Multimodal Technologies and Interaction

MDPI

Article

Interactive Content Retrieval in Egocentric Videos Based on Vague Semantic Queries

Linda Ablaoui [1,2,*], Wilson Estecio Marcilio-Jr [3], Lai Xing Ng [2,4], Christophe Jouffrais [2,5] and Christophe Hurter [1,2]

1 Ecole Nationale de l'aviation Civile, 7 Avenue Edouard Belin, CS 54005, CEDEX 4, 31055 Toulouse, France; christophe.hurter@enac.fr
2 CNRS, IPAL, 15 Computing Drive, Singapore 117418, Singapore; ng_lai_xing@i2r.a-star.edu.sg (L.X.N.); christophe.jouffrais@cnrs.fr (C.J.)
3 Faculty of Sciences and Technology, São Paulo State University (UNESP), Presidente Prudente 19060-900, SP, Brazil; wilson.marcilio@unesp.br
4 Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, #21-01, Connexis South Tower, Singapore 138632, Singapore
5 Institut de Recherche en Informatique de Toulouse (IRIT), Centre National de la Recherche Scientifique (CNRS), 31062 Toulouse, France
* Correspondence: linda.ablaoui@enac.fr

Abstract

Retrieving specific, often instantaneous, content from hours-long egocentric video footage based on hazily remembered details is challenging. Vision–language models (VLMs) have been employed to enable zero-shot textual-based content retrieval from videos. But, they fall short if the textual query contains ambiguous terms or users fail to specify their queries enough, leading to vague semantic queries. Such queries can refer to several different video moments, not all of which can be relevant, making pinpointing content harder. We investigate the requirements for an egocentric video content retrieval framework that helps users handle vague queries. First, we narrow down vague query formulation factors and limit them to ambiguity and incompleteness. Second, we propose a zero-shot, user-centered video content retrieval framework that leverages a VLM to provide video data and query representations that users can incrementally combine to refine queries. Third, we compare our proposed framework to a baseline video player and analyze user strategies for answering vague video content retrieval scenarios in an experimental study. We report that both frameworks perform similarly, users favor our proposed framework, and, as far as navigation strategies go, users value classic interactions when initiating their search and rely on the abstract semantic video representation to refine their resulting moments.

Keywords: human–computer interaction; zero-shot content retrieval; multimodal querying; egocentric videos

Academic Editor: Wei Liu

Received: 29 April 2025
Revised: 10 June 2025
Accepted: 16 June 2025
Published: 30 June 2025

Citation: Ablaoui, L.; Marcilio-Jr, W.E.; Ng, L.X.; Jouffrais, C.; Hurter, C. Interactive Content Retrieval in Egocentric Videos Based on Vague Semantic Queries. *Multimodal Technol. Interact.* **2025**, *9*, 66. https://doi.org/10.3390/mti9070066

1. Introduction

First-person, or egocentric, videos are ubiquitous, serving as a medium for human–robot interactions [1] and advanced action recognition in virtual reality [2], with many applications in assistive technologies [3,4], education and security [3], and social media [5]. They differ from third-person videos in several ways: these videos often change camera angles [6], include a first-person view of hands [7,8] which makes occlusions more frequent [9], and have a limited field of view, making global context and scene understanding difficult [10] and relevant content within them can be sparse as relevant events or

objects appear in short time segments within footage that can span hours [11]. To efficiently navigate the ever-expanding volume of egocentric video content, navigation methods must continually improve to meet the demands of increasingly diverse and complex data. Initial approaches to video navigation involved manually tagging videos with keywords for direct access to specific video parts. Advancements in computer vision and deep learning have enabled the automatic tagging of videos [12], significantly reducing the manual effort required. And yet, tags may fail if the user is looking for content that is not covered by them. The emergence of zero-shot content retrieval [13] resolved this issue by enabling natural language querying of videos without prior knowledge of the specific content. This particular approach leverages vision–language models (VLMs) to bridge the gap between visual and textual data, allowing for zero-shot video searches. Moreover, many datasets with annotated egocentric footage [8,14,15] exist for the purpose of pre-training VLMs and specializing them in egocentric content retrieval tasks [16–18]. However, zero-shot content retrieval approaches often suffer from a lack of accuracy when queries do not exactly match the target content [19]. This can occur when users formulate vague semantic queries [20].

Consider a person who waters their plants daily. One evening when they get home, they wonder "Have I watered the plants in my bedroom yet?". They can formulate a query like "Did I water the plants?". The term "plant" is ambiguous as it can refer to an herb, a flower, a tree or a cactus, a wild plant or a potted plant. Depending on the footage, several subcategories of the term "plant" can be present and can thus be relevant to a VLM-based retrieval system which interprets terms based on visual features [21]. A better formulated query, like "Did I water my potted plants?" can still be confusing, as a retrieval system would retrieve all moments where a potted plant appears, regardless of the location. Even a more specific query like "Did I water the potted plants in my bedroom?" can still be vague because of the concept of possession in "*my* bedroom". The system would need to understand what uniquely makes a bedroom the user's (location, specific furniture, presence of personal items). If multiple bedrooms with plants exist in the footage, the system may not differentiate between them. This example illustrates how queries can be inherently vague and confound current retrieval systems. Many studies show that human users frequently formulate such vague queries [22–24], which degrade the performance of VLM-based zero-shot content retrieval approaches. While recent research has explored vague queries in web search [25] and Q&A [26], the impact of semantic vagueness on video retrieval—especially in egocentric videos—remains underexplored.

This paper investigates the requirements for a content retrieval system that supports navigation in egocentric videos based on such vague semantic queries. Specifically, this paper investigates the following research questions:

RQ1: What interaction techniques and design guidelines can assist users in effectively retrieving relevant content from egocentric videos based on vague semantic queries?

RQ2: How do users navigate egocentric video content based on vague semantic queries?

First, we identify the factors that cause vague query formulation and narrow them down to incompleteness and ambiguity. Second, we propose interaction methods that leverage a VLM capable of zero-shot content retrieval and integrate them in a user-centered video content retrieval framework. The framework is based on the idea of incremental refinement of queries to resolve their vagueness (narrowing down specific content). Third, we evaluate the proposed framework against a baseline video navigation framework in an experimental study where eight participants navigated videos to locate specific moments aligned with six predefined scenarios, with a simulated sense of egocentricity. Our framework achieves an average completion time of 9.375 min—comparable to 9.042 min for standard navigation—and a mean precision of 63.2% (vs. 63.0%), with trade-offs in recall

(39.8% vs. 51.4%). Likert-scale-based assertions answered by participants show that they favor our proposed framework (average 6/7 vs. 5/7) and consider it more efficient (5.5/7 vs. 5/7). Furthermore, we analyzed participants' navigation strategies and found that they tend to initiate their search with standard navigation methods, such as textual search and sequential browsing, and filter out non-relevant moments by refining their results with semantic visualizations of video frames and interactions utilizing this visualization.

The contributions of this work are as follows:

- We propose novel video interaction methods leveraging zero-shot content retrieval and incremental query building within a user-centered framework to support navigation in egocentric videos based on vague semantic queries;
- We extract user strategies for vague scenario-based video navigation based on interaction logs and structured observations and derive design implications that highlight the importance of hybrid navigation patterns, precision-first behaviors, and semantic visual feedback for supporting ambiguity and incompleteness in user queries.

The remainder of this paper is organized as follows: Section 2 presents a literature review of existing video content retrieval methods, namely interactive, zero-shot-based, egocentric- driven, and vague-query-supporting methods. Section 3 presents our proposed zero-shot, user-centered video content retrieval framework. Section 4 details the experimental study that we conducted to (1) validate our framework when dealing with vague queries and (2) explore user strategies when retrieving information in videos based on vague scenarios. Section 5 reports the results of our experimental study. We discuss our work in Section 6 and conclude with Section 7.

2. Literature Review

2.1. Interactive Video Navigation

Interactive video navigation distinguishes itself from traditional navigation by emphasizing user engagement, control, and direct manipulation of the video content. Several works use video content-dependent keywords called "tags" to pilot the content in videos. These tags are either manually assigned by users or automatically assigned tags [27–29], directly usable to move to specific parts of a video [28–30] or used in a content-based search mechanism [27]. Other works allow users to navigate video content by directly engaging with objects within the video (direct manipulation video navigation systems) [31–33]. Another category of works uses domain-specific hypotheses to tailor the video navigation approach and interactions for domain specific tasks [29,34–36].

Several of these approaches are effective for navigating a video with predefined context. However, in long videos where users are partially familiar with the video content and can only formulate queries based on their hazy knowledge, such approaches may not be suitable as the content users seek may not align with the content highlighted by the given tags or interactions. In this work, we propose a navigation framework that adapts to ambiguity and/or incompleteness in user queries that is independent from the exact content of the egocentric videos through incremental refinement of query results.

2.2. Zero-Shot Content Retrieval in Video Tasks

Zero-shot content retrieval relies on encoding data in an auxiliary format that encodes relevant information. This concept enables models to recognize classes that were not observed in training [37]. Several video analysis tasks benefit from the adaptability of zero-shot content retrieval models such as video question answering [38–40], video moment retrieval [40–44], video captioning [45], video summarization [46,47], action recognition [48,49], and object detection and tracking [50,51].

While zero-shot based approaches demonstrate advancements over traditional methods for video analysis in terms of generalization and semantic understanding, VLMs like CLIP [52] that are capable of zero-shot retrieval are very sensitive to textual input, as ambiguous textual input reduces the relevance of the output [19]. In this work, we investigate the design requirements needed to leverage such zero-shot approaches in order to support vague textual prompts for optimal outputs and reveal key insights on interactions built on zero-shot learning models that are favored by users to navigate egocentric videos based on vague queries.

2.3. Content Retrieval in Egocentric Videos

Egocentric videos present unique challenges for content retrieval related to the first-person perspective. Most works in this egocentric video content analysis focus on object detection [15,53–55], hand tracking [7,9,56], hand–object interactions [8,57–59], human–human interactions [14,60,61], human action recognition [62,63], and action recommendation [64]. Beyond these perceptual tasks, works have also focused on visual question answering [14,65–67] and moment retrieval [68–70] in egocentric videos.

Particularly, the concept of episodic-memory based retrieval is addressed in [14,66,68], where the video is considered to be a surrogate memory of the user and their queries/questions are episodic memory oriented, i.e., linked to a specific moment, physically grounded, highly personal, and include the users themselves (for example, "Where did I put my keys after returning home?"), rather than factual questions (for example, "What color is the car?"). However, the annotations employed in these tasks are mostly well-defined and specific questions paired with exact video timestamps. As queries/questions are shaped by the episodic memory, they can be formulated vaguely due to episodic memory occlusions (Section 3.1). To our knowledge, little research has explicitly addressed egocentric video retrieval when queries are vaguely formulated. Thus, we investigate the design requirements for a video content retrieval framework that supports vague semantic queries.

