# Structured Entity Extraction from Travel Videos Using Vision-Language Models

Kevin Dela Rosa[1]

[1]Aviary Labs, New York City, New York

## Abstract

In this study, we explore the potential of leveraging vision-language models (VLMs) for structured entity extraction from travel videos, a critical task in enhancing tourism recommendations and improving travel analytics. Travel videos are rich in diverse content, covering aspects such as locations, transportation, dining, and cultural experiences, making them ideal candidates for structured information extraction. We propose a zero-shot prompting framework that utilizes VLMs to extract entities and their attributes across multiple modalities, including visual content, on-screen text, and spoken language. Our approach is demonstrated through a pilot study involving thousands of travel-oriented YouTube videos, showing that over 95% of scenes contain extractable entities. The extracted structured information is subsequently integrated into a video AI agent-based chatbot, capable of interacting with users to provide video-enriched, contextually relevant recommendations. Additionally, this structured information enhances search and discovery capabilities, offering users more precise and relevant access to travel content. This work highlights the potential of structured entity extraction in improving video search, discovery, and sentiment analysis, and lays the groundwork for future research in fine-tuning multimodal models for specific video domains.

## Keywords

Travel Video Analytics, Video AI Agents, Video-Based Recommender Systems, Multimodal Entity Extraction

## 1. Introduction

Videos have become an increasingly influential medium, heavily relied upon for purchase decisions and inspiration. Travel videos, in particular, play a crucial role in inspiring individuals to plan trips [1, 2] and in promoting points of interest across various destinations [3]. These videos are typically information-dense, covering a wide range of aspects such as points of interest, modes of transportation, event recommendations, and culinary experiences. Understanding this content is crucial for various downstream applications, including activity recommendation, discovery, trip planning, location-based search, and sentiment analysis.

Video understanding in the era of Vision-Language Models (VLMs) remains a relatively nascent field. Much of the focus has been on object or visual entity recognition (e.g., [4]) and video question answering (e.g., [5]), with less attention given to the extraction of fine-grained structured information, such as entity or attribute extraction, particularly from non-vision modalities relevant to videos. Text information from visual or spoken modalities has historically proven useful in applications such as video retrieval [6] and video recommendation [7]. In the tourism domain, information extraction has been explored in various studies, particularly in the natural language space. For instance, travel-related entities have been extracted from travel blogs [8] and emails [9], and more recently, aspect-based sentiment analysis has been applied to detect food-related videos [10]. Additionally, vision and language foundation models have been utilized in event recommendation [11].

Structured information extraction is valuable for many downstream applications. For example, having a multifaceted view of different entities can aid in filtering the most relevant data segments for a user segment or retrieval query (e.g., filtering by location or availability based on the time of year). The interplay of multiple modalities in video content makes structured entity extraction an intriguing

*Workshop on Recommenders in Tourism (RecTour 2024), October 18th, 2024, co-located with the 18th ACM Conference on Recommender Systems, Bari, Italy.*
kdr@aviaryhq.com (K. D. Rosa)
0009-0006-0506-3707 (K. D. Rosa)

challenge across domains. Given the critical role of video content in a wide array of applications, particularly within the travel domain, it becomes essential to delve into specific methodologies for extracting meaningful entities from this rich multimedia source, thereby enabling more effective analysis and interaction with video data. In this paper, we propose a zero-shot prompting-based approach for video entity recognition and structured information extraction. We conduct a pilot study involving thousands of travel-oriented YouTube videos to demonstrate the wealth of information that can be extracted per scene or key frame, and we showcase example outputs. We then demonstrate how this extracted information can be leveraged by a Video AI Agent-based chatbot to answer travel-related queries with video-enriched responses, providing relevant structured information for comprehensive user interactions.

## 2. Domain-Specific Video Entity Extraction via Vision-Language Models

Entity extraction is the task of identifying key objects or settings that occur within a document. In the context of video, these entities can be present in various modalities: they may appear visually within the scene, as text displayed on the screen, or be mentioned during dialogue. For example, in the travel domain, relevant entities may include locations, food, transportation modes, accommodations, or activities.

In structured entity extraction, the goal extends beyond identifying entities to extracting key attributes or aspects related to them. For instance, in a food review video, relevant attributes might include the type of dish, the cuisine, flavor profile, presentation, or even subjective perceptions such as the narrator's sentiment or commentary.

