain valuable insights into their similarities and differences, enhancing our understanding of AI in visual and language co-processing tasks. The choice of appropriate weighting methods is crucial for aligning AI models more closely with human observation patterns. Additionally, the integration of diverse image features into AI models is a significant step towards better simulating human-like performance and attention mechanisms, presenting expansive opportunities for research and development in the field of machine vision.

5.1.1 More discussion based on the metric results

In Table 4.1, the superior correlation (ρ) and Jensen-Shannon Divergence (jsd) scores observed with region-based features can be attributed to the pre-trained model's inherent knowledge of image details like region positions, objects, and bounding boxes. Despite

employing the same scientific methods as VQA-MHUG Sood et al., 2021b, variations in metrics could result from differences in the number of training epochs and other subtle implementation parameters. To validate this hypothesis and gain a clearer understanding, a comprehensive analysis comparing the metrics of neural attention maps across various epochs in both implementations could be insightful. Such a comparative study could reveal how model training intricacies influence the alignment of neural attention with human attention.

In Table 4.2, the observation that neural attention maps with grid features outperform those with region-based or combined features in terms of correlation (ρ) and Jensen-Shannon Divergence (jsd) resonates with the findings of Jiang et al., 2020. Although these high values suggest a significant correlation between neural and human attention, the underlying reasons remain intriguing and not entirely explained by existing research.

Upon reviewing the visualized neural attention maps, it appears that maps with region and grid features exhibit a sparser distribution of focus points compared to those using solely grid features. This sparser distribution could potentially enhance model performance while paradoxically yielding lower scores in conventional neural attention map metrics. If this observation is accurate, it suggests the need for more nuanced metrics that better align with actual model performance, surpassing the capabilities of traditional Spearman's rank correlation and Jensen-Shannon Divergence in analyzing neural attention maps. Such new metrics would be instrumental in providing a more accurate assessment of how closely AI models' attention patterns align with human attention, especially in complex visual question answering tasks.

The close alignment of metric values in the implementations of this thesis on both VQA-MHUG, Sood et al., 2021b, and AiR-D, Chen et al., 2022, datasets suggests a high correlation between them. This consistency indicates that both datasets are robust and reliable for further analysis in studying the parallels and differences between human and neural attention mechanisms. The similar outcomes across these two distinct datasets reinforce their validity as tools for in-depth attention analysis in VQA research.

5.2 Experiments on CQA

The effort of extending the analysis pipeline to Chart Question Answering (CQA) models, inspired by Masry et al., 2022, which introduces a benchmark along with a dedicated dataset and two CQA models utilizing attention mechanisms for feature encoding-decoding and answer generation, is also conducted. The intuitive attention flow extraction method is used in for ChartQA models,Masry et al., 2022 focusing on the

attention weights from the last layer. The unique aspect of CQA models, where image features are extracted via a ChartOCR model and converted into tabular data, posed challenges for attention extraction and reconstruction, complicating the analysis process. Re-implementing the ChartQA models presents its own set of challenges as well, complicating the efforts to replicate and analyze the models effectively.

Beyond the ChartQA benchmark, Masry et al., 2022, I also explore the MatCha-ChartQA model[1], which is an adaptation of the broader MatCha framework, Liu et al., 2022, tailored for CQA tasks. Although time constraints hindered a thorough examination of the MatCha framework and its specific mechanisms for attention extraction, I recognize its potential. The MatCha framework, with its capacity for fine-tuning and application across diverse vision-language tasks, presents a valuable avenue for future research, warranting deeper investigation into its capabilities and methodologies.

The exploration into attention mechanisms and model explainability within Chart Question Answering (CQA) tasks has been limited, highlighting a potentially underdeveloped area in CQA research. This gap presents an opportunity for significant advancements in understanding and improving CQA models through deeper investigation into their interpretability and decision-making processes.

5.3 Future Work

This thesis is constrained by time, limiting the attention analysis to the VQA-MHUG, Sood et al., 2021b, and AiR-D datasets, Chen et al., 2022 using the MCAN_small model, Yu et al., 2019b. This methodology could extend to other models within the MULAN framework, Sood et al., 2021a, and even to additional VQA models like LXMERT, Tan et al., 2019, offering broader insights into neural network attention mechanisms and enhancing model trust and explainability.

In addition to the focus on VQA models, the attention mechanisms and model explainability within Chart Question Answering (CQA) remain underdeveloped. Advancing research in this area is crucial for a deeper understanding of how machines process and interpret both images and text in CQA tasks. Exploring these mechanisms can provide valuable insights into the cognitive processes that AI models emulate when analyzing data representations like charts, thereby broadening the comprehension of AI's capabilities in interpreting complex visual and textual information. This line of research could lead to more sophisticated and interpretable CQA models, enhancing their utility in various applications.

[1] https://huggingface.co/google/matcha-chartqa

6 Conclusion

This thesis replicates the MCAN model implementation, Yu et al., 2019b, within the MULAN framework, Sood et al., 2021a, on the VQA-MHUG dataset, Goyal et al., 2017, as a starting point. The accuracy and correctness of the re-implementation are verified against official versions. Using the MULAN toolkit, $1D$ attention weights are transformed into $2D$ neural attention maps. These maps are then compared with VQA-MHUG's human-gazing heatmaps using Area Under ROC Curve auc, Spearman's Rank Correlation ρ, and Jensen-Shannon Divergence jsd metrics.

The approach is then extended to the AiR-D dataset, Chen et al., 2022, involving training on the GQA dataset, Hudson et al., 2019, extracting neural attention maps, and applying the same metrics to compare with human gazing heatmaps.

The thesis concludes with a discussion on the differences between the official and re-produced implementations, along with insights into the interpretability of neural attention in VQA models.

Bibliography

Abnar, S., W. Zuidema (2020). *Quantifying Attention Flow in Transformers*. arXiv: 2005.00928 [cs.LG] (cit. on p. 3).

Agrawal, A., J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh (2016). *VQA: Visual Question Answering*. arXiv: 1505.00468 [cs.CL] (cit. on pp. 8, 9).

Akoglu, H. (2018). "User's guide to correlation coefficients." In: *Turkish Journal of Emergency Medicine* 18.3, pp. 91–93. ISSN: 2452-2473. DOI: https://doi.org/10.1016/j.tjem.2018.08.001. URL: https://www.sciencedirect.com/science/article/pii/S2452247318302164 (cit. on pp. 39, 40).

Anderson, P., X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang (2018). *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*. arXiv: 1707.07998 [cs.CV] (cit. on p. 18).

Bahdanau, D., K. Cho, Y. Bengio (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv: 1409.0473 [cs.CL] (cit. on pp. 5, 6).

Chau, S. L., R. Hu, J. Gonzalez, D. Sejdinovic (2022). *RKHS-SHAP: Shapley Values for Kernel Methods*. arXiv: 2110.09167 [stat.ML] (cit. on pp. 12, 13).

Chen, S., M. Jiang, J. Yang, Q. Zhao (2022). *Attention in Reasoning: Dataset, Analysis, and Modeling*. arXiv: 2204.09774 [cs.CV] (cit. on pp. 2, 3, 10, 27, 28, 31, 37, 40, 41, 44, 45, 47, 53, 56).

Devlin, J., M.-W. Chang, K. Lee, K. Toutanova (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL] (cit. on p. 7).

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs.CV] (cit. on pp. 1, 6).

Goyal, Y., T. Khot, D. Summers-Stay, D. Batra, D. Parikh (2017). *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering*. arXiv: 1612.00837 [cs.CV] (cit. on pp. 2, 3, 9, 21, 26, 35, 47).

Graves, A., G. Wayne, I. Danihelka (2014). *Neural Turing Machines*. arXiv: 1410.5401 [cs.NE] (cit. on p. 5).

Hochreiter, S., J. Schmidhuber (Nov. 1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 1).

Hudson, D. A., C. D. Manning (2019). *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*. arXiv: 1902.09506 [cs.CL] (cit. on pp. 2, 3, 9, 28, 30, 31, 35, 36, 47).

Jiang, H., I. Misra, M. Rohrbach, E. Learned-Miller, X. Chen (2020). *In Defense of Grid Features for Visual Question Answering*. arXiv: 2001.03615 [cs.CV] (cit. on pp. 8, 20, 26, 27, 44).

Johnson, J., B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, R. Girshick (2016). *CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning*. arXiv: 1612.06890 [cs.CV] (cit. on p. 10).

Kafle, K., B. Price, S. Cohen, C. Kanan (2018). *DVQA: Understanding Data Visualizations via Question Answering*. arXiv: 1801.08163 [cs.CV] (cit. on p. 11).

Kahou, S. E., V. Michalski, A. Atkinson, A. Kadar, A. Trischler, Y. Bengio (2018). *FigureQA: An Annotated Figure Dataset for Visual Reasoning*. arXiv: 1710.07300 [cs.CV] (cit. on p. 11).

Krajna, A., M. Kovac, M. Brcic, A. Šarčević (2022). "Explainable Artificial Intelligence: An Updated Perspective." In: *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pp. 859–864. DOI: 10.23919/MIPRO55190.2022.9803681 (cit. on pp. 3, 12).

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, F.-F. Li (2016). *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*. arXiv: 1602.07332 [cs.CV] (cit. on pp. 9, 18, 19).

Lin, T.-Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár (2015). *Microsoft COCO: Common Objects in Context*. arXiv: 1405.0312 [cs.CV] (cit. on p. 9).

Liu, F., F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, J. M. Eisenschlos (2022). *MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering*. arXiv: 2212.09662 [cs.CL] (cit. on p. 45).

Lu, P., H. Li, W. Zhang, J. Wang, X. Wang (2017). *Co-attending Free-form Regions and Detections with Multi-modal Multiplicative Feature Embedding for Visual Question Answering*. arXiv: 1711.06794 [cs.CV] (cit. on pp. 8, 31).

Malinowski, M., M. Fritz (2015). *A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input*. arXiv: 1410.0210 [cs.AI] (cit. on p. 9).

