

Analysis of Customer Journey Video Data Using Eye-Tracking and Multimodal AI

Sami Heikkinen^[0000-0002-2283-1345], Jaani Väisänen^[0009-0000-8099-9552], Hannu Kaikonen,
Sami Makkula, and Petteri Markkanen

LAB University of Applied Sciences, Lahti, Finland
sami.heikkinen@lab.fi

Abstract. This study introduces a novel methodological approach for analyzing customer journey data using eye-tracking technology and multimodal AI. Customer journey research faces significant limitations due to reliance on resource-intensive qualitative methods that are difficult to scale. We address this gap by employing vision language models (VLMs) to automate the interpretation of eye-tracking data, eliminating the manual coding bottleneck while enabling analysis of larger and more diverse datasets. Our research evaluates five locally-deployable VLMs to determine the optimal balance between semantic accuracy and computational efficiency. Using data collected with Tobii eye-tracking glasses, we developed an analytical pipeline that synchronizes gaze location data with verbal think-aloud protocols to create a comprehensive multimodal dataset. Results demonstrate that Gemma3 (4B parameters) achieved 100% semantic accuracy on our test set while maintaining reasonable processing efficiency (43.26 seconds per image). When validated against human coding across the complete dataset, the model achieved a 74.2% recall rate. The integration of eye-tracking and verbal data revealed distinctive attention patterns including "navigational uncertainty," "confirmatory scanning," and "socially-mediated attention" throughout the customer journey. Our approach provides objective behavioral evidence of visual attention that complements traditional self-reported measures, enabling more comprehensive touchpoint analysis while aligning with event-driven perspectives from process mining research. This methodology offers promising applications for service design by identifying discrepancies between reported and actual customer attention patterns and providing a foundation for developing automated behavioral indicators to detect moments of customer confusion, decision-making, or confirmation throughout service journeys.

Keywords: Customer journey analysis; Eye-tracking; Vision language models; Service design; Customer experience.

This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/978-3-032-12478-4_1

Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use <https://www.springernature.com/gp/open-research/policies/acceptedmanuscript-terms>

1 Introduction

Halvorsrud et al. [1] observe that the lack of formalism and tools to automatically capture the actual journeys for analytical purposes hinders its operational use, highlighting the need for more systematic and automated approaches.

Eye-tracking technology offers a promising data source for understanding customer journeys, providing objective behavioral evidence of visual attention throughout. Customer journeys have gained widespread recognition for portraying customer behaviors and experiences from the perspective of human end-users [2]. The analysis of these journeys provides critical insights for service design, but current methodological approaches face significant limitations. Customer journey research has traditionally relied heavily on qualitative data collection methods and manual coding processes that are resource-intensive and difficult to scale [3].

The challenge of analyzing large volumes of customer journey data has been well-documented in the literature. Bernard and Andritsos [4] note that current research typically limits the number of journeys to be compared to less than ten, making the overall process relatively straightforward, but this approach fails to capture the diversity and complexity of customer experiences at scale. Similarly,

service interactions. However, the analysis of eye-tracking data has traditionally required labor-intensive manual annotation of areas of interest (AOIs), creating a resource bottleneck that limits its application in large-scale studies [5]. This constraint has prevented researchers and practitioners from fully leveraging the rich insights that eye-tracking can provide about customer attention patterns across touchpoints.

Recent advances in artificial intelligence, particularly vision language models (VLMs), present an opportunity to overcome these limitations. These models can potentially automate the interpretation of eye-tracking data, eliminating the manual coding bottleneck while enabling the analysis of larger and more diverse datasets. Kobialka et al. [6] highlight the value of systematic, data-driven methods for deriving user journeys, noting the shift from manually constructed journey maps toward automated analysis approaches.

This study aims to give an example of how vision language models can be utilized to automate the analysis of visual attention patterns across customer journey touchpoints. The approach presented here offers potential for scaling up customer journey analysis while providing deeper insights into the visual dimension of customer experience—a dimension that has been challenging to capture systematically with traditional methods.

2 Background

Understanding human visual attention is essential in numerous fields, including computer vision, human-computer interaction, psychology, and customer experience research. Eye movements provide rich information about real-time human visual attention and cognition, offering valuable insights into where people look and how they process visual information [7].

Customer journeys have gained widespread recognition for portraying customer behaviors and experiences from the perspective of human end-users [2]. A customer

journey consists of a sequence of steps called touchpoints, each representing an instance of communication or interaction with a service. Due to digitalization, most touchpoints leave digital traces in IT systems, although some touchpoints may remain invisible to service providers, such as face-to-face conversations or interactions with external actors [2]. These actions in the real world require an approach that provides a methodology to help organizations understand their customers' needs and ways of perceiving service offerings.

2.1 Conceptualizing Customer Experience and the Customer Journey

Lemon and Verhoef [8] provide a comprehensive framework for understanding customer experience throughout the customer journey. They define customer experience as "a multidimensional construct focusing on a customer's cognitive, emotional, behavioral, sensorial, and social responses to a firm's offerings during the customer's entire purchase journey" (p. 74). This definition emphasizes the holistic and dynamic nature of experience, which aligns with our multimodal approach to analyzing customer experience data. The customer journey perspective is closely linked to customer experience, with research using customer journeys not only as a means to understand the customer's point of view but also as a way to gain insight into their experiences [3]. This connection reflects the growing recognition that customer experience is shaped during interactions between customers and service providers [9, 10], making the systematic analysis of these interactions crucial for experience management.

