G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios

Citation: Wang, Zeyu, Shi, Yuanchun, Wang, Yuntao, Yao, Yuchen, Yan, Kun et al. 2024. "G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8 (2).

As Published: 10.1145/3659623

Publisher: Association for Computing Machinery

Persistent URL: https://hdl.handle.net/1721.1/155208

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios

ZEYU WANG, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China
YUANCHUN SHI*, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China and Intelligent Computing and Application Laboratory of Qinghai Province, Qinghai University, China
YUNTAO WANG*, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China
YUCHEN YAO, Tsinghua University, China
KUN YAN, Microsoft Research Asia, China and SKLSDE Lab, Beihang University, China
YUHAN WANG, Beijing University of Posts and Telecommunications, China
LEI JI, Microsoft Research Asia, China
XUHAI XU, Massachusetts Institute of Technology, United States
CHUN YU, Tsinghua University, China

Modern information querying systems are progressively incorporating multimodal inputs like vision and audio. However, the integration of gaze — a modality deeply linked to user intent and increasingly accessible via gaze-tracking wearables — remains underexplored. This paper introduces a novel gaze-facilitated information querying paradigm, named G-VOILA, which synergizes users' gaze, visual field, and voice-based natural language queries to facilitate a more intuitive querying process. In a user-enactment study involving 21 participants in 3 daily scenarios (p = 21, scene = 3), we revealed the ambiguity in users' query language and a gaze-voice coordination pattern in users' natural query behaviors with G-VOILA. Based on the quantitative and qualitative findings, we developed a design framework for the G-VOILA paradigm, which effectively integrates the gaze data with the in-situ querying context. Then we implemented a G-VOILA proof-of-concept using cutting-edge deep learning techniques. A follow-up user study (p = 16, scene = 2) demonstrates its effectiveness by achieving both higher objective score and subjective score, compared to a baseline without gaze data. We further conducted interviews and provided insights for future gaze-facilitated information querying systems.

*Corresponding authors.

Authors' Contact Information: Zeyu Wang, wang-zy23@mails.tsinghua.edu.cn, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Haidian Qu, Beijing Shi, China; Yuanchun Shi, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Haidian Qu, Beijing Shi, China and Intelligent Computing and Application Laboratory of Qinghai Province, Qinghai University, Xining, Qinghai, China, shiyc@tsinghua.edu.cn; Yuntao Wang, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Haidian Qu, Beijing Shi, China, yuntaowang@tsinghua.edu.cn; Yuchen Yao, Tsinghua University, Haidian Qu, Beijing Shi, China, yaoyc19@mails.tsinghua.edu.cn; Kun Yan, Microsoft Research Asia, Haidian Qu, Beijing Shi, China and SKLSDE Lab, Beihang University, Haidian Qu, Beijing Shi, China, kunyan@buaa.edu.cn; Yuhan Wang, Beijing University of Posts and Telecommunications, Haidian Qu, Beijing Shi, China, yigetianluoturanjiu@gmail.com; Lei Ji, Microsoft Research Asia, Haidian Qu, Beijing Shi, China, leiji@microsoft.com; Xuhai Xu, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States, orson.xuhai.xu@gmail.com; Chun Yu, Tsinghua University, Haidian Qu, Beijing Shi, China, chunyu@tsinghua.edu.cn.

ACM 2474-9567/2024/6-ART78
https://doi.org/10.1145/3659623

CCS Concepts: • **Human-centered computing → Natural language interfaces**; **Ubiquitous and mobile computing systems and tools**; • **Information systems → Information retrieval query processing**.

Additional Key Words and Phrases: information retrieval, gaze tracking, smart glasses, large language models, information query

ACM Reference Format:
Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. 2024. G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 78 (June 2024), 33 pages. https://doi.org/10.1145/3659623

1 INTRODUCTION

As the cyber and physical spaces quickly merge, people exhibit a significant demand for information retrieval (IR) anywhere and anytime in their daily lives [10, 14, 15, 22, 56], no longer confined to a specific device or location. With advancements in the computational capabilities of wearable devices, the incorporation of a virtual assistant that can provide on-demand, in-situ answers to users' inquiries has the potential to greatly facilitate the interaction with surrounding targets [68, 75] and enhance the naturalness of the user's information retrieval experience [3]. Specifically, smart glasses with gaze tracking open new possibilities for natural information retrieval techniques in daily scenarios by combining the voice and gaze modalities [35, 62].

Prior research has investigated the potential of utilizing smart glasses [29, 67] equipped with gaze tracking to streamline information retrieval related to physical objects [3, 15, 26, 39, 52, 71, 74]. These approaches enhanced information retrieval by superimposing digital widgets around physical entities, allowing users to access information through interacting with the widgets in a multi-stage manner [3, 14]. While this has been a significant stride, the ultimate goal is to enable users to retrieve information using natural expressions, eliminating the need for these intermediary widgets.

This paper introduces G-VOILA[1], a future information querying paradigm that combines users' gaze data, visual field and voice-based natural language queries. However, a challenge arises when users employ multi-modal, often ambiguous expressions during interactions with virtual assistants tailored for information querying in everyday settings [12, 41]. The central issue lies in the precise interpretation of such ambiguous expressions, particularly in the presence of data from other modalities, such as gaze. Previous work have shown the relation between eye movements pattern and daily activities [13, 35], there remains a gap in understanding how gaze is intricately linked to daily querying, especially in the context of mouth-eye coordination.

To gain insights into users' natural expression patterns, anticipated use cases, and potential engagement with G-VOILA, we conducted a user-enactment study involving 21 participants. Our findings indicate that users frequently omitted specific context-related details in their expressions, assuming that the system could inherently understand such contexts. The study also unveiled distinct patterns of mouth-eye coordination in user expressions, particularly arising from pronoun usage and gaze fixation on items relevant to their queries. Figure 1 illustrates how the user naturally interacts with G-VOILA paradigm in a supermarket for ingredient selection.

Based on the identified natural expression patterns from the user-enactment study, we proposed a design framework for implementing an intelligent assistant within the G-VOILA paradigm. Specifically, we established an inferential pipeline between the user's gaze pattern and the often ambiguous intent behind their inquiries.

To evaluate the efficacy of the G-VOILA paradigm and proposed framework, we implemented a preliminary version of VOILA-G[2] assistant to conduct a proof-of-concept user study, with 16 participants using Pupil Labs

[1] We refer to our proposed paradigm as G-VOILA because it can be an abbreviation of **G**aze-facilitated **VOI**ce and **VI**sual-based natural **L**anguage querying **A**ssistant. Also, "Voila" is used to indicate or draw attention to something that is being presented or brought to someone's notice in French, meaning "there it is" in English, which can represent the case of using pronouns as described in Section 4.2

[2] We put the letter "G" backward to separate the G-VOILA paradigm from our preliminary implementation, also to be in consist with other baselines mentioned in Section 6.2

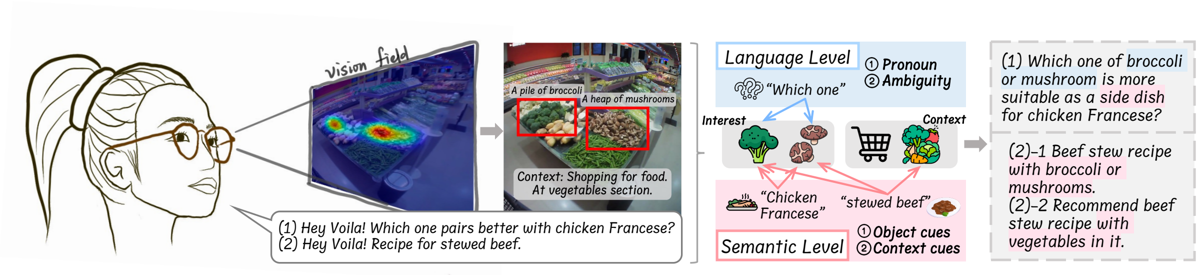

vision field
A pile of broccoli
A heap of mushrooms
Context: Shopping for food. At vegetables section.
(1) Hey Voila! Which one pairs better with chicken Francese?
(2) Hey Voila! Recipe for stewed beef.
Language Level
① Pronoun
② Ambiguity
"Which one"
Interest
Context
"Chicken Francese"
"stewed beef"
① Object cues
② Context cues
Semantic Level
(1) Which one of broccoli or mushroom is more suitable as a side dish for chicken Francese?
(2)-1 Beef stew recipe with broccoli or mushrooms.
(2)-2 Recommend beef stew recipe with vegetables in it.

Fig. 1. Illustrating G-VOILA use cases in a shopping scenario. A user wearing smart glasses is shopping for dinner. She reaches the vegetable aisle and desires detailed information about specific vegetables. Here is how an assistant within the G-VOILA paradigm might assists her: (1) **Query Initiation**: She posts queries through natural voice commands. (2) **Contextual Analysis**: The assistant analyzes the user's field of view and gaze to discern specific areas of interest. (3) **Query Response**: By further aligning the situational contexts with the posted questions, the assistant deduces her precise query intent and delivers a clear response.

Invisible gaze tracker [64] in 2 daily life settings, covering 12 specific inquiry tasks. We compared VOILA-G to baseline methods in an ablation manner. The results indicated that the answers generated by VOILA-G achieved an 89% objective recall score, surpassing the baseline by 11%, illustrating the increment of incorporating gaze in an information querying system. VOILA-G also received a significantly higher level of preference score than the baseline in terms of subjective evaluation matrices, including matchness between reconstructed query expression and user intent, as well as satisfaction, precision, usefulness for the generated response, etc. We gained insights in an afterwards interview session to further understand user's experience and anticipation for VOILA-G.

Overall, our contributions can be summarized as four-fold:

- We proposed a future information querying paradigm that combines users' gaze data, visual field and voice-based natural language question as input, namely the G-VOILA paradigm.
- We conducted an user-enactment study that reveals users' natural query behavior in a mouth-eye coordination manner within the G-VOILA paradigm, as well as revealed qualitative and quantitative findings.
- We proposed a design framework for intelligent assistant within the G-VOILA paradigm upon insights gained from the user-enactment study. We implemented a proof-of-concept assistant under the proposed framework, namely VOILA-G.
- We conducted a user study that validated VOILA-G's effectiveness in inferring users' query intent and showed significantly better subjective outcomes across several metrics including satisfaction, usefulness, and mental demand etc.

2 RELATED WORKS

In this section, we describe the related work, including information retrieval for daily scenarios, contextual computing for interaction, and multi-modal large models. We discuss the difference between G-VOILA and existing works at the end of each subsections.

2.1 Information Retrieval for Daily Scenarios

Information Retrieval (IR) has evolved from manual document searches to keyword-based online search engines, yet still encounters query reformulation issue [10, 16]. Recently, generative AI has empowered search engines to understand more natural and expressive queries [1]. Meanwhile, multi-modality-based search tools have started to leverage diverse information types as input [79]. Despite these advancements, contemporary IR techniques remain predominantly restricted to traditional screen-based interfaces.

With the advancement of Augmented Reality (AR) devices, researchers have integrated IR closely with physical world through Head-Mounted Displays (HMDs) [3, 14, 27, 39, 52, 71, 76], which is defined as reality-based information retrieval (RBIR) by [15]. RBIR adheres to the principle of "find what you need with zero query terms or less" [7], adopting a proactive strategy to impose information on certain objects in order to reduce inquiry costs [27, 39, 76]. However, this approach introduces the issue of information overload and challenges in precisely addressing users' queries.

To tackle information overload in AR, Piening et al [52] proposed gaze-adaptive interfaces to modulate the display of information panels. Meanwhile, researchers [3, 14, 15] introduced multi-round IR operations leveraging AR's interactivity to facilitate precise information acquisition, including selection and search refinement. For example, a user interested in avocados at a grocery store may select "nutrition details" rather than "recipes" on avocados' information panel. However, these proactive display strategies either struggle to align with user intent or require extensive interaction, impeding the immediate recognition of information utility[56]. In contrast, natural language queries can convey intentions more accurately, thus G-VOILA returns to question-answering interaction paradigm enriched with contextual information.

2.2 Contextual Computing for Interaction

Context-awareness can enhance interactive technologies' ability to discern user intentions. Advancements in wearable devices brought by integration of various sensors, exemplified by Pupil Labs Invisible glasses [64], have enabled electronic devices to offer increasingly intelligent services.

2.2.1 *Gaze-based Intention Reasoning.*

Gaze has been applied across various tasks related to cognitive reasoning, including collaborative robotics [8, 21, 38], interactive technology inputs [51, 53, 69], reading assistant [4, 23], cognitive load measurement [50], and intent classification [24, 31, 45, 72]. Eye movements is a reliable indicator of user intent - "eyes rarely visit objects that are irrelevant to the action" [35, 62], which suggests a strong correlation between eye movements and inquiry intent in daily scenarios. Current methods for gaze-based intent prediction typically employ deep learning techniques, such as CNNs, for predefined classifications [2, 57]. Gaze data also serve as key indicator of interest in RBIR [3, 52]. To date, no research has integrated eye movements with natural language to deduce user intent.

2.2.2 *In-situ voice-based interaction.*

Research in context-aware, command-based interaction systems has been extensively conducted. Studies on user programming of smart home appliances revealed a preference for ambiguous and multimodal command expression, such as combining gesture with spoken pronouns [12, 41]. To our best knowledge, there is a lack of research on both qualitative language patterns and quantitative pattern for gaze-enhanced voice interactions. This paper aims to fill this gap by presenting a user-enactment study leveraging G-VOILA's context-aware capabilities to identify patterns in user's in-situ expressions.

2.3 Multi-Modal Large Models

Recent advancements in large language models (LLMs), including ChatGPT [47], GPT-4 [46], and others [19, 63, 65], have catalyzed research in multimodal information integration with substantial relevance to Human-Computer Interaction (HCI). These investigations can be primarily classified into two main approaches: Integrated Systems and Unified Multimodal Models. Each approach promotes fluid interaction between humans and computers, bridging vision and language to enhance user experience and efficiency in various applications.

