Fine-Grained Action Understanding with Tools in Instructional Videos

Saelyne Yang[1] Jaesang Yu[1] Jae Won Cho[2] Juho Kim[1]

[1]KAIST [2]Sejong University

Abstract

Instructional videos provide step-by-step information on how to achieve a task. Previous research in video understanding has advanced comprehension of the hierarchical structure of procedural information in videos, such as goals and steps, by introducing several datasets and models. While the emphasis has largely been on the 'what' aspects of actions (e.g., frying eggs, cutting a carrot), it is also important to understand the 'how' behind these actions. In this research, we focus on tools (e.g., comb, hair clip) used to perform tasks for a fine-grained understanding of actions, as tools have direct relevance to how actions are performed and drive the changes of the final outcome of the task. To achieve this, we developed an annotation schema that identifies the tools used in a step and how the tool is used for which purpose. Based on the schema, we annotated 48 video clips across 12 domains, each corresponding to a step in COIN [23], an instructional video dataset with taxonomy-based step labels. Our collected dataset reveals the detailed granularity of actions, demonstrating diversity even within identical steps. Through an analysis of the dataset, we demonstrate the significance of tool-centered annotation for fine-grained action understanding.

1. Introduction

Instructional videos provide visual demonstrations of how to achieve tasks, such as '*Cooking pasta*', often complemented by verbal instructions. These videos provide step-by-step guidance toward achieving task goals, containing hierarchical and procedural knowledge. To facilitate procedural video learning, various datasets have been introduced [10, 14, 21, 23, 26, 27, 32, 34, 35]. These datasets are annotated with temporal segment boundaries and the actions performed within each segment, enabling a range of video understanding tasks such as video or moment retrieval [2, 12, 14, 29], video captioning [1, 9, 13, 19, 25, 30], and action recognition or localization [3–5, 8, 11, 16, 18, 24].

While these datasets have advanced video understanding, they primarily focus on the '*what*' aspects of actions rather than the '*how*' aspects. Comprehending *how* actions are performed is vital for understanding the actions and their outcomes as this lies at the core of skill acquisition and human intelligence [20, 22, 28]. Improved perception of the nuances in actions would facilitate both humans and robots to replicate these actions better [6]. Recent efforts in fine-grained action understanding have incorporated learning verb-adverb relationships to distinguish between actions such as 'slice *slowly*' and 'slice *quickly*' [6, 7, 15]. However, they focus on elaborating the action solely through the use of adverbs, but there are much more detailed aspects of how an action is performed. For instance, an action like '*Cutting a carrot*' could entail details such as the tool used or the shape of the cut.

In this research, we propose that incorporating *tools*—the equipment used in task execution—can serve as a means of understanding how actions are performed in more detail. We focus on instructional videos that involve physical demonstrations such as cooking and crafting, which often employ tools to accomplish tasks [31]. The use of tools, whether directly or indirectly, results in changes to the target object of the task, thereby enabling us to understand the step in action units that drive these changes. For example, understanding the action '*Cutting a carrot*' in terms of the tools used would enable us to discern the specific tools employed (such as knives or scissors) and how they are utilized.

To enhance the fine-grained understanding of instructional videos with tools, we introduce a tool-oriented annotation schema and curate a dataset containing action-level descriptions. Based on the COIN dataset [23], an instructional video caption dataset with task and step-level annotations, as a source of videos and captions, we choose two of the step labels from distinct tasks from each of the 12 domains and pair each step with two distinct video clips. With our annotation schema, we produce tool-oriented descriptions for each step, beginning with identifying tools used, specifying the actions performed with each tool, and elaborating on how each tool is used, resulting in a free-form text annotation within a syntactic structure.

