RAVEN: An Agentic Framework for Multimodal Entity Discovery from Large-Scale Video Collections

Kevin Dela Rosa

Aviary Inc.
San Francisco, CA
kdr@aviaryhq.com

Abstract

We present RAVEN (**R**ecognition and **A**daptation of **V**ideo **EN**tities), an adaptive AI agent framework designed for multimodal entity discovery and retrieval in large-scale video collections. Synthesizing information across visual, audio, and textual modalities, RAVEN autonomously processes video data to produce structured, actionable representations for downstream tasks. Key contributions include (1) a category understanding step to infer video themes and general-purpose entities, (2) a schema generation mechanism that dynamically defines domain-specific entities and attributes, and (3) a rich entity extraction process that leverages semantic retrieval and schema-guided prompting. RAVEN is designed to be model-agnostic, allowing the integration of different vision-language models (VLMs) and large language models (LLMs) based on application-specific requirements. This flexibility supports diverse applications in personalized search, content discovery, and scalable information retrieval, enabling practical applications across vast datasets.

1 Introduction

The exponential growth of video content across platforms necessitates intelligent systems for organizing and retrieving information at scale. Video collections, spanning domains such as education, entertainment, and instructional content, present unique challenges due to their multimodal nature—needing to integrate visual, auditory, and textual data.

Recent advances in large language models (LLMs) and vision-language models (VLMs) enable new opportunities for multimodal understanding (Zhang, Li, and Bing 2023; Maaz et al. 2024; Lin et al. 2023). However, these methods typically process videos in isolation focusing on individual video comprehension, but lack mechanisms for collection-wide analysis. This capability is crucial for applications requiring a cohesive understanding of video collections rather than isolated clips.

RAVEN addresses these gaps with a model-agnostic design, allowing the integration of different VLMs and LLMs to suit domain-specific needs. This ensures adaptability to application-specific constraints such as computational efficiency or context length requirements. Our contributions include:

- A modular agentic architecture for video category cannonicalization and multimodal entity extraction.
- A synthetic example and schema-guided mechanism for contextual prompting.
- Demonstration of high-quality structured entity extraction across large-scale video datasets using popular off the shelf LLM/VLMs.

2 RAVEN Framework Overview

RAVEN operates as an adaptive agent for structuring multimodal video data, allowing one to operate on a collection of videos without necessarily having deep knowledge of the content contained within the collection but still allow for rich and consistent extraction of structured information across the provided videos. Our framework is comprised of two core stages: **Category Understanding**, and **Rich Domain Specific Entity Extraction**

Throughout these stages, as illustrated in Figure 1, RAVEN uses a vision-language model for video-based tasks such as categorization and entity extraction, and a large language model for category name consolidation and generation. This frame work is designed to be model-agnostic as long as they handle comparable context lengths and support structured output JSON representations. For this study we used Gemini 1.5 Flash (DeepMind 2023) as our VLM (using both visual video and audio input) and GPT-4o (OpenAI 2023) as our general text LLM.

2.1 Category Understanding & Schema Generation

Video clips are first processed by the VLM to infer *video categories* and if desired extract *general-purpose entities* such as people, objects, and locations in same prompt. This process can optionally include user-provided prompts to steer categorization toward specific goals. Then the top occurring raw category names produced by the VLM are fed through an LLM to normalize and dedupe similar concepts to produce a canonical list of categories.

Using the canonical categories, the LLM generates a list of typical entities expected in that domain and produces corresponding *domain-specific entity schemas*. For each category, the generated schema includes:

- A list of *typical entities*.

