Automated Generation of Eye-Tracking Trajectory Descriptions Using Vision-Language Models

Anna Bezenkova

August 4, 2025

Abstract

This work explores the application of vision-language models (VLMs) for automated generation of textual descriptions from eye tracking data visualizations. We compare several Image2Text and VLM architectures, including GIT [Wang et al., 2022], BLIP [Li et al., 2022], PaliGemma [Beyer et al., 2024], Qwen2-VL [Bai et al., 2024], and Gemma 3 [DeepMind, 2025], to determine the most effective approach to transform gaze trajectory images into structured natural language descriptions. Our results show that compact VLMs like Gemma 3 and Qwen2-VL achieve superior performance over classical Image2Text models in both semantic similarity and answer accuracy metrics. A dataset of annotated eye-tracking visualization images was created and used for model evaluation. The code are available at: GitHub repo.

1 Introduction

Eye-tracking technology enables the analysis of human visual attention by recording gaze trajectories during interaction with visual stimuli. Interpretation of such data typically requires manual annotation, which is time-consuming and subjective. In this work, we propose an automated solution using vision-language models (VLMs) capable of generating structured textual descriptions from eye-tracking visualizations.

Our approach differs from previous work in several aspects:

- Unified processing of fixation points, saccades, and pupil dilation changes within a single model.

- Use of compact VLMs suitable for deployment on limited hardware resources.

- Evaluation based on both semantic similarity and task-specific answer accuracy metrics.

1.1 Team

Anna Bezenkova designed the overall structure of the experiment, curated the dataset, and prepared this document.

2 Related Work

Vision-language pre-training has been widely explored in recent years.

BLIP [Li et al., 2022] introduced a bootstrapping framework for improving both understanding and generation tasks using noisy web data.

GIT [Wang et al., 2022] simplified the architecture by combining image encoder and text decoder under a unified language modeling task.

More recently, Qwen-VL [Bai et al., 2023] and its improved version Qwen2-VL [Bai et al., 2024] demonstrated strong performance across multiple vision-language benchmarks due to their dynamic resolution mechanism and multimodal rotary position embedding.

PaliGemma [Beyer et al., 2024] provides a versatile open-source base model trained on diverse vision-language tasks, while Gemma 3 [DeepMind, 2025] extends lightweight language models with vision capabilities and supports long context lengths up to 128K tokens.

In this work, we focus on applying such vision-language models to the specific task of generating textual descriptions from eye-tracking visualizations. Initially, a collection of images with annotated gaze trajectories was compiled, where fixations, saccades, and pupil diameter changes were clearly marked. After a thorough analysis of these visualizations, key aspects requiring description were identified, including the initial fixation point, gaze movement patterns, and the final selected answer.

Based on this analysis, a structured prompt was designed for description generation, incorporating a detailed legend that explains all eye-tracking elements (circle colors for different pupil states, arrow types for saccades). Then a dataset of labeled images was then created, where each image was paired with a textual description generated by a large language model (LLM) using the defined prompt.

To objectively evaluate the quality of the generated descriptions, a custom metric was developed, combining two key parameters: semantic similarity (measured using a SentenceTransformer model [Reimers and Gurevych, 2019]) and accuracy in identifying the final answer (assessed via a keyword-based system). Initial experiments with classical Image2Text models such as BLIP [Li et al., 2022], GIT [Wang et al., 2022], and ViT-GPT2 [Desai and Kumar, 2022] showed insufficient accuracy in interpreting eye-tracking data. This motivated us to test compact Vision-Language Models such as QwenVL2.5 3B [Bai et al., 2023] and Gemma3 4B [DeepMind, 2025], which demonstrated more adequate understanding of the relationship between visual eye-tracking elements and their textual interpretation.

