






A Schema-Based Approach to the Linkage of Multimodal Learning Sources with Generative AI

Christine Kwon¹ , James King², John Carney² , and John Stamper¹ 

¹ Carnegie Mellon University, Pittsburgh, PA 15213, USA
ckwon2@andrew.cmu.edu, jstamper@cmu.edu

² MARi LLC, Alexandria, VA 22314, USA
{james.king, john.carney}@mari.com

Abstract. Learning how to execute a complex, hands-on task in a domain such as auto maintenance, cooking, or guitar playing while relying exclusively on text instruction from a manual is often frustrating and ineffective. Despite the need for multimedia instruction to enable the learning of complex, manual tasks, learners often rely exclusively on text instruction. However, through widespread usage of user-generated content platforms, such as YouTube and TikTok, learners are no longer limited to standard text and are able to watch videos from easily accessible platforms to learn such procedural tasks. As YouTube consists of a large corpus of diverse instructional videos, the accuracy of videos on sensitive and complex tasks has yet to be validated in comparison to “golden standard” manuals. Our work provides a unique LLM-based multimodal pipeline to interpret and verify task-related key steps in a video within organized knowledge schemas, in which demonstrated video steps are automatically extracted, systematized, and validated in comparison to a text manual of official steps. Applied to a dataset of twenty-four videos on the task of flat tire replacement on a car, the LLM-based pipeline achieved high performance on our metrics, identifying an average of 98% of key task steps, with 86% precision and 92% recall across all videos.

Keywords: Generative AI · Large Language Models (LLMs) · Multimodal Learning · Video Training

1 Introduction

The execution of complex tasks from steps in written form from a manual can be difficult for novice learners. Pedagogical models of demonstration and apprenticeship have long served as the foundation for teaching complex hands-on skills, reflecting the nuanced interplay between observation, practice, and feedback. This pedagogical approach, deeply rooted in the cognitive apprenticeship theory, emphasizes learning in context [5], where the learner engages in authentic tasks with the guidance of a more knowledgeable expert. The effectiveness of these models in complex skill acquisition is largely attributed to the scaffolded support they offer, allowing learners to gradually develop competence through iterative engagement and personalized feedback [13].

In recent times, the surge in digital technologies has paved the way for innovations to these traditional learning models. Platforms such as YouTube and TikTok are now widely used for the demonstration and teaching of complex tasks in areas such as auto maintenance, sports, or learning to play a musical instrument. While these videos can provide better observation than a written manual, students may still need to follow and understand the steps in a manual to achieve mastery. In this research, we recognize the multimodal aspects of this kind of learning and are building the foundation of a technical infrastructure to allow learners to seamlessly move between multiple modes of learning by organizing complex tasks into written steps that can then be instantly attached to video or audio content at the correct time point.

To accomplish this, we have proposed a structural ontology around the concept of a knowledge object (KO) [7]. We define a KO around a specific task that is composed of one or more steps and the metadata associated with each step. There are several reasons why the idea of defining KOs is important for task-based learning. The primary reason is to easily organize tasks for learning purposes. A secondary and equally important reason is the ability to search for tasks and examples on demand. Scenarios where KOs excel are mechanical tasks like automotive repair (changing oil, spark plugs, tire, etc.) where finding the task currently is quite easy, but a specific instance for a particular vehicle may be much harder to find and may have details specific to that instance. Having the right metadata in the right structure is critical, especially where the metadata links multimodal datastreams for a task. It is possible to define a hierarchical structure to a set of KOs where parent-child relationships exist connecting tasks at multiple levels, although for the purpose of broad search we are most interested in the higher level KOs.

A noteworthy development in KOs is the integration of Large Language Models (LLMs), such as GPT-4, and frameworks like LangChain, which have shown potential to revolutionize the way we approach the organization and interpretation of multimodal data in educational contexts [14]. These advanced LLMs, with both generative capabilities and nuanced understanding of natural language, offer unprecedented support in structuring learning experiences around complex tasks. They not only facilitate the organization of diverse data types—ranging from textual instructions and procedural guidelines to visual demonstrations and interactive simulations—but also enhance the interactive learning experience by providing real-time, contextually relevant support and feedback.

The incorporation of LLMs to link demonstration and apprenticeship models in KOs enables a more dynamic, adaptive, and personalized learning environment. By leveraging the power of LangChain schemas to interlink and contextualize multimodal information, educators and learners can now navigate complex hands-on tasks with greater ease and efficiency. This paper aims to explore the implications of this integration, highlighting the potential of LLMs to augment learning support and reshape the pedagogical approaches to complex task training.