The different modalities present in video content offer unique clues for entity extraction. Speech, for instance, can provide detailed descriptions of attributes such as cultural significance or accessibility. On-screen text might directly present the entity or location name as an overlay, while visual analysis can offer insights into color, material, or demographic information. Visual cues can also suggest attributes like popularity, inferred from indicators such as the level of crowding in a scene.

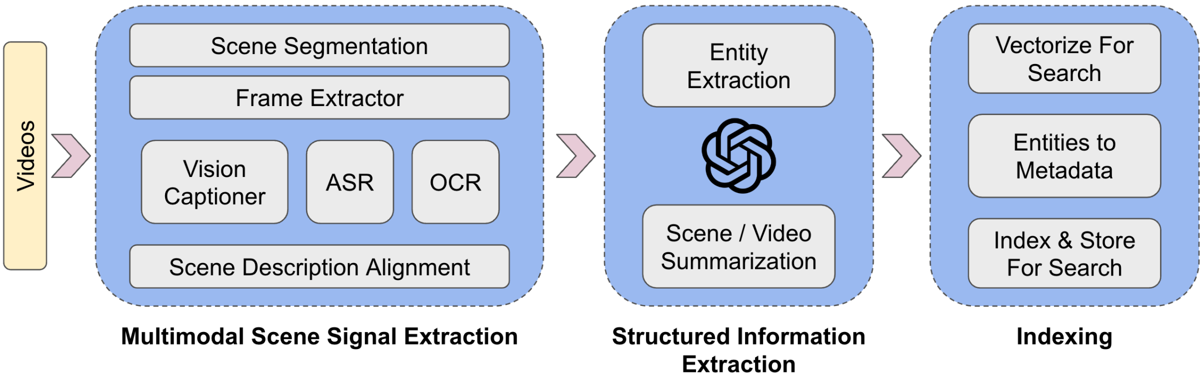


**Figure 1:** Overview of the structured information extraction process from videos, detailing multimodal signal extraction, entity extraction, and indexing for search.

In this study, we propose a general framework that leverages vision-language models to perform structured information extraction from videos in a zero-shot prompting fashion, as outlined in Figure 1.

The overall processing system can be broken down into three stages:

1. **Multimodal Scene Signal Extraction**: This initial stage processes videos by performing segmentation to extract meaningful time boundaries, identifying key frames, and running the content through a set of general signal extractors. In our case, these extractors include a dense vision captioner, a speech recognizer, and an optical character recognizer (OCR) to transcribe text visible in the scene. The output of this stage is an aligned scene transcript—a representation that extends

the work of [12] to include OCR text, inspired by works like [13] which have found OCR useful for structured information extraction, and other signals extracted per temporally aligned scene.

2. **Structured Information Extraction**: The aligned scene transcripts from each scene are fed into a vision-language model (VLM), such as OpenAI's GPT-4, with a prompt requesting the extraction of specific types of entities along with their corresponding attributes, as per a predefined schema. An example of an abbreviated prompt is shown in Figure 2. Additionally, concatenated aligned scene transcripts for all scenes within a video are fed into the VLM to produce video-level summaries, categorize topics, and extract primary named entities (e.g., Points of Interest in travel videos), with the top occurring video locations shown in Figure 7
3. **Indexing**: This final stage varies by use case but generally involves indexing the structured information and various text blobs or multimodal content so they can be queried in the desired format. Typically, this involves vectorizing relevant pieces of information for search, parsing out attributes or entity names to use as metadata or filters (useful if your database representation is SQL-like or can leverage filters, as demonstrated in [14]), and storing the resulting vectors and metadata in an appropriate datastore.

```
Given the textual description of a video scene, provide a list of entities
that appear or are mentioned in the scene, along with structured information
for each:

- Locations: a list of uniquely identifiable locations
- Activities: a list of travel-related activities
...

For each entity, extract the following attributes if present in this JSON format:
{
"Locations": [
{"Value": "Canonical name of the entity; should be uniquely identifiable"},
{"Category": "Category of the location (e.g., city, landmark, natural feature)"},
{"Commentary": "Narrator's review or commentary on the entity"},
{"Sentiment": "Narrator's general sentiment towards the entity"},
{"GeographicalFeatures": "Notable physical characteristics (e.g., mountains,
    rivers)"},
{"CulturalSignificance": "Cultural or historical importance (e.g., historical
    site)"},
{"Accessibility": "Ease of access (e.g., public transport, walking distance)"},
{"Popularity": "Tourist popularity (e.g., tourist hotspot, hidden gem)"}
],
"Activities": [...],
...
}
```

**Figure 2:** Snippet of Vision Language Model Prompt for Extracting Entities from Aligned Scene Transcript

## 3. Travel Video Use Cases

Travel videos are often information-dense, making them an ideal use case for structured information extraction. Structured information extracted from a large video collection can unlock several valuable applications.

