ChartGaze: Enhancing Chart Understanding in LVLMs with Eye-Tracking Guided Attention Refinement

Ali Salamatian[1] **Amirhossein Abaskohi**[1] **Wan-Cyuan Fan**[1,2]
Mir Rayat Imtiaz Hossain[1,2] **Leonid Sigal**[1,2,3] **Giuseppe Carenini**[1]

[1]University of British Columbia [2]Vector Institute for AI [3]CIFAR AI Chair
alisalam@student.ubc.ca
{aabaskoh, wancyuan, rayat137, lsigal, carenini}@cs.ubc.ca

Abstract

Charts are a crucial visual medium for communicating and representing information. While Large Vision-Language Models (LVLMs) have made progress on chart question answering (CQA), the task remains challenging, particularly when models attend to irrelevant regions of the chart. In this work, we present ChartGaze, a new eye-tracking dataset that captures human gaze patterns during chart reasoning tasks. Through a systematic comparison of human and model attention, we find that LVLMs often diverge from human gaze, leading to reduced interpretability and accuracy. To address this, we propose a gaze-guided attention refinement that aligns image-text attention with human fixations. Our approach improves both answer accuracy and attention alignment, yielding gains of up to 2.56 percentage points across multiple models. These results demonstrate the promise of incorporating human gaze to enhance both the reasoning quality and interpretability of chart-focused LVLMs[1].

1 Introduction

Charts are a common visual medium for communicating structured information and supporting analysis, comparison, and decision-making across domains. With the advancement of large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023; Chiang et al., 2023; Grattafiori et al., 2024; Anil et al., 2023; Liu et al., 2024a) and large vision-language models (LVLMs) (Liu et al., 2023; Team et al., 2023; Zhou et al., 2024; Chen et al., 2024c,b; Masry et al., 2025b), Chart Question Answering (CQA) has emerged as a key research challenge at the intersection of language, vision and data understanding (Masry et al., 2022). Early CQA approaches often converted charts into structured data or templates, but recent efforts have increasingly focused on leveraging LVLMs directly on chart images (Han et al., 2023; Masry et al., 2025b; Tian et al., 2025).

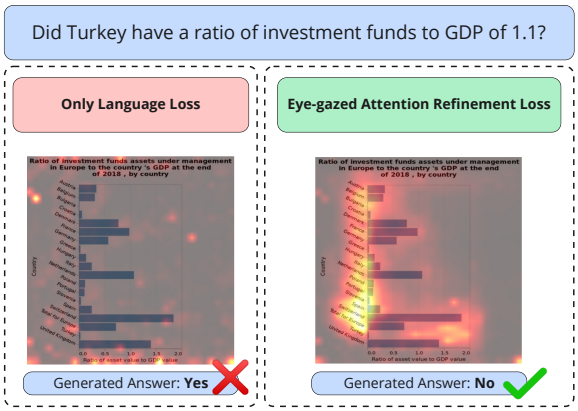

Figure 1: **Motivation.** Example of the model's attention maps. The left model, trained with language loss only, attends inconsistently and gives the wrong answer. The right model, trained with our proposed attention refinement loss, focuses more meaningfully on relevant regions and answers correctly.

Automated chart understanding can significantly aid evidence-based decision-making, by supporting fact-checking (Akhtar et al., 2023) or improving accessibility for visually impaired users through textual or spoken descriptions (Choi et al., 2019). As these models continue to advance and find increasing deployment, there is an increasing interest in understanding their internal mechanisms, especially their attention dynamics. Attention between visual and textual input modalities plays a vital role in how LVLMs form inductive biases and make an inference. Additionally, image-text attention serves as a window into the model's reasoning process, offering a potential avenue for interpretability and transparency. This is particularly important in high-stakes domains such as finance, medicine, and scientific research, where the ability to interpret model decisions is critical.

[1]Code and dataset are publicly available at ChartGazeCode and ChartGazeData, respectively.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12105–12125
November 4-9, 2025

However, recent studies have shown that image-text attention maps in LVLMs often fail to align with the regions humans focus on when answering questions, a phenomenon known as attention misalignment (Li et al., 2024; An et al., 2025; Han et al., 2025; Shu et al., 2025). This misalignment can cause models to fixate on irrelevant text or visual elements, leading to incorrect predictions and reduced interpretability. In chart understanding, this issue can be even more problematic, as key information, such as a specific bar, point, or legend entry, is often small or densely packed. As a result, the decision-making process of LVLM-based CQA systems becomes untrustworthy, limiting their use where interpretability is critical.

Hence, in this paper, we analyze the *attention misalignment* in the context of CQA. We observe that LVLMs, even those instruction tuned on charts, frequently attend to irrelevant chart elements, leading to incorrect responses and a reduced interpretability. We draw inspiration from human visual attention, as prior work (Gao et al., 2022; Yan et al., 2024) has shown that human gaze tends to align with perceived importance. We leverage eye-tracking data as explicit supervision to guide the attention maps of LVLMs. Specifically, we collect eye-tracking data from human participants responding to chart reasoning questions. Using the human gaze data, we train models for CQA to focus on regions where humans typically fixate, thereby improving both alignment and interpretability. As shown in Figure 1, models trained with our approach produce more interpretable and human-aligned attention maps, leading to more accurate answers. Empirical results show that aligning LVLM attention with human gaze improves CQA accuracy by up to 2.56 percentage points compared to fine-tuning with language loss alone.

To summarize, our key contributions are:

- **Eye-Tracking Chart Dataset:** We introduce a new eye-tracking dataset for CQA, capturing regions users look at while answering chart-related questions, serving as the ground-truth.
- **Analysis of LVLM Attention on Charts:** To the best of our knowledge, we are the first to conduct a systematic study of attention patterns of LVLMs on chart understanding and analyze how they compare with human gaze.
- **Gaze-Guided Attention Refinement:** We develop a training approach that aligns LVLM attention with human gaze, using a gaze-supervised loss.

2 Related Work

Chart Question Answering Datasets: Document understanding, particularly scientific chart understanding, has gained significant attention in the machine learning community. As a result, various datasets and benchmarks have been developed to accelerate progress and evaluate models in chart understanding, including summarization (Kantharaj et al., 2022b), question answering (Masry et al., 2022), explanation generation (Kantharaj et al., 2022a), and fact-checking (Akhtar et al., 2023). Among these, CQA has become a focal point, driven by the rapid advancements of LVLMs. Early benchmarks such as STL-CQA (Singh and Shekhar, 2020), LEAF-QA (Chaudhry et al., 2020), FigureQA (Kahou et al., 2018), and DVQA (Kafle et al., 2018) relied on synthetic charts or templated questions. Later efforts like PlotQA (Methani et al., 2020) and ChartQA (Masry et al., 2022) introduced charts from real-world sources, improving the diversity and realism of the data. Recent benchmarks pushed the evaluation into open domains (Masry et al., 2025a; Wang et al., 2024b; Liu et al., 2024b) and more complex, reasoning-intensive understanding tasks (Fan et al., 2024; Xia et al., 2024; Xu et al., 2023). Different from prior work, our benchmark uses eye-tracking annotations alongside chart question pairs to measure how well the LVLMs' attention aligns with that of humans. This human-centric design is crucial for bridging the gap between model performance and interpretability.