Masry, A., X. L. Do, J. Q. Tan, S. Joty, E. Hoque (May 2022). "ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning." In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 2263–2279. DOI: 10.18653/v1/2022.findings-

acl.177. URL: https://aclanthology.org/2022.findings-acl.177 (cit. on pp. 6, 11, 44, 45).

Nathan Silberman Derek Hoiem, P. K., R. Fergus (2012). "Indoor Segmentation and Support Inference from RGBD Images." In: *ECCV* (cit. on p. 9).

Pennington, J., R. Socher, C. Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by A. Moschitti, B. Pang, W. Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162 (cit. on pp. 1, 2).

Peters, B., V. Niculae, A. F. T. Martins (2019). *Sparse Sequence-to-Sequence Models*. arXiv: 1905.05702 [cs.CL] (cit. on p. 5).

Ren, S., K. He, R. Girshick, J. Sun (2016). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. arXiv: 1506.01497 [cs.CV] (cit. on pp. 2, 18–20, 22, 23, 28).

Ribeiro, M. T., S. Singh, C. Guestrin (2016). "*Why Should I Trust You?*": *Explaining the Predictions of Any Classifier*. arXiv: 1602.04938 [cs.LG] (cit. on pp. 12, 13).

Rozemberczki, B., L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, R. Sarkar (2022). *The Shapley Value in Machine Learning*. arXiv: 2202.05594 [cs.LG] (cit. on pp. 12, 13).

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra (Oct. 2019). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In: *International Journal of Computer Vision* 128.2, 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7 (cit. on p. 1).

Sood, E., F. Kögel, P. Müller, D. Thomas, M. Bace, A. Bulling (2021a). *Multimodal Integration of Human-Like Attention in Visual Question Answering*. arXiv: 2109.13139 [cs.CV] (cit. on pp. 2, 3, 9, 20–22, 28, 38, 45, 47).

Sood, E., F. Kögel, F. Strohm, P. Dhar, A. Bulling (2021b). *VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering*. arXiv: 2109.13116 [cs.CV] (cit. on pp. 2, 10, 15, 21, 25–27, 37–40, 43–45, 53, 56).

Tan, H., M. Bansal (2019). *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. arXiv: 1908.07490 [cs.CL] (cit. on pp. 8, 45).

Thomee, B., D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li (Jan. 2016). "YFCC100M: the new data in multimedia research." In: *Communications of the ACM* 59.2, 64–73. ISSN: 1557-7317. DOI: 10.1145/2812802. URL: http://dx.doi.org/10.1145/2812802 (cit. on p. 9).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin (2017). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL] (cit. on pp. 1, 5, 6, 16–18).

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio (2016). *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. arXiv: 1502.03044 [cs.LG] (cit. on p. 5).

Yang, X. (2020). "An Overview of the Attention Mechanisms in Computer Vision." In: *Journal of Physics: Conference Series* 1693.1, p. 012173. DOI: 10.1088/1742-6596/1693/1/012173. URL: https://dx.doi.org/10.1088/1742-6596/1693/1/012173 (cit. on p. 43).

Yang, Z., X. He, J. Gao, L. Deng, A. Smola (2016). *Stacked Attention Networks for Image Question Answering*. arXiv: 1511.02274 [cs.LG] (cit. on p. 8).

Yu, Z., Y. Cui, Z. Shao, P. Gao, J. Yu (2019a). *OpenVQA*. https://github.com/MILVLG/openvqa (cit. on pp. 30, 31, 33).

Yu, Z., J. Yu, Y. Cui, D. Tao, Q. Tian (2019b). *Deep Modular Co-Attention Networks for Visual Question Answering*. arXiv: 1906.10770 [cs.CV] (cit. on pp. 2, 3, 6, 8, 16–23, 28, 39–41, 45, 47).

Yu, Z., J. Yu, J. Fan, D. Tao (2017). *Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering*. arXiv: 1708.01471 [cs.CV] (cit. on p. 8).

A Appendix

A.1 Region and Gaussianized Region attention maps

This section provides some examples of the region and Gaussianized region machine attention maps on the VQA-MHUG, Sood et al., 2021b, and AiR-D dataset, Chen et al., 2022. From left to right, each figure shows the original image, human gazing heatmap, region masked image, and Gaussianized region masked image.

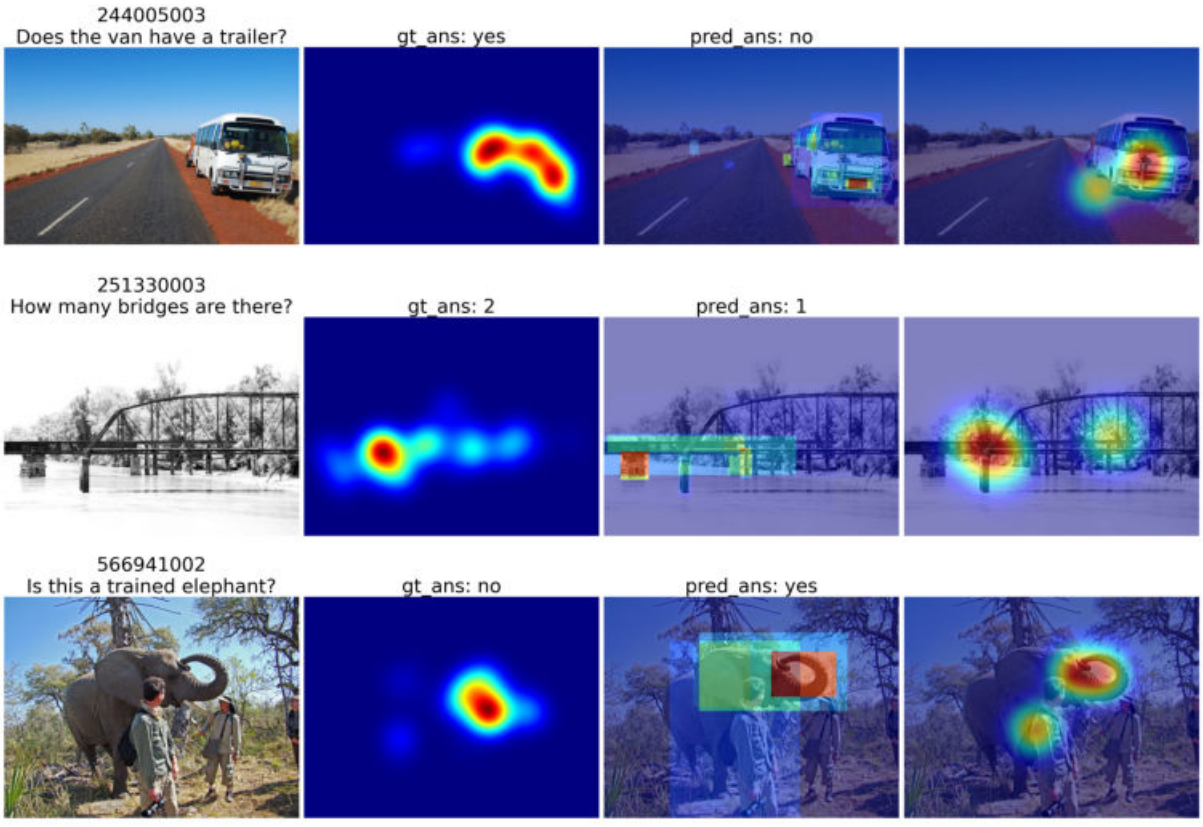

Figure A.1: Region and Gaussianized Region attention maps on VQA-MHUG dataset.

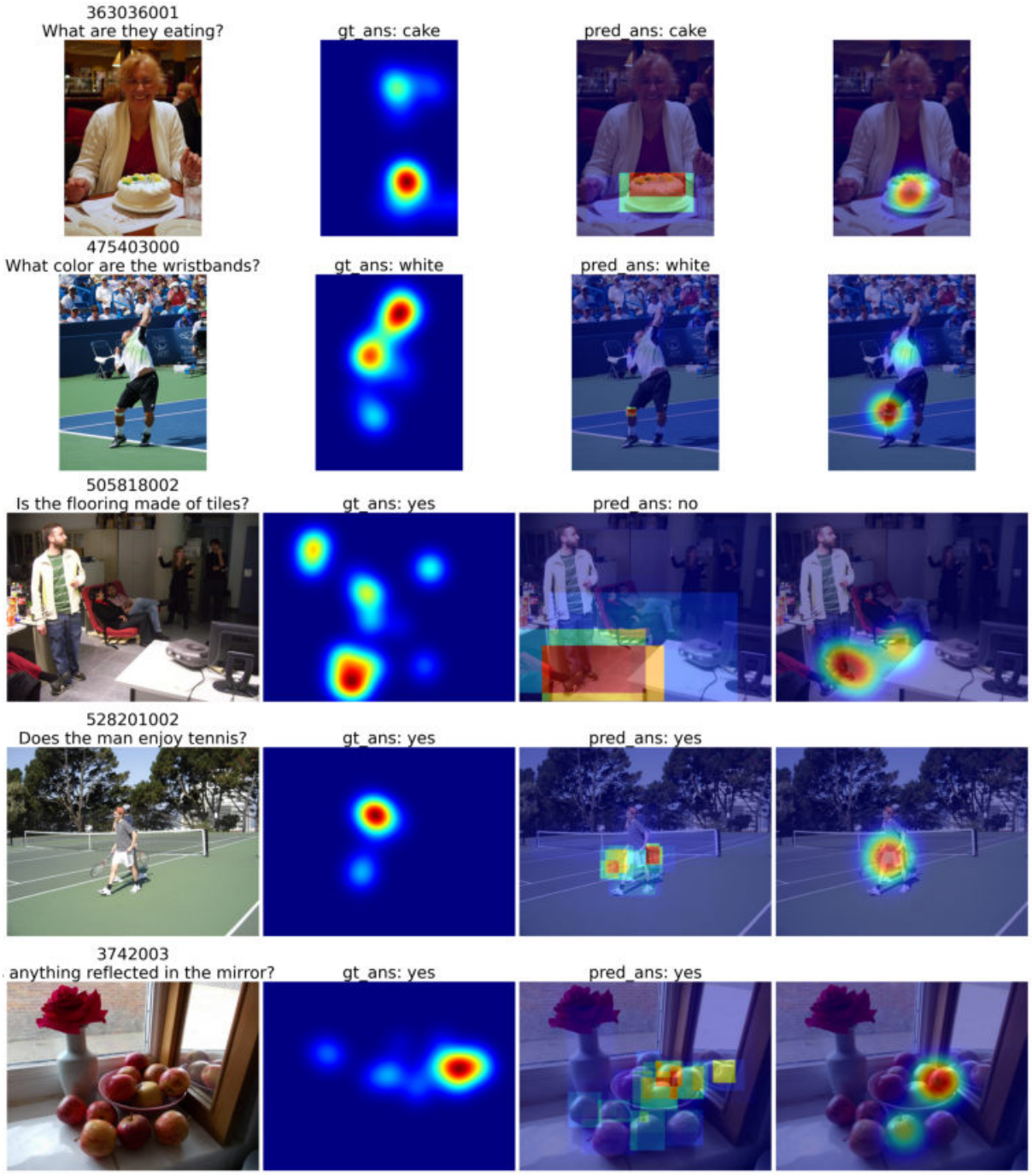

Figure A.2: Region and Gaussianized Region attention maps on VQA-MHUG dataset.

