



Universiteit
Leiden

Gaze-Driven Efficient Inference for Vision-Language Models

Wang Lei (s3995526)

Supervisors:
Chen Qinyu & Qi Jiawen

BACHELOR THESIS– DSAI

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

August 21, 2026

Abstract

Vision-Language Models (VLMs) have achieved strong performance on multimodal reasoning tasks such as Visual Question Answering (VQA). However, their inference cost is largely dominated by the processing of dense visual tokens, making efficient visual computation an important challenge. Existing approaches typically reduce computational cost through architectural modifications or token-selection strategies, which may require additional training or specialized model designs.

This thesis investigates whether human gaze can serve as an effective external guidance signal for efficient VLM inference. To support this study, a VQA dataset containing human gaze data was collected using a Pupil Labs Neon eye tracker, comprising 268 real-world images, 718 question-answer pairs, and synchronized gaze traces. Based on this dataset, a gaze-guided inference framework was implemented to compare full-image, ROI-based, and two-scale inference strategies without modifying or retraining the underlying models.

Experiments were conducted on four open-source Vision-Language Models from the Qwen and LLaVA families. The proposed ROI strategy consistently reduced the visual-token budget by approximately 74%–82%, resulting in estimated prefill FLOP reductions of 84%–91% and substantially lower prefill latency while maintaining competitive answer accuracy. Additional analyses, including non-gaze cropping baselines, threshold sensitivity, question-type analysis, and GPT-5 semantic evaluation, further demonstrate that gaze-guided region selection provides a more effective allocation of visual computation than generic cropping strategies.

Overall, the results suggest that human gaze provides an effective and practical guidance signal for improving the computational efficiency of Vision-Language Models without requiring changes to the underlying models, offering a simple and general approach for reducing visual inference cost while largely preserving answer quality.

Contents

1	Introduction	1
1.1	Vision-Language Models and Efficient Inference	1
1.2	Human Gaze as Task-Driven Attention	1
1.3	Research Questions	2
2	Related Work	3
2.1	Vision-Language Models	3
2.2	Efficient Inference and Token Pruning	3
2.3	Human Visual Attention and Eye Tracking	4
2.4	Gaze-Guided Multimodal Inference	4
3	Dataset Construction	5
3.1	Overview	5
3.2	Gaze Data Collection	5
3.3	Reference Scene Images	7
3.4	Gaze Data Mapping	8
3.5	Gaze Segment Selection	9
3.6	Question-Answer Pair Construction	10
3.7	Data Filtering and Quality Control	11
3.8	Dataset Summary	11
4	Methodology	12
4.1	Overview	12
4.2	Gaze Heatmap Construction	13
4.3	Gaze-Driven ROI Extraction	13
4.4	Multimodal Input Variants	14
4.5	Evaluation Perspective	16
5	Experimental Setup	16
5.1	Dataset and Task Setting	16
5.2	Evaluated Models	16
5.3	Inference Conditions	16
5.4	Hardware and Processor Settings	17
5.5	Evaluation Protocol	17
5.5.1	Efficiency Metrics	17
5.5.2	Automatic Answer-Match Accuracy	18
5.5.3	GPT-5 Pairwise Evaluation	18
5.6	Experimental Baselines for ROI Comparison	19
5.7	Gaze-Mass Threshold Analysis	20
6	Results and Analysis	21
6.1	Model Performance Evaluation	21
6.1.1	Accuracy and Computational Cost	21
6.1.2	Prefill Latency	22

6.2	Comparison with Non-Gaze Crops	23
6.3	GPT-5 Pairwise Evaluation	23
6.4	Effect of the Processor Resolution Budget	24
6.5	Effect of the Gaze-Mass Threshold	25
6.6	Question-Type Analysis	27
6.6.1	Accuracy by Question Type	27
6.6.2	Case Analysis	27
7	Conclusion and Future Work	29
7.1	Summary of Findings	29
7.2	Main Contributions	29
7.3	Future Work	30

1 Introduction

1.1 Vision-Language Models and Efficient Inference

Vision-Language Models (VLMs) integrate visual perception with natural language understanding in a single framework and have rapidly become a dominant paradigm for multimodal artificial intelligence [1]. Representative models, including BLIP-2, LLaVA, InstructBLIP, Qwen-VL, and Qwen2.5-VL, have demonstrated strong performance on a broad range of multimodal tasks such as image captioning, visual question answering (VQA), and multimodal interaction [2, 3, 4, 5, 6].

Among these tasks, VQA has become one of the most representative settings for evaluating multimodal reasoning. In a typical VQA task, the model receives an image together with a natural-language question and must produce an answer based on visual evidence and contextual reasoning. This requires more than object recognition: the model must also identify spatial relations, relevant scene details, and question-dependent evidence within the image.

Despite these advances, inference efficiency remains a practical limitation. Modern VLMs usually represent images as dense sequences of visual tokens derived from image patches. As image resolution increases, the number of visual tokens grows rapidly and so does the computational burden at inference time [7, 8, 9]. Compared with text tokens, visual tokens are especially expensive because they encode dense spatial information, much of which may be irrelevant to a given question.

This issue becomes more pressing in resource-constrained environments such as augmented reality (AR) and virtual reality (VR), where latency and energy use matter directly. Similar efficiency concerns also appear in on-device and long-context multimodal systems, where visual input can quickly dominate the runtime budget [10, 11]. Existing work shows that visual computation can account for much of the inference cost in multimodal systems. Reducing redundant visual tokens while preserving task performance has therefore become a central goal in efficient VLM research [7, 8, 9].

Several recent approaches improve efficiency through token pruning, sparse attention, or dynamic token selection. Methods such as SparseVLM, MMTok, and AdaFV aim to reduce computation by retaining only informative visual regions or by improving multimodal alignment under tighter compute budgets [7, 9, 12]. However, these methods often rely on additional architectural components, learned saliency estimation, or intermediate feature analysis, all of which can introduce extra overhead. More importantly, model-generated saliency is not explicitly guided by human attention during task execution and may therefore identify different visual regions from those humans rely on for answering a question.

1.2 Human Gaze as Task-Driven Attention

Human visual attention offers a different starting point for this efficiency problem. During visual reasoning, gaze behavior reflects how perceptual and cognitive resources are allocated. Rather than processing every part of a scene equally, humans tend to focus on task-relevant regions while ignoring much of the irrelevant background. Previous studies in cognitive science have shown that eye movements are closely related to perception, semantic understanding, and reasoning processes [13, 14, 15, 16].

Unlike static saliency maps, human gaze is strongly task-dependent. The same image can produce very different gaze patterns depending on the question being asked. A question about

location may concentrate attention on specific spatial regions, while a question about attributes, text, or counting may draw attention to different visual details. Gaze therefore indicates not only where attention is directed, but also which evidence is likely to matter for the task. This close coupling between visual attention and language-guided interpretation has also been observed in psycholinguistic studies of situated comprehension [17].

Recent studies suggest that gaze information can help identify task-relevant visual regions and provide useful insight into attention behavior during multimodal tasks [18, 19]. At the same time, advances in wearable eye-tracking technology have made it increasingly feasible to collect gaze data in natural environments [20], creating new opportunities to incorporate human attention signals into multimodal systems.

Motivated by these observations, this thesis investigates whether gaze can be used as a guidance signal for efficient multimodal inference. Instead of asking the model to process the entire image at full visual cost, gaze information can be used to identify Regions of Interest (ROIs) that contain the evidence most likely to matter for the question. The central idea is that allocating more computation to gaze-relevant regions may reduce redundant visual processing while maintaining answer quality.

1.3 Research Questions

This thesis examines whether human gaze can improve the inference efficiency of Vision-Language Models in visual question answering. More specifically, it asks whether gaze can be used as an external attention signal to guide the selection of task-relevant visual regions during inference.

The main research question is formulated as follows:

Can human gaze signals improve the inference efficiency of Vision-Language Models while maintaining answer quality in visual question answering tasks?

To address this research question, the following sub-questions are investigated:

- To what extent can gaze-guided region selection reduce visual token usage and computational cost during inference?
- How does gaze-guided inference affect answer quality compared with conventional full-image inference?
- What trade-offs arise between computational efficiency and reasoning performance under different gaze-guided settings?

These questions are examined empirically on the gaze-aligned VQA dataset constructed in this thesis and on multiple open-source Vision-Language Models under a controlled inference protocol. The aim is therefore to determine whether gaze-guided region selection is effective within this concrete evaluation scope, rather than to claim universal effectiveness across all users, tasks, or model families.

2 Related Work

2.1 Vision-Language Models

Vision-Language Models (VLMs) integrate visual perception and natural language understanding within a unified framework, enabling reasoning over both images and text. Recent progress in multimodal learning has improved performance on tasks such as image captioning, visual question answering (VQA), visual reasoning, and multimodal dialogue.

Most modern VLMs contain three main components: a vision encoder, a multimodal projection module, and a large language model (LLM). The vision encoder transforms an image into a sequence of visual tokens, which are projected into a shared representation space and then processed together with textual inputs by the language model. This design allows VLMs to connect visual content with linguistic concepts and to perform multimodal reasoning.

Representative models such as LLaVA, InstructBLIP, Qwen-VL, and Qwen2.5-VL show how large-scale multimodal training improves visual understanding and language generation [3, 4, 5, 6]. These models can answer image-based questions, describe scenes, and perform tasks that require contextual visual reasoning.

Despite these capabilities, modern VLMs often incur substantial inference cost. Because images are represented as dense visual-token sequences, the computational burden increases rapidly with image resolution. Similar pressure appears in long-context multimodal systems such as long-video understanding, where large visual input streams make selective processing increasingly necessary [11]. This has motivated a growing body of work on efficient inference and visual-token reduction.

2.2 Efficient Inference and Token Pruning

Existing approaches to efficient VLM inference mainly reduce redundant visual computation through token pruning, sparse attention, token compression, and alignment-aware acceleration strategies [7, 8, 9, 12].

SparseVLM introduces language-guided visual token sparsification to reduce the number of processed visual tokens during inference [7]. By leveraging textual information, the model attempts to retain task-relevant visual regions while discarding less informative patches. Similarly, MMTok formulates token selection as a multimodal coverage optimization problem and selects informative token subsets based on multimodal similarity estimation [9].

Other approaches focus on visual token compression. OmniVLM reduces computational cost by compressing visual representations into compact token embeddings through lightweight projection modules [8]. AdaFV argues that efficiency can also be improved by rethinking visual-language alignment itself, rather than treating acceleration only as a post-hoc token filtering problem [12]. Additional work explores sparse computation schedules and dynamic token-selection strategies to reduce multimodal inference overhead.

