

Article

Approach to Eye Tracking Scanpath Analysis with Multimodal Large Language Model

Xiangdong Li ^{1,*} , Kailin Yin ^{1,†} and Yuxin Gu ^{2,†}¹ College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; klyin@zju.edu.cn² College of Software Technology, Zhejiang University, Hangzhou 310027, China; 22251368@zju.edu.cn

* Correspondence: axli@zju.edu.cn

† These authors contributed equally to this work.

Abstract

Eye tracking scanpaths encode the temporal sequence and spatial distribution of eye movements, offering insights into visual attention and aesthetic perception. However, analysing scanpaths still requires substantial manual effort and specialised expertise, which limits scalability and constrains objectivity of eye tracking methods. This paper examines whether and how multimodal large language models (MLLMs) can provide objective, expert-level scanpath interpretations. We used GPT-4o as a case study to develop eye tracking scanpath analysis (ETSA) approach which integrates (1) structural information extraction to parse scanpath events, (2) knowledge base of visual-behaviour expertise, and (3) least-to-most and few-shot chain-of-thought prompt engineering to guide reasoning. We conducted two studies to evaluate the reliability and effectiveness of the approach, as well as an ablation analysis to quantify the contribution of the knowledge base and a cross-model evaluation to assess generalisability across different MLLMs. The results of repeated-measures experiment show high semantic similarity of 0.884, moderate feature-level agreement with expert scanpath interpretations ($F1 = 0.476$) and no significant differences from expert annotations based on the exact McNemar test ($p = 0.545$). Together with the ablation and cross-model findings, this study contributes a generalisable and reliable pipeline for MLLM-based scanpath interpretation, supporting efficient analysis of complex eye tracking data.



Academic Editor: Wei Gao

Received: 15 October 2025

Revised: 1 December 2025

Accepted: 3 December 2025

Published: 10 December 2025

Citation: Li, X.; Yin, K.; Gu, Y.Approach to Eye Tracking Scanpath Analysis with Multimodal Large Language Model. *Modelling* **2025**, *6*, 164. <https://doi.org/10.3390/modelling6040164>**Copyright:** © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license

(<https://creativecommons.org/licenses/by/4.0/>).**Keywords:** scanpath analysis; multimodal large language model; eye tracking; visual attention behaviour

1. Introduction

An eye tracking scanpath is the recorded sequence of fixations and saccades that occur as a person views a visual stimulus such as an image, interface, or product. Fixations mark points where the gaze lingers, reflecting attentional focus, while saccades represent rapid eye movements between these points. Together, they form a temporal-spatial trace of visual exploration that reveals not only where and for how long attention is allocated, but also how it shifts over time [1]. As such, scanpaths provide a window into visual attention, search strategies, and cognitive processes, making their analysis an important method in HCI, visual cognition, and design research.

Building on this value, scanpath analysis has been widely adopted across diverse domains. In interface evaluation and web design, it illuminates how users navigate layouts and interact with content [2,3]. In industrial design, scanpaths reveal how consumers

explore form, colour, and functional features, and these offer insights that support design decisions to improve usability and aesthetic appeal [4–6]. However, interpreting scanpaths remains labour-intensive, expert-dependent, and difficult to scale [7]. Analysts must manually trace gaze sequences, identify meaningful patterns, and map them to design elements. This process is slow, subjective, and often irreproducible, limiting the integration of eye tracking methods into iterative design workflows and broader adoption in HCI research.

Prior efforts in scanpath analysis predominantly rely on machine learning or deep learning methods that compute spatial or sequential similarity (e.g., ScanMatch [8], Multi-Match [9], sequence-edit metrics [10]). While these techniques effectively quantify structural correspondence, they are not designed to explain the cognitive or visual-behavioural meaning of those patterns. They lack the ability to integrate domain knowledge and thus cannot replace expert reasoning. This creates a critical research gap: how to achieve both computational efficiency and expert-level interpretive depth in scanpath analysis.

Recent advances in multimodal large language models (MLLMs) present a promising opportunity. Vision-centric MLLMs, such like BLIP [11], LLaVA [12], and GPT-4V [13], can perform fine-grained visual description, identify spatial relationships, and leverage sequential and semantic cues [14]. However, these capabilities remain too generic to capture the specialised temporal-spatial structures of scanpaths or to translate these patterns into actionable design insights.

To fill this gap, we explore whether a carefully designed MLLM-based framework can support reliable, interpretable, and scalable scanpath analysis. We use GPT-4o as a case study because it is a representative model with strong multimodal reasoning ability and support for stepwise structured explanations, which are essential for reconstructing visual attention patterns. Accordingly, we formulate the following research questions:

- **RQ1:** How to build a MLLM-based framework to achieve coherent and design-relevant scanpaths interpretations?
- **RQ2:** How does the proposed approach generate consistent interpretations across repeated analyses of the same stimuli?
- **RQ3:** How closely do interpretations generated by the proposed approach align with expert designers' scanpath analyses?

To answer these questions, we introduce ETSA, eye tracking scanpath analysis, an approach for adapting MLLMs to scanpath interpretation in the context of industrial product evaluation. Using GPT-4o as a case study, the approach integrates three components: (i) a structural information extraction module that decomposes scanpaths into fine-grained events; (ii) a knowledge base encoding visual-behaviour expertise; and (iii) prompt engineering strategies employing least-to-most [15] and few-shot chain-of-thought reasoning [16] to scaffold the model's outputs. Furthermore, we evaluate the approach through two complementary experiments: a repeated-measures experiment assessed the reliability of the approach's outputs, and a user study compared the approach's visual-feature mappings with expert designer annotations. In addition, we conducted an ablation analysis to quantify the contribution of the knowledge base and a cross-model evaluation to assess generalisability across different MLLMs.

This paper makes three contributions:

- (1) ETSA, a knowledge-grounded approach for scanpath interpretation with MLLMs. We introduce, to our knowledge, the first approach that systematically adapts MLLMs to the specialised tasks of interpreting eye tracking scanpaths. By combining structural parsing of fixation-saccade sequences, a knowledge base of visual-behaviour expertise, and prompt engineering procedures, the approach turns general-purpose vision-language models into task-specific analytic tools while remaining model-agnostic.

- (2) A methodological integration of structural and semantic reasoning. Unlike prior computational approaches that rely on simplified metrics or pattern-matching, our approach decomposes scanpaths into fine-grained events and embeds them in expert-informed prompts, which enable sequence-aware, design-relevant reasoning. The resulting pipeline produces interpretable, auditable explanations and offers a transferable strategy for other domain-specific visual reasoning tasks in HCI.
- (3) Empirical evidence of reliability and expert alignment. Across a repeated-measures study and a comparative user study, the ETSA yields high within-approach consistency (0.884), moderate feature-level agreement with expert scanpath interpretations ($F1 = 0.476$) and no significant differences from expert annotations based on the exact McNemar test ($p = 0.545$). Together with the ablation and cross-model findings, they suggest that the method can provide reasonably dependable support for scanpath interpretation, contributing to more scalable and consistent analysis practices in eye-tracking research and industrial design workflows.

Table 1 summarise these three contributions.

Table 1. Overview of contributions.

Contributions	Focus	Summary
ETSA	Framework design	Proposes a knowledge-grounded approach for scanpath interpretation with MLLMs.
Structured Reasoning	Methodological insight	Combines structural and semantic reasoning to support scanpath analysis.
Empirical assessment	Evaluation	Demonstrates reliability and expert alignment through experiments.

The remainder of the paper is structured as follows: Section 2 reviews related work on eye tracking, usability evaluation, and scanpath analysis. Section 3 details the proposed ETSA methodology. Section 4 reports the evaluation studies and results. Section 5 discusses the findings and their implications, and Section 6 concludes the paper.

2. Related Work

2.1. Eye Tracking Scanpath

A eye tracking scanpath (Figure 1) is the sequential trajectory of eye movements generated when a person views a stimulus, typically represented as a sequence of fixations and saccades [17]. Fixations, defined as clusters of gaze points that remain within a small region, indicate where visual attention is maintained, while saccades are rapid movements that shift gaze between fixations [1]. The spatial distribution of fixations reflects how broadly an observer samples visual information, whereas their temporal organisation reveals patterns of exploration and cognitive processing [18]. Unlike heatmaps, which capture only the aggregate duration and location of gaze, scanpaths encode the order of gaze shifts, thereby revealing revisit patterns, attentional priorities, and search strategies [19,20]. This multidimensional representation provides richer evidence for understanding how users navigate and process complex visual environments.