2.3.1 Integrated Systems Approach. This approach employs ChatGPT [47] as a central orchestrator, connecting specialized models designed for distinct visual tasks. Language prompts serve as bridges to invoke expert visual-language models, such as VisualChatGPT [70], HuggingGPT [59], Cola [18], X-GPT [84], MM-REACT [77], and ViperGPT [61]. Despite offering a modular and adaptable solution for multimodal integration, this method presents challenges in individual model's training limitations and increased API query costs, potentially affecting the efficiency and scalability of HCI systems like G-VOILA.

2.3.2 Unified Vision-Language Model Approaches. Efforts to integrate language models with visual processing have led to the creation of models like Flamingo, OpenFlamingo, and BLIP-2 [5, 9, 37]. The release of GPT-4 [46] further catalyzed the development of advanced enterprise models including GPT-4 vision-language model (OpenAI), PaLM-E (Google), ERNIE (Baidu), Tongyi Qianwen (Alibaba), and SenseNova (Sensetime) [6, 11, 25, 46, 58]. In academia, contributions include LLaMA-Adapters, Mini-GPT4, LLaVA, Otter, and Voila-A [28, 36, 40, 74, 83]. LLaMA-Adapters extends LLaMA with an adapter module and multimodal prompts [65]. Mini-GPT4 adapts BLIP-2's architecture, integrating Vicuna to replace the language decoder [19, 37]. LLaVA utilizes a trainable projector matrix for text-visual modality connections, which increases training demands [40]. Otter, inspired by Flamingo, utilizes cross-gated attention layers for vision-language integration while freezing the vision encoder and language decoder to simplify training [5, 36]. Voila-A [74] takes a further step, discusses how to align vision-language models with user's gaze attention.

Despite these models are capable of engaging in image-related dialogues, they are limited by their dependence on users providing detailed queries and context. This reliance often leads to misunderstandings of user intent, thus limiting their practicality and real-world impact.

2.3.3 Multimodal Integration for Alignment. Efforts to enhance fine-grained grounding and align with human intentions have incorporated diverse modalities, including bounding boxes [17, 20, 81, 82], patches [33], coordinate tokens [48, 66, 78], and traces [54, 73]. Despite advancements in multimodal integration, challenges persist in accurately interpreting context and intentions to provide pertinent, timely responses. This gap underscores the necessity for continued research to refine models that can adeptly navigate contextual nuances and human intentions, thereby enhancing the efficacy and intuitiveness of human-computer interactions as seen in G-VOILA.

3 THE USER-ENACTMENT STUDY

In this section, we first introduce the G-VOILA paradigm that combines gaze and voice inputs for information querying in daily scenarios. Then, we describe the prototype for an enactment experiment, which explores user's querying behavior under the G-VOILA paradigm.

3.1 Introducing G-VOILA: A Gaze-Facilitated Querying Paradigm

We propose G-VOILA as a novel ubiquitous querying paradigm that combines users' gaze data, visual field and voice-based natural language question as input. The design concerns of G-VOILA can be summarized in three folds:

(1) **Hands-free:** G-VOILA is designed to be deployed on wearable devices with gaze sensors, such as smart glasses and VR/AR systems, allowing users to post questions verbally without the need to operate the device with hands.
(2) **In-context:** G-VOILA's ability to capture visual field provides environmental clues relevant to users' queries, and in turn, users may formulate questions closely related to their current situation.
(3) **Multi-modality:** G-VOILA enables multi-modality data inputs that can complement each other to produce a more comprehensive understanding of users' questions. Consequently, users can express queries in a more flexible manner.

In traditional mobile device usage, users typically focus on relevant context in their surroundings before inputting necessary data into the application. However, given G-VOILA's properties, users can express in multiple modalities simultaneously, unlike mobile systems that necessitate sequential input of natural language or media. This raises a unique research question: do users' gaze patterns show consistency with their spoken natural language, and if so, how do they relate spatiotemporally?

3.2 The User-Enactment Study

To investigate user's querying behavior with G-VOILA, we conducted a user-enactment study where participants submit queries in any manner they believed G-VOILA should be able to comprehend and respond to, given its sensing capabilities.

3.2.1 Scenarios and Participants. Each participant partake in one of three daily scenarios: supermarket shopping, museum visits, and domestic living. For supermarket and museum, we selected publicly accessible and popular locations within local community. For domestic living, we established an independent room furnished with at least 10 pieces of furniture and 20 common household items. We enrolled 21 participants (8, 5, 8 for each scene), consisting of 13 males and 8 females aged from 19 to 30 ($std = 3.18$). The duration of participation ranged from 90 to 150 minutes, with a compensation rate of 30 USD per hour.

3.2.2 Hardware Setup. The Pupil Labs Invisible [64] [3] is a pair of gaze-tracking smart glasses widely used in research. Featuring gaze sensors, egocentric camera, and microphone, it meets the needs for G-VOILA. Participants were instructed to wear the glasses without any headdress that might cause occlusion to the sensors. Data was continuously recorded as participants engaged in their chosen scenario.

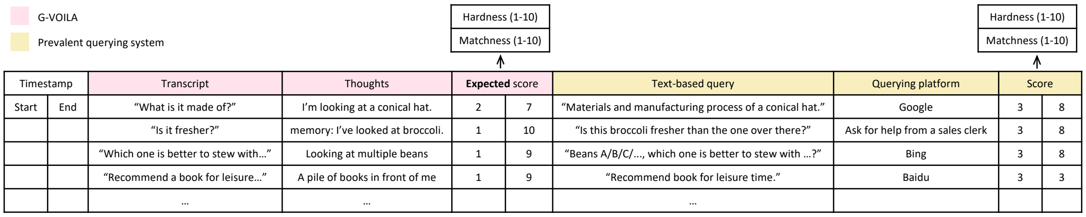
3.2.3 Procedure. The study comprises three sessions: an introduction session, an enactment session where participants pose queries within their chosen scenario, and a discussion session for reviewing their queries and exploring insightful queries through dialogue with the experimenter.

The introduction session aims to ensure that participants understood G-VOILA's sensing capabilities and encourage them to come up with diverse query formats. We provided an analogical script of conversation with friends, illustrating how multiple expression modalities can be employed in daily life. Then participants put on the Invisible glasses and viewed a real-time egocentric video with overlaid gaze trace, which facilitated their understanding of the gaze-tracking. Each scenario was given a distinct purpose: purchasing ingredients for a lavish dinner, exploring an exhibition for intriguing knowledge, and spending a day at home or working remotely. This general purpose was assigned to prevent participants from feeling aimless and confused in the scene, since they wouldn't receive further instructions or interaction with the experimenter during the enactment session. We also briefed participants on subsequent study procedures.

In the enactment session, participants were required to wear the Invisible glasses and engaged with their chosen scenario. They were encouraged to ask any questions related to the scene. Participants did not receive any responses during this session. A minimum of 10 questions was required. All sensors' data were recorded throughout this session.

In the discussion session, participants reviewed the recorded video (overlaid with gaze traces) from the enactment session and completed a form for each query, as depicted in Figure 2. For each query, they reformulated their queries to search with prevalent IR systems until they received satisfied information or encountered frustration. Based on their daily problem-solving preferences, participants can choose from digital search tools to human interactions (consulting sales clerks) as their querying platform.

[3] The Invisible glasses rely on a connection to a mobile phone to be driven; thus we advised participants to place the phone in their pockets to minimize potential distractions.


G-VOILA
Prevalent querying system
Hardness (1-10)
Matchness (1-10)
Hardness (1-10)
Matchness (1-10)

Timestamp		Transcript	Thoughts	**Expected** score		Text-based query	Querying platform	Score	
Start	End	“What is it made of?”	I’m looking at a conical hat.	2	7	“Materials and manufacturing process of a conical hat.”	Google	3	8
		“Is it fresher?”	memory: I’ve looked at broccoli.	1	10	“Is this broccoli fresher than the one over there?”	Ask for help from a sales clerk	3	8
		“Which one is better to stew with...”	Looking at multiple beans	1	9	“Beans A/B/C/..., which one is better to stew with ...?”	Bing	3	8
		“Recommend a book for leisure...”	A pile of books in front of me	1	9	“Recommend book for leisure time.”	Baidu	3	3
		...	...			...			

Fig. 2. Discussion chart for the user-enactment study. Headers and four illustrative examples were shown in the chart. For the first example, participant supposed G-VOILA could understand “it” to be a conical hat because (s)he was looking at it, whereas there is no clue showing that (s)he also wanted to know the manufacture process, thus G-VOILA should result in a lower matchness score. But when querying with her daily used platform google, she might rephrase her question to be more clarified.

For each query, Participants rated (1) the hardness in finding desired answer, including query formulation and information scanning, and (2) the matchness of the information retrieved. The final query text and utilized querying systems were documented. Although G-VOILA could not respond during the study, participants still evaluated it in terms of hardness and matchness. Matchness scores were assigned based on participants' perceptions of whether G-VOILA could comprehend their verbal question and provide a satisfactory response, given its sensing capabilities.

After reviewing all queries, the experimenter identified queries where significant discrepancies exist between the verbal and textual queries, discussing with participants about how they assumed G-VOILA could understand and respond to such question. Notable insights were recorded in the “thoughts” column. Timestamps were later added to the form by the experimenter.

4 ANALYSIS AND FINDINGS

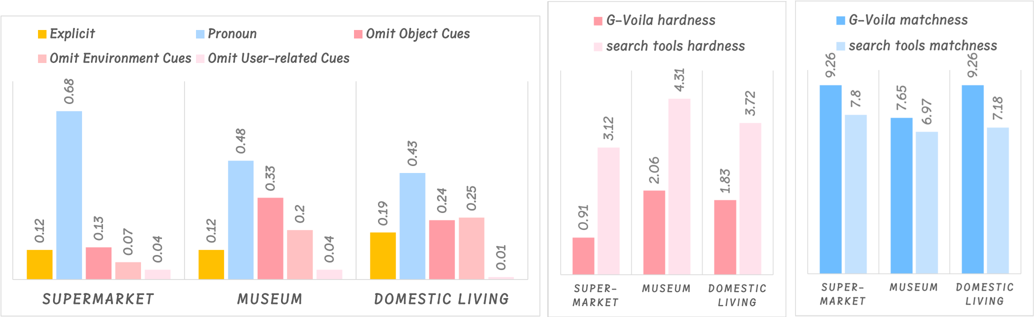
In three predetermined daily life scenarios, we collected a total of 418 valid inquiry data and annotation pairs. The collected data was analyzed to comprehend users' multi-modal inquiry form-factors within the G-VOILA paradigm. As discussed in Section 3.1, we aimed to answer three research questions:

(1) **How do users perceive and anticipate an IR system under the G-VOILA paradigm?**
(2) **What distinctive language expression patterns do users exhibit with G-VOILA?**
(3) **How are users' gaze patterns spatiotemporally related to their spoken query language?**

4.1 User's Comprehension and Expectation of G-VOILA

To understand users' comprehension and expectations of G-VOILA, we computed average hardness and matchness scores for each scenario, as shown in Figure 3b. Lower hardness and higher matchness scores suggest users anticipate G-VOILA to better interpret casual queries. Qualitative analysis of the “thoughts” content yielded three key aspects for G-VOILA to enhance its functionality:

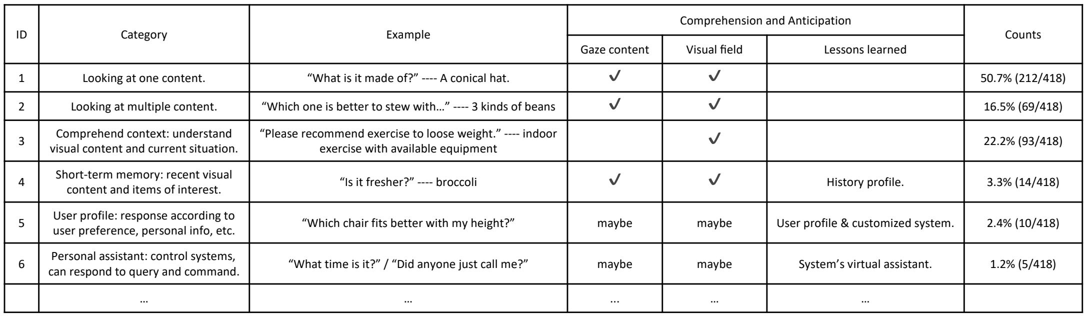
(1) **User interests indicated by gaze content:** When users do not provide sufficient descriptions or indications of their query objects, G-VOILA should be capable of extracting this information by examining users' gaze data.
(2) **Contextual information derived from visual field:** G-VOILA should infer situational context and answer constraints from the shared visual environment when user descriptions are insufficient.
(3) **Incorporating features from other systems:** Users expect G-VOILA to incorporate features characteristic of other systems they have experienced, such as controlling appliance or memorizing personal preferences.



(a) Category counts

(b) Hardness and Matchness

Fig. 3. Statistic results for the formative study, shown for all scenarios. (a) A bar plot displaying statistic count of queries for each category. All queries are classified it's ambiguity, which taxonomy is presented in Section 4.2 and Figure 5. (b) Under the user-envisaged G-VOILA and text-based IR approaches, the hardness of seeking solutions and answer's matchness to user's query intent.



ID	Category	Example	Comprehension and Anticipation			Counts
			Gaze content	Visual field	Lessons learned	
1	Looking at one content.	"What is it made of?" ---- A conical hat.	✓	✓		50.7% (212/418)
2	Looking at multiple content.	"Which one is better to stew with..." ---- 3 kinds of beans	✓	✓		16.5% (69/418)
3	Comprehend context: understand visual content and current situation.	"Please recommend exercise to loose weight." ---- indoor exercise with available equipment		✓		22.2% (93/418)
4	Short-term memory: recent visual content and items of interest.	"Is it fresher?" ---- broccoli	✓	✓	History profile.	3.3% (14/418)
5	User profile: response according to user preference, personal info, etc.	"Which chair fits better with my height?"	maybe	maybe	User profile & customized system.	2.4% (10/418)
6	Personal assistant: control systems, can respond to query and command.	"What time is it?" / "Did anyone just call me?"	maybe	maybe	System's virtual assistant.	1.2% (5/418)
	...	...	...	...	...	