Although these approaches improve efficiency, several limitations remain. Many require additional architectural modifications, intermediate feature analysis, or learned saliency estimation at inference time. In addition, informative regions are usually selected from model-generated signals rather than from human task focus. As a result, visual evidence relevant to the question may still be discarded during token reduction.

2.3 Human Visual Attention and Eye Tracking

Human visual attention has long been studied in cognitive science and computer vision as a window into perception and reasoning. Previous research shows that eye movements are closely related to semantic understanding, attention allocation, and decision making during visual tasks [13, 14, 15, 16].

Eye-tracking technologies provide direct measurements of gaze behavior by recording gaze locations and eye-movement trajectories. Studies in cognitive psychology show that humans selectively attend to task-relevant regions while ignoring much of the irrelevant background. Unlike static saliency estimation, gaze behavior is strongly task-dependent and changes with the reasoning objective [13, 14, 15]. The interaction between visual attention and language understanding has also been demonstrated in studies of situated comprehension [17].

Recent advances in wearable eye-tracking devices have made gaze collection in natural real-world environments increasingly practical. Devices such as Pupil Labs Neon allow researchers to capture high-resolution gaze data during everyday activities outside laboratory settings [20]. This development creates new opportunities to incorporate human attention signals into multimodal artificial intelligence systems.

Because gaze behavior reflects where humans allocate attention during visual reasoning, it may provide a useful signal for identifying semantically important regions. Gaze has also been explored in machine learning applications beyond traditional computer vision, including interpretability, user alignment, and human-computer interaction [21]. However, its use as a direct guidance mechanism for efficient multimodal inference remains relatively underexplored.

2.4 Gaze-Guided Multimodal Inference

Recent studies have explored the connection between human gaze and multimodal reasoning. Their findings suggest that gaze can provide useful cues for identifying task-relevant visual regions and for understanding attention behavior during visual tasks [18, 19, 22].

Unlike conventional saliency-based approaches, gaze more naturally reflects task-driven attention and semantic priorities during visual reasoning [18].

VOILA demonstrates that human gaze can improve multimodal reasoning by providing task-relevant attention cues for visual understanding [18]. G-VOILA extends this idea to interactive information querying in everyday scenarios, enabling users to retrieve information through gaze-based interaction rather than focusing on computational efficiency [19]. More recently, GazeVLM explores gaze-guided efficient inference by using gaze information to identify regions of interest (ROIs), reducing redundant visual computation while preserving global contextual information through a two-scale image representation [22].

Although these studies demonstrate the potential of gaze-guided multimodal inference, the field is still at an early stage. Existing work has primarily been evaluated on benchmark datasets or controlled experimental settings [18, 19, 22]. The use of wearable eye tracking to collect real-world gaze data and evaluate gaze-guided efficient inference under natural viewing conditions remains comparatively underexplored.

Building upon the above literature, this thesis investigates whether gaze signals collected from real-world wearable eye-tracking sessions can effectively guide efficient inference in Vision-Language Models. This work focuses on the trade-off between computational efficiency and answer quality

when gaze information is used to guide visual input selection.

3 Dataset Construction

3.1 Overview

This chapter describes the dataset used in this thesis. The dataset consists of reference scene images, gaze recordings collected using a Pupil Labs Neon wearable eye tracker, and VQA annotations, comprising 268 images and 718 question-answer pairs.

The construction pipeline involves several main steps. First, gaze data were collected in real-world scenes using the Pupil Labs Neon device. Second, the recorded gaze data were mapped onto the corresponding reference scene images using Pupil Cloud. Finally, the recorded questions were paired with their corresponding answers and organized as VQA annotations. Together, these steps produced a dataset that links visual content, gaze information, and VQA annotations.

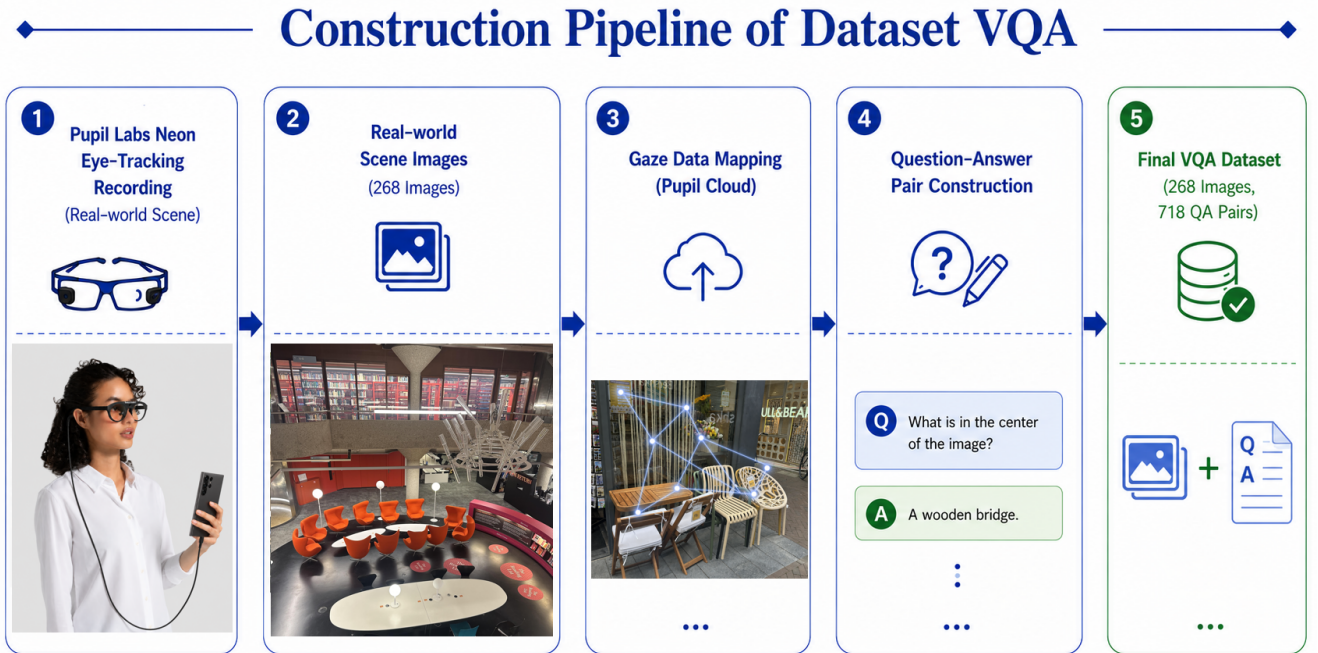


Figure 1: Overview of the dataset construction pipeline.

Figure 1 provides an overview of the complete data collection and annotation workflow. It illustrates how gaze recordings, reference scene images, and VQA annotations are integrated into a dataset for the experiments presented in this thesis.

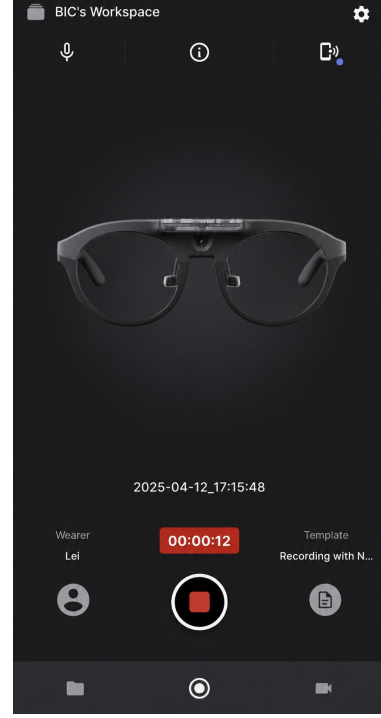
3.2 Gaze Data Collection

Gaze data were collected using the Pupil Labs Neon wearable eye-tracking system in real-world environments. The wearable eye tracker enabled gaze recording during natural scene exploration, making it suitable for capturing visual attention under everyday viewing conditions.

Figure 2 illustrates the gaze recording setup. The participant wore the Pupil Labs Neon eye-tracking glasses, which were connected to a smartphone via a USB Type-C cable. The smartphone ran the Pupil Labs Companion application, which was used to establish the device connection, initiate recording sessions, and monitor the recording status throughout data collection.



(a) Pupil Labs Neon recording setup



(b) Pupil Labs Companion application

Figure 2: Example of the gaze data collection setup. (a) The Pupil Labs Neon eye tracker worn during data collection. (b) The Pupil Labs Companion application interface.

After the connection was established, the participant wore the eye-tracking glasses and naturally explored different real-world environments, including university buildings, libraries, supermarkets, retail stores, parks, and other public spaces. Whenever a visually informative scene suitable for VQA was encountered, a new recording session was initiated. Each recording lasted approximately 30–45 seconds, with a maximum duration of one minute.

During each recording session, the eye tracker synchronously recorded the first-person scene video, binocular gaze coordinates sampled at 200 Hz, blink events, eye-state information, and inertial measurement unit (IMU) measurements. Together, these synchronized signals captured both the visual content of the environment and the participant’s gaze behaviour throughout the recording.

Figure 3 presents three representative frames extracted from a gaze recording together with the corresponding gaze points. The examples illustrate how the participant’s visual attention shifts between different objects and regions of interest during natural scene observation. The recorded gaze points capture the participant’s visual attention at different moments throughout the recording. These recordings provide the input for the subsequent gaze mapping process.



Figure 3: Example frames from a gaze recording with corresponding gaze points.

After recording, all sessions were automatically uploaded to Pupil Cloud through the Pupil Labs Companion application and exported in the native Pupil Labs recording format. The exported recordings were then processed during the gaze mapping stage, where the recorded gaze data were aligned with the corresponding reference scene images to establish the visual-gaze correspondence required for the proposed VQA dataset. The detailed mapping procedure is described in Section 3.4.

3.3 Reference Scene Images

For each valid recording, a corresponding scene image was captured using a smartphone camera. The scene image captures the same real-world scene as the corresponding recording and serves as the reference image for the subsequent gaze mapping process. In total, 268 reference scene images were collected, each corresponding to one valid recording and having a resolution of 4032×3024 pixels.

Although the scene camera of the eye tracker simultaneously recorded the corresponding scene video during each recording session, the reference scene images were captured separately using a smartphone camera rather than extracted from the recorded video. The use of smartphone images provides several advantages. First, smartphone images provide higher resolution and better image quality, preserving fine-grained visual details such as small objects, textual information, and object attributes. Second, compared with the scene camera, smartphone images are less affected by lens distortion and therefore provide a more faithful representation of the real-world scene. Finally, smartphone cameras generally produce more stable image quality under varying lighting conditions, reducing the occurrence of overexposed or underexposed images. These advantages make smartphone images more suitable as reference scene images for gaze mapping and subsequent VQA annotation.