2.4. Vague Semantic Queries in Content Retrieval

The term "vague query" is a vague term itself. Many works in content retrieval employ it to refer to the opposite of a specific query. Other works employ the term "ambiguous query" to refer to polysemous textual inputs [71]. The problem of natural language ambiguity is a well-documented one in content retrieval. Several methods exist to resolve the polysemy of natural language terms, such as query reformulation [72] and query augmentation [73]. Multimodal querying, like sketch-based querying [74], image-based or composed querying [70], and motion-based interactions [75], can also be employed to overcome natural language ambiguity. In conversational agent setups, suggestion mechanisms and clarification questions [76,77] are employed when ambiguity is detected in a user query and are used to involve the user in the disambiguation process.

In video content retrieval particularly, the concept of vague queries is less studied. Zakra et al. [25] propose a fuzzy ontology to reinforce video semantic interpretations. However, such an approach relies on stored annotations which may not necessarily cover the topics sought by users. VLM-based solutions for moment retrieval and question answering in videos are known to be sensitive to textual input [19] and most datasets have specific annotations [64,78–80]. We tackle this gap by putting design guidelines for vague-query-supporting video content retrieval in egocentric videos and evaluate this video content retrieval framework in a vague-scenario-based video content retrieval experiment.

3. Materials and Methods

Video content retrieval refers to the task of locating one or more moments within a video that match a user's intent, expressed through a query. If we treat the video as a collection of discrete moments, shots, or images and restrict queries to text (as this is one of the most intuitive modality for human users [81]), we can model retrieval as a function mapping a textual query to a set of candidate moments where three levels of query certainty can be defined as follows:

Specific query: all retrieved moments are relevant;

Vague query: the retrieved set contains a mix of relevant and irrelevant moments; and

Erroneous query: None of the retrieved moments are relevant.

In this paper, we focus on egocentric videos, which assume that the user has at least partial knowledge of the video content. Therefore, we concentrate specifically on vague queries and semantic vagueness. Erroneous queries fall outside the scope of this work.

3.1. Factors of Vagueness

Several works on information retrieval [82–84] employ the term "ambiguous queries" to refer to textual keywords or questions asked to a retrieval system (document retrieval framework powered by an AI (Artificial Intelligence) that are polysemous, i.e., which can refer to more than one concept or concrete object and which confuses the retrieval system. Several techniques exist to resolve this confusion.

While natural language ambiguity is one factor that leads to vagueness, several other factors can make remembering an event or expressing it in natural language difficult. Song et al. [23] classified queries as specific, ambiguous, and broad. Keyvan et al. [85] list different factors that can make a query ambiguous or unclear in a conversational search setup: poor definition, lack of specifics or structure, a large set of returned results, multi-domain returned results, or reference to a prior context.

Moreover, as egocentric videos are videos that the users shot themselves, they are linked to the users' memories. One may consider them a surrogate memory as prior works have shown their use in augmentative memory systems [86,87]. Accordingly, the formulation of search queries over egocentric content is closely tied to how well users remember the captured events and any failure in their memory can have an impact on the queries they formulate. Following Schacter's classification of common memory failures into seven categories or "seven sins of memory" [88], we can *infer* the factors that hinder the process of specific query formulation:

- Absent-mindedness (lack of attention) and transience (memory fading over time) may lead to the occlusion of certain information on events and thus to under-specified, incomplete queries.
- Blocking (memory retrieval failure) may prevent users from finding the exact terms to describe a concept or an event, leading to incomplete queries if no specific terms are employed or ambiguous queries when synonyms and paraphrasing are employed.
- Misattribution and bias may cause ambiguous queries referring to several scenes or even erroneous queries referring to incorrect or irrelevant scenes.

Considering every factor of inducing vagueness in query expression when designing a single framework is an overly complex task. Therefore, we focus on two specific aspects that are theoretically linked to memory failures in this work, incompleteness and ambiguity where an

Ambiguous query represents a polysemous natural language query, i.e., a query that can describe more than one single concrete object/action/scene/concept depicted in the

target video. (For example,"Where did I put my keys after returning home?", possible ambiguity on "keys" if several exist in the footage—house keys, car keys, etc.);

Incomplete query represents a natural language query that does not describe the intended concrete object/action/ scene/concept depicted in the target video sufficiently to identify it in a unique manner (for example, "Where are my car keys?" requires specification if the keys appear several times in the footage, but only the most recent occurrence is truly relevant).

3.2. Video Content Retrieval Approach

The proposed approach for matching queries to video frames was inspired by the zero-shot VLM moment retrieval process proposed by [89]. All frames of the video are encoded using CLIP's vision encoder beforehand when the video is loaded. Then, the user input (image or text) is encoded using either CLIP's vision or textual encoder, respectively. Finally, a cosine similarity is computed between the encoded input embedding and each encoded video frame embedding. This gives a similarity score between the query and each video frame. Unlike what occurred in [89], no post-processing is performed on the similarity scores. Instead, the results (similarity scores and embedding vectors) are displayed to the user and it is through several interactions based on the representation of these results that the user decides on which moments to select. The user-centered framework is depicted in Figure 1.

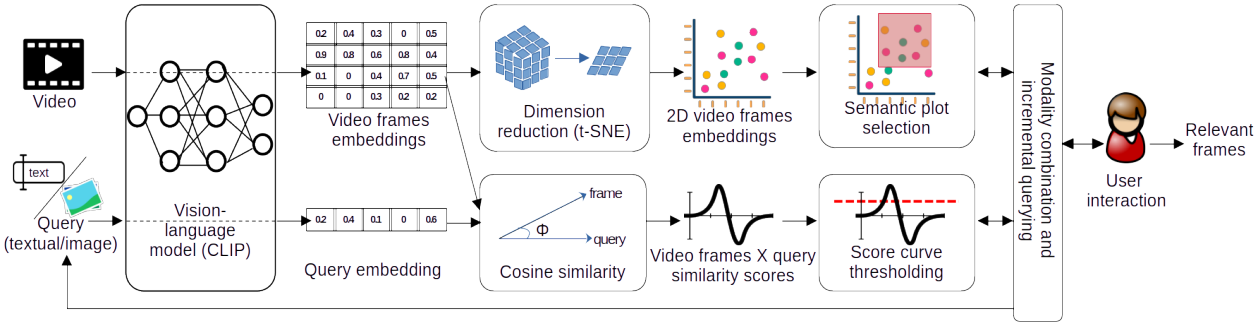


Figure 1. Vague-query-based video information retrieval framework. Video frames are encoded using the VLM vision encoder (CLIP). After encoding the query (modality chosen by the user), a cosine similarity is applied. Users can make decisions on which frames to select using the visualizations and interactions that are built on the cosine similarity and encoded embeddings' vectors.

Several approaches rely on an end-to-end process that adds a moment proposal step after computing the cosine similarity [38,42,44,89]. Resolving vagueness can be performed through disambiguation processes (Section 3.1). While these approaches have shown promising results in specific domains, they can only resolve vagueness to some degree and specifically address the problem of natural language ambiguity. Vagueness is directly tied to relevance and relevance is a subjective concept that varies with users. So, it makes sense that the best qualified entity to resolve vagueness would be the user, which is why we propose a user-centered framework that helps users make the decision on what moments are relevant, through different interactions and video data representations, instead of a framework that decides on relevant moments in the video.

3.3. Interaction Design

The idea is to provide the user with the necessary interaction methods to search for content based on different levels of certainty. If the user knows exactly what they seek, textual search coupled with score thresholding can be used, while if they do not, they

can use broad navigation methods like temporal or semantic navigation to identify some content first and refine their textual query incrementally.

3.3.1. Sequential Navigation of Videos

Navigating sequentially in a video is a common interaction method, but we wanted to avoid users relying solely on it and neglecting other framework features. To address this, we disabled video playback and instead allowed users to inspect frames by either clicking a physical rectangular timeline indicating timestamps or using left/right arrow keys to move frame by frame. The keystroke option compensates for the timeline's coarse precision, enabling users to either explore broad video segments or pinpoint specific frames to check subtle changes in the scene.

3.3.2. Textual Search

Textual search is widely used in applications like search engines and file explorers. To prevent over-reliance on it, we limited the textual query field to accept only one sentence (subject + verb + object) at a time. If users employ keywords, the system precedes them with "a photo of" or "something" depending on whether the keyword is a noun or a verb, respectively, for better performance by CLIP [90]. For more complex queries, users could add additional sentences, each paired with its own similarity score. This design ensures users can pinpoint content in a video based on a specific query while also encouraging the combination of textual search with other modalities.

3.3.3. Similarity Score Interactions

The cosine similarity score is presented as a standard 2D curve. Users can adjust a horizontal slider to set a threshold, above which frames are automatically selected. This feature enables a semi-automatic moment selection process, complementing the manual dragging interaction available on the timeline.

3.3.4. Image Search

Users can work with two types of images: external images and image crops. Image crops support incremental querying by allowing users to search for ambiguous forms or colors and then "screenshot" an object from the video to refine their search, effectively performing disambiguation of polysemous terms. External images, meanwhile, serve as an alternative to textual queries, extending the expression power of users (for example, using a picture of a specific model of bicycle is more specific than using the keyword "bicycle").

3.3.5. Interaction with Hyperspace Representation of Encoded Video Frames

This interaction relies on embedding vectors generated through the encoding of video frames by the VLM (CLIP). Each frame produces a high-dimensional vector (512 dimensions for clip-vit-base-patch16), representing the semantic distribution of video frames in the hyperspace of the VLM. To simplify visualization and manipulation, these vectors are reduced to 2D space using the t-SNE [91] method.

The 2D projection allows users to visually explore the semantic distribution of frames, with each point corresponding to a frame and preserving relative semantic relationships. Users can identify clusters of similar frames or unique outliers. In order to help understand this 2D projection and how it ties to video frames, we cluster the frames using DBSCAN [92], take the frame at the center of each cluster, and display a small version of it on the sides of the 2D projection space. To further aid navigation, the 2D projection space includes tools for zooming, panning, selection, and four color maps.

Selection (base): highlights the frames selected by users. It can help users filter outliers out of a selection based on a vague semantic query.

Clusters: highlights DBSCAN computed clusters. It can help users identify global semantic areas in the video.

Timestamps: applies a two color gradient on the 2D projection where darker points indicate frames at the (temporal) beginning of the video recording and lighter points indicates frames at the end of the recording. It helps users link the temporal and semantic representation of the video.

Scores: if scores are computed, applies a two color gradient on the 2D projection where green points show frames with higher scores and red points show frame with lower scores. It can help users filter outliers through high score frame clusters.

3.4. Mechanisms of Vagueness Resolution

Resolving vague queries requires adapting the system behavior to the type of vagueness at hand, choosing between the different meanings of an ambiguous query and enriching an incomplete query. In this work, this process is not performed in an end-to-end manner, rather this role is given to the user and our framework is meant to facilitate this role. The user iteratively interacts with semantic data representations to isolate the relevant frames and incrementally refine their query.