Gaze Datasets: Several studies have collected human gaze data to understand visual attention on charts. Borkin et al. (2016) collected a dataset of 393 visualizations along with participants' fixation locations during encoding and recognition of the visualizations. Polatsek et al. (2018) analyzed human visual attention during task-solving on 30 charts. More recently, Shin et al. (2022) gathered human attention on 10,960 chart images for the task of chart type recognition using webcam-based eye tracking. In the context of chart question-answering, Wang et al. (2024a) used BubbleView (Kim et al., 2017) to crowd-source mouse-click approximations of attention over 3,000 visualizations. While these datasets have contributed valuable insights, many either remain relatively small in size or rely on indirect, lower-fidelity methods such as mouse clicking

or webcam-based tracking. We chose to use high-precision eye-tracking equipment for our dataset, as prior work has demonstrated that eye tracking yields more accurate and consistent attention maps than mouse tracking. For example, Tavakoli et al. (2017) showed that visual congruency across participants is considerably higher with eye-tracking data, noting that even a large volume of mouse-tracking data from 90 participants could not match the performance of eye-tracking data from just 15 participants. Moreover, eye tracking captures immediate, cognitively grounded attention, whereas mouse-based methods introduce delays due to the slower nature of cursor movement. Prior work has shown that gaze data reliably reflects natural visual behavior and is widely used in saliency prediction, cognitive studies, and attention-aware interface designs (Jacob and Karn, 2003; Judd et al., 2012; Nielsen and Pernice, 2009; Majaranta and Bulling, 2014). Therefore, for a deeper analysis of LVLM and human attention and to fine-tune models for improved interpretability, we provide a high-quality, large-scale eye-tracking and CQA dataset (4,638 attention maps).

Visual Attention in LVLMs: In LVLMs, the attention mechanism between visual and textual tokens is critical not only for enabling cross-modal interaction but also for providing interpretability into how models integrate visual and textual information (Aflalo et al., 2022; Ben Melech Stan et al., 2024). Despite this, recent studies have shown that LVLMs often exhibit unintuitive visual attention patterns (Arif et al., 2025; Woo et al., 2025). Specifically, LVLMs tend to assign disproportionately high attention weights to specific tokens which are irrelevant to the text query (Zhang et al., 2025a; Kang et al., 2025). Moreover, prior works (Liu et al., 2024c; Chen et al., 2024a; Tong et al., 2024) have found that LVLMs under-utilize the visual inputs, leading to text-biased reasoning and weak visual grounding.

Several techniques have been proposed to address these issues. VAR (Kang et al., 2025) redistributes attention from irrelevant to relevant visual tokens. Visual contrastive decoding (Leng et al., 2024) encourages reliance on visual inputs by contrasting outputs with and without images. Other methods (Zhu et al., 2025; Zhang et al., 2024) explicitly boost attention to visual tokens to improve grounding. Unlike these approaches, we propose a novel approach that guides LVLM attention in training, using human gaze as an implicit supervisory signal, offering a cognitively grounded prior that enhances both reasoning and interpretability.

3 ChartGaze Dataset

To address attention misalignment in LVLMs (Shu et al., 2025), we introduce ChartGaze, a novel dataset that captures how humans visually process charts. By capturing high-precision human gaze patterns for chart-based questions, ChartGaze enables detailed comparisons between human and model attention. This opens up new insights into how human gaze can serve as an implicit training signal to enhance LVLM performance on chart understanding. We describe the dataset's curation process and statistics in the following subsections.

3.1 Dataset Construction

ChartGaze builds on chart images from the VisText and ChartQA datasets (Tang et al., 2023; Masry et al., 2022), which feature real-world charts sourced from platforms such as Statista and Pew Research. These charts span a broad range of topics, making them suitable for diverse question generation and visual reasoning. In what follows, we detail our two-stage dataset construction process: question-answer generation and gaze data collection.

3.1.1 QA Generation

While ChartQA already includes queries alongside each chart, for the VisText subset, we generated 3–5 question–answer pairs per chart caption using the pipeline illustrated in Figure 2. Because VisText summaries are human-authored, detailed, and semantically rich, they provide the necessary context for LLMs to generate questions that are meaningfully grounded in chart content. We used few-shot prompting with GPT-4o to generate questions regarding descriptive statistics, point-wise comparisons, and trend analysis. To ensure high-quality supervision, we instructed the model to:

- Use diverse phrasing in the questions,
- Balance True/False answers equally,
- Return output in a strict `JSON` format to facilitate automatic parsing and validation.

The complete prompt can be found in Appendix A.

For detailed error analysis and to better understand our dataset's composition, we categorized

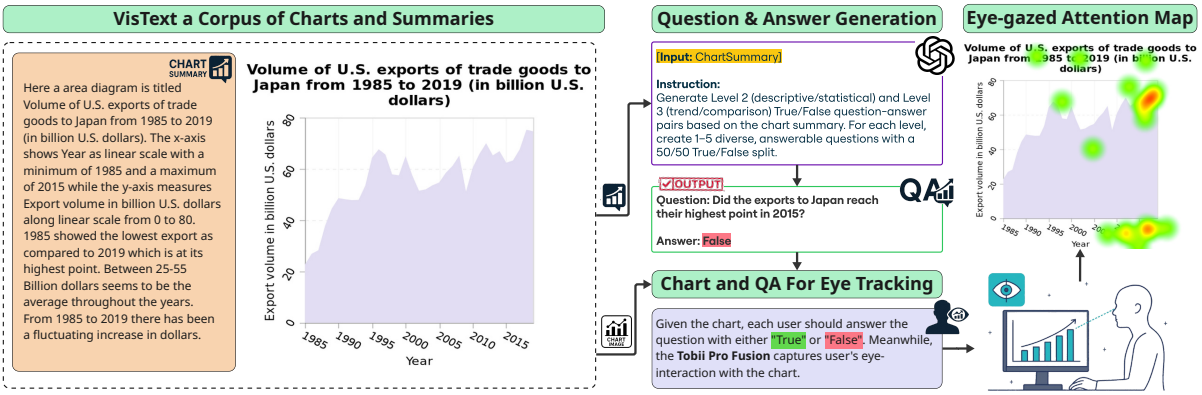

Figure 2: **Overview of the dataset creation pipeline** for generating chart-question-answer pairs with human gaze.

each question into one of six semantic types: Trend Analysis (TA), Finding Extremum (FE), Filtering (F), Comparison (CP), Retrieving Value (RV), and Computing Derived Value (CV). We defined each category with examples and incorporated them into the GPT-4o prompt (Appendix B).

3.1.2 Gaze Collection

We first detail the UI layout and then the gaze collection process.

Each chart was displayed at the bottom center of the screen, with the corresponding question in the top-left corner. This encouraged participants to read the question first and then shift their gaze to the chart, which helped reduce noise from repeated top-to-bottom transitions. A visualization of our UI is provided in Appendix C.

We captured gaze data using the Tobii Fusion Pro eye-tracker, which records gaze positions at the pixel level with microsecond temporal resolution. We then extracted fixation points and aggregated them into gaze maps via the following steps:

- Compute total fixation duration for each pixel.
- Apply a logarithmic non-linearity to smooth out sharp differences in fixation time.
- Apply Gaussian filter on the fixation map to simulate human visual receptive field.

3.2 Quality Control

We performed a two-step quality control process on our collected data to ensure reliability and accuracy. This included filtering both the generated questions-answers, and the collected gaze data.

3.2.1 QA Pair Quality Control

To ensure the correctness of our generated questions, we had human annotators flag ambiguous or incorrect questions during data collection. Of 4,811 questions, only 98 were flagged (a 2.0% error rate), which is lower than the 3.7% error rate found in the 593 selected ChartQA questions (22 errors). All flagged instances were removed.

We also verified the quality of the GPT-generated answers by measuring agreement with human annotators. A random sample of 100 examples showed a 93% agreement between GPT-4o and the majority vote of our seven human annotators. The highest pairwise agreement among human annotators was 92%. For this subset, Fleiss' Kappa was 0.7098, with an average pairwise agreement of 83.75%. These high agreement scores confirm the quality of both the GPT-4o answers and the human annotations.

3.2.2 Gaze Data Quality Control

We implemented a calibration and filtering process to ensure high-quality gaze data. For each participant, we calibrated the device and manually validated its accuracy, proceeding only when there was 0% data loss, resulting in an average accuracy of 0.42 degrees (16.8 px).