Figure 4 illustrates the distribution of scene categories in the dataset. The dataset covers three major scene groups: Indoor Public Spaces, Indoor Commercial Spaces, and Outdoor Environments, including libraries, academic buildings, museums, supermarkets, clothing stores, retail stores, furniture stores, cosmetics stores, record stores, parks, architectural scenes, and vehicle-related scenes.

Scene Category Distribution

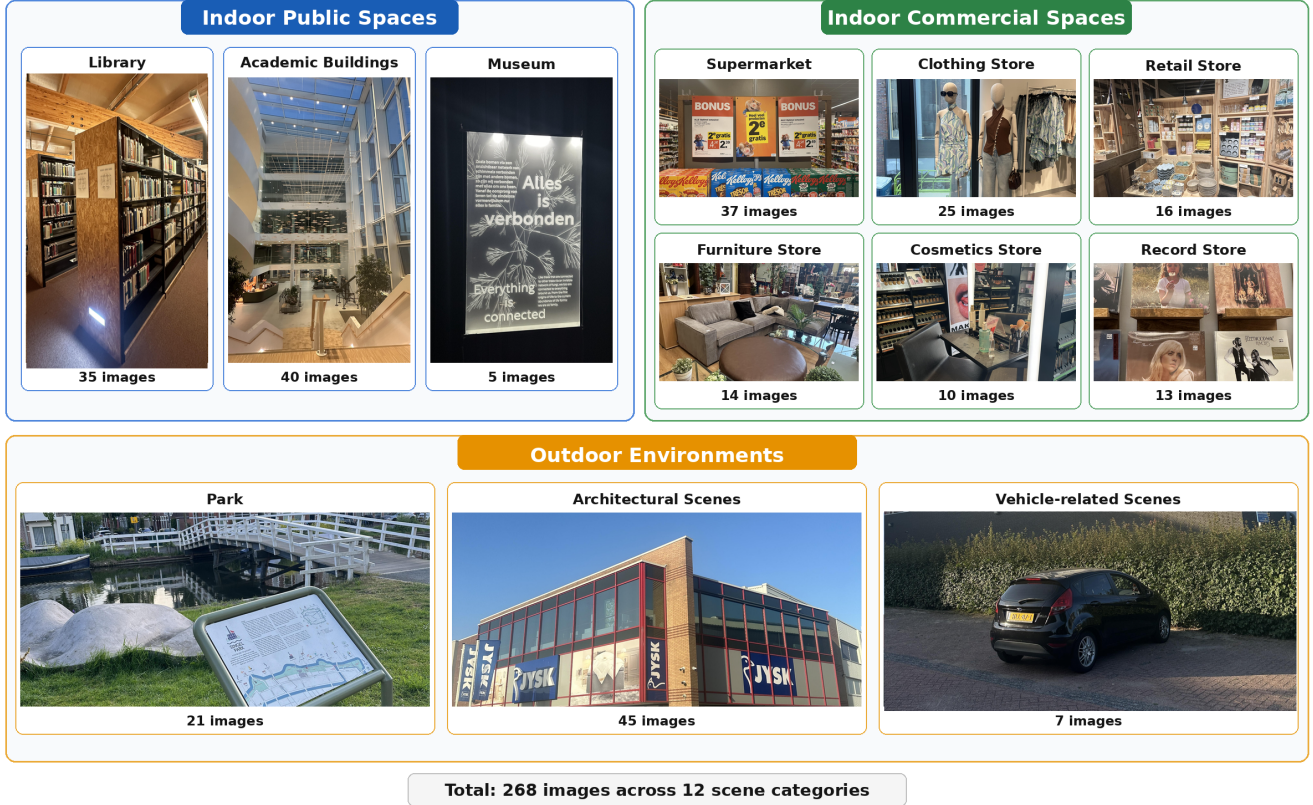


Figure 4: Distribution of scene categories in the dataset and representative examples.

The diversity of scene categories provides a broad range of visual contexts, supporting VQA tasks involving object recognition, text understanding, attribute recognition, spatial reasoning, and scene understanding.

3.4 Gaze Data Mapping

After gaze data collection and reference image acquisition, the recorded gaze data were mapped onto the corresponding reference scene images. The gaze coordinates produced by the eye tracker are defined with respect to the scene video recorded by the forward-facing camera, whereas the VQA dataset is constructed using the separately captured reference scene images. Therefore, an additional mapping step was required to establish the spatial correspondence between the recorded gaze data and the reference scene images.

The mapping process was performed using the Reference Image Mapper enrichment provided by Pupil Cloud [23]. For each scene, three inputs were required: an eye-tracking recording, a corresponding reference scene image, and a scanning recording. The eye-tracking recording contained both the scene video and the corresponding gaze data collected during natural observation. The reference scene image was captured separately using a smartphone camera, as described in Section 3.3.

The scanning recording was captured in the same environment using the scene camera of the

Pupil Labs Neon glasses and contained only scene video without gaze data. During scanning, the eye-tracking glasses were slowly moved to observe the scene from different directions, viewing angles, and distances, providing sufficient visual coverage of the environment. Depending on the size and complexity of the scene, the scanning recording typically lasted between one and three minutes, with a maximum duration of three minutes.

The eye-tracking recording, scanning recording, and reference scene image were uploaded to Pupil Cloud and processed using the Reference Image Mapper enrichment. Based on visual feature matching across the three inputs, the platform automatically established the correspondence between the recorded scene video and the reference scene image and projected the recorded gaze samples onto its coordinate system. Figure 5 shows an example of the gaze mapping result.



Figure 5: Example of the gaze mapping result. (a) Reference scene image. (b) The corresponding reference scene image with mapped gaze.

After enrichment, the mapped gaze data were exported from Pupil Cloud in CSV format. The exported `gaze.csv` file contains the timestamp of each gaze sample together with its corresponding coordinates on the reference scene image. Although the enrichment also exports additional information, such as gaze events and area-of-interest statistics, only the mapped gaze coordinates and timestamps were used to construct the VQA dataset.

The mapped gaze distribution reflects the participant’s visual attention throughout the recording and provides the gaze information required for the subsequent VQA annotation process.

3.5 Gaze Segment Selection

After gaze data mapping, each reference scene image was associated with gaze samples covering the entire eye-tracking recording. Since each recording typically lasted 30–45 seconds, the mapped gaze data contained visual attention across multiple objects and scene regions. As each question-answer pair refers to a specific object or scene region, only the gaze samples associated with the annotated question were retained. For each question-answer pair, a gaze segment of approximately one second

was extracted from the mapped gaze data. The selected segment corresponds to the period during which the participant attended to the relevant object or scene region, reducing the influence of unrelated visual attention.

Since a single reference scene image may contain multiple objects or scene regions, multiple question-answer pairs can be associated with the same image. In the final dataset, each reference scene image is typically associated with two to three question-answer pairs, with a maximum of three questions per image. Each question-answer pair is linked to an independent gaze segment corresponding to the queried object or scene region.

3.6 Question-Answer Pair Construction

To support visual question answering tasks, question-answer pairs were collected from questions that naturally arose while the participant interacted with real-world daily environments during gaze recording. Rather than being manually designed after data collection, the questions reflect the participant’s spontaneous information needs at the moment they occurred. Whenever the participant’s gaze focused on a particular object, attribute, text element, or scene component, the corresponding question was recorded and subsequently paired with an answer based on the corresponding reference scene image.

In total, 718 question-answer pairs were collected across the dataset.

The questions were categorized into the following five types according to the primary reasoning capability required to answer them:

- **Spatial Reasoning:** Requires localizing the target object through spatial relationships (e.g., left, right, above, below, or behind) before answering.
- **Attribute Recognition:** Focuses on identifying visual properties such as color, shape, material, or object category.
- **OCR/Text Recognition:** Involves directly reading or interpreting textual information in the image without additional spatial reasoning.
- **Existence:** Determines whether a specified object or entity is present in the image.
- **Semantic Reasoning:** Requires commonsense knowledge or semantic understanding beyond direct visual perception.

Table 1 summarizes the distribution of question types in the evaluation dataset and gives one example question for each type. Spatial reasoning constitutes the largest category, followed by attribute recognition and OCR/text recognition. Existence questions account for a moderate proportion of the dataset, whereas semantic reasoning represents the smallest category.

Table 1: Question type distribution and representative examples

Question Type	Count	Percentage	Example Question
Spatial Reasoning	228	31.75%	Which picture contains a Stegosaurus and a Tyrannosaurus Rex? Please answer only using the “middle” or “top”.
Attribute Recognition	171	23.82%	What is the dominant color of the cover of the magazine ‘Geographical’?
OCR/Text Recognition	164	22.84%	What is the name of the journal with a red grid pattern on the cover?
Existence	109	15.18%	Is there a yellow pepper visible in the picture?
Semantic Reasoning	46	6.41%	What phenomenon is illustrated by the triangle on the “The Dark Side of the Moon” album cover?
Total	718	100.00%	

3.7 Data Filtering and Quality Control

Quality control was integrated throughout the dataset construction process rather than performed as a separate post-processing step. During gaze mapping, gaze segment selection, and manual VQA annotation, each sample was reviewed to ensure consistent correspondence between the reference scene image, the mapped gaze data, and the associated question-answer pair.

Approximately 400 eye-tracking recordings were collected during data acquisition. Recordings that did not satisfy the quality requirements were discarded throughout the construction process. After gaze mapping and manual inspection, 268 valid reference scene images were retained for the final dataset.

The primary reasons for discarding recordings included unsuccessful gaze mapping, incomplete or unstable gaze traces, and incorrect correspondence between the eye-tracking recording and the reference scene image.

Question-answer pairs were also reviewed during annotation. Annotations that were ambiguous, unreliable, duplicated, or insufficiently supported by the corresponding gaze data were discarded. This process ensured that each question-answer pair remained consistently grounded in both the reference scene image and the participant’s gaze.

3.8 Dataset Summary

The following table summarizes the key statistics of the dataset.

Property	Value
Reference scene images	268
Question-answer pairs	718
Image resolution	4032×3024 pixels
Scene categories	12
Question-answer pairs per image	2–3
Data modalities	Image, gaze trace, question, answer

4 Methodology

4.1 Overview

This chapter presents the gaze-guided inference framework used in this thesis. The framework operates entirely at inference time without any additional model training. It uses human gaze as a task-driven signal to reduce the visual computation required by Vision-Language Models (VLMs) while preserving answer quality in visual question answering.

For each image-question pair, the framework takes an input image I , an aligned gaze segment P , and a textual question T . It then proceeds in three steps. First, the gaze segment is converted into a dense heatmap that represents the spatial distribution of visual attention. Second, the heatmap is used to extract a gaze-guided Region of Interest (ROI). Third, different multimodal input variants are constructed using the full image, the ROI crop, and, when required, a low-resolution global view. Across all variants, the question remains the same, and only the visual input changes. Concrete runtime settings and experimental parameters are described in Chapter 5.

