Beyond Instructions: A Taxonomy of Information Types in How-to Videos

Saelyne Yang
saelyne@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Sangkyung Kwak*
skkwak9806@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Juhoon Lee*
juhoonlee@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Juho Kim
juhokim@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

ABSTRACT

How-to videos are rich in information—they not only give instructions but also provide justifications or descriptions. People seek different information to meet their needs, and identifying different types of information present in the video can improve access to the desired knowledge. Thus, we present a taxonomy of information types in how-to videos. Through an iterative open coding of 4k sentences in 48 videos, 21 information types under 8 categories emerged. The taxonomy represents diverse information types that instructors provide beyond instructions. We first show how our taxonomy can serve as an analytical framework for video navigation systems. Then, we demonstrate through a user study (n=9) how type-based navigation helps participants locate the information they needed. Finally, we discuss how the taxonomy enables a wide range of video-related tasks, such as video authoring, viewing, and analysis. To allow researchers to build upon our taxonomy, we release a dataset of 120 videos containing 9.9k sentences labeled using the taxonomy.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

How-to Videos; Information Type; Video Content Analysis

ACM Reference Format:
Saelyne Yang, Sangkyung Kwak, Juhoon Lee, and Juho Kim. 2023. Beyond Instructions: A Taxonomy of Information Types in How-to Videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3544548.3581126

1 INTRODUCTION

How-to videos provide procedural information about performing tasks such as cooking, makeup, and crafting. They explain how to perform a task by visually demonstrating workflows while providing verbal explanations. Due to their detailed explanations, how-to videos have been a popular source of help when performing a task [11, 26].

There is diverse information beyond instructions intertwined in how-to videos. In addition to instructional information about how to perform each step, instructors share their strategies for choosing supplies [12] or give additional commentary [55]. They also share their personal tips or pitfalls [10], or even ideas not directly related to the task, such as greetings or jokes [35].

From the sea of information, each user requires different information that caters to their specific purpose or situation of watching videos. Depending on their needs, users might want to see only relevant instructions [26], ingredients or tools used, or check the final outcome of a video [39]. To help users find the content of interest, the most common approach has been to enable chapter-based navigation where it segments the video into coherent subtopics [11, 18, 28, 39, 45, 48, 55, 57, 62]. It allows users to navigate videos based on subtopics in videos and locate a section of interest.

However, the diverse information within a video is scattered throughout, making it difficult for users to identify information that meets their needs. Even a chapter contains various types of information. Moreover, the diverse kinds of information are intertwined in no particular order. The author may proceed to offer their rationale, describe intermediate outcomes, or even promote their channels in the middle of giving instructions at any part of the video. The unpredictability of a video's structure makes it even more difficult for users to retrieve the information they need.

We propose that a comprehensive taxonomy that identifies and categorizes the types of information shared in how-to videos can serve as a foundation for supporting users in navigating videos. It provides a structural basis for analyzing and understanding users' navigational behavior. It facilitates the understanding of useful information types for different user needs arising from distinct settings such as the purpose of watching or the domain of the video. Understanding how users leverage information types to navigate

[*] Both authors contributed equally to this research.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581126

videos will ultimately lead to better designs of video navigation systems that suit users' needs.

To this end, we investigated verbal utterances from how-to videos to identify and organize information types in how-to videos. We focused on verbal utterances as the primary source of information because they often contain explicit explanations of what instructors demonstrate [12, 36], sometimes giving additional information that is not visually available. Thus, we presume that verbal information would cover a wide range of information delivered in how-to videos.

To construct the taxonomy, we selected 120 videos from the HowTo100M dataset, a large-scale dataset of narrated how-to videos that covers 12 different genres (e.g., Cooking, Arts, Sports) [36]. We performed an iterative open coding of 4k sentences from 48 videos to generate a taxonomy of information types in how-to videos. From the analysis, 21 information types emerged under 8 categories: *Greeting*, *Overview*, *Method*, *Supplementary*, *Explanation*, *Description*, *Conclusion*, and *Miscellaneous*.

To validate the taxonomy, we applied the taxonomy to a total of 120 how-to videos containing 9.9k sentences which we contribute as a dataset, HTM-Type[1]. From the analysis of the dataset, we found that `Method`, the core information required to complete the task, makes up 47.5% of the video time on average. We also found that the task type (i.e., Creating, Fixing, or Using) and narration style (i.e., Real-time or Dubbing) affect the distribution of information types, and that certain categories have a temporal tendency.