Recent literature has emphasized the need for more sophisticated frameworks to analyze customer experience (CX). [11] introduced the Touchpoints, Context, Qualities (TCQ) nomenclature to provide structure and clarity to the customer experience field. While this framework offers conceptual clarity, measuring these components objectively remains challenging. Traditional research methods often rely on self-reported data, which may be subject to recall bias and cannot capture real-time reactions. One particularly challenging aspect is understanding visual attention during customer journeys - specifically, what customers look at during touchpoint interactions on their customer journey. Lemon and Verhoef [8] highlight that customer experience measurement remains fragmented, with no strongly established scales for measuring the overall customer experience across the journey. This measurement gap presents an opportunity for eye-tracking methodologies that can capture objective behavioral data across the customer journey.

2.2 Customer Journey Modeling and Analysis

An important distinction exists between planned journeys (design-time) and actual journeys that materialize during execution (run-time) [2]. Bogale et al. [12] frame this distinction as "patient pathways" representing the planned blueprint of the care process, while "patient journey" denotes an individual's actual experience revealed retrospectively. This conceptualization can be applied to customer journeys in general, distinguishing between the designed service process and the lived customer experience.

Traditional approaches to analyzing customer journeys typically fall into two categories: customer journey mapping (analyzing existing service processes "as is") and customer journey proposition (generative design activities toward possible services "to

be") [3]. Both approaches focus heavily on understanding what customers think and feel throughout their service interactions but often rely on self-reported data subject to recall bias.

The Customer Journey Modeling Language (CJML) has emerged as a domain-specific modeling approach that distinguishes between planned journeys (design-time) and actual journeys that materialize during execution (run-time). CJML also differentiates between objective elements that can be observed and logged, and subjective components encompassing personal customer experiences [2].

Beyond descriptive modeling, recent research has explored predictive approaches to customer journey analysis. Hassani and Habets [13] demonstrated that a customer's next touch point can be predicted using machine learning techniques applied to historical interaction data. Knowing what a customer will do next gives the company the ability to proactively provide support to the customer [13], potentially improving customer satisfaction while reducing service costs.

While traditional customer journey mapping typically relies on qualitative methods, recent research has explored more data-driven approaches. Arias et al. [14] demonstrated how process mining techniques can be applied to systematically map customer journeys in healthcare settings. Their approach treats each touchpoint as a process activity and uses event logs to reconstruct the actual paths customers follow, revealing variations that might not be captured through traditional mapping methods.

Process discovery, trace clustering, and decision mining methods have been used for customer journey analysis [15]. Process discovery creates representational graphs that encapsulate actions observed in an event log, capturing the main flow of customer activities during interactions. Trace clustering addresses the complexity challenge by clustering event logs into smaller groups based on similar properties, making the resulting process models more comprehensible. Decision mining enriches these models by identifying decision points and applying machine learning algorithms to understand what factors influence customer choices at each stage of the journey.

The concept of journey mapping extends beyond marketing and into other domains such as healthcare, where similar methodologies are employed to understand patient experiences. Davies et al. [16] define patient journey mapping as 'a patient-oriented project that has been undertaken to better understand barriers, facilitators, experiences, interactions with services and/or outcomes for individuals and/or their carers and family members as they enter, navigate, experience and exit one or more services in a health system by documenting elements of the journey to produce a visual or descriptive map' (p. 84). This definition emphasizes the holistic nature of mapping experiences across multiple touchpoints, which aligns with our multimodal approach to analyzing customer journey data.

Customer journeys have emerged as a significant concept in service design and management, with various approaches developed to analyze and visualize these journeys. Bernard and Andritsos [4] introduced CJM-explorer (CJM-ex), a web-based approach that bridges customer journey mapping with process mining techniques. Recent innovations in customer journey analysis have expanded beyond traditional qualitative approaches. Halvorsrud et al. [1] propose a framework that combines formal representations of user journeys with process mining techniques to automate the capture and analysis of user interactions extending across service systems. Their approach emphasizes creating a semantic journey database from which patterns that optimize

user experience can be identified. This aligns with Bernard and Andritsos's [17] genetic approach to customer journey discovery, which considers both sequential activities and contextual data when measuring similarities between journeys. Their method addresses the challenge of summarizing large collections of customer journeys into representative patterns that effectively capture the essential characteristics of the observed behaviors.

Various visual approaches to journey mapping have been identified in recent literature. Davies et al. [16] categorized journey mapping approaches into six primary types: (1) mapping key experiences and recommendations over the duration of a journey; (2) mapping by location; (3) mapping by event; (4) mapping roles, input, and experiences of key stakeholders; (5) mapping from multiple perspectives; and (6) mapping timelines of events.

Recent advances in user journey analysis have expanded beyond traditional qualitative approaches. There is a need for a shift from manually constructed journey maps toward automated, data-driven analysis, addressing the current problem that hinders their operational use in complex services.

2.3 Eye-Tracking in Customer Journey Research

Despite the widespread adoption of customer journey approaches, there are significant methodological limitations in current practice. Most notably, customer journey research relies heavily on self-reported data and subjective assessments that may be subject to recall bias and cannot capture real-time reactions [3]. One particularly challenging aspect is understanding visual attention during customer journeys—specifically, what customers actually look at during interactions with touchpoints.

Eye-tracking technology offers a promising solution to this limitation by providing objective behavioral data across the customer journey. Previous research has demonstrated the value of eye-tracking in retail environments, showing that visual attention metrics such as fixation count and total visit duration can be meaningfully correlated with consumer decision-making [18]. Similarly, Clement [19] used eye-tracking in supermarket settings to reveal distinct phases in visual decision processes, challenging traditional linear models of buying behavior.

The integration of IoT sensors—such as eye-tracking glasses in our context—with business process analysis allows for finer-grained automated monitoring and analysis [20]. This integration is particularly relevant for customer journey analysis, where understanding visual attention at different touchpoints can provide deeper insights into customer behavior and experience. [20] identify event mining from sensors as one of the four primary topics in IoT-BPM integration research. This directly relates to our approach of extracting meaningful patterns from eye-tracking sensor data.