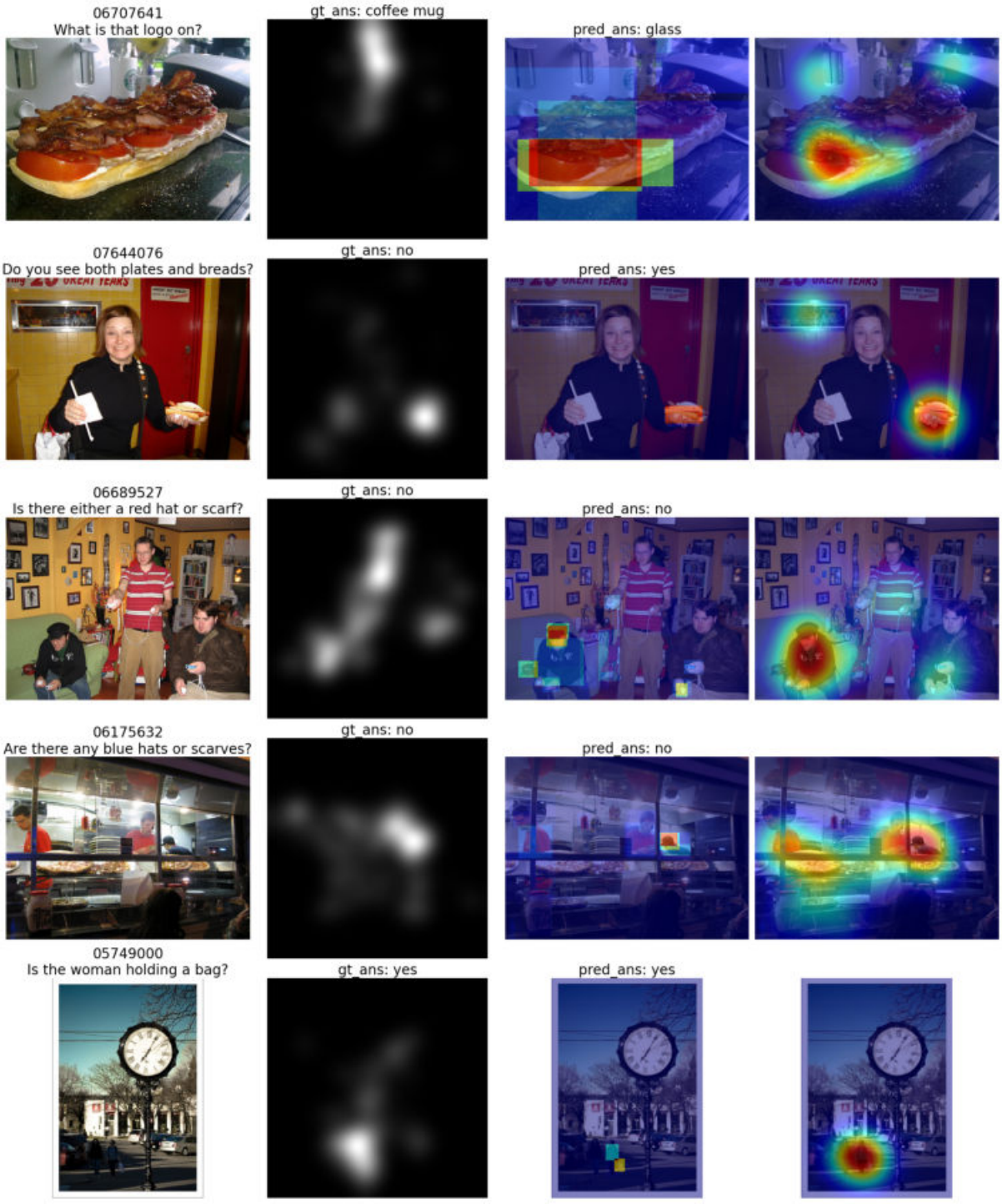

Figure A.3: Region and Gaussianized Region attention maps on AiR-D dataset.

A.2 More examples of machine attention maps

This section provides more examples of the human and neural attention maps on the VQA-MHUG dataset, Sood et al., 2021b, and AiR-D dataset, Chen et al., 2022. From left to right, each figure has the original image, human gazing heatmap, neural attention map, and masked image.

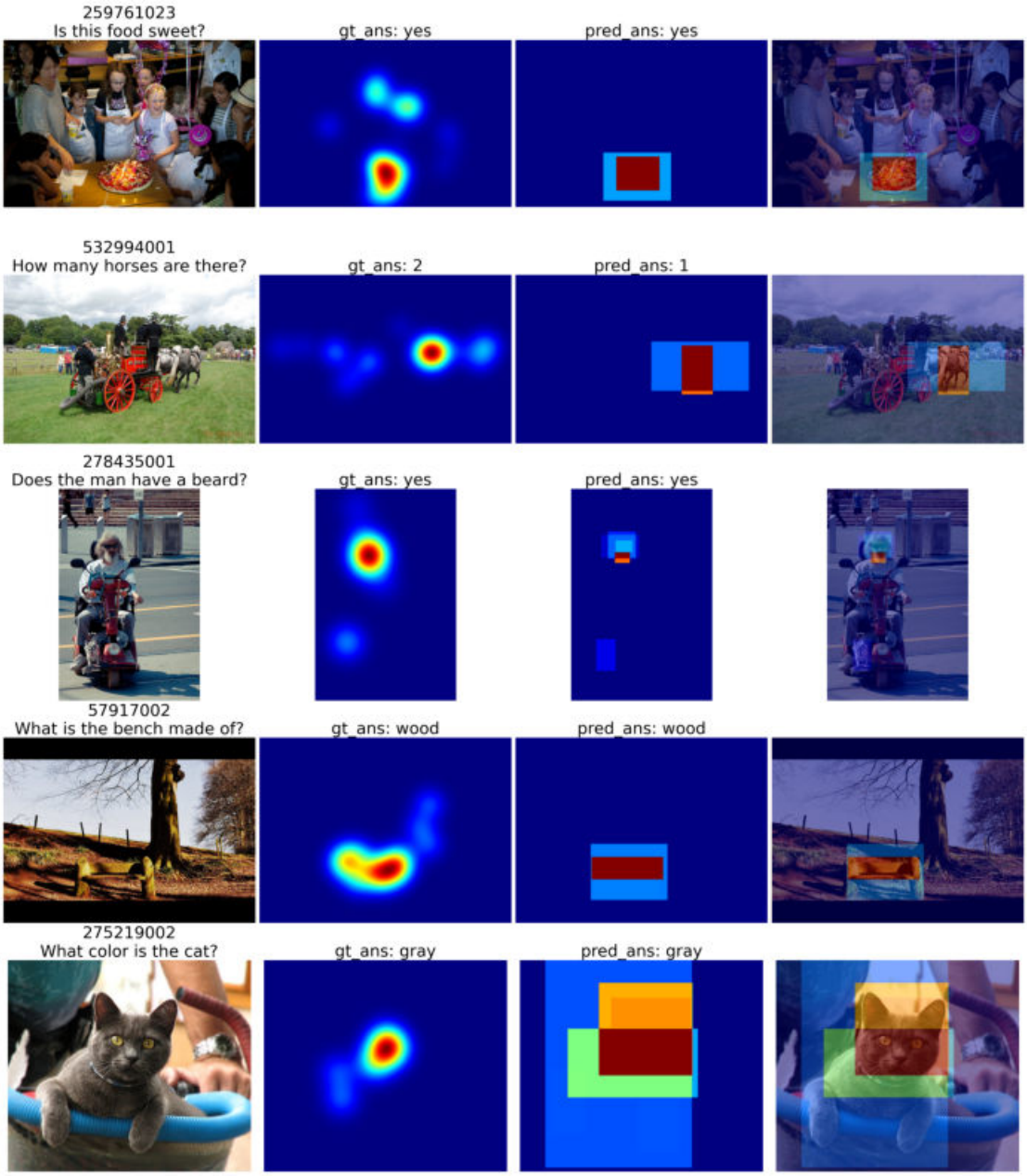

Figure A.4: More example of Region attention maps on VQA-MHUG.