Figure 6 illustrates the overall pipeline on a representative scene image. The highlighted attention pattern and ROI are included for illustration only and do not represent quantitative analysis results. This presentation makes the transition from the original image to the ROI-based model input easier to understand than an abstract flowchart.

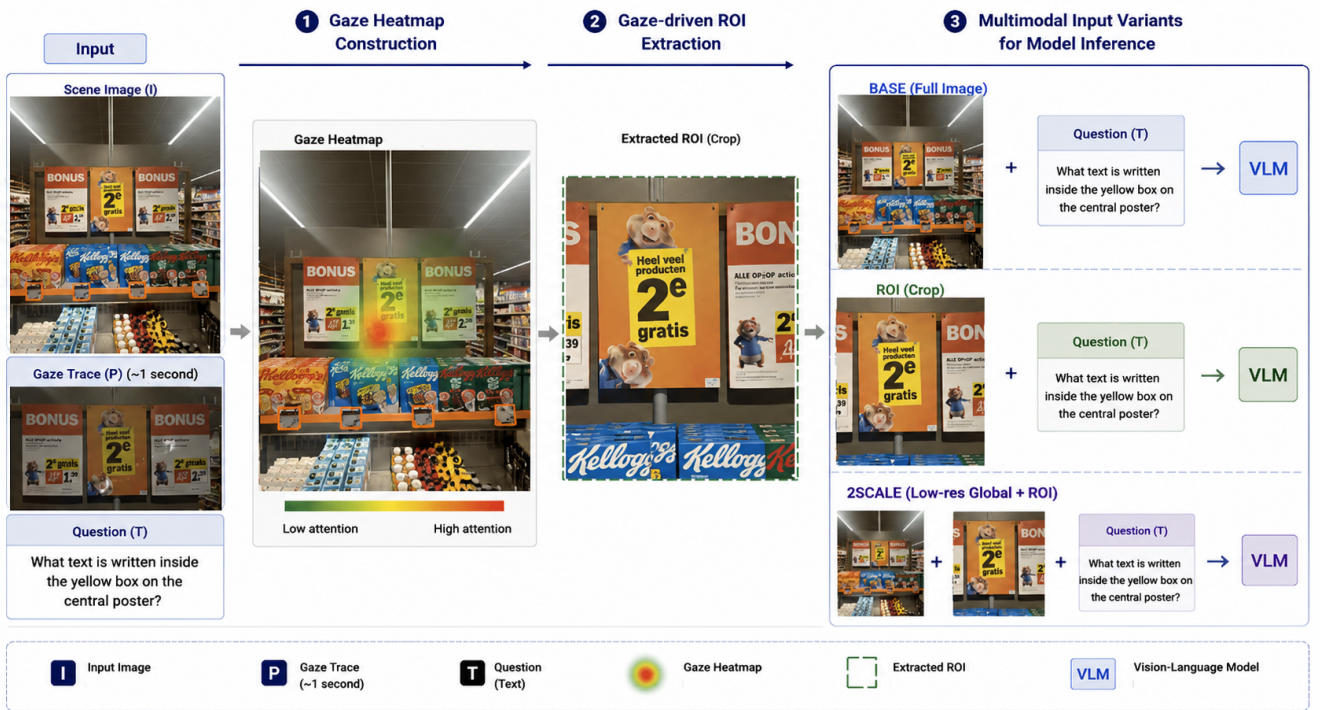


Figure 6: Overview of the gaze-guided inference framework

4.2 Gaze Heatmap Construction

The gaze traces collected using the Pupil Labs Neon eye-tracking device provide the primary source of attention information. Each trace consists of a sequence of normalized image coordinates

$$P = \{(u_i, v_i)\}_{i=1}^N,$$

where $u_i, v_i \in [0, 1)$ denote the horizontal and vertical locations of each gaze sample on the image plane.

To obtain a stable attention representation, the raw gaze trace is projected onto a fixed 224×224 canvas. Let

$$M \in \{0, 1\}^{224 \times 224}$$

denote the resulting discrete gaze map. In the implementation used in this thesis, each gaze sample is first clipped to the valid range and then mapped onto the canvas according to

$$x_i = \lfloor 224u_i \rfloor, \quad y_i = \lfloor 224v_i \rfloor.$$

The corresponding grid cell is then set to 1. If multiple gaze samples fall into the same cell, the cell value remains 1 rather than accumulating counts, resulting in a sparse binary gaze map.

The binary gaze map is subsequently smoothed using a Gaussian filter with kernel size 199×199 , so that each gaze sample influences a local neighborhood rather than a single grid cell. This produces a smoother spatial attention distribution. Let the smoothed gaze map be denoted by $\tilde{G}$. The normalized heatmap G is then computed as

$$G(x, y) = \frac{\tilde{G}(x, y) - \min(\tilde{G})}{\max(\tilde{G}) - \min(\tilde{G})},$$

whenever

$$\max(\tilde{G}) > \min(\tilde{G}).$$

Otherwise, the normalized heatmap is defined as the zero matrix. At this stage, the normalized heatmap satisfies

$$G \in [0, 1]^{224 \times 224},$$

where larger values indicate stronger gaze concentration.

Before ROI extraction, the normalized heatmap is resized using bilinear interpolation to match the spatial resolution of the image used for inference. In the default setting of this thesis, this coincides with the original image resolution unless an optional long-edge downscaling is applied beforehand. This allows the ROI to be extracted directly in the coordinate space of the image passed to the subsequent inference pipeline. The resized heatmap is then used for the gaze-guided ROI extraction described in the next subsection.

4.3 Gaze-Driven ROI Extraction

After the gaze heatmap has been resized to the spatial resolution of the image used for inference, a gaze-driven ROI is extracted to identify the visually important part of the scene. Let $\bar{G} \in [0, 1]^{H \times W}$

denote the resized heatmap and let $\rho \in (0, 1]$ denote the retained gaze-mass ratio. The pixels of $\bar{G}$ are sorted in descending order of attention intensity, and the smallest pixel set Ω_ρ satisfying

$$\sum_{p \in \Omega_\rho} \bar{G}(p) \geq \rho \sum_p \bar{G}(p)$$

is selected. In other words, Ω_ρ is the smallest set of highly attended pixels whose cumulative weight covers a proportion ρ of the total gaze mass. A binary support mask is then constructed from Ω_ρ , and the smallest axis-aligned bounding box covering this mask is extracted.

Let the resulting raw bounding box be

$$B_\rho = (y_0, y_1, x_0, x_1).$$

If no valid support region is obtained, the implementation falls back to the full image extent. In the current implementation, a zero-valued heatmap effectively leads to a full-image selection mask, so the fallback behavior is consistent with using the whole image.

Directly cropping the raw box may produce unstable results when the attended region is narrow or fragmented. For this reason, the box is further expanded with a context-preserving margin. Let

$$w_B = x_1 - x_0, \quad h_B = y_1 - y_0.$$

The horizontal and vertical padding terms are computed as

$$m_x = \max(\lfloor 0.1 w_B \rfloor, \max(\lfloor a/2 \rfloor, 1)), \quad m_y = \max(\lfloor 0.1 h_B \rfloor, \max(\lfloor a/2 \rfloor, 1)),$$

where a is a runtime alignment hyperparameter. In the current experiments, this parameter is fixed to

$$a = 20.$$

The expanded box is then clipped to the valid image boundary.

The resulting crop is denoted by

$$I_r = I[\hat{B}_\rho],$$

where $\hat{B}_\rho$ denotes the expanded and clipped bounding box. After cropping, the ROI may optionally be resized again. In the current implementation, this post-crop resizing is performed in two possible stages: first, the crop may be upsampled if its short edge is smaller than a specified minimum threshold; second, it may be downsampled if its long edge exceeds a specified maximum threshold. Both operations preserve aspect ratio and are optional runtime settings.

In the default setting of the main pipeline, no additional short-edge upscaling or long-edge downscaling is applied unless these options are explicitly enabled for a particular experiment. The resulting ROI is then passed to the downstream visual processor and inference pipeline.

4.4 Multimodal Input Variants

Once the ROI has been extracted, the next question is how visual evidence should be presented to the VLM. Using only the full image preserves complete scene context, but it also forces the model to spend computation on irrelevant background regions. Using only the ROI reduces visual redundancy, but it may remove global cues such as scene layout, object co-occurrence, or spatial

relations outside the attended crop. To balance these trade-offs, the framework defines three multimodal input variants.

Let I_b denote the image used for inference after any optional image-level preprocessing, such as long-edge downscaling, and let I_r denote the gaze-guided ROI crop obtained from the previous subsection.

Full-Image Inference. The full-image variant provides the inference image together with the question to the multimodal processor. If $\mathcal{P}$ denotes the multimodal processor that converts an ordered image list and a text query into model inputs, this variant is written as

$$X_{\text{base}} = \mathcal{P}([I_b], T).$$

This input preserves the complete scene context and serves as the reference setting.

ROI-Based Inference. The ROI-only variant replaces the full image with the gaze-guided crop:

$$X_{\text{roi}} = \mathcal{P}([I_r], T).$$

This reduces the amount of visual content that must be encoded while keeping the textual question unchanged.

Two-Scale Inference. The third variant combines a coarse global view with the fine-grained ROI crop. In the current implementation, the global image is formed by resizing the inference image to a square thumbnail of fixed edge length:

$$I_g = \text{Resize}(I_b; s_g, s_g),$$

where s_g denotes the global thumbnail edge length. The resulting two-scale input is

$$X_{2\text{scale}} = \mathcal{P}([I_g, I_r], T).$$

Thus, the global view preserves coarse scene structure at low cost, while the ROI view retains local detail inside the attended region.

In the current implementation, the two images are passed to the processor in a fixed order, with the global view first and the ROI second. The only difference between conditions is the visual evidence supplied to the model.

This design is based on a simple observation: both local detail and global context can matter for visual reasoning, but they do not need to be represented at the same resolution. The two-scale variant therefore allocates high visual fidelity to the attended crop while preserving a compressed representation of the full scene.

In the main experiments, s_g is set to 28, meaning that the global image is first resized to a 28×28 thumbnail before being passed to the multimodal processor. The processor subsequently applies its own model-specific preprocessing and tokenization.