1. **Search & Discovery**: Structured information enables the provision of unsupervised tags, allowing users to browse or filter video collections by fine-grained concepts. For example, users can filter

by place types (e.g., national parks), specific locations (e.g., Miami, Florida), dishes (e.g., biryani), or modes of transportation (e.g., auto rickshaw).

2. **Sentiment Analysis**: By extracting general sentiment associated with specific entity types, it is possible to determine which foods, activities, or locations are generally reviewed positively or negatively.

To demonstrate the utility of our approach, Section 3.1 discusses an initial pilot study for extracting structured information from travel YouTube videos. Further, Section 3.2 showcases the utility of this information in a retrieval-augmented generation-based video AI agent that interacts with travel videos.

### 3.1. Structured Entity Extraction from Travel YouTube Videos

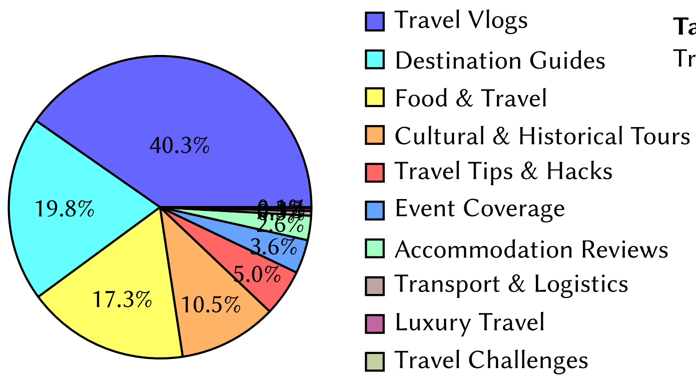


**Figure 3**
Distribution of Travel Video Content Categories

**Table 1**
Travel YouTube Video Dataset

| Dimension | Count |
|---|---|
| Videos | 1720 |
| Total Video Seconds | 1,322,249 |
| Scenes | 188,209 |
| Scenes w/ Entities | 179,096 |

In this study, we collected 1,720 YouTube videos from the Travel category, amounting to roughly 370 hours of footage. We extracted scene-level entities across five travel-specific categories: Accommodations, Activities, Food, Locations, and Transportation. A summary of the video dataset is presented in Table 1. The overall count of scenes containing specific entity types, along with the distribution of these extracted entities, is illustrated in Figures 5 and 6, respectively. The high proportion of scenes with one or more entities (approximately 95%) underscores the potential for fine-grained retrieval and content discovery tasks. To further illustrate the utility of the extracted structured information, we provide a sample output for entities generated using the procedure described in Section 2, as shown in Figure 4. Scene segmentation was performed using PySceneDetect [15] with adaptive visual detection and a minimum clip length of 2 seconds. The vision language model employed for structured information extraction was OpenAI's GPT-4o [16].

Additionally, we used the VLM to automatically categorize the videos using the aligned scene transcripts into subcategories, as shown in Figure 3, to understand the distribution of topics covered in the videos. We further utilized the concatenated video-level aligned scene transcripts as input to extract video-level summaries and primary named entities, which were then employed in the Video AI agent described in Section 3.2.

### 3.2. Leveraging Structured Information in Video AI Agents

To demonstrate the potential real-world applications of the methods and dataset described in this study, we developed a prototype Vision AI agent-based chatbot for interacting with the collection of Travel videos described in Section 3.1. Example outputs for two user queries are shown in Figure 8.