Scanpath analysis is widely used across HCI and applied research fields such as web usability, information visualisation, and education [2,3]. In industrial design, scanpath analysis help designers understand how users visually explore prototypes or products, highlighting attentional priorities, information-seeking patterns, and potential usability bottlenecks [21,22]. For example, fixation sequences can show whether critical affordances

attract timely attention [23], while temporal-spatial patterns highlight overlooked features or distracting elements that affect clarity [24]. Such evidence supports designers in diagnosing perceptual load, optimising visual hierarchy, and testing aesthetic or functional hypotheses. Thus, incorporating scanpath analysis into industrial design workflows offers behaviour-based insight that complements subjective feedback, improves decision making, and helps reduce costly redesign cycles [25].

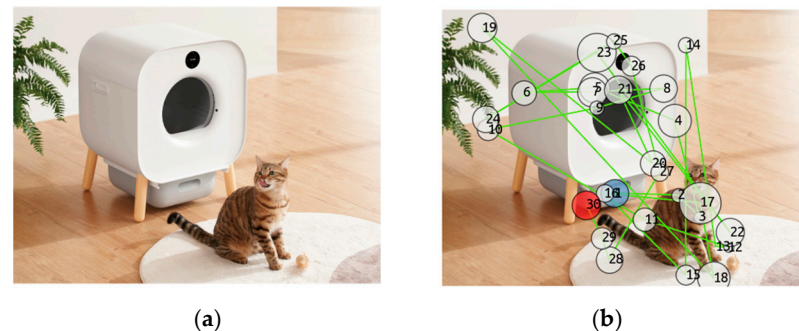


Figure 1. Example of (a) an original image and (b) the corresponding eye tracking scanpath, with numbered circles indicate fixation order.

2.2. Approach for Interpreting Scanpath

Existing approaches to interpreting eye tracking scanpaths can be broadly categorised into expert-driven subjective interpretation and metrics-based quantitative analysis [26]. Expert-driven approaches rely on domain specialists to interpret scanpaths using their knowledge and experience. For example, in multimedia reading research, experts identify characteristic patterns such as linear reading, figure referencing, and cross-modal integration to infer distinct attentional behaviours [27]. While such analyses can capture subtle visual strategies, they lack standardised criteria, leading to variability both across experts and within the same expert over time [28]. This inconsistency, combined with the high demands of time and labour, limits scalability and risks introducing cognitive bias.

Quantitative approaches instead divide scanpaths into measurable attributes in temporal, spatial, and directional dimensions. Common metrics include such like total duration and path length (processing fluency and efficiency), spatial density and convex hull area (distribution and coverage of fixations), transition matrices and path regularity (structural gaze patterns), and directional measures (strategic preferences such as reading order) [3,29,30]. These methods enable efficient, reproducible evaluation and rapid feature extraction but often overlook the sequential dependencies among fixations that are critical to capturing the dynamic flow of attention. Moreover, computational algorithms such as ScanMatch, MultiMatch, and string-edit-distance methods align fixation sequences across dimensions including shape, direction, and position to quantify similarity between scanpaths. Although they provide a structured basis for comparing scanpaths, they still operate at the level of sequence matching rather than interpretation, and thus cannot explain why two sequences differ, what visual features drive the observed behaviour, or how attentional transitions relate to design elements.

Despite their differences, both categories of approaches face a fundamental tension between depth of interpretation and efficiency of analysis. Sequence similarity metrics, for example, can identify structural resemblance between two scanpaths but cannot determine whether such resemblance reflects comparable cognitive processes [31]. Similarly, the definition of areas of interest (AOIs) still requires subjective expert judgement tailored to task context and domain knowledge [32]. As a result, current analytic pipelines often combine structural feature extraction with expert interpretation, remaining heavily dependent on specialised expertise and offering limited scalability.

These limitations highlight the need for approaches that combine the rigour of expert-level interpretation with the efficiency and reproducibility of automated analysis. Recent advances in MLLMs offer a promising opportunity. They can integrate spatial, temporal, and semantic information in ways that conventional methods cannot, and it opens a pathway towards more scalable, transparent, and expert-aligned scanpath interpretation.

2.3. MLLMs for Image Interpretation

Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in visual description and semantic reasoning, enabling them to identify objects, spatial relations, and high-level patterns across diverse images [33]. These strengths have led to applications in areas such as automated text evaluation (e.g., G-EVAL, GPTSCORE) and UI analysis, where MLLMs help detect layout issues, suggest refinements, and assess consistency and accessibility [34–36]. These advances underscore the growing potential of MLLMs for tasks requiring integrated visual and semantic understanding.

However, despite their broad utility, the multimodal architectures underlying current MLLMs face critical challenges that are directly relevant to scanpath analysis. Most MLLMs use visual encoders that produce spatial feature tokens for a single static image and fuse them with text features through cross-modal attention. Because the visual input is treated as an atemporal representation, these architectures are not suited for modelling temporal gaze sequences such as fixation–saccade transitions. Consequently, MLLMs struggle with fine-grained visual parsing, which leads to missed or misinterpreted details in complex gaze sequences [37].

Furthermore, because these architectures are trained on general-purpose image–text datasets, their reasoning is not grounded in domain-specific knowledge. Without embedded constructs from visual attention theory or design principles, their outputs frequently remain shallow and overlook key factors such as visual hierarchy, perceptual grouping, and design-related cues [38]. In addition, existing multimodal fusion pipelines have limited support for multi-stage or structured reasoning, making it difficult for MLLMs to integrate spatial, temporal, and semantic information into coherent interpretations of gaze behaviour [39]. These limitations reveal a clear gap between the general-purpose capabilities of MLLMs and the specialised demands of rigorous scanpath interpretation, underscoring the need for methods that combine structured event-level parsing with domain-informed reasoning.

2.4. Lesson Learned

The reviewed literature highlights three lessons. First, despite the widespread use of visual scanpaths in industrial design and HCI, their interpretation still relies heavily on expert judgement, limiting scalability and reproducibility. Second, neither automated metrics nor expert-only approaches are sufficient for effective scanpath interpretation. Third, while MLLMs promise richer semantic reasoning, their limitations in fine-grained parsing, spatio-temporal understanding, and domain grounding constrain their reliability.

Together, these insights point to the need for a hybrid approach that integrates structured scanpath parsing with knowledge-guided reasoning. These insights form the conceptual foundation for our ETSA framework described in Section 3.

3. Development of ETSA Approach

To answer **RQ1**, we used GPT-4o as a case study and proposed ETSA, a knowledge-grounded approach structured around three complementary components (Figure 2). This approach is design to address the limitations of existing scanpath analysis methods and

adapt MLLMs to this specialised task, and the design of this approach is guided by three rationales derived from the literature.

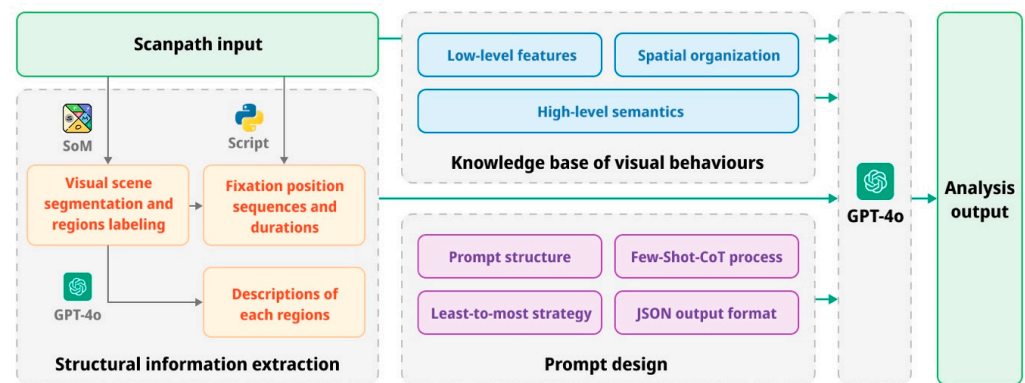


Figure 2. Overview of ETSA framework.