After creating and validating the taxonomy, we demonstrate the utility of the taxonomy in both analyzing users' navigational behavior and supporting their navigation in how-to videos. We first show how our taxonomy can serve as an analytical framework for existing video systems that were built to support video navigation. We observed that the systems utilized different information types to meet users' specific needs. To further investigate how users leverage information types in various navigation tasks, we built a research probe that enables users to navigate using the information types within the video. Through a user study with nine participants, we observed that the participants effectively used different information types for finding specific information needed to perform each of the Search, Summarize, and Follow tasks. We further discuss how our taxonomy can enable a number of applications in video authoring, viewing, and analysis.

This paper makes the following main contributions:

- A taxonomy of information types in how-to videos
- HTM-Type, a dataset of 9.9k sentences from 120 videos labeled according to the taxonomy
- Empirical findings on how people use information types in navigating videos

2 RELATED WORK

Our work proposes a taxonomy of information types in how-to videos and shows how the taxonomy can assist users in video-related tasks. We review related work in the taxonomy of video content, video navigation techniques, and existing approaches to leverage information types in videos.

2.1 Taxonomy of Video Content

Instructional videos, which are videos made to facilitate learning, include several types of videos such as tutorials or lecture videos [16]. There have been many approaches to understanding lecture videos. For example, researchers have classified the video production styles to understand instructional patterns [13, 20, 49]. They have organized video styles according to the level of human presence and the type of instructional media [13] and communication styles [49]. Researchers have further investigated how each presentation style affects students' learning performance, engagement, and attitudes [9, 20, 23, 42].

While the aforementioned work provide a landscape of lecture videos, several approaches investigated the types of content within a video. Espino examined MOOC videos and proposed a taxonomy of verbal information according to their functions and types [16]. Among the video production type, Sugar et al. have focused on screencast instructional videos and offered a screencasting framework that encompasses structural components and instructional strategies that appear in a video [51]. Morain and Swarts also have focused on software instructional videos, proposing a framework for assessing videos based on modalities and information design [38].

Although both how-to videos and lecture videos contain instructional content, they have differences in that how-to videos are more focused on demonstrating procedural knowledge, while lecture videos are more focused on delivering declarative knowledge. While the HCI and Education community mostly focused on lecture videos for generating taxonomies, the Computer Vision community has investigated how-to videos to classify in-video content. Pieces of work classified each sentence of transcripts from narrated how-to videos according to their visual relevancy, which describes how relevant each spoken sentence is to their visual content [21, 35]. Our work builds a foundation for understanding how-to videos in more depth by investigating the types of information shared in a video. To our knowledge, our work is the first approach to creating a taxonomy of information types in how-to videos.

2.2 Video Navigation Techniques

How-to videos provide rich explanations of how to complete a task. However, the linear nature of the video makes it difficult for users to navigate or skim through the content [11, 17, 45]. For example, it is hard to locate a specific point of interest in videos without navigating over a time scale. Researchers have proposed several approaches to overcome such limitations. One of the popular approaches is to segment a video into meaningful sections and create a table of contents [11, 18, 28, 39, 45, 48, 55, 57, 62]. It helps users navigate the video based on semantics and locate a section of interest. To further improve the utility of segmentation, Truong et al. have introduced two-level hierarchical makeup videos, where they organize a set of actions into spatial locations [55]. Similarly, VideoWhiz organized steps in recipe videos by reflecting the dependencies between the steps [39].

Another approach is to identify conceptual objects introduced in videos, which allows users to navigate a video based on objects or concepts of interest [6, 31, 37]. Specifically, RubySlippers [6]

[1] videomap.kixlab.org

focused on a setting where users' hands are occupied with physical activities, which it supports with keyword-based voice commands for navigating videos. A data-driven approach has been introduced as well to improve video navigation. Researchers found that interaction traces of other users help identify points of importance or confusion [27] or the difficulty of each step [61]. Finally, script-based navigation approaches have allowed users to efficiently search the content [27, 43], give feedback on videos [44], or edit videos [3, 15, 22, 54].

In summary, existing methods for video navigation are based on the script, conceptual objects, section, or interaction traces. While the script and conceptual objects allow users to navigate in a finer-grained way, it lacks in supporting navigation in a holistic view. On the other hand, while section and interaction traces allow users to see the overall flow of videos, it does not support detailed navigation. Our research presents a novel unit for video navigation, information types, which allows users to see the overall composition of videos as well as navigate at a shorter segment level. We present findings on how information types enable efficient navigation through a research probe.