Fig. 4. Breakdown of user queries by their anticipation of G-VOILA usage. "Lessons learned" are short for "Lessons learned from other well-designed system". At least one query example are provided for each category.

As depicted in Figure 4, user queries were classified according to their anticipation of G-VOILA's functionality. The majority of queries (67.2%, n=281) requires G-VOILA to enrich responses by interpreting users' eye movements to identify objects of interest, focusing on either single (50.7%, n=212) or multiple objects (16.5%, n=69). The shared visual field also serves as another crucial input stream (22.2%, n=93) by providing either the context of the activity (e.g., "can I drink it now" in a supermarket) or the surroundings layout (e.g., "where to place my glasses" at home). Fewer queries (6.9%, n=29) draw on insights gained from other interaction systems, involving elements like history and user profiles or operating system assistance. These occasionally depend on G-VOILA's sensing ability, such as constructing a reliable and privacy-aware history profile based on gaze and egocentric views.

Take-aways: The majority of user queries anticipate G-VOILA to leverage gaze data to pinpoint objects of interest (67.2%) and visual data to supplement context information (22.2%).

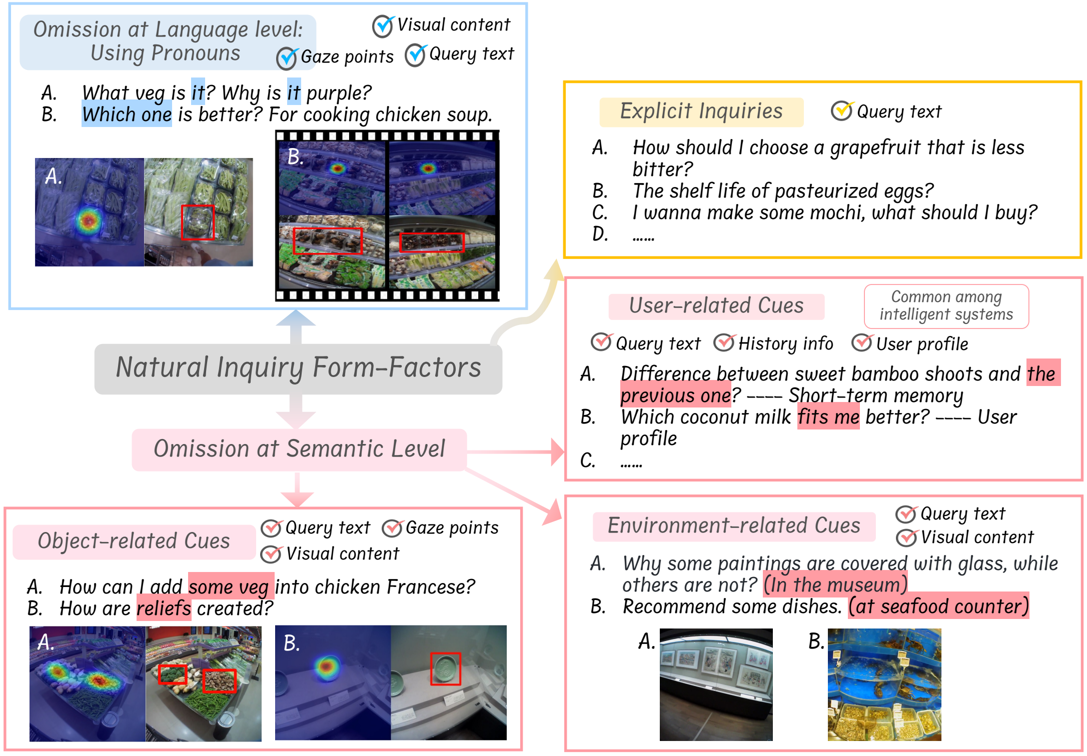

Omission at Language level: Using Pronouns
Visual content
Gaze points
Query text
A. What veg is it? Why is it purple?
B. Which one is better? For cooking chicken soup.
A.
B.
Explicit Inquiries
Query text
A. How should I choose a grapefruit that is less bitter?
B. The shelf life of pasteurized eggs?
C. I wanna make some mochi, what should I buy?
D. ……
User-related Cues
Common among intelligent systems
Query text
History info
User profile
A. Difference between sweet bamboo shoots and the previous one? ---- Short-term memory
B. Which coconut milk fits me better? ---- User profile
C. ……
Natural Inquiry Form-Factors
Omission at Semantic Level
Object-related Cues
Query text
Gaze points
Visual content
A. How can I add some veg into chicken Francese?
B. How are reliefs created?
A.
B.
Environment-related Cues
Query text
Visual content
A. Why some paintings are covered with glass, while others are not? (In the museum)
B. Recommend some dishes. (at seafood counter)
A.
B.

Fig. 5. Exploring Inquiry Form-Factors with G-VOILA. The identified categories were shown in each card, with a checkmark indicating the required data for refilling user's query intent. Keywords in the query text are highlighted, and representative frames are selected from each query video.

4.2 User's Verbal Expression Patterns with G-VOILA

We compared user's verbal questions with their reformulated queries textto discern language expression patterns with G-VOILA, as illustrated in Figure 5. We identified the following categories considering the diction and semantic gaps:

- **Explicit Inquiries.** Intelligent assistants could understand and respond to inquiries using only the verbal question and common knowledge. Natural language's ambiguity and the discrepancy between spoken and written expression present challenges, which can be mitigated by LLMs as evidenced by prior research.
- **Omission with Language Indicator – Using Pronouns.** Intelligent agents could identify the existence of ambiguity by examining the query text, often indicated by the usage of pronouns. By replacing the pronouns with referred objects, the agent can convert it into an explicit inquiry. Gaze data can clarify the referred objects, although deciphering a user's gaze pattern to pinpoint when they begin to look at the targeted items remains challenging.
- **Omission at Semantic Level.** Queries may seem complete linguistically but lack specific descriptive information, leading to a broader answer scope. Including:

– Object-related cues. Such omission may be discovered by aligning gaze content and query text, e.g. “relief on porcelain plates” and “some veg” referring to specific kinds of vegetables.
– Environment-related cues, where lack of context may yield answers unsuited to the current setting, e.g., suggesting painting placement in a museum instead of an art classroom or other locations.
– User-related cues. Since G-VOILA’s uniqueness lies in its gaze sensing and shared visual field, we designated other types of omissions that cannot be compensated by gaze and visual content as “user-related cues”, such as historical information or user preferences.

In accordance with this taxonomy, three researchers examined each query record and reviewed the corresponding video clips (overlaid with gaze data) to classify all queries, eventually reaching a consensus. Figure 3a displays the statistical results for each category.

Take-aways: The majority of user queries involve usage of pronouns (68%, 48%, 43%) and omission of certain constraint (20%, 40%, 39%).

4.3 Gaze Patterns and their Spatiotemporal Relationship to Spoken Query Language

We conducted quantitative analysis on the spatiotemporal correlation between users’ gaze patterns and their spoken questions, revealing a mouth-eye coordination when querying with G-VOILA, as depicted in Figure 6.

4.3.1 Data preprocessing. We leveraged fixation annotations $F : \{f_n\}$ provided by Pupil Labs, where each fixation f_i consists a tuple of $(t^i_{start}, t^i_{end}, x_i, y_i)$, representing the start and end times, and the fixation coordinates. Each spoken query $q \in Q$ is represented by $(t^q_{start}, t^q_{end}, W_q)$, with $W_q = \{w_{q1}, w_{q2}, ..., w_{qn}\}$ denoting the sequence of spoken words. Time segments T_q for each word $w_{qj} \in W_q$ were generated using [30, 43, 55], with $t_{qj} : (t^{qj}_{start}, t^{qj}_{end})$ specifying the timing of each word.

Fixations were labeled with it’s gaze content by manually inspecting video clips (overlaid with gaze trace and fixation points). We established a binary mapping function $g : F \times Q \longrightarrow \{0, 1\}$, where $g(f_i, q) = 1$ if the gaze content of f_i is relevant to query q, and $g(f_i, q) = 0$ otherwise. Relevancy was determined by whether the fixation’s content matched the expected answer or was mentioned in the query.

We mapped each word to sequential fixations using $a : W \longrightarrow F$, where $f_i \in a(w_{qj})$ if $[t^i_{start}, t^i_{end}] \cap [t^{qj}_{start}, t^{qj}_{end}] \neq \emptyset$. This function indicates the concurrent of word w_{qj} and fixation f_i.

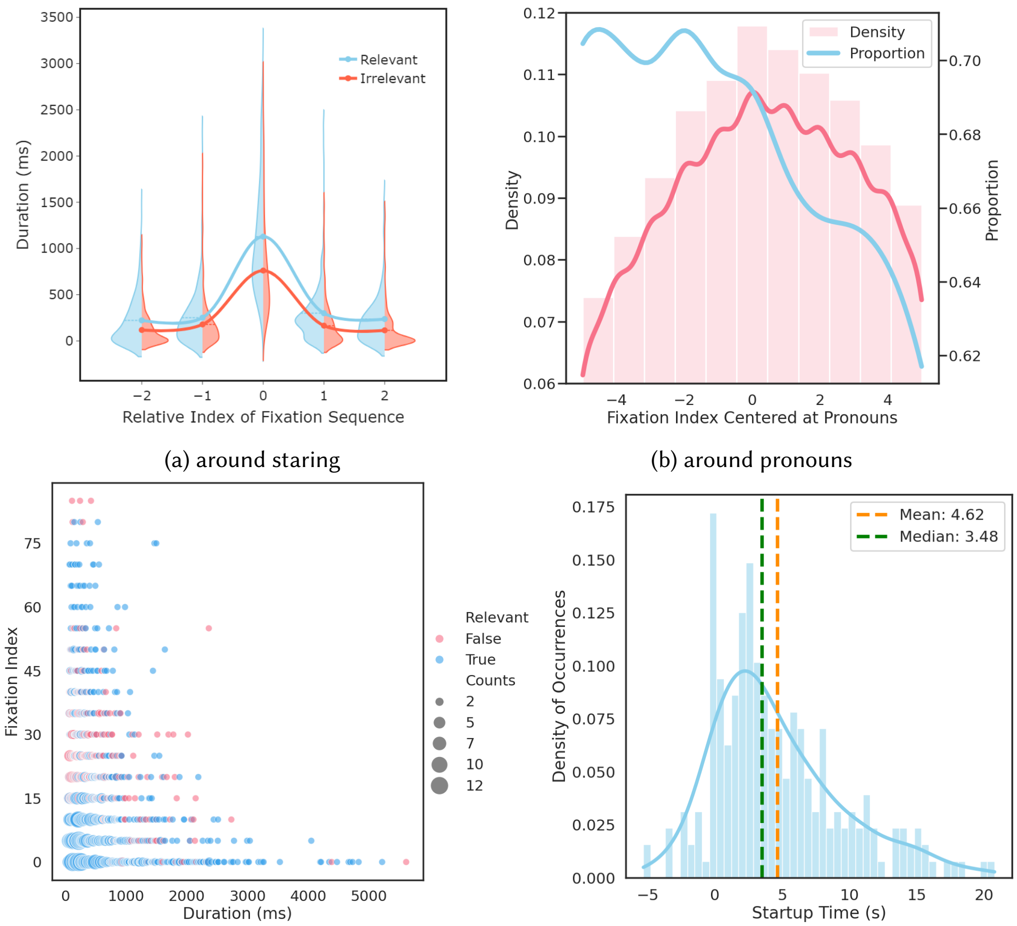
4.3.2 ***Users tend to focus more on relevant objects than irrelevant ones when conveying queries.*** We compared the longest relevant fixation to the irrelevant one. For each query q with fixation sequence $\{f_n\}$, we identified the indices of the longest relevant fixation f_i and the longest irrelevant fixation f_j, i.e.

$$i = \arg\max_i \{t^i_{end} - t^i_{start} \mid g(f_i, q) = 1\}, j = \arg\max_j \{t^j_{end} - t^j_{start} \mid g(f_j, q) = 0\}$$

Both i and j were denoted as relative zero index, then the distribution and mean duration of adjacent fixations were plotted in Figure 6a.

The results show that relevant fixation are significantly longer than irrelevant ones, indicating user’s propensity to concentrate on targeted objects rather than unrelated surroundings. This finding supports the idea that the increased focus on relevant objects may facilitate better understanding and communication of the query, as well as potentially indicate the user’s cognitive processing in relation to the question being asked.

Take-aways: Although users may have wandering gaze on irrelevant surroundings when expressing queries, they tend to focus longer on relevant objects($mean \approx 1.1s$) than irrelevant objects ($mean \approx 0.75s$).



(a) around staring

(b) around pronouns

(c) absolute distribution

(d) distribution of startup time

Fig. 6. Relevant fixation distribution. (a) A bar plot and approximation curve illustrating relevant fixation density centering around pronoun during querying. (b) A violin plot comparing duration of relevant and irrelevant fixation peak. (c) A scatter plot demonstrating fixation's occurrence order during querying. (d) A distribution plot of startup time for all querying.

4.3.3 ***Co-occurrence exists between using pronouns and looking at referred items.*** Pronouns w_{qj} were located within each query's word sequence W_q, with j_{pron} representing the pronoun's index. For each query's pronoun, we quantified relevant fixations on nearby words (within a window of $r \in [-5, 5]$ words) as

$$c(r) = Count\{f_i \mid f_i \in a(w_{q(j_{pron}+r)}), g(f_i, q) = 1, q \in Q\}$$

where r is the relative word position. Based on the calculation of $c(r)$, the density of relevant fixations with relative indices (r) was plotted as the "density" distribution in Figure 6b. Additionally, we calculated the proportion of relevant to total fixations around each pronoun as follows, and presented as the “proportion” plot in Figure 6b.

$$p(r) = \frac{c(r)}{c_{all}(r)}, c_{all}(r) = Count\{f_i \mid f_i \in a(w_{q(j_{pron}+r)}), q \in Q\}$$

The “Density” plot peaks at the pronoun’s index (relative zero index), indicating a higher occurrence of relevant fixations when users are speaking the pronoun. Conversely, the “Proportion” plot shows a decline following the spoken of pronoun. This suggests a shift in focus away from the referent to the next objects of interest or irrelevant stimuli, which happens right after the pronouns are spoken. This pattern may reflect a transition in cognitive focus post-referent identification.