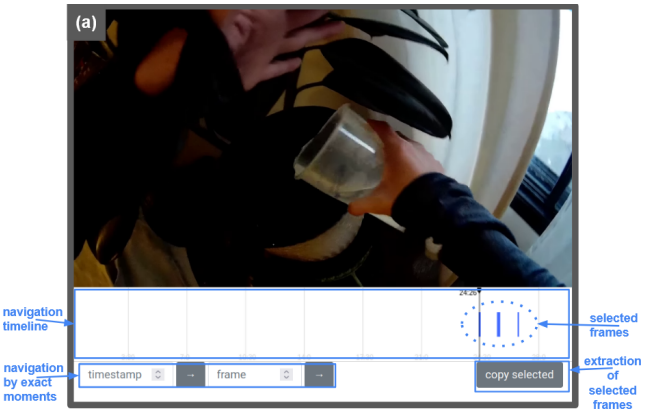
An example scenario—finding if the bedroom plants got watered (detail in Table 1)—is depicted with the graphical user interface of the implementation of the framework in Figure 2 and showcases this concept of incremental querying. A combination of several modalities was used to find the target frame in Figure 2a. First, textual search (Figure 2c) was employed to find frames that relate to "potted plants" and to "bedroom". Then, a threshold (Figure 2b) was applied to the obtained scores. Finally, the selected area was viewed and refined in the semantic plot (Figure 2e). For more complex queries, users can add additional sentences, each paired with its own score (detailed in Appendix B).

Table 1. Scenarios used in the experimental setup. Participants are presented with the title first, followed by the contextual paragraph, then a timeline showing descriptions and representative frames of each segment of the associated video as annotated in the Ego4d dataset, and finally the instructions needed to be matched with moments from the video.

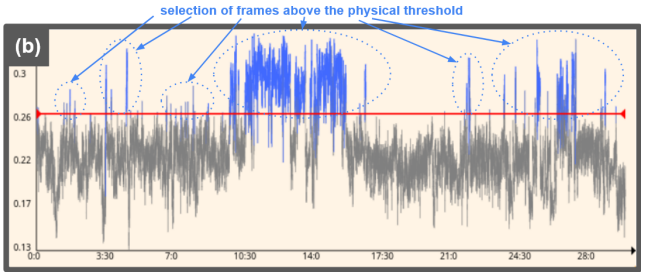
n°	Title	Expected Query Type	Context	Instruction
0	Camping	Specific	A group of friends went camping. Afterwards, the person who brought the barbecue cannot find the bottle of oil and wonders if a friend took it by mistake.	- Find all moments where a bottle of oil is visible.
1	Fire alarm triggers	Specific	A security inspector is responsible for seeing if security protocols are being followed in an office. Among the things they must check is to ensure that fire extinguishers and fire alarm triggers are in place.	- Find all moments that contain fire alarm triggers in the workplace only.
2	Water the plants	Vague	A person regularly waters their plants at home. When they got home, this person forgot whether they had already watered all their plants or not, especially those in their bedroom.	- Find all moments where the person waters the plants. - Did they water the plants in the bedroom?
3	Where are my papers?	Vague	A person prints papers for later use. When it comes to using the papers (after returning home), the person no longer remembers where they put the papers.	- Find the last location of the printed papers. - When did they end up there?
4	Who won?	Vague	A group of friends gathered one evening for a party where they played games. While reminiscing about the event, they cannot agree on who won.	- Find when the first game ended. - Who won? The person filming or their opponent?

Table 1. *Cont.*

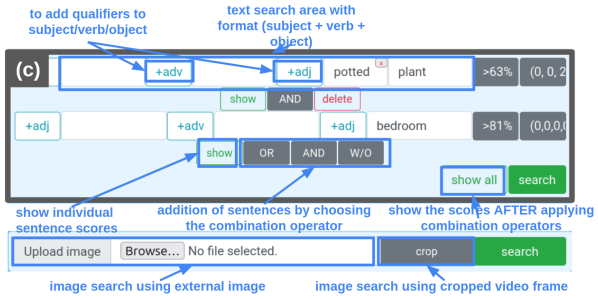
n°	Title	Expected Query Type	Context	Instruction
5	Diet	Vague	A person goes to the doctor and is informed that they have to undergo a low-sugar diet. The person struggles to remember every high-sugar substance they regularly consume.	- Find all moments that show high-sugar food that the person consumes or will consume.
6	Warmups	Vague	A person particularly enjoys the warmup exercises they performed before climbing. The person would like to reproduce them in another course.	- Find all moments where the group performs warmups before climbing.



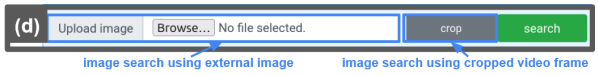
(**a**) Timeline video navigation



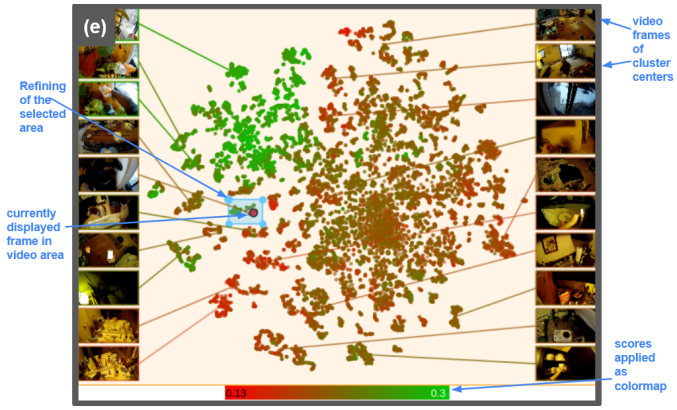
(**b**) Similarity score curve



(**c**) Textual search inputs



(**d**) Image search inputs



(**e**) Semantic video representation plot

(**a**)	video cannot be read, navigation is limited to timeline pinpointing or exact moment goto. Selection with drag and drop
(**b**)	visualizes the cosine similarity between the video frames embeddings sand a textual prompt (c) or an image prompt (d). It enables users to set a threshold on scores (physical horizontal bar) that highlights (blue) all frames with scores above it.
(**c**)	textual prompts are formatted as (adjectives +) Subject + (adverbs +) Verb + (adjectives +) Object sentences. Each sentence has its own score. These scores can be combined with the boolean operators OR, AND, DIFFERENCE (detail in Appendix B).
(**d**)	image inputs can either be external or screenshots from the currently displayed frame in (a).
(**e**)	displays the video frames embeddings after being reduced using t-SNE from 512D to 2D. Several color maps can be applied to (e): cluster map (DBSCAN), timestamp map and score map (current). Users can select frames with a sliding window.

(**f**) Description of features

Figure 2. Video content retrieval system interface components for video navigation and video analysis leveraging vision–language model CLIP. (**a**), (**b**), (**c**), (**d**), and (**e**) represent the features of the system. Their functioning is explained in (**f**). An example video content retrieval is highlighted where textual input paired with score selection and semantic refinement is used to find moments where a person waters their bedroom plants.

3.4.1. Disambiguation

The two principal interactions that help disambiguating a polysemous query are image-based search and semantic plot navigation. When a term like "plant" can refer to several types of flowers or trees, an image gives not only the exact species that a user may look for but also the exact instance, differentiating between plants in a bedroom and on a balcony, for instance.

If no image can be provided, isolating initial results through textual search and score thresholding, then visualizing said results in a semantic representation of the video shows the different instances of the "plant" entity through different semantic clusters of video frames.

3.4.2. Incompleteness Refinement

The key principle to resolve incompleteness in our proposed framework is incremental querying and modality combination. Indeed, users tend to ask more than a single query when dealing with search tasks and reformulate their queries to be more precise over time [93].

Egocentric videos, by nature, give users an initial hint; users often remember approximate moments, environments, or objects, even if they lack precise terms, giving a starting point for textual querying. These can be refined by complementary exploration through temporal and semantic navigation to identify further content and ask more precise textual queries or direct image queries from video screenshots of objects of interest.

4. Experimental Validation

The experimental setup was designed with two goals in mind, (1) to extract user strategies for navigating videos with vague queries and (2) to evaluate the usability of the proposed framework by comparing it to baseline navigation methods.

Participants were tasked with finding specific content in videos based on textual scenarios, each linked to one video. They identified the moments that best matched the scenarios. Each participant used two frameworks: the proposed content retrieval framework and a standard video navigation framework (similar to video players like YouTube (https://www.youtube.com/, accessed on 10 June 2025) or VLC (https://www.videolan.org/vlc/, accessed on 10 June 2025)). Participants were divided into two groups, with the order of framework usage varying between groups.

Figure 3 outlines the experiment process. After providing written consent, participants received an overview of the task and a short video introducing the framework based on the session and group. Then, they practiced using each framework with a familiarization scenario (S0). This served two purposes: (1) familiarize participants with the nature of the task and (2) train them on each framework. Testing followed, with three scenarios per session (S1–S3 in session 1; S4–S6 in session 2) presented in a fixed order to measure participants' adaptation to the sequence of scenarios. The experiment concluded with a feedback questionnaire on framework and feature appreciation.

A total of eight participants (M = 7, F = 1), aged 18 to 34 (mean = 25.7), were recruited through word of mouth. Most participants were master's or PhD students, except for one research engineer and one fishmonger. No compensation was offered for participation.

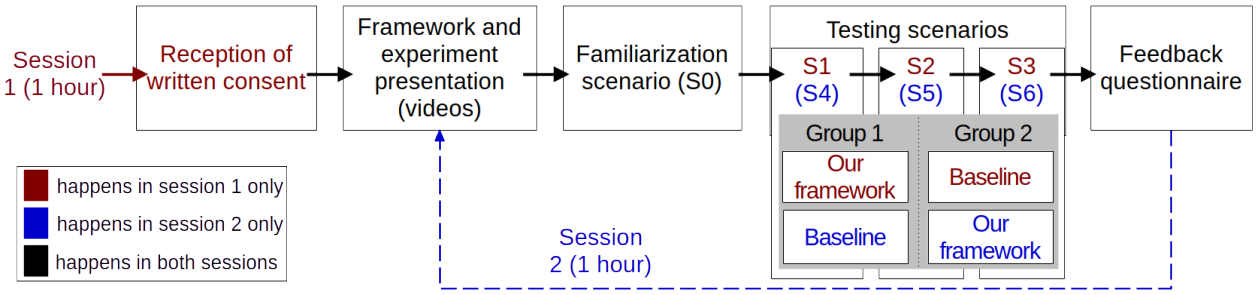


Figure 3. Unfolding of the experimental study for each participant. After reception of written consent, the study is performed in two sessions of one hour each. Each session starts with a presentation of the experiment and the framework to be used. Participants have a non-tested scenario (S0) to become familiar with the framework and experiment modality (video navigation to retrieve content based on a written scenario). Then, participants answer three scenarios in a fixed order (S1 to S3 in session 1, S4 to S6 in session 2). Each session concludes with a feedback questionnaire.