2.3 Leveraging Information Types in Videos

To make videos more useful, researchers have leveraged various information types in videos. First of all, many approaches have identified subgoals in instructional videos to segment a video and help users navigate [28, 55, 57]. Moreover, some approaches have identified tools used in how-to videos to incorporate into a segmentation pipeline [55] or explicitly let users know about the required equipment [39]. In the educational context, ViZig [59] has identified types of visual anchor points in slide-based lecture videos such as figures and tables to help learners localize the points.

Another line of work has supported video authoring using information types. For example, DemoCut [12] allows users to annotate a how-to video with five types of markers indicating the type of scene, which are then used for automatic video editing. Promptiverse [29] identifies the types of relations between concepts in lecture videos to generate scalable prompts for learners. In our work, we propose a framework of information types in how-to videos that can enable a number of applications in video authoring, viewing, and analysis.

3 TAXONOMY OF INFORMATION TYPES IN HOW-TO VIDEOS

To examine the diverse information types present in how-to videos, we conducted a content analysis on how-to videos. The goal of our analysis was to identify information types, which are the intent behind the units of content in videos. We chose verbal utterances as the primary source of information in our research scope. This is because instructors often explicitly describe the visual content such as what they are doing or what is happening [12, 36], sometimes giving additional information that is not visually available. However, we also considered visual information as an additional factor to take context into account, because sometimes it is hard to identify the type of information the instructor is delivering just from the textual description. For example, when the instructor uses pronouns such as *"it"* or *"this"*, it is hard to know what they are referring to (e.g., tool, method, or situation). Also, it is hard to tell if a sentence is a joke or an instruction without watching the actual situation (e.g., *"What do you do with the half you have leftover? Dip it in some hummus, of course."*). Below we describe our approach to generating the taxonomy and present the results.

3.1 Methods

3.1.1 *Data Collection.*

We selected videos from the HowTo100M dataset, a large-scale dataset of narrated how-to videos [36]. The dataset covers 12 different genres of how-to videos, organized according to the categories in WikiHow [58]: *Arts and Entertainment, Cars and Other Vehicles, Computers and Electronics, Education and Communications, Food and Entertaining, Health, Hobbies and Crafts, Holidays and Traditions, Home and Garden, Personal Care and Style, Pets and Animals, and Sports and Fitness.* To ensure that we cover a wide range of topics, we selected 10 videos from each of the 12 genres, resulting in 120 videos in total.

We first filtered for videos that were longer than 5 minutes to ensure a sufficient amount of content and that were produced within the last five years (that is, 2017 or later) to reflect the most recent and relevant production trends in how-to videos. To acquire the duration and publication date of the videos, we used youtube-dl [64], open-source software for downloading videos and the related metadata. Then, we went through each of the filtered videos and selected 10 videos from each of the 12 genres that 1) are narrated in English, 2) have one person demonstrating, and 3) are in the scope of "how-to videos", namely explaining how to get a task done[2]. After selecting the videos, we transcribed them using Microsoft Azure Speech-to-text API [2], which transcribes the spoken language in videos with timestamps of each word using Automatic Speech Recognition. Then, we used a BERT-based punctuation model [41] to split the transcripts into sentences.

3.1.2 *Constructing the Taxonomy.*

After selecting the videos, three of the authors performed an iterative open coding for the content analysis of the videos. We individually coded each sentence based on the type they believed it to be conveying. We watched the videos while identifying the types to make sure we incorporated the exact context of each sentence and clarify any errors in the transcript. Also, we split a sentence if it contained two or more information types so that each sentence only contains one information type. The total number of split sentences was around 1% of all sentences. Then, we resolved each conflict through a discussion between the three authors and merged the codes every six videos.

To ensure the validity of our taxonomy, we set two criteria for its construction following the practice in taxonomy development [40]: (1) All elements in the taxonomy should be mutually exclusive (i.e., no overlapping between elements) and (2) the taxonomy should be collectively exhaustive (i.e. cover everything). First, to verify that all elements are mutually exclusive, we convened every session to discuss the discovered information types and whether they were mutually exclusive or could be divided into smaller parts or merged. If there were any ambiguous sentences that could be interpreted as multiple types, we handled those cases by figuring out what factors caused the ambiguity. We divided the types into smaller

[2] HowTo100M dataset occasionally contains videos that are not exactly instructional, such as playing with toys or comparing two products.