A structured approach for translating low-level sensor data into higher-level process insights is presented by [21]. Their framework for process discovery from sensor data addresses the challenge that, the event data generated during the execution of a process is at a much lower level of abstraction. This mirrors our challenge of interpreting continuous eye-tracking data within the context of customer journeys.

Traditionally, eye tracking research has relied on specialized hardware operated in controlled laboratory environments. This approach, while accurate, presents several limitations: it is expensive, time-consuming, and difficult to scale for larger participant pools [7]. These constraints have historically limited the size of eye tracking datasets

compared to those available for other visual recognition tasks, restricting the potential for data-intensive machine learning approaches to gaze analysis.

The challenge of translating low-level sensor data into higher-level, discrete insights is not unique to eye-tracking research. [21] present a framework for process discovery from sensor data that addresses the challenge of lower level abstraction. Their approach discretizes observed sensor data into activities by applying unsupervised learning through clustering techniques. This conceptual approach is relevant to eye-tracking analysis, where continuous gaze data must similarly be translated into meaningful discrete events representing visual attention patterns.

Recent innovations have explored more accessible methods for collecting eye tracking data. For example, [7] developed a webcam-based eye tracking system that supports large-scale, crowdsourced collection of eye gaze data. Their approach demonstrated that carefully designed algorithms and protocols could obtain eye tracking data comparable to traditional lab settings but at significantly lower cost and effort. The researchers employed various techniques to address common challenges in remote eye tracking, including managing lighting conditions, head movements, and ensuring participant attention.

Gaze data has been effectively utilized in various applications including object detection, tracking, and referring expressions in videos. Vasudevan et al. [22] investigated the problem of object referring in videos using both language descriptions and human gaze data. They highlight that humans naturally gaze at objects when issuing referring expressions, and this gaze information can be a valuable cue for object localization. Their research demonstrated that combining multiple modalities—including appearance, motion, gaze, and spatio-temporal context—significantly improved the accuracy of object localization compared to using visual data alone.

Several research groups have developed datasets and analytical frameworks for various eye tracking applications. For example, Judd et al. [23] created a dataset of natural images viewed by 15 subjects, while others have focused on specific domains such as object search [24] or emotional stimuli [25]. These efforts have established important benchmarks for evaluating the performance of computational models of visual attention [7].

Recent innovations in eye-tracking analysis have moved beyond traditional approaches that rely on manual annotation of areas of interest (AOIs). Castner et al. [5] developed a novel approach to gaze scanpath comparison that incorporates convolutional neural networks (CNNs) to process scene information at the fixation level. Their method, called DeepScan, extracts image patches linked to respective fixations as input for a CNN, with the resulting feature vectors providing temporal and spatial gaze information necessary for accurate scanpath similarity comparison. By utilizing deep learning to automatically extract semantic information from gaze fixation points, their method enables more objective analysis of eye-tracking data across varying stimuli. The DeepScan approach is particularly relevant for real-world applications where stimuli vary naturally, such as customer journey analysis.

3 Rationale and Aim of the Study

Our research builds upon these foundations while addressing the specific challenge of analyzing customer journey data through eye-tracking information captured in video recordings. Rather than focusing on static images or controlled visual stimuli, we examine dynamic real-world interactions where gaze patterns must be understood in relation to verbal exchanges and environmental context. This approach requires integrating techniques from video processing, artificial intelligence-based image analysis, natural language processing, and semantic analysis to effectively identify and categorize patterns of visual attention throughout the customer journey.

While our research employs specialized eye-tracking glasses rather than webcam-based solutions, the analytical challenges addressed in previous research remain relevant. The processing and interpretation of gaze data, particularly the identification of fixations (stable gaze on a specific point) versus saccades (rapid eye movements between fixations), requires sophisticated computational approaches [7]. Similarly, the challenge of mapping raw gaze data to meaningful semantic information about what participants are looking at remains a complex problem across all eye tracking methodologies.

While Huddleston et al. [18] focused on static retail displays rather than dynamic customer journey videos, their methodological approach to defining areas of interest and analyzing visual attention patterns provides valuable precedent for our research. Our work extends this capability by applying similar principles to dynamic interactions throughout the customer journey. Particularly relevant to our research is their methodological approach, where they extracted two primary visual metrics—fixation count and total visit duration—to measure attention.

The user-centric perspective highlighted by Kobialka et al. [6] aligns with our multimodal approach to analyzing customer journey video data. While their focus on system logs as their primary data source, they acknowledge the limitations of this approach, noting that journey diagrams are usually generated by hand, and the user perspective is derived from interviews with experts and users. Our methodology extends this work by incorporating eye-tracking data as an additional layer of objective behavioral evidence, providing insights into visual attention that cannot be captured through system logs alone.

This research extends current customer journey analysis methods by applying an event-driven approach to eye-tracking data. While Halvorsrud et al. [1] emphasize that the lack of formalism and tools to automatically capture the actual journeys, our methodology addresses this gap by providing a structured framework for analyzing visual attention throughout the customer journey. Similar to how Berendes et al. [26] propose a domain-specific modeling language to represent complex customer journeys in retail settings, our approach offers a specialized technique for analyzing customer journey data through the lens of visual attention, thereby contributing to the emerging field of data-driven customer journey mapping.

This study addresses the identified gaps by introducing a novel methodological approach that combines three key elements: a) the use of locally-deployable VLMs to automate the analysis of eye-tracking data, which resolves the scalability and resource issues of manual coding while addressing data privacy concerns, b) the integration of multimodal data, by synchronizing eye-tracking data with think-aloud protocols,

enabling a deeper understanding of the relationship between visual attention and cognitive processes during service interactions, and c) the generation of objective behavioral data, which complements traditional self-reported measures and establishes a foundation for data-driven customer journey analysis in line with event-driven perspectives from process mining research. The research questions for the current study are:

RQ1: How accurately can locally-run vision language models (VLMs) identify and interpret gaze locations in eye-tracking video data from customer journeys?