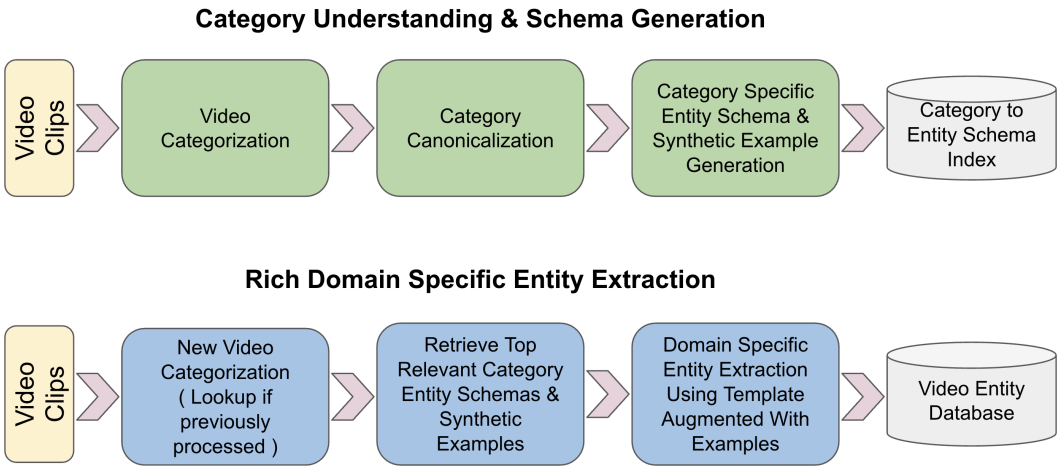

Figure 1: RAVEN Framework Overview. The two main LLM / VLM driven flows: 1) Category Understanding & Schema Generation, 2) Rich Domain Specific Entity Extraction

- Attributes for each entity, with *descriptions* and *example* values.

The resulting entity schema lists per category are indexed by category name for later retrieval/lookup to help guide the actual entity extraction process. This modular design ensures adaptability for domain-specific needs, reducing manual intervention in schema development.

2.2 Rich Domain Specific Entity Extraction

Finally, video clips are processed again (as well as any desired additional clips) through the VLM. Using the assigned category from the first agentic flow, the system retrieves the most relevant schema based on *semantic similarity* of the original unnormalized category to the top matching canonical category name. The schema is integrated into a prompt template with example values, facilitating in-context learning to extract entities and attributes aligned with the schema's structure. The results are then persisted and indexed according to the needs of the downstream application.

3 Experiments & Evaluation

In this section we evaluate RAVEN's performance in its two core stages. We explore RAVEN's ability to infer video categories and generate schemas in Section 3.1, and in Section 3.2 investigate RAVEN's entity extraction capabilities.

3.1 Category Understanding & Schema Generation Analysis

To demonstrate RAVEN's effectiveness in category understanding and schema generation, we applied RAVEN on 1.5 million video clips (over 5000 hours of footage) from the *Aligned Video Captions* dataset (Delarosa 2024), and report results from this large scale qualitative exploration.

In Figure 2 we show the distribution of video clips by inferred canonical category, the individual categories and spread align nicely with the source dataset which was sampled equally from from YouTube's 15 top level categories.

Category	Entity → Attribute	Top Values
Generic	Person → Role	speaker, host, child, listener, narrator, chef
Generic	Background → Setting	kitchen, living room, studio, office, city street
History	Event → Description	world war ii, apollo 11, vietnam war
History	Figure → Name	neil armstrong, abraham lincoln, adolf hitler
How-To	Tools & Materials → Type	knife, pot, bowl, camera, fishing rod
How-To	Techniques → Type	cutting, cooking, mixing, installation
Travel	Destination → Location	bangkok, new york city, tokyo, london

Table 1: Sample entity values extracted for generic and domain-specific entities, showcasing RAVEN's versatility and domain adaptation

Figure 3 presents the distribution of generic entity types and attributes, highlighting the flexibility of our framework to capture widely applicable generic entities with their associated attributes.

In Figure 4, we demonstrate the framework's capacity for domain-specific extractions, exemplified by **How-To** and **History** videos, where we show the generated entity types and attributes, and their distributions in the dataset. Table 1 lists the most frequently extracted entity value for sample entity types in different domains.