Take-aways: User’s visual attention on relevant objects peaks at the mention of pronouns and then shifts away, indicating the existence of mouth-eye coordination but with gaze movements ahead.

4.3.4 ***Users tend to look at objects that are relevant to their query content at the beginning of the query.*** We draw a distribution plot of fixation relevancy $g(f_i, q)$ for all queries, in terms of its fixation index i and duration $t^q_{end} - t^q_{start}$, as illustrated in Figure 6c. As observed from the plot, blue dots representing relevant fixations dominate the lower left corner and tend to have larger radius at lower sequential index. This indicates a concentration of relevant fixations with longer durations in the initial stages of query formulation, revealing the fact that users are more focused on relevant objects when starting a query. The term “focus” here can be interpreted as either longer duration or higher frequency.

This trend aligns with the concept that “gaze often moves on before the last act is complete” [35], implying that user’s attention is more focused at the onset of a task. Furthermore, considering the next act, gaze moving on before the last act is complete means that gaze may move to the next act before it starts, which inspired us to ask the research question: Do users tend to look at items of interest before starting to ask questions? If so, how long?

Take-aways: Users demonstrate a pronounced focus on relevant objects at the beginning of query formulation.

4.3.5 ***Users tend to look at items of interest before starting to ask questions.*** We introduce the concept of “startup time” to denote the interval during which users visually engage with their query subject before verbalizing the question. Upon analyzing adjacent query intervals, we established a maximum threshold of 20 seconds for the startup time. For each query q, the startup fixation f_{i_s} and corresponding startup time $t^q_{startup}$ are determined by:

$$i_s = \arg\max_i \{t^q_{start} - t^i_{start} \mid t^q_{start} - t^i_{start} \le 20, g(f_i, q) = 1\}, t^q_{startup} = t^q_{start} - t^{i_s}_{start}$$

Figure 6d displays the distribution of startup times across all queries, highlighting an average startup time of 4.62 seconds and a median of 3.48 seconds. We further identified a behavior of “turning back”, where users briefly disengage from an object and then redirect their attention back to it when formulating a related query.

Take-aways: Users typically engage with objects of interest before initiating a related query, with an average start up time of 4.62s.

5 DESIGN FRAMEWORK FOR G-VOILA PARADIGM

In this section, we proposed a design framework for intelligent assistants within the G-VOILA paradigm. Our design framework tackles critical issues of incorporating gaze data into the query response process, considering

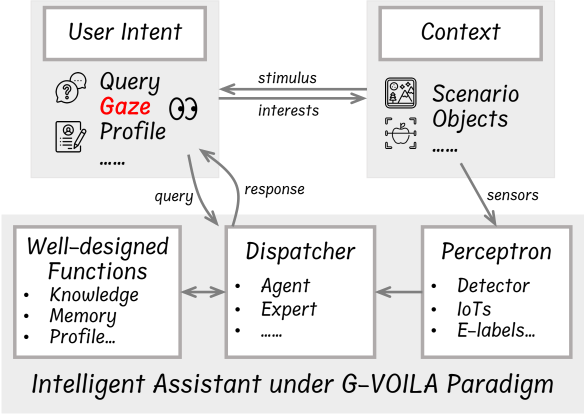


Fig. 7. A concept map illustrating the information flow between user, environment, and G-VOILA during one query.

(1) spatial indicators, (2) temporal relationships to the query expression, and (3) semantic connections between gaze content and query intent.

Building upon the user comprehension and expectations of G-VOILA (Section 4.1), as shown in Figure 7, we have delineated a modular envision for intelligent systems within the G-VOILA paradigm. This encompasses three core elements: the user, the surrounding context, and the system itself. The interaction procedure of conducting a single query can be decomposed as follow: triggered by a specific scenario, the user manifests interest in particular contextual information and subsequently initiates a query to the intelligent assistant; given the propensity for ambiguity in user's language expressions (Section 4.2), the intelligent assistant is tasked with deducing user's explicit query intention by leveraging information perceived from gaze trackers and egocentric camera (Figure 4, Section 4.3); meanwhile, the user's query may necessitate the activation of specialized functionalities (Section 4.1), where the intelligent assistant should dispatch the appropriate functions to efficiently address the query; by then, a satisfactory response can be delivered to the user.

The information divergence between the user and the intelligent assistant primarily stems from the assistant's lack of insight into (1)**when** the user become interested, (2)**what** sparked their interest, and (3)**how** their curiosity is specifically oriented. Hence, addressing this gap necessitates a joint analysis of G-VOILA's multi-modal sensory inputs to accurately deduce the user's query intent, which should be consistently reflected throughout the response mechanism. In light of this, as shown in Figure 8, we advocate a novel design principle for IR systems within the G-VOILA paradigm, which partitions the response process into two distinct phase: initially, multi-modal inputs is leveraged to discern and augment the user's intent; subsequently, responses are generated meticulously aligned with the refined understanding of the user's intent. Details will be discussed in the upcoming subsections.

5.1 Gaze-Facilitated: Deciphering In-Situ Query Intention

Contemporary text-based Information Retrieval (IR) techniques, including search engines and Large Language Models (LLMs), have demonstrated robust proficiency in addressing explicit query texts and mitigating ambiguities inherent in natural language. However, within the G-VOILA paradigm, **a novel challenge emerges: how to explicate user's ambiguous query expression through the integration of gaze and visual data?** To clarify the framework for deciphering in-situ query intentions, we dissect this process into three distinct components, each elucidated in detail. While these components are presented as separate entities for conceptual clarity, they can be, in practical applications, implemented by a single multi-modal model.

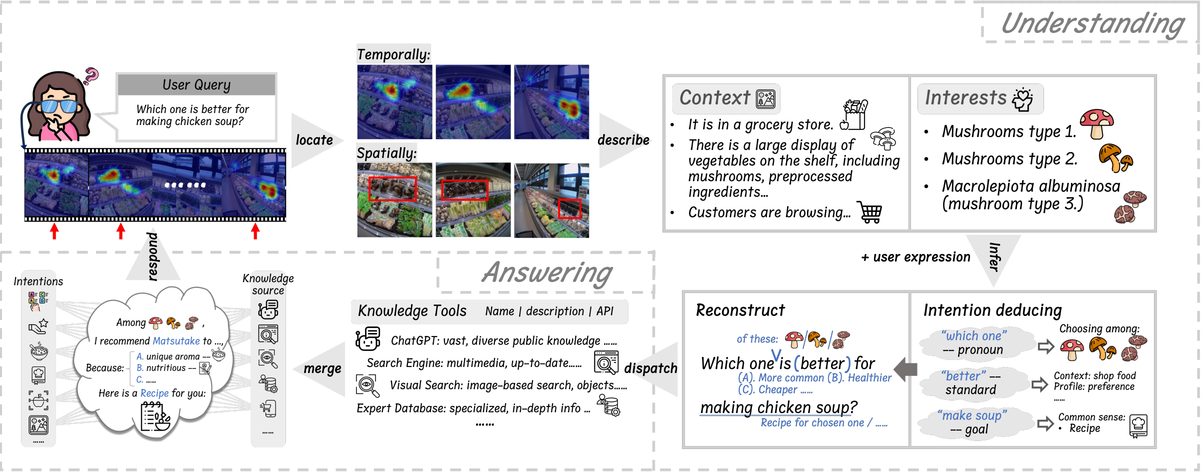

Understanding
User Query
Which one is better for making chicken soup?
locate
Temporally:
Spatially:
describe
Context
It is in a grocery store.
There is a large display of vegetables on the shelf, including mushrooms, preprocessed ingredients...
Customers are browsing...
Interests
Mushrooms type 1.
Mushrooms type 2.
Macrolepiota albuminosa (mushroom type 3.)
+ user expression
Infer
Reconstruct
of these:
Which one is (better) for
(A). More common (B). Healthier
(C). Cheaper ......
making chicken soup?
Recipe for chosen one / ......
Intention deducing
"which one" -- pronoun
Choosing among:
"better" -- standard
Context: shop food
Profile: preference
......
"make soup" -- goal
Common sense:
Recipe
dispatch
Answering
Knowledge Tools
Name | description | API
ChatGPT: vast, diverse public knowledge ......
Search Engine: multimedia, up-to-date......
Visual Search: image-based search, objects......
Expert Database: specialized, in-depth info ...
......
merge
Intentions
......
Knowledge source
......
Among
I recommend Matsutake to ...,
Because:
A. unique aroma --
B. nutritious --
C. ......
Here is a Recipe for you:
respond

Fig. 8. G-VOILA's inference and generate pipeline. Each step is indicated by a gray arrow and label.

5.1.1 ***Temporal and Spatial Localization of Interest.*** The data stream within the G-VOILA paradigm can be conceptualized as a chronological series of video and eye gaze data. This necessitates a meticulous cross-referencing approach, both temporally and spatially, to pinpoint elements within the three-dimensional scope of the video data (time dimension and two-dimensional image frames) that are pertinent to the user's query intent.

Discussion in Section 4.3 underscores the significance of temporal localization: **users are not always looking at their intended focus; instances of wandering gaze or moving on in advance are commonplace.** A poorly devoted temporal localization strategy risks emphasizing video frames that correspond to irrelevant gaze points. Such a misalignment can lead to a disastrous deviation from the user's original intent, given that gaze coordinates should be leveraged for spatial localization. Conversely, one distinctive feature of G-VOILA usage is the synchronicity between eye movements and verbal expressions. As corroborated in Section 4.3.3, **temporal alignment exists between the semantics of gaze content and the semantics of the user's spoken words**, which indicates a robust correlation between gaze and voice modalities. To conclude, the selection of the appropriate data frames for intent analysis should consider gaze pattern's intrinsic nature, as well as can be potentially guided by the timeline of spoken words.

Although gaze coordinates can indicate the spatial location of a user's interest (Section 4.3.2), they fall short in predicting the interest range. This limitation becomes particularly salient in scenarios where multiple objects are spatially nested. In the case where user's gaze is directed at a flower within a vase on a shelf, the specific object of interest becomes equivocal: the flower solely, the flower-vase ensemble, or the shelf. Our user-enactment study occurred several instances of such nested structures within the user's field of vision. A notable example involved a user standing in front of a pile of water tank filled with all sorts of seafood and querying about a specific kind of fish.

To address this ambiguity, we propose two indicative cues for disambiguating the precise range of interest. **The first cue is the user's linguistic expressions, which can used to align with the semantic of items at different granularity.** For example, a request like "recommend a dish for this" would most likely pertain to a specific type of fish, whereas "recommend a dish" might refer to either a category of seafood or a specific variety. **The second cue is enhanced robustness by analyzing the gaze trace over a brief duration.** In this context, a broad sweep across the water tank would suggest an interest in the general category of seafood, while a relatively steady gaze on a specific section might indicate interest in a particular fish variety.

5.1.2 Semantically Understand Visual Context and Gaze Content. In the pursuit of offering a precise response to a user's query, a system must primarily grasp the visual context, encompassing elements like the background and location. One effective approach to achieving this is by imbuing descriptions with an appropriate level of semantic understanding. Simultaneously, this necessitates the system's capability to differentiate between coarse and fine levels of semantic granularity. A coarse semantic description might simply categorize an object as a "plate", while a fine-grained description would delve into its material, color, pattern, and distinctive attributes, such as labeling it as a "ceramic plate adorned with a relief of a dragon". The depth of semantic understanding should harmonize with the nature of the question being addressed. In essence, this constitutes a multimodal understanding challenge that is independent of our paradigm design. Every enhancement in this domain can seamlessly integrate into our paradigm, resulting in improved performance.

5.1.3 ***Deduce Intention and Reconstruct Query Expression.*** Guided by the taxonomy established in Section 4.2, query expressions that inadequately convey the user's intended meaning typically suffer from two deficiencies: the utilization of ambiguous terms, such as pronouns, or the exclusion of pivotal information with no discernible hints in the expression. Addressing the former necessitates the substitution of ambiguous language with contextual information acquired from alternative modalities, as elaborated upon in Sections 4.3.2 and 4.3.3, with a particular focus on gaze data. Conversely, resolving the latter challenge involves a strategic integration of additional information, sourced either from perceptual inputs or user-specific contexts. For both cases, the deployment of a meticulously formulated mechanism described in Section5.1.1, is essential to provide dependable cues that reflect the user's actual intentions.

5.2 Intention-Oriented: Constructing Query Response

Upon restoring the user's query intent, the ensuing task is to formulate a response that is congruent with this discerned intent. Considering the widespread adoption of gaze-facilitated VR in tourism [76], it is plausible to envision IR systems developed within G-VOILA paradigm could be effectively applied for such use cases that requisite professional knowledge. Moreover, a small segment of users has expressed their anticipation for G-VOILA to incorporate various features inspired by other well-established systems, as shown in Section 4.1. In response to these insights, we advocate for a bifurcated approach in guiding system design, which encompasses an open design space for mounting such desired functionalities.

5.2.1 Tools-Driven Knowledge Retrieval. While addressing the professional use cases mentioned above necessitates specialized knowledge domains, everyday scenarios also demand specific expertise. For instance, in a shopping context, comparing prices to online products requires access to an e-commerce database. Thus, we propose that a mature system within the G-VOILA paradigm should incorporate a module of diverse functionalities and a corresponding dispatcher. A particularly notable function is short-term memory, which, when integrated with G-VOILA's unique gaze modality, distinguishes it from similar functions in other paradigms. The interaction between gaze data and this short-term memory module—specifically, the criteria for triggering data recording and the format of such records—presents a complex and independent research topic, as well as the privacy issue raised by its always-on property.

5.2.2 Intention-Guided Information Integration. Upon dispatching specialized functionalities, a critical subsequent step involves refining the output through a process of filtering and rephrasing, tailored to align with the user's specific query intentions.

In the development of our comprehensive design framework, we adhered to two fundamental principles that consistently guided every aspect of it. The first principle is a steadfast focus on the user's intention. This user-centric approach ensures that every component of the system is designed and optimized to cater to the user's specific needs and goals. The second principle is the utilization of gaze tracking as a primary tool for uncovering user's query intentions, which is a unique feature introduced by G-VOILA paradigm. By leveraging gaze data, the system gains a nuanced understanding of what the user is focused on, thereby providing insights into their underlying intentions and interests. These principles are not just theoretical guidelines but are deeply integrated into the framework's architecture, influencing everything from understanding questions to answering questions.