Across all three variants, the framework uses the same textual question and the same multimodal prompting interface for each image-question pair. The model is always used in an inference-only setting: no model parameters are updated, and no additional training objective is introduced. Thus, the only difference across conditions lies in the ordered set of visual inputs supplied to the processor. This design ensures that any observed difference in model behavior can be attributed to the visual input construction strategy rather than to changes in the textual prompt or model adaptation procedure. Concrete decoding options, output-length limits, hardware conditions, and model-specific runtime settings are described in Chapter 5, since these belong to the experimental protocol rather than to the method definition itself.

4.5 Evaluation Perspective

The proposed framework is evaluated from two complementary perspectives: computational efficiency and answer quality.

From the efficiency perspective, the central question is whether replacing the full image with a gaze-guided ROI, or augmenting that ROI with a compressed global view, reduces the computational cost of multimodal inference.

From the answer-quality perspective, the central question is whether these reduced visual inputs still preserve sufficient scene evidence to support reliable answers.

The specific evaluation metrics, judging procedures, and aggregation rules are described in Chapter 5. Together, they operationalize these two evaluation perspectives and provide the basis for the empirical analysis presented in Chapter 6.

5 Experimental Setup

This chapter describes the experimental setup used throughout this thesis. It introduces the evaluation dataset, the evaluated vision-language models, the inference conditions, the hardware and processor settings, and the evaluation protocol. The described settings support the experiments reported in Chapter 6, including the comparison with non-gaze cropping strategies, the GPT-5 pairwise evaluation, and the gaze-mass threshold analysis.

5.1 Dataset and Task Setting

All experiments in this thesis are conducted on the visual question answering dataset described in Chapter 3. The dataset contains 268 real-world images and 718 question-answer pairs. Each sample consists of an image, the corresponding gaze trace collected during human observation, a question, and a reference answer.

5.2 Evaluated Models

The experiments evaluate four vision-language models: Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and LLaVA-OV-1.5-8B-Instruct [6, 24, 25]. The selected models cover different parameter scales and two representative VLM families, enabling evaluation across diverse models.

5.3 Inference Conditions

The experiments compare three inference conditions: BASE, ROI, and 2SCALE.

Full-Image Inference (BASE). The full image is provided as the visual input.

ROI-Based Inference (ROI). The gaze-guided ROI crop is provided as the visual input.

Two-Scale Inference (2SCALE). A low-resolution global image together with the gaze-guided ROI crop is provided as the visual input.

5.4 Hardware and Processor Settings

All experiments are conducted on a single NVIDIA A100-SXM4-40GB GPU on the Snellius computing cluster.

Qwen2.5-VL-3B-Instruct, **Qwen2.5-VL-7B-Instruct**, and **Qwen3-VL-8B-Instruct** were evaluated using their default processor configurations [6, 24].

LLaVA-OneVision-1.5-8B-Instruct [25]. The model was evaluated under two processor configurations: the checkpoint-default setting (`min_pixels=3136`, `max_pixels=3240000`) and a higher-resolution setting that increased the upper pixel budget to `max_pixels=12192768` while keeping the same lower bound.

5.5 Evaluation Protocol

The experiments evaluate both computational efficiency and answer quality.

5.5.1 Efficiency Metrics

Efficiency is measured using visual token count, total token count, prefill FLOPs, and prefill latency. Visual token count refers to the number of visual placeholder tokens that enter the language-model input context. Total token count refers to the total multimodal input-context length, including both visual and textual context tokens but excluding generated output tokens. Prefill latency is measured under the same single-GPU setting for all conditions.

For each inference condition $c \in \{\text{BASE}, \text{ROI}, \text{2SCALE}\}$, let T_c^{prefill} denote the prefill forward-pass time over the full multimodal input context. Over a dataset of M samples, the reported average prefill latency is

$$\bar{T}_c^{\text{prefill}} = \frac{1}{M} \sum_{i=1}^M T_{c,i}^{\text{prefill}}.$$

When the thesis reports prefill-latency reduction relative to the full-image baseline, it is computed as

$$R_c^{\text{prefill}} = \frac{\bar{T}_{\text{BASE}}^{\text{prefill}} - \bar{T}_c^{\text{prefill}}}{\bar{T}_{\text{BASE}}^{\text{prefill}}} \times 100\%.$$

In this thesis, FLOPs refer specifically to the prefill stage and do not include decode-side autoregressive generation. Let L_{ctx} denote the total input-context length, d the hidden size, N the number of transformer layers, n_h the number of attention heads, d_h the head dimension, and d_{ff} the feed-forward dimension. The reported prefill FLOPs are estimated as

$$F_{\text{prefill}} = N \left(4n_h L_{\text{ctx}}^2 d_h + 4L_{\text{ctx}} d^2 + 2L_{\text{ctx}} d d_{ff} \right).$$

The first term corresponds to self-attention, the second to projection layers, and the third to the feed-forward network. Decode-side autoregressive computation is excluded from the efficiency comparison reported in this thesis. Although the raw inference logs also record a decode-inclusive total-FLOP field, all thesis tables and discussions use the prefill-only estimate defined above.

5.5.2 Automatic Answer-Match Accuracy

The primary answer-quality metric used in this thesis is automatic answer-match accuracy. Since the dataset introduced in this thesis is self-collected and no official benchmark evaluator is available, a deterministic rule-based answer-matching procedure is adopted to provide a consistent and reproducible quantitative evaluation across all experiments.

For each image-question pair, the generated answer is compared with the corresponding reference answer. Before comparison, both answers are normalized to reduce superficial formatting differences, including Unicode normalization, lowercasing, whitespace normalization, dash normalization, and the removal of non-essential punctuation and quotation marks.

After normalization, a prediction is counted as correct if at least one of the following conditions is satisfied:

- the normalized prediction exactly matches the normalized reference answer;
- the reference answer is **yes** or **no** and appears as a standalone token in the prediction;
- the reference answer is numeric or price-related and its compact normalized form appears in the compact normalized prediction;
- the normalized reference answer contains at least three characters and appears as a substring of the normalized prediction.

Over a dataset containing M samples, the reported automatic answer-match accuracy is computed as

$$\text{Accuracy} = \frac{1}{M} \sum_{i=1}^M \mathbf{1}[\text{match}(\hat{y}_i, y_i)] \times 100\%.$$

where $\hat{y}_i$ denotes the generated answer for sample i , y_i denotes the corresponding reference answer, and $\text{match}(\cdot)$ represents the deterministic answer-matching procedure described above.

This metric provides a deterministic and fully reproducible evaluation that can be applied consistently across all evaluated models and inference settings. In contrast, because it relies on lexical matching rather than full semantic equivalence, it may fail to recognize semantically equivalent answers expressed using different wording and may occasionally accept lexically similar but semantically incorrect matches. Therefore, a supplementary GPT-5 pairwise evaluation is conducted for selected experiments to provide a complementary semantic assessment of answer quality.

5.5.3 GPT-5 Pairwise Evaluation

To complement the automatic answer-match accuracy, a supplementary GPT-5 pairwise evaluation is conducted for representative experiments. Unlike the rule-based automatic evaluation, GPT-5 compares the semantic quality of two candidate answers with respect to the question and the reference answer, making the evaluation less sensitive to lexical variations.

For each image-question pair, two candidate answers are compared in a pairwise setting. The judge receives the question, the reference answer, the two candidate responses, and the corresponding

image caption when available. To reduce answer-order bias, each pair is evaluated twice with reversed answer order, and the final comparison result is obtained by aggregating the two evaluations.

In addition to an overall preference label, GPT-5 assigns four fine-grained scores, namely coverage, accuracy, details, and fluency, each on a scale from 0 to 10. Following the evaluation protocol adopted in the gaze-guided VLM study [22], the overall score is computed as

$$S = 0.40S_{\text{coverage}} + 0.40S_{\text{accuracy}} + 0.15S_{\text{details}} + 0.05S_{\text{fluency}}.$$

For each sample, the reported score for a candidate answer is obtained by averaging its dimension scores across the forward-order and reversed-order judgments and then computing the weighted score above. The reported GPT-5 score is the mean of these per-sample weighted scores across all evaluated samples. In addition, Win, Tie, and Loss rates are reported to indicate how often one inference setting is preferred over another after aggregating the two orderings. These GPT-5 results are intended as a supplementary semantic evaluation rather than a replacement for the automatic answer-match accuracy.

5.6 Experimental Baselines for ROI Comparison

To determine whether the performance advantage of ROI arises from gaze-guided region selection rather than from cropping alone, several comparison baselines are constructed under the same per-sample visual budget.

For each image-question pair, the standard gaze-guided pipeline first produces an ROI with width w_r and height h_r . These resulting dimensions are then reused to construct alternative visual inputs with different region-selection strategies. Figure 7 illustrates the resulting comparison baselines.

CENTER places a crop of the same size at the image center, **RANDOM** places the crop at a valid random location, and **DOWNSAMPLE** retains the full image and resizes it to the same final input size as the corresponding ROI image. In addition, an image-only saliency baseline **Saliency-ROI** is constructed using a pretrained DeepGaze IIE model, which replaces the human gaze signal while preserving the same ROI dimensions.



Figure 7: Illustration of the ROI comparison baselines

All comparison conditions use the same downstream multimodal inference pipeline, including identical crop resizing, model-specific visual processing, prompting, decoding, and evaluation settings. Consequently, the comparison isolates the effect of the region-selection strategy while keeping the visual budget and all other experimental settings unchanged.

5.7 Gaze-Mass Threshold Analysis

In addition to the default configuration, the experimental pipeline also supports a supplementary analysis with different gaze-mass thresholds. The threshold ρ controls the proportion of normalized gaze mass retained during ROI extraction. Smaller values produce tighter crops, whereas larger values preserve more surrounding visual context. The default setting used in the main experiments is $\rho = 0.5$.

To examine the sensitivity of the proposed method to this parameter, additional experiments are performed with

$$\rho \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60\}.$$

For each threshold, only the ROI extraction threshold is changed. The dataset, model weights, processor settings, prompting, decoding configuration, and evaluation protocol remain unchanged. Consequently, any observed differences can be attributed solely to the amount of retained gaze mass.

The BASE condition is independent of ρ because it always uses the full image, whereas the ROI and 2SCALE conditions are recomputed for each threshold value.

The threshold sweep is analyzed using the same metrics as the main experiments, including automatic answer-match accuracy, visual token count, total token count, prefill FLOPs, and prefill

latency. This analysis characterizes the trade-off between computational efficiency and answer quality under different levels of gaze-guided visual compression.

6 Results and Analysis

This chapter presents the experimental evaluation of gaze-guided visual input selection across multiple vision-language models. The evaluation considers both answer quality and computational efficiency. Additional analyses investigate the effectiveness of gaze-guided region selection, its generalizability across different models, and its robustness under different experimental settings.

6.1 Model Performance Evaluation

This section evaluates the performance of the evaluated vision-language models under the BASE, ROI, and 2SCALE settings. The analysis first compares answer accuracy and computational efficiency, followed by measured prefill latency.