First, accurate scanpath interpretation requires models to handle both spatial and temporal dependencies, which are often lost in conventional metrics. To meet this need, the structural information extraction module decomposes scanpaths into fine-grained events, so to enable precise parsing of fixation sequences and transitions.

Second, generic MLLMs lack the domain-specific grounding that is necessary for reliable reasoning about visual behaviour. To overcome this, the knowledge base module encodes principles of visual attention and design expertise, which allows the ETSA approach to anchor its interpretations in established theoretical constructs rather than surface-level patterns.

Third, effective scanpath interpretation demands multi-stage, structured reasoning rather than superficial, one-shot outputs. The prompt design module therefore employs advanced prompt engineering strategies, such as least-to-most prompting and few-shot chain-of-thought (CoT), to support the approach's reasoning process. This ensures outputs that are both coherent and aligned with expert analytical practices.

Together, these components form a pipeline that combines the scalability of automation with the interpretive depth of expert reasoning, thereby effectively addressing the key challenges of objectivity, reproducibility, and efficiency in scanpath analysis.

3.1. Structural Information Extraction

Effective scanpath interpretation requires the approach to receive inputs that preserve both the spatial context of visual scenes and the temporal dynamics of gaze behaviour. To this end, we developed a structural information extraction pipeline that transforms raw scanpath images into semantically rich, machine-readable inputs.

First, the original stimulus image is segmented into distinct regions using the Set-of-Mark (SoM) model [40]. We experimented with multiple levels of segmentation granularity and ultimately adopted a granularity level of 1.8, which provides an effective balance between fine-grained region density and semantic interpretability. As shown in Figure 3a, each segmented region is automatically assigned a label to support subsequent contextual mapping. All SoM segmentation results were manually inspected by researchers, and images that exhibited noticeable errors (e.g., incorrect splitting of objects or overlapping region labels) were removed from downstream analysis to ensure clean structural input. Subsequently, GPT-4o is employed to generate detailed descriptions of each visual object region, capturing attributes such as colour, size, location, texture, and functionality. These descriptions supply the approach with regional semantic information that would otherwise remain implicit in the raw image.

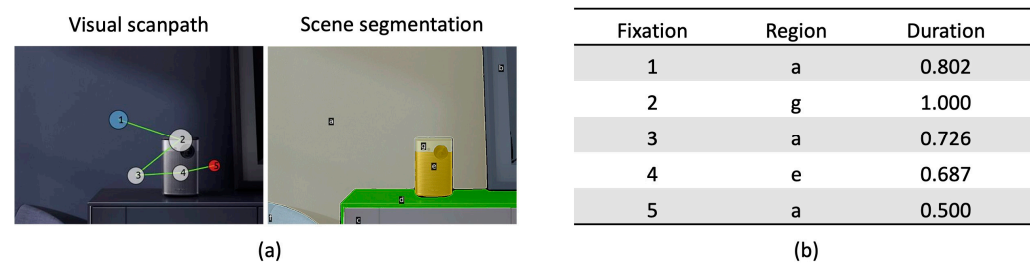


Figure 3. Examples of structural information extraction, including (a) example scene segmentation with labelled regions, where numbered circles indicate fixation order, and (b) fixation mapped region and durations.

To incorporate the temporal dimension of gaze behaviour, a Python script (Python 3.10) extracts fixation sequences and durations from the scanpath overlay. As shown in Figure 3b, fixation coordinates are mapped to their corresponding object regions, which produce a temporally ordered sequence of region-level gaze events. Fixation duration is calculated from the radius of the fixation circles depicted in the scanpath image, with a linear scale in which radius corresponds to duration normalized to a 0–1 interval. This method ensures that both attentional focus and transition dynamics are represented. Coordinates of each fixation point were then mapped to their corresponding SoM-defined visual regions, producing an ordered sequence of region-level gaze events and associated normalized fixation durations.

The final structured output includes: (i) visual scene segmentation with labelled regions; (ii) semantic descriptions of each region; (iii) fixation sequences with mapped object references and durations; and (iv) the original stimulus and scanpath images for reference. Together, these elements provide the approach with an enriched representation that captures both regional context and gaze dynamics, which enable more accurate and interpretable scanpath analysis in subsequent stages.

3.2. Knowledge Base Construction

To strengthen the explanatory capacity of the MLLMs and ensure its outputs reflect established principles of visual attention, we constructed a knowledge base that embeds well-documented domain expertise into the scanpath interpretation process. This resource synthesises theoretical and empirical findings from visual attention research, cognitive psychology, and the eye-tracking literature, with particular focus on studies examining task-free image viewing [41–44]. By grounding the approach in naturalistic patterns of visual exploration, the knowledge base ensures that its interpretations remain theoretically consistent and practically relevant.

The knowledge base is organised as a three-level visual hierarchy that captures the progression of attentional processing from low-level perceptual features, through intermediate principles of spatial organisation, to high-level semantic understanding. Each level is linked to characteristic gaze behaviours observed in scanpath research, so to create a structured foundation for reasoning about why particular regions attract or sustain attention. This hierarchical organisation allows the approach to connect visual content to the temporal-spatial patterns of gaze observed in scanpaths, thus providing interpretability that goes beyond surface metrics.

- Low-level features capture early perceptual cues (e.g., brightness, colour, size, contrast, texture) that drive bottom-up saliency and initial fixations [41,42].
- Spatial organisation encodes mid-level compositional principles (e.g., repetition, layout, proximity, alignment, balance, focal points), which shape how viewers traverse a visual scene [41,42].

- High-level semantics reflect top-down attentional drivers (e.g., faces, symbols, tools, finger direction, conceptual relationships) that anchor attention to meaningful objects and guide search strategies [43,44].

Table 2 summarises the hierarchical structure, its attributes, and their associated behavioural implications. By embedding these features, the knowledge base equips the approach with a structured, theory-driven lens for interpreting scanpaths, reducing reliance on arbitrary heuristics and aligning outputs with expert-level reasoning.

Table 2. Summary of hierarchical structure and related visual behaviours.

Hierarchy	Attributes	Descriptions
Low-level features	Brightness	Lighting can be quickly distinguished from surface brightness at early visual processing levels [41].
	Detail	Items with a high level of detail possess higher saliency. Observers prefer objects and areas with richer details [41].
	Size	Larger text tends to override smaller text, making it a potential successful entry point for attention [42].
	Colour	Elements with high saturation, contrast, and brightness are more likely to attract attention. Extreme and contrasting colours are given higher visual priority.
	Contrast	Clear and distinct contrast is crucial; otherwise, elements may appear similar and create compositional conflict.
	Form	Visually unusual or unexpected forms tend to have higher visual priority.
	Texture	Components with deep textures in a design can attract significant attention [42].
Spatial organisation	Repetition	When elements are repeated across different parts of a design, the viewer's eyes tend to follow them [42].
	Layout	Combining multiple layout patterns guides users to browse page content in a specific sequence [42].
	Focal point	The point where a pair of extended lines intersect automatically attracts attention [41].
	Implied network	Ensuring elements are aligned, negative space is organized, elements are properly grouped, and the overall composition is balanced helps unify the poster and facilitates visual guidance.
	Centre of gravity	Objects placed near the centre of gravity of a display attract more attention [41].
	Alignment	When objects are aligned, an invisible line connects them, enhancing visual cohesion.
	Negative space	Negative space enhances surrounding elements, providing visual breathing room and allowing the eyes to rest.
	Proximity	The principle of proximity groups related items together, making them appear cohesive.
	Balance	Balance maintains a comfortable and stable composition within a design.
	Spatial overlay	Spatial overlay involves identifying physical overlapping relationships between elements.

Table 2. *Cont.*

Hierarchy	Attributes	Descriptions
High-level semantics	Interesting objects	Humans often allocate their gaze to interesting objects within a scene, with most fixations occurring near the centre of such objects [43].
	Faces	A face attracts attention more strongly than other objects.
	Words and symbols	Attention is drawn to personal names, emotionally charged words, or symbolic elements [42].
	Tools and displays	Elements designed to attract attention or facilitate human interaction (e.g., viewability, operability) serve as effective attentional cues.
	Finger direction	Images of fingers guide the viewer’s attention toward the pointed object or area [44].
	Conceptual relationships	Elements within a poster can be unified through an overarching idea that conceptually connects them.