RQ2: What are the trade-offs between computational efficiency and semantic accuracy when using different VLMs for automated eye-tracking analysis?

RQ3: How can the integration of eye-tracking data and verbal think-aloud protocols enhance our understanding of customer attention patterns throughout service interactions?

4 Methods

This study employed a computational approach to analyze customer journey data through eye-tracking information captured in video recordings. The methodology integrated video processing, vision language models (VLMs), and semantic analysis to systematically identify and categorize patterns of visual attention in relation to verbal think-aloud data. Our analytical pipeline, developed in R [27], follows a structured workflow: (1) data collection, (2) video pre-processing, (3) automated gaze location analysis, (4) validation of AI interpretations, and (5) integration with verbal data for comprehensive journey analysis.

4.1 Data Collection

Data was collected using Tobii eye-tracking glasses, which simultaneously recorded three data streams: (1) the participant's field of view, (2) a red marker indicating precise gaze location in real time, and (3) audio of participant think-aloud verbalizations. For this pilot study, we recorded one customer journey wherein a participant completed a predefined task within a physical servicescape environment [28]. The participant was instructed to verbalize their thoughts naturally while navigating the environment, providing concurrent verbal data alongside the eye-tracking measurements. This approach aligns with established protocols for capturing multimodal customer experience data [29]. Following the recording session, the audio component was professionally transcribed and time-stamped to enable synchronization with the visual data during analysis.

Our analytical approach shares conceptual similarities with frameworks developed for processing continuous sensor data in other domains. Janssen et al. [30] propose a four-step methodology for translating sensor event data into meaningful process models: (1) event correlation, (2) activity discovery, (3) event abstraction, and (4) process discovery. While developed for different sensor types, this framework offers a useful structure for interpreting continuous eye-tracking data to analyze run-time [2] customer journeys. Our approach similarly aims to transform low-level gaze events into higher-level semantic understanding through algorithmic processing and pattern identification.

4.2 Video Pre-processing

The continuous video footage was decomposed into sequential still images using the `av` package [31], enabling frame-by-frame analysis of gaze location. After evaluating various sampling rates, we configured the extraction at 10 frames per second (fps), which provides a temporal resolution of 100ms between images. This sampling rate was selected as optimal because it is sufficient to detect meaningful gaze fixations, which typically require a minimum duration of 200ms to indicate cognitive processing [18, 25], while maintaining computational efficiency. Higher sampling rates (e.g., 30fps) would have created redundant data given the established minimum fixation duration threshold, while lower rates might have missed brief but significant fixations. The extraction process yielded 1,770 still images for subsequent analysis.

4.3 LVM Evaluation with Gaze Location Analysis

The core innovation in our approach was the utilization of vision-capable language models to automatically identify gaze focus points in each frame, eliminating the need for manual coding. This automation addresses a significant limitation in eye-tracking research, where manual annotation of areas of interest (AOIs) typically creates a resource-intensive bottleneck [5].

To address RQ1 (accuracy of VLMs in identifying gaze locations) and RQ2 (trade-offs between computational efficiency and semantic accuracy), we conducted a systematic evaluation of multiple vision language models using a controlled testing protocol.

Privacy considerations for customer journey data necessitated using locally-deployable large language models rather than cloud-based solutions. We implemented the Ollama framework [32] to host and run various vision language models (VLMs). To gain a comprehensive understanding of performance trade-offs, five distinct VLMs were selected for this study based on a set of specific criteria. The selection criteria were local deployability, variations in model sizes, architectural diversity, and open availability. To adhere to privacy considerations for customer journey data, all models had to be executable locally within the Ollama framework. The selected models, with parameter counts ranging from 3.8B to 11B, represent different computational scales, allowing for an evaluation of the trade-off between processing efficiency and semantic accuracy. The models represent different model families (such as Llama, Gemma, and Minicpm) to avoid over-reliance on a single architecture. All chosen models were openly available and well-supported. Five models were systematically evaluated: Llama3.2-vision (11B parameters), Minicpm-v (8B parameters), Llava-phi3 (3.8B parameters), Llava-llama3 (8B parameters), and Gemma3 (4B parameters). This evaluation directly addresses RQ2 by examining how model size and architecture affect the balance between processing efficiency and semantic accuracy.

To assess model performance, we randomly selected five diverse still images from the dataset and prompted each model to describe the object at the gaze marker location. Each model was evaluated on two criteria: (1) semantic accuracy – the ability to correctly identify the object at the gaze location, and (2) computational efficiency – the processing time required per image. This evaluation approach follows established methods for assessing VLM performance in specialized visual analysis tasks [33].

4.4 Standardized Prompting Protocol

Based on the evaluation results, Gemma3 was selected as the optimal model for semantic analysis of the still images. We then conducted prompt engineering trials to determine the most effective instructions for consistent gaze location identification. The finalized prompt was: "The red circle shows eye gaze. Describe the exact location using format: [Object] at [position] (e.g., 'Monitor at upper right', 'Patient chart at center')." This structured prompt format was designed to produce consistent, parsable outputs that could be efficiently processed in subsequent analysis stages.

To maximize computational efficiency, we implemented parallel processing via the parallel package, distributing the image analysis tasks across multiple CPU cores. This parallelization approach reduced total processing time by approximately 65% compared to sequential processing.

4.5 Validation of AI Interpretations

To validate the reliability of the VLM-generated interpretations, we conducted a systematic comparison between AI and human interpretations. It is important to clarify that this process was not a formal cross-validation but rather a validation of the model's performance (recall) against a human-coded "ground truth" benchmark. Human annotation is the bedrock for training and evaluating AI models in supervised learning paradigms, providing factual and verified data obtained through direct observation. As the goal of this pilot study was to assess the model's general performance across the dataset rather than optimize its predictive power, validation based on a random sample provided a sufficient and resource-effective approach. Even though discrepancies between the experts' subjective assessment of a knowledge domain and AI's more know-what approach have been reported [34], this bias was mitigated by meticulously training the observers to use shared terminology and similar points of interest.