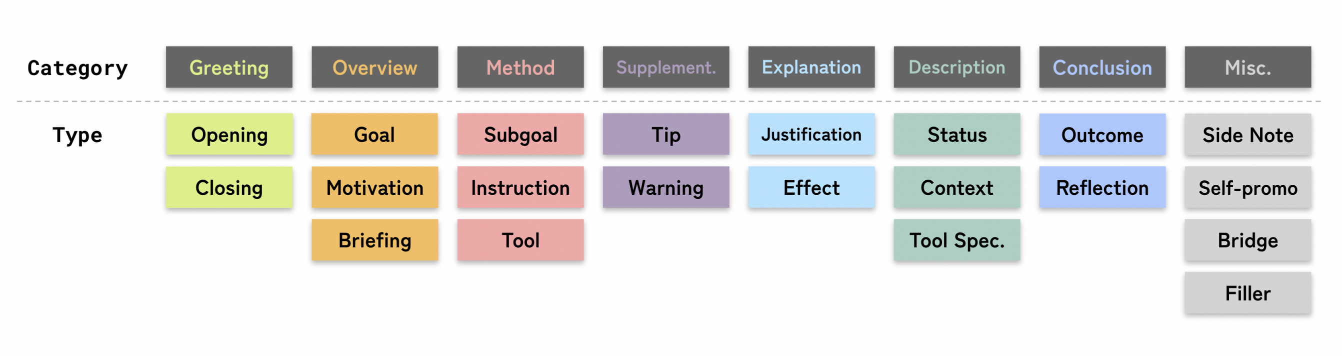


Figure 1: Taxonomy of information types in how-to videos.

components when the types covered multiple intents or merged if the types were redundant.

To make sure the taxonomy covered all information in how-to videos, we checked if any sentence contained information that could not be covered by the existing taxonomy. If so, we added additional types that encompassed the sentence and other similar content. After resolving conflicts and defining new information types, the new taxonomy would be used to reexamine the entire dataset.

Among the entire dataset of 120 videos, we started from an initial set of six videos and repeated the process until convergence was reached; (1) no new types were added and (2) no types were merged or split in the last iteration [40]. If these conditions were not met, we added six additional videos to the investigation. This resulted in an analysis of 48 videos to create the taxonomy. We show that the 48 videos used in constructing the taxonomy are representative of the 120 videos in Appendix A.5.

3.2 Taxonomy

Through the iterative open coding, 21 **types** of information were identified. We further grouped the types into eight **categories** based on what function the types perform in a video. Below we explain the eight categories and the information types under each category in detail. For ease of reading, we denote the various hierarchies as follows: *Category*, and `Type`.

3.2.1 Greeting.

Greeting category offers statements to start and end the video, such as hellos, channel introductions, Intro and Outro, with `Opening` and `Closing`, respectively. `Opening` includes beginning remarks and instructor/channel introductions, such as *"Welcome back to my channel!"* On the other hand, `Closing` gives parting remarks and wrap-up sentences, such as *"I hope you guys enjoyed this video, see you guys next time!"*

3.2.2 Overview.

Overview category discusses the overall structure and information about the video. `Goal` is the main purpose of the video and its descriptions. For example, `Goal` of a cooking video may be, *"Today, we'll be making potato soup."* *Overview* also includes `Motivation`, which is the reasons or background information on why the video was created, such as *"Because everyone is getting a cold these days!"*. Finally, `Briefing` covers a quick rundown of how the goal will be achieved, such as *"I'll be doing a two-step process in this demonstration"*.

3.2.3 Method.

Method provides core information required to complete the task. `Subgoal` outlines the objective of a subsection of the video, such as *"Now, let's prepare all our vegetables."*, without detailing specific directions that the user can follow. Rather, `Instruction` is the action that the instructor performs to complete the task that directly informs the user what they must do, such as *"Now, cut this rubber sleeve off."* `Tool` includes sentences that introduce or show the materials, ingredients, and equipment that will be used during the task, such as *"What we get usually is some cooking aluminum foil."*

3.2.4 Supplementary.

Supplementary information suggests additional instructions or knowledge that aid the core instructions. `Tip` is information given to make the instructions easier, faster, or more efficient, such as *"This step is easiest to complete if you lower the headrest all the way down."* They are typically optional, but helpful advice that arises from the instructor's experience or knowledge. Meanwhile, `Warning` alerts the user on actions that should be avoided to prevent negative consequences, such as *"Don't get too wild with a hammer on there."*

3.2.5 Explanation.

Explanation elaborates on the reasons or consequences of the instruction to help users understand it more clearly. `Justification` is the reason why the instruction was performed. For example, the instructor may decide to use chicken breast because *"it has less fat than chicken thighs."* `Effect` refers to statements that explain the consequences of an action, such as *"Adding this activator will make the slime harden."*