The knowledge base is incorporated into the prompt as a structured, hierarchical reference, where each level is presented as an explicit hierarchy followed by a list of attributes and corresponding behavioural descriptions. This structure is preserved so that the model receives the knowledge base in a machine-readable, logically ordered form. Notably, the descriptions are more detailed than that in Table 2 to ensure the model receives complete and structured guidance during reasoning. An example of this representation is shown in Figure 4. Additionally, no dynamic feature selection is performed. The entire knowledge base is supplied in each run to maintain stable and reproducible reasoning across images.

3.3. Prompt Engineering Design

A key challenge in adapting MLLMs to scanpath interpretation lies in ensuring that the approach follows a structured and transparent reasoning process rather than producing arbitrary or hallucinated outputs. To address this, we developed a prompt engineering framework that explicitly decomposes the task into stages, constrains the approach’s role and inputs, and enforces consistency in its outputs.

As shown in Figure 4, the framework consists of seven components: role specification, task instruction, input description, workflow, rules, knowledge base, and output format. The role is set as that of “an expert in visual attention analysis.” The task instruction directs the approach to “analyse the visual scanpath of the given image to evaluate the effectiveness of visual guidance based on the following knowledge, rules and workflow.” The input description includes the original stimulus, the scanpath overlay, the segmentation map, fixation sequences with durations, and detailed descriptions of visual objects. By pairing visual and textual representations, the input design ensures multimodal alignment and provides the approach with both semantic and structural context.

The workflow part employs a least-to-most prompting strategy [15], which breaks the overall analysis into four subtasks: (1) scanpath recapitulation and fixation duration aggregation; (2) interpretation guided by the visual-attention hierarchy in the knowledge base; (3) holistic synthesis of attentional patterns; and (4) multi-dimensional scoring of visual guidance effectiveness. This stepwise structure enables the approach’s reasoning and mirrors the process typically followed by human experts. To further enhance domain fidelity, the knowledge base of visual-behaviour expertise is injected as contextual information during analysis.



Figure 4. Framework of prompt engineering design. Symbols ## denotes section headers used in the prompt, and ** indicates emphasis markers included in the prompt template.

The rules require the approach to adopt a step-by-step Chain-of-Thought (CoT) reasoning model [16], which ensures that intermediate reasoning steps are explicit and logically coherent. Finally, outputs are constrained to a standardised JSON template with predefined fields and automated field-population mechanisms, thus enhancing reproducibility and enabling systematic comparison across analyses.

By combining role specification, multimodal input structuring, stepwise task decomposition, and explicit reasoning rules, this prompt engineering framework provides a principled mechanism for aligning MLLM outputs with expert analytical practices. It ensures not only more accurate interpretations of scanpaths but also greater transparency and consistency than ad hoc prompting approaches.

3.4. Implementation Details

GPT-4o was accessed through the official API using the vision-enabled model endpoint. All analyses were conducted with a fixed prompt template to ensure consistency across images. The temperature parameter of GPT-4o was set to 0, ensuring deterministic decoding in which the model always selects the highest-probability token. All analyses were performed with identical model parameters and prompt structure. All calls were made using Python 3.10 with the OpenAI client library, and raw outputs were stored without post-generation modification other than formatting for analysis. This setup ensures a controlled and reproducible runtime environment for evaluating the model's interpretive behaviour.

The overall ETSA workflow is summarised in Figure 5, which illustrates how the modules of our framework interact.

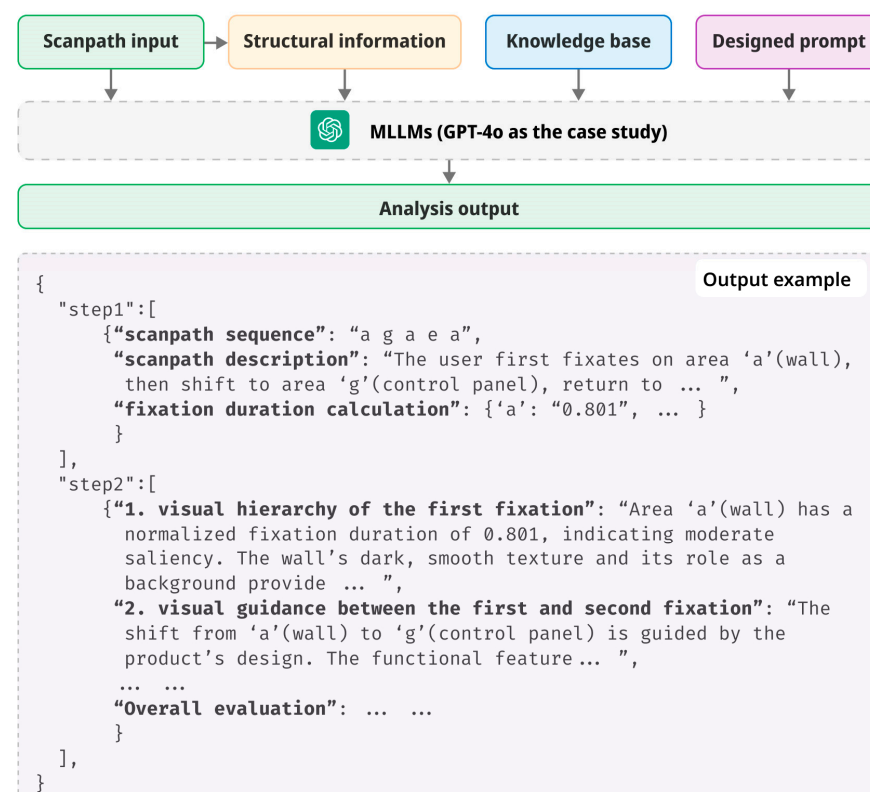


Figure 5. Schematic overview of the ETSA workflow and an example of output.

4. Evaluation of Scanpath Analysis Approach

To answer RQ2 and RQ3, we designed an evaluation strategy targeting two core requirements for scanpath analysis: reliability and effectiveness, to assess the validity of our ETSA approach. Reliability refers to the consistency of outputs across repeated runs of the model—a critical property for reproducibility and trust in automated pipelines. Effectiveness captures the extent to which the approach outputs align with expert interpretations, so to confirm that automated analysis retains the depth and validity of professional judgement. Together, these two dimensions address the key limitations of prior approaches, which have struggled to combine objectivity with expert-level interpretive fidelity.

For empirical testing, we constructed a dataset of exemplary industrial design works. Specifically, 140 industrial product posters from global design Awards (past five years) were selected. These works were chosen because they exemplify professional, high-quality design practice, and offer clear communicative intent, diverse visual compositions, and strong thematic expression. The selection criteria included high image clarity, balanced composition, and identifiable design elements. These ensure that the dataset provides

challenging yet representative cases for scanpath interpretation. As shown in Figure 6, the collected dataset spans multiple industrial design presentation types, including full product views, usage scenarios, detail shots, interaction depictions, and multi-angle representations, thereby it supports a robust evaluation across a broad spectrum of design contexts.



Figure 6. Examples of collected dataset across five industrial design presentation types.

The scanpaths for these images were generated using the well-established IOR-ROI LSTM model [20], which simulates human-like gaze behaviour and generates visual scanpath. Each generated scanpath includes fixation positions and fixation-circle radii, which we extract from the visualization for structural parsing and relative temporal weighting. Using the IOR-ROI LSTM model ensures a valid and controlled input source for our scanpath interpretation framework.

4.1. Reliability Evaluation

To evaluate the reliability of the ETSA approach outputs, we conducted a reliability study using 50 representative images drawn from the dataset. Each associated scanpath image was submitted to the model 20 independent times, yielding 20 textual interpretations per image. This repeated-sampling design ensures that the evaluation captures both within-image variability and consistency across multiple categories of design stimuli.

For analysis, we measured the semantic similarity of outputs using sentence embeddings derived from the Bidirectional Encoder Representations from Transformers (BERT) model [45]. Sentence BERT embeddings provide a high-dimensional representation of linguistic meaning, making them well-suited for detecting underlying semantic consistency despite superficial lexical differences.