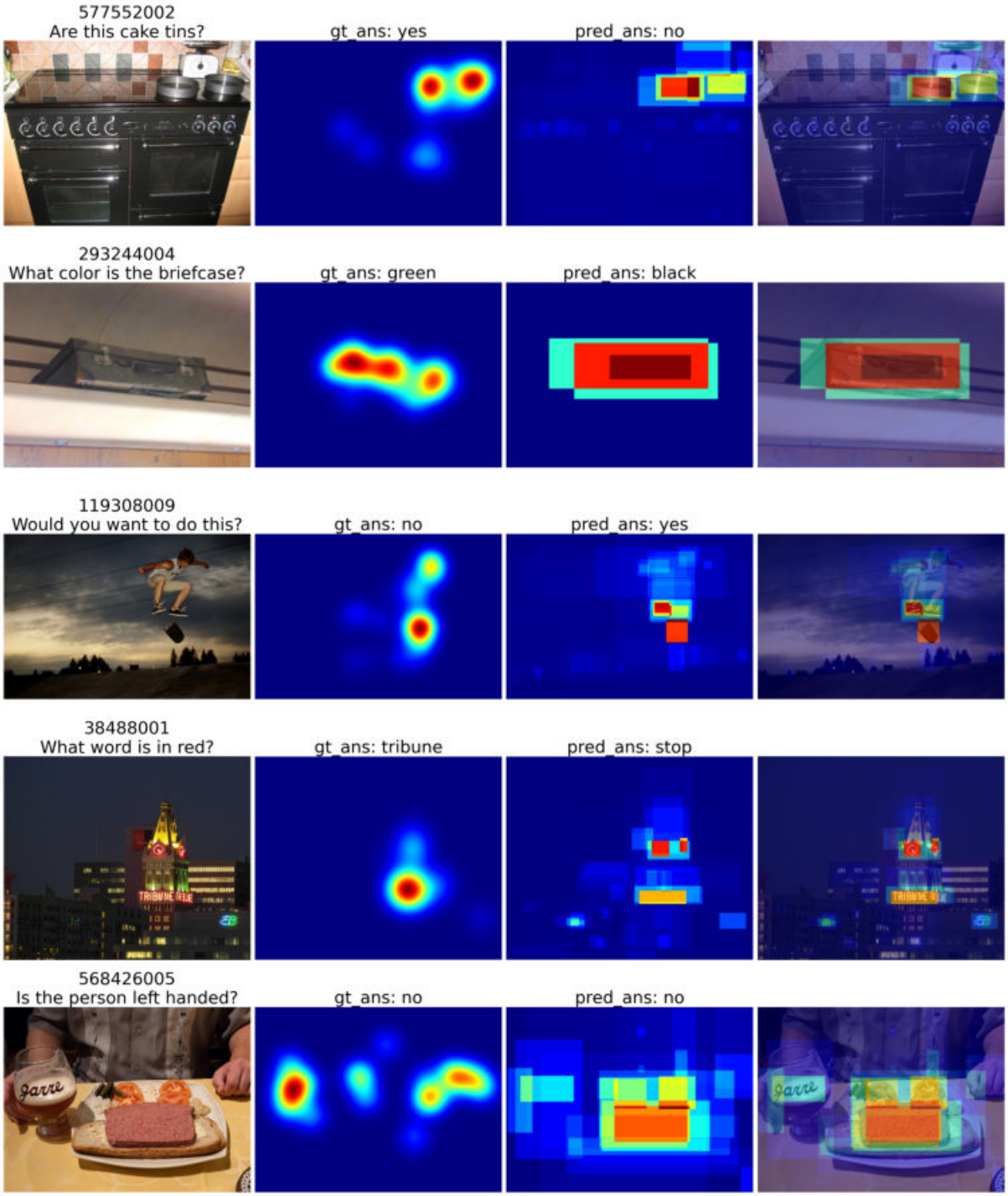

Figure A.5: More example of Region attention maps on VQA-MHUG.

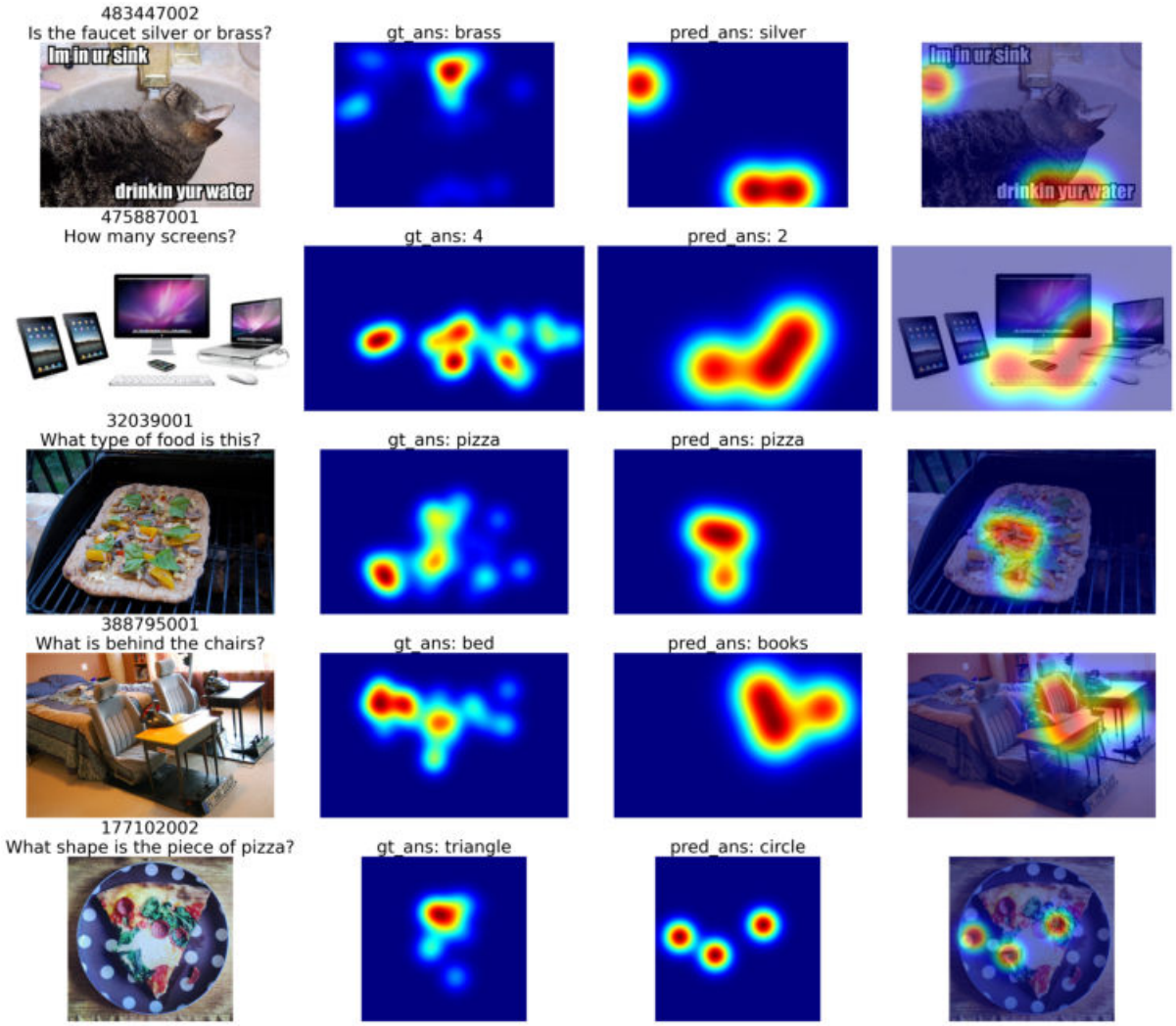

Figure A.6: More example of Gaussianized Region attention maps on VQA-MHUG.

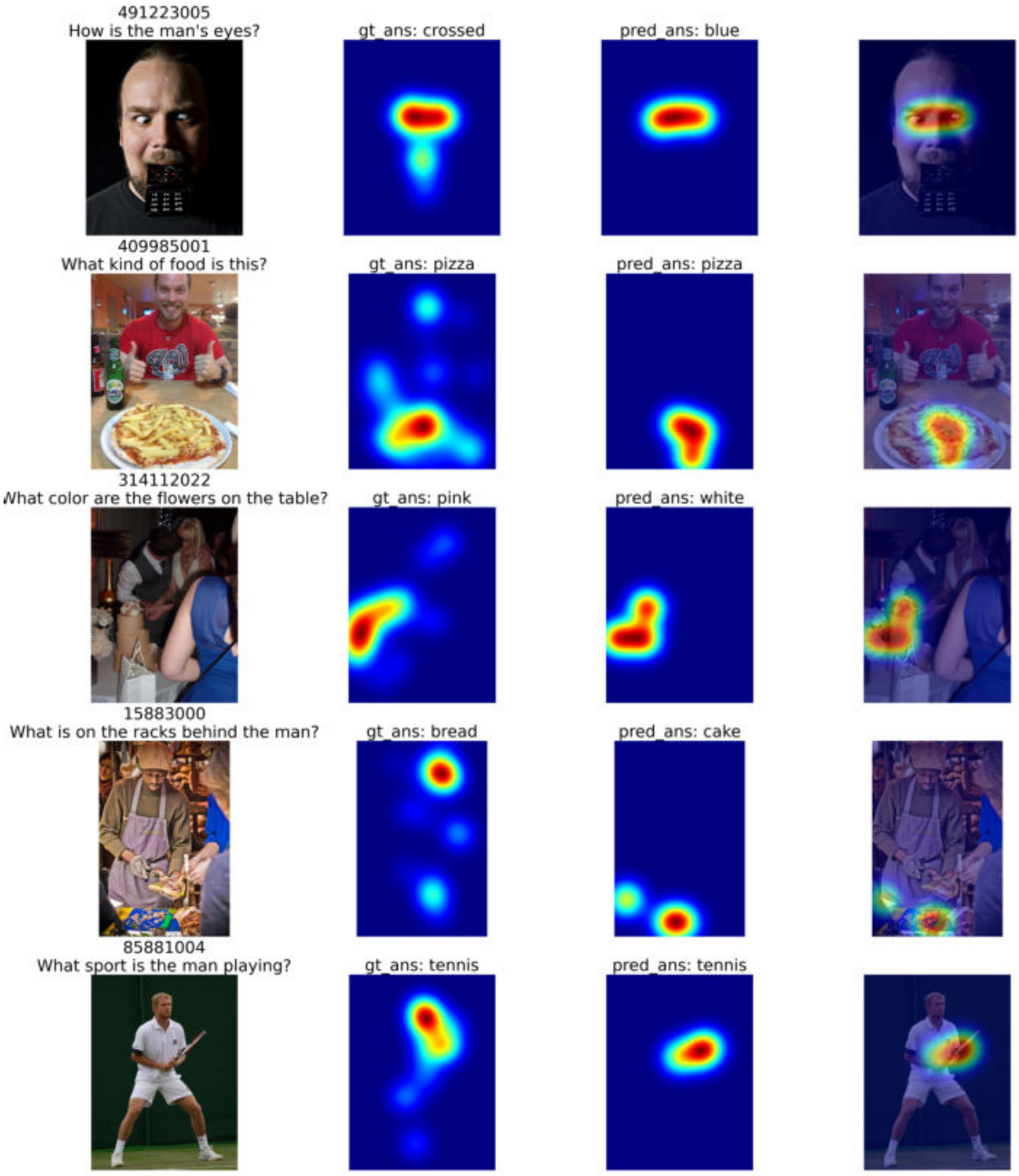

Figure A.7: More example of Gaussianized Region attention maps on VQA-MHUG.