3.2.6 Description.

Description adds descriptions regarding the information relevant to the task, such as the state of the objects or the context of an action. `Status` describes the current state of the object or the target of the task. Sentences such as *"The car is making less noise."* is reporting on how the car is behaving currently and is thus `Status`. `Context` is the description of the method or the setting. For the method, the instructor may point out how arduous a task may be or explain how long it might take, such as *"It will take a while to come up."* For the setting, the instructor could mention, *"The room was really humid, so it took a while to dry."* Lastly, `Tool Specification` adds details and descriptions about the materials,

Category	Type	Definition	Example from Dataset
Greeting	Opening	Starting remarks and instructor/channel introductions	*"Hey, what's up you guys, Chef [...] here."*
	Closing	Parting remarks and wrap-up	*"Stay tuned, we'll catch you all later."*
Overview	Goal	Main purpose of the video and its descriptions	*"Today, I'll show you a special technique which is totally special and about image pressing."*
	Motivation	Reasons or background information on why the video was created	*"[...] Someone is making a very special valentine's day meal for another certain special someone."*
	Briefing	Rundown of how the goal will be achieved	*"I'm pretty sure that just taking a pencil and putting it over the front and then putting a bunch of rubber bands around the pencil [...] that's going to do it."*
Method	Subgoal	Objective of a subsection	*"Now for the intricate layer that will give me the final webbing look."*
	Instruction	Actions that the instructor performs to complete the task	*"We're going to pour that into our silicone baking cups."*
	Tool	Introduction of the materials, ingredients, and equipment to be used	*"I'm also going to use a pair of scissors, a glue stick, some fancy tape or some regular tape."*
Supplementary	Tip	Additional instructions or information that makes instructions easier, faster, or more efficient	*"I find that it's easier to do just a couple of layers at a time instead of all four layers at a time."*
	Warning	Actions that should be avoided	*"I don't know but I would say avoid using bleach if you can."*
Explanation	Justification	Reasons why the instruction was performed	*"Because every time we wear our contact lenses, makeup and even dirt particles [...] might harm our eyes directly."*
	Effect	Consequences of the instruction	*"And these will overhang a little to help hide the gap."*
Description	Status	Descriptions of the current state of the target object	*"Something sticky and dirty all through the back seat."*
	Context	Descriptions of the method or the setting	*"[...] The process of putting on a tip by hand [...] takes a lot of patience but it can be done if you're in a pinch."*
	Tool Specification	Descriptions of the tools and equipment	*"These are awesome beans, creamy texture, slightly nutty loaded with flavor."*
Conclusion	Outcome	Descriptions of the final results of the procedure	*"And now we have a dinosaur taggy blanket that wrinkles, so a fun gift for any baby on your gift giving list."*
	Reflection	Summary, evaluation, and suggestions for the future about the overall procedure	*"However, I am still concerned about how safe rubbing alcohol actually is to use so maybe next time, I will give vodka a try."*
Miscellaneous	Side Note	Personal stories, jokes, user engagement, and advertisements	*"Tristan is back from basketball - He made it on the team so it's pretty exciting."*
	Self-promotion	Promotion of the instructor of the channel (i.e. likes, subscription, notification, or donations)	*"So if you like this video, please give it a thumbs up and remember to subscribe."*
	Bridge	Meaningless phrases or expressions that connect different sections	*"And we're going to go ahead and get started."*
	Filler	Conventional filler words	*"Whoops."*

Table 1: Definition and examples of information types in our taxonomy. Minor errors from Speech-to-Text results in example sentences are corrected.

ingredients, and equipment that may be mentioned in `Tool` or other parts of the video. The difference between the two types is that `Tool` merely establishes the usage of a tool (*"We'll be using some resin."*) while `Tool Specification` supplies other information or characteristics about the tool (*"This resin emits a lot of fumes."* or *"I'll leave a link of where I got it below."*).

3.2.7 Conclusion. Conclusion wraps up the video by showing the final outcome of the task and reflecting on the overall procedure. `Outcome` describes the final results of the procedure, such as *"Look how beautiful our cake turned out."* `Reflection` focuses on the summary, evaluation, and suggestions for the future. The following sentences, *"We made the batter, baked and iced it, and finally decorated it with some fruit."*, *"The process was so easy that even kids can do it."*, *"Next time, let's try using some honey instead of sugar."*, all fall under `Reflection`.