6 A PRELIMINARY IMPLEMENTATION

The aforementioned analysis based on a user-enactment study (Section 4) already provides abundant insights for a system design framework within the G-VOILA paradigm. However, further implementation of a functional system (henceforth referred to as VOILA-G) in a user study could offer additional understanding of user interactions with such a system. More importantly, an evaluation study is instrumental in proving the concept of integrating gaze tracking in the G-VOILA paradigm and the effectiveness of our proposed design framework.

6.1 VOILA-G Implementation

VOILA-G[4] collects user query data using Pupil Labs Invisible eye-tracking glasses and posts requests to a remote server equipped with NVIDIA V100 for processing responses. As shown in Figure 13, users can initiate queries by speaking to VOILA-G, with the audio being transcribed using the Whisper [55]. The local gaze-based temporal localization process computes key video frames, which, along with the query and gaze data, are sent to the remote server. Responses, comprising reconstructed queries, direct answers, and inferential thoughts, are returned and presented on a computer screen for users to evaluate.

We use SAM [34] to segment users' interest objects at various scales according to gaze points. Caption generation is handled by KosMos-2 [48], while OWL-ViT [44] is used for object detection. KosMos-2 aids in depicting the overall context. Leveraging the advanced capabilities of large language models (LLMs) in creating natural language-based agents, we selected GPT-4 from the ChatGPT API for its exceptional comprehension, inference, and instruction-following abilities.VOILA-G's prompt structure is based on the instruction-examples format, elucidating implicit inquiry form-factors (Figure 5) within the instruction part and further exemplified. Prompts can be found in Appendix A, detailed implementation specifics can be found in Appendix B.

6.2 Baseline Implementation

To reiterate, the distinctiveness of both the G-VOILA paradigm and VOILA-G's implementation stems from the incorporation of gaze information. Consequently, our primary comparisons involve systems that eliminate the influence of gaze. In VOILA-G, gaze has been instrumental in reconstructing user query intentions, specifically for temporally and spatially pinpointing users' interests. We implemented three baseline systems in an ablation manner, described as follows:

(1)**VOILA.** In this system, we eliminate gaze input. The gaze-driven spatial localization process is replaced with a confident object detection result, while the gaze-driven temporal localization process is substituted by randomly selecting unblurred video frames.

(2) **VOILA-T.** In this system, we retain the gaze-driven temporal localization process compared to VOILA. VOILA-T served as a baseline throughout our evaluation study.

(3) **VOILA-S.** In this system, we keep the gaze-driven spatial localization process compared to VOILA.

We further considered two alternatives to the incorporation of gaze:

(4)**VOILA-center.** In this system, we consider user's facing direction as their interest indicator, which is the center of camera's visual field.

[4] Open source code at: https://github.com/Sky-Wang326/gvoila.git

Supermarket shopping		**Domestic living**	
Task	Description	Task	Description
Comparison	Query for details to make a choice among several products	Appliance Malfunction	Seek solutions to rectify the malfunction
Completing Recipe	Figure out how to incorporate an ingredient into a cherished dish	Activity & Health	Get advice for certain health concerns or activities
Recommend	Dishes with ingredients in front of you	Snack & Fruits	Choose from snacks and fruits for a specific purpose
Knowledge	Query for a specific attribute of a product	Dressing Advice	Pick out the right outfit for tomorrow's plan
Decision Making	Request VOILA-G to decide whether to buy a product based on certain criteria	Entertainment	Recommend recreational activities at home
Strengthen Decision	Request VOILA-G to strengthen your resolve to resist a certain temptation	Small Talk	Suddenly desire intriguing information about a specific object

Table 1. Guiding Tasks for Evaluation Experiment

(5)**VOILA-saliency.** Considering the prevalent of saliency prediction techniques, where user's gaze trace on an image can be predicted by a deep learning model, we incorporated a saliency model to substitute gaze data. Implementation details can be found in the Appendix C.

7 EVALUATION STUDY

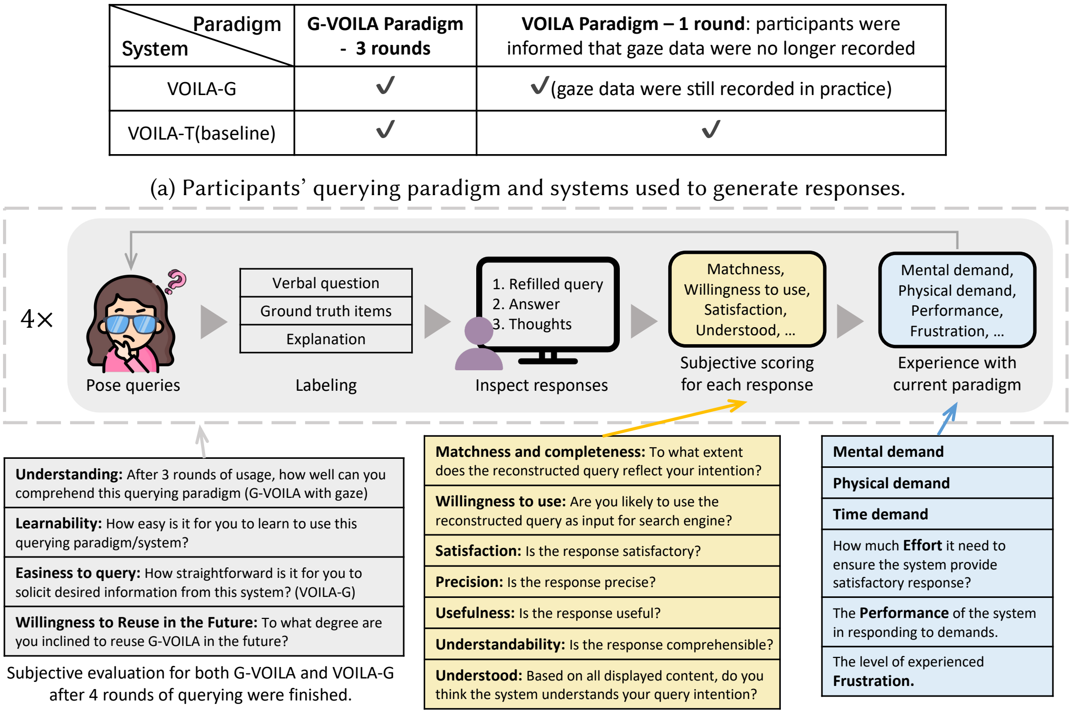
With the preliminary implementation of VOILA-G, we designed an evaluation study to prove the effectiveness of integrating gaze and the guideline of such integration proposed in the design framework.

7.1 Evaluation Study Design

7.1.1 Scenarios and Participants. The user study involves two daily-life scenarios: supermarket shopping and domestic living. We excluded the museum visiting scenario, previously considered in the user-enactment study, due to the challenges in integrating a specialized database, which we were not able to acquire. As described in Table 1, six tasks for each scenario were predefined, drawing upon the queries identified during the formative study. We designed these tasks with a twofold purpose: (1) To encourage participants to ask questions requiring varied applications of gaze tracking and egocentric views, as exemplified in Figure 4. For instance, the "Comparison" task prompts participants to engage with multiple items simultaneously. (2) To offer a structured guideline for participants, assisting them in formulating queries, particularly in instances where they might struggle to think of questions.

We included 16 participants (8 for each scene and have no overlap with previous enactment study), comprising 8 males and 8 females, aged from 20 to 30 (with a standard deviation of 2.06). Each participant required approximately 240 minutes to complete the study, and they received a compensation of 30 USD per hour for their participation.

7.1.2 Procedure. Similar to the user-enactment study, the evaluation study comprises three sessions: an introduction session, a querying session in which participants engage in four rounds of using and evaluating, and a subsequent interview session. The introduction session follows the same format as the user-enactment study. We emphasized VOILA-G's uniqueness of gaze tracking compared to other querying paradigms and encouraged participants to actively try out various phrasing strategies that can allow them to express themselves in a more convenient manner.



Paradigm System	**G-VOILA Paradigm - 3 rounds**	**VOILA Paradigm – 1 round**: participants were informed that gaze data were no longer recorded
VOILA-G	✓	✓(gaze data were still recorded in practice)
VOILA-T(baseline)	✓	✓

(a) Participants' querying paradigm and systems used to generate responses.

Understanding: After 3 rounds of usage, how well can you comprehend this querying paradigm (G-VOILA with gaze)

Learnability: How easy is it for you to learn to use this querying paradigm/system?

Easiness to query: How straightforward is it for you to solicit desired information from this system? (VOILA-G)

Willingness to Reuse in the Future: To what degree are you inclined to reuse G-VOILA in the future?

Subjective evaluation for both G-VOILA and VOILA-G after 4 rounds of querying were finished.

Matchness and completeness: To what extent does the reconstructed query reflect your intention?

Willingness to use: Are you likely to use the reconstructed query as input for search engine?

Satisfaction: Is the response satisfactory?

Precision: Is the response precise?

Usefulness: Is the response useful?

Understandability: Is the response comprehensible?

Understood: Based on all displayed content, do you think the system understands your query intention?

Mental demand

Physical demand

Time demand

How much **Effort** it need to ensure the system provide satisfactory response?

The **Performance** of the system in responding to demands.

The level of experienced **Frustration.**

(b) Evaluation pipeline for each round of usage. Repeated for 4 rounds.

Fig. 9. Procedure of evaluation study.

In the querying session, each participant takes part in four rounds of querying, as depicted in Figure 9a. Within the continuous three-round usage of the G-VOILA paradigm, participants are made aware of G-VOILA's ability to capture eye movements and are encouraged to experiment with different strategies for phrasing their questions. Simultaneously, participants are required to utilize the VOILA paradigm for one round. The VOILA paradigm is a modified version of the G-VOILA paradigm, with the gaze modality eliminated. Therefore, during the one-round usage of the VOILA paradigm, participants are informed that gaze data are no longer recorded; however, gaze data are still collected to support the response process of VOILA-G. The order of different paradigms is counterbalanced among participants.

The evaluation pipeline for each round is illustrated in Figure 9b. At the beginning, participants are asked to pose one query for each predetermined task. Subsequently, they are instructed to identify the objects they are inquiring about, which are later employed as ground truth to objectively calculate response recall and precision. For each query, both VOILA-G and VOILA-T(baseline) provide responses consisting of (refilled query expression, answer, thought). After examining the responses from both systems, which are displayed anonymously in a randomized order, participants are requested to subjectively rate the responses from both systems, with the understanding that waiting time should be disregarded. Finally, participants are required to rate their experience of querying within the current paradigm. After the inspecting and scoring were done for current round, participants can move on to the next round of querying. Such "querying, labeling, checking answer, scoring" process will be repeated for 4 rounds. With such repetition, participants may become more familiar with G-VOILA paradigm and VOILA-G system.

		VOILA-G	VOILA-S	VOILA-T	VOILA	VOILA-center	VOILA-saliency
Recall	**All***	89.1%	83.84%	79.25%	76.91%	78.44%	78.76%
	Explicit	97.28%	94.12%	94.12%	93.6%	92.81%	93.86%
	Ambiguous*	80.11%	72.54%	62.91%	58.57%	62.67%	62.19%
Precision	**All***	83.83%	78.4%	69.83%	67.62%	71.8%	72.86%
	Explicit	95.79%	92.73%	93.44%	91.23%	91.6%	93.03%
	Ambiguous*	70.7%	62.65%	43.9%	41.7%	50.06%	50.7%

Table 2. Objective Response Evaluation. Recall rate and precision score for both VOILA-G and the baseline models. Queries related to specific objects in the scenario are classified into explicit and ambiguous categories based on whether the user mentions the target object's name in their query expression. The Wilcoxon signed-rank test was employed for all comparative scoring, with significant differences denoted by **bold*** labels. We discussed the results of VOILA-saliency in Appendix C.

In the interview section, participants were asked to respond to the same questions posed in the subjective evaluation of the querying paradigm, along with some additional inquiries as outlined in Figure 9b's list (bottom-left).

7.1.3 Evaluation Metrics. For the objective evaluation, we employed recall score and precision score. The ground truth objects were identified by participants in each round of usage. We recognized items emphasized in the responses as predicted values. Given that the predicted values are in open-vocabulary, we cannot calculate the objective accuracy score.

Participants rated the subjective metrics on a 7-point Likert scale. For each query response, the metrics are outlined in Figure 9b's list (bottom-middle).

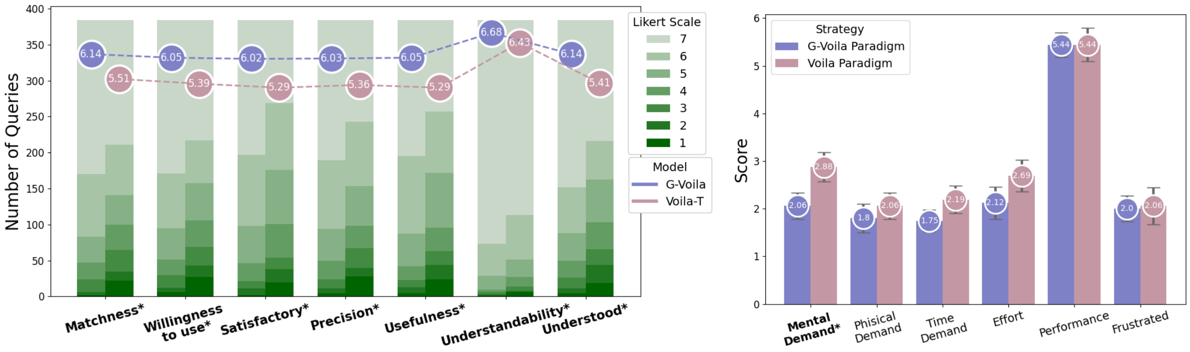
For the experience with the querying paradigm, the metrics are outlined in Figure 9b's list (bottom-right).