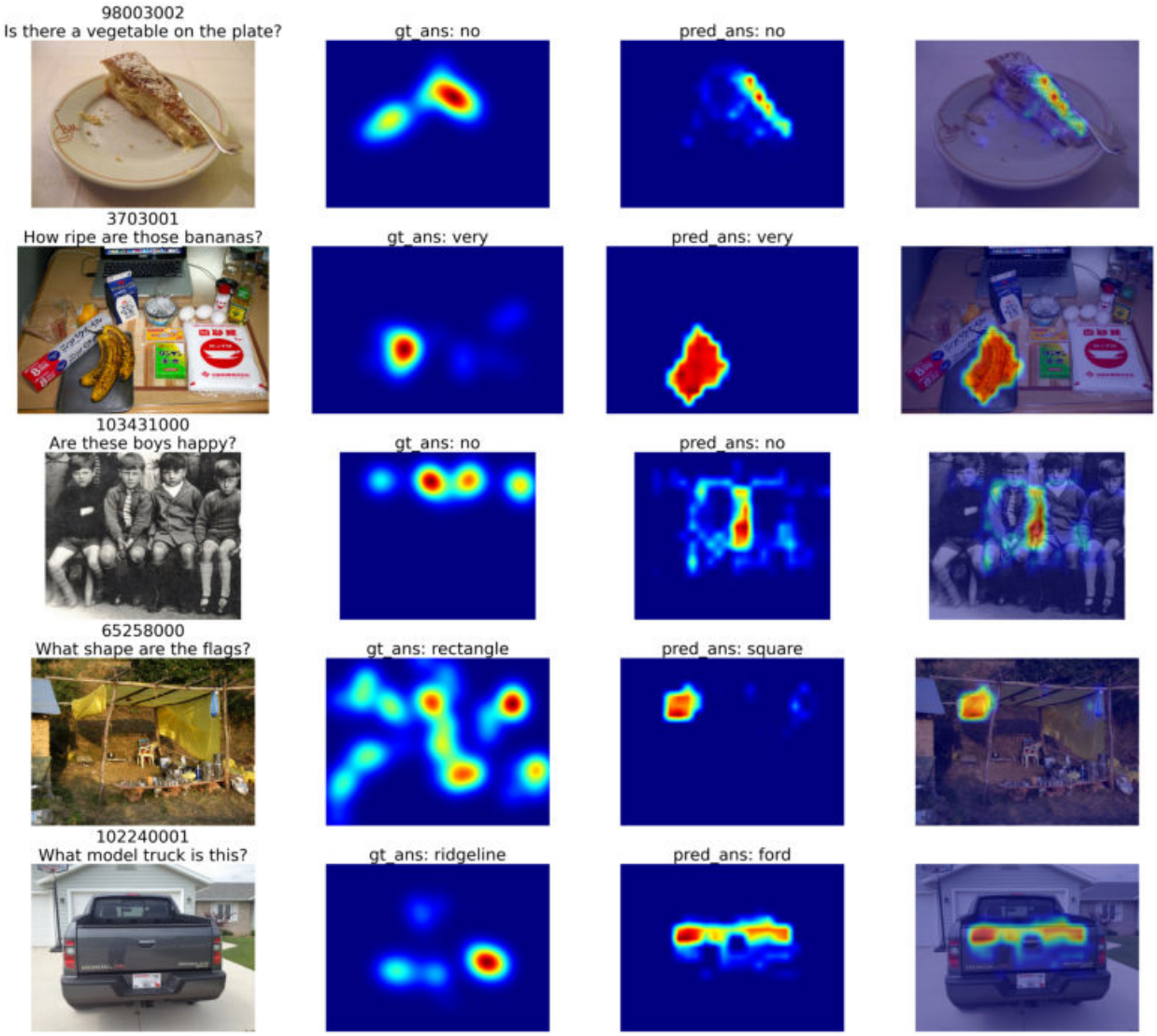

Figure A.8: More example of Grid attention maps on VQA-MHUG.

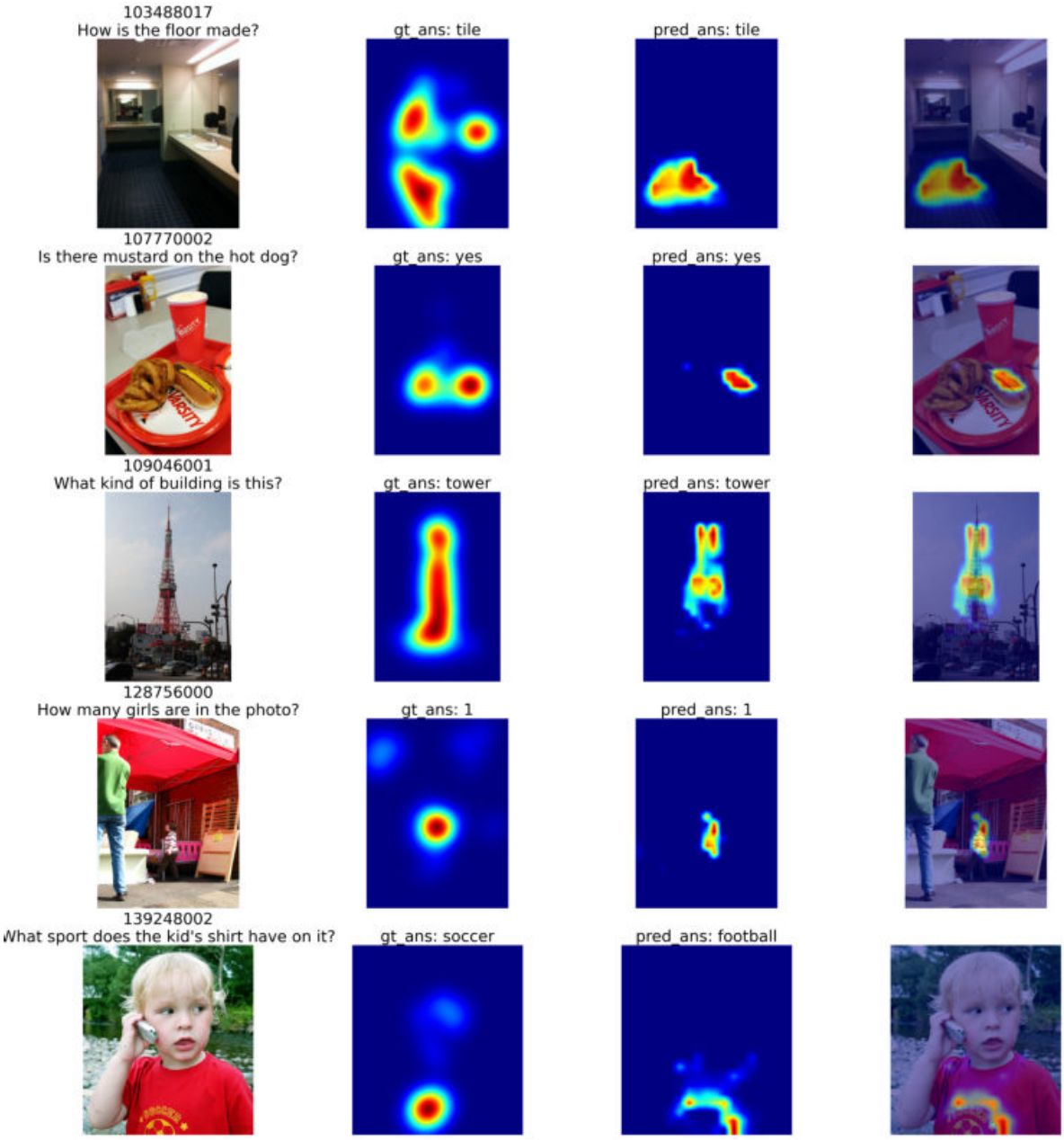

Figure A.9: More example of Grid attention maps on VQA-MHUG.