Ultimately, we collected 186 action-level descriptions,

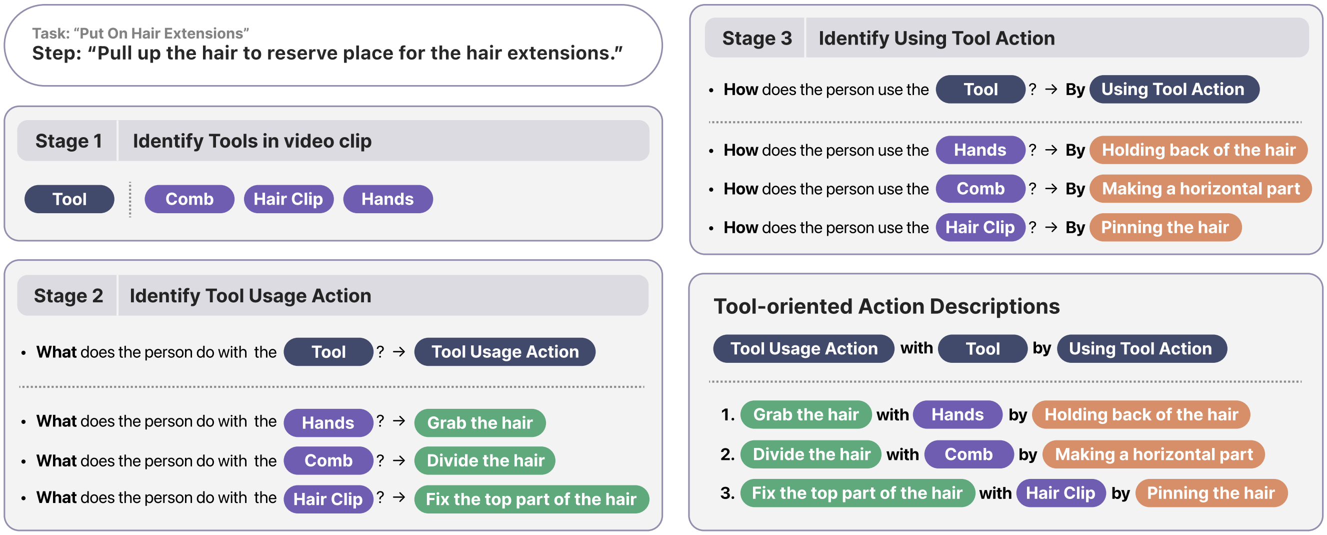

Task: "Put On Hair Extensions"
Step: "Pull up the hair to reserve place for the hair extensions."
Stage 1 Identify Tools in video clip
Tool
Comb
Hair Clip
Hands
Stage 2 Identify Tool Usage Action
What does the person do with the Tool ? → Tool Usage Action
What does the person do with the Hands ? → Grab the hair
What does the person do with the Comb ? → Divide the hair
What does the person do with the Hair Clip ? → Fix the top part of the hair
Stage 3 Identify Using Tool Action
How does the person use the Tool ? → By Using Tool Action
How does the person use the Hands ? → By Holding back of the hair
How does the person use the Comb ? → By Making a horizontal part
How does the person use the Hair Clip ? → By Pinning the hair
Tool-oriented Action Descriptions
Tool Usage Action with Tool by Using Tool Action
1. Grab the hair with Hands by Holding back of the hair
2. Divide the hair with Comb by Making a horizontal part
3. Fix the top part of the hair with Hair Clip by Pinning the hair

Figure 1. Annotation schema designed to capture fine-grained actions in instructional videos. Starting with the step label from COIN [23], Stage 1 asks to identify the tools used in the step in the video clip, Stage 2 asks to identify *what* the demonstrator does with the tool (*Tool Usage* action), and Stage 3 asks to describe the action of *how* the demonstrator uses the tool for its usage (*Using Tool* action). Finally, each tool-oriented action description is crafted by integrating the *Tool Usage* action and the *Using Tool* action alongside the specified tool.

leading to a total of 1.7K words[1]. There is an average of 3.9 actions per step label and an average of 9 words within each annotated action description. Through extensive analysis, we demonstrate the significance of tool-centered annotation for fine-grained procedural video understanding.

2. Dataset Collection

We constructed a dataset that contains action information based on tools used in instructional videos. Below, we describe how we selected the videos and discuss our tool-oriented annotation schema.