3 Model Description

Vision-Language Models (VLMs) are designed to jointly process and reason over both visual and textual data. The core workflow of a VLM begins by tokenizing the input image into a set of visual tokens, typically by dividing the image into patches and projecting each patch into a vector embedding using a visual encoder such as a Vision Transformer (ViT). Simultaneously, the input text is tokenized into subword units and mapped to text embeddings. These visual and textual tokens are then concatenated into a single sequence and passed through a multimodal transformer, which enables the model to attend across modalities and capture complex relationships between image regions and textual elements. The output is generated in an autoregressive manner, where the model produces natural language descriptions or answers based on the fused visual-textual context.

We selected two state-of-the-art Vision-Language Models (VLMs) — Qwen2.5-VL [Bai et al., 2024] and Gemma 3 4B [DeepMind, 2025] — as the core components of our pipeline. These models were chosen for their compact size, multimodal architecture, and strong performance on vision-language understanding and generation tasks. In particular, Qwen2.5-VL builds upon its predecessor with enhanced visual encoder capabilities and a refined dynamic resolution mechanism that allows the model to process high-resolution images with complex layouts more effectively. It also utilises a multimodal rotary position embedding scheme, enabling precise spatial reasoning and contextual interpretation of visual elements such as fixations and saccades [Bai et al., 2024].

Qwen2.5-VL. Qwen2.5-VL is an advanced vision-language model that integrates a Vision Transformer (ViT) as its visual encoder and a Qwen2-based language model for text processing. The model employs dynamic resolution processing, allowing it to handle images and videos of arbitrary resolutions by converting them into a variable number of visual tokens. This is achieved through the use of 2D rotary position embeddings (2D-RoPE), which provide robust spatial reasoning capabilities [Bai et al., 2024]. The visual encoder is further optimized with efficient attention mechanisms and normalization layers, resulting in fast and accurate inference. Qwen2.5-VL also supports structured output generation, enabling the extraction and localization of key visual elements within complex images.

Gemma 3. Gemma 3 4B is part of Google's lightweight multimodal model family, designed for efficient and robust vision-language reasoning. It features a custom SigLIP-based vision encoder that processes images at a fixed resolution, using a Pan&Scan algorithm to adaptively crop and resize inputs while preserving important details [DeepMind, 2025]. The model leverages grouped-query attention and QK-norm for improved accuracy and supports long context windows, which is beneficial for tasks involving extended sequences or multiple images. Gemma 3 4B is optimized for deployment in resource-constrained

environments, offering fast inference and strong few-shot and zero-shot generalization abilities, making it particularly suitable for domain-specific applications such as eye-tracking trajectory interpretation.

Prior to selecting these models, we conducted an initial evaluation using classical Image2Text architectures such as BLIP [Li et al., 2022], GIT [Wang et al., 2022], and ViT-GPT2 [Desai and Kumar, 2022]. However, these approaches demonstrated limited capability in interpreting the structured visual elements of eye-tracking visualizations [Li et al., 2022], particularly in recognizing saccade patterns and final selection markers [Wang et al., 2022]. Their inability to capture semantic relationships between visual components led us to focus on more advanced VLMs capable of contextual reasoning [Desai and Kumar, 2022].

Both Qwen2.5-VL and Gemma 3 were fine-tuned on a custom dataset of eye-tracking visualization images annotated with distinct visual elements corresponding to the following gaze-related features:

- **Fixations**, represented as colored circles;
- **Saccades**, depicted as arrows connecting consecutive fixation points;
- **Final answer**, indicated by an enlarged circular marker.

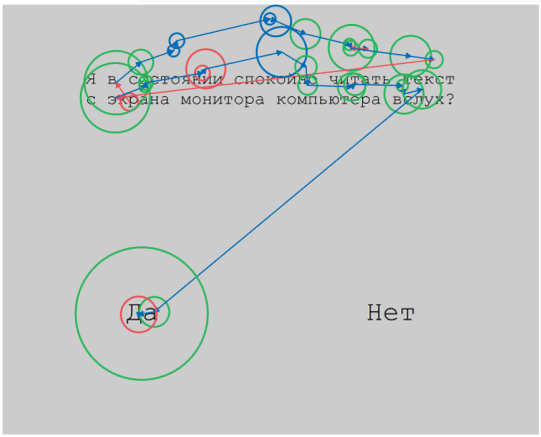

Figure 1: Example of eye-tracking visualization used as input to the model.