2 Related Work

There has been a large body of prior work in the area of video-based learning for procedural tasks, as well as comparisons of video-based learning to text-based learning. We are also interested in previous work that investigates various methods of extraction of learning materials, including LLM-based methods, from instructional videos.

2.1 Comparing Video-Based and Text-Based Learning

Using YouTube-like videos has become a core source of learning for many students, yet there is still an insufficient understanding of how video-based learning juxtaposes with text-based learning, especially for complex procedural tasks. A prime example of a highly sensitive task that benefits from video-based learning is surgical procedures [3, 11]. In particular, prior research compared video and text-based learning material to investigate which modality of learning promotes higher student learning in basic laparoscopic suturing and clinical procedures and skills [3, 11]. Sonnenfeld et al. investigated training procedures in electronic and distance learning approaches in the context of flight crew training, in which they found that video-based learning was effective for procedure-based training in providing flight crew trainees with interactive opportunities and formal training content [12]. While video-based learning has the potential to teach and deliver information that is difficult to convey from standard text material [10], it is especially important to understand how to conduct effective video-based learning for high-risk tasks that are dependent on a correct ordering of steps.

2.2 Video-Based Step-By-Step Feature Extraction

Instructional videos are extremely diverse in content, especially in the hierarchy of steps and key moments in learning how to conduct a certain task. Hence, there has been a noticeable growing interest in improving the searchability and indexing of these instructional videos by extracting key informational features from these videos [16]. The detection of key steps within instructional videos is difficult, which rely on both the chronological and temporal ordering of steps in accomplishing a task [18]. Recent studies employ a joint method that evaluates text-based extractions in conjunction with visual feature detectors [1, 8]. Additional methods require extracted visual actions and objects to have a high degree of semantic relatedness with the textual information attached to a procedural video [9]. These extraction methods require an examination of the relevance of these features to an empirical measure of comparison. Our work aims to automate this comparative process by using an LLM-based pipeline to accurately extract task step-related features and systematize the hierarchy of steps verified by an empirical measure of comparison to support more complex task-based learning.

2.3 LLM-Based Step-By-Step Information Extraction.

Currently, Large Language Models (LLMs) are showing their immense potential in efficiently carrying out diverse NLP tasks [2, 19]. However, we have yet to completely

rely on LLM-based text generation as little is known about the ability of LLMs to verify their content output [17]. To improve the ability of complex reasoning and content generation of LLMs, Wei et al. introduced the “chain of thought” (CoT) prompting method, which induces these language models to deconstruct a problem into multiple in-depth reasoning steps [15]. However, verification methods using prompt chaining may not be completely efficient for extracting and verifying multiple ordered steps of a task, especially if the task data is unstructured. On the other hand, another prior study used an LLM-assistance pipeline to extract organized annotations and informational features from unstructured clinical data [6]. While these prior studies separately investigated the validation of LLM-generated content and organized extraction of information from unstructured data, we have yet to see work that combines these processes, especially to verify task-based video content.

3 Methodology

```

step_schema = {
  "properties": {
    "Start Time: [step description]": {"type": "string"},
    "End Time: [step description]": {"type": "string"},
    "Transcript context: [step description]": {"type": "string"}
  }
}

```

```

step_schema = {
  "properties": {
    "Start Time: apply parking brake": {"type": "string"},
    "End Time: apply parking brake": {"type": "string"},
    "Transcript context: apply parking brake": {"type": "string"}
  }
  ...
  "Start Time: tighten the lug nuts": {"type": "string"},
  "End Time: tighten the lug nuts": {"type": "string"},
  "Transcript context: tighten the lug nuts": {"type": "string"}
}

```

Fig. 1. The left figure is the outline of the schema input, which consists of three features: the start time and end times of the demonstrated step and the transcript context used to determine the time interval of each step occurrence. The right figure is the schema input used to extract key task steps for each video on replacing a flat tire on a car.

We employed LangChain, a language model integration framework powered by Large Language Models (LLMs) to develop and extricate organized schemas from any text source [4]. We specifically used LangChain in conjunction with GPT 4.0 to create and extract schemas that align the retrieved steps with affirmed key steps from any official manual.

Using the LangChain framework, we created an input schema, labeled “step_schema,” shown in Fig. 1, that lists fundamental key steps on a procedural task, in this case, car tire replacement, from an official manual within an organized knowledge metastructure. Within the schema, we listed the desired extracted features for each step, which include the starting timestamp of when a step is first mentioned, the ending timestamp when a step is no longer mentioned, and the transcript context used to extract each step occurrence from a video narration. The extracted features for each step of a task are also explained in Table 1. Figure 1 also presents the specific schema input used in our approach to extract the key steps listed in Table 2 from videos on car flat tire replacement.