4.1. Scenario Design and Use Cases

Participants were exposed to a total of seven scenarios: one familiarization scenario (referred to as S0) and six testing scenarios. Each scenario is paired with a video. All videos were selected from the Ego4d dataset [14] based on their content (no two scenarios with the same tasks) and their length (less than 5 min for S0 and 20 to 50 min for S1–6, with the possibility to cut some videos). Scenarios were presented in the following way:

1. The title of the scenario;
2. A short, textual description to describe the context of the scenario;
3. A timeline showing a few snapshots of the video. The snapshots were representative images taken from the middle of each annotated segment (the description annotations of the Ego4d dataset); and
4. The instruction that needs to be matched with video moments.

A mixture of specific- and vague-query-inducing scenarios were used to test the incremental querying framework's usability. The exact formulation of each scenario and its expected type of queries can be found in Table 1 and detail of the associated videos can be found in Appendix A.

This setup aimed to simulate both egocentricity of the videos and vagueness for participants.

Egocentricity and simulated memory: We aim to understand how human users navigate through egocentric videos based on vague semantic queries. For that, ideally, participants would navigate their own egocentric video recordings. However, such a setup introduces uncontrollable variability as to the level of detail remembered by each participant as well as the content and length of the videos, which complicates comparison. Thus, we aimed for a more controlled experimental setup where we simulate the same level of familiarity for all participants with already existing egocentric videos from the Ego4d dataset through contextual cues and image snapshots that come with each scenario.

Vagueness in query formulation and navigation: Without any knowledge of the video, navigation would be entirely random, as participants would have no clue about where or how to search. Conversely, perfect awareness of the video content would eliminate ambiguity, allowing participants to directly locate the matching moment. To ensure this balance, participants were asked beforehand about their familiarity with the dataset, and all confirmed they had never encountered it.

To create a sense of vagueness, participants were provided with approximate video segments, representative images, and rough segment descriptions, encouraging strategic exploration rather than precise targeting or random navigation.

A multi-pass process was applied to annotate the moments for each scenario. The first pass consisted of initial annotations based on scenarios. Then, a second pass of correcting of missing and false moments was performed. Instantaneous moments are relaxed by turning them into intervals while long intervals are compressed and/or subdivided in a third pass. A fourth pass of verification concluded the annotation process.

4.2. Data Collection and Analysis

The data considered for evaluation are as follows:

Completion Time: Automatically recorded via framework logs. Only interaction time is accounted for (any loading pre-computation time is ignored for both frameworks)

Retrieved Moments: Participants submitted responses via a web form for comparison with expert annotations.

Participants' feedback: Participants used a seven-point Likert scale to evaluate intuitiveness, efficiency, satisfaction, and likelihood of recommending each framework, as well as the ease of use, speed, efficiency, and satisfaction of each feature.

To evaluate whether the proposed incremental querying framework offers a practical advantage over standard navigation approaches, we measure effectiveness (precision, recall), efficiency (completion time) and user satisfaction (user feedback). Precision and recall are computed from comparing participants moments with expert annotations and descriptive statistics are performed on completion time, precision, recall and user feedback.

To compare features and extract user strategies, descriptive statistics on user feedback were performed and a semi-structured approach, combining in-session notes and log analysis, was employed to understand participants' strategies. Participants' workflows were visualized in subway-graph format to analyze user strategies (Section 5.2).

4.3. Implementation Details and Computational Cost

The experiment was performed on an ASUS TUF A15 FA507XI laptop (accompanied with a Lenovo MOJUUO mouse and Iiyama XUB2529WSU screen) with the following specifications: an Nvidia GeForce RTX 4070 (8GB GRAM) GPU, an AMD Ryzen 9 7940HS w/ Radeon 780M Graphics × 16 CPU, 16 GB of RAM and an SSD-type internal storage of 512 GB. The employed operating system was Ubuntu 22.04.03.

Our proposed framework was implemented as a Python backend to provide the VLM and video processing and an HTML5/Vanilla JS frontend to interface with the backend and provide video interactions to the users. Python (version 3.8), Cuda (version 12.1), and Torch (version 3.8) were used.

For comparability, the standard video navigation framework was also implemented in HTML5/Vanilla JS. Its features included reading the video with both mouse and keystrokes (play/pause), speeding up/slowing down the video with keystrokes, pinpointing frames with the mouse, and selecting video frames on a physical timeline, similar to what is depicted in Figure 2a.

Table 2 details the computational resources and mean time consumption when querying and processing the video of Scenario 0 (duration of 4 min 8 s, fixed to 10 FPS) using our proposed incremental querying video navigation framework. Frontend interactions (click/drag and drop on similarity threshold, click/drag and drop on timeline, click drag and drop/selection on semantic plot) using this same video take around 5–10 ms to execute,

making them usable in real time. This time increases the longer the video is but it still remains under 50ms even for the longest video of our selected scenarios.

Table 2. Computational resources and time consumption for querying video of Scenario 0 in the implementation of our proposed incremental querying video navigation framework. The values are averaged from 100 runs.

Consumption in	Video Loading	Model Loading	Video Embeddings	Text/Image Embedding	Inference
Time	60 ms	0.35 s	55.85 s	3.503 ms	0.813 ms
Memory	5.75 MB (RAM) 183 MB (SSD)	1.18 GB (GRAM)	2.456 GB (RAM)	8 MB (RAM)	/

5. Results

This section presents and analyzes the findings of our user study. We first compare the proposed incremental querying framework to a baseline video player using standard performance metrics (completion time, precision, recall) and user-reported satisfaction. The aim of this comparison is to validate the incremental querying framework as a vague-query-based egocentric video navigation framework. Then, we analyze how participants engaged with the system's features, on one hand, and what global strategies they employ when dealing with vague video content retrieval scenarios, on the other hand, based on interaction logs and structured feedback.

5.1. Quantitative Results

Table 3 shows the mean values for performance metrics between both our proposed incremental querying framework and the baseline. The average completion time for both frameworks (9.375 min for incremental querying, 9.042 min for standard) is almost equal. Average precision is nearly identical for both frameworks (0.632 for incremental querying, 0.63 for standard), but recall is markedly higher for standard navigation (0.398 for incremental querying, 0.514 for standard). This suggests a key behavioral pattern for participants using the incremental querying framework: they prioritize precision over recall, filtering out potentially relevant frames that they perceive as less certain, likely due to their interaction with the features and the model's limitations.

We further examine the details for individual scenarios.

Table 3. Mean performance metrics for both video navigation frameworks (our proposed interactive and incremental framework VS standard video navigation as baseline). Mean value is computed across participants and scenarios.

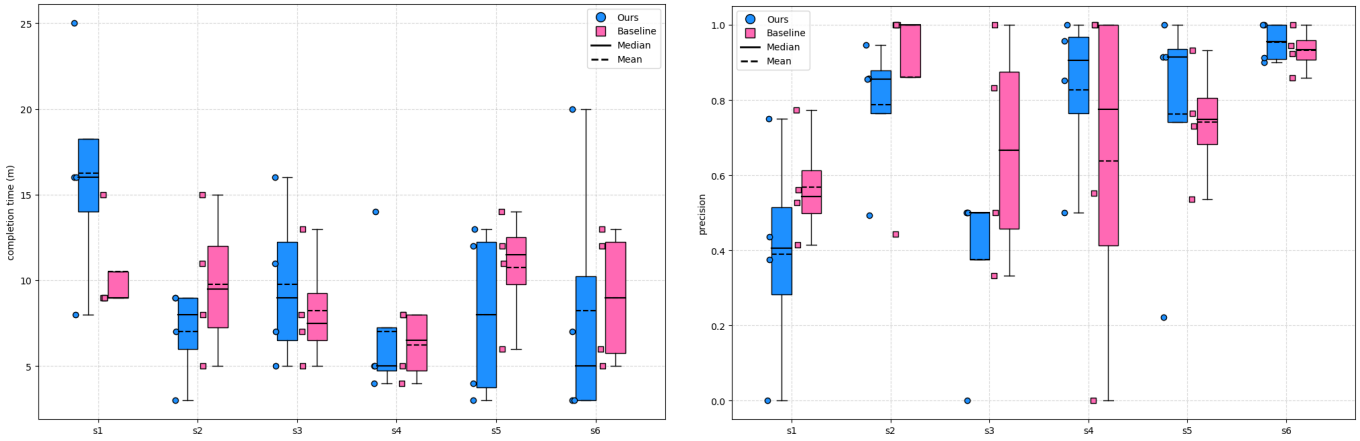
Framework	Completion Time (m)	Precision	Recall
Ours	9.375	0.632	0.398
Baseline	9.042	0.63	0.514

5.1.1. Completion Time

Figure 4a shows the mean completion time for both frameworks across all users for each scenario. For scenarios S2, S5, S6, incremental querying performs better than the standard framework, while for Scenarios S4, S3 and S1, the opposite trend is noticed, with a huge gap of time completion for Scenario S1. Such a trend for S1 is due to learning effect, as S1 was presented first for all participants, the distribution of moments in (widespread and instantaneous compared to other scenarios) and the video length (43 min 25 s, the longest presented video).

5.1.2. Precision and Recall

Figure 4b,c display mean precision and mean recall, respectively, for each scenario based on both frameworks. The incremental querying framework outperforms the standard framework in precision and recall for Scenario S4, a complex task that requires identifying the winner a card game. This highlights the strength of the framework in addressing vague, multi-step scenarios. In contrast, Scenario S3, another multi- step and vague scenario, sees standard navigation performing better, especially in recall. The instructions distinguished between finding an object and identifying when an action occurs on said object, which confused several participants (P1, P2, P4, P6).



(**a**) Completion time (**b**) Precision

(**c**) Recall

Figure 4. Box plots of mean (across scenarios) metrics for comparison between incremental querying (ours) and baseline video navigation frameworks. (**a**) Both frameworks perform similarly, with the exception of S1 where the baseline outperforms our proposed framework by far, likely due to learning effect, distribution of moments in the video and video length. In terms of precision (**b**), both frameworks perform comparably, and overall recall (**c**) is higher for standard navigation, highlighting a trend for users to prioritize precision over recall (filter less relevant results).

5.1.3. Participants' Feedback for Framework Comparison

Figure 5 shows the mean and standard deviation values of the answers for each statement by participants. Overall, participants think that the standard navigation framework is more intuitive than the incremental querying one (over one point difference), a result of the familiarity with standard frameworks for video navigation. When it comes to efficiency,

satisfaction and whether participants would recommend a framework, there is a tendency of preferring the incremental querying framework over the standard navigation framework.

5.1.4. Participants' Feedback for Feature Comparison

Figures 6 and 7 show the mean and standard deviation of the answers to all statements (*feature* is easy to use/ fast/efficient/satisfying) for each feature of the incremental querying framework. The average values (Figure 6) show that the image search feature is the least-liked feature. The detailed values for each category (Figure 7) show that participants think that it is quite intuitive (over 5) but are overall not that satisfied with the performance (average 4), highlighting that the feature is not working as intended and requires further refining.