The input-output interface was adapted to interpret these visual elements and generate structured textual descriptions encompassing the following key aspects:

- The initial fixation point indicating the starting location of visual attention;
- A sequential account of observed visual elements;

- Identification of the final selected option;
- Interpretation of pupil dynamics in relation to cognitive or emotional states.

The model's operation can be formally described as follows:

Input: I = Image containing fixations, saccades, and pupil change indicators

Output: T = Natural language description of the observed gaze trajectory

An example of such an input image is shown in Figure **??**, illustrating the typical structure of the eye-tracking visualizations used during model training and evaluation.

4 Dataset

The dataset comprises real-world eye-tracking visualizations collected from psychological experiments involving reading comprehension. Each visualization represents a participant's gaze trajectory, including fixations, saccades, and changes in pupil diameter over time.

Key characteristics of the dataset include:

- **Fixations**: Represented as colored circles — blue (pupil constriction), red (dilation), green (stable size).
- **Saccades**: Shown as directional arrows between fixation points.
- **Final selection**: Indicated by a large circular marker, representing the participant's choice.

Each image was manually annotated with a structured textual description of the gaze trajectory, capturing the sequence of visual attention, cognitive load indicators (via pupil size), and the final decision.

	Train	Valid	Test
Images	167	19	21
Avg. Description Length	114 tokens	116 tokens	115 tokens
Vocabulary size		1,477	

Table 1: Dataset statistics showing training/validation/test splits and average description lengths.

Due to confidentiality agreements with participating research institutions, the dataset cannot be publicly released. However, the annotation process, including detailed guidelines, preprocessing steps, and prompt templates used for generating textual descriptions, is documented and available on GitHub to ensure transparency and facilitate reproducibility.

5 Experiments

5.1 Metrics

We evaluated the generated descriptions using:

- **Semantic Similarity**: Measured with Sentence-BERT [Reimers and Gurevych, 2019]
- **Answer Accuracy**: Binary match of the final selected option
- **Total Similarity**: Combined score of semantic similarity and answer accuracy

5.2 Experiment Setup

Each model was fine-tuned for 3 epochs using the AdamW optimizer with a learning rate of 2×10^{-4} and a batch size of 4. Low-Rank Adaptation (LoRA) was applied to efficiently adapt the pre-trained vision-language models to our specific task, significantly reducing both memory consumption and training time while preserving performance [Hu et al., 2021].

Early stopping was implemented based on validation loss to prevent overfitting. Training was conducted on a single NVIDIA T4 GPU, leveraging mixed-precision training to further optimize resource utilization.

5.3 Baselines

We compared against classical Image2Text models:

- BLIP [Li et al., 2022]
- GIT [Wang et al., 2022]
- ViT-GPT2 [Desai and Kumar, 2022]

These baselines showed poor performance on our specific task, indicating the need for more advanced VLM architectures.

6 Results

Model	Total Sim.	Sem. Sim.	Answer Acc.
BLIP	-0.03	0.11	0.00
GIT	-0.05	0.08	0.00
ViT-GPT2	-0.02	0.13	0.00
Qwen25-VL	0.68	0.85	0.33
Gemma 3	**0.80**	**0.89**	**0.67**

Table 2: Comparison of different models on test set. Gemma 3 outperforms other models in all metrics.

These results confirm that compact VLMs like Gemma 3 and Qwen2-VL significantly outperform classical Image2Text models in both semantic under-

standing and answer recognition. Notably, Gemma 3 achieved a total similarity score of 0.80 and correctly identified the final answer in 67% of cases.