3.2.8 Miscellaneous. Miscellaneous refers to trivial information or phrases devoid of relevant information to the task. `Side Note` includes any sentences that mention personal stories, jokes, and advertisements or try to engage and communicate with the user, such as *"Comment down below what you think about this new look."* `Self-promotion` is the promotion of the instructor or the channel through the encouragement of likes, subscription, notification, or donation features common on creator-based video-streaming platforms, such as *"Please give it a thumbs up."* `Bridge` is meaningless phrases or expressions that connect different sections or phrases, such as *"Let's move onto the next part."* Finally, `Filler` is the conventional filler words prevalent in spoken language, such as *"um"*, *"uh"*, or *"well."*

4 DATASET

To validate the taxonomy, we applied the taxonomy to the remaining 72 videos and contribute the type-labeled 120 videos as a dataset. The dataset can be used to model automatic type detection pipelines or be leveraged to explore various system design opportunities that apply our taxonomy. This section describes the dataset and the following section describes the analysis we performed on the dataset to investigate how videos are structured.

4.1 Method

We applied the taxonomy to the remaining 72 videos (5.9k sentences) to validate the taxonomy and contribute a dataset. Two external fluent English-speaking annotators coded 72 videos based on the taxonomy (6 videos each from 12 genres), where they independently coded the sentences with their types and merged the labels into agreed-upon final labels. Similar to the taxonomy construction process, the annotators watched the videos while labeling the type of each sentence to understand the context behind each sentence and to clarify any errors in the transcript. The annotators were asked to split the sentence if they thought it contained more than one information type. The total number of split sentences was around 1% of all sentences. The two annotators and one of the authors met regularly to discuss ambiguous cases and resolve conflicts. For the last 42 videos (3.4k sentences, with the remaining videos used for training), the two annotators had Cohen's Kappa score of 0.78, which shows a satisfactory level of agreement [1]. After the score was calculated, conflicts were resolved by a discussion between the two annotators and one of the authors. The coding process took approximately 70 hours per coder.

4.2 Dataset: HTM-Type

We release a dataset, HTM-Type[3], which contains a total of 9,918 type-labeled sentences (mean=82.65, SD=21.8) from 120 videos selected from the HowTo100M dataset [36]. It consists of 10 videos from each of the 12 genres identified by HowTo100M. All videos are longer than 5 minutes and published within the last five years (2017 and onward). The average length of the videos is 7 minutes 3 seconds (SD=1 min 35 sec, min=5 min 1 sec, max=14 min 49 sec), totaling 14.1 hours. The average portion of spoken language is 82.4%, representing the average portion of the entire video in which the author talks (min=50.5%, max=97.6%). The dataset denotes for each sentence the id, publication date, duration, and genre of its video, as well as start and end time stamps, and type and category categorization.

5 ANALYSIS

To understand the structure of how-to videos, we analyzed the HTM-Type dataset in three different aspects: **(1)** how each information type is distributed across the dataset, **(2)** how the video style affects the type distribution, and **(3)** how information type distribution relates to time.

5.1 Method

For all three analyses, we first identified the proportion of each information type in a video by calculating the start and end timestamps of each labeled sentence. Afterward, we divided the time portion of each type by the total time of the video containing narration to obtain the final proportion.

(1) The first analysis aims to observe how the information types are distributed throughout the how-to videos. We calculated the average distribution of each type across the entire dataset by dividing the total time proportion of each type by the number of videos.

(2) The second analysis examines how the video characteristics affect the information distribution along two different attributes: task type and narration style. We chose task type and narration style specifically as the analysis axes as they require different strategies by the instructor in providing the information. For example, explaining how to fix a car likely attributes a larger portion of the video to describing the situation in comparison to baking cookies.

To compare whether video characteristics affect the distribution of the information type, we performed the Kruskal-Wallis test for each of the two attributes with an α value of 0.05 for each category. We further performed the Kruskal-Wallis test on types within the different categories if the category showed a significant difference. To confirm which specific video characteristics differed from one another, we further performed post-hoc Dunn's test with Bonferroni adjustment on significantly different categories or types.