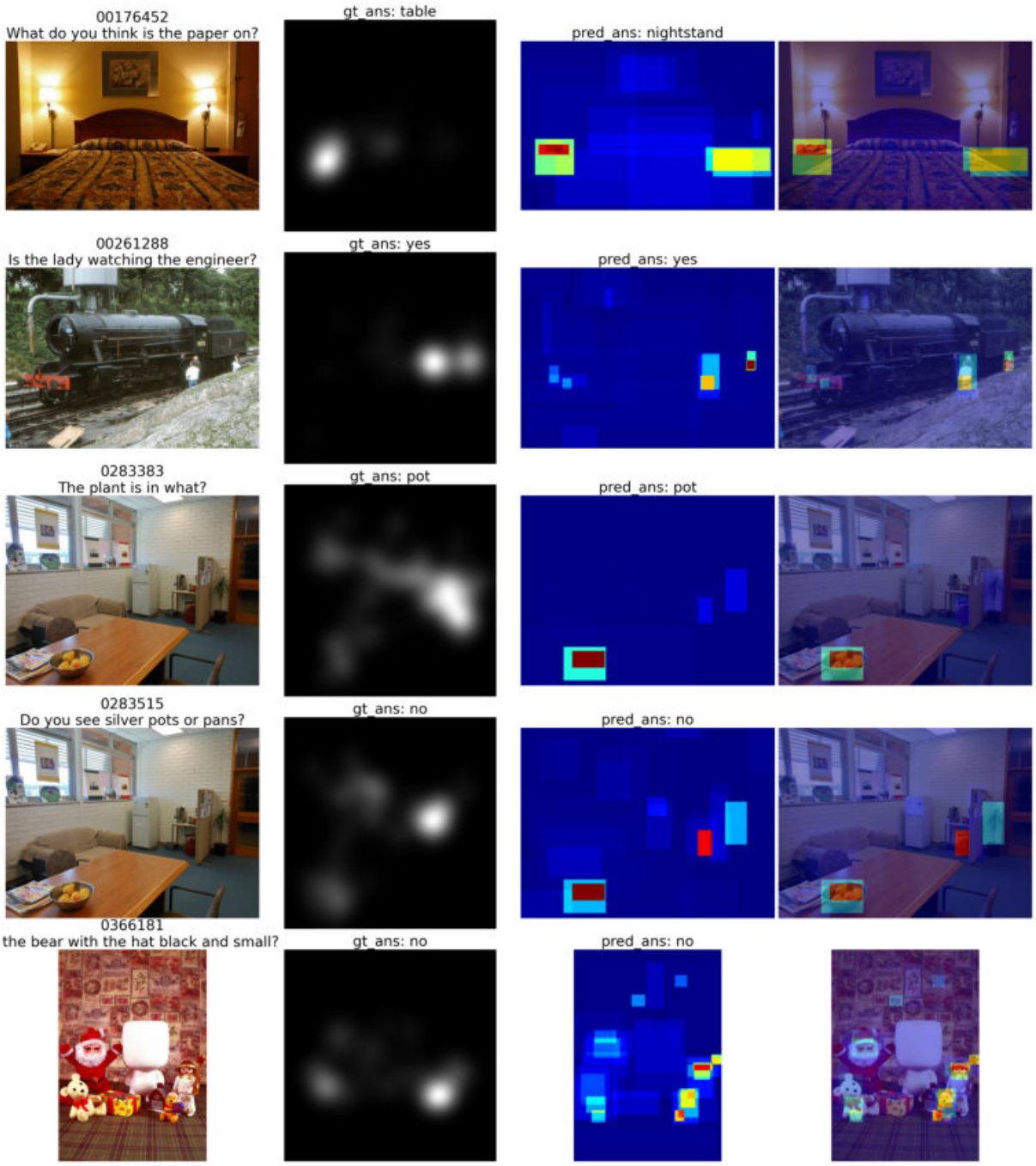

Figure A.10: More example of Region attention maps on AiR-D.

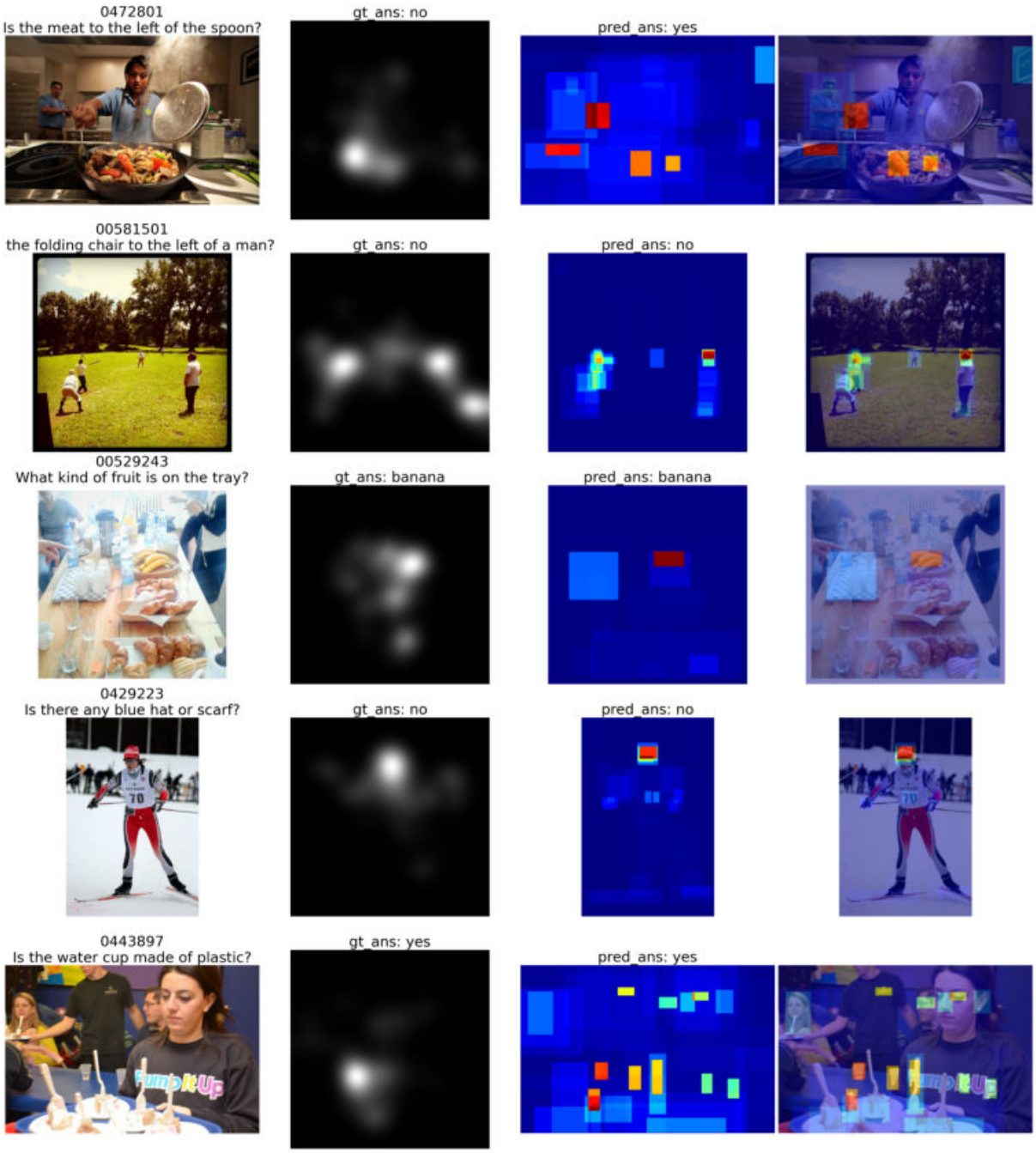

Figure A.11: More example of Region attention maps on AiR-D.

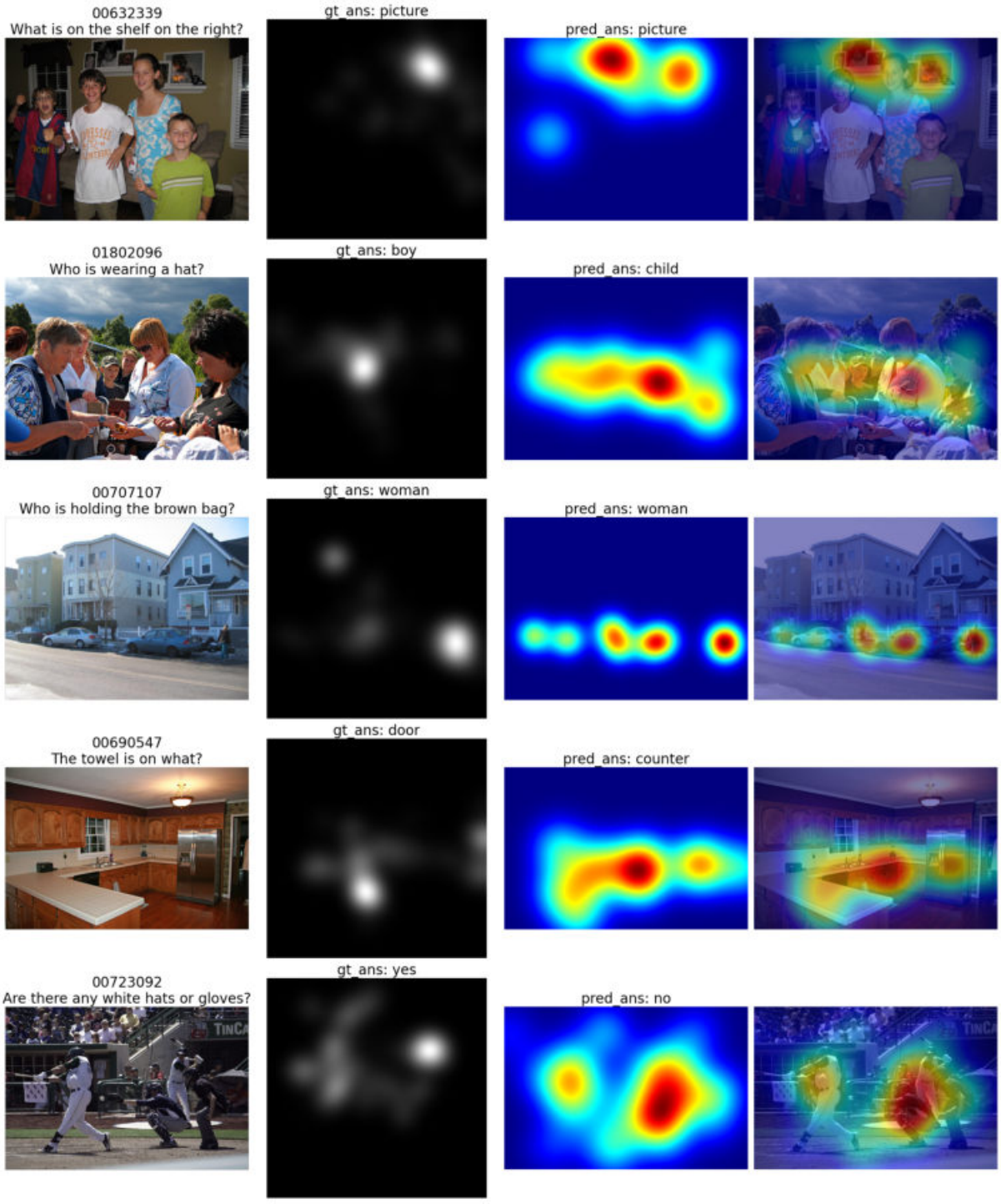

Figure A.12: More example of Gaussianized Region attention maps on AiR-D.