Table 1. This table presents the unique features that are listed in the properties for each task step of the input schema. Each feature is listed along with its type and definition.

Feature Name	Feature Type	Feature Definition
Start Time: [step description]	string	Starting timestamp on first narrated mention of step
End Time: [step description]	string	Ending timestamp on last narrated mention of step
Transcript context: [step description]	string	The portion of text on the full occurrence of step

Table 2. This table lists 12 official manual steps of flat tire replacement on a car used to list and verify the retrieved steps within the extracted schema.

Manual Steps (1–6)	Manual Step	Manual Steps (7–12)	Manual Step
1	Apply parking brake	7	Remove lug nuts
2	Remove the spare tire from the car	8	Remove flat tire
3	Use wheel chocks to block the wheels opposite of the wheel you’re changing	9	Place the spare tire
4	Loosen the lug nuts from the tire	10	Screw on the lug nuts
5	Loosen the jack	11	Use the jack to lower the car
6	Use the jack to lift up the car	12	Tighten the lug nuts

In providing an input schema with key task steps, this LLM-based pipeline can structure an output schema isomorphic to the input schema. By systematizing task steps and their features within an organized schema, crucial step information is easily accessible and directly aligned with authenticated task steps from a verified source.

4 Preliminary Results and Discussion

We collected 24 videos from YouTube on replacing a flat tire on a car. YouTube-generated transcripts from each video, which included toggled timestamps, were collected as input data for our LLM-based pipeline. This pipeline extracted an isomorphic schema to the input schema shown in Fig. 1 for each video. To evaluate the performance of the pipeline, two human evaluators determined a correctly identified extracted task step of replacing a flat car tire if its starting and ending timestamps overlap with the time interval of the

same step in each video. The LLM-based pipeline achieved high-performance metrics, identifying an average of 98% of key task steps with 86% precision and 92% recall across all videos. The pipeline was highly successful in identifying missing steps for most videos. However, we also encountered some drawbacks with this LLM-based approach. Though the LLM-based pipeline retrieved and identified 100% of key task steps for approximately 92% of videos in the dataset, at least some steps were incorrectly identified for 73% of the videos. For instance, while the pipeline was successful in recognizing most task steps in each video, it also was prone to misidentify task steps not demonstrated in videos. Additionally, there were instances where the pipeline incorrectly identified the sections of the video in which a task step was demonstrated. We intend to address these limitations by evaluating and improving the LLM-based pipeline performance on diverse video datasets.

The ability to auto-identify the start and stop time of each required step in a “how-to” video lays the groundwork for instructional videos to be systematically checked for accuracy at scale. The benefits of automatically extracting accurate content are most important in mission-critical spaces like defense, aviation, and healthcare. When a novice is learning to execute a lower-stakes procedure such as car dent removal, for example, perhaps the wisdom of the crowd conveyed via the most-liked videos on YouTube is a sufficient filter to ensure useful content reaches learners. However, when learning a high-stakes procedure like aircraft repair, execution of each step in the required order is critical; the learner must be able to trust the accuracy of a single video.

The schema-driven approach that links related instructional content to a golden standard official procedure can also streamline updates to instructional content. For example, consider the official procedures and documentation for AI/ML tools that change frequently. If the plethora of YouTube tutorial videos describing how to use said AI/ML tools was linked with “gold standard” documentation, relevant moments of the videos could be automatically flagged as deprecated when appropriate. Importantly, the current implementation described above compares only a video’s time-stamped transcript against the official procedure. Therefore, the step of the procedure must be described or referred to aloud in the video for the algorithm to identify its presence. The below Future Work section describes the extensibility of the transcript-procedure comparison approach as well as plans to parse objects and actions detected in videos.

5 Conclusion and Future Work

Automatic alignment of crowdsourced instructional videos, like those on YouTube, and official “gold standard” procedure steps, that are found in written manuals, will enable learners of tasks in diverse domains to quickly jump from a step in a manual to an instructional video moment detailing a task step of interest. Mission-critical and machine repair contexts such as aviation and medicine often involve reliance on official step-by-step procedure documents that lack the detail needed to convey a task’s proper execution. Linkage of an official procedure step and related multimedia segments could enable learners to toggle between a text and video depiction of the same task, empowering the learner to access detailed instruction from a trusted source.

Further, the automatic tagging of multimedia content segments with their associated official procedure steps will enable the recommendation of video moments to individual

learners or groups based on observed task and/or procedure proficiency. For example, a content recommendation system could leverage tags to serve a video moment that depicts “How to replace Part X on a given aircraft” to a learner with relevant task proficiency below a threshold. In this way, the automatic alignment of official procedure steps and video moments will enable targeted content pushes that preempt a learner’s proactive query. The ability to proactively push tutorial content segments to learners could yield performance benefits during cases of “unconscious incompetence”, in which the learner does not know that they lack the ability to effectively execute a task.

