Supporting AI-assisted Task Learning with Hierarchical Representation of Procedural Knowledge

ANONYMOUS AUTHOR(S)

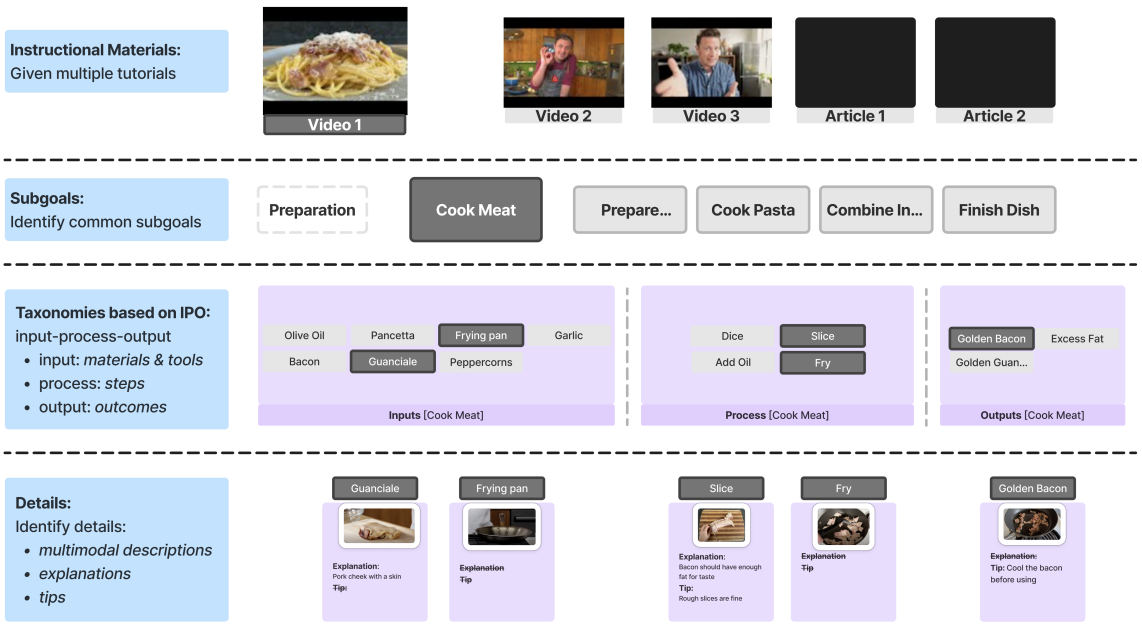

Fig. 1. Hierarchical procedure representation for (1) breaking down, (2) aligning, and (3) retrieving procedural information from multiple instructional materials (e.g., tutorial articles, how-to videos).

Tutorial articles and how-to videos are widely used for learning procedural tasks such as cooking, programming, and software use. However, these static resources cannot adapt to different user contexts and constraints. On the other hand, recent Generative AI tools—such as ChatGPT and Gemini—can offer tailored step-by-step instructions, enabling AI-assisted procedural task learning. Yet, they often struggle with coherent procedural reasoning and factual accuracy necessary for providing quality and effective instructional guidance. This work explores leveraging existing instructional materials (e.g., tutorial articles, how-to videos) to enhance AI-assisted task learning. We introduce a hierarchical representation of procedural knowledge that integrates instructional content from multiple sources, allwoing generative models to retrieve, refine, and generate more grounded instructional content. Additionally, we propose an automatic vision-language model (VLM)-based pipeline for constructing this representation and discuss its potential applications in AI-assisted task learning.

Manuscript submitted to ACM

ACM Reference Format:
Anonymous Author(s). 2018. Supporting AI-assisted Task Learning with Hierarchical Representation of Procedural Knowledge. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Procedural knowledge refers to "how-to" knowledge crucial for achieving specific goals [1]. This knowledge is fundamental to many human activities, including cooking, crafting, and programming. With the availability of online instructional resources—such as tutorial articles and how-to videos—access to procedural knowledge has expanded significantly. Users engage with these resources for practical guidance, inspiration, and knowledge validation [7, 15].

However, most instructional materials are designed as standalone guides, presenting fixed procedures that assume an audience with particular prior knowledge and circumstances. Yet, procedural tasks often involve varying situational constraints, and users come with different contexts and experiences [4]. As a result, these static materials may not fully address individual needs, requiring users to manually navigate multiple fragmented sources [4, 6], which is cognitively demanding and inefficient.

Meanwhile, recent Generative AI tools like ChatGPT, Gemini, and Claude have become useful for a wide range of creative and reasoning tasks [5, 8]. They can flexibly generate text that aligns with user instructions (i.e., prompts), which enable many useful applications such as creative writing, summarization, question-answering, and problem-solving. They can also produce instructional content tailored to specific user requests, facilitating personalized guidance that adapts to diverse needs and constraints. However, generative models often struggle with coherent procedural reasoning, structured planning [17, 18], and factual accuracy [11], which limits their reliability for procedural task learning.