(3) The third analysis aims to investigate any specific patterns that may appear in the temporal distribution of each category. To do so, we normalized video time to [0, 1000] seconds to align all

[3] Abbreviated from HowTo100M-Type

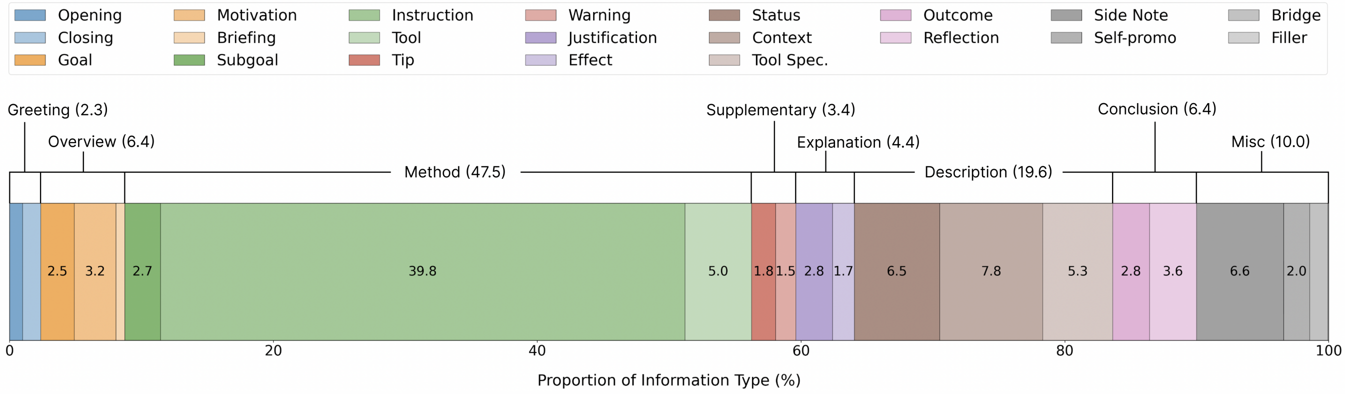


Figure 2: Distribution of Categories and Types of all videos in HTM-Type. Categories are denoted above the types using group brackets. Only proportions greater than 1.5% are written in text. `Instruction` makes up 39.8% of the total video, suggesting that the majority of the video contains information that does not directly give actions for the user to follow. The results illustrate the large diversity of information types in how-to videos.

the videos in the dataset. Then, we counted each type occurrence across all 120 videos for every second on the normalized timeline. As none of the videos in the dataset are longer than 1000 seconds, the normalization will not drop any labels. Afterward, we calculated the range on the normalized timeline that contains data points between the 5th and the 95th quantile for category.

5.2 Results

5.2.1 Information Distribution in How-To Videos. We first investigated the composition of the dataset to look into how the diverse information is distributed over how-to videos. The results for categories and types are shown in Figure 2. The average number of types in a video is 7.25 for category and 14.57 for type, signifying that the videos comprise a wide variety of information. Additionally, the large variance of the types suggests diverse variations in how the information is composed within instructional videos (Appendix A.1).

On average, the results show that almost half of the video comprises *Method* (47.5%, SD=16.9%). Looking at the type level, `Instruction` makes up 39.8% of the total video, meaning that the majority of the video contains information that does not directly give actions for the user to follow. The ratio shows a resemblance to the percentage of visually alignable narration as explained by Han et al. [21] (30%), which is a narration that is visually demonstrated or shown in the video. As instruction usually entails the majority of the visual information, the similarity may imply some correlation.

5.2.2 Information Distribution Based on Video Characteristics. We then analyzed how the video characteristics (i.e. task type and narration style) affect the information distribution. Through the analysis, we found that the composition of information types in a video differed by its characteristics, which we describe below.

Task Type. The first aspect examined is the type of task completed. Through an iterative process, we found three different task types: Creating, Fixing, and Using. Creating refers to tasks whose primary goal is to craft or make a final product, such as cooking or woodworking. Fixing tasks address a problem and improve the state of an object or a situation. Using tasks aim to demonstrate how a tool or equipment is supposed to be used. Our dataset contains 82 videos for Creating, 27 videos for Fixing, and 11 videos for Using.

The results of the Kruskal-Wallis test show significant differences between the tasks for *Description* (H(3)=21.696, p<0.001) and *Miscellaneous* (H(3)=10.435, p=0.015). Further performing the Kruskal-Wallis test on the types in the *Description* and *Miscellaneous* categories reveals that `Status`, `Context`, and `Side Note` are significantly different.

Further performing post-hoc Dunn's test with Bonferroni adjustment showed that Creating-Fixing and Using-Fixing pairs for `Status` and Creating-Fixing for `Context` are significantly distinct in their distributions ((Z=-2.680, p=0.022), (Z=3.126, p=0.005), and (Z=-2.443, p=0.043) respectively). Fixing (10.0%) has a greater proportion of `Status` than Creating (5.7%) and Using (3.3%). For `Context`, Fixing (11.6%) is greater than Creating (6.2%) by 5.4% (Appendix A.2). Such differences can be explained by the tendency for Fixing tasks to require more descriptions of the target object. Conveying `Status` in Fixing videos lays the necessary foundation to communicate the instructions effectively. Likewise, Fixing has more explanations than Creating about the method and the setting because the user needs to fully grasp the current circumstances before they can improve upon them.