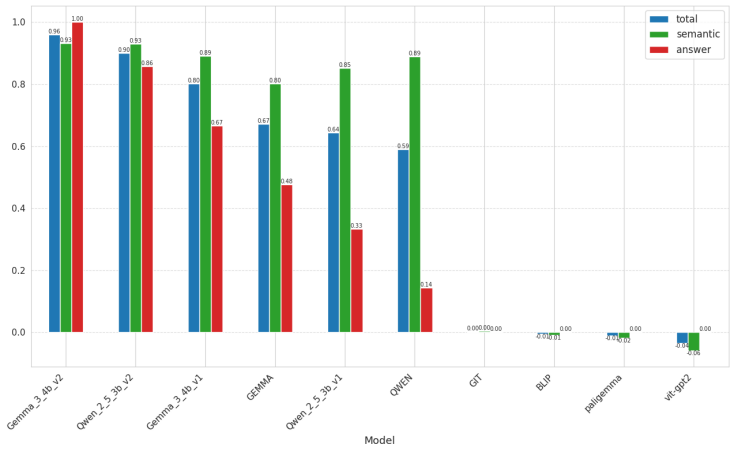

Figure 2: Comparison of vision-language models by evaluation metrics: total similarity, semantic similarity, and answer accuracy.

Having achieved acceptable performance metrics after fine-tuning the VLMs, we confirmed the models' ability to recognize the relationship between visual elements — such as fixation circles and saccades arrows — and their corresponding textual interpretation.

However, while the models demonstrated strong general understanding of gaze trajectories, there was still room for improvement in explicitly identifying and reporting the final selected answer. To enhance the accuracy of this critical output, we introduced an explicit instruction into the prompt, directly indicating the human-selected response. This modification aimed to guide the model toward a more consistent and precise identification of the final decision point in the trajectory.

After improving the prompt by explicitly including the final human-selected answer, we observed a significant boost in model performance, particularly in identifying the final decision point of the gaze trajectory. The updated evaluation results are presented in Table 3.

7 Conclusion

In this study, we explored the use of vision-language models (VLMs) for automated interpretation of eye-tracking visualizations. Our results demonstrate that compact VLMs such as Gemma 3 and Qwen 2-VL significantly outperform classical Image2Text models in both semantic similarity and task-specific accuracy metrics. The comparison of model performance, summarized in Figure 2,

Model	Total Sim.	Sem. Sim.	Answer Acc.
BLIP	-0.03	0.11	0.00
GIT	-0.05	0.08	0.00
ViT-GPT2	-0.02	0.13	0.00
Qwen25-VL	0.68	0.85	0.33
Gemma 3	0.80	0.89	0.67
Qwen25-VL_v2	0.90	0.93	0.86
Gemma 3_v2	**0.96**	**0.93**	**1.00**

Table 3: Comparison of different models on the test set after prompt modification. Gemma 3_v2 achieves perfect accuracy in identifying the final answer and outperforms other models across all metrics.

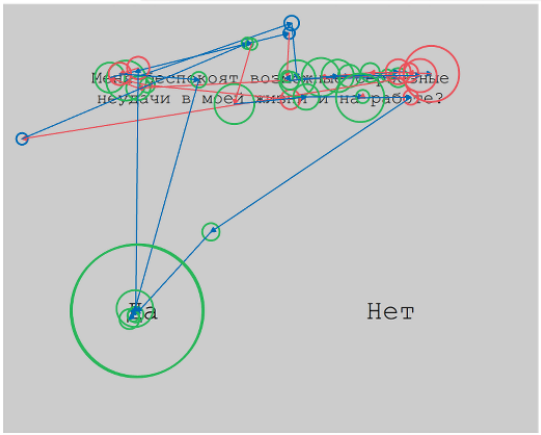

Figure 3: Example of an eye-tracking visualization from the test set used as input to the fine-tuned model.

highlights the superiority of fine-tuned VLMs, particularly Gemma 3_v2, which achieved perfect answer accuracy (1.0) and a total similarity score of 0.96.

We created a custom dataset of annotated eye-tracking visualizations and proposed a structured methodology for evaluation, including prompt engineering to improve final answer recognition. Future work includes expanding the dataset to include more complex visual layouts, exploring domain adaptation techniques for medical and further optimizing model inference.