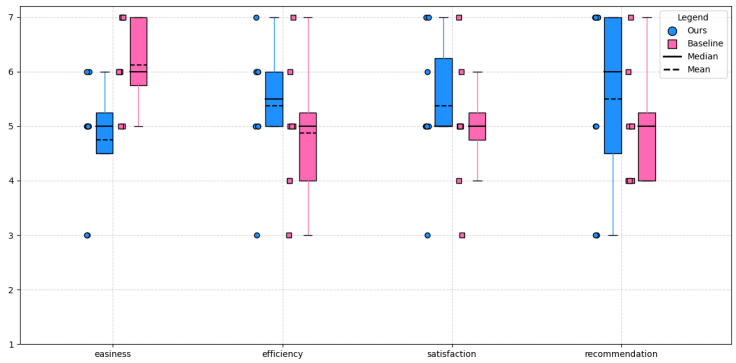
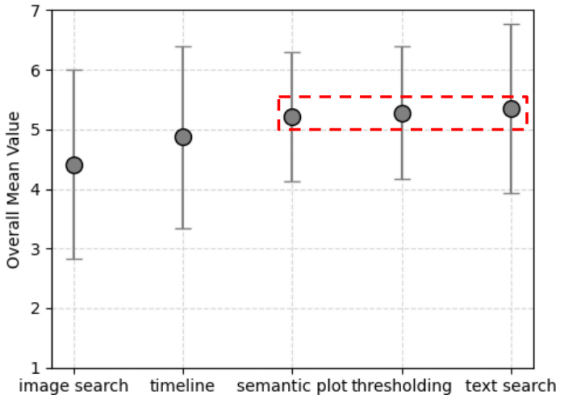


Figure 5. Box plot of participants' global appreciation for comparison of the different features of the incremental querying framework (7-point Likert scale). While participants think that the baseline framework is easier to use due to their familiarity with standard video navigation/playing, participants overall appreciate our framework over baseline methods for navigating egocentric videos.

The three most-liked features are textual search, thresholding and semantic plot navigation. The latter shows high scores of performance appreciation (both speed and efficiency) which displays the value participants put into this feature.



(**a**) Mean scores with std. deviation bars

Features	Mean	Std. Deviation
Image search	4.406	1.588
Timeline	4.875	1.528
Semantic plot	**5.219**	1.084
Thresholding	**5.281**	1.117
Text search	**5.344**	1.421

(**b**) Corresponding numerical values

Figure 6. Comparison of participants' global appreciation of the different features of the incremental querying framework (7-point Likert scale). (**a**) Mean scores with standard deviation bars for each feature, highlighting the top three (text search, thresholding, and semantic plot navigation) within a dashed red box. Notably, semantic plot navigation received comparable appreciation to text-based and thresholding interactions despite its relative complexity. (**b**) Corresponding numeric values with means and standard deviations, confirming the close clustering of top-rated features and the lower appreciation for image search. These values were obtained by computing the mean values per feature in Figure 7.

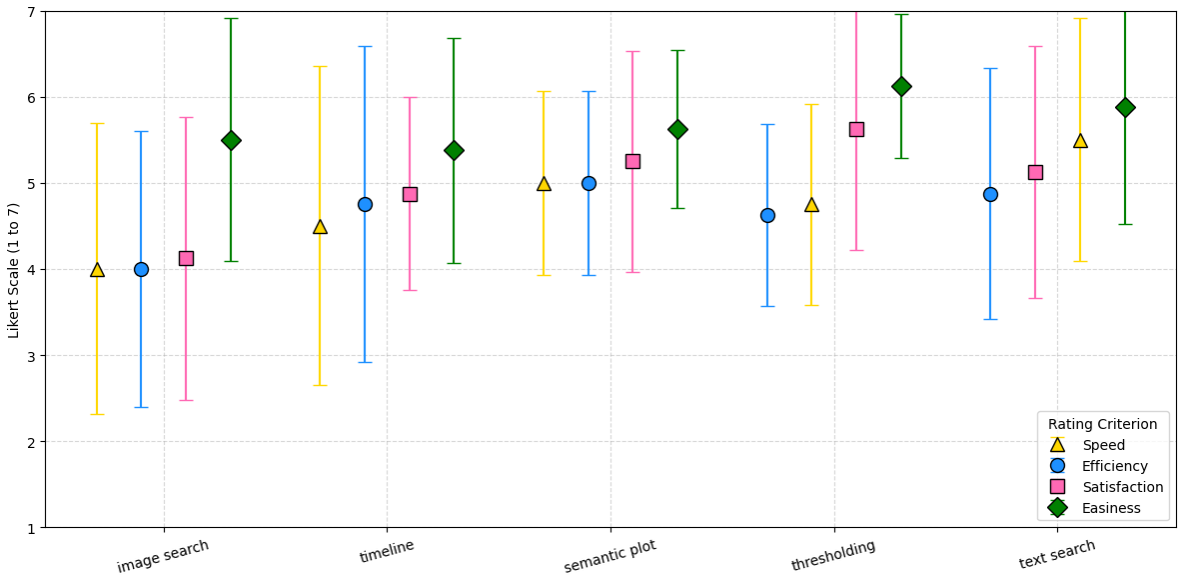


Figure 7. Comparison of participants' category-based appreciation of features in the incremental querying framework (7-point Likert scale). The graph represents mean and std. deviation values for the appreciation of each feature based on 4 categories: ease of use (green), speed (yellow), efficiency (blue) and satisfaction (pink)

5.2. Analysis of Participants Search Strategies

Figure 8 displays the extracted participants' workflow for each scenario. Several key patterns can be identified:

Search initiation: Participants started 79% of scenarios with textual querying while the other 21% of scenarios were initiated with temporal navigation. While the Graphical User Interface (GUI) may have influenced participants' initial choice, this shows the immediacy of textual querying and its accessibility in the design.

Query refinement: Several strategies were employed for refining the initially entered textual queries.

- Thresholding + temporal navigation: a strategy used in 42% of the workflows (across participants and scenarios) that consisted of a back and forth between setting a manual threshold on scores and temporally navigating the frames with scores above said threshold. This strategy was sufficient for P2 for answering Scenario 1, it was used alongside iterative textual querying by P1 for answering Scenario 3 and P5 for answering scenarios 5 and 6 and P7 for answering Scenario 6, and it was augmented by semantic navigation in the other cases.
- Checking scores + temporal navigation: Two participants chose to manually infer the frames to temporally navigate and select from the shape of the score curve instead of using the horizontal bar tool to set a physical threshold that automatically highlights the frames with high scores as shown in Figure 2b. While P4 did not understand the feature, P6 made a deliberate choice to bypass it. This suggests that the manual curve inspection held perceived advantages in clarity or control over the more automated threshold mechanism.
- Semantic navigation: This emerged as a key strategy in 45% of workflows. Some participants used it to supplement temporal methods (P5, P6, P7 and P8), while others (P2 and P4) completely abandoned the temporal score curve for the semantic representation to visualize scores on frames and search for the objects specific to each scenario.

Full temporal navigation: In Scenario 4, P7 solely relied on temporal navigation, i.e., they navigated the video in a standard way (no incremental querying), while P8 attempted a textual query but later relied only on temporal navigation.

Moment selection: Four strategies were observed for moment selection before copying them to the answering web page:

1. Full reliance on manual, temporal-guided moment selection for all relevant moments, then copying (P2, P4, P6, P7, P8);
2. Using the automatic threshold selection function, then refining the results through addition or deletion of specific moments (mainly deletion), then copying (P1, P2 in S1, P5, P7 in S6, P8 is S6);
3. Using the semantic plot selection feature, then refining the results through addition or deletion of specific moments (mainly deletion), then copying (P3, P5);
4. Full reliance on manual, temporal-guided moment selection for a portion of relevant moment, then copying, then repeating the process every time a new textual query and query refinement iteration is performed (P1 in S1).

These varied moment selection strategies reflect differing levels of trust in the system's automatic highlighting and the importance of providing a flexible and multimodal framework for exploratory and verification tasks.

Reliance on temporal navigation: It persisted in 62% of interactions, often in conjunction with other features. P3 thinks that a hybrid video navigation method that combines features of both frameworks would be more efficient, while P6 would have preferred to sequentially navigate the video for some scenarios. This highlights the influence standard video navigation has on users even after the familiarization scenario and the limitations put in the GUI to minimize this effect.

Strategies of semantic navigation: Semantic navigation was used at least once for each scenario with diverse strategies: while P2 and P4 used the semantic plot to pinpoint specific moments or objects, P3 selected areas with a high density of high scores, P5 relied on the semantic plot clusters to identify similar frames, P6 used it as a progression checkpoint, and P8 tried to identify frames with a specific camera angle. This highlights the expression power of the semantic space.

Under-reliance on image querying: Scenario S0 aside, participants hardly ever used image querying, which corroborates the results of Section 5.1.4. P1 and P2 think that it is counter-intuitive, while P5 and P7 were dissatisfied with the returned scores. This confirms that the feature needs to be improved for better disambiguation efficiency.

The observed user strategies reveal notable patterns in addressing query vagueness. For disambiguation challenges, on one hand participants relied heavily on semantic navigation to differentiate between possible interpretations of ambiguous terms—visually exploring clusters in the semantic plot to distinguish between different instances of objects, namely the plants in Scenario 2 and sugary food in Scenario 5, which aligns with our proposed disambiguation mechanism (Section 3.4.1). On the other hand, participants underused the image search feature despite its potential for "instantiating" ambiguous terms.

For incompleteness refinement, the predominant "thresholding + temporal navigation" strategy demonstrates how users iteratively build understanding through incremental exploration, which aligns with our proposed mechanism (Section 3.4.2). The persistence of temporal navigation as an initial entry point for the search while being later supplemented by score and semantic representations suggests participants appreciate the enhancement of familiar interactions by multimodal interactions when dealing with vague scenarios.

These patterns indicate that effective vagueness resolution benefits from multimodal input flexibility and hybrid navigation approaches, confirming the need for frameworks that support incremental query enhancement rather than expecting precise initial formulations.