For each image in this validation set, we compared the AI-generated interpretation with the human coder's interpretation using object identification agreement, whether both identified the same primary object. This validation approach is consistent with recommended practices for evaluating AI visual interpretation systems [35].

For this study involving 1770 still images, the ground truth was determined from a 160-image random sample. The sample size was determined to estimate the likelihood of the AI model correctly identifying an object within the target population with a specified level of confidence and precision. We calculated the required sample size using the standard formula for estimating a population proportion based on a desired confidence level (95%). Due to the preliminary stage of the research, as well as time and resource limitations, the chosen margin of error was determined as 7.5% instead of the more traditional 5%. As there was no reliable prior estimate for the population proportion, we used $p=0.5$. This conservative value maximizes the required sample size, ensuring the study is adequately powered for the desired precision regardless of the true proportion. Given that our population size was finite at $N=1770$, the Finite Population Correction was applied, yielding a result of 156.

Since the AI model was not classifying objects into predetermined categories but rather attempting to interpret the content of the gaze point in a still image, it only produced True Positive (correct identification) and False Negative (incorrect

identification) results. To evaluate the validity of AI interpretations, the following procedure took place: 1) Ground Truth definition by human verifiers, 2) The generation of AI output, 3) Comparison, if the Ground Truth object was mentioned in AI output, 4) Calculation of the Recall figure as follows: (Number of GT objects mentioned by AI) / (Total number of GT objects).

Recall for the model's performance was 74,2%, meaning that for every 100 Ground Truth objects, the model would successfully recognize 74 of them. To get a better understanding of how well the AI model would rank different categories of images, the Ground Truth images were categorized into the following five categories: 1) a car (n=19), 2) a human figure (n=27), 3) a building (n=30), 4) entrances into spaces. (n=29), 5) human-made guidance systems, i.e. roads, signs, etc. (n=49). The reviewers deemed six observations unintelligible and, therefore, omitted from the analysis. A similar procedure to calculate the Recall value was conducted for all categories. The recall value was consistent across all categories, ranging from 68,4% (category 1) to 79,3% (category 4).

One clarification to our validation scheme needs further addressing for transparency. Our human observers might detect several objects within the subject's gaze point when determining the Ground Truth. This would result in Ground Truth depictions, such as "human figure by a door," effectively combining categories 2 and 4. In such cases, the category was determined by which of the categories appeared first in the description. The category decision was based on the assumption of visual salience, where more animate objects in a scene will be produced earlier in a sentence, an approach partially supported by Montag and MacDonald [36].

4.6 Integration and Semantic Analysis

To address RQ3 regarding the integration of eye-tracking and verbal data, we developed a multimodal analysis approach. Transcript data containing time-stamped verbal exchanges was imported and synchronized with the visual gaze data based on temporal alignment. Each gaze description was paired with any verbal data occurring within ± 1 second of the frame timestamp, creating a comprehensive multimodal dataset.

To identify patterns of sustained attention on specific objects across multiple frames, we developed a semantic similarity analysis pipeline with three components. First, we performed lexical normalization with object references were normalized through lemmatization and stop-word removal to reduce surface-level variation (e.g., "the large computer screen" and "computer monitor" would both be normalized to "computer screen"). Second, we recognized synonyms by creating domain-specific synonym groups for common objects in the servicescape environment. This allowed the system to recognize that terms like "display," "monitor," and "screen" often refer to the same physical object. Third, we determined similarity threshold through iterative testing. We established a semantic similarity threshold of 0.75 (on a scale of 0-1) using cosine similarity measures applied to word embeddings. Frame sequences with similarity scores above this threshold were classified as continuous attention on the same object, even when terminology varied slightly across frames.

This semantic analysis approach builds upon methodologies developed for customer journey analytics by Bernard and Andritsos [4], adapted specifically for eye-tracking

data. By detecting semantic similarity between consecutive gaze descriptions, our system could identify both sustained attention on single objects and transitions between different objects of interest, mapping the visual journey through the servicescape.

The final output consisted of a structured dataset containing frame numbers, timestamps, gaze focus descriptions, transcript text, speaker identification, and semantic similarity scores between consecutive frames. This integrated dataset enabled both quantitative analysis of attention patterns (e.g., duration of fixation on specific objects, frequency of transitions between objects) and qualitative assessment of the relationship between verbal communication and visual attention.

5 Results

5.1 Model Accuracy Results (RQ1)

Fig. 1 shows an example of a still image captured from the eye-tracking video, with the red marker indicating the participant's gaze focus point. Using our standardized prompt on this image and others, we evaluated each model's ability to correctly identify the object being viewed.



Fig. 1. An example of a still image captured from the video shows the red marker illustrating the focus of the viewer.

Table 1 presents example responses from five different VLMs for the image shown in Figure 1, along with human evaluation of semantic correctness. As shown in the "Semantically correct" column, four of the five models correctly identified that the gaze was focused on a computer monitor, although with varying levels of descriptive detail.

5.2 Model Efficiency Results (RQ2)

Table 2 summarizes the performance of all five evaluated models across our test set of images, showing both semantic accuracy and processing time requirements.

The relationship between model size, accuracy, and processing time revealed interesting trade-offs. The largest model (Llama3.2-vision:11b) achieved perfect semantic accuracy but required the longest processing time (average 55.48 seconds per image). In contrast, smaller models like Llava-phi3:3.8b processed images much faster (4.51 seconds) but with substantially lower accuracy (40%).