We also filtered the collected data to remove noise. Specifically, we removed invalid gaze samples and non-fixations caused by blinks or head movements. To ensure reliable gaze maps, we also discarded the bottom 3% of charts based on total viewing time.

3.3 Dataset Statistics

After filtering, our dataset consists of 4,638 attention maps derived from 1,620 unique chart images. These maps were collected from 476 yes/no QA pairs sourced from ChartQA and an additional 4,162 pairs generated from 1,144 VisText captions, averaging 3.6 QA pairs per caption. The dataset

was divided into a training set of 3,716 maps (80%) and a validation set of 922 maps (20%). The eye-tracking data was gathered from 32 student participants (11 volunteers and 21 paid contributors at $21/hour). The attention maps are distributed across chart types as follows: 2,470 bar charts, 1,100 line charts, 968 area charts, and 100 pie charts. Figure 6 shows the distribution of question categories. The prominence of TA and FE questions underscores the dataset's focus on level 2 and 3 semantic reasoning (Lundgard and Satyanarayan, 2021). Finally, the average question length was 12.2 words.

4 Gaze-Guided Refinement Method

LVLMs typically process and integrate information from both visual and textual inputs through multi-head self-attention layers. In this work, we investigate the attention pattern of these models during output generation and tune them to be more interpretable. As shown in Figure 3, the attention maps from LVLM are extracted and aligned with human gaze using a joint training objective that combines standard language modelling loss and gaze-guided attention alignment loss. We explain each component in the subsequent sections.

4.1 Extracting the Attention Maps

To obtain the image-text attention maps that capture how the model attends to image patches based on the text prompt, we first extract the maps from the first M layers. Our choice to focus on the first M layers is based on a qualitative analysis that aligns with the findings of Zhang et al. (2025b), which shows that earlier layers of LVLMs are crucial for this type of interaction and that information flow converges in these shallow layers. The relevant part of the attention matrices have a dimension of $\mathbb{R}^{M\times N_h\times T\times I}$, where M is the number of initial layers selected, N_h is the number of heads, T is the number of text tokens, and I is the number of image patches.

To simplify the representation while preserving the core attention structure, we followed the work of Jiang et al. (2025); Zhang et al. (2025b); Helbling et al. and averaged the attention maps across the text tokens, the N_h heads, and the first M layers. We obtained a single aggregated attention score for each image token, represented as $\mathbf{A}' \in \mathbb{R}^{1\times I}$. This map is then reshaped to the original image dimensions, creating a visual saliency map that is spatially aligned with the gaze data. For more information on our analysis and the specific values of M used for each model, see Appendix E.

4.2 Model Training

Following the attention extraction process, we train the model to align its visual attention with human gaze patterns while maintaining its language modeling capabilities. To this end, we jointly optimize two objectives: a language modeling loss and a gaze-guided attention alignment loss.

Let $x = \{x_1, \ldots, x_T\}$ denote the sequence of input tokens representing the question and answer. The standard language modeling loss is defined as:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{T}\sum_{t=1}^{T}\log P(x_t \mid x_{<t}) \tag{1}$$

To align model attention with human gaze, we used a weighted mean squared error (W-MSE) loss over the flattened attention maps (Bruckert et al., 2021). Let $A \in \mathbb{R}^{H\times W}$ denote the model's normalized attention map over the image, and $G \in \mathbb{R}^{H\times W}$ be the corresponding normalized gaze map. We flatten both maps into vectors of length $N = H \times W$: $A = [A_1, A_2, \ldots, A_N]$, $G = [G_1, G_2, \ldots, G_N]$. For each pixel i, we define the weight:

$$w_i = \frac{1}{\alpha - G_i}, \quad \text{with } \alpha = 1.1 \tag{2}$$

Here, α is a tunable parameter. We followed (Bruckert et al., 2021) in setting α to 1.1; this weighting emphasizes areas with higher human gaze values, which we want to prioritize. The W-MSE loss is then:

$$\mathcal{L}_{\text{W-MSE}} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot (G_i - A_i)^2 \tag{3}$$

Finally, we combine both objectives (i.e., equations (1) and (3)) into a single loss function:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{LM}} + \lambda_2\mathcal{L}_{\text{W-MSE}} \tag{4}$$

Here, λ_1 and λ_2 are tunable parameters that we set to 1. This joint training encourages the model to produce gaze-aligned visual attention while preserving its ability to generate accurate and fluent responses.

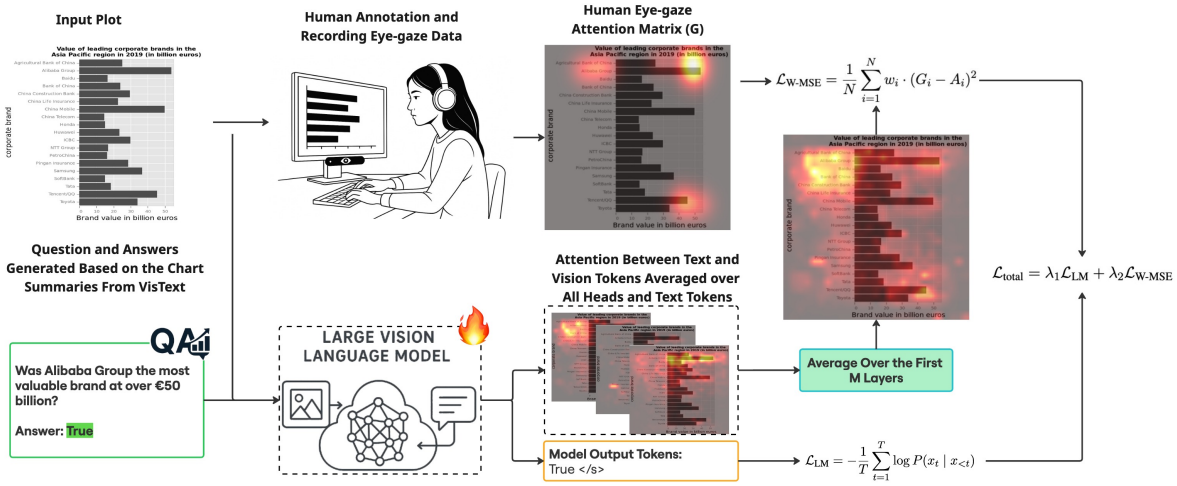

Figure 3: **Overview of our attention refinement training**, including attention map extraction and loss computation.

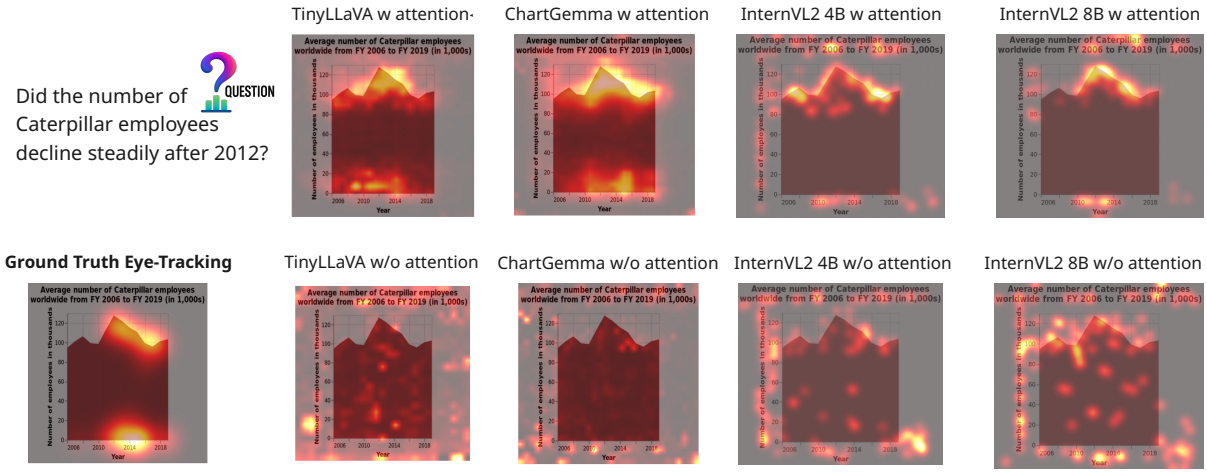

Figure 4: Comparison of attention maps from models trained with attention refinement loss vs. language loss only