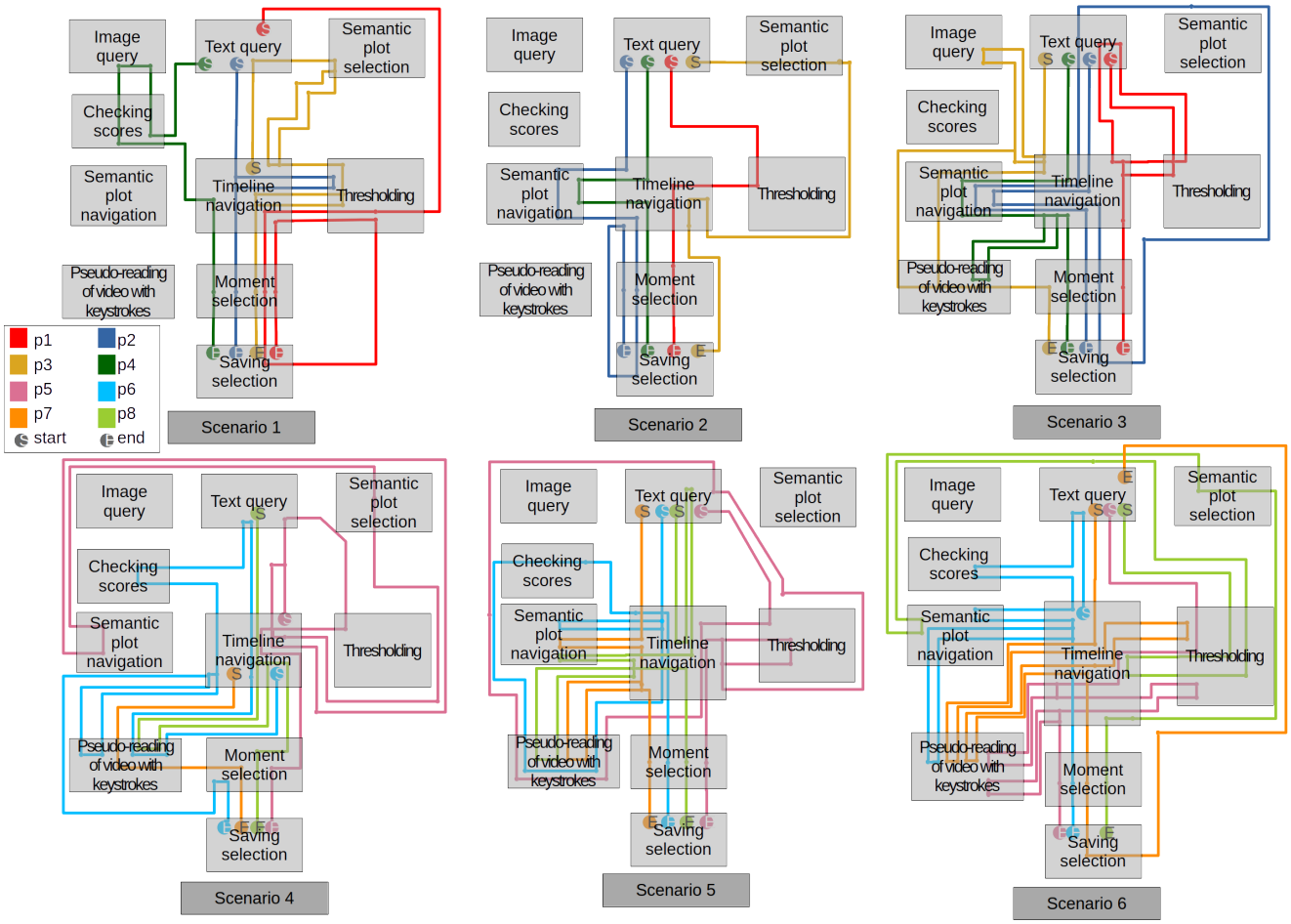


Figure 8. Participants' strategies for answering experimental video navigation scenarios reconstructed from logs and in-session expert notes. Each colored path represents the sequential interaction workflow for each participant (P1–P8) through macro operations, including (but not limited to) the features of our proposed video navigation framework (e.g., text query, timeline navigation, semantic plot selection, etc.). Start and end points of each participant's interaction are marked with 'S' and 'E', respectively. The diagram reveals diverse usage patterns across scenarios, including repeated module visits, reliance on specific query modalities, and variations in navigation behavior. This representation was constructed through an analysis of logged tool interaction events during task performance, piloted by in-session expert notes and user feedback.

6. Discussion

In this work, we defined vagueness in queries based on the nature of the queries themselves as well as the resulting moments. Also, we limited the factors that induce vagueness in queries to either incompleteness or ambiguity. This definition helped us consider a sub-problem of the general concept of vagueness. However, by treating results as a discrete set of moments, our current framing simplifies the temporal nature of video content, which is inherently continuous. While this simplification enabled practical reasoning and implementation, future work could refine the definition to better capture the continuity of moments, as well as consider erroneous queries and other types of vague queries (such as abstract or complex queries) in the design of the egocentric video navigation frameworks.

6.1. Framework Design

Our approach leverages the power of zero-shot content retrieval, using an off-the-shelf vision–language model in an interactive video navigation framework to enable egocentric video navigation based on vague semantic queries. This method proved effective, as evidenced by the performance of our framework. Future works can delve into improving the way the similarity scores are obtained in several ways.

- Engineering prompts to improve the accuracy of queries, especially image-based queries. Multi-sentence textual queries can be divided into several smaller queries and an aggregated score can be computed ([43]). Image queries can benefit from upscaling or transformation into text with generative AI to ensure same modality encoding.
- Testing other vision-language models such as Yolo-world [94] or EgoCLIP [18] (which was pre-trained on egocentric videos) to determine if they improve similarity scores.
- Sound modality: Most videos are audio-visual and a great deal of information is found within the audio modality. The framework would benefit from encoding sound and including it when querying videos.
- Refining score computing using, for instance, image segmentation techniques like SAM [95] to identify and encode individual objects within video frames separately and then computing an aggregated score measure.
- Improving interaction design such as image search or thresholding to function as intended or removing limits on timeline navigation to have a hybrid framework, as mentioned by participants.

Privacy and Security Concerns

While our framework treats the VLM (CLIP) as a frozen, unmodified black box and does not engage with model training or fine-tuning, it is important to acknowledge potential security risks. Vision–language models have been shown to be vulnerable to adversarial prompts and clean-label backdoor attacks in other domains [96], although such threats are not directly addressed in our work. Given that our framework places moment selection in the user's control via interactive refinement, we expect some robustness to these attack vectors in practice, but this remains untested.

More critically, our framework deals with egocentric and potentially personal or sensitive video data, which raises serious concerns regarding data security, consent, and storage. Our current implementation was designed to locally process videos when extracting embedding vectors, but we did not test its scalability for very large videos (4 h+). Also, our current implementation does not address encrypted storage and strong access controls to protect user privacy. Future deployments of such systems must consider secure and scalable solutions for data storage and processing.

6.2. Experimental Validation

Eight participants navigated videos using the proposed incremental querying video navigation framework and the standard video navigation framework. Quantitative results show that the incremental querying framework performs comparably to the standard framework in terms of time and precision. Participants prefer the incremental querying framework, highlighting its usability and effectiveness. The textual search, score thresholding, and semantic plot interactions of the proposed framework are highly valued. Participants' preference for semantic plot interactions indicates effective use of the framework without over-reliance on temporal navigation. Qualitative results and log analysis reveal that participants favor traditional navigation methods (sequential navigation, textual search) for initiating their search. They also tend to take coarse retrieved moments and eliminate outliers over time, benefiting from the expression power of semantic video

representations for this process of moment refinement. Overall, the results provide insights into user approaches to video navigation based on vague information.

These results are preliminary as testing was performed on only eight participants who were all French and consisted mainly of young (18–34 years old) students. This low number was chosen for the sake of facilitating behavior and strategy analysis over comparing methods. Future research should conduct extensive validation with a larger, more diverse participant group in order to apply robust statistical tests (ANOVA [97]) to validate this proposed incremental querying video navigation framework.

Also, our controlled setup had participants navigate egocentric videos that were not their own with simulated knowledge of said videos through contextual and visual ques. Future testing could explore semi-controlled setups where participants film their own first-person videos while performing predefined controlled tasks and are asked to navigate these videos.

Finally, we chose a standard video player as a baseline for evaluating the usability of our solution. Future tests can compare our interactive video content retrieval framework to end-to-end zero-shot video content retrieval frameworks in terms of efficiency in handling vague semantic queries in video content retrieval tasks.

7. Conclusions

This paper investigated user strategies in navigating egocentric videos based on vague semantic queries. First, we limited vagueness to incompleteness and ambiguity. Then, we designed several VLM powered video interactions integrated in a user-centered video content retrieval framework that supports incrementally refining such vague queries. Next, we extracted user strategies for video navigation and evaluated the usability of this framework with standard video navigation frameworks in an experimental video navigation setup. We found that users employ diverse strategies to answer each scenario of video navigation. Overall, they favor usual navigation methods such as sequential navigation and textual search to obtain initial results, favor precision over recall and adopt an elimination process strategy for refining their query's resulting moments, and highly value interactions based on the semantic proximity of frames for both content exploration and result refining. This user strategy analysis highlights the need and the added value of flexible video interactions for incremental query enhancement. We hope this encourages future works to conceive interactive and incremental query-refining video navigation systems, especially works that deal with egocentric video footage obtained from the processing of wearable devices (i.e., cameras, smart glasses, etc.).

Future research could also explore enhancing the framework by incorporating the audio modality, among others, refining the interaction design to further improve usability, expanding the framework to handle a broader range of vague queries and conducting extensive validation with a larger, more diverse participant pool.

Author Contributions: L.A.: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft preparation, and writing—review and editing; W.E.M.-J.: methodology, software, and writing—review and editing; L.N.: conceptualization, validation, writing—review and editing, and supervision; C.J.: conceptualization, validation, resources, writing—review and editing, supervision, project administration, and funding acquisition; C.H.: conceptualization, methodology, validation, resources, writing—review and editing, supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee for Research of the University of Toulouse "CERNI" (project code 2024_880, 8 July 2024).

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The implementation of our solution is accessible in the following Github repository: https://github.com/l-ablaoui/IQVN/releases/tag/v0.0.1, accessed on 28 April 2025. Further inquiries can be directed to the corresponding author.

Acknowledgments: We thank all participants who took part in our experiments. The icons used in Figure 1 were obtained from Flaticon, http://www.flaticon.com, accessed on 22 April 2025) and Pixabay, http://www.pixabay.com, accessed on 22 April 2025). We acknowledge the use of GPT-4 for syntactical/grammatical corrections and reformulations when writing this manuscript. We have reviewed and edited the output and take full responsibility for the content of this publication.

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

Appendix A. Experimental Setup

Table A1 shows the association between the experimental study scenarios (explicated in Table 1) and the Ego4d dataset videos. All videos were used as is, aside from the video of Scenario 6, which was truncated to its first 37 min.