Gemma3:4b emerged as the optimal balance point, achieving 100% semantic accuracy on our test set while requiring only 43.26 seconds per image on average. When scaled to the full dataset of 1,770 images, this represented approximately 21 hours of processing time, which our parallel processing implementation reduced to 7.6 hours of wall clock time using an 8-core CPU.

Table 1. The comparison between LVMs based on their responses, responses' semantical correctness, and time elapsed getting the responses.

LVM	Response	Semantically correct	Time elapsed (sec)
Llama3.2-vision:11b	The focus of the person is on a computer screen displaying a room with multiple monitors and computers, likely indicating that they are working or interacting with digital content related to the space.	yes	84.5
Minicpm-v:8b	The red circle indicates that a person's attention or gaze is focused on another individual who appears to be seated and possibly engaged with equipment, as seen through multiple monitors within what seems like an office environment dedicated to technical work such as video editing or broadcasting control room operations. The exact nature of the interaction between this central figure and others in the scene cannot be determined from the image alone without additional context.	no	42.6
Llava-phi3:3.8b	The focus of the person is on a computer screen with a red circle on it.	yes	4.3
Llava-llama3:8b	The focus of the person is on the computer monitor, which has a red dot on it.	yes	7.8
Gemma3:4b	The image presents a somewhat unsettling and surreal scene. It appears to be a live video feed being displayed on multiple monitors, and the subject of that feed is a chaotic, dimly lit control room. The overall feeling is one of surveillance and potential tension.	yes	43.95

Table 2. The overall comparison between LVMs based on the proportion of semantically correct responses and average time elapsed for getting the responses.

LVM	Semantically correct	Time elapsed
Llama3.2-vision:11b	1	55.48
Minicpm-v:8b	0.6	42.49
Llava-phi3:3.8b	0.4	4.51
Llava-llama3:8b	0.4	7.64
Gemma3:4b	1	43.26

5.3 Integration Results (RQ3)

We analyzed the relationship between visual attention patterns and concurrent verbal data throughout the customer journey. Our dataset comprised 1,770 frames of synchronized eye-tracking and verbal data, capturing a rich multimodal representation of the participant's experience through the servicescape environment.

The data revealed distinct patterns in both visual attention and verbal reporting. The contributed utterances on 1,031 instances (58.2% of frames) with the remaining frames (739 instances, 41.8%) containing no verbal data. This differential pattern of verbalization suggests varying levels of engagement throughout the journey, with certain servicescape elements prompting more extensive verbalization than others.

Analysis of eye movements revealed 850 gaze transitions across the dataset, with an average fixation duration of 1.08 frames (approximately 108ms) between transitions. This relatively short average fixation time suggests an active visual scanning pattern throughout the journey, with the participant's attention frequently shifting between different environmental elements. The longest sustained fixation observed was 24 frames (2.4 seconds) focused on a road segment, during which the participant made verbal observations about wayfinding considerations.

The top five objects of visual attention were: Building (135 instances), Monitor (128 instances), Person (90 instances), Window (73 instances), and Sign (70 instances). This distribution reflects the environmental context of the journey, with architectural elements and informational displays receiving priority attention.

We identified 979 service interaction patterns where eye gaze on specific objects was accompanied by concurrent verbalization. These patterns represent moments where visual attention and verbal reporting were integrated, potentially indicating heightened cognitive processing at key touchpoints. Analysis of the verbal content revealed that verbalizations were predominantly related to wayfinding and navigational decision-making (e.g., "So I left the car there and now I should go this way."). Interestingly, direct verbal-visual alignment—where the object being viewed was explicitly mentioned in concurrent verbalization—was relatively infrequent. This suggests that participants often verbalized higher-level cognitive processes or contextual considerations rather than simply narrating their views.

The integration of eye-tracking and verbal data revealed several noteworthy patterns in how participants processed and responded to the servicescape environment:

Navigational uncertainty. Extended visual fixations on signs and directional cues were frequently accompanied by verbalizations expressing uncertainty or decision-making processes ("And from this direction, I don't really know anything other than guessing I'll go this way.").

Confirmatory scanning. After verbal statements indicating a decision, participants often exhibited a pattern of brief fixations on environmental elements that could confirm their chosen direction, suggesting an integration of verbal commitment and visual verification.

Socially-mediated attention. Following verbal exchanges between speakers, gaze patterns showed responsiveness to suggestions or directions provided by the other speaker, indicating that verbal interaction shapes visual attention allocation.

6 Discussion

6.1 Key Findings

This study demonstrates the viability of using locally deployable VLMs to automate the analysis of eye-tracking data in customer journey research. Our evaluation of five different VLMs revealed significant variations in both semantic accuracy and computational efficiency. Llama3.2-vision (11B parameters) and Gemma3 (4B parameters) both achieved 100% semantic accuracy on our test set, but with substantial differences in processing time. Gemma3 emerged as the optimal model, balancing high accuracy with reasonable processing efficiency (43.26 seconds per image on average).

When applied to the full dataset, our validation against human coding showed that the Gemma3 model achieved a 74.2% recall rate across all object categories, with relatively consistent performance across different types of objects (ranging from 68.4% for cars to 79.3% for entrances). These results suggest that locally-run VLMs can provide reasonably accurate automated interpretation of gaze locations in eye-tracking videos, though with some limitations in recognition accuracy compared to human coders.

The integration of eye-tracking and verbal data revealed several noteworthy patterns in how participants interacted with the servicescape environment. We identified specific behavioral patterns including "navigational uncertainty" (extended fixations on directional cues accompanied by verbalized uncertainty), "confirmatory scanning" (brief fixations verifying previously stated decisions), and "socially-mediated attention" (gaze patterns responsive to verbal exchanges). These patterns demonstrate how visual attention and verbalization interact throughout the customer journey, providing deeper insights into the cognitive processes underlying service interactions. However, it is important to note that these findings are based on a single-participant pilot study, and the generalizability of the results requires broader validation.