For each image, we computed pairwise cosine similarity across the 20 outputs (190 pairs in total) and averaged these values to produce an image-level similarity score. Next, category-level reliability were derived by calculating the mean and standard deviation of these image-level scores within each category. The overall reliability score was obtained by averaging the 50 image-level scores across the full dataset.

The results indicate a high degree of semantic stability. Across all 50 images, the overall mean similarity was 0.884 (SD = 0.025, 95% CI [0.869, 0.901]), confirming that repeated approach outputs preserved the same underlying content. Category-level analysis showed comparably high values: overall product views (M = 0.874, SD = 0.012, 95% CI [0.867, 0.882]), usage scenarios (M = 0.876, SD = 0.025, 95% CI [0.860, 0.892]), detail shots (M = 0.881, SD = 0.027, 95% CI [0.864, 0.898]), interaction depictions (M = 0.908, SD = 0.022, 95% CI [0.895, 0.922]), and multi-angle views (M = 0.882, SD = 0.025, 95% CI [0.866, 0.898]). Notably, interaction depictions achieved the highest stability, which suggest that the approach is particularly robust when processing gaze data involving human–object interactions.

Taken together, these findings demonstrate that the approach consistently generates semantically coherent interpretations, even when prompted multiple times with the same scanpath input. Minor lexical variation was observed across runs, but the preservation of meaning and logical structure confirms the reliability and reproducibility of the approach under repeated conditions (Figure 7).

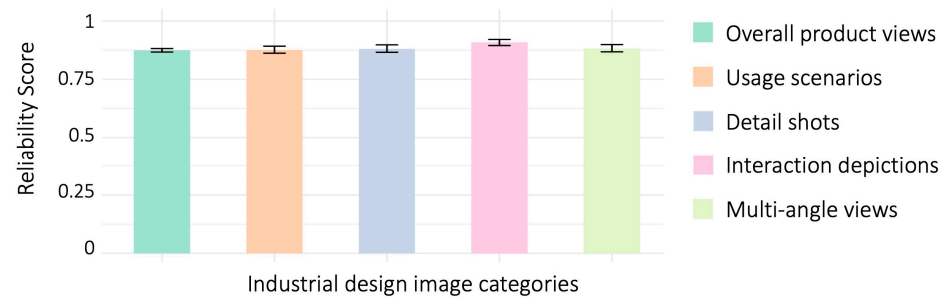


Figure 7. Comparing semantic similarity score across five image categories. Error bar represents the 95% confidence intervals.

4.2. Effectiveness Evaluation

To evaluate the effectiveness of the ETSA approach, we conducted a user study comparing approach-generated scanpath analyses with assessments made by professional designers. This study addressed the key question of whether the approach's interpretations align with expert reasoning in identifying the visual features underlying fixations and saccades.

4.2.1. Experiment Design and Implements

The study was implemented on a web-based platform (Figure 8) and administered via online sessions. Each participant completed two tasks per image: (1) annotate at least three visual features that contributed to the observed fixations and saccades, and (2) rate the accuracy of the corresponding model-generated analysis on a 0–3 scale.

Original image

Scene segmentation

Scanpath

Choose file

3.7

Extracted visual scanpath sequence

e e b e b

Visual features

size, colour, brightness, detail, shape, white space, repetition, layout, position, and semantic information.

Please select at least three visual features that have the most significant impact on the area corresponding to each fixation point or visual scanpath.

fixation points	corresponding region	size	colour	brightness	detail	shape	white space	repetition	layout	position	semantic info
1	e										
1-2	e-e										
2	e										
2-3	e-b										
3	b										
3-4	b-e										
4	e										
4-5	e-b										
5	e										

Please rate the accuracy of model generation results.

3 - Very accurate

JavaScript

```
{
  "1.Visual hierarchy of the first fixation point": "Area e has a high fixation duration of 0.9347, indicating its importance. The circular shape and central position of area e, combined with its dark frame and contrast against the light background, make it visually dominant.",
  "1-2.Visual guidance between the first fixation point and the second fixation point": "The repeated fixation on area e suggests its strong visual saliency due to its central position, functional significance, and high contrast. The circular morphology and alignment with th
```

Copy Caption

Figure 8. Example interface of the experimental platform for user studies.

To ensure clarity and consistency in feature annotations, the experimenters refined the initial set of 22 visual features identified in the literature review. Redundant or overlapping descriptors were removed, and semantically similar terms (e.g., alignment, proximity, overlap) were merged. This process yielded 10 distinct features: size, colour, brightness, detail, shape, white space, repetition, layout, position, and semantic information.

Six design experts ($N = 6$, 3 female, 3 males; $M_{age} = 23.8$, $SD_{age} = 1.8$) were recruited through social media. All participants had substantial design expertise, each having received major international design awards (e.g., iF, Red Dot, K-Design). Although the number of experts is relatively small because of high standard, this sample size is consistent with common practice in design research and expert-judgment studies [46], where evaluations often rely on small but highly specialized expert groups. Before the experiment, participants were informed about the research purpose, provided written consent, and completed a brief demographic questionnaire. Next, standardized task instructions were given: participants were first asked to examine the scanpath visualization and identify the visual features contributing to each fixation and saccade by selecting at least three relevant features from a predefined list (annotation task), and then to read the corresponding model-generated interpretation and rate its accuracy on a 0–3 scale (evaluation task). During the formal experiment, each participants completed five independent trials, with each trial involving one randomly selected image from a different design category. The whole procedure took approximately one hour per participant, yielding expert annotations and ratings for a total of 30 images.

4.2.2. Data Analysis

To quantitatively assess the alignment between ETSA approach-generated analyses and expert annotations, we developed a semantic mapping and evaluation pipeline that converts textual outputs into a unified feature-based representation suitable for comparison. This pipeline ensures that conceptual correspondence—a critical requirement for analysing descriptive visual reasoning, rather than surface lexical overlap, is evaluated.

Text preprocessing and tokenisation. Approach outputs were tokenised and filtered using an expanded English stop-word list, which is customised to account for domain-specific terminology. Multi-word expressions (e.g., spatial focus) were treated as fixed units to preserve semantic coherence, while compound phrases (e.g., an object designed to be seen or touched) were encoded using sentence-level embeddings. This process improved linguistic precision and avoided token fragmentation that could distort semantic meaning.

Semantic embedding and feature vector construction. Each token was represented as a weighted combination of its contextualised embedding (capturing sentence-level semantics) and its standalone word vector (capturing independent meaning). A 2:8 weighting ratio was adopted to prioritise stable word-level representation while maintaining contextual awareness—an empirically balanced configuration found to enhance semantic matching in pilot tests. For each target visual feature (e.g., colour, balance), a composite feature vector was constructed by averaging its base vector with three semantically related terms, thereby capturing lexical variation and synonymy.

Feature matching using cosine similarity. Cosine similarity was calculated between each token vector and the target feature vectors. Tokens with similarity scores ≥ 0.8 were considered semantically equivalent to the feature, based on prior findings that this threshold reliably indicates conceptual correspondence in embedding space. This process yielded a binary presence/absence vector for each feature, enabling direct comparison between model and expert results.

Evaluation metrics. Model performance was evaluated using standard multi-label classification metrics:

- Hamming Loss (HL) measures the proportion of incorrectly predicted labels to the total number of labels:

$$HL = \frac{1}{NL} \sum_{i=1}^N \sum_{j=1}^L (y_{ij} \oplus \hat{y}_{ij}) \quad (1)$$

where N is the total number of samples, L is the total number of labels, y_{ij} and $\hat{y}_{ij}$ denote the true and predicted values of the j^{th} label for the i^{th} sample, respectively, and $\oplus$ represents the XOR operation.

- Precision (P) is defined as the ratio of correctly predicted positive labels to the total number of predicted positive labels:

$$P = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|} \quad (2)$$

where Y_i denotes the set of true labels for the i^{th} sample, and $\hat{Y}_i$ represents the set of predicted labels for the i^{th} sample.

- Recall (R) is the ratio of correctly predicted positive labels to the total number of true positive labels:

$$R = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \cap \hat{Y}_i|}{|Y_i|} \quad (3)$$

- F1-score (F1) is the harmonic mean of Precision and Recall, providing a balanced assessment of the model's overall performance:

$$F1 = \frac{2}{N} \sum_{i=1}^N \frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i| + |Y_i|} \quad (4)$$

These metrics collectively provide a rigorous, interpretable assessment of the approach's alignment with expert reasoning, which capture both its ability to identify the correct visual features (precision) and its completeness in recovering all relevant features (recall).