6.1.1 Accuracy and Computational Cost

Table 2 summarizes the answer accuracy and computational efficiency of the evaluated models under the BASE, ROI, and 2SCALE settings. The reported metrics include answer accuracy, visual token count, total token count, and estimated prefill FLOPs.

Table 2: Comparison of accuracy and computational efficiency.

Model	Mode	Acc. (%)	Visual tok	Total tok	Prefill TFLOPs	Visual ↓	TFLOPs ↓
Qwen2.5-VL-3B	BASE	73.54	15.55k	15.59k	106.5	–	–
	ROI	82.03	2.46k	2.50k	7.7	84.2%	92.8%
	2SCALE	82.17	2.46k	2.50k	7.7	84.2%	92.8%
Qwen2.5-VL-7B	BASE	75.63	15.55k	15.59k	179.6	–	–
	ROI	78.69	2.46k	2.50k	16.1	84.2%	91.1%
	2SCALE	77.99	2.46k	2.50k	16.1	84.2%	91.0%
Qwen3-VL-8B	BASE	80.78	11.84k	11.87k	156.1	–	–
	ROI	83.70	1.88k	1.91k	14.7	84.1%	90.6%
	2SCALE	83.98	1.94k	1.98k	15.1	83.6%	90.3%
LLaVA-OV-1.5-8B	BASE	70.06	4.07k	4.11k	35.0	–	–
	ROI	75.21	2.39k	2.43k	18.5	41.3%	47.1%
	2SCALE	74.65	2.39k	2.44k	18.6	41.3%	46.9%

Across the evaluated models, ROI consistently maintains or improves answer accuracy while reducing computational cost, although the magnitude of the efficiency gains varies across different model families.

For the three Qwen-family models, ROI consistently improves or maintains answer accuracy while reducing the visual token count by approximately 84% and the estimated prefill FLOPs by more than 90% compared with the BASE setting. The 2SCALE strategy achieves answer accuracy

comparable to ROI across all three models. Although it performs slightly better for Qwen2.5-VL-3B and Qwen3-VL-8B, ROI achieves the highest accuracy on Qwen2.5-VL-7B, with all differences remaining below one percentage point. These results suggest that the gaze-guided ROI already preserves most task-relevant visual information, while incorporating an additional low-resolution global image provides only marginal improvements under the current experimental setting.

A similar trend is observed for LLaVA-OV-1.5-8B. Compared with the BASE setting, ROI improves answer accuracy while reducing the visual token count by 41.3% and the estimated prefill FLOPs by 47.1%. These reductions are smaller than those observed for the Qwen-family models because the default processor of LLaVA-OneVision-1.5 already imposes an upper limit on the input image resolution, resulting in a substantially lower visual token budget in the BASE setting. Consequently, the potential for further token reduction is inherently smaller. Nevertheless, ROI still achieves a favorable trade-off between answer accuracy and computational efficiency. The influence of the processor budget is analyzed in more detail in Section 6.4.

These findings demonstrate that the proposed gaze-guided input selection strategy generalizes well across different vision-language models. Although the magnitude of the computational savings varies across models, the results consistently show that gaze-guided regions preserve the task-relevant visual information required for answering the associated questions.

6.1.2 Prefill Latency

Table 3 reports the measured prefill latency under the BASE, ROI, and 2SCALE settings. The results presented here measure the actual runtime during the prefill stage.

Table 3: Comparison of prefill latency under different models.

Model	BASE prefill	ROI prefill	2SCALE prefill	ROI red.	2SCALE red.
Qwen2.5-VL-3B	3.20s	0.45s	0.46s	85.8%	85.8%
Qwen2.5-VL-7B	3.76s	0.54s	0.54s	85.7%	85.6%
Qwen3-VL-8B	3.60s	0.31s	0.32s	91.4%	91.1%
LLaVA-OV-1.5-8B	1.01s	0.48s	0.49s	52.5%	51.5%

Across all evaluated models, ROI substantially reduces the measured prefill latency. The three Qwen-family models achieve latency reductions of 85.7%–91.4%, closely matching the reductions observed in visual tokens and estimated prefill FLOPs. Among them, Qwen3-VL-8B achieves the largest speedup, reducing the prefill latency from 3.60s to 0.31s.

LLaVA-OV-1.5-8B exhibits a similar trend. Compared with the BASE setting, ROI reduces the measured prefill latency by 52.5%, consistent with the reductions in visual tokens and estimated prefill FLOPs reported in the previous subsection. As discussed in Section 6.4, the smaller latency reduction compared with the Qwen-family models is mainly due to the default processor budget, which already limits the input image resolution in the BASE setting.

Overall, the measured latency results validate that the computational savings introduced by the proposed gaze-guided ROI strategy translate into practical runtime improvements during inference.

6.2 Comparison with Non-Gaze Crops

The previous experiments demonstrate that gaze-guided ROI substantially improves computational efficiency while maintaining answer quality. To determine whether these gains arise from gaze-guided region selection rather than cropping alone, ROI is compared with four region-selection strategies under the same crop size: CENTER, RANDOM, DOWNSAMPLE, and Saliency-ROI.

Figure 8 compares the answer accuracy of ROI and four alternative region-selection strategies across the evaluated models.

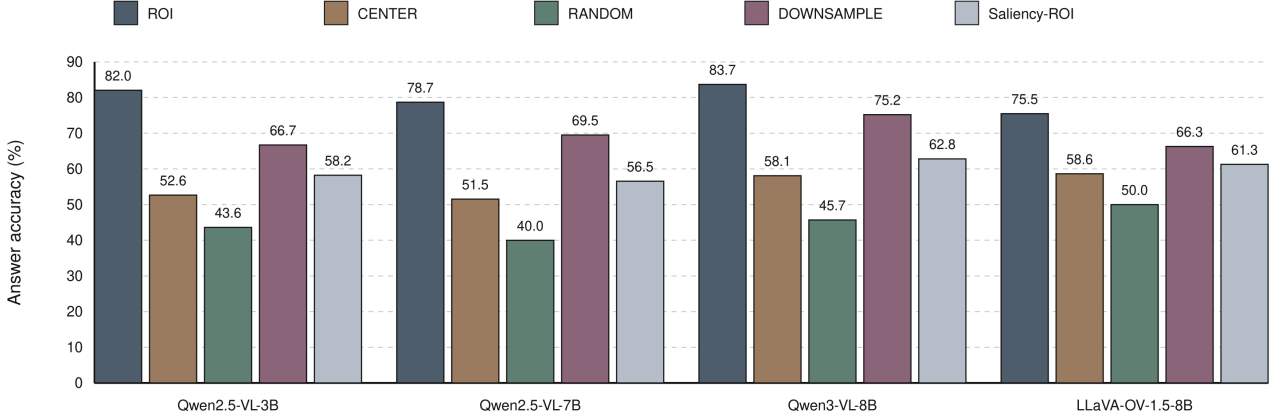


Figure 8: Answer accuracy under different region-selection strategies across the evaluated models.

Across all evaluated models, ROI consistently achieves the highest answer accuracy. Among the comparison baselines, DOWNSAMPLE performs best, followed by Saliency-ROI, CENTER, and RANDOM, and this ranking remains consistent across all evaluated models.

Compared with DOWNSAMPLE, ROI improves answer accuracy by 15.32, 9.19, 8.49, and 9.19 percentage points for Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-8B, and LLaVA-OV-1.5-8B, respectively. Compared with Saliency-ROI, the corresponding improvements are 23.81, 22.14, 20.89, and 14.21 percentage points. These results show that the performance gain is not simply a consequence of cropping or image-only saliency. Although preserving global context through downsampling is more effective than selecting arbitrary local regions, gaze-guided region selection consistently provides the highest answer quality.

These results indicate that the effectiveness of the proposed approach arises from gaze-guided region selection rather than cropping itself. Under the same visual budget, human gaze provides a more informative region-selection signal than image-only saliency or heuristic cropping strategies.

6.3 GPT-5 Pairwise Evaluation

The automatic answer-match metric provides a deterministic measure of answer correctness but may underestimate semantically equivalent responses expressed using different wording. To provide a complementary semantic evaluation, GPT-5 was used to perform pairwise comparisons between the ROI and BASE settings across all evaluated models.

Table 4: GPT-5 pairwise comparison between ROI and BASE across the evaluated models.

Model	ROI Score	BASE Score	ROI Advantage	Win (%)	Tie (%)	Loss (%)
Qwen2.5-VL-3B	8.41	7.79	+0.62	27.72	59.47	12.81
Qwen2.5-VL-7B	8.18	7.88	+0.30	25.21	55.29	19.50
Qwen3-VL-8B	8.08	7.81	+0.27	36.77	33.15	30.08
LLaVA-OV-1.5-8B	8.00	7.65	+0.35	19.64	64.48	15.88

Table 4 shows that ROI achieves higher average semantic evaluation scores than BASE across all evaluated models, with score improvements ranging from 0.27 to 0.62. The distributions of Win, Tie, and Loss vary across models. In particular, Qwen3-VL-8B exhibits the lowest tie rate together with the highest win and loss rates, suggesting that GPT-5 more readily distinguishes semantic differences between the two inference settings for this model.

Although Qwen3-VL-8B achieves the highest automatic answer-match accuracy among the evaluated models, it does not exhibit the largest improvement in the GPT-5 semantic evaluation. This difference likely reflects the distinct criteria used by the two evaluation methods. The automatic metric mainly tests whether the reference answer can be matched after normalization, whereas GPT-5 additionally evaluates semantic completeness, factual accuracy, level of detail, and fluency. In addition, Qwen3-VL-8B tends to generate substantially longer responses than the two smaller Qwen models, which allows it to preserve the key answer while also increasing the chance of introducing redundant or slightly imprecise information. As a result, some answers that are counted as correct by the automatic matcher may still receive lower semantic scores under GPT-5 evaluation. Taken together, the GPT-5 evaluation complements the automatic answer-match results, indicating that gaze-guided ROI inference preserves semantic answer quality while substantially improving computational efficiency.

6.4 Effect of the Processor Resolution Budget

To investigate the effect of the processor resolution budget, LLaVA-OV-1.5-8B was evaluated under two processor configurations while keeping all other experimental settings identical. The default configuration uses `min_pixels` = 3136 and `max_pixels` = 3240000. The enlarged configuration increases the upper bound to `max_pixels` = 12192768, allowing the original 4032×3024 images to be processed at a substantially higher effective resolution.

Table 5: Comparison of LLaVA-OV-1.5-8B under the different resolution budgets.