6.2 Comparison with Existing Knowledge

Our findings extend the current body of knowledge in both methodological and theoretical dimensions. Methodologically, our approach addresses limitations identified by Halvorsrud et al. [1]. Whereas they noted that the lack of formalism and tools to automatically capture actual journeys hinders operational use, this study

presents a concrete methodological solution that applies VLMs to solve this challenge specifically for the dimension of customer visual attention. By automating visual attention analysis through VLMs, our methodology provides a structured framework for analyzing an important dimension of customer experience that has previously been difficult to capture at scale.

Our results regarding eye movement patterns align and extend with previous eye-tracking research in retail environments. Our work extends the findings of Huddleston et al. [18] and Clement [19] from static retail displays to dynamic, multi-touchpoint customer journeys. Furthermore, by integrating verbal data, we not only identify phases of visual decision-making but also offer explanations for the cognitive processes behind them, such as the "confirmatory scanning" pattern we identified.

While our integration approach shares conceptual similarities with Castner et al.'s [5] methodology in moving beyond manual annotation of AOIs, our work shifts the focus from scanpath comparison for expertise classification to the semantic interpretation of gaze locations and their relationship with verbal data. This enables a content-focused analysis of the customer experience itself, rather than just the classification of behavior.

6.3 Applications for Customer Journey Analysis

The methodology developed in this study offers several promising applications for enhancing customer journey analysis. First, our approach provides objective behavioral data that complements traditional self-reported measures. By capturing what customers actually look at during service interactions, organizations can identify discrepancies between what customers say they notice versus what they actually attend to. For example, a retailer can analyze whether customers actually look at a new promotional sign that they reported noticing in an interview. If their gaze does not land on the sign, its placement or design has likely failed to capture attention effectively. This addresses a significant limitation identified by Følstad & Kvale [3] regarding the reliance on self-reported data in customer journey research.

Second, the integration of eye-tracking and verbal data enables more comprehensive touchpoint analysis. Lemon and Verhoef [8] highlighted that customer experience measurement remains fragmented, with no strongly established scales for measuring overall customer experience across the journey. Our methodology contributes to addressing this gap by providing objective measurements of visual attention that can be quantified and compared across touchpoints.

Third, our approach aligns with the event-driven perspective advocated by process mining researchers [4, 14]. By capturing discrete attention events throughout the customer journey, our methodology enables the analysis of visual attention patterns using similar analytical frameworks to those applied to system-generated event logs. This creates potential for integrating eye-tracking events with other digital trace data to create more comprehensive customer journey models.

Fourth, the patterns we identified (e.g., "navigational uncertainty," "confirmatory scanning") provide a foundation for developing behavioral indicators that could be used to automatically detect moments of customer confusion, decision-making, or confirmation throughout the service journey. For instance, a service provider could use the "navigational uncertainty" indicator to identify locations in a physical servicescape

(like a hospital or airport) where signage is unclear. By detecting spots where a customer's gaze wanders between signs for an extended period while they express verbal uncertainty, the provider can implement targeted improvements to wayfinding and measure their impact. These indicators could help service providers identify and address pain points in the physical servicescape design.

6.4 Future Research

Several promising directions for future research emerge from this study. First and foremost, future research must address the primary limitation of this study: the single-participant dataset. The methodology must be validated with a larger and more diverse sample, including multiple participants across different service contexts and with varying behaviors. This is essential to establish the generalizability of the identified attention patterns and to assess the robustness of the VLM-based interpretation approach.

Second, refinement of VLM prompting strategies could improve semantic accuracy. Our current prompt engineering focused on structured output format, but more sophisticated prompting approaches that incorporate domain-specific knowledge about servicescape environments might yield higher accuracy rates.

Third, integration with other sensor data would provide a more comprehensive understanding of customer experience. Combining eye-tracking with other physiological measures (e.g., heart rate variability, electrodermal activity) could reveal emotional responses alongside visual attention patterns, addressing the multidimensional nature of customer experience defined by Lemon and Verhoef (2016).

Fourth, development of more accessible technological approaches would enable wider adoption. Future research should explore whether similar analysis could be conducted using webcam-based eye-tracking or even smartphone-based solutions, reducing the hardware barriers to implementation.

Fifth, longitudinal studies could reveal how visual attention patterns evolve across multiple service interactions. Understanding how customers' visual attention changes as they become more familiar with a service environment could provide insights into learning processes and habit formation throughout the customer lifecycle.

Finally, comparative studies across different VLMs as the technology evolves will be valuable. The field of vision language models is rapidly advancing, and newer models may offer improved semantic accuracy with lower computational requirements, potentially making this approach more accessible and efficient for service providers.

7 Conclusion

In conclusion, this research demonstrates the potential of integrating eye-tracking data and VLM-based analysis to provide deeper insights into customer journeys. While technical and methodological challenges remain, this approach offers a promising path toward more objective, comprehensive understanding of customer experience across service touchpoints.