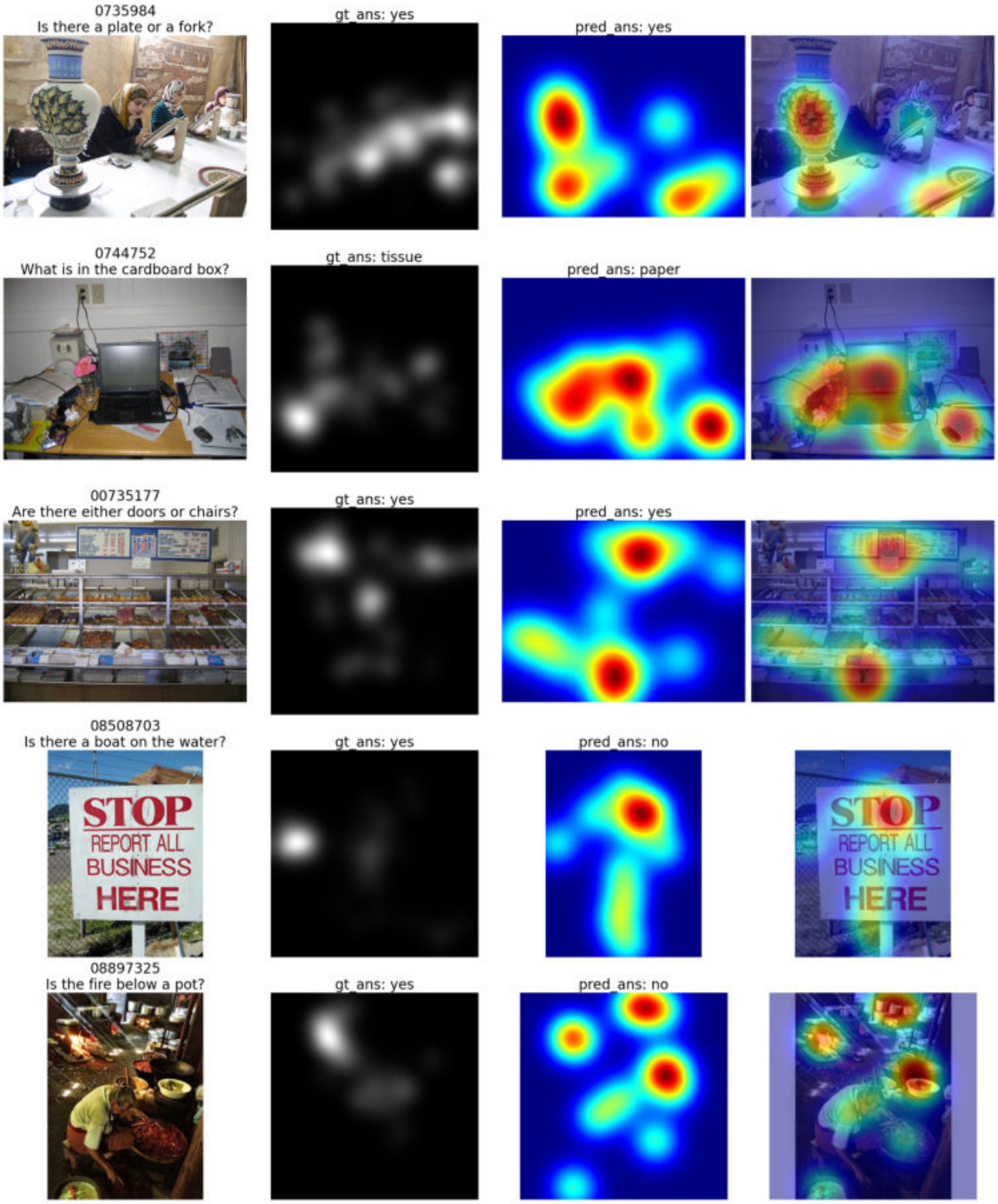

Figure A.13: More example of Gaussianized Region attention maps on AiR-D.

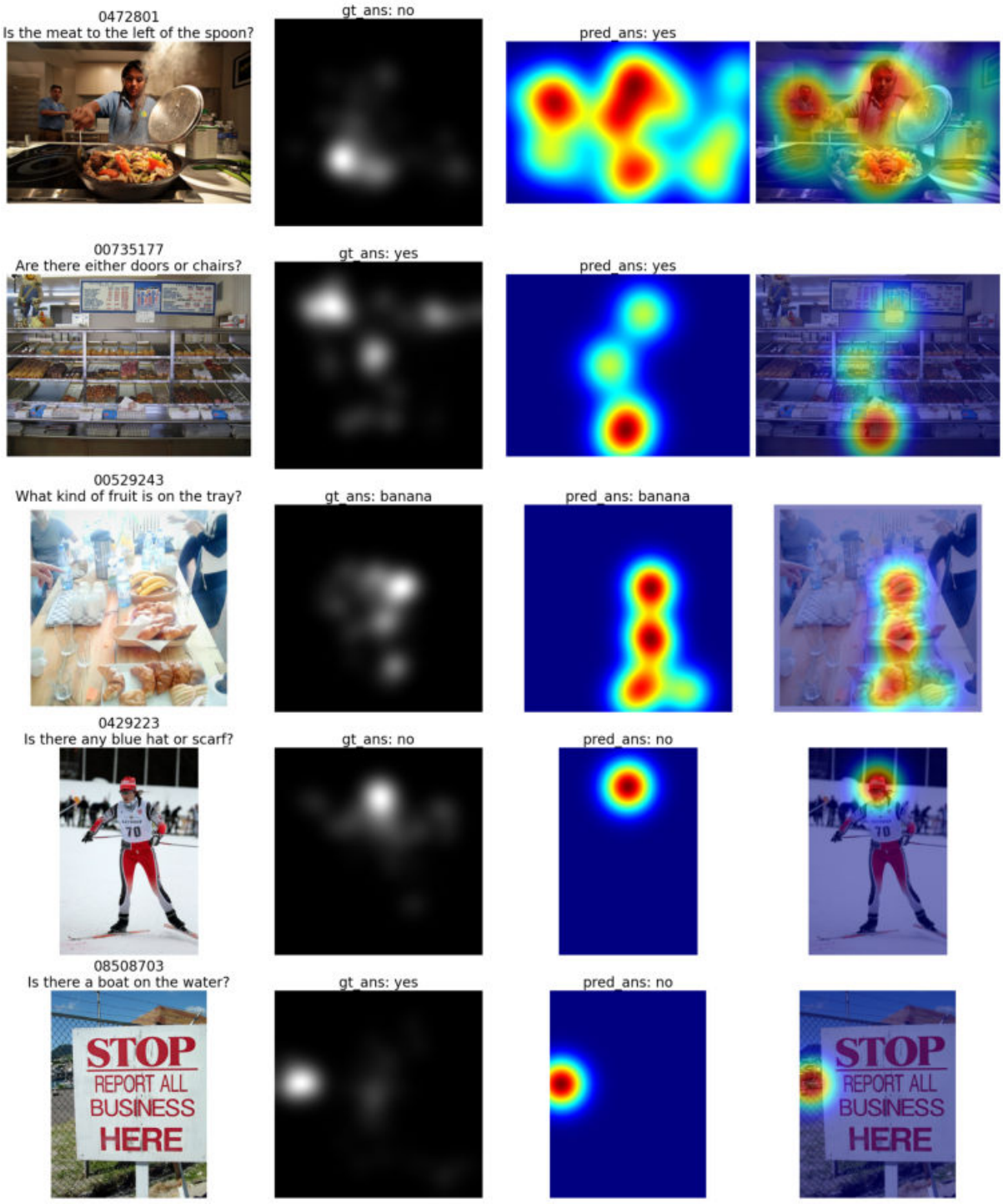

Figure A.14: More example of Grid attention maps on AiR-D.