5 Experiments and Results

LVLMs Evaluated: We evaluated the effectiveness of our approach on four models with diverse architectures, training strategies, and sizes: TinyLLaVA-450M, InternVL2-4B, InternVL2-8B, and ChartGemma-3B. TinyLLaVA-450M is the smallest model in the TinyLLaVA family, which has shown competitive results despite its size (Zhou et al., 2024). InternVL2 is a state-of-the-art vision-language foundation model with strong performance in visual question answering tasks (Chen et al., 2024c). We include both the 4B and 8B variants to assess the impact of scale. Finally, ChartGemma-3B is an instruction-tuned model built on PaliGemma and currently represents the state-of-the-art in CQA (Masry et al., 2025b).

Metrics: Since our dataset consists of yes/no questions, we use accuracy, defined as the percentage of questions correctly answered, to compare chart understanding performance across different models. To assess the similarity between the LVLM attention maps and the human attention maps, we used three standard metrics commonly used in saliency prediction (Meur and Baccino, 2012; Bruckert et al., 2021). Pearson's Correlation Coefficient (CC), Kullback–Leibler (KL) divergence, and histogram intersection (also known as similarity or SIM metric) between the ground-truth human gaze distribution and the model's attention distribution over the chart image.

5.1 Performance Improvement Across Models

We compare model performance under three conditions: zero-shot, fine-tuning with language loss only, and fine-tuning with our proposed attention-guided loss. As shown in Table 1, models fine-tuned with attention-guided loss consistently outperform those trained solely with language loss.

We set the temperature to zero to ensure deterministic zero-shot results. All other experiments were repeated three times, and we report the mean and standard deviation of each evaluation metric.

Notably, TinyLLaVA-450M, InternVL2-4B, and InternVL2-8B achieved statistically significant improvements of 1.19%, 1.54%, and 2.56% respectively. In contrast, ChartGemma-3B showed a marginal improvement of 0.18%, which was not statistically significant. We hypothesize that this is due to its extensive prior exposure to a corpus of 122,857 charts during instruction tuning (Masry et al., 2025b), which may have resulted in an effective, though less interpretable, attention structure.

In addition to accuracy gains, our approach improves the alignment between model attention and human gaze. As shown in Table 1, models fine-tuned with our method produce higher CC and SIM scores and lower KL divergence, indicating more human-aligned and interpretable attention. Furthermore, we observe a consistent positive correlation between interpretability metrics and QA performance, both within repeated runs of the same model and across different model architectures as shown in Figures 9 and 10 (in Appendix).

Figure 4 compares attention maps produced by different models trained with and without our proposed attention refinement loss. Models fine-tuned with language loss alone exhibit noisy attention, often failing to align with salient chart regions. In contrast, models trained with attention supervision display sharper, more human-like focus patterns. ChartGemma, in particular, closely aligns with human fixation maps, while TinyLLaVA similarly produces coherent attention, often emphasizing trend-relevant regions. InternVL2 variants also demonstrate focused activation on key visual elements such as chart peaks and axis labels. These qualitative results support the effectiveness of our approach in guiding models to attend to the most relevant regions of the chart.

5.2 Error Analysis

We analyze error rates by question categories and chart types for both TinyLLaVA and ChartGemma. As shown in Figure 5, TinyLLaVA struggles the most with computing derived value (CV) questions, which require multi-step reasoning, while it performs best on trend analysis (TA), likely due to the high number of TA examples in the training set.

Across chart types, TinyLLaVA performs relatively consistently, with the exception of pie charts, where it shows notably higher error. This may be attributed to limited exposure; only 77 pie chart examples were included in training.

ChartGemma outperforms TinyLLaVA in most categories, except for TA and comparison (CP) questions. This may be because its extensive prior exposure to over 120K charts during instruction tuning limited its responsiveness to the specific supervision in our setting.

5.3 Ablation Studies: Masked Inference

To evaluate the role of human-aligned attention in the model's reasoning process, we conducted controlled ablation experiments using two intervention strategies: masking and blurring human-attended regions. We used the fine-tuned InternVL2-8B model on our human gaze data in this experiment.

For the masking condition, we generated a binary mask from human gaze annotations, where pixels attended by humans were set to 0. For the blurring condition, we directly modified the input images by applying a Gaussian blur (kernel size = 15, $\sigma = 5$) to the regions with high human attention density, thus degrading the visual information in those critical areas. These experiments allow us to assess whether the model merely mimics human attention or leverages it for accurate reasoning.

As shown in Table 2, performance dropped significantly for the attention-tuned models, with a 7.08% decrease for blurring and an 8.00% decrease for masking. In contrast, the model trained only with a language loss saw a smaller drop of 4.34% and 5.22%, respectively. This suggests that the attention-tuned model relies more heavily on these semantically meaningful areas to generate its answers. Blurring the non-human-attended regions, on the other hand, resulted in a relatively minor performance decrease (1.90%), reinforcing that the model is not only aligned with human gaze visually but also actively uses that information during reasoning.

5.4 Ablation Studies: Loss Function

We investigated the role of different loss functions in aligning TinyLLaVA-450M attention maps with human gaze data. Specifically, we chose our losses from two broad categories of loss functions: pixel-based and distribution-based.

For pixel-based loss, we adopted W-MSE as defined in Section 4.2, which prioritizes regions with high fixation density. Among distribution-based losses, we evaluated KL Divergence (KLD), Focal

Training	Model	Test Acc.	CC ↑	KL ↓	SIM ↑
Zero-shot	TinyLLaVA-450M	46.64	-0.078	1.810	0.267
	InternVL2-4B	49.86	-0.060	1.722	0.282
	InternVL2-8B	50.93	-0.054	1.681	0.296
	ChartGemma-3B	52.39	0.100	1.559	0.323
Without Attn	TinyLLaVA-450M	62.58 ± 0.27	-0.048 ± 0.005	1.705 ± 0.031	0.288 ± 0.004
	InternVL2-4B	63.91 ± 0.20	-0.028 ± 0.004	1.532 ± 0.010	0.301 ± 0.004
	InternVL2-8B	65.36 ± 0.22	-0.017 ± 0.003	1.487 ± 0.009	0.312 ± 0.004
	ChartGemma-3B	72.49 ± 1.69	0.092 ± 0.004	1.594 ± 0.026	0.316 ± 0.003
With Attn	TinyLLaVA-450M	63.77 ± 0.54	0.391 ± 0.007	1.132 ± 0.015	0.439 ± 0.002
	InternVL2-4B	65.45 ± 0.23	0.402 ± 0.006	1.072 ± 0.008	0.451 ± 0.004
	InternVL2-8B	67.92 ± 0.15	0.417 ± 0.006	1.036 ± 0.007	**0.468 ± 0.005**
	ChartGemma-3B	**72.67 ± 1.24**	**0.436 ± 0.011**	**1.033 ± 0.014**	0.452 ± 0.005

Table 1: Performance of models trained with and without attention loss. ↑ / ↓ indicates higher / lower is better.