4.2.3. Results

Overall performance. Figure 9 summarises the effectiveness of ETSA approach with different models (GPT-4o with and without knowledge base and Claude Sonnet 4) across four multi-label metrics (HL, P, R, and F1). The results of GPT-4o with knowledge base showed that the average Hamming Loss was 0.364 (SD = 0.080, 95% CI [0.335, 0.414]), which indicate a controlled error rate given the multi-feature, multi-label setting. Precision (M = 0.488, SD = 0.102, 95% CI [0.424, 0.552]) and Recall (M = 0.491, SD = 0.104, 95% CI [0.426, 0.556]) were closely balanced, suggesting the approach neither over-predicts spurious features nor systematically misses expert-identified ones. Their harmonic mean, the F1-score, averaged 0.476 (SD = 0.092, 95% CI [0.417, 0.533]), reflecting moderate, balanced alignment with expert annotations. In practical terms, the ETSA approach recovers most of the visual features highlighted by experts while avoiding excessive false positives—an important trade-off for design analysis where both over- and under-prediction can undermine interpretability.

Ablation study. To examine the contribution of the knowledge base, we compared the performance of the same model (GPT-4o) with and without the proposed knowledge base. As illustrated in Figure 9, removing the knowledge base substantially decreased performance across the metrics, with notable reductions in Precision (M = 0.395, SD = 0.147, 95% CI [0.304, 0.486]), Recall (M = 0.308, SD = 0.147, 95% CI [0.217, 0.398]) and F1-score (M = 0.320, SD = 0.129, 95% CI [0.241, 0.401]). These results demonstrate that the knowledge base provides clear incremental value, strengthening the model's ability to retrieve

expert-relevant visual features and domain cues. The ablation confirms that the observed improvements stem from the integration of structured domain knowledge rather than stochastic variation or prompt-level artefacts.

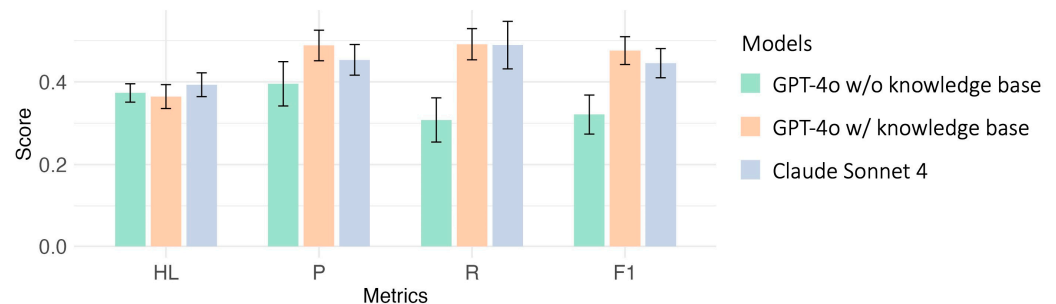


Figure 9. Performance comparison of GPT-4o (with and without knowledge base) and Claude Sonnet 4 across four metrics: Hamming Loss (HL), Precision (P), Recall (R), and F1-score (F1). Error bars represent 95% confidence intervals.

Cross-model generalisability. To evaluate whether the proposed ETSA approach depends on a specific MLLM, we further compared GPT-4o with Claude Sonnet 4 (<https://www.anthropic.com/news/claude-4>, accessed on 26 November 2025) under identical prompt structures and input formats. As illustrated in Figure 9, both models exhibit similar performance on the four metrics: Hamming Loss ($M = 0.393$, $SD = 0.079$, 95% CI [0.364, 0.442]), Precision ($M = 0.453$, $SD = 0.102$, 95% CI [0.390, 0.516]), Recall ($M = 0.489$, $SD = 0.157$, 95% CI [0.391, 0.586]) and F1-score ($M = 0.445$, $SD = 0.097$, 95% CI [0.385, 0.505]). These results indicate that the framework transfers well across different state-of-the-art MLLMs. The consistently similar scores across models suggest strong generalisability and robustness to architectural variations, demonstrating that its effectiveness is not tied to a single MLLM implementation.

Alignment with experts. To test whether the ETSA approach and expert labels differed at the image level, we applied exact McNemar’s test on a 2×2 contingency table of agreements/disagreements. The result ($\chi^2 = 0.367$, $p = 0.545$) indicates no statistically significant difference between the two label sets. The corresponding effect size, defined as the difference in proportions, was 0.020 with a 95% CI of [−0.042, 0.082]. Because values of effect size close to zero indicate balanced disagreement, this small effect size suggests that ETSA exhibits no meaningful directional bias relative to expert annotations. While a non-significant result does not prove equivalence, it does provide no evidence of systematic divergence between ETSA outputs and expert judgements.

Perceived accuracy. Design experts rated the ETSA approach’s analyses on a 0–3 scale. The mean score was 2.267 ($SD = 0.691$) (Figure 10), which indicates that experts generally regarded the outputs as reliable and practically valuable for interpreting scanpaths.

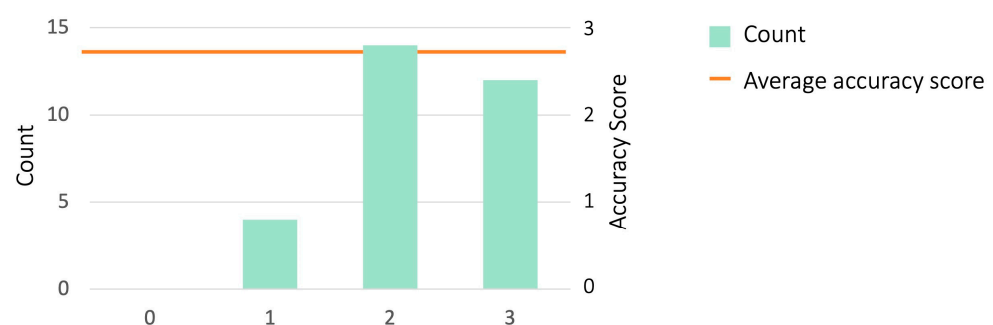


Figure 10. Distribution of subjective accuracy assessment scores from design experts and the corresponding average accuracy score.

5. Discussion

This study introduced and evaluated a MLLM-based framework for interpreting eye tracking scanpaths in industrial design. Through two complementary evaluations assessing its reliability and effectiveness, the ETSA approach demonstrated high semantic stability, moderate alignment with expert feature identification, clear benefits from the integrated knowledge base, and strong cross-model generalisation. The approach also shows no detectable distributional difference from expert labels, and was positively appraised by expert designers. Table 3 summarised the results from these experiments.

Table 3. Summary of the experimental results.

Evaluation	Experiments	Metrics	Main Results
Reliability evaluation	Stability study by repeated runs	Semantics similarity	Mean = 0.884
Effectiveness evaluation	Overall performance (GPT-4o w/knowledge base)	HL, P, R, F1	HL = 0.364, P = 0.488, R = 0.491, F1 = 0.476
	Ablation study (GPT-4o w/o knowledge base)	HL, P, R, F1	HL = 0.373, P = 0.395, R = 0.308, F1 = 0.320
	Cross-model generalisability (Claude Sonnet 4)	HL, P, R, F1	HL = 0.393, P = 0.453, R = 0.489, F1 = 0.445
	Alignment with experts	Exact McNemar test	$p = 0.545$
	Perceived accuracy	Questionnaire (0–3 scale)	Mean = 2.267

Across both evaluations, our findings highlight several mechanisms through which ETSA achieves dependable and interpretable scanpath analysis. First, the ETSA approach produced highly consistent interpretations, indicating that the structured prompt design and domain-grounded reasoning effectively reduced stochastic variability in multimodal generation. This stability is particularly significant given that generative models often exhibit random fluctuations in output, which has historically limited their use in scientific and professional applications. The ETSA approach’s consistency suggests that transparent reasoning, such as least-to-most prompting and chain-of-thought procedures, can meaningfully improve the dependability of large language models for analytic tasks in HCI.