Budget	Mode	Acc. (%)	Visual tok	Total tok	Prefill TFLOPs	Prefill latency
<code>max_pixels</code> = 3240000	BASE	70.06	4.07k	4.11k	35.0	1.01s
–	ROI	75.21	2.39k	2.43k	18.5	0.48s
–	2SCALE	74.65	2.39k	2.44k	18.6	0.49s
<code>max_pixels</code> = 12192768	BASE	61.84	15.55k	15.59k	237.7	9.70s
–	ROI	75.35	2.46k	2.50k	19.3	0.52s
–	2SCALE	74.65	2.46k	2.50k	19.4	0.52s

Table 5 shows that increasing the processor resolution budget does not improve the performance of this checkpoint. In the BASE condition, increasing `max_pixels` raises the visual token count

from 4.07k to 15.55k, increases the estimated prefill cost from 35.0 to 237.7 TFLOPs, and increases the measured prefill latency from 1.01 s to 9.70 s, while reducing answer accuracy from 70.06% to 61.84%. In contrast, ROI and 2SCALE remain largely unaffected: answer accuracy changes by at most 0.14 percentage points, and the increases in visual tokens, prefill FLOPs, and latency are all marginal.

Because the gaze-guided crop already occupies only a small portion of the original image, increasing the processor budget has little influence on the cropped inputs. These results suggest that increasing the processor resolution budget beyond the default setting introduces substantially higher computational cost without improving answer quality. One possible explanation is that the additional high-resolution visual details make it more difficult for the model to focus on task-relevant evidence, although further investigation is required to verify this hypothesis. Overall, the default processor configuration provides a better trade-off between answer quality and computational efficiency, while the relative ranking of BASE, ROI, and 2SCALE remains unchanged.

6.5 Effect of the Gaze-Mass Threshold

To evaluate the effect of the gaze-mass threshold, experiments were conducted using different values of ρ . The threshold determines how much accumulated gaze mass is enclosed by the extracted ROI and therefore directly controls its size. A larger threshold retains more visual information but also increases the number of visual tokens and the associated computational cost. This analysis aims to identify the threshold that provides the best trade-off between computational efficiency and answer quality.

Table 6: Gaze-mass threshold sweep for the evaluated models.

(a) Qwen2.5-VL-3B

ρ	ROI tok	Tok↓	FLOPs (G)	FLOP↓	accR	acc2S
0.05	194	98.8%	549	99.5%	44.6	45.7
0.10	401	97.4%	1,065	99.0%	54.6	54.0
0.20	829	94.7%	2,212	97.9%	66.4	66.4
0.30	1,311	91.6%	3,647	96.6%	74.8	74.0
0.40	1,844	88.1%	5,399	94.9%	78.6	78.8
0.50	2,456	84.2%	7,619	92.8%	82.0	82.2
0.60	3,178	79.6%	10,550	90.1%	79.7	80.1

BASE: 15.5k tokens, 106.4 TFLOPs, accuracy 73.5%.

(b) Qwen2.5-VL-7B

ρ	ROI tok	Tok↓	FLOPs (G)	FLOP↓	accR	acc2S
0.05	194	98.8%	1,266	99.3%	43.7	43.3
0.10	401	97.4%	2,428	98.6%	52.4	51.7
0.20	829	94.7%	4,934	97.2%	67.4	67.3
0.30	1,311	91.6%	7,952	95.6%	73.1	73.1
0.40	1,844	88.1%	11,515	93.6%	77.4	77.7
0.50	2,456	84.2%	15,887	91.1%	78.7	78.0
0.60	3,178	79.6%	21,472	88.0%	79.5	78.0

BASE: 15.5k tokens, 179.2 TFLOPs, accuracy 75.6%.

(c) Qwen3-VL-8B

ρ	ROI tok	Tok↓	FLOPs (G)	FLOP↓	accR	acc2S
0.05	180	98.5%	1,872	98.8%	47.4	46.4
0.10	339	97.1%	2,892	98.1%	58.9	59.6
0.20	664	94.4%	5,049	96.8%	73.7	72.7
0.30	1,035	91.3%	7,699	95.1%	80.2	78.3
0.40	1,444	87.8%	10,827	93.1%	83.0	82.5
0.50	1,911	83.9%	14,683	90.6%	83.7	84.0
0.60	2,465	79.2%	19,616	87.4%	84.0	82.7

BASE: 11.8k tokens, 156.1 TFLOPs, accuracy 80.8%.

(d) LLaVA-OneVision-1.5-8B

ρ	ROI tok	Tok↓	FLOPs (G)	FLOP↓	accR	acc2S
0.05	194	95.2%	1,574	95.5%	49.6	51.9
0.10	401	90.1%	2,941	91.6%	56.8	57.7
0.20	827	79.7%	5,890	83.1%	67.0	67.8
0.30	1,300	68.1%	9,421	73.0%	72.0	71.9
0.40	1,814	55.4%	13,539	61.3%	75.5	75.5
0.50	2,391	41.3%	18,526	47.0%	75.2	74.7
0.60	3,033	25.5%	24,509	29.9%	74.8	75.2

BASE: 4.1k tokens, 35.0 TFLOPs, accuracy 70.1%.

Tables 6 exhibit a consistent trend across all evaluated models. Increasing the gaze-mass threshold enlarges the extracted ROI, resulting in more visual tokens and higher estimated prefill FLOPs. Consequently, the computational savings gradually decrease. In contrast, answer accuracy improves rapidly at lower threshold values before gradually reaching a plateau. At very small

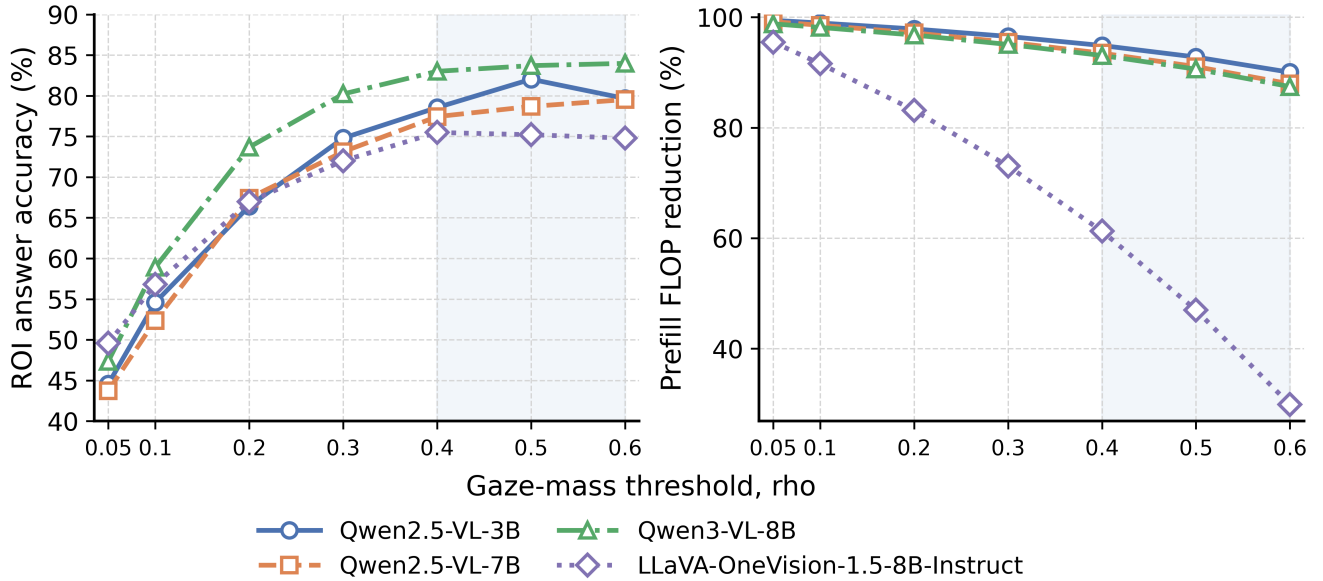


Figure 9: ROI answer accuracy (left) and prefill FLOP reduction (right) under different gaze-mass thresholds.

thresholds ($\rho = 0.05$ and 0.10), the extracted ROI often excludes relevant visual information, leading to noticeable performance degradation despite the substantial computational savings. Increasing the threshold to approximately $\rho = 0.40$ – 0.50 recovers most of the baseline accuracy while still preserving substantial reductions in visual tokens and prefill computation. A similar trend is observed for both ROI and 2SCALE inference, indicating that the threshold affects the two inference strategies consistently.

The saturation point varies slightly across models. Both Qwen2.5-VL-3B and Qwen2.5-VL-7B provide the best overall trade-off at approximately $\rho = 0.50$. Qwen3-VL-8B reaches its highest ROI accuracy at $\rho = 0.60$; however, the improvement over $\rho = 0.50$ is only 0.3 percentage points while requiring substantially higher computational cost. In contrast, LLaVA-OneVision-1.5-8B reaches its highest accuracy at $\rho = 0.40$, with larger thresholds providing little or no additional benefit.

The results indicate that thresholds between $\rho = 0.40$ and $\rho = 0.50$ provide a robust balance between answer quality and computational efficiency across different models. Therefore, $\rho = 0.50$ is adopted as the default threshold in the remaining experiments.

Figure 9 summarizes the relationship between answer accuracy and computational efficiency as the gaze-mass threshold increases. As shown in Figure 9(left), ROI accuracy increases rapidly at lower threshold values before gradually reaching a plateau. This trend is consistent across all evaluated models, indicating that enlarging the extracted ROI initially preserves more task-relevant visual information, while further increases yield only marginal accuracy improvements.

Figure 9(right) illustrates the corresponding reduction in estimated prefill FLOPs. As the gaze-mass threshold increases, the extracted ROI becomes larger, leading to higher computational cost and consequently smaller prefill FLOP reductions. Although the reduction decreases for all models, the decline is considerably steeper for LLaVA-OV-1.5-8B than for the three Qwen-family models, reflecting the larger increase in its visual-token budget as the ROI expands.

Taken together, the two plots demonstrate a clear trade-off between answer quality and

computational efficiency. Thresholds around $\rho = 0.40$ – 0.50 recover most of the attainable accuracy while still preserving substantial computational savings, making $\rho = 0.50$ a suitable operating point for the remaining experiments.

6.6 Question-Type Analysis

6.6.1 Accuracy by Question Type

To better understand which types of visual reasoning benefit most from gaze-guided image processing, model performance is further analyzed across five manually defined question categories (Section 3.6). Table 7 reports the category-wise accuracy of the three Qwen models under the BASE, ROI, and 2SCALE settings using the same evaluation criterion.

Table 7: Accuracy by question type across the three Qwen models.