Condition	Acc. ↑	CC ↑	KL ↓	SIM ↑
Language loss only:	65.36	-0.017	1.487	0.312
— Blur human gaze areas	61.02	-0.139	1.681	0.236
— Mask human gaze areas	60.14	-0.124	1.713	0.221
— Blur non-gaze areas	64.10	0.112	1.392	0.298
— Mask non-gaze areas	62.85	0.064	1.456	0.274
Gaze supervision + language loss:	**67.92**	**0.417**	**1.036**	**0.468**
— Blur human gaze areas	60.84	-0.174	1.794	0.201
— Mask human gaze areas	59.92	-0.152	1.752	0.188
— Blur non-gaze areas	66.82	0.284	1.218	0.395
— Mask non-gaze areas	63.72	0.203	1.314	0.356

Table 2: Ablations showing the importance human-aligned attention towards the model's reasoning process. The experiments are grouped by training setup. Each header row reports the unperturbed model's performance; indented rows apply perturbations to probe model's reliance on attended vs. non-attended regions.

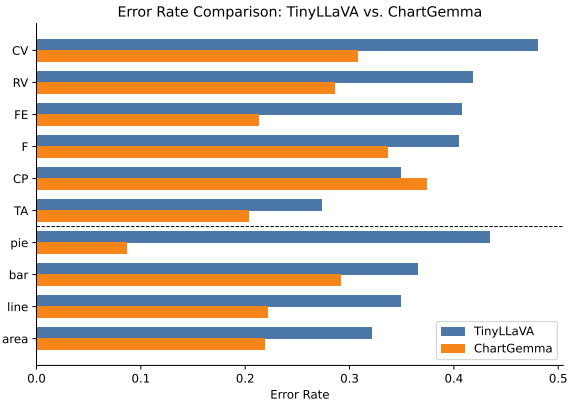

Figure 5: Comparison of error rates for TinyLLaVA and ChartGemma across question categories (top) and chart types (bottom), sorted by TinyLLaVA's error rates. A dashed line separates the two groups. ChartGemma consistently shows lower error rates, particularly on chart types and reasoning-heavy categories.

Loss (Lin et al., 2020), and a combined Dice + Binary Cross Entropy (BCE) loss. KLD encourages similarity between the extracted and ground-truth human attention distributions, while Focal Loss and Dice + BCE emphasize salient regions with higher ground-truth attention values. Detailed information is provided in Appendix D.

Figure 8 in Appendix presents qualitative comparisons of attention maps extracted from models trained with each loss function. As can be seen, W-MSE matches the ground truth most closely. KL and Focal loss have resulted in uniformly high attention on the axis and top of the bar and BCE + Dice had a too focused attention map that misses some critical points.

Table 4 reports test accuracy along with three attention evaluation metrics. W-MSE yields the best performance, achieving both the highest task accuracy and the most faithful attention alignment with human gaze. Notably, we observe a consistent trend across loss types: improvements in attention quality are correlated with gains in task accuracy.

5.5 Effect of Dataset Size

In domains where interpretability is essential, such as medical applications, training data is often limited. To evaluate the effectiveness of our attention supervision approach under low-data settings, we conducted experiments using randomly selected 25% and 50% of the ChartGaze dataset in addition to our previous result on the entire dataset. We trained InternVL2-8B with three different random

Training Setup	Test Acc.	CC ↑	KL ↓	SIM ↑
Without Attn Supervision				
MODEL-25%	60.21 ± 0.73	-0.045 ± 0.005	1.602 ± 0.012	0.274 ± 0.006
MODEL-50%	63.58 ± 0.24	-0.028 ± 0.004	1.530 ± 0.010	0.295 ± 0.005
MODEL-100%	65.36 ± 0.22	-0.017 ± 0.003	1.487 ± 0.009	0.312 ± 0.004
With Attn Supervision				
MODEL-25%	64.07 ± 0.26	0.297 ± 0.008	1.174 ± 0.011	0.402 ± 0.006
MODEL-50%	66.51 ± 0.20	0.396 ± 0.007	1.065 ± 0.009	0.454 ± 0.005
MODEL-100%	**67.92 ± 0.15**	**0.417 ± 0.006**	**1.036 ± 0.007**	**0.468 ± 0.005**

Table 3: Performance of models trained with and without attention loss across different dataset sizes.

Loss Function	Test Acc.	CC ↑	KL ↓	SIM ↑
W-MSE	**64.53**	**0.386**	**1.145**	**0.438**
KLD	62.36	0.306	1.209	0.380
Focal Loss	61.06	0.339	1.188	0.388
Dice + BCE	60.41	0.194	4.174	0.183

Table 4: Impact of loss functions on performance.

Fixation σ	Test Acc.	CC ↑	KL ↓	SIM ↑
20	62.89	0.277	1.875	0.272
40	**63.77**	0.391	1.132	0.439
80	61.49	**0.490**	**0.588**	**0.612**

Table 5: Performance of models trained with attention loss across different σ values.

seeds for each data setting to account for variability, which tends to be amplified in low-data regimes. This is reflected in the higher standard deviations reported in Table 3. Our method consistently outperforms models trained with standard language supervision, with performance gains becoming more pronounced as data availability decreases. Specifically, our approach achieves accuracy improvements of 3.86, 2.93, and 2.56 percentage points for the 25%, 50%, and 100% subsets, respectively. This trend highlights the value of attention supervision in low-resource settings.

Moreover, we again observe a strong linear correlation between attention quality and task accuracy, reinforcing the link between interpretable attention mechanisms and overall model performance.

5.6 Gaze Map Post-Processing Analysis

As part of our gaze map post-processing pipeline, we applied a Gaussian filter with a fixed standard deviation of $\sigma = 40$ pixels, corresponding to the spatial spread of visual attention around fixation points. To investigate the sensitivity of our model to this hyperparameter, we trained TinyLLaVA-450M with two different values of σ. Table 5 presents the results of this analysis. Our choice of sigma yields the best accuracy. Interestingly, $\sigma = 80$ scores really well on the attention quality metrics, but has a lower accuracy. This is because choosing such a high sigma results in high attention in a very large radius, therefore the model learns to have high attention in many regions, which is not beneficial for distinguishing the relevant parts of the image. Thus, high attention on a large part of the image as shown in Figure 11, results in a good metric value but in this case it is not same as learning which regions to attend to and hence not improving the accuracy.

6 Conclusion and Future Work

As chart understanding models are increasingly deployed in real-world applications, it is critical to ensure their interpretability. In this work, we introduced a novel attention refinement method and demonstrated its effectiveness on the newly collected ChartGaze dataset. Our results show that attention supervision not only improves alignment with human gaze (making the model attend to the interpretable parts of the chart) but also leads to performance gains in non-instruction-tuned models. While our method proved effective across multiple architectures, including TinyLLaVA and InternVL, it yielded only marginal gains on ChartGemma. This may be due to ChartGemma's prior instruction tuning on chart-related tasks, which could result in attention patterns that are either already well-formed or less responsive to additional supervision. Future work could explore strategies to better integrate attention refinement into instruction-tuned models, as well as extend this work to more diverse chart types and free-form question formats to better understand how attention varies with task complexity.

Limitations

This study focuses on Yes/No questions and relatively simple chart types (bar, line, and pie). These choices enabled large-scale data collection while allowing for accurate capture of participant attention during reasoning. However, these decisions may limit the generalizability of our findings to more complex visualizations and open-ended questions, where attention behavior may differ.

Ethical Considerations

This study was reviewed and approved by the University of British Columbia behavioural research ethics board. All participants provided informed consent prior to participation. Participants were recruited based on normal vision and hearing criteria and completed a chart interpretation task while eye-tracking data were collected. To protect privacy, all collected data were de-identified and securely stored on encrypted devices. Only anonymized data were used for analysis and model training. Participants were compensated at a rate of $7 per 20-minute session, up to a maximum of $21. They were informed of their right to withdraw at any time without penalty. Any public release of the dataset ensures participant anonymity and complies with open-access research guidelines. Moreover, our use of ChartQA and VisText is consistent with their intended purpose as open research datasets for chart understanding. For the ChartGaze dataset we created, we specify its intended use is for research on chart understanding, question-answering and interpretability. Model trained using our approach may still have over-reliance on superficial visual-textual correlations, leading to plausible-sounding but incorrect answers. We caution against deploying this system in high-stakes environments without robust safeguards and proper fact checking. We used generative AI tools, including ChatGPT, to support editing, formatting, and idea refinement during the research and writing process. All intellectual contributions, experimental designs, and analyses were developed and validated by the authors. No AI-generated content was included without human review and revision.