7.2 Results and Analysis

7.2.1 Objective Response Score. With labeled querying objects for each question, we calculated the recall rate and precision for VOILA-G and the baseline. We examined whether the key objects in the ground truth label appeared in the response and identified semantically parallel objects as detected keywords. We excluded queries that were not related to specific content in the scenario, then categorized the remainder into **Explicit Query** and **Ambiguous Query**based on whether key object names were spoken in the query. Table 2 presents the average recall and precision for each category. VOILA-G statistically outperforms VOILA in ambiguous queries(recall: $Z = 184.5$, $p < .001$; precision: $Z = 497.5$, $p < .001$) and all queries(recall: $Z = 295.5$, $p < .001$; precision: $Z = 873.0$, $p < .001$), while slightly surpassing VOILA in explicit queries(recall: $Z = 14.0$, $p = .013$; precision: $Z = 49.0$, $p = .002$).

Our ablation study indicates that using gaze as a spatial cue substantially improves VOILA-G's performance. Notably, when leveraging gaze solely as a spatial indicator, incorporating its temporal properties further refines VOILA-G's ability to comprehend ambiguous queries (from 72.5% to 80.1%). An improvement in both recall and precision score (VOILA-G vs. VOILA-T, VOILA-S vs. VOILA) demonstrates the efficacy of integrating gaze's spatial localization process. However, (VOILA-G vs. VOILA-S)exhibits a more substantial performance growth than (VOILA-T vs. VOILA), which suggests that the advantage conferred by incorporating gaze's temporal localization process is more pronounced when gaze serves as a spatial indicator.

The precision improvement (29%) with VOILA-G is notably higher than recall enhancement (22%), especially for ambiguous queries. This suggests that VOILA-G effectively leverages gaze data to discern user interest among multiple potential query-related items in the visual field. For example, the implemented systems may occasionally



(a) Subjective scoring for queries

(b) Subjective scoring for strategies

Fig. 10. Subjective Scoring Analysis. The Wilcoxon signed-rank test was utilized for all comparative scoring, with significant differences indicated by **bold*** labels. (a) A stacked bar plot representing each query's score, where a higher value signifies better performance. (b) A bar plot comparing gaze and no-gaze querying strategies, with lower values indicating superior results (excluding the performance bar).

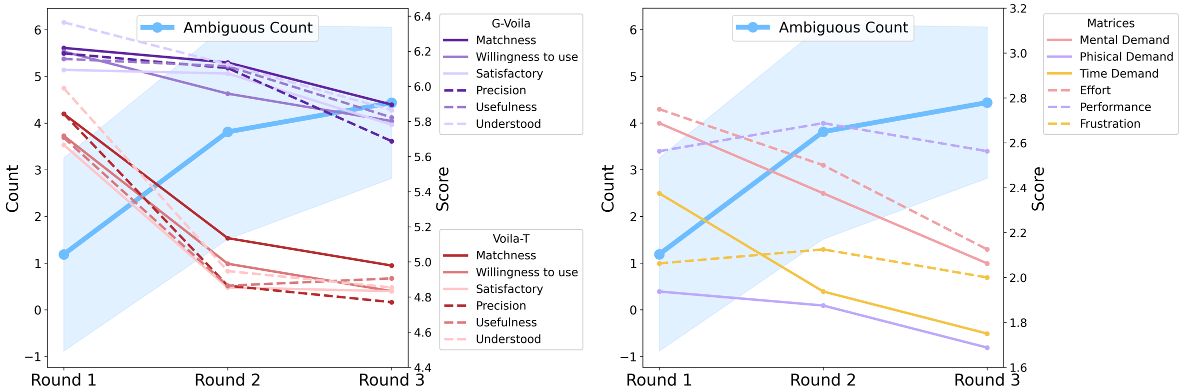
include extra items they presume to be of users' interest for explicit questions. Furthermore, when dealing with highly ambiguous queries, such as simply stating "calorie content" without using restrictive pronouns like "which one between these two", baseline systems may enumerate the calorie content for all detected food in the visual content. However, VOILA-G can pinpoint objects of interest to reduce the occurrence of both issues.

VOILA-center shows enhanced performance compared to VOILA, indicates the correlation between eye and head movements, with the latter serving as a reasonable proxy for inferring user intent. Nevertheless, head direction is a less precise indicator of intent compared to gaze (VOILA-center vs. VOILA-S). See Section C for a detailed discussion on VOILA-saliency results.

Take-aways: Incorporating gaze as a temporal and spatial indicator can both enhance the ability to discern user's interest. A synergistic catalytic effect exists when incorporating both temporal and spatial properties. Head direction can be used as a less precise indicator of user's intent. Existing saliency models fall short for G-VOILA's use case (see Appendix C).

7.2.2 Subjective Score for Queries. The subjective score for VOILA-G and VOILA-T, as presented in Figure 10a, indicate a statistically significant preference for VOILA-G via a Wilcoxon signed-rank test across all metrics. Participants favored VOILA-G's query reformulations, citing a higher concordance with their intended meaning and a stronger propensity to use these queries as text-based search input. VOILA-G was also rated higher in satisfaction, utility, and perceived accuracy over the baseline. However, the understandability of VOILA-G's outputs showed only a slight increase, as LLMs are already adept at producing intelligible text, which gaze data can contribute minimally to. Overall, when evaluating answers and thoughts collectively, participants felt that VOILA-G more effectively captured their query intentions with the aid of gaze data.

7.2.3 Subjective Score for Paradigms. To ensure impartial evaluation, the order of response presentation from both models was randomized and anonymized. Participants rated the used paradigm for each round based on their usage experience during querying and displayed responses. We selected the score for the final round of applying G-VOILA paradigm compared to the only round of applying VOILA paradigm, as shown in Figure 10b. Participants'



(a) Ambiguous count and query score

(b) Ambiguous count and round score

Fig. 11. User's scoring changes of three-round usage of G-VOILA. Both graphs show ambiguous query count for each round. (a) illustrates how 6 questions' average scoring changes for each evaluation matrix. (b) illustrates scoring change for G-VOILA's usage. The "performance" matrix is subtracted by 8 to better fit in the plot.

preferences for the two strategies varied individually, with G-VOILA holding an overall advantage. Subsequently, we conducted interviews with each participant, and the results thereof were discussed in Section 7.3.1.

7.2.4 Three-round Usage of G-VOILA paradigm. Figure 11 shows the trajectory of participant ratings over three rounds usage of G-VOILA paradigm. Following each round, our researchers communicated with the participants to facilitate a better understanding of G-VOILA's functionality and encouraged them to express queries more naturally, in accordance with their daily habits.

Figure 11a displays the ambiguous question count and the average query scores for both models across the three rounds. We omit understandability from the analysis due to its insignificance. A trend emerges where participants, growing accustomed to G-VOILA, posed their queries with increased ambiguity. The heightened ambiguity had a minimal impact on VOILA-G's performance compared to baseline, benefiting from gaze-derived interest information.

Figure 11b illustrates the ambiguous question count and scores for each round's overall experience with G-VOILA. It should be noted that the original rating for performance was higher-is-better; however, to better fit the plot, we modified the performance as $8 - x$. As participants became more familiar with G-VOILA and used more arbitrary expressions, they tended to rate G-VOILA higher in all aspects. This finding aligns with the learning process typically observed with the adoption of any convenient new technology.

Take-aways: As users became more familiar with G-VOILA, they increasingly leveraged its capabilities by posing more ambiguous queries, which have lower performance degradation on VOILA-G and better user experience.

7.3 Feedback and Interviews

In the interview, participants were asked about their rating criteria and thoughts on each metric when evaluating G-VOILA usage. They also provided overall user experience scores and explanations.

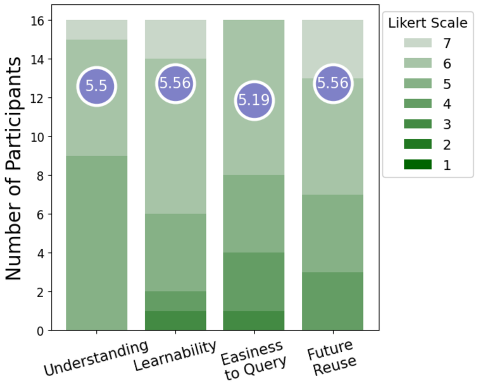


Fig. 12. Overall Evaluation. A stacked bar plot illustrating participants' scores on a Likert scale regarding their overall experience with G-VOILA paradigm and VOILA-G, as reported in the follow-up interview. In this plot, a higher value signifies a better result.

7.3.1 Rating for Each Round. Participants were interviewed about their experiences with each round usage of G-VOILA, covering cognitive, physical, and time demands, effort, performance, and frustration.

Most participants regarded effort as a combination of cognitive, physical and time demands, which was reduced as they adapted to the querying style during the experiment (or in potential long-term use) and formed a habit. Compared to the traditional text-based query approach, participants often cited two advantages: (1) simplified queries due to casual conversational and omitted information, and (2) on-demand querying without extra hand operations, such as taking out a device or operating software. These factors directly led to lower demands. However, P14 noted that, compared to search engines that provide multiple options, G-VOILA and chatbot systems offer more targeted answers, which may increase cognitive pressure to ask questions precisely. Meanwhile, P10, P11, and P14 found that controlling their eye movements physically demanding, while others regarded gaze as a natural behavior.

Similarly, performance is considered a combination of effort and response accuracy, while frustration is related to expectation and misidentification. The response accuracy of VOILA-G dropped slightly when queries became more ambiguous and participants adjusted their expectations as well. P7 and P13 stated that they were expecting delightful surprises brought by unexpected associations rather than overly focusing on context information for those questions where VOILA-G successfully identified their fixation content. Furthermore, P15 mentioned that **G-VOILA**'s attempts to combine context and provide more comprehensive answers could alleviate frustration when misidentification occurred.

7.3.2 Rating for G-VOILA paradigm and VOILA-G. As shown in Figure 12, participants exhibited a positive preference for VOILA-G. The majority of participants reported that after 2-3 rounds of use, they were able to better understand VOILA-G's capabilities and limitations, as well as master the collaboration between eye movements and query expression. Their inquiry style became more aligned with colloquial and daily expressions, which differs significantly from that of commonly used search engines, making it relatively easy to form a question. P3 and P10 mentioned that bearing the risk of incorrectness might increase the difficulty of asking questions. P14 noted that information is substantially fully presented to VOILA-G in either natural language or gaze behavior, providing a choice for output channels.

Due to G-VOILA's on-demand nature and its ability to alleviate the burden of daily questioning, most participants expressed a high willingness to reuse it in the future, yet they had certain expectations for its development.

P16 mentioned that VOILA-G's advantage lies in its ability to comprehend spatial knowledge, and he hoped that this capability could reach the same level as ChatGPT's text comprehension ability. Furthermore, P3 expressed concerns about the popularity of head-mounted devices, as VOILA-G relies on such a platform. Other suggestions beyond gaze and vision were given, including the creation of user profiles or integration with smart home control systems.

8 DISCUSSION AND FUTURE WORK

In this section, we discuss G-VOILA's major results, findings, and limitations. We also provide insights for future work.

8.1 A Summary and Discussion of Findings in User-Enactment Study

8.1.1 Ambiguity: Reasons for users omitting certain information in their query expressions. The taxonomy introduced in Section 4 indicates various categories of ambiguity that may arise during the query formulation process with G-VOILA. Statistical results, as depicted in Figure 3a, reveals a significant incidence of ambiguous queries, with a notably high frequency of pronoun usage exceeding our initial expectation. We further discuss several underlying factors that this tendency to generate queries lacking explicit environmental context and precise intent may stem from.

(1) **It is time-consuming or cognitively demanding to express certain information.** When users are unfamiliar with the precise name of an item, they tend to use pronouns and object categories as a way to describe it. For instance, in a museum setting, they might refer to a ceramic piece as "this porcelain" instead of reading the name on the exhibit label. Similarly, when referring to multiple items, users prefer to use pronouns to indicate the range of objects they are comparing. For example, in a supermarket, they might say "these few" when comparing different types of fish rather than listing specific names.
(2) **Users struggle to create well-organized expressions for their inquiry intent.** In some cases, users may not know what the target object is and can only use pronouns to refer to it. A kitchen novice wanting to know how to eat red amaranth might describe it as a spinach-like vegetable with purple meridians on the leaf, without knowing the proper name. Additionally, users may associate the inquiry target with existing knowledge, leading to the use of incorrect or inaccurate vocabulary in their descriptions. For instance, when asking how a "relief" is made on a porcelain plate in a museum, it is confusing because traditionally understood relief is on architects or wooden artifacts.
(3) **Users become immersed in a situation and neglect to describe certain constraint information.** When looking for activities at home, users may overlook keywords like "indoor" but still expect to receive suggestions suitable for their situation, such as not wanting outdoor activities like skiing or swimming. In another scenario, users may not be aware that their current situation imposes constraints on the information retrieval process. An ESL user in a museum, for example, might select the most appropriate definition of an unfamiliar English word on an exhibit label based on the context when using a dictionary without realizing the importance of the context in their search.

8.1.2 Eye-movement pattern throughout each querying and it's coordination with mouth. Through rigorous quantitative analysis, we have delineated several patterns that are congruent with established gaze-related behaviors during task execution, as well as novel patterns specific to G-VOILA. Users tend to focus more on the relevant objects than irrelevant ones when conveying queries, whereas the distribution of this relevancy does not remain uniform throughout the querying process. Analogous to the phenomenon where "gaze often moves on before the last act is complete" [35], users initially direct their gaze toward objects that are pertinent to the content of their query. As the query progresses, we observe a greater deviation in gaze towards the end.

Conversely, our findings indicate that users typically shift their attention to items of interest prior to initiating their queries, with a pre-emptive focus averaging 4.62 seconds. Most notably, a distinctive gaze pattern associated with the querying task is the existence of eye-mouth coordination, where users are most probably looking at their intended objects before the pronouns were spoken.

Beyond the general quantitative analysis, our researchers have noted a "turning back" behavior among most participants when annotating data. This behavior is characterized by users redirecting their gaze to inquire about objects they had previously disregarded. Additionally, variations were observed in participants' behaviors at the conclusion of their queries. While the majority of participants remain stationary until the entire question is articulated, a subset exhibits a tendency to move away promptly after formulating their question.