In this system, user messages are processed by an agent that orchestrates sub-prompts to generate the final output:

1. The query is analyzed to determine the appropriate type of structured output to return (e.g., points of interest, activity, food review, etc.).

```
"Value": "Hainanese Chicken Rice",
"Category": "dish",
"Commentary": "A well-known dish featuring tender chicken and fragrant rice.",
"Sentiment": "positive",
"Cuisine": "Chinese",
"Presentation": "served at food stalls",
"Ingredients": "chicken, rice, herbs",
"FlavorProfile": "savory"

"Value": "Shinkansen",
"Category": "train",
"Commentary": "The speaker is about to take the Shinkansen to Nagoya.",
"Sentiment": "neutral",
"Capacity": "50+",
"Speed": "fast",
"Accessibility": "public",
"Cost": "medium"
```

**Figure 4:** Example of Structured Information Extracted from Travel Video Scenes for Food & Transportation

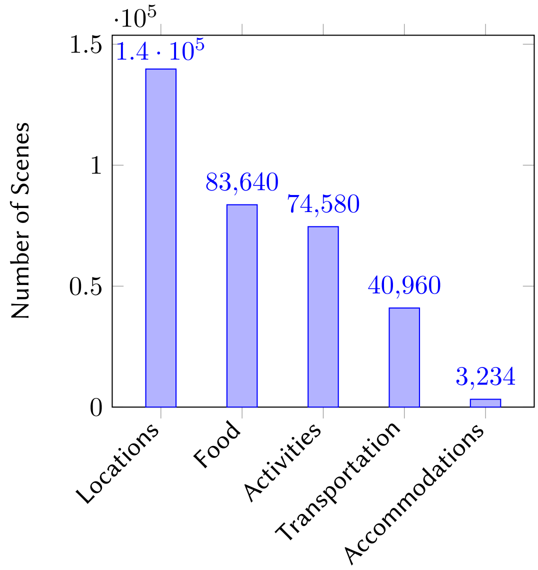


**Figure 5:** Number of Scenes Containing Entity Types

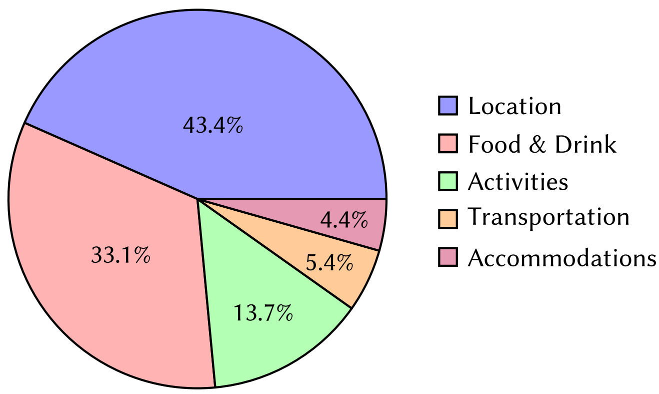


**Figure 6:** Unique Extracted Entities by Type

2. The query is sent to a retriever to find relevant video scenes. This retriever uses chunked text from the aligned scene transcript for semantic search, converts it to embeddings for semantic similarity search, and returns hits in a JSON blob containing the matching video(s). Each hit includes a corresponding video summary, primary entities with extracted attributes, and text blurbs (speech, scene text, VLM-generated visual caption) for the most relevant scenes in the parent video.
3. A prompt is sent to the VLM containing the retrieved context to perform retrieval-augmented generation, requesting the language model to use only the provided context from Step 2 to answer the query. The results include a blurb summarizing the findings and the desired structured output in the format predicted in Step 1.
4. If the output contains a recognizable point of interest, information to generate a map (e.g., latitude/longitude) is retrieved.
5. The frontend renders the chat message response and records history for follow-up questions.

| | | | | | |
|---|---|---|---|---|---|
| 19 | bangkok, thailand | 10 | india | 7 | singapore |
| 19 | new york city | 10 | indonesia | 7 | sri lanka |
| 18 | tokyo, japan | 9 | bali, indonesia | 7 | scotland |
| 16 | london | 9 | seoul, south korea | 6 | united states |
| 14 | japan | 9 | philippines | 6 | chiang mai, thailand |
| 12 | vietnam | 9 | thailand | 6 | mexico |
| 10 | iceland | 8 | los angeles, california | 6 | hawaii |
| 10 | hong kong | 7 | phuket, thailand | 6 | australia |
| 5 | europe | 5 | taiwan | 5 | disney world |
| 5 | kyrgyzstan | 5 | morocco | 5 | mekong delta, vietnam |
| 5 | kuala lumpur, malaysia | 5 | walt disney world | 5 | los angeles |