Acknowledgments

We gratefully acknowledge the support of Google for providing computing credits used in this work. This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs, NSERC Canada Research Chair (CRC), and NSERC Discovery. Resources used in preparing this research were provided, in part, by the Digital Research Alliance of Canada and by John R. Evans Leaders Fund CFI grant.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. 2025. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1773–1781.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8182–8187.

Michelle A. Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2016. Beyond memorability: Visualization recognition and

recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexandre Bruckert, Hamed R Tavakoli, Zhi Liu, Marc Christie, and Olivier Le Meur. 2021. Deep saliency models: The quest for the loss function. *Neurocomputing*, 453:693–704.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. 2019. Visualizing for the non-visual: Enabling the visually impaired to use visualization. In *Computer graphics forum*, volume 38, pages 249–260. Wiley Online Library.

Wan-Cyuan Fan, Yen-Chun Chen, Mengchen Liu, Lu Yuan, and Leonid Sigal. 2024. On pre-training of multimodal language models customized for chart understanding. *Adaptive Foundation Models Workshop at Advances in Neural Information Processing Systems*.

Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *Preprint*, arXiv:2311.16483.

Zhongyi Han, Gongxu Luo, Hao Sun, Yaqian Li, Bo Han, Mingming Gong, Kun Zhang, and Tongliang Liu. 2025. Alignclip: navigating the misalignments for robust vision-language generalization. *Machine Learning*, 114(3):1–19.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In *Forty-second International Conference on Machine Learning*.

Robert Jacob and Keith Karn. 2003. *Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises*, volume 2, pages 573–605.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014.

Tilke Judd, Frédo Durand, and Antonio Torralba. 2012. A benchmark of computational models of saliency to predict human fixations.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. In *ICLR (Workshop)*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022a. OpenCQA: Open-ended question answering with charts. In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, pages 11817–11837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022b. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Nam Wook Kim, Zoya Bylinskii, Michelle A. Borkin, Krzysztof Z. Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. Bubbleview: An interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction*, 24(5):1–40.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(02):318–327.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024b. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1287–1310.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1073–1083.

Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human–Computer Interaction*, pages 39–65. Springer London, London.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025a. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria. Association for Computational Linguistics.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2025b. ChartGemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643, Abu Dhabi, UAE. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Olivier Le Meur and Thierry Baccino. 2012. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45:251 – 266.

Jakob Nielsen and Kara Pernice. 2009. *Eyetracking Web Usability*, 1st edition. New Riders Publishing, USA.

Patrik Polatsek, Manuela Waldner, Ivan Viola, Peter Kapec, and Wanda Benesova. 2018. Exploring visual attention and saliency modeling for task-based visual analysis. *Computers & Graphics*, 72:26–38.

Sungbok Shin, Sunghyo Chung, Sanghyun Hong, and Niklas Elmqvist. 2022. A scanner deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data. *IEEE Transactions on Visualization and Computer Graphics*, 29:396–406.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *Preprint*, arXiv:2501.01346.

Hrituraj Singh and Sumit Shekhar. 2020. Stl-cqa: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Hamed R Tavakoli, Fawad Ahmed, Ali Borji, and Jorma Laaksonen. 2017. Saliency revisited: Analysis of mouse movements versus fixations. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 1774–1782.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. 2025. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 31(3):1731–1745.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1927–1951, Vienna, Austria. Association for Computational Linguistics.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *ArXiv*, abs/2402.12185.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. 2024. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in Neural Information Processing Systems*, 37:1890–1918.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2025b. From redundancy to relevance: Enhancing explainability in multimodal large language models. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. 2024. Prompt highlighter: Interactive control for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13215–13224.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *Preprint*, arXiv:2402.14289.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633.

A Prompt for Question-Answer Generation

Question and Answer Generation Prompt

Generate a set of question-answer pairs based on the following summary of the chart in JSON format. Focus on the following semantic levels of questions:

- **Level 2 (L2):** Descriptive statistics, extrema, outliers, and correlations (binary: True/False)
- **Level 3 (L3):** Point-wise comparisons, complex trends, pattern synthesis (binary: True/False)

For **Level 2** and **Level 3**, make decisions on how to generate the questions to produce specific True/False answers. Ensure that the generated questions:

- Have a **50/50 chance** of being True or False.
- For False answers, use values that are **domain-appropriate and plausible** based on the chart data.

Use diverse language and ensure that the questions are relevant to the key points in the text and that the answers are accurate and concise. Only include questions that have a clear answer in the text. Aim to generate 1–5 questions for each level.
IMPORTANT: Return your response as a valid JSON array where each question follows this exact format:

```
[
  {
    "question": "...?",
    "answer": true/false,
    "level": 2/3
  },
  ...more questions...
]
```

VERY IMPORTANT: DO NOT use markdown code blocks (“‘json or “‘). Return ONLY the raw JSON.
There are five chart summary, QA pairs below as examples. Note that level 1 is given for the context of the chart and we don’t want to create question answer pairs for this level.
—
First example:
Summary:
L1: This line diagram is titled Canadian imports of bauxite from 2005 to 2019 (in 1,000 metric tons). The x-axis measures Year on linear scale from 2006 to 2018 while the y-axis shows Imports in thousand metric tons on linear scale of range 0 to 4,000.
L2L3: The year with the lowest import of bauxite was 2009. Although the results fluctuate from year to year, the graph tends to show a general increase in bauxite imports over time, with the exception of 2009, where there was a large decrease in imports.
QAs (in JSON format):

```
[
  {
    "question": "Was the year with the lowest import of bauxite 2009?",
    "answer": true,
    "level": 2
  },
  {
```

```
    "question": "Was the year with the highest import of bauxite 2008?",
    "answer": false,
    "level": 2
  },
  {
   "question": "Did the Canadian imports of bauxite tend to show a general increase
    over time?",
    "answer": true,
    "level": 3
  },
  {
    "question": "Was there a large decrease in imports in 2015?",
    "answer": false,
    "level": 3
  }
]
```

—

Second example:

Summary:

L1: Poverty rate for families in the United States from 1990 to 2019 is a line chart. The y-axis plots Poverty rate while the x-axis plots Year.

L2L3: Poverty in the USA was at its highest in the early nineties, dropping in 2000 to its second lowest point on the chart. Then poverty rose gradually again, almost hitting the same peak in 2010 and staying level for a few years before dropping sharply in the mid 2000's to well below the levels of the 90s.

QAs (in JSON format):

```
[
  {
    "question": "Was poverty at its highest in the early nineties?",
    "answer": true,
    "level": 2
  },
  {
    "question": "Was the second lowest poverty rate recorded in 2005?",
    "answer": false,
    "level": 2
  },
  {
    "question": "Did the poverty rate rise from 2000 to 2010?",
    "answer": true,
    "level": 3
  },
  {
    "question": "Was there a rise in poverty rates right after 2010?",
    "answer": false,
    "level": 3
  },
  {
   "question": "Was the poverty level in 90s way higher than that of mid 2000s?",
    "answer": true,
```

```
    "level": 3
  },
  {
    "question": "Did the poverty in the USA fluctuate over the years but had an
    overall upward trend?",
    "answer": false,
    "level": 3
  }
]
```

—

Third example:

Summary:

L1: This bar diagram is labeled Top 10 U.S. states based on production value of principal fresh and processing market vegetables in 2019 (in 1,000 U.S. dollars). There is a categorical scale with Arizona* on one end and Washington at the other on the x-axis, marked State. Production value in thousand U.S. dollars is plotted on the y-axis.