References

1. Halvorsrud, R., Mannhardt, F., Johnsen, E.B., Tapia Tarifa, S.L.: Smart journey mining for improved service quality. In: 2021 IEEE International Conference on Services Computing (SCC). pp. 367–369. IEEE (2021).
2. Halvorsrud, R., Mannhardt, F., Prillard, O., Boletsis, C.: Customer journeys and process mining – challenges and opportunities. ITM Web Conf. 62, 05002 (2024).
3. Følstad, A., Kvale, K.: Customer journeys: a systematic literature review. *Journal of Service Theory and Practice*. 28, 196–227 (2018).
4. Bernard, G., Andritsos, P.: CJM-ex: Goal-oriented exploration of customer journey maps using event logs and data analytics. *Business Process Management*. (2017).
5. Castner, N., Kübler, T., Scheiter, K., Richter, J., Eder, T., Hüttig, F., Keutel, C., Kasneci, E.: Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing. (2020).
6. Kobialka, P., Lizeth Tapia Tarifa, S., Bergersen, G.R., Johnsen, E.B.: User journey games: automating user-centric analysis. *Softw. Syst. Model*. 23, 605–624 (2024).
7. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking, <http://arxiv.org/abs/1504.06755>, (2015).
8. Lemon, K.N., Verhoef, P.C.: Understanding customer experience throughout the customer journey. *J. Mark.* 80, 69–96 (2016).
9. Berry, L.L., Wall, E.A., Carbone, L.P.: Service clues and customer assessment of the service experience: Lessons from marketing. *Acad. Manag. Perspect.* 20, 43–57 (2006).
10. Berry, L., Carbone, L., Haeckel, S.H.: Managing the total customer experience. *MIT Sloan Management Review*. 43, 85–89 (2002).
11. De Keyser, A., Verleye, K., Lemon, K.N., Keiningham, T.L., Klaus, P.: Moving the customer experience field forward: Introducing the touchpoints, context, qualities (TCQ) nomenclature. *J. Serv. Res.* 23, 433–455 (2020).
12. Bogale, B., Vesinurm, M., Lillrank, P., Celius, E.G., Halvorsrud, R.: Visual modeling languages in patient pathways: Scoping review. *Interact. J. Med. Res.* 13, e55865 (2024).
13. Hassani, M., Habets, S.: Predicting next touch point in A customer journey: A use case in telecommunication. *Eur Conf Model Simul*. 48–54 (2021).
14. Arias, M., Rojas, E., Aguirre, S., Cornejo, F., Munoz-Gama, J., Sepúlveda, M., Capurro, D.: Mapping the patient’s journey in healthcare through process mining. *Int. J. Environ. Res. Public Health*. 17, 6586 (2020).
15. Nguyen Chan, N., Nguyen Vo, D.L., Pham-Nguyen, C., Le Dinh, T., Dam, N.A.K., Pham Thi, T.T., Vu Thi, M.H.: Design and deployment of a customer journey management system: the CJMA approach. In: *The 5th International Conference on Future Networks & Distributed Systems*. ACM, New York, NY, USA (2021).
16. Davies, E.L., Bulto, L.N., Walsh, A., Pollock, D., Langton, V.M., Laing, R.E., Graham, A., Arnold-Chamney, M., Kelly, J.: Reporting and conducting patient journey mapping research in healthcare: A scoping review. *J. Adv. Nurs.* 79, 83–100 (2023).
17. Bernard, G., Andritsos, P.: Contextual and behavioral customer journey discovery using a genetic approach. In: *Advances in Databases and Information Systems*. pp. 251–266. Springer International Publishing, Cham (2019).
18. Huddleston, P., Behe, B.K., Minahan, S., Fernandez, R.T.: Seeking attention: an eye tracking study of in-store merchandise displays. *Int. J. Retail Distrib. Manag.* 43, 561–574 (2015).
19. Clement, J.: Visual influence on in-store buying decisions: an eye-track experiment on the visual influence of packaging design. *J. Mark. Manag.* 23, 917–928 (2007).

20. De Luzi, F., Leotta, F., Marrella, A., Mecella, M.: On the interplay between Business Process Management and Internet-of-Things: A Systematic Literature Review. *Bus. Inf. Syst. Eng.* 67, 265–288 (2025).
21. Koschmider, A., Janssen, D., Mannhardt, F.: Framework for process discovery from sensor data. *EMISA*. 32–38 (2020).
22. Vasudevan, A.B., Dai, D., Van Gool, L.: Object referring in videos with language and human gaze. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4129–4138. IEEE (2018).
23. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 2106–2113. IEEE (2009).
24. Papadopoulos, D.P., Clarke, A.D.F., Keller, F., Ferrari, V.: Training object class detectors from eye tracking data. In: *Computer Vision – ECCV 2014*. pp. 361–376. Springer International Publishing, Cham (2014).
25. Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., Chua, T.-S.: An eye fixation database for saliency detection in images. In: *Lecture Notes in Computer Science*. pp. 30–43. Springer Berlin Heidelberg, Berlin, Heidelberg (2010).
26. Berendes, C.I., Bartelheimer, C., Betzing, J.H., Beverungen, D.: Data-driven Customer Journey Mapping in Local High Streets: A Domain-specific Modeling Language. In: *Thirty Ninth International Conference on Information Systems*. aisel.aisnet.org (2018).
27. R Core Team, R., Others: R: A language and environment for statistical computing. (2013).
28. Rosenbaum, M.S., Massiah, C.: An expanded servicescape perspective. *J. Serv. Manag.* 22, 471–490 (2011).
29. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., de Weijer J., V.: *Eye tracking: A comprehensive guide to methods and measures*. (2011).
30. Janssen, D., Mannhardt, F., Koschmider, A., van Zelst, S.J.: Process model discovery from sensor event data. In: *Lecture Notes in Business Information Processing*. pp. 69–81. Springer International Publishing, Cham (2021).
31. Ooms, J.: av: Working with Audio and Video in R. (2025).
32. Ollama: Run large language models locally, <https://ollama.com/>, last accessed 2025/05/14.
33. Denner, S., Bujotzek, M.R., Bounias, D., Zimmerer, D., Stock, R., Maier-Hein, K.: Visual Prompt Engineering for Vision Language Models in Radiology. In: *Medical Imaging with Deep Learning* (2025).
34. Lebovitz, S., University of Virginia, Levina, N., Lifshitz-Assa, H., New York University, New York University: Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts’ know-what. *MIS Q.* 45, 1501–1526 (2021).
35. Steyvers, M., Griffiths, T.: Probabilistic topic models. *Handbook of latent semantic analysis*. (2007).
36. Montag, J.L., MacDonald, M.C.: Visual salience modulates structure choice in relative clause production. *Lang. Speech.* 57, 163–180 (2014).