**Figure 7:** Most Frequently Occurring Video Level Primary Location

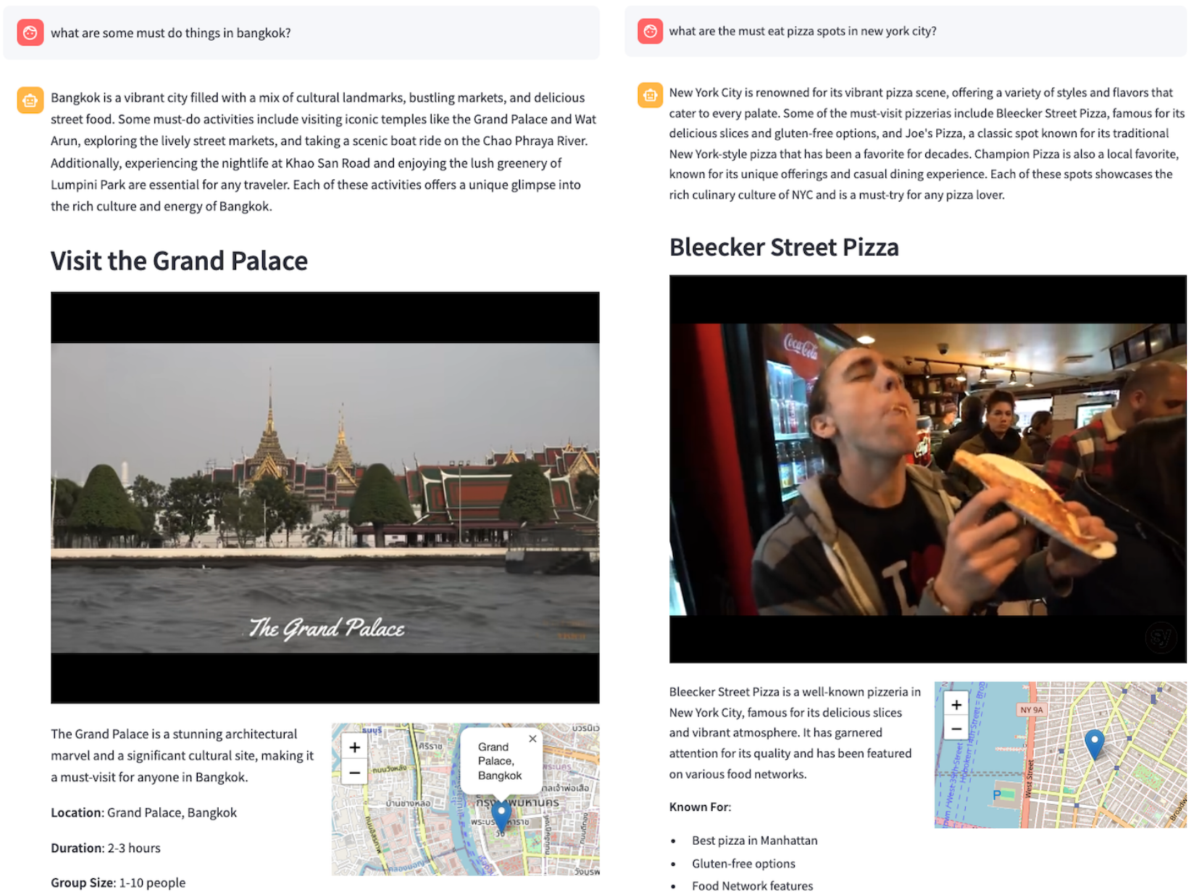


**Figure 8:** Demo screenshots of the AI chatbot leveraging structured information extracted from the Travel YouTube video collection. The left screenshot demonstrates the use of "Activity" information in response to "what are some must do things in Bangkok?," while the right shows "Location" information in response to "what are the must eat pizza spots in New York City"

## 4. Conclusion

This study presents a zero-shot prompting framework for extracting structured entity information from travel videos. We highlight the utility of this structured information in search, discovery, and sentiment analysis use cases. Furthermore, we demonstrate its practical application in a video AI agent, enabling users to interact with a video collection and receive video-enriched responses augmented by structured information.

This approach to structured entity extraction, while demonstrated in the travel domain, has the potential to be applied across various sectors such as e-commerce, healthcare, and education, where video content is increasingly being used for information dissemination, customer engagement, and personalized recommendations. The generalizability of our methodology underscores its relevance for enhancing video analysis and interaction in diverse fields.

Future work will focus on fine-tuning large multimodal language models for structured entity extraction, formalizing methodologies for constructing exemplar schemas for target entity types, exploring unsupervised entity taxonomy generation, and benchmarking these techniques on a larger and more diverse set of travel videos and as well as in different video domains.