Question Type	Count	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen3-VL-8B
Spatial Reasoning	228	73.25 / 85.96 / 86.40	75.00 / 82.89 / 82.02	78.51 / 84.21 / 85.53
Attribute Recognition	171	76.61 / 88.30 / 87.72	79.53 / 78.95 / 80.12	85.38 / 87.13 / 85.96
OCR/Text Recognition	164	66.46 / 73.17 / 73.78	73.78 / 74.39 / 74.39	76.22 / 79.88 / 78.05
Existence	109	91.74 / 97.25 / 96.33	85.32 / 89.91 / 84.40	96.33 / 97.25 / 99.08
Semantic Reasoning	46	43.48 / 34.78 / 36.96	47.83 / 41.30 / 45.65	54.35 / 50.00 / 52.17

Note: Values in each model column are reported as BASE / ROI / 2SCALE (accuracy in %).

First, gaze-guided inference consistently improves performance on question types that rely primarily on localized visual evidence. Spatial Reasoning shows the most consistent gains across all three models, with both ROI and 2SCALE outperforming BASE in every case. These results suggest that directing visual computation toward the gaze-guided region is particularly effective for questions involving local spatial relations, text, or object presence.

Second, the effect on Attribute Recognition is less consistent. Qwen2.5-VL-3B and Qwen3-VL-8B both benefit from gaze-guided inference, whereas Qwen2.5-VL-7B shows little change across the three settings. This indicates that the usefulness of cropping for attribute recognition depends more strongly on the underlying model.

Third, Semantic Reasoning remains the most challenging category for all three models. Neither ROI nor 2SCALE consistently improves performance, and BASE achieves the highest accuracy in all three models. Notably, 2SCALE consistently outperforms ROI, suggesting that retaining low-resolution global context partially mitigates the loss of contextual information caused by aggressive local cropping. This indicates that questions requiring broader contextual or commonsense reasoning depend more heavily on global scene information than on localized visual evidence.

6.6.2 Case Analysis

To complement the quantitative results, representative Qwen3-VL-8B cases in which BASE is correct but ROI is incorrect were qualitatively analyzed to identify recurring failure patterns. Some of these failures are recovered by 2SCALE, suggesting that part of the ROI degradation is caused by the loss of global context.

Figure 10 presents five representative Qwen3-VL-8B cases, one from each question type. Each row includes the original image with the gaze-guided ROI, the ROI crop, the ground truth, and the outputs produced by BASE, ROI, and 2SCALE.

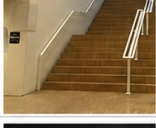
Question Type	Question	Original + ROI Box	ROI Crop	GT	BASE Answer	ROI Answer	2SCALE Answer
Spatial Reasoning	Is there a bicycle rack visible in the foreground?			Yes	Yes	No	Yes
OCR / Text Recognition	What is the title of the book with many colorful small rectangles on the cover?			10 Jaar Onderzoek & Uitgelicht	10 jaar Onderzoek uitgelicht	10 jaar WO2 Onderzoek uitgelicht	10 jaar WO2 Onderzoek uitgelicht
Attribute Recognition	How many black bar stools are visible?			4	4	3	4
Existence	Is there a handrail visible on both sides of the staircase?			Yes	Yes	No	Yes
Semantic Reasoning	What category does the journal 'CHASQUI' most likely belong to?			Literature and Culture	Latin American / Indigenous literature and culture	Academic journal	Academic / scholarly journal

Figure 10: Representative BASE-Correct, ROI-Incorrect Cases by Question Type

Figure 10 highlights several representative failure patterns. In the Spatial Reasoning and Existence examples, ROI mainly fails because important global spatial context is missing. It misses the foreground bicycle rack and the fact that handrails are visible on both sides of the staircase, whereas both BASE and 2SCALE remain correct. In Attribute Recognition, ROI undercounts the visible bar stools, indicating that even a relevant local crop can weaken multi-object estimation.

The OCR/Text Recognition and Semantic Reasoning examples illustrate two additional failure modes. In the OCR example, the crop covers the target book, but ROI produces the additional token *WO2*. In the Semantic Reasoning example, BASE uses the subtitle of *CHASQUI* to infer the specific category *Literature and Culture*, whereas ROI and 2SCALE predict the broader category *Academic journal*.

Taken together, these qualitative examples suggest that ROI failures mainly arise from the loss of global context, distraction from nearby text, and reduced semantic specificity under aggressive local cropping. Some automatically identified failures are also influenced by the strict lexical matching criterion.

7 Conclusion and Future Work

7.1 Summary of Findings

This thesis investigated whether human gaze can serve as an external guidance signal for improving the inference efficiency of Vision-Language Models in visual question answering. To this end, a real-world gaze-based VQA dataset was constructed, and a gaze-guided inference framework was developed and evaluated on multiple open-source Vision-Language Models.

The experimental results demonstrate that gaze-guided region selection substantially improves the quality–efficiency trade-off of Vision-Language Models. Across the evaluated models, the proposed framework consistently reduces computational cost while maintaining, and in several cases improving, answer accuracy. Similar trends were observed across both the Qwen family and LLaVA-OneVision-1.5, suggesting that the proposed framework is applicable to different Vision-Language Models within the scope of the evaluated experiments.

Additional analyses further support these findings. Compared with non-gaze cropping strategies, including CENTER, RANDOM, DOWNSAMPLE, and Saliency-ROI, gaze-guided ROI consistently achieves the highest answer accuracy under comparable computational budgets. The question-type analysis indicates that gaze guidance is particularly beneficial for tasks requiring localized visual evidence, such as spatial reasoning, OCR, and attribute recognition. Qualitative case studies further suggest that the remaining failure cases mainly arise when broader scene-level context is required, highlighting the complementary role of global visual information.

Overall, the results demonstrate that human gaze provides an effective guidance signal for allocating visual computation during Vision-Language Model inference. Within the scope of the evaluated dataset and models, the findings suggest that aligning model computation with human visual attention offers a practical approach for improving inference efficiency while maintaining competitive answer quality.

7.2 Main Contributions

The main contributions of this thesis are summarized as follows.

- A real-world gaze-based visual question answering dataset was constructed using a wearable eye tracker. The dataset contains 268 reference scene images, 718 question-answer pairs, synchronized gaze traces, and manually annotated question types, providing a benchmark for studying gaze-guided Vision-Language Model inference in real-world environments.
- A training-free gaze-guided inference framework was proposed to investigate whether human gaze can serve as an effective guidance signal for improving the computational efficiency of Vision-Language Models without requiring model retraining or architectural modification. The framework was systematically evaluated under three inference settings on multiple open-source Vision-Language Models.
- Extensive experiments demonstrate that human gaze provides an effective guidance signal for visual input selection. Compared with non-gaze cropping strategies under comparable computational budgets, the proposed framework consistently achieves a more favorable balance between computational efficiency and answer accuracy across different Vision-Language Models.

7.3 Future Work

Several directions can be explored in future research.

First, future work could expand the dataset by collecting gaze data from a larger and more diverse group of participants. This would improve the robustness of the dataset and enable the evaluation of inter-subject variability in gaze-guided inference.

Another direction for future work is to evaluate the proposed method across a wider range of VLM architectures using model-specific image preprocessing configurations. In this work, a unified high-resolution input setting was adopted to ensure experimental consistency. However, different VLMs are optimized for different image resolutions and processor configurations. Exploring model-specific input resolutions could improve fairness across models and enable a broader evaluation of the proposed method.

Third, this work extracts a fixed one-second gaze segment for each question. Future work could investigate adaptive or longer temporal windows to determine whether richer gaze information improves ROI extraction and downstream reasoning performance.

Future work could investigate task-aware gaze prediction models that incorporate both visual and textual information. Unlike image-only saliency prediction models, question-conditioned gaze prediction may better capture human attention during multimodal reasoning and further improve ROI selection.

Overall, this work provides an initial step toward gaze-guided efficient vision-language inference. By combining richer gaze datasets, adaptive preprocessing strategies, and task-aware gaze prediction models, future research could further improve both computational efficiency and reasoning performance, making gaze-guided multimodal inference more robust and applicable to a broader range of real-world vision-language tasks.

References

- [1] Ryan Hartsock and Ghulam Rasool. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2405.17247*, 2024.
- [2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.
- [3] Haotian Liu et al. Visual instruction tuning. In *NeurIPS*, 2023.
- [4] Zheng Dai et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.
- [5] Jinze Bai et al. Qwen-vl: Enhancing vision-language model’s perception of the world. *arXiv preprint arXiv:2308.12966*, 2023.
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

- [7] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.
- [8] Wei Chen, Zhiyuan Li, and Shuo Xin. Omnivlm: A token-compressed, sub-billion-parameter vision-language model for efficient on-device inference. *arXiv preprint arXiv:2412.11475*, 2024.
- [9] Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multi-modal coverage maximization for efficient inference of vision-language models. *arXiv preprint arXiv:2508.18264*, 2025.
- [10] Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*, 2024.
- [11] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024.
- [12] Jiayi Han, Liang Du, Yiwen Wu, Xiangguo Zhou, Hongwei Du, and Weibo Zheng. Adafv: Rethinking of visual-language alignment for vlm acceleration. *arXiv preprint arXiv:2501.09532*, 2025.
- [13] Alfred L. Yarbus. *Eye Movements and Vision*. Springer, 1967.
- [14] Marcel A. Just and Patricia A. Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354, 1980.
- [15] James E. Hoffman and Baskaran Subramaniam. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795, 1995.
- [16] Peter König, Niklas Wilming, Tim C. Kietzmann, Jose P. Ossandón, Selim Onat, Benedikt V. Ehinger, Ricardo R. Gameiro, and Kai Kaspar. Eye movements as a window to cognitive processes. *Journal of Eye Movement Research*, 9(5):25, 2016.
- [17] Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, 1995.
- [18] Ke Yan et al. Voila-a: Aligning vision-language models with user’s gaze attention. *arXiv preprint arXiv:2401.09454*, 2024.
- [19] Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. G-voila: Gaze-facilitated information querying in daily scenarios. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–33, 2024.
- [20] Pupil Labs. Pupil labs neon eye tracking glasses. <https://pupil-labs.com/products/neon>. Accessed: 2026-02-20.

- [21] Angela Lopez-Cardona, Sebastian Idesis, and Ioannis Arapakis. Integrating cognitive processing signals into language models: A review of advances, applications and future directions. *arXiv preprint arXiv:2504.06843*, 2025.
- [22] Qin Chen and Jian Qi. Eye gaze tells you where to compute: Gaze-driven efficient vlms. *arXiv preprint arXiv:2509.16476*, 2025.
- [23] Pupil Labs. Reference image mapper. <https://docs.pupil-labs.com/neon/pupil-cloud/enrichments/reference-image-mapper/>, 2025. Accessed: 2026-07-15.
- [24] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- [25] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.