L2L3: California has the most production value of principal fresh and processing market vegetables at nearly 8,000,000 US dollars whereas New Jersey has the least production value of principal fresh and processing market vegetables at only approximately 100,000 US dollars.

QAs (in JSON format):

```
[
  {
    "question": "Was California the state with the highest production value?",
    "answer": true,
    "level": 2
  },
  {
    "question": "Was the production value of New Jersey 500,000 US dollars?",
    "answer": false,
    "level": 2
  },
  {
    "question": "Did New Jersey have the second-highest production value?",
    "answer": false,
    "level": 3
  },
  {
   "question": "Was the difference between the highest and lowest production values
    around 7,900,000 US dollars?",
    "answer": true,
    "level": 3
  }
]
```

—

Fourth example:

Summary:

L1: This bar diagram is titled Egypt: National debt from 2015 to 2025 in relation to gross domestic product (GDP). The y-axis shows Year on a categorical scale with 2015 on one end and 2025* at the other. Along the x-axis, National debt in relation to GDP is measured with a linear scale of range 0.0 to 1.0.

L2L3: The National debt in Egypt between 2015 to 2025 has been pretty consistent with a slight rise in 2017.
QAs (in JSON format):


```
[
  {
    "question": "Was there a slight rise in national debt in 2017?",
    "answer": true,
    "level": 2
  },
  {
    "question": "Was national debt at its highest in 2017?",
    "answer": false,
    "level": 2
  },
  {
    "question": "Has the national debt in Egypt been consistent overall?",
    "answer": true,
    "level": 3
  },
  {
    "question": "Was there a sharp drop in national debt in 2018?",
    "answer": false,
    "level": 3
  }
]
```


—

Fifth example:
Summary:
L1: This is an area plot called Number of passengers arriving and departing at airport terminals in the United Kingdom (UK) from 1992 to 2019 (in millions). On the x-axis, Year is measured. Passengers in millions is plotted on the y-axis.
L2L3: The number of passengers at UK airports has risen steadily since 1992 to 2019 with the exception of a period of a few years in 2007–2010 where numbers fell. Overall, numbers have nearly tripled from just over 100 million passengers in 1992 to almost 300 million in 2019. The rate of increase has been relatively steady with the exception of the 2007–2010 period.
QAs (in JSON format):


```
[
  {
    "question": "Was the number of passengers in 1992 just over 100 million?",
    "answer": true,
    "level": 2
  },
  {
    "question": "Did passenger numbers fall sharply from 2015 to 2019?",
    "answer": false,
    "level": 2
  },
  {
    "question": "Did the number of passengers double from 1992 to 2019?",
    "answer": false,
```

```
    "level": 3
  },
  {
    "question": "Was the period of 2007-2010 an outlier with falling numbers?",
    "answer": true,
    "level": 3
  }
]
```

—

Now generate question-answer pairs from the following summary in the specified JSON format:
{caption}

B Prompt for Question Category Annotation

Question Category Prompt

Classify the following questions into one of the six categories based on its intent:

- CP (Comparison): The question compares two or more data points (e.g., "Which country has more X than Y?").
- CV (Computing Derived Value): The question involves computing a value from others (e.g., difference, average, ratio). - RV (Retrieving Value): The question asks for the value(s) of specific data points or attributes.
- FE (Finding Extremum): The question asks for the maximum or minimum value in the data.
- F (Filtering): The question asks for data points that satisfy multiple specified conditions.
- TA (Trend Analysis): The question asks about changes over time, patterns, increases, decreases, or stability.

Output format: Return only the category abbreviation without additional text in a list.
Questions: {questions}

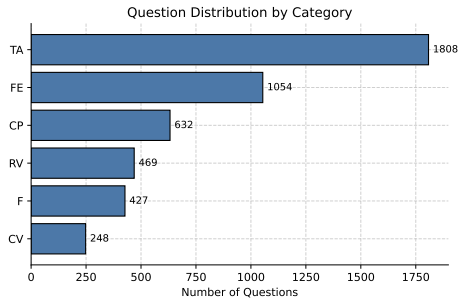

Figure 6: Distribution of questions across categories used in the prompt. TA and FE are the most common, reflecting the reasoning-heavy nature of the dataset.

C User Interface Setup and Participant Demographics

Figure 7 shows the interface used for our chart-based Yes/No question answering task. All participants were students who reported interacting with charts on a daily to weekly basis. The cohort was gender-balanced and represented a range of academic backgrounds, including engineering, computer science, and statistics.

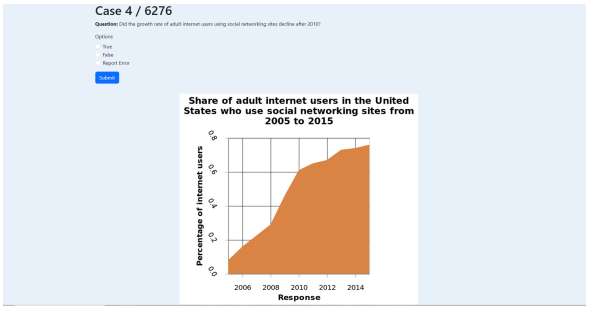

Figure 7: Example user interface setup used for collecting gaze data.

D Attention Loss Function Details

Focal Loss. To address the foreground-background imbalance in gaze prediction, we adopt the Focal Loss introduced by Lin et al. (2020). This loss down-weights well-classified examples and focuses training on hard examples. It is defined as:

$$\mathcal{L}_{\text{FL}}(P,Q) = -\sum_{i=1}^{H\times W}\Big[P_i(1-Q_i)^{\gamma}\log Q_i + (1-P_i)Q_i^{\gamma}\log(1-Q_i)\Big] \quad (5)$$

where i indexes $H \times W$ pixels, P_i is the ground truth, Q_i is the predicted value, and γ is a focusing parameter (set at 2 in our experiments).

Layers used	Accuracy ↑
All layers	57.27%
First 10 layers	**63.77%**
Last 10 layers	60.20%

Table 6: Effect of layer selection in TinyLLaVA. Consistent with (Zhang et al., 2025b), focusing on early layers yields the highest accuracy, indicating that information critical for attention map extraction is concentrated in the initial stages.

Dice + BCE Loss. To improve overlap between predicted and true gaze maps, we combine Dice Loss with Binary Cross-Entropy (BCE). The total loss is:

$$\mathcal{L}_{\text{BCE+Dice}}(P,Q) = \lambda_{\text{Dice}}\cdot\mathcal{L}_{\text{Dice}}(P,Q) + \lambda_{\text{BCE}}\cdot\mathcal{L}_{\text{BCE}}(P,Q) \quad (6)$$

where λ_{Dice} and λ_{BCE} are scalar weights. We use $\lambda_{\text{Dice}} = 100$ and $\lambda_{\text{BCE}} = 1.0$. The Dice loss is defined as:

$$\mathcal{L}_{\text{Dice}}(P,Q) = 1 - \frac{2\sum_i P_iQ_i + \varepsilon}{\sum_i P_i + \sum_i Q_i + \varepsilon} \quad (7)$$

where $\varepsilon = 10^{-8}$ is added for numerical stability. The BCE loss is given by:

$$\mathcal{L}_{\text{BCE}}(P,Q) = -\sum_{i=1}^{H\times W}\Big[P_i\log Q_i + (1-P_i)\log(1-Q_i)\Big] \quad (8)$$

This composite loss ensures both accurate per-pixel predictions and global shape alignment.

Note that to ensure comparability across losses, we apply a scaling coefficient to each attention loss term so that their magnitudes are roughly aligned; this coefficient can be treated as a tunable hyperparameter.