2.1. Video Selection

Existing datasets on instructional videos have been annotated using predefined steps [23, 35] or via free-from text [32, 34]. We chose to start with a set of videos annotated with predefined steps and further elaborate on the steps with free-form text annotations, thereby creating a hierarchical structure while ensuring diverse and high-quality captions. We selected videos from the COIN dataset [23], given its coverage of various domains and its hierarchical organization of tasks (*e.g., 'Replace a Bulb'*) and steps (*e.g., 'Take out the old bulb'*) in each task.

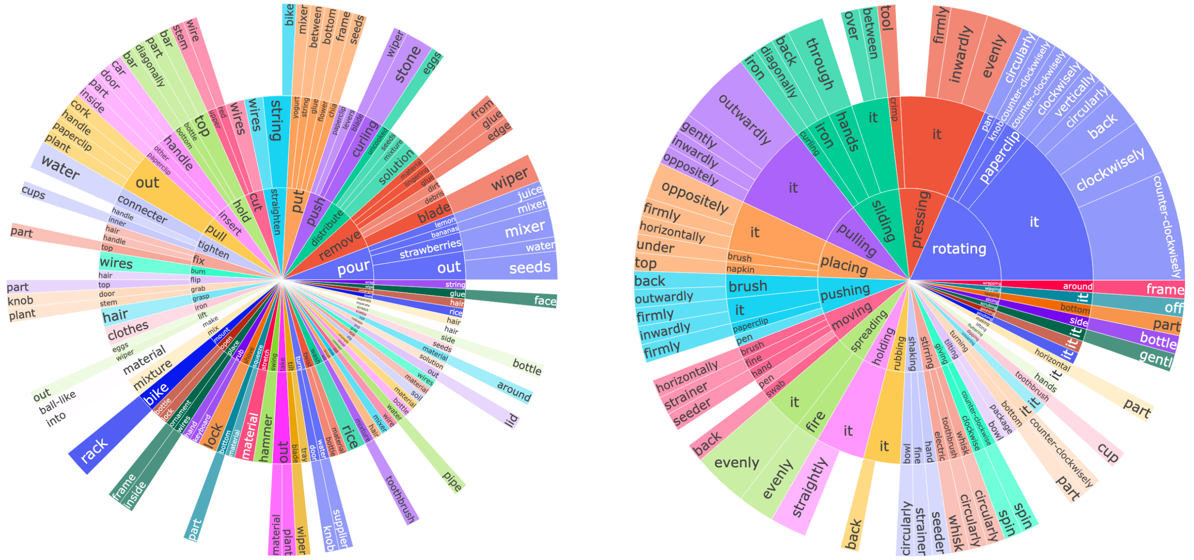
We first randomly sampled two tasks from each of the 12 domains in the COIN dataset. For each task, we select one of the predefined step labels. Here, we avoid selecting steps that are too broad, such as *'Make the detergent'*. To further analyze the differences in a step with more granularity, we chose two video clips for each step label. This resulted in a collection of 48 video clips, spanning a diverse range of tasks and domains. Table 1 provides details regarding the step labels we select.

2.2. Tool-oriented Annotation Schema

For fine-grained action annotation, we focus on the tools used and their role in each step. Since the annotation framework significantly impacts response quality, each component should be designed with substantial consideration [17]. In this work, as a first step, we as the authors have taken on the role of annotators and iteratively revised its design. As a result, we have developed a cascaded annotation schema comprising three stages, with each stage building upon the previous one. Our goal is to keep each step straightforward, enabling annotators to concentrate and build upon each response, and therefore induce enriched annotations.