3.2 Domain Specific Entity Richness Analysis

For understanding the quality of entities extracted in the framework, we evaluated RAVEN on a sample of 300 video

Table 2: Case study of extracted entities by method. This table illustrates the qualitative differences in entity extraction across methods for a sample video (e.g., a historical documentary about Abraham Lincoln)

Entity → Attribute	Ours	Speech	OCR	Caption	YOLO
Class Agnostic Generic Entity					
Person → Role	Abraham Lincoln → President	Abraham Lincoln; President Lincoln	PRESIDENT LINCOLN	Abraham Lincoln	Person
Person → Gender	Male	-	-	Man	-
Person → Age	Mid 50s	-	-	-	-
Person → Appearance	Wearing a Dark Suit	-	-	-	-
Person → Mood	Sad Reflective	-	-	-	-
Object → Type	Train Car	Train	-	Train Car	Train
Object → Color	Black & White	-	-	-	-
Object → Size	Large	-	-	-	-
History & Documentary Specific Entities					
Historical Event → Description	Surrender of the Army of Northern Virginia	Battle of Appomattox courthouse	-	-	-
Historical Event → Date	April 9, 1865	-	-	-	-
Historical Event → Location	Appomattox Courthouse, Virginia	-	-	-	-
Historical Event → Key Figures	Robert E. Lee, Ulysses S. Grant	-	-	-	-
Historical Site → Location	Lincoln Memorial → Washington, D. C.	Lincoln Memorial	-	-	-
Historical Site → Era	Early 20th Century	-	-	-	-
Historical Site → Architectural Features	Marble structure, neoclassical design	-	-	-	-

Figure 2: Inferred canonical video category distribution

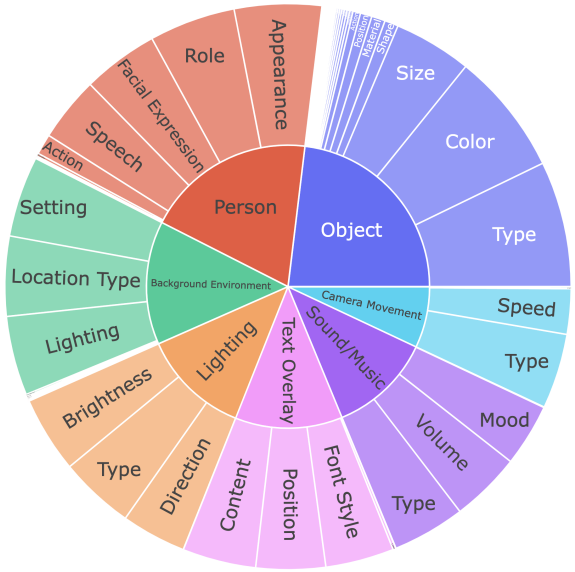

Figure 3: Distribution of extracted structured attributes for each generic entity type

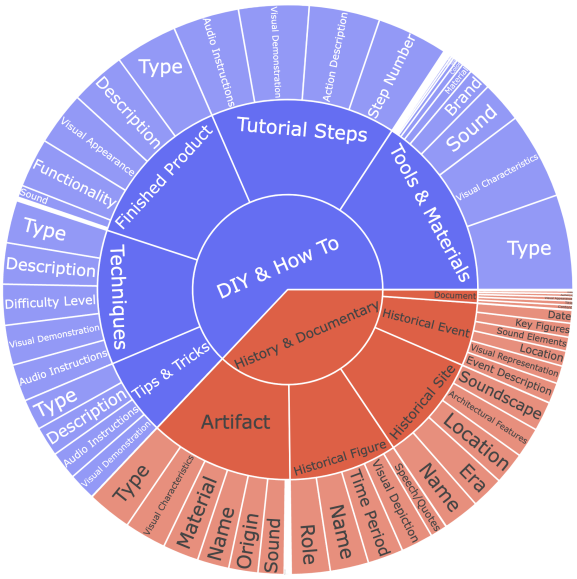

Figure 4: Distribution of domain-specific entity types & attributes for How-To and History videos

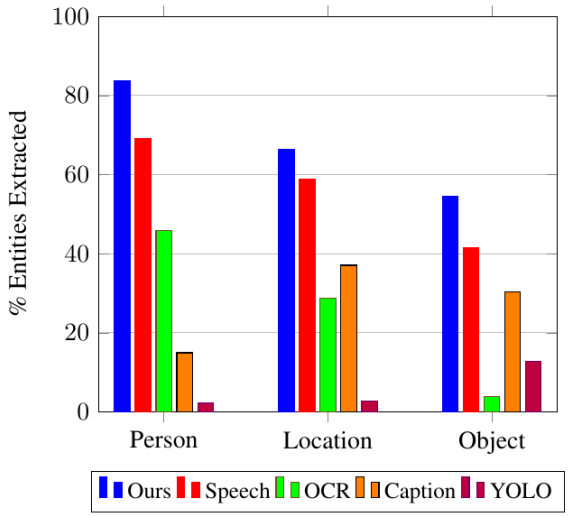

Figure 5: Entity Recall by method. Each bar represents the proportion of entities extracted by each method for the Person, Location, and Object entity types respectively