E Implementation details.

E.1 Attention Map Extraction Layer Justification

We extract attention maps capturing the alignment between visual and textual modalities from three representative LVLM, TinyLLaVA (Zhou et al., 2024), InternVL2 (Chen et al., 2024c), and Chart-Gemma (Masry et al., 2025b), which serve as the basis for gaze-guided attention refinement.

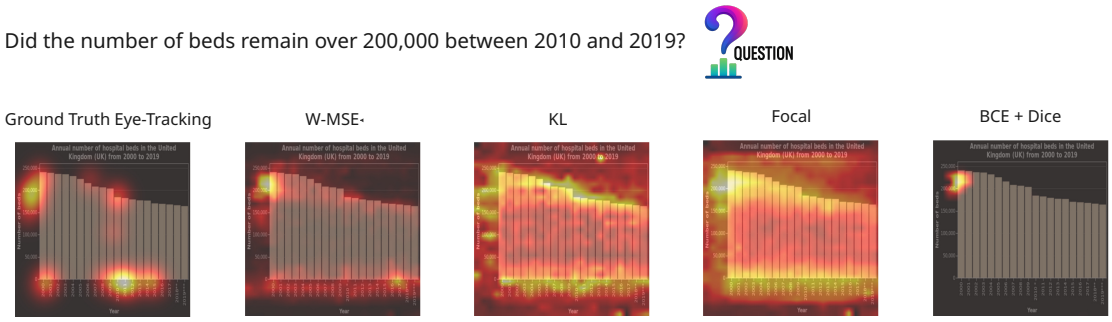

Figure 8: Comparison of attention maps from models trained with different attention refinement loss functions.

Following the work of (Zhang et al., 2025b), which identified that the initial layers of LVLMs like LLaVA play a key role in cross-modal interaction, we conducted an analysis on TinyLLaVA to determine which layers were most effective for our task. As seen in Table 6, our findings showed that focusing on the first 10 layers resulted in an accuracy of 63.77%, a significant improvement over using all layers (57.27%) or only the last 10 layers (60.195%). This demonstrates that the information crucial for attention map extraction is primarily concentrated in the model's earlier layers.

We adopt a similar strategy for InternVL2. Due to its deeper architecture and higher-resolution visual encoding, we extracted attention from twelve layers (instead of ten) for both the 4B and 8B versions. For ChartGemma, we aggregated attention maps from the first six layers, as it is shallower than both TinyLLaVA and InternVL2.

E.2 Training Details

We finetuned TinyLLaVa using Low-Rank Adaptation (LoRA) on 15 epochs with learning rate of 1×10^{-4} with batch size 4, gradient accumulation steps 8, LoRA rank of 32 and LoRA α of 64. Moreover, we used RTX 3090 GPU with 24G VRAM. The fine-tuning was fast and took approximately 4 hours only. For ChartGemma, since the model has already instruction tuned on charts, we fine-tuned using LoRA on 7 epochs with learning rate of 5×10^{-5}, and batch size of 1 due to memory constraints with the same LoRA rank and α and with early stopping. For ChartGemma we used A100 with 40G VRAM. For InternVL2, we performed LoRA fine-tuning using 2× A100 GPUs with 80GB VRAM each. We used a learning rate of 5×10^{-5}, batch size 2, gradient accumulation steps 16, LoRA rank of 32, and LoRA α of 64.

F Correlation Between Attention Interpretability and Accuracy

As can be seen in Figures 9 and 10, there is a positive correlation between attention interpretability and accuracy.

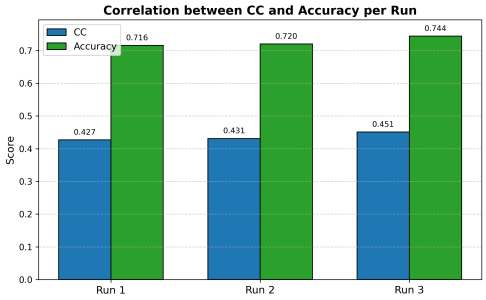

Figure 9: Correlation between CC (Correlation Coefficient) and accuracy across three independent runs of ChartGemma. Higher CC is consistently associated with higher QA accuracy, suggesting that more interpretable attention benefits task performance.

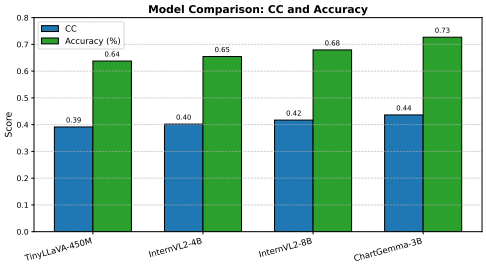

Figure 10: Model-level comparison of CC and QA accuracy across four models. Models with higher CC (e.g., InternVL2 and TinyLLaVA) tend to achieve higher QA accuracy, reinforcing the link between attention alignment and task effectiveness.

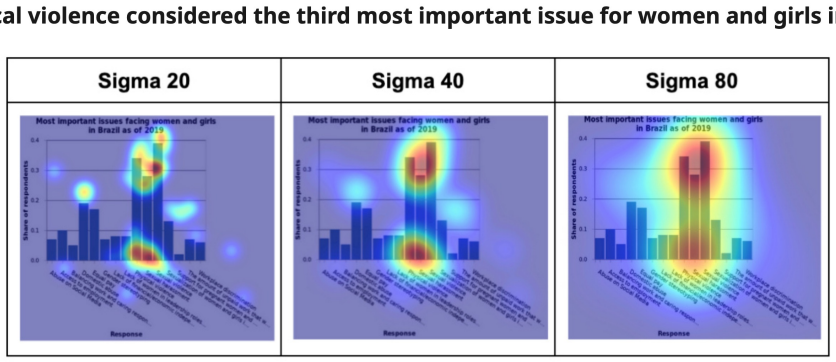

Figure 11: Comparison of different σ values and its effect on the human gaze map.

Model	Relaxed Accuracy
No Fine-tuning	
InternVL2-4B (Pre-trained)	81.52
InternVL2-8B (Pre-trained)	83.28
Fine-tuned on ChartGaze, lang loss only	
InternVL2-4B	49.15
InternVL2-8B	50.28
Fine-tuned on ChartGaze, lang+attention loss	
InternVL2-4B	51.42
InternVL2-8B	53.08

Table 7: Model performance on the ChartQA test set after fine-tuning with different strategies on our dataset.

Model	Relaxed Accuracy
No Fine-tuning	
InternVL2-4B (Pre-trained)	81.52
InternVL2-8B (Pre-trained)	83.28
Fine-tuned on ChartGaze, lang loss only	
InternVL2-4B	76.40
InternVL2-8B	77.25
Fine-tuned on ChartGaze, lang+attention loss	
InternVL2-4B	78.64
InternVL2-8B	79.85

Table 8: Model performance on the ChartQA test set after fine-tuning with different strategies on our modified dataset.

G Generalization

We conducted preliminary experiments on 150 open-ended ChartQA examples and observed an improvement of about 1% in relaxed accuracy. This indicates that the method may generalize beyond Yes/No tasks, though further study is required to assess cost-effectiveness and scalability. We also evaluated InternVL2 models finetuned on our dataset using the ChartQA test set. As shown in Table 7, accuracy drops significantly because the models largely lose their language ability when trained only on Yes/No questions. Nevertheless, models trained with our loss still achieve better test performance. To further mitigate this issue, we re-trained the models on a version of our dataset where answers included full sentences (generated by GPT-4o from the question and Yes/No response) and then tested them on ChartQA. As shown in Table 8, this strategy reduces the performance drop, indicating that language ability is preserved to a greater extent. Once again, our trained model outperforms the baseline, suggesting that our attention refinement method yields models with stronger generalizability.