Second, the effectiveness evaluation results indicated a moderate predictive performance of the ETSA approach, and a balanced precision and recall. The findings demonstrated that ETSA achieved identifying the features highlighted by experts while limiting spurious predictions. While the F1 score is moderate (0.476), it should be interpreted in the context of a highly subjective and complex multi-label interpretation task and achieving near-perfect scores would be unlikely even between human experts. The balance between precision and recall is particularly encouraging, as it suggests the model avoids common failure modes of being overly verbose or overly conservative.

Third, the ablation study illustrates the central role of the knowledge base. Remove the knowledge base greatly degraded interpretive quality, demonstrating that domain grounding is essential for aligning model reasoning with expert visual cognition. This effect is also evident in the qualitative behaviour of the model. For example, a representative reasoning excerpt states: “The user’s visual attention switched multiple times between the product and the background, which could be related to factors such as contrast, colour and shading, suggesting . . .” This reflects how multiple feature-related rules in the knowledge base jointly inform the model’s interpretation.

Moreover, the cross-model comparison suggests that ETSA's behaviour is not tied to a specific MLLM architecture, underscoring that the effectiveness of ETSA primarily from its structured workflow, including structural information, knowledge base integration, and prompt design, rather than single model's internal capabilities. Therefore, the ETSA approach functions as a dependable "second reader" in design practice—delivering repeatable, interpretable textual descriptions that correspond with expert judgement while maintaining computational scalability.

Beyond the quantitative results, the findings carry broader implications for human–computer interaction and the design research community. The high semantic consistency across repeated runs supports the notion of hybrid intelligence, where large language models serve as collaborative analytic partners rather than replacements for human expertise. In practical design settings, ETSA could support activities such as providing initial scanpath interpretations during design reviews, highlighting atypical or unexpected gaze transitions for further inspection, or generating baseline analyses that help designers converge on shared interpretations of visual behaviour. For example, during iterative concept development, ETSA could offer rapid assessments of whether different design variants attract attention as intended, allowing experts to focus on higher-level interpretive reasoning. The workflow aligns with the HCI community's growing interest in collaborative and explainable AI—systems that amplify, rather than automate, human cognition.

The ETSA approach also advances the methodological reproducibility of scanpath analysis. Traditional approaches, whether expert-driven or metric-based, often produce results that are difficult to replicate across studies or practitioners. By encoding visual-behaviour theory into a structured knowledge base and constraining the reasoning process through prompt engineering, the proposed ETSA approach transforms a subjective interpretive task into a transparent and standardised analytical procedure. This methodological consistency contributes to resolving long-standing reproducibility issues in behavioural HCI research and opens the possibility of creating benchmark protocols for visual attention analysis.

Looking ahead, ETSA could be extended into an interactive analytic tool by incorporating real-time expert feedback. Such a system would allow designers to correct, approve, or refine intermediate reasoning steps as the model processes scanpaths, enabling iterative co-analysis. Real-time interaction would also support adaptive knowledge-based expansion as experts contribute new visual cues or design heuristics during use. This interactive loop would deepen the hybrid-intelligence paradigm by enabling mutual adaptation between expert insight and model-level reasoning, creating a collaborative analytic workflow that is both explainable and incrementally improving.

At a conceptual level, the study illustrates how general-purpose MLLMs can be adapted into domain-specific analytic instruments. By grounding the ETSA approach's reasoning in established visual attention theory, we demonstrate that it is possible to bridge data-driven analysis and cognitive interpretation. This approach may inspire future HCI research on constructing structured reasoning frameworks for other multimodal analytic tasks—such as interface evaluation, information visualisation, or interaction design—where interpretive depth and computational efficiency must coexist.

Nevertheless, several limitations should be acknowledged. First, the evaluation dataset focused primarily on poster-based industrial design stimuli, which may constrain the generalisability of the findings. Potential dataset biases such as cultural conventions, stylistic norms, or differences in visual complexity may also influence model interpretations. Extending the dataset to include sketches, 3D products, physical prototypes, and culturally diverse design materials would provide a more comprehensive validation. Second, the number of experts involved in the effectiveness evaluation was relatively small, future work would help capture a broader range of expert perspectives across different levels of

expertise. Third, while the knowledge base was tailored to industrial design, its adaptability to other domains such as UI/UX or architectural design remains untested. Additionally, the framework relies on upstream models such as SoM for segmentation, whose inaccuracies may have a cascading impact on the downstream analysis.

Beyond technical enhancement, future research should explore interactive and explainable extensions of the ETSA approach. Integrating real-time expert feedback could enable iterative reasoning refinement, transforming the model from a static interpreter into an interactive analytic collaborator. Such developments would align with emerging paradigms in participatory and explainable AI, allowing experts to query, critique, and guide model reasoning directly. More broadly, these advances would support the design of AI systems that operate not as black-box predictors but as transparent cognitive partners, capable of co-constructing meaning with human users.

In sum, this research demonstrates that structured prompting, knowledge grounding, and explicit reasoning constraints can transform general-purpose MLLMs into interpretable analytic assistants for complex eye tracking scanpath analysis. The ETSA approach contributes to both methodological rigour and conceptual understanding in HCI by showing how AI systems can participate meaningfully in the interpretation of human visual perceptual behaviour. By bridging computational scalability and expert interpretive depth, this ETSA approach represents a step toward more transparent, reproducible, and collaborative forms of human–AI interaction in design research and practice.

6. Conclusions

This paper introduced ETSA approach, a MLLM-based framework for automated interpretation of eye tracking scanpaths in industrial design. ETSA integrates structural information extraction, a theory-grounded knowledge base, and a structured prompting strategy to generate systematic, transparent scanpath interpretation. Our empirical evaluation demonstrated that ETSA achieves high reliability and strong expert alignment. Therefore, the ETSA not only enhances the reproducibility of scanpath interpretation but also opens new avenues for explainable AI systems that support design reasoning rather than replace it.

Beyond empirical performance, ETSA also illustrates how structured prompting and domain grounding can repurpose general-purpose MLLMs into specialised instruments for scanpath reasoning. This positions ETSA as a hybrid-intelligence pipeline that combines human expertise with machine reasoning, enabling scalable yet meaningful interpretation of visual behaviour. Therefore, ETSA provides a scalable and theory-grounded framework for transparent, standardized scanpath analysis in human–AI collaborative design research.

Author Contributions: Conceptualisation: X.L. and Y.G.; methodology and experiment conduction: Y.G.; formal analysis: K.Y.; data curation, Y.G.; writing and visualisation: K.Y.; supervision, project administration and funding acquisition: X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Zhejiang Provincial Key R&D Programme, grant number 2023C01045 and 2024C01210.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to ethical restriction.

Acknowledgments: During the preparation of this manuscript, the authors used GPT4o for the purposes of language proofreading and grammar checking. The authors have reviewed and edited the outputs and take full responsibility for the contents of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MLLM	Multimodal Large Language Model
LLM	Large Language Model
HCI	Human-Computer Interaction