Table A1. Scenarios used in the experimental setup alongside the Ego4D dataset identifier of their associated video.

n°	Title	Ego4D Video ID
0	Camping	b8b4fc74-e036-4731-ae03-4bfab7fd47b9
1	Fire alarm triggers	1bec800a-c3cf-431f-bf0a-7632ad53bcb7
2	Water the plants	24a27df2-892b-4630-909a-6e1b1d3cf043
3	Where are my papers?	1bec800a-c3cf-431f-bf0a-7632ad53bcb7
4	Who won?	915a7fd2-7446-4884-b777-35e6b329f063
5	Diet	e837708f-8276-4d2b-88b3-319cddabbf74
6	Warmups	6395f964-0122-4f16-b036-337c28488504

Appendix B. Textual Query Combination Logic

Textual-query-based search is a common interaction in information retrieval systems. To ensure users do not over-rely on this interaction in the proposed video content retrieval framework (Figures 1 and 2) and to promote the usage of incremental query building, we limit the input to one sentence with the format subject + verb + object with the possibility to add qualifiers to each sentence component. If users want to express more complex concepts, they can add sentences and combine them pair-wise with ensemble operators AND, OR, and DIFFERENCE. When computing the query scores, each sentence is processed individually. When visualizing a score curve, the values reflect the scores of the focused sentence alone. The operators are applied on the frames selected for each score curve. Let there be two sentences E^1 = "potted plants" and E^2 = "bedroom". Let the scores for each sentence be S^1 and S^2, respectively. If the operator to combine both sentences is AND, the resulting selected frames F are the intersection of the selected frames from E^1 using a first threshold t^1 and the selected frames from E^2 using a second threshold t^2, $F = \text{select}(S^1, t^1) \cap \text{select}(S^2, t^2)$. The combination logic is explained in Figure A1.

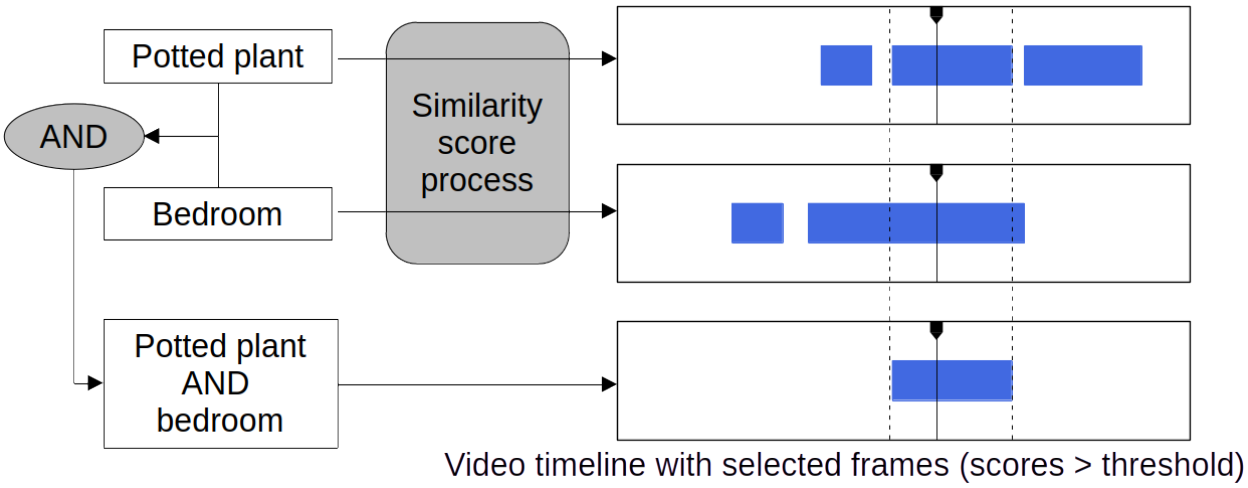


Figure A1. Combination logic for textual queries. Each entered sentence is processed alone and produces its own score curve. A threshold is applied on each score curve and an ensemble operator is applied on the selected frames sets.

References

1. Rodin, I.; Furnari, A.; Mavroeidis, D.; Farinella, G.M. Predicting the future from first person (egocentric) vision: A survey. *Comput. Vis. Image Underst.* **2021**, *211*, 103252. [CrossRef]
2. Huang, S.; Wang, W.; He, S.; Lau, R.W. Egocentric hand detection via dynamic region growing. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *14*, 1–17. [CrossRef]
3. Thakur, S.K.; Beyan, C.; Morerio, P.; Del Bue, A. Predicting gaze from egocentric social interaction videos and imu data. In Proceedings of the 2021 International Conference on Multimodal Interaction, Ottawa, ON, Canada, 18–22 October 2021; pp. 717–722.
4. Legel, M.; Deckers, S.R.; Soto, G.; Grove, N.; Waller, A.; Balkom, H.v.; Spanjers, R.; Norrie, C.S.; Steenbergen, B. Self-Created Film as a Resource in a Multimodal Conversational Narrative. *Multimodal Technol. Interact.* **2025**, *9*, 25. [CrossRef]
5. Yin, H. From Virality to Engagement: Examining the Transformative Impact of Social Media, Short Video Platforms, and Live Streaming on Information Dissemination and Audience Behavior in the Digital Age. *Adv. Soc. Behav. Res.* **2024**, *14*, 10–14. [CrossRef]
6. Patra, S.; Aggarwal, H.; Arora, H.; Banerjee, S.; Arora, C. Computing Egomotion with Local Loop Closures for Egocentric Videos. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 454–463. [CrossRef]
7. Zhang, L.; Zhou, S.; Stent, S.; Shi, J. Fine-Grained Egocentric Hand-Object Segmentation: Dataset, Model, and Applications. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 127–145.
8. Damen, D.; Doughty, H.; Farinella, G.M.; Furnari, A.; Ma, J.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis. (IJCV)* **2022**, *130*, 33–55. [CrossRef]
9. Mueller, F.; Mehta, D.; Sotnychenko, O.; Sridhar, S.; Casas, D.; Theobalt, C. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017; pp. 1154–1163.
10. Liu, X.; Zhou, S.; Lei, T.; Jiang, P.; Chen, Z.; Lu, H. First-Person Video Domain Adaptation With Multi-Scene Cross-Site Datasets and Attention-Based Methods. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7774–7788. [CrossRef]
11. Duane, A.; Zhou, J.; Little, S.; Gurrin, C.; Smeaton, A.F. An Annotation System for Egocentric Image Media. In *MultiMedia Modeling*; Amsaleg, L., Guomundsson, G.B., Gurrin, C., Jónsson, B.B., Satoh, S., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 442–445.
12. Wang, M.; Ni, B.; Hua, X.S.; Chua, T.S. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv. (CSUR)* **2012**, *44*, 1–24. [CrossRef]
13. Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C.P.; Wang, X.Z.; Wu, Q.J. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4051–4070. [CrossRef]
14. Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18995–19012.

15. Zhu, C.; Xiao, F.; Alvarado, A.; Babaei, Y.; Hu, J.; El-Mohri, H.; Culatana, S.; Sumbaly, R.; Yan, Z. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 20110–20120.
16. Xue, Z.; Song, Y.; Grauman, K.; Torresani, L. Egocentric video task translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2310–2320.
17. Dai, G.; Shu, X.; Wu, W.; Yan, R.; Zhang, J. GPT4Ego: Unleashing the potential of pre-trained models for zero-shot egocentric action recognition. *IEEE Trans. Multimed.*, **2024**, *27*, 401–413. [CrossRef]
18. Lin, K.Q.; Wang, J.; Soldan, M.; Wray, M.; Yan, R.; Xu, E.Z.; Gao, D.; Tu, R.C.; Zhao, W.; Kong, W.; et al. Egocentric video-language pretraining. *Adv. Neural Inf. Process. Syst.*, **2022**, *35*, 7575–7586.
19. Hong, H.; Wang, S.; Huang, Z.; Wu, Q.; Liu, J. Why only text: Empowering vision-and-language navigation with multi-modal prompts. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Jeju, Republic of Korea, 3–9 August 2024; pp. 839–847.
20. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* **2023**, arxiv:2309.01219.
21. Esfandiarpoor, R.; Menghini, C.; Bach, S. If CLIP Could Talk: Understanding Vision-Language Model Representations Through Their Preferred Concept Descriptions. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 9797–9819.
22. Furnas, G.W.; Landauer, T.K.; Gomez, L.M.; Dumais, S.T. The vocabulary problem in human-system communication. *Commun. ACM* **1987**, *30*, 964–971. [CrossRef]
23. Song, R.; Luo, Z.; Wen, J.R.; Yu, Y.; Hon, H.W. Identifying ambiguous queries in web search. In Proceedings of the 16th International Conference on World Wide Web, New York, NY, USA, 8–12 May 2007; pp. 1169–1170.
24. Asai, A.; Choi, E. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In Proceedings of the ACL-IJCNLP 2021—59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021. [CrossRef]
25. Zarka, M.; Ben Ammar, A.; Alimi, A.M. Fuzzy reasoning framework to improve semantic video interpretation. *Multimed. Tools Appl.*, **2016**, *75*, 5719–5750. [CrossRef]
26. Chae, J.; Kim, J. Uncertainty-based Visual Question Answering: Estimating Semantic Inconsistency between Image and Knowledge Base. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Pajua, Italy, 18–23 July 2022; pp. 1–9. [CrossRef]
27. Chang, M.; Huh, M.; Kim, J. Rubyslippers: Supporting content-based voice navigation for how-to videos. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–14.
28. Pérez Ortiz, M.; Bulathwela, S.; Dormann, C.; Verma, M.; Kreitmayer, S.; Noss, R.; Shawe-Taylor, J.; Rogers, Y.; Yilmaz, E. Watch less and uncover more: Could navigation tools help users search and explore videos? In Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, New York, NY, USA, 14–18 March 2022; pp. 90–101.
29. Zhou, Z.; Ye, L.; Cai, L.; Wang, L.; Wang, Y.; Wang, Y.; Chen, W.; Wang, Y. Conceptthread: Visualizing threaded concepts in MOOC videos. *IEEE Trans. Vis. Comput. Graph.* **2024**, *31*, 1354–1370. [CrossRef]
30. Fong, M.; Miller, G.; Zhang, X.; Roll, I.; Hendricks, C.; Fels, S.S. An Investigation of Textbook-Style Highlighting for Video. In Proceedings of the Graphics Interface, Victoria, BC, Canada, 1–3 June 2016; pp. 201–208.
31. Dragicevic, P.; Ramos, G.; Bibliowitcz, J.; Nowrouzezahrai, D.; Balakrishnan, R.; Singh, K. Video browsing by direct manipulation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 5–10 April 2008; pp. 237–246.
32. Clarke, C.; Cavdir, D.; Chiu, P.; Denoue, L.; Kimber, D. Reactive video: Adaptive video playback based on user motion for supporting physical activity. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, 20–23 October 2020; pp. 196–208.
33. Lilija, K.; Pohl, H.; Hornbæk, K. Who put that there? temporal navigation of spatial recordings by direct manipulation. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–11.
34. Wu, Y.; Xie, X.; Wang, J.; Deng, D.; Liang, H.; Zhang, H.; Cheng, S.; Chen, W. Forvizor: Visualizing spatio-temporal team formations in soccer. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 65–75. [CrossRef] [PubMed]
35. Chan, G.Y.Y.; Nonato, L.G.; Chu, A.; Raghavan, P.; Aluru, V.; Silva, C.T. Motion Browser: Visualizing and understanding complex upper limb movement under obstetrical brachial plexus injuries. *IEEE Trans. Vis. Comput. Graph.* **2019**, *26*, 981–990. [CrossRef]
36. He, J.; Wang, X.; Wong, K.K.; Huang, X.; Chen, C.; Chen, Z.; Wang, F.; Zhu, M.; Qu, H. Videopro: A visual analytics approach for interactive video programming. *IEEE Trans. Vis. Comput. Graph.* **2023**, *30*, 87–97. [CrossRef]
37. Xian, Y.; Schiele, B.; Akata, Z. Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4582–4591.

38. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Zero-Shot Video Question Answering via Frozen Bidirectional Language Models Cross-modal Training. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022.
39. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Learning to Answer Visual Questions from Web Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *47*, 3202–3218. [CrossRef]
40. Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Chen, G.; Pei, B.; Zheng, R.; Wang, Z.; Shi, Y.; et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 396–416.
41. Ma, K.; Zang, X.; Feng, Z.; Fang, H.; Ban, C.; Wei, Y.; He, Z.; Li, Y.; Sun, H. LLaViLo: Boosting Video Moment Retrieval via Adapter-Based Multimodal Modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 2798–2803.
42. Panta, L.; Shrestha, P.; Sapkota, B.; Bhattarai, A.; Manandhar, S.; Sah, A.K. Cross-modal Contrastive Learning with Asymmetric Co-attention Network for Video Moment Retrieval. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 4–8 January 2024.
43. Luo, D.; Huang, J.; Gong, S.; Jin, H.; Liu, Y. Zero-Shot Video Moment Retrieval from Frozen Vision-Language Models. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2024.
44. Jiang, X.; Zhou, Z.; Xu, X.; Yang, Y.; Wang, G.; Shen, H.T. Faster video moment retrieval with point-level supervision. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 1334–1342.
45. Ma, Y.; Qing, L.; Li, G.; Qi, Y.; Sheng, Q.Z.; Huang, Q. Retrieval Enhanced Zero-Shot Video Captioning. *arXiv* **2024**, arxiv:2405.07046
46. Wu, G.; Lin, J.; Silva, C.T. IntentVizor: Towards Generic Query Guided Interactive Video Summarization. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 10493–10502. [CrossRef]
47. Pang, Z.; Nakashima, Y.; Otani, M.; Nagahara, H. Unleashing the Power of Contrastive Learning for Zero-Shot Video Summarization. *J. Imaging* **2024**, *10*, 229. [CrossRef]
48. Ahmad, S.; Chanda, S.; Rawat, Y.S. Ez-clip: Efficient zeroshot video action recognition. *arXiv* **2023**, arXiv:2312.08010.
49. Yu, Y.; Cao, C.; Zhang, Y.; Lv, Q.; Min, L.; Zhang, Y. Building a Multi-modal Spatiotemporal Expert for Zero-shot Action Recognition with CLIP. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 9689–9697.
50. Zhao, X.; Chang, S.; Pang, Y.; Yang, J.; Zhang, L.; Lu, H. Adaptive multi-source predictor for zero-shot video object segmentation. *Int. J. Comput. Vis.* **2024**, *132*, 3232–3250. [CrossRef]
51. Chu, W.H.; Harley, A.W.; Tokmakov, P.; Dave, A.; Guibas, L.; Fragkiadaki, K. Zero-Shot Open-Vocabulary Tracking with Large Pre-Trained Models. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 4916–4923. [CrossRef]
52. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Westminster, UK, 18–24 July 2021; PMLR, Proceedings of Machine Learning Research, Volume 139, pp. 8748–8763.
53. Tang, H.; Liang, K.J.; Grauman, K.; Feiszli, M.; Wang, W. Egotracks: A long-term egocentric visual object tracking dataset. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 75716–75739.
54. Akiva, P.; Huang, J.; Liang, K.J.; Kovvuri, R.; Chen, X.; Feiszli, M.; Dana, K.; Hassner, T. Self-supervised object detection from egocentric videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 5225–5237.
55. Hao, S.; Chai, W.; Zhao, Z.; Sun, M.; Hu, W.; Zhou, J.; Zhao, Y.; Li, Q.; Wang, Y.; Li, X.; et al. Ego3DT: Tracking Every 3D Object in Ego-centric Videos. In Proceedings of the MM 2024 32nd ACM International Conference on Multimedia. Association for Computing Machinery, Melbourne, Australia, 28 October–1 November 2024; pp. 2945–2954. [CrossRef]
56. Cartas, A.; Dimiccoli, M.; Radeva, P. Detecting Hands in Egocentric Videos: Towards Action Recognition. In *Computer Aided Systems Theory—EUROCAST 2017*; Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 330–338.
57. Liu, S.; Tripathi, S.; Majumdar, S.; Wang, X. Joint hand motion and interaction hotspots prediction from egocentric videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3282–3292.
58. Xu, B.; Wang, Z.; Du, Y.; Song, Z.; Zheng, S.; Jin, Q. Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions? In Proceedings of The Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.

59. Xu, Y.; Li, Y.L.; Huang, Z.; Liu, M.X.; Lu, C.; Tai, Y.W.; Tang, C.K. Egopca: A new framework for egocentric hand-object interaction understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 5273–5284.
60. Ng, E.; Xiang, D.; Joo, H.; Grauman, K. You2me: Inferring body pose in egocentric video via first and second person interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9890–9900.
61. Khirodkar, R.; Bansal, A.; Ma, L.; Newcombe, R.; Vo, M.; Kitani, K. Ego-humans: An ego-centric 3d multi-human benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 19807–19819.
62. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the IEEE 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
63. Kong, Y.; Fu, Y. Human action recognition and prediction: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 1366–1401. [CrossRef]
64. Abreu, S.; Do, T.D.; Ahuja, K.; Gonzalez, E.J.; Payne, L.; McDuff, D.; Gonzalez-Franco, M. Parse-ego4d: Personal action recommendation suggestions for egocentric videos. *arXiv* **2024**, arXiv:2407.09503.
65. Fan, C. Egovqa-an egocentric video question answering benchmark dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
66. Bärmann, L.; Waibel, A. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1560–1568.
67. Di, S.; Xie, W. Grounded question-answering in long egocentric videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 12934–12943.
68. Jiang, H.; Ramakrishnan, S.K.; Grauman, K. Single-stage visual query localization in egocentric videos. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 24143–24157.
69. Pramanick, S.; Song, Y.; Nag, S.; Lin, K.Q.; Shah, H.; Shou, M.Z.; Chellappa, R.; Zhang, P. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 5285–5297.
70. Hummel, T.; Karthik, S.; Georgescu, M.I.; Akata, Z. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 1–17.
71. Sanderson, M. Ambiguous queries: Test collections need more sense. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 20–24 July 2008; pp. 499–506.
72. Dhole, K.D.; Agichtein, E. Genqrensemble: Zero-shot llm ensemble prompting for generative query reformulation. In Proceedings of the European Conference on Information Retrieval, Glasgow, UK, 24–28 March 2024; pp. 326–335.
73. Hambarde, K.A.; Proença, H. Information Retrieval: Recent Advances and Beyond. *IEEE Access* **2023**, *11*, 76581–76604. [CrossRef]
74. Woo, S.; Jeon, S.Y.; Park, J.; Son, M.; Lee, S.; Kim, C. Sketch-based video object localization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2024; pp. 8480–8489.
75. Miyanishi, T.; Hirayama, J.i.; Kong, Q.; Maekawa, T.; Moriya, H.; Suyama, T. Egocentric video search via physical interactions. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
76. Lee, D.; Kim, S.; Lee, M.; Lee, H.; Park, J.; Lee, S.W.; Jung, K. Asking Clarification Questions to Handle Ambiguity in Open-Domain QA. In Proceedings of The 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023.
77. Nakano, Y.; Kawano, S.; Yoshino, K.; Sudoh, K.; Nakamura, S. Pseudo ambiguous and clarifying questions based on sentence structures toward clarifying question answering system. In Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, Dublin, Ireland, 26 May 2022; pp. 31–40.
78. Gao, J.; Sun, C.; Yang, Z.; Nevatia, R. Tall: Temporal activity localization via language query. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017; pp. 5267–5275.
79. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
80. Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; Pinkal, M. Grounding action descriptions in videos. *Trans. Assoc. Comput. Linguist.* **2013**, *1*, 25–36. [CrossRef]
81. Rothe, A.; Lake, B.M.; Gureckis, T.M. Do people ask good questions? *Comput. Brain Behav.* **2018**, *1*, 69–89. [CrossRef]
82. Liu, S.; Yu, C.; Meng, W. Word sense disambiguation in queries. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM '05, New York, NY, USA, 31 October–5 November 2005; pp. 525–532. [CrossRef]
83. Wang, Y.; Agichtein, E. Query ambiguity revisited: Clickthrough measures for distinguishing informational and ambiguous queries. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA, 2–4 June 2010; pp. 361–364.

84. Aliannejadi, M.; Zamani, H.; Crestani, F.; Croft, W.B. Asking clarifying questions in open-domain information-seeking conversations. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 475–484.
85. Keyvan, K.; Huang, J.X. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *ACM Comput. Surv.* **2022**, *55*, 1–40. [CrossRef]
86. Hodges, S.; Williams, L.; Berry, E.; Izadi, S.; Srinivasan, J.; Butler, A.; Smyth, G.; Kapur, N.; Wood, K. SenseCam: A retrospective memory aid. In Proceedings of the UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006, Orange County, CA, USA, 17–21 September 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 177–193.
87. Gurrin, C.; Smeaton, A.F.; Doherty, A.R. Lifelogging: Personal big data. *Found. Trends® Inf. Retr.* **2014**, *8*, 1–125. [CrossRef]
88. Schacter, D.L. The seven sins of memory: Insights from psychology and cognitive neuroscience. *Am. Psychol.* **1999**, *54*, 182. [CrossRef] [PubMed]
89. Diwan, A.; Peng, P.; Mooney, R. Zero-shot Video Moment Retrieval with Off-the-Shelf Models. In Proceedings of the Transfer Learning for Natural Language Processing Workshop, PMLR, 2023; pp. 10–21. Available online: https://proceedings.mlr.press/v203/diwan23a.html (accessed on 28 April 2025).
90. Sultan, M.; Jacobs, L.; Stylianou, A.; Pless, R. Exploring CLIP for Real World, Text-based Image Retrieval. In Proceedings of the 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 27–29 September 2023; pp. 1–6. [CrossRef]
91. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
92. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the KDD, Portland, OR, USA, 2–4 August 1996; Volume 96, pp. 226–231.
93. Chen, J.; Mao, J.; Liu, Y.; Zhang, F.; Zhang, M.; Ma, S. Towards a better understanding of query reformulation behavior in web search. In Proceedings of the Web Conference, Ljubljana, Slovenia, 19–23 April 2021; pp. 743–755.
94. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. Yolo-world: Real-time open-vocabulary object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 16901–16911.
95. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 4015–4026.
96. Huang, X.; Gao, Y.; He, S.; Xie, X.; Wang, Y.; Zhang, Y.; Liu, X.; Li, B.; Liu, Y. Exploring Clean-Label Backdoor Attacks and Defense in Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 12345–12354.
97. St, L.; Wold, S. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* **1989**, *6*, 259–272.