Figure 1 shows the overall annotation schema we devised, which focuses on annotating fine-grained actions with tool information. Starting from Stage 1, annotators are asked to identify the tools used in the step given a video clip. Here, tools are defined as objects directly used by the demonstrator to execute the step, while materials that are integrated to produce the final target object, such as ingredients, are not considered as tools. Then in Stage 2, annotators are prompted to specify what the demonstrator does with each tool (*Tool Usage* action). Note that the same tool might be used in different actions, and an action might require the use of multiple tools. Stage 3 asks annotators to describe how the demonstrator uses the tool, thereby elabo-

[1] https://anonymous.4open.science/r/how2how2-2225/



(a) *Tool Usage* action

(b) *Using Tool* action

Figure 2. Action descriptions in our dataset can be decomposed into three components: (1) Tool, (2) *Tool Usage* (what the demonstrator does with the tool), and (3) *Using Tool* (how the person uses the tool). The figures describe distributions of natural language descriptions for *Tool Usage* and *Using Tool* actions by their first three words. We removed stop words and stemmed the remaining words. The stemmed words appearing multiple times in one video were counted as unique to analyze the general distribution of the dataset without any bias.

rating the tool usage (*Using Tool* action). Here, we ask them to describe these aspects from a first-person perspective to ensure consistency within the descriptions. Ultimately, each elaborated action description is formed by combining the *Tool Usage* action and the *Using Tool* action alongside the specified tool. Our schema leverage the advantages of both fixed-set [23, 35] and free-form annotations [32, 34], which allow for efficient learning for various video understanding tasks while ensuring elaboration of descriptions. After finalizing each description, annotators are asked to provide the corresponding time segment.

3. Dataset Analysis

With the selected steps and annotation schema, we have gathered a total of 186 action-level descriptions, amounting to 1.7K words, as a preliminary result reflecting our initial steps. On average, each step label contains 3.9 actions, with each action description comprising 9 words. Our annotation schema, designed in a cascaded fashion, facilitated annotation with a broad vocabulary. Furthermore, the schema allows free-form natural language descriptions, resulting in a dataset consisting of 128 unique sentences and a vocabulary of 228 words. Calculating the ratio of the number of sentences to the size of vocabulary as a measure of *sentence diversity* [33], we found that YouCook2 [34] and HiREST [32], which contain action-level descriptions, have averages of 5.4 and 2.3 sentences per word, respectively. Whereas our dataset exhibits an average of 0.8 sentences per word, indicating a high degree of sentence diversity compared to previous instructional video datasets.

Our annotation schema emphasizes the elaboration of actions, focusing on both the 'what' and 'how' aspects of tool usage. These are categorized as *Tool Usage* actions, detailing what the person does with the tool, and *Using Tool* actions, explaining how the person employs the tool for its intended purpose. As a result, each action description in our dataset follows a structure, comprising three parts: the tool itself, a description of its usage, and a description of how it is used. This breakdown enables a comprehensive analysis of how tools are integrated into task demonstrations within an instructional context.

Figure 2 shows how the first three words of our action descriptions vary, indicating that tool-oriented descriptions have a diverse vocabulary within a fixed syntactic framework. Specifically, the *Tool Usage* verbs (Fig. 2a), which detail the purpose of the tools, convey high-level actions and the intentions behind their use in achieving the conceptual goal (*e.g., remove, distribute, straighten, etc.*). Conversely, the *Using Tool* verbs (Fig. 2b), which illustrate the method of using the tools, are associated with lower-level actions, focusing on the fundamental atomic actions undertaken with the tool (*e.g., pulling, pushing, pressing, etc.*).

It is also important to understand both *what* the tools are used for and *how* they are used. For a given tool, there may be various methods of utilization. To examine the diversity in how each tool is employed, we have analyzed three components for each action-level annotation: the tool used, the