References

[Bai et al., 2023] Bai, S., Liu, Y., Hu, S., Xu, W., Tang, J., Liu, Z., Sun, M., Chen, Y., Fu, J., Zhou, J., Zhang, P., Liu, Y., Bi, B., Ma, Z., Zhao, Y., Wen, B., Zhang, M., Liang, X., Li, L., Zou, J., Sun, X., Wang, B., Zhang, Q., Wang, X., Yang, H., Yin, J., Zhang, Q., Hu, Z., Zhou, H., Li, C., Qian, Y.,

1. Текст вопроса: Меня беспокоят возможные серьезные неудачи в моей жизни и на работе?
2. Начало движения: Взгляд начинает движение с первого слова вопроса: "Меня".
3. Изменение диаметра зрачка:
- *Сужение зрачка:* На словах "беспокоят возможные серьезные неудачи".
- *Расширение зрачка:* На словах "в моей жизни и на работе?".
- *Стабильный размер:* На остальной части вопроса.
4. Движение взгляда: Взгляд начинается с левой части экрана, фиксируется на слове "Меня", затем следует кратковременная фиксация на слове "беспокоят". После этого происходит прямая саккада к слову "возможные", за которой следует длительная фиксация на словах "серьезные неудачи". Затем взгляд перемещается ко второй строке вопроса по прямой траектории.
5. Выбор ответа: Человек выбрал ответ "Да" длительной фиксацией взгляда в области соответствующего варианта.

Table 4: Generated description of eye-tracking trajectory from test sample.

Fu, J., Men, R., Liu, X., Sun, C., Shang, L., Sun, C., Xie, G., Liu, X., Qiu, X., Zhong, Y., Ye, D., Liu, Y., Su, Y., Zhou, J., Qin, H., Wu, Y., Zhu, J., Huang, B., Liu, X., Li, X., Lan, Y., Lin, W., Cheng, X., Ye, Q., Huang, F., Liu, T.-Y., Pei, J., Huang, M., Li, H., Zhang, W., He, J., Zhang, Y., Shen, X., and Zhou, J. (2023). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

[Bai et al., 2024] Bai, S., Liu, Y., Hu, S., Xu, W., Tang, J., Liu, Z., Sun, M., Chen, Y., Fu, J., Zhou, J., Zhang, P., Liu, Y., Bi, B., Ma, Z., Zhao, Y., Wen, B., Zhang, M., Liang, X., Li, L., Zou, J., Sun, X., Wang, B., Zhang, Q., Wang, X., Yang, H., Yin, J., Zhang, Q., Hu, Z., Zhou, H., Li, C., Qian, Y., Fu, J., Men, R., Liu, X., Sun, C., Shang, L., Sun, C., Xie, G., Liu, X., Qiu, X., Zhong, Y., Ye, D., Su, Y., Zhou, J., Qin, H., Wu, Y., Zhu, J., Huang, B., Liu, X., Li, X., Lan, Y., Lin, W., Cheng, X., Ye, Q., Huang, F., Liu, T.-Y., Pei, J., Huang, M., Li, H., Zhang, W., He, J., Zhang, Y., Shen, X., and Zhou, J. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

[Beyer et al., 2024] Beyer, L., Mustafa, B., Bornschein, J., Djolonga, J., Duckworth, D., Gelly, S., Houlsby, N., Zhai, X., Minderer, M., and Kolesnikov, A. (2024). PaliGemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

[DeepMind, 2025] DeepMind, G. (2025). Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

[Desai and Kumar, 2022] Desai, C. V. S. A. and Kumar, P. (2022). Vit-gpt2: Vision transformer based encoder-decoder for multimodal tasks.

[Hu et al., 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

[Li et al., 2022] Li, J., Li, D., Xiong, C., and Hoi, S. C. H. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4356–4366.

[Wang et al., 2022] Wang, J., Yang, Z., Hu, X., Gan, Z., Hoi, S. C. H., Chua, T.-S., Feng, J., Feris, R., and Cai, D. (2022). Git: A generative image-to-text transformer for vision and language.