8.2 Response Precision and User Experience

The precision and recall metrics for VOILA-G when addressing ambiguous queries, as depicted in Figure **??**, do not achieve a production level. This shortfall is primarily due to the reliance on open-source models for constructing the visual system of VOILA-G without much engineering effort. The limitations are twofold: (i) the performance of the visual understanding models does not meet the product standards for real-world applications, and (ii) there is a discrepancy between the distribution of the training dataset (internet-sourced images) and our operational environment (snapshots from egocentric videos). Notably, approximately half of the study participants (n=7) expressed potential frustration with instances where VOILA-G erroneously identified objects. This necessitates a discussion on the potential impact of suboptimal performance on user experience and the interpretation of experimental results.

However, VOILA-G has proven its conceptual validity by successfully integrating gaze tracking, as evidenced by both objective and subjective evaluations presented in Table 2 and Figure 10a. Participants rated VOILA-G more favorably on all subjective measures compared to the VOILA-T baseline with statistically significant differences. This suggests that even with room for performance enhancement, VOILA-G effectively serves as a proof-of-concept prototype.

8.2.1 User Preferences and Trust in Response Precision. Participants showcased divergent preferences concerning the precision of responses to similar questions. When inquirying about objects, VOILA-G sometimes provide answers that exceeded the scope of the user's intention, particularly for ambiguous queries. Some users had higher expectations for a precise response, while others demonstrated higher tolerance for redundant replies, as long as the target information was prominent and could be quickly located. For example, Participant 13 states his/her preference for more divergent answers, rather than being limited to the specific question and visual content.

One possible factor revealed in our afterwards interview is that individual's information retrieval habits may influence their expectations for precision. Participant 11, accustomed to converstional interface like ChatGPT, paid more attention in phrasing queriesto elicit precise responses. Conversely, Participant 15, who predominantly uses search engines, feels satisfied with a broader explanations provided by VOILA-G when it attempts to encompass his/her intention more fully.

Moreover, the expression style users presented in their final round of querying was related not only to their personal linguistic style but also to their trust in the system's capabilities. Some participants abandoned the use of complete sentences and resorted to terse phrases such as "vitamin content", "want to exercise abs", and "eyes feel dry". In subsequent interviews, these participants had high praise for VOILA-G's eye-tracking and visual abilities, indicating a correlation between their trust in the system and their expression style.

8.3 Future Deployment Potential

VOILA-G relies on large-scale computer vision and language models, and as a result, is subject to their limitations. Firstly, although large language models have demonstrated impressive understanding capabilities, the text generation speed of the API is the primary factor contributing to waiting times, as mentioned by participants in their feedback. Secondly, the understanding capabilities of computer vision models have not yet reached a high level, which limits the accuracy of our system in recognizing objects that users gaze at. Moreover, there is still a discrepancy between the training datasets of currently available large-scale vision models and real-world ego-centric data, such as the recognition of fruits in a supermarket being significantly affected when wrapped in plastic film. Furthermore, to provide services in high IR demand scenarios, such as museums, integration with expert knowledge is necessary. Last but not least, the modules employed in this article to handle gaze, vision, and text modalities are relatively independent from each other, resulting in insufficient accuracy in spatially and temporally locating objects of interest to users and occasionally leading to deviations in understanding user intent. To better integrate these data, a multi-modal large model with excellent performance is needed.

8.4 Beyond Gaze and Voice

G-VOILA currently relies on gaze and voice input for query-based information retrieval, whereas we have discovered other data modalities that might help for natural querying expression. One typical data modality is gesture, which has been widely used in Human-Computer Interactions and can also serve a pointing purpose. Activity data captured by the inertial motion unit (IMU) may indicate the user's movements and head motion, such as stopping at some place, may indicate an increase of interest and can also help to stabilize gaze points. Moreover, facial expressions can also reveal the user's attitude towards currently looking content. We expect to add more insightful data modalities to G-VOILA in the future.

9 CONCLUSION

To conclude, we first envisioned a future information querying paradigm, namely G-VOILA, which combines user's gaze data, visual field and voice-based natural language queried. We revealed users inherent ambiguous and colloquial phrasing manner when querying with the G-VOILA paradigm in a user-enactment study. Based on quantitative analysis of collected data, we discovered a mouth-eye coordination in users' expression behavior. Inspired by this study, we proposed a design framework for G-VOILA which addresses the key aspect of gaze incorporation. We implemented a proof-of-concept G-VOILA assistant by harnessing the advanced deep learning technique and employed it for a controlled user study. We demonstrated the effectiveness of G-VOILA by achieving 89% recall rate and 84% precision score, surpassing the baseline method in both objective and subjective scoring metrics.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (NSFC) under Grant No. 62132010, Young Elite Scientists Sponsorship Program by CAST under Grant No.2021QNRC001, Tsinghua University Initiative Scientific Research Program, Beijing Key Lab of Networked Multimedia, Institute for Artificial Intelligence, Tsinghua University, Undergraduate / Graduate Education Innovation Grants, Tsinghua University, and Beijing National Research Center for Information Science and Technology (BNRist).

REFERENCES

[1] [n. d.]. Microsoft Bing. https://www.bing.com/. Accessed: 2023-09-07.

[2] Henny Admoni and Siddhartha Srinivasa. 2016. Predicting user intent through eye gaze for shared autonomy. In *2016 AAAI Fall Symposium Series*.

[3] Antti Ajanki, Mark Billinghurst, Toni Järvenpää, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, Teemu Ruokolainen, et al. 2010. Contextual information access with augmented reality. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 95–100.

[4] Antti Ajanki, David R Hardoon, Samuel Kaski, Kai Puolamäki, and John Shawe-Taylor. 2009. Can eyes reveal interest? Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction* 19 (2009), 307–339.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[6] Alibaba. 2023. Tongyi Qianwen. (2023).

[7] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. In *Acm sigir forum*, Vol. 46. ACM New York, NY, USA, 2–32.

[8] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. *International Conference on Human Factors in Computing Systems* (2017). https://doi.org/10.1145/3025453.3026033

[9] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. *OpenFlamingo*. https://doi.org/10.5281/zenodo.7733589

[10] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.

[11] Baidu. 2023. ERNIE Bot: Enhanced Representation through Knowledge Integration. (2023).

[12] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

[13] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. 2009. Eye movement analysis for activity recognition. In *Proceedings of the 11th international conference on Ubiquitous computing*. 41–50.

[14] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. 2018. Demonstrating Reality-Based Information Retrieval. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–4.

[15] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. 2018. Here and Now: Reality-Based Information Retrieval: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 171–180. https://doi.org/10.1145/3176349.3176384

[16] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the web conference 2021*. 743–755.

[17] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195* (2023).

[18] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2023. Language Models are Visual Reasoning Coordinators. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*. https://openreview.net/forum?id=kdHpWogtX6Y

[19] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://vicuna.lmsys.org

[20] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*. PMLR, 1931–1942.

[21] Sarah D'Angelo and Darren Gergle. 2016. Gazed and Confused: Understanding and Designing Shared Gaze for Remote Collaboration. *International Conference on Human Factors in Computing Systems* (2016). https://doi.org/10.1145/2858036.2858499

[22] Ellysse Dick. 2021. The promise of immersive learning: Augmented and virtual reality's potential in education. *Information Technology and Innovation Foundation* (2021).

[23] Jiexin Ding, Bowen Zhao, Yuqi Huang, Yuntao Wang, and Yuanchun Shi. 2023. GazeReader: Detecting Unknown Word Using Webcam for English as a Second Language (ESL) Learners. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 149, 7 pages. https://doi.org/10.1145/3544549.3585790

[24] Anup Doshi and Mohan M. Trivedi. 2009. Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions. *IEEE Intelligent Vehicles Symposium* (2009). https://doi.org/10.1109/ivs.2009.5164397

[25] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).

[26] Mats Ole Ellenberg, Marc Satkowski, Weizhou Luo, and Raimund Dachselt. 2023. Spatiality and Semantics - Towards Understanding Content Placement in Mixed Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 254, 8 pages. https://doi.org/10.1145/

3544549.3585853

[27] Mats Ole Ellenberg, Marc Satkowski, Weizhou Luo, and Raimund Dachselt. 2023. Spatiality and Semantics - Towards Understanding Content Placement in Mixed Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 254, 8 pages. https://doi.org/10.1145/3544549.3585853

[28] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010* (2023).

[29] Ziqi Gao, Yuntao Wang, Jianguo Chen, Junliang Xing, Shwetak Patel, Xin Liu, and Yuanchun Shi. 2023. MMTSA: Multi-Modal Temporal Segment Attention Network for Efficient Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–26.

[30] Toni Giorgino. 2009. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software* 31, 7 (2009). https://doi.org/10.18637/jss.v031.i07

[31] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* (2015). https://doi.org/10.3389/fpsyg.2015.01049

[32] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1072–1080.

[33] Woojeong Jin, Subhabrata Mukherjee, Yu Cheng, Yelong Shen, Weizhu Chen, Ahmed Hassan Awadallah, Damien Jose, and Xiang Ren. 2023. GRILL: Grounded Vision-language Pre-training via Aligning Text and Image Regions. *arXiv preprint arXiv:2305.14676* (2023).

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV]

[35] Michael F Land. 2006. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research* 25, 3 (2006), 296–324.

[36] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv preprint arXiv:2305.03726* (2023).

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[38] Mingyang Li, Yulin Xu, and Aolei Yang. 2021. Collaborative Robot Grasping System Based on Gaze Interaction. *Intelligent Equipment, Robots, and Vehicles* (2021). https://doi.org/10.1007/978-981-16-7213-2_8

[39] Tica Lin, Yalong Yang, Johanna Beyer, and Hanspeter Pfister. 2021. Labeling out-of-view objects in immersive analytics to support situated visual searching. *IEEE Transactions on Visualization and Computer Graphics* (2021).

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[41] Xiaoyi Liu, Yingtian Shi, Chun Yu, Cheng Gao, Tianao Yang, Chen Liang, and Yuanchun Shi. 2023. Understanding In-Situ Programming for Smart Home Automation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–31.

[42] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. 2022. TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing* (2022). https://doi.org/10.1016/j.neucom.2022.04.080

[43] Jérôme Louradour. 2023. whisper-timestamped. https://github.com/linto-ai/whisper-timestamped.

[44] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. arXiv:2205.06230 [cs.CV]

[45] Joshua Newn, Ronal Singh, Fraser Allison, Prashan Madumal, Eduardo Velloso, and Frank Vetere. 2019. Designing Interactions with Intention-Aware Gaze-Enabled Artificial Agents. *null* (2019). https://doi.org/10.1007/978-3-030-29384-0_17

[46] OpenAI. 2023. GPT-4 Technical Report. (2023).

[47] OpenAI. 2023. Introducing ChatGPT. (2023).

[48] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv:2306.14824 [cs.CL]

[49] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.

[50] Lisa Perkhofer and Othmar Lehner. 2019. Using gaze behavior to measure cognitive load. In *Information Systems and Neuroscience: NeuroIS Retreat 2018*. Springer, 73–83.

[51] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang, and Hans Gellersen. 2015. Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze. *ACM Symposium on User Interface Software and Technology* (2015). https://doi.org/10.1145/2807442.2807460

[52] Robin Piening, Robin Piening, Ken Pfeuffer, Augusto Esteves, Tim Mittermeier, Sarah Prange, Philippe Schröder, and Florian Alt. 2021. Looking for Info: Evaluation of Gaze Based Information Retrieval in Augmented Reality. *IFIP TC13 International Conference on Human-Computer Interaction* (2021). https://doi.org/10.1007/978-3-030-85623-6_32

[53] Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. 2023. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *Comput. Surveys* (2023). https://doi.org/10.1145/3491207

[54] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 647–664.

[55] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[56] B. J. Rhodes and P. Maes. 2000. Just-in-time information retrieval agents. *IBM Systems Journal* 39, 3.4 (2000), 685–704. https://doi.org/10.1147/sj.393.0685

[57] Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2020. Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. *Neurocomputing* 387 (2020), 369–382.

[58] SenseTime. 2023. Sense Nova. (2023).

[59] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580* (2023).

[60] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. 2023. A Unified Transformer Framework for Group-based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection. *IEEE Transactions on Multimedia* (2023).

[61] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128* (2023).

[62] Vildan Tanriverdi and Robert JK Jacob. 2000. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 265–272.

[63] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[64] Marc Tonsen, Chris Kay Baumann, and Kai Dierkes. 2020. A High-Level Description and Performance Evaluation of Pupil Invisible. *arXiv preprint arXiv:2009.00508* (2020).

[65] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[66] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*. PMLR, 23318–23340.

[67] Yuntao Wang, Zirui Cheng, Xin Yi, Yan Kong, Xueyang Wang, Xuhai Xu, Yukang Yan, Chun Yu, Shwetak Patel, and Yuanchun Shi. 2023. Modeling the Trade-off of Privacy Preservation and Activity Recognition on Low-Resolution Images. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[68] Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. FaceOri: Tracking head position and orientation using ultrasonic ranging on earphones. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–12.

[69] Yushi Wei, Rongkai Shi, Difeng Yu, Yihong Wang, Yue Li, Lingyun Yu, and Hai-Ning Liang. 2023. Predicting Gaze-based Target Selection in Augmented Reality Headsets based on Eye and Head Endpoint Distributions. *International Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581042

[70] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).

[71] Xuhai Xu, Anna Yu, Tanya R. Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, Te-Yen Wu, Junxiao Shen, Ting Zhang, Narine Kokhlikyan, Fulton Wang, Paul Sorenson, Sophie Kim, and Hrvoje Benko. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 202, 30 pages. https://doi.org/10.1145/3544548.3581500

[72] Xuhai Xu, Chun Yu, Yuntao Wang, and Yuanchun Shi. 2020. Recognizing unintentional touch on interactive tabletop. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–24.

[73] Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, and Shuai Ma. 2021. Control image captioning spatially and temporally. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2014–2025.