To address these challenges, we investigate how existing instructional materials can support AI-assisted procedural learning. These materials can serve as reliable sources of procedural knowledge, with the web containing vast amounts of information on even a single task [2, 3, 12]. These resources include diverse procedures that require different sets of materials, methods, and expected outcomes. At the same time, they include insignificant but inevitable variations in terminology, procedural structure, and information presentation, which can hinder integration.

To address these inconsistencies while preserving valuable procedural variations, we propose a structured approach to procedural knowledge representation. Specifically, we introduce a hierarchical model that facilitates the alignment, comparison, and retrieval of instructional content across multiple sources about the same task (i.e., resources retrieved by a single query like "how to cook pasta" or "how to remove objects from an image"). This representation organizes procedural knowledge into three layers:

- **Subgoals**, which break down tasks into meaningful, modular units.
- **Input-Process-Output (IPO)** structures, which standardize procedural components by decomposing subgoals into required materials, actions, and expected outcomes.
- **Details**, which capture explanations, tips, and multimodal descriptions for each IPO element.

With a structural representation of procedural knowledge, we can enable AI-powered systems to retrieve, refine, and generate instructional content in a grounded manner, ensuring more coherent and adaptable procedural guidance. We propose a VLM-based pipeline that can processes a set of instructional materials about the same task and construct the hierarchical representation automatically. We then describe two key applications of this representation to support AI-assisted personalized task learning.

2 Hierarchical Representation

We developed a hierarchical representation for procedural knowledge that facilitates consistent alignment between different instructional resources for the same task. The design of the representation was informed by prior research on task learning and a formative study (N=10) on how users consume and interact with instructional materials from multiple sources.

2.1 Subgoals: Breaking down the task

The first layer of our representation, subgoals, breaks down tasks into meaningful and modular stages. Research in instructional design has shown that segmenting procedures into subgoals enhances learning efficiency and ease of navigation [9]. This structured approach also enables the alignment and comparison of instructional materials, even when different sources present information at varying levels of detail or in different sequences. Furthermore, by focusing on one subgoal at a time, we can meaningfully compare resources that may not fully correspond across all subgoals.

The granularity of subgoals depends on the domain and application. For example, in cooking tutorials, subgoals might represent key milestones such as preparing ingredients or cooking meat (Figure 1), whereas in makeup tutorials, they may correspond to specific facial areas [16].

2.2 Input-Process-Output (IPO): Breaking Down Subgoals

The second layer of the representation employs an Input-Process-Output (IPO) framework to decompose each subgoal into its constituent elements. Inputs refer to materials, tools, or ingredients required to perform the subgoal; processes represent the actions or transformations applied to those inputs; and outputs capture the results or intended effects. Our formative study revealed that users naturally engage with instructional materials by comparing inputs, methods, and expected outcomes. When engaging with different resources, they frequently focus on how variations in materials or techniques influence results. For example, in a cooking tutorial, one source might specify guanciale, while another substitutes bacon (Figure 1). Similarly, different materials may employ distinct techniques, such as dicing vs. slicing meat, which ultimately affects the texture and cooking process.

Beyond facilitating comparisons, the IPO structure enhances the clarity of procedural dependencies. Many procedural tasks involve steps where an output from one stage serves as an input for another. For instance, frying guanciale with different methods influences both the rendered fat and the crispiness, which in turn affects the texture of the final dish. The IPO framework can capture these dependencies explicitly. Additionally, since different resources may use varied terminologies for similar actions, IPO allows to standardize vocabulary by categorizing inputs, processes, and outputs within a taxonomy. This standardization ensures that equivalent components—such as “dice” and “chop”—can be recognized as functionally similar, improving cross-material alignment.

2.3 Details: Capturing Supporting Information

The third layer, details, focuses on capturing explanations, tips, and multimodal descriptions associated with specific inputs, processes, or outputs. While subgoals and IPO establish a structured foundation for organizing procedural content, the details layer ensures that relevant supplementary knowledge is explicitly recorded. This layer accommodates different types of information, including rationales for particular techniques, contextual tips that enhance performance, and visual or textual descriptions of different inputs, processes, and outputs.

2.4 Pipeline for building the representation

We construct subgoals using a two-step process similar to the LUSE framework [13], which employs a VLM model to extract and aggregate procedural steps. First, we extract subgoals from instructional materials by prompting GPT-4o [1] to describe them in a way that is independent of specific inputs, methods, or outputs. This approach ensures that the subgoal descriptions remain generalized. Next, we refine these subgoals by clustering them using hierarchical algorithms [10]. Once the subgoals are established, we extract the corresponding inputs, processes, and outputs from each instructional material, along with any explanations, tips, and descriptions. Finally, we organize these components into taxonomies by applying hierarchical clustering algorithms.