# References


[1] X. Fang, C. Xie, J. Yu, S. Huang, J. Zhang, How do short-form travel videos trigger travel inspiration? identifying and validating the driving factors, Tourism Management Perspectives 47 (2023) 101128. URL: https://www.sciencedirect.com/science/article/pii/S2211973623000569. doi:https://doi.org/10.1016/j.tmp.2023.101128.

[2] P. Nguyen, L. Pham, K. Tran, T. Giang, A systematic literature review on travel planning through user-generated video, Journal of Vacation Marketing 30 (2023) 135676672311529. doi:10.1177/13567667231152935.

[3] J. Gan, S. Shi, R. Filieri, W. K. Leung, Short video marketing and travel intentions: The interplay between visual perspective, visual content, and narration appeal, Tourism Management 99 (2023) 104795. URL: https://www.sciencedirect.com/science/article/pii/S0261517723000778. doi:https://doi.org/10.1016/j.tourman.2023.104795.

[4] D. Li, J. Li, H. Li, J. C. Niebles, S. C. Hoi, Align and prompt: Video-and-language pre-training with entity prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4953–4963.

[5] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, L. Yuan, Video-llava: Learning united visual representation by alignment before projection, 2023. URL: https://arxiv.org/abs/2311.10122. arXiv:2311.10122.

[6] N. Radha, Video retrieval using speech and text in video, in: 2016 International Conference on Inventive Computation Technologies (ICICT), volume 2, 2016, pp. 1–6. doi:10.1109/INVENTIVE.2016.7824801.

[7] B. Yang, T. Mei, X. Hua, L. Yang, S. Yang, M. Li, Online video recommendation based on multimodal fusion and relevance feedback, in: ACM International Conference on Image and Video Retrieval, 2007. URL: https://api.semanticscholar.org/CorpusID:8293328.

[8] E. Haris, K. H. Gan, Extraction and visualization of tourist attraction semantics from travel blogs, ISPRS International Journal of Geo-Information 10 (2021). URL: https://www.mdpi.com/2220-9964/10/10/710. doi:10.3390/ijgi10100710.

[9] D. Kaushik, S. Gupta, C. Raju, R. Dias, S. Ghosh, Making travel smarter: Extracting travel information from email itineraries using named entity recognition, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 354–362. URL: https://doi.org/10.26615/978-954-452-049-6_047. doi:10.26615/978-954-452-049-6_047.

[10] H. Nanba, S. Fukuda, Automatic detection of geotagged food-related videos using aspect-based sentiment analysis, in: J. Neidhardt, W. Wörndl, T. Kuflik, D. Goldenberg, M. Zanker (Eds.), Proceedings of the Workshop on Recommenders in Tourism co-located with the 17th ACM Conference on Recommender Systems (RecSys 2023), Singapore and Online, September 19, 2023, volume 3568 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 10–17. URL: https://ceur-ws.org/Vol-3568/paper2.pdf.

[11] H. Halimeh, F. Freese, O. Müller, Event recommendations through the lens of vision and language foundation models, in: J. Neidhardt, W. Wörndl, T. Kuflik, D. Goldenberg, M. Zanker (Eds.), Proceedings of the Workshop on Recommenders in Tourism co-located with the 17th ACM Conference on Recommender Systems (RecSys 2023), Singapore and Online, September 19, 2023,

volume 3568 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 39–60. URL: https://ceur-ws.org/Vol-3568/paper5.pdf.
[12] K. D. Rosa, Video enriched retrieval augmented generation using aligned video captions, 2024. URL: https://arxiv.org/abs/2405.17706. arXiv:2405.17706.
[13] S. An, Y. Liu, H. Peng, D. Yin, Vkie: The application of key information extraction on video text, 2024. URL: https://arxiv.org/abs/2310.11650. arXiv:2310.11650.
[14] LangChain, Self-query retriever, 2024. URL: https://python.langchain.com/v0.1/docs/modules/data_connection/retrievers/self_query/, accessed: 2024-08-16.
[15] Breakthrough, Pyscenedetect: Video scene detection in python, https://github.com/Breakthrough/PySceneDetect, 2024. Accessed: 2024-08-16.
[16] OpenAI, Gpt-4: Generative pre-trained transformer, https://openai.com/research/gpt-4, 2023. Accessed: 2024-08-16.