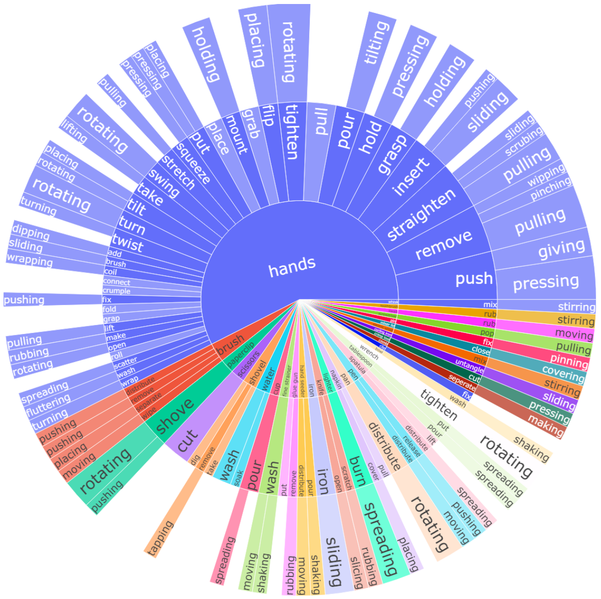


Figure 3. Distribution of *Tool*, *Tool Usage* verb, and *Using Tool* verb. The same tool can be used for different purposes (*Tool Usage*), and even when tools are used for the same purpose, there can be different methods of usage (*Using Tool*).

verb describing the tool's purpose (*Tool Usage* verb), and the verb detailing how the tool is used (*Using Tool* verb). These components were combined to create a unique set, and we counted the occurrences of each set across all tasks and domains. Each set was treated as unique for each video clip to focus on the general distribution. Figure 3 illustrates the overall distribution in the gathered set. This visualization demonstrates that the same tool can be utilized for different purposes, and even when tools are employed for the same purpose (*Tool Usage*), there can be various methods of usage (*Using Tool*). Moreover, the figure highlights the significant role of *hands* as a tool in instructional videos. While *hands* are the most frequently used tool in our dataset, they also serve a diverse range of purposes and use cases. By providing detailed information about the tools, the annotation from our proposed schema enables a fine-grained understanding of instructional videos.

4. Discussion

As we conducted annotation for two distinct video clips corresponding to the same step label within the same task, we were able to analyze how our annotation schema can effectively distinguish between them by breaking down the step into tool-oriented actions. In our dataset, for instance, consider the step label *'Open the bottle carefully'*, within the *'Open Champagne Bottle'* task. One video clip depicts a bottle being opened using a knife, with the demonstrator vigorously scratching the side of it back and forth, while the other video clip shows a bottle being opened using hands, with the demonstrator gently twisting the bottom part of the bottle. Since each video employs different tools to perform the task, this results in a completely different set of actions. Therefore, we argue that capturing the tools used to perform a task or action is a crucial aspect in designing an instructional video dataset for fine-grained action understanding.

Another issue worth discussing is that our annotation schema has been designed to adhere to a fixed syntactic structure for annotations. The action descriptions in our dataset mostly follow a consistent grammar format: *Tool Usage* action (verb + noun) + with + Tool (noun) + by + *Using Tool* action (verb + noun + adverb) (e.g., 'clean the floor with a brush by distributing the solution outwardly'), except a few samples that also include prepositional information (e.g., "put the string *in between the tuning post* with hands by pressing it evenly"), and some without a description of how the tool is used (*Using Tool*). While we have observed that some step-level descriptions are indivisible, resulting in a single action description, this single action description serves to elaborate upon the original step label. For example, the original step label "*push curling*" has been expressed as a single action description "*push the curling stone with hands by giving a clockwise spin*". Thus, our tool-oriented schema was able to facilitate rich elaborations at all levels of step complexity in the COIN dataset.

5. Conclusion and Future Directions

Our tool-oriented elaborated annotations offer valuable resources for advancing language-based procedural video understanding. The atomic action-level descriptions, enriched with a diverse vocabulary from our proposed annotation schema, have the potential to enhance video understanding in various downstream tasks, such as dense video captioning, action localization, and multi-label action recognition. Additionally, our tool-oriented dataset opens up opportunities for novel tasks in tool learning for instructional tasks. Given the crucial role of tools in instructional actions, as highlighted by our analysis, we plan to propose tasks such as tool usage recognition and learning, further enhancing the use of procedural videos.