References

- Goldberg, J.H.; Helfman, J.I. Visual scanpath representation. In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, Austin, TX, USA, 22–24 March 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 203–210. [\[CrossRef\]](#)
- Lopez-Cardona, A.; Emami, P.; Idesis, S.; Duraisamy, S.; Leiva, L.A.; Arapakis, I. A Comparative Study of Scanpath Models in Graph-Based Visualization. In Proceedings of the 2025 Symposium on Eye Tracking Research and Applications, Tokyo, Japan, 26–29 May 2025; Association for Computing Machinery: New York, NY, USA; p. 89. [\[CrossRef\]](#)
- Wang, Y.; Băce, M.; Bulling, A. Scanpath Prediction on Information Visualisations. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 3902–3914. [\[CrossRef\]](#)
- Ilhan, A.E.; Togay, A. Use of eye-tracking technology for appreciation-based information in design decisions related to product details: Furniture example. *Multimed. Tools Appl.* **2023**, *83*, 8013–8042. [\[CrossRef\]](#)
- Yang, W.; Su, J.; Qiu, K.; Zhang, X.; Zhang, S. Research on Evaluation of Product Image Design Elements Based on Eye Movement Signal. In *Engineering Psychology and Cognitive Ergonomics*; Springer International Publishing: Cham, Switzerland, 2019; pp. 214–226. [\[CrossRef\]](#)
- Wu, B.; Zhu, Y.; Yu, K.; Nishimura, S.; Jin, Q. The effect of eye movements and cultural factors on product color selection. *Hum.-Centric Comput. Inf. Sci.* **2020**, *10*, 48. [\[CrossRef\]](#)
- Frame, M.E.; Warren, R.; Maresca, A.M. Scanpath comparisons for complex visual search in a naturalistic environment. *Behav. Res. Methods* **2019**, *51*, 1454–1470. [\[CrossRef\]](#)
- Cristino, F.; Mathôt, S.; Theeuwes, J.; Gilchrist, I.D. ScanMatch: A novel method for comparing fixation sequences. *Behav. Res. Methods* **2010**, *42*, 692–700. [\[CrossRef\]](#)
- Dewhurst, R.; Nyström, M.; Jarodzka, H.; Foulsham, T.; Johansson, R.; Holmqvist, K. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behav. Res. Methods* **2012**, *44*, 1079–1100. [\[CrossRef\]](#)
- Dolezalova, J.; Popelka, S. ScanGraph: A Novel Scanpath Comparison Method Using Visualisation of Graph Cliques. *J. Eye Mov. Res.* **2016**, *9*, 21. [\[CrossRef\]](#)
- Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, Hawaii, 23–29 July 2023; pp. 19730–19742. Available online: <https://proceedings.mlr.press/v202/li23q.html> (accessed on 2 December 2025).
- Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916.
- Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A. Gpt-4o system card. *arXiv* **2024**, arXiv:2410.21276. [\[CrossRef\]](#)
- Xu, Z.; Shi, S.; Hu, B.; Wang, L.; Zhang, M. MultiSkill: Evaluating large multimodal models for fine-grained alignment skills. In Proceedings of the Findings of the Association for Computational Linguistics, Miami, FL, USA, 12–16 November 2024; pp. 1506–1523. [\[CrossRef\]](#)
- Zhou, D.; Scharli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Bousquet, O.; Le, Q.; Chi, E.H. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv* **2022**, arXiv:abs/2205.10625.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Curran Associates Inc.: Red Hook, NY, USA, 2022; p. 1800. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf (accessed on 2 December 2025).
- Noton, D.; Stark, L. Scanpaths in eye movements during pattern perception. *Science* **1971**, *171*, 308–311. [\[CrossRef\]](#)
- Anderson, N.C.; Anderson, F.; Kingstone, A.; Bischof, W.F. A comparison of scanpath comparison methods. *Behav. Res. Methods* **2015**, *47*, 1377–1392. [\[CrossRef\]](#)
- Foulsham, T. Scenes, saliency maps and scanpaths. In *Eye Movement Research: An Introduction to its Scientific Foundations and Applications*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 197–238. [\[CrossRef\]](#)

20. Sun, W.; Chen, Z.; Wu, F. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 2101–2118. [\[CrossRef\]](#)
21. Starke, S.D.; Baber, C. The effect of four user interface concepts on visual scan pattern similarity and information foraging in a complex decision making task. *Appl. Ergon.* **2018**, *70*, 6–17. [\[CrossRef\]](#)
22. Bigne, E.; Simonetti, A.; Guixeres, J.; Alcaniz, M. Visual attention and product interaction: A neuroscientific study on purchase across two product categories in a virtual store. *Int. J. Retail. Distrib. Manag.* **2024**, *52*, 389–406. [\[CrossRef\]](#)
23. Zhang, N.; Zhang, J.; Jiang, S.; Ge, W. The Effects of Layout Order on Interface Complexity: An Eye-Tracking Study for Dashboard Design. *Sensors* **2024**, *24*, 5966. [\[CrossRef\]](#)
24. Zhou, Q.; Cheng, Y.; Liu, Z.; Chen, Y.; Li, C. The Layout Evaluation of Man-Machine Interface Based on Eye Movement Data. In Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), Florence, Italy, 26–30 August 2018; Springer International Publishing: Cham, Switzerland, 2019; pp. 64–75. [\[CrossRef\]](#)
25. Guo, F.; Ding, Y.; Liu, W.; Liu, C.; Zhang, X. Can eye-tracking data be measured to assess product design?: Visual attention mechanism should be considered. *Int. J. Ind. Ergon.* **2016**, *53*, 229–235. [\[CrossRef\]](#)
26. Eraslan, S.; Yesilada, Y.; Harper, S. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *J. Eye Mov. Res.* **2015**, *9*, 2. [\[CrossRef\]](#)
27. Holsanova, J.; Holmberg, N.; Holmqvist, K. Reading information graphics: The role of spatial contiguity and dual attention guidance. *Appl. Cogn. Psychol.* **2009**, *23*, 1215–1226. [\[CrossRef\]](#)
28. Renshaw, J.A.; Finlay, J.E.; Tyfa, D.; Ward, R.D. Understanding visual influence in graph design through temporal and spatial eye movement characteristics. *Interact. Comput.* **2004**, *16*, 557–578. [\[CrossRef\]](#)
29. Mohamed Selim, A.; Barz, M.; Bhatti, O.S.; Alam, H.M.T.; Sonntag, D. A review of machine learning in scanpath analysis for passive gaze-based interaction. *Front. Artif. Intell.* **2024**, *7*, 1391745. [\[CrossRef\]](#)
30. Faghihi, N.; Vaid, J. Reading/writing direction as a source of directional bias in spatial cognition: Possible mechanisms and scope. *Psychon. Bull. Rev.* **2023**, *30*, 843–862. [\[CrossRef\]](#)
31. Fahimi, R.; Bruce, N.D. On metrics for measuring scanpath similarity. *Behav. Res. Methods* **2021**, *53*, 609–628. [\[CrossRef\]](#)
32. Busjahn, T.; Tamm, S. A deeper analysis of AOI coverage in code reading. In *ACM Symposium on Eye Tracking Research and Applications*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–7. [\[CrossRef\]](#)
33. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Curran Associates Inc.: Red Hook, NY, USA, 2023; p. 2020. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf (accessed on 2 December 2025).
34. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023. [\[CrossRef\]](#)
35. Fu, J.; Ng, S.-K.; Jiang, Z.; Liu, P. GPTScore: Evaluate as You Desire. In *North American Chapter of the Association for Computational Linguistics*; Elsevier: Amsterdam, The Netherlands, 2023.
36. Duan, P.; Warner, J.; Hartmann, B. Towards Generating UI Design Feedback with LLMs. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023; Association for Computing Machinery: New York, NY, USA, 2023; p. 70. [\[CrossRef\]](#)
37. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [\[CrossRef\]](#)
38. Tan, Q.; Ng, H.T.; Bing, L. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv* **2023**, arXiv:2306.08952. [\[CrossRef\]](#)
39. Blascheck, T.; Kurzahls, K.; Raschke, M.; Strohmaier, S.; Weiskopf, D.; Ertl, T. AOI hierarchies for visual exploration of fixation sequences. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, Charleston, South Carolina, 14–17 March 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 111–118. [\[CrossRef\]](#)
40. Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.-y.; Gao, J. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv* **2023**, arXiv:abs/2310.11441.
41. Rensink, R.A. The management of visual attention in graphic displays. In *Human Attention in Digital Environments*; Roda, C., Ed.; Cambridge University Press: Cambridge, UK, 2011; pp. 63–92. [\[CrossRef\]](#)
42. Eldesouky, D.F.B. Visual Hierarchy and Mind Motion in Advertising Design. *J. Arts Humanit.* **2013**, *2*, 148–162.
43. Xu, J.; Jiang, M.; Wang, S.; Kankanhalli, M.S.; Zhao, Q. Predicting human gaze beyond pixels. *J. Vis.* **2014**, *14*, 28. [\[CrossRef\]](#)
44. Borji, A.; Sihite, D.N.; Itti, L. What stands out in a scene? A study of human explicit saliency judgment. *Vis. Res.* **2013**, *91*, 62–77. [\[CrossRef\]](#)

45. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2–7 June 2019; Volume 1, pp. 4171–4186. [\[CrossRef\]](#)
46. Castner, N.; Arsiwala-Scheppach, L.; Mertens, S.; Krois, J.; Thaqi, E.; Kasneci, E.; Wahl, S.; Schwendicke, F. Expert gaze as a usability indicator of medical AI decision support systems: A preliminary study. *npj Digit. Med.* **2024**, *7*, 199. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.