A next step of our current approach, which compares a video transcript against an official procedure, is integrating automated video intelligence. Detecting on-screen objects and actions will enable the content-procedure alignment algorithm to identify a step included in a video that is not mentioned aloud. Development of object and action recognition video intelligence methods is a high priority going forward and will involve fine-tuning existing large models with data relevant to each use case (e.g., car tire replacement, de-icing aircraft). As the use of mixed reality learning content to address skill gaps in hands-on domains grows, the same content-procedure alignment method will also be applied to mixed reality (MR) content.

Acknowledgments. This work was supported by US Navy STTR #N68335-21-C-0438.

References

1. Alayrac, J.B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., Lacoste-Julien, S.: Unsupervised learning from narrated instruction videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4575–4583 (2016)
2. Ampel, B.M., Yang, C.H., Hu, J., Chen, H.: Large language models for conducting advanced text analytics information systems research (2023). arXiv preprint [arXiv:2312.17278](https://arxiv.org/abs/2312.17278)
3. Buch, S.V., Treschow, F.P., Svendsen, J.B., Worm, B.S.: Video- or text-based e-learning when teaching clinical procedures? A randomized controlled trial. *Adv. Med. Educ. Pract.* 257–262 (2014)
4. Chase, H.: LangChain. <https://langchain.com/>. Accessed on 1 Aug 2023
5. Dennen, V.P., Burner, K.J.: The cognitive apprenticeship model in educational practice. In: *Handbook of research on educational communications and technology*, pp. 425–439. Routledge (2008)
6. Goel, A., et al.: LLMS accelerate annotation for medical information extraction. In: *Machine Learning for Health (ML4H)*, pp. 82–100. PMLR (2023)
7. Kwon, C., Stamper, J., King, J., Lam, J., Carney, J.: Multimodal data support in knowledge objects for real-time knowledge sharing. In: *Proceedings of CROSSMMLA Workshop at the 13th International Conference on Learning Analytics & Knowledge (2023)*
8. Malmaud, J., Huang, J., Rathod, V., Johnston, N., Rabinovich, A., Murphy, K.: What’s cookin’? interpreting cooking videos using text, speech and vision (2015). arXiv preprint [arXiv:1503.01558](https://arxiv.org/abs/1503.01558)
9. Manju, A., Valarmathie, P.: Organizing multimedia big data using semantic based video content extraction technique. In: *2015 International Conference on Soft-Computing and Networks Security (ICSNS)*, pp. 1–4. IEEE (2015)
10. Navarrete, E., Nehring, A., Schanze, S., Ewerth, R., Hoppe, A.: A closer look into recent video-based learning research: a comprehensive review of video characteristics, tools, technologies, and learning effectiveness (2023). arXiv preprint [arXiv:2301.13617](https://arxiv.org/abs/2301.13617)

11. Routh, D., Rao, P.P., Sharma, A., Arunjeet, K.: To compare the effectiveness of traditional textbook-based learning with video-based teaching for basic laparoscopic suturing skills training—a randomized controlled trial. *Medical Journal of Dr. DY Patil University* (2023)
12. Sonnenfeld, N., Nguyen, B., Boesser, C.T., Jentsch, F.: Modern practices for flightcrew training of procedural knowledge. In: 84th International Symposium on Aviation Psychology, p. 303 (2021)
13. Stamper, J., Barnes, T., Croy, M.: Enhancing the automatic generation of hints with expert seeding. In: *Intelligent Tutoring Systems: 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14–18, 2010, Proceedings, Part II* 10, pp. 31–40. Springer (2010)
14. Topsakal, O., Akinci, T.C.: Creating large language model applications utilizing langchain: a primer on developing LLM apps fast. In: *International Conference on Applied Engineering and Natural Sciences*, vol. 1, pp. 1050–1056 (2023)
15. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural. Inf. Process. Syst.* **35**, 24824–24837 (2022)
16. Zala, A., et al.: Hierarchical video-moment retrieval and step-captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23056–23065 (2023)
17. Zhang, X., Gao, W.: Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method (2023). arXiv preprint [arXiv:2310.00305](https://arxiv.org/abs/2310.00305)
18. Zhong, Y., Yu, L., Bai, Y., Li, S., Yan, X., Li, Y.: Learning procedure-aware video representation from instructional videos and their narrations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14825–14835 (2023)
19. Zhu, Y., et al.: Large language models for information retrieval: A survey (2023). arXiv preprint [arXiv:2308.07107](https://arxiv.org/abs/2308.07107)