Moving forward, we are in the process of curating a comprehensive instructional video dataset utilizing our tool-oriented annotation framework. This dataset will cover various domains within instructional videos and include videos that feature a wide range of tools. To demonstrate the effectiveness of our curated dataset and proposed annotation schema, we intend to design a hierarchical benchmark comprising multi-level descriptions. Additionally, we aim to introduce a novel video-tool learning task, which will pave the way for a deeper understanding of procedural tasks through videos.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, and et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1

[4] Joao Carreira and Andrew Zisserman. Action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[5] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with long-memory transformer. In *ECCV*, 2022. 1

[6] Hazel Doughty and Cees G. M. Snoek. How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[7] H. Doughty, I. Laptev, W. Mayol-Cuevas, and D. Damen. Action modifiers: Learning from adverbs in instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 865–875, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 1

[8] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017. 1

[9] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, , and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1

[11] Chi-Hsi Kung, Shu-Wei Lu, Yi-Hsuan Tsai, and Yi-Ting Chen. Action-slot: Visual action-centric representations for multi-label atomic activity recognition in traffic scenes. In *CVPR*, 2024. 1

[12] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 1

[13] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, , and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, 2022. 1

[14] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 1

[15] Davide Moltisanti, Frank Keller, Hakan Bilen, and Laura Sevilla-Lara. Learning Action Changes by Measuring Verb-Adverb Textual Relationships. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[16] Ioanna Ntinou, Enrique Sanchez, and Georgios Tzimiropoulos. Multiscale vision transformers meet bipartite matching for efficient single-stage action localization. In *CVPR*, 2024. 1

[17] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkila, and S. Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14277–14286, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2

[18] Junting Pan, Siyu Chen, Mike Zheng Shou, Jing Shao Yu Liu, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *CVPR*, 2021. 1

[19] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, 2022. 1

[20] Yoav Shoham. *Reasoning about change: time and causation from the standpoint of artificial intelligence*. MIT Press, Cambridge, MA, USA, 1988. 1

[21] Yale Song, Gene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1

[22] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007. 1

[23] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[24] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 1

[25] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, 2021. 1

[26] Weiying Wang, Yongcheng Wang, Shizhe Chen, and Qin Jin. Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. In *EMNLP-IJCNLP*, 2019. 1

[27] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *ICLR*, 2024. 1

[28] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 1

[29] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1

[30] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 1

[31] Saelyne Yang, Sangkyung Kwak, Juhoon Lee, and Juho Kim. Beyond instructions: A taxonomy of information types in how-to videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. Association for Computing Machinery. 1

[32] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *CVPR*, 2023. 1, 2, 3

[33] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Title generation for user generated videos. In *ECCV*, 2016. 3

[34] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. 1, 2, 3

[35] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 1, 2, 3

Fine-Grained Action Understanding with Tools in Instructional Videos

Supplementary Material

6. Selected Step Information

Domain	**Task**	**Step**
Nursing and Care	PutOnHairExtensions	pull up the hair to reserve place for the hair extensions
	RemoveBlackheadsWithGlue	wipe the glue to a layer
Sport	PlayCurling	push curling
	ThrowHammer	pre-swing
Vehicle	InstallBicycleRack	mount the bracket to the back of the car
	ReplaceAWiperHead	take out the wiper
Science and Craft	MakeSlimeWithGlue	rub and drag the materials
	MakeFlowerCrown	stick or bind flower to the frame
Electrical Appliance	CleanLaptopKeyboard	clean the inside of the button
	MakeRJ45Cable	cut a certain length
Furniture and Decoration	InstallClosestool	connect the water pipe
	ReplaceDoorKnob	install the new door knob
Gadgets	MakeWirelessEarbuds	process the copper wire inside the earphone cable
	OpenALockWIthPaperclips	twist the paperclips by hands
Drink and Snack	MakeHomemadeIceCream	stir the mixture
	MakeStrawberrySmoothie	put strawberries and other fruits into the juicer
Leisure and Performance	ChangeGuitarStrings	fix the new string on the head of the guitar
	OpenChampagneBottle	open the bottle carefully
Housework	IronClothes	iron the cloths with the iron
	CleanCementFloor	clean the floor
Dish	CookOmelet	fry eggs
	UseRiceCookerToCookRice	soak and wash the rice
Pets and Fruit	Transplant	take out the plant
	Sow	sow on the soil

Table 1. Selected task and step labels within each domain from the COIN dataset