3 Applications

Beyond using our structured representation as useful data model for storing, indexing, and retrieving procedural information, it can enable generative AI applications that provide more adaptive and personalized instructional experiences. With the representation, AI systems can track progress, dynamically adjust instruction, and experiment with new procedural knowledge in a grounded and verifiable manner. We highlight two primary applications: (1) customizing instruction based on user constraints and progress, and (2) synthesizing novel procedures beyond learning.

3.1 Adaptive and Personalized Instruction

Novice users often struggle to articulate their needs, leading to misaligned AI-generated instructions that are too advanced or irrelevant. Our representation can enable an AI assistant coherently assess users' prior knowledge and needs by evaluating familiarity with critical subgoals and IPO elements. For example, an experience with knife is crucial for any task that requires cutting, chopping, or dicing. And if users are not familiar, an AI assistant can either adjust the guidance to involve more detailed explanations and visuals for using the knife, or suggest potential alternatives based on other instructional materials. Due to the structured manner of the representation, the detailed information and potential alternatives are adjacent to each other, and are readily accessible.

Beyond initial customization, our representation can enable AI to track user progress and adapt instructions in real time. Instead of treating interactions as isolated queries, AI can maintain procedural context by recognizing completed subgoals, already used inputs (e.g., tools, materials) and methods, as well as achieved outcomes. Based on this context, AI can dynamically adjust the future guidance, for example by reducing redundant explanations about already used methods or selecting subsequent path that aligns the best with the tools the user has. This state-aware procedural guidance can foster dynamic, personalized learning experiences.

3.2 Synthesizing and Optimizing Procedures

Generative AI can leverage this structured representation to explore and refine procedural knowledge. Research has shown that these models demonstrate decent commonsense reasoning abilities [14], enabling them to propose novel approaches or alternative methods by recombining subgoals or different IPO elements. For example, in cooking, AI might suggest using sous-vide instead of pan-frying, or in crafting, it could explore potential material substitutions. While the generated procedures still need validation, our structured representation serves as a foundation for systematic exploration and iterative refinement, acting as a guide for grounded experimentation.

[1] gpt-4o-2024-08-06

References

[1] Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition.* Addison Wesley Longman, Inc.

[2] Minsuk Chang, Léonore V Guillain, Hyeungshik Jung, Vivian M Hare, Juho Kim, and Maneesh Agrawala. 2018. Recipescape: An interactive tool for analyzing cooking instructions at scale. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[3] Minsuk Chang, Ben Lafreniere, Juho Kim, George Fitzmaurice, and Tovi Grossman. 2020. Workflow graphs: A computational model of collective task strategies for 3D design software. In *Graphics Interface 2020*.

[4] Bogeum Choi, Jaime Arguello, and Robert Capra. 2023. Understanding Procedural Search Tasks "in the Wild". In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 24–33.

[5] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. Generative ai. *Business & Information Systems Engineering* 66, 1 (2024), 111–126.

[6] Luanne Freund, Elaine G Toms, and Julie Waterhouse. 2005. Modeling the information behaviour of software engineers using a work-task framework. *Proceedings of the American Society for Information Science and Technology* 42, 1 (2005).

[7] Ben Lafreniere, Andrea Bunt, Matthew Lount, and Michael Terry. 2013. Understanding the roles and uses of web tutorials. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7. 303–310.

[8] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439* (2023).

[9] Lauren E Margulieux, Mark Guzdial, and Richard Catrambone. 2012. Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications. In *Proceedings of the ninth annual international conference on International computing education research*. 71–78.

[10] Frank Nielsen and Frank Nielsen. 2016. Hierarchical clustering. *Introduction to HPC with MPI for Data Science* (2016), 195–211.

[11] Vipula Rawte, A. Sheth, and Amitava Das. 2023. A Survey of Hallucination in Large Foundation Models. *ArXiv* abs/2309.05922 (2023). https://api.semanticscholar.org/CorpusID:261696947

[12] Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Unsupervised semantic parsing of video collections. In *Proceedings of the IEEE International conference on Computer Vision*. 4480–4488.

[13] Chuyi Shang, Emi Tran, Medhini Narasimhan, Sanjay Subramanian, Dan Klein, and Trevor Darrell. [n. d.]. LUSE: Using LLMs for Unsupervised Step Extraction in Instructional Videos.

[14] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *arXiv preprint arXiv:2201.05320* (2022).

[15] Cristen Torrey, Elizabeth F Churchill, and David W McDonald. 2009. Learning how: the search for craft knowledge on the internet. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1371–1380.

[16] Anh Truong, Peggy Chi, David Salesin, Irfan Essa, and Maneesh Agrawala. 2021. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[17] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:249889477

[18] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. TravelPlanner: A Benchmark for Real-World Planning with Language Agents. *ArXiv* abs/2402.01622 (2024). https://api.semanticscholar.org/CorpusID:267406800