[74] Kun Yan, Lei Ji, Zeyu Wang, Yuntao Wang, Nan Duan, and Shuai Ma. 2023. Voila-A: Aligning Vision-Language Models with User's Gaze Attention. arXiv:2401.09454 [cs.CV]

[75] Yukang Yan, Haohua Liu, Yingtian Shi, Jingying Wang, Ruici Guo, Zisu Li, Xuhai Xu, Chun Yu, Yuntao Wang, and Yuanchun Shi. 2023. ConeSpeech: Exploring Directional Speech Interaction for Multi-Person Remote Communication in Virtual Reality. *IEEE Transactions on*

Visualization and Computer Graphics 29, 5 (2023), 2647–2657.

[76] Felix Yang, Saikishore Kalloori, Ribin Chalumattu, and Markus Gross. 2022. Personalized Information Retrieval for Touristic Attractions in Augmented Reality. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1613–1616. https://doi.org/10.1145/3488560.3502194

[77] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381* (2023).

[78] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. PEVL: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169* (2022).

[79] Belinda Zeng. 2022. *Go beyond the search box: Introducing multisearch.* https://blog.google/products/search/multisearch/

[80] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. 2020. Gradient-Induced Co-Saliency Detection. In *European Conference on Computer Vision (ECCV)*.

[81] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 16793–16803.

[82] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. 2023. RegionBLIP: A Unified Multi-modal Pre-training Framework for Holistic and Regional Comprehension. *arXiv preprint arXiv:2308.02299* (2023).

[83] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592* (2023).

[84] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. 2022. Generalized Decoding for Pixel, Image, and Language. *arXiv preprint arXiv:2212.11270* (2022).

A VOILA-G PROMPT

The prompt used for VOILA-G is shown in Table 3.

B VOILA-G IMPLEMENTATION SPECIFICATIONS

B.1 Device and Workflow

We separated the devices as **[Local device]** and **[Remote device]**, which is more compatible with future deployment. Local devices are tasked with executing algorithms of lower computational complexity, such as data preprocessing, while the remote server is designated for algorithms that are more computationally intensive. An additional critical aspect of our setup is the optimization of data transfer between the Local device and the Remote device to facilitate real-time response capabilities in future iterations, which means transmitting entire video clips is impractical. Therefore, we temporally located at most three representative frames on the local devices and transmitted them to the server for further processing.

VOILA-G's implementation is shown in Figure 13.

- **[Local device]** The Pupil Lab Invisible glasses and a companion smartphone (OnePlus 8T O).
  - Usage details: The Pupil Lab Invisible glasses require a connection to a companion smartphone to be driven, which is not depicted in this illustration figure. To reduce potential distractions, participants were advised to carry the smartphones in their pockets throughout the experiment.
  - Data transfer: Upon completion of each round of queries, the collected data were uploaded to Pupil Cloud and subsequently downloaded to our experiment laptop.
- **[Local device]** Experiment laptop - MacBook Pro with Intel processor and running the Ventura OS.
  - Device requirements: our experimental setup is compatible with a broad range of laptops and OS.
  - Usage and data transfer: We preprocess the data on the local devices and send the data to a remote server equipped with high-performance GPUs for further analysis.

➤ Overall template:	VOILA is designed to be able to assist with visual question answering task. With the input of a series of ego-centric snapshot description and user’s query question, VOILA can (1) answer the question directly and (2) provide query input for web search engine. VOILA is able to understand large amounts of images and videos, then answer user’s question with its knowledge and the help of search engine. VOILA can not directly read images or videos, but a list of visual understanding tools will provide textual information about the visual content, based on which VOILA can infer knowledge about the visual content. The textual information provided may be vague and general, but VOILA should do its best to organize the information to understand the visual content. The whole visual assitant is able to capture ego-center snapshots and user’s eyegaze attention coordinate, based on which several visual understanding tools will generate textual information. When the user asks questions knowing that VOILA has already understood the visual content and his/her eyeygaze interest, so his/her question may be vague by using pronoun or skipping intent words. VOILA should do its best to infer user’s query intent by fulfilling the missing information. With the fulfilled query content, VOILA should (1) answer the questions directly and (2) generate proper and unambiguous query input for web search engine. When inferring user’s query intent, Voila should follow steps below: (1) figure out whether user’s query contains ambiguous words, such as pronouns or words can not be understand depend solely on query text. (2) if yes, figure out the possible meaning of the ambiguous words based on the textual information provided by visual understanding tools and user’s query question. (3) if no, check out whether the context or interest information might be related to user’s query. If yes, combine it with query text properly. If no again, infer user’s query intent solely based on the query text. Overall, VOILA is a powerful visual dialogue assistant tool that can integrate visual understanding content and user’s ambiguous query to answer user’s question and generate proper query input for web search engine.
	——
	VOILA takes the following information type as input: inputs
	——
	VOILA output a response in format (MUST BE IN JSON FORMAT):
	“‘json {
	"thought": string \\Explain the process of inferencing user’s unambiguous query intent, use the textual information provided by visual understanding tools and user’s query question
	"answer": string \\The answer to user’s question
	"query": string \\The query input for web search engine
	}

➤ Input type	→Context Caption: A textual description of the whole visual content, generated by the visual captioning tool. → Interest Caption: A textual description of the user’s eye gaze interest, generated by a visual captioning tool – a list of textual descriptions of the object that the user is looking at. The recognition result might be inaccurate, and the input is the top 3 descriptions with the highest confidence. → OCR: Extracted text from the whole vision field, generated by OCR tool, the recognition result can be considered highly inaccurate except for understandable phrases. → User Query: The user’s query question, may be vague by using pronouns or skipping intent words. Query text is a transcript using a speech recognition tool, which may be inaccurate if you find some words hard to understand.

Table 3. **Prompt Template of Voila-G**

  - Response display: The final results returned by the remote server are presented to participants through a streamlined user interface, which is developed using REACT.
- **[Remote device]** GPU server - either two NVIDIA V100 32GB GPUs or a single NVIDIA A100 80GB GPU.
  - Device requirements: The total GPU memory requirement for VOILA-G is estimated to be around 23GB.
  - Usage and data transfer: The server can handle the computational demands of multiple complex deep-learning models. After completing the generation of textual transcripts and descriptions for the submitted data, VOILA-G formulates a query utilizing the prompt outlined in Table 3 for the GPT-4 API. The server then proceeds to cleanse the data format of GPT-4’s output and relays the refined response back to the local laptop.

B.2 Data Preprocessing Workflow

- A critical component of voice-interactive systems is the detection of the initiation and conclusion of user interactions, such as keyword spotting and push buttons. In our experimental framework, participants activate queries by directly speaking their questions. Thus, our researchers manually annotated the start

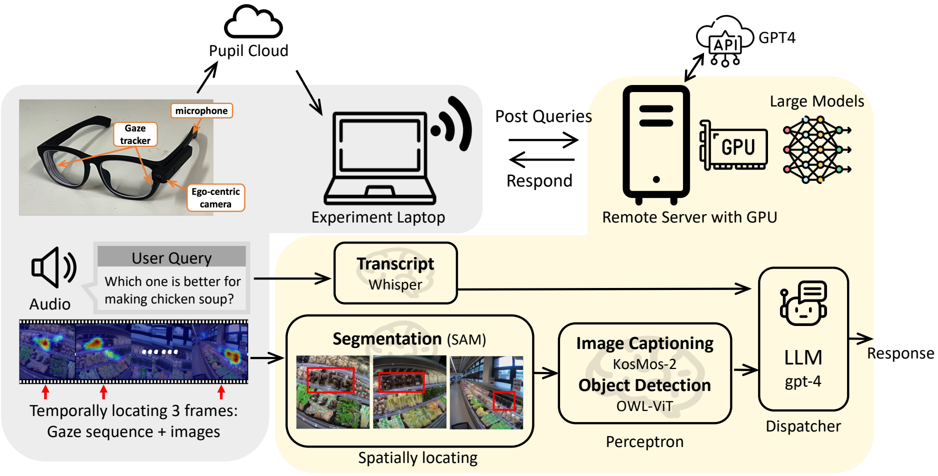


Fig. 13. A Preliminary Implementation: VOILA-G's pipeline.

and end points of each query using the audio editing software Audacity. For the prospect of real-time interaction—a feature not inherently supported by the Invisible Glasses due to the lack of an input mechanism—we integrated a WiFi-based button and other supportive (not inherently supported by the Invisible Glasses) for the prospect of future real-time interaction, not utilized in the current experiment to maintain simplicity.

- With the interval flag of each query, we programmed the local devices to selectively extract three frames for each query and crop out the associated gaze data sequences. The chosen images, gaze data, and the audio of the user's query were then posted to the server. The frame selection is guided by the criteria detailed in Section 4.3, where we prioritize high-resolution frames from fixation points that are (1) notably longer in duration and (2) located proximate to the query's initiation timestamp. We refrained from using pronouns as a cue for frame selection, as generating transcripts to identify such cues would necessitate the employment of Whisper and additional Natural Language Processing (NLP) models. Given their substantial computational demands, these models are impractical to run on local devices.

B.3 Remote Algorithm Pipeline

In the remote processing phase, we leverage deep learning models whose configurations are documented in Table 4. The generation of context descriptions is provided by KosMos-2 [48] wrapped up with gradio since Kosmos-2's requirements are complicated and more suitable to run in docker. Meanwhile, interest descriptions are managed by OWL-ViT [44]. We choose potential interest objects by (1) sorting bounding boxes based on the Intersection over Union (IoU) and (2) filtering them by a confidence score threshold. Due to the risk of misidentification in object detection tasks, our pipeline outputs three candidate labels for each object of interest, which increases the likelihood that the user's intended object is among the identified candidates (ensure recall).

C VOILA-SALIENCY

Saliency models aim to predict eye fixation points, which are the locations where a person's gaze would typically rest when looking at a scene. Applying saliency models to substitute the usage of a gaze tracker can be an

	version	specified parameters	input data	output data
Whisper [55]	large, multilingual, v2	-	audio file	transcribed query text
SAM [34]	vit_h(default)	-	image + gaze	b-box
KosMos-2 [48]	patch14-224, 1.6B params	description_type="Detail"	image	caption
OWL-ViT [44]	large, patch14	-	image	(type, b-box, score)

GPT-4 [46]	model0613, api0515	max_tokens=1500,
		temperature=0, top_p=1,
		frequency_penalty=0,
		presence_penalty=0.6
			see prompt	see prompt

Table 4. Deployed deep learning models and API for VOILA-G. "b-box" is short for bounding box.

	Recall			Precision		
	All	Explicit	Ambiguous	All	Explicit	Ambiguous
saliency-image	78.76%	93.86%	62.19%	72.86%	93.03%	50.7%
saliency-video	78.83%	94.12%	62.04%	73.6%	93.31%	51.96%

Table 5. Comparison between VOILA system built on image saliency prediction and video saliency prediction.

alternative for the VOILA querying paradigm. This section presents the implementation of a saliency-based visual querying system and discusses the experiment results.

C.1 Implementation

Saliency models can be roughly categorized into two types: image and video saliency prediction. For image saliency, we utilized TranSalNet [42], and for video saliency, we adopted UFO [60], both leading in the "Saliency Detection" task[5].

- **Image Saliency Prediction:** TranSalNet [42] was used to determine saliency areas in selected frames, with the most salient area's center designated as gaze coordinates.
- **Video Saliency Prediction:** UFO's [60] video saliency capabilities allowed for frame-by-frame gaze coordinate prediction, creating a gaze trace analogous to gaze sensor data. This predicted gaze trace was processed using the same pipeline as VOILA-G.

Table 2 exclusively presents image-based saliency results, as our system and baselines are implemented with image-level deep learning techniques. As detailed in Appendix B, transferring entire video clips to **[Remote Devices]** or executing video-level deep learning on **[Local Devices]** is impractical. To evaluate video-based saliency models, we pre-transferred all data to our server and conducted preprocessing there.

C.2 Results and Discussion

Results presented in Tables 2 and 5 indicate that the performance of VOILA-center, VOILA-saliency (image), and VOILA-saliency (video) are comparable. VOILA-saliency (video) marginally outperforms VOILA-saliency (image),

[5] https://paperswithcode.com/area/computer-vision

with the latter showing a slight advantage over VOILA-center, in terms of overall recall and precision. Contrary to expectations, utilizing gaze predictions from saliency models offers a negligible improvement over using head direction (center) to infer interest points. Additionally, temporal saliency analysis (video) does not significantly surpass saliency prediction from isolated frames. We discussed possible reasons, as well as highlighted the gap between current saliency models and the specific demands of G-VOILA's querying task.

- Why do saliency models underperform compared to using head direction?
  - **Dataset discrepancy:** Saliency models, trained on datasets like SALICON [32] and CoCA [80], often feature images with eye-catching objects or subjects isolated from the background. In contrast, daily scenarios such as supermarkets present closely arranged homogeneous items, making it hard to identify the user's interested object. Similarly, at-home and street scenarios are composed of complex structure and object placement. Additionally, these models are trained on high-quality photos, whereas our use case involves snapshots from egocentric videos.
  - **Semantic gap in gaze for saliency prediction and G-VOILA use case:** Even though both use cases aim to predict user interest, saliency models emphasize visually striking objects whereas G-VOILA indicates objects that users want information for. Therefore, a saliency model might focus on objects with vibrant colors, unique shapes and located in well-illuminated areas, etc. For instance, in the home setting, our employed saliency model tends to highlight a Minions toy over other objects placed on the table. Therefore, as a companion movement to the user's eye movements, the head direction can gain comparable performance with VOILA-saliency.
- Why temporal saliency analysis (video) underperforms frame-level prediction (image):
  - The image and video saliency models stem from different research and are not directly comparable.
  - **Video saliency prediction requires a more sophisticated temporal locating algorithm because the predicted gaze trace does not share similar properties with real gaze.** Gaze data's feature alone can be used to temporally locate frames that contain the user's interest, as shown in Section 7.2.1. However, predicted gaze traces cannot accurately pinpoint interest frames because of both prediction inaccuracies and lower sampling rates. A video saliency model tailored to querying tasks may contribute to this problem.
  - **Static-dynamic mismatch for objects of interest:** Video saliency model typically highlights dynamic objects against a static background [49]. However, in scenarios like supermarkets or workspaces, the object of interest is often placed static and captured within a dynamically moving egocentric view.