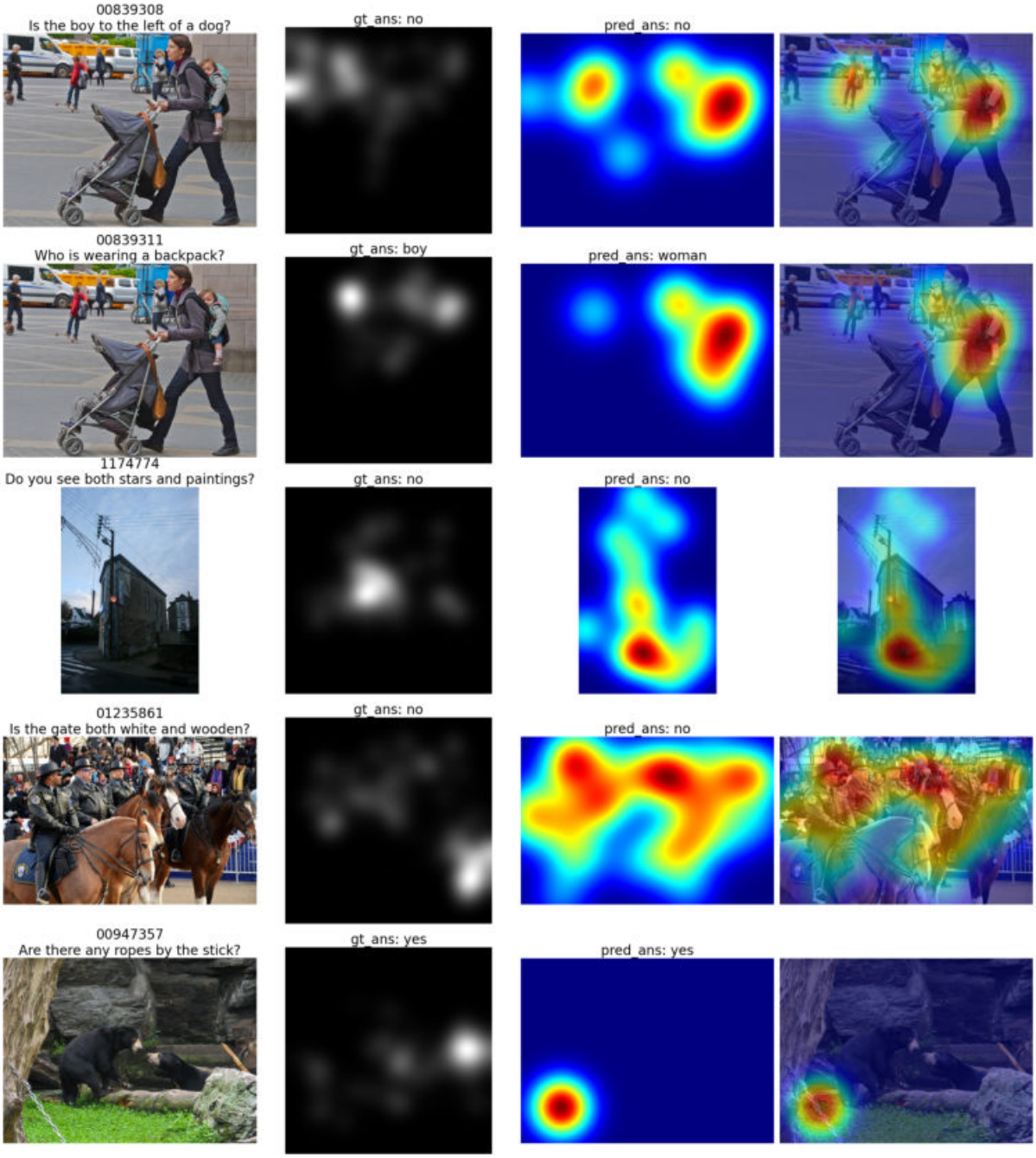

Figure A.15: More example of Grid attention maps on AiR-D.

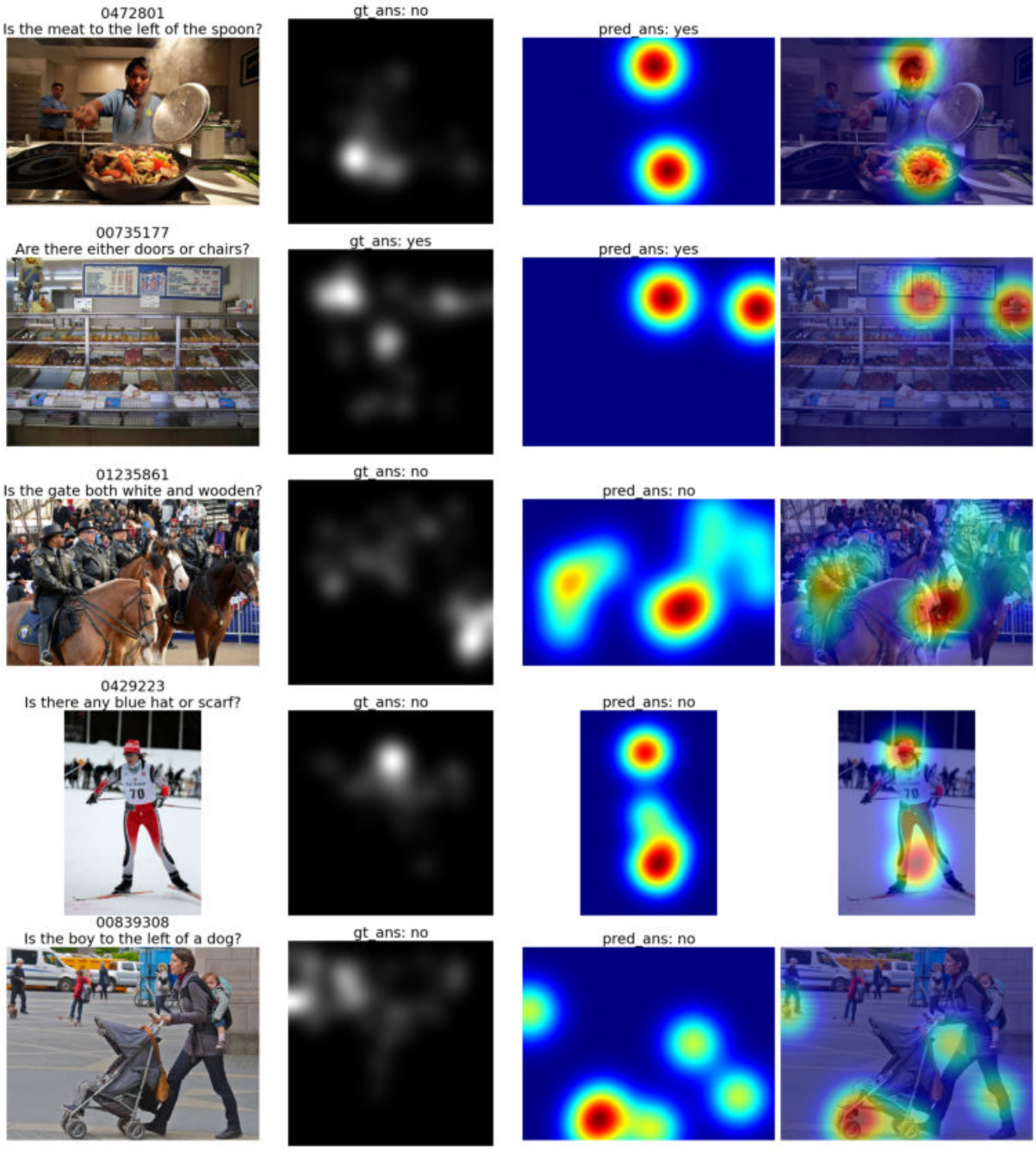

Figure A.16: More example of Region + Grid attention maps on AiR-D.

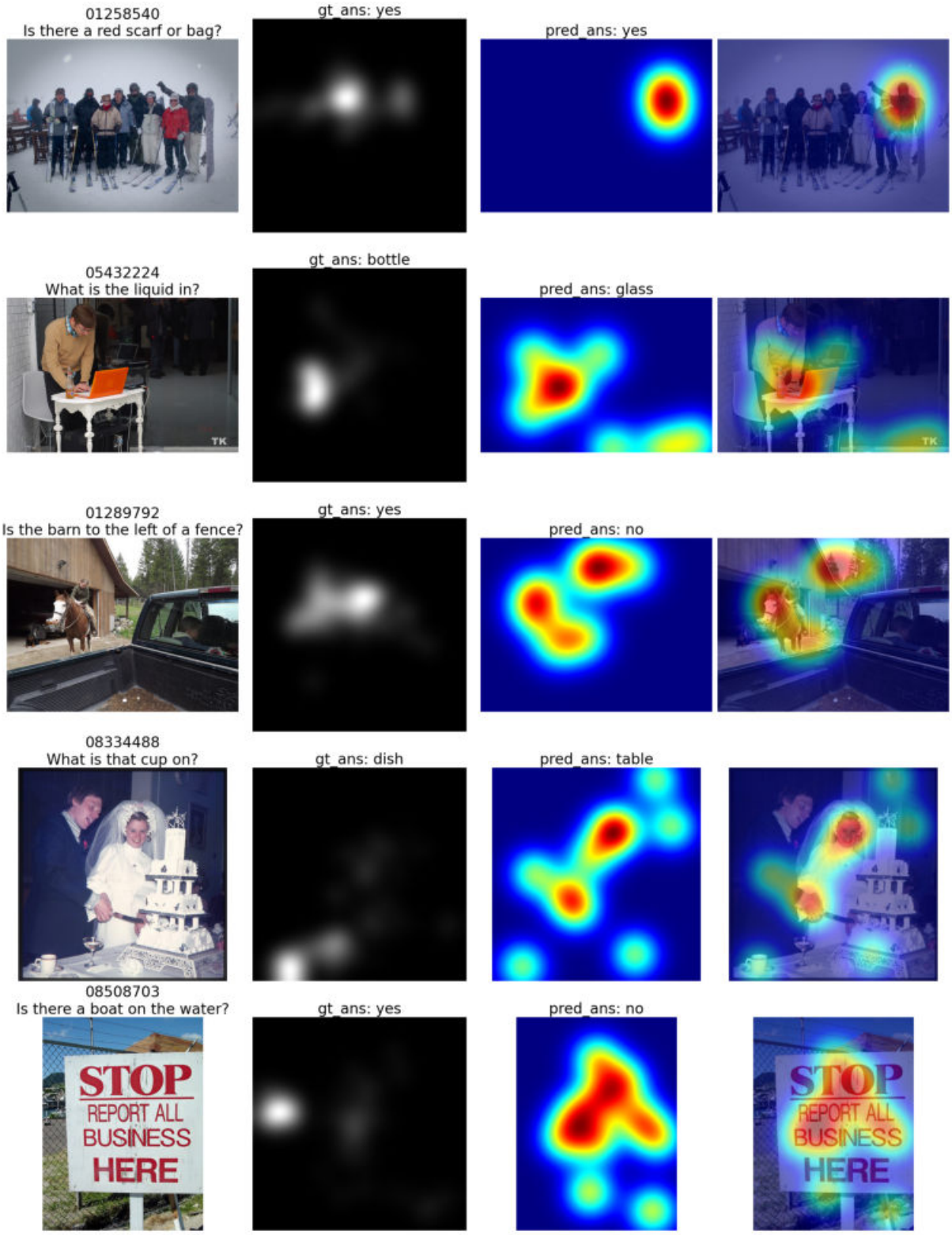

Figure A.17: More example of Region + Grid attention maps on AiR-D.