clips (∼10 seconds each) drawn from the *Aligned Video Captions* dataset (Delarosa 2024). The clips were selected from diverse categories, including *Travel*, *Education*, *History*, and *Instructional Content*, to ensure broad domain coverage. Each clip was processed through the RAVEN pipeline, starting with category inference and schema generation, followed by rich entity extraction.

To benchmark performance, we compared RAVEN against the following baselines using standard configurations:

- **NER on Speech:** Identifies entities from transcribed video speech, extracted automatically via AssemblyAI Speech-to-Text (AssemblyAI 2023). The named entities were extracted from the speech transcript using GLiNER (Zaratiana et al. 2023).
- **OCR on Scene Text:** Extracts visible text from frames, capturing entities like place names, using EasyOCR (JaidedAI 2020).
- **Keyword Extraction from Visual Caption** Extracts keywords from the visual captions provided in Panda-70M (Chen et al. 2024).
- **YOLO Object Detection:** Detects general objects in frames, labeling them without contextual structure, using YOLOv10 (Wang et al. 2024) pretrained on COCO dataset (Lin et al. 2014).

The baseline methods were selected to evalate different aspects of entity extraction. NER on Speech assesses linguistic entity recognition, OCR captures text from visual data, and visual captioning provides descriptive context. These baselines highlight specific extraction limitations that our framework addresses comprehensively by operating in a multimodal fashion.

Figure 5 visualizes the ability of each method to extract class agnostic generic entities from various videos. Our framework shows a strong ability to extract *Person*, *Location*, and *Object* entities. The baselines often lacked multimodal context and finer grained vocabulary to successfully extract entities presented in the videos. Figure 5 demonstrates RAVEN's ability to synthesize multimodal context, significantly improving recall rates over unimodal baselines

Futhermore, we present a case study on a historical documentary to illustrate the qualitative depth of extracted entities, for both the class agnostic generic setting and the domain specific setting. Table 2 shows that our framework not only extracts named entities but also attributes, descriptions, and relationships (e.g., *Person* → *Role, Event* → *Location*). Baseline methods produce isolated labels limiting their utility in structured retrieval. The qualitative analysis (Table 2) demonstrates RAVEN's capability to extract fine-grained, contextual attributes compared to baseline methods, which often fail to capture relationships and attribute.

4 Conclusion

We presented RAVEN, an agentic AI framework for multimodal entity discovery and retrieval. By integrating category understanding, schema generation, and retrieval augmented & example-guided extraction, RAVEN addresses the challenges of structuring unstructured video content. RAVEN demonstrates the ability to extract structured, domain-specific representations, advancing multimodal information retrieval. Future work will explore extending RAVEN's capabilities to support entity relationships and optimizations for schema generation & signal extraction process. RAVEN's scalability and modularity position it as a versatile solution for evolving multimodal retrieval challenges.

References

AssemblyAI. 2023. AssemblyAI Speech-to-Text. Accessed: 2023-11-07.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; and Tulyakov, S. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

DeepMind. 2023. Gemini (Flash). Accessed: 2024-11-03.

Delarosa, K. 2024. Video Enriched Retrieval Augmented Generation Using Aligned Video Captions. arXiv:2405.17706.

JaidedAI. 2020. EasyOCR: Ready-to-use Optical Character Recognition with 80+ Supported Languages. Accessed: 2023-11-04.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv:2311.10122.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'a r, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

OpenAI. 2023. GPT-4. Accessed: 2024-11-03.

Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; and Ding, G. 2024. YOLOv10: Real-Time End-to-End Object Detection. *arXiv preprint arXiv:2405.14458*.

Zaratiana, U.; Tomeh, N.; Holat, P.; and Charnois, T. 2023. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. arXiv:2311.08526.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858.