Narration Style. The second aspect is the narration style of the video. Videos were classified by how the instructor provided verbal information — whether the narration was spoken in real-time with the action or dubbed afterward. We found 78 videos are real-time narrated and 42 are dubbed videos.

The results of the Kruskal-Wallis test on the categories showed that *Method* and *Description* show significant differences between the narration styles ((H(1)=6.602, p=0.01) and (H(1)=7.036, p=0.008), respectively). To figure out how each type distribution differs within the two categories (*Method*, *Description*), we further performed the

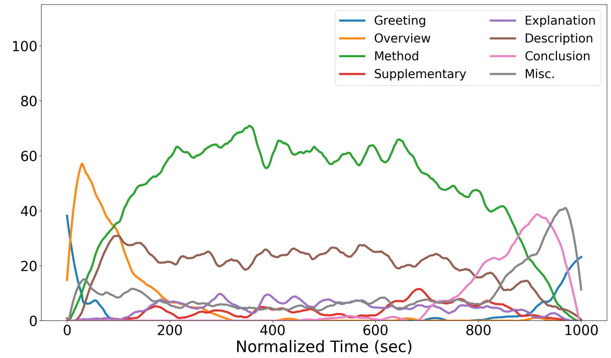


Figure 3: The number of labels for the category along normalized time. *Greeting, Overview, Conclusion,* and *Miscellaneous* show clear positional preferences while *Method, Supplementary, Explanation* and *Description* are widely distributed.

Kruskal-Wallis test for each type in the categories. Instruction and Tool specification have significant differences in their distributions ((H(1)=7.568, *p*=0.006) and (H(1)=4.043, *p*=0.04), respectively). When comparing the absolute value of each type proportion on average, for Instruction, dubbed videos (45.0%) contain an 8.1% greater portion than real-time narration videos (36.9%). On the other hand, for Tool Specification, real-time narration videos (4.2%) have more than dubbed videos (5.9%) (Appendix A.3).

The differences show that video styles can affect the distribution of information. Real-time narrated videos contain a larger portion of descriptions such as Tool Specification, Status, and Context. One possible reason may be that the instructor dedicates more time to explaining the current status quo as they actually perform the task.

5.2.3 Information Distribution Based on Time. We then analyzed the temporal distribution of each category to see if they showed any specific patterns. We report the resulting quantiles and standard deviations for each category in Appendix A.4. We visualized the data with a time-series graph (Figure 3).

The results show that certain categories have a positional preference. *Greeting* shows skewed distributions towards both ends of the video. Such a trend reflects the tendency for instructors to begin or end their videos by greeting their audiences. *Overview* occupies the first (23.8%) of the video, as it covers the overall structure or encompassing details of the video. Meanwhile, *Conclusion* lies in the last (28.0%) of the video. In contrast, *Method* (11.1% to 85.3%), *Supplementary* (16.9% to 86.3%), *Explanation* (16.8% to 87.2%) and *Description* (8.5% to 86.9%) are relatively evenly distributed towards the middle of the video. Finally, *Miscellaneous* extends throughout the video (4.8% to 98.0%) with a noticeable increase at the end (Figure 3), attributed to the abundance of self-promotion and side notes (e.g., outtakes).

6 TAXONOMY AS ANALYTICAL FRAMEWORK

In this section, we demonstrate how our taxonomy can serve as a conceptual and analytical framework for understanding existing systems that support video navigation. Existing video navigation systems are designed to address specific user needs. Our taxonomy provides an opportunity to analyze the information types that each system focuses on. Such an analysis can be used to identify important information types that best fit the users' context and also reveal information types that are underexplored by existing systems.

For instance, ToolScape [28] and MixT [11] have identified step-by-step information (Subgoal) with representative images for each step (Status) to allow users to navigate videos based on important milestones. To better support navigation in a specific video genre, VideoWhiz [39] has extracted ingredients (Tool) and intermediate outcomes (Status) in food recipe videos, and Truong et al. [55] has leveraged makeup tools (Tool) in makeup tutorial videos. To support users navigating videos in a setting where they use voice commands, RubySlippers [6] has allowed users to refer to objects (Tool) and actions (Instruction) that appear in the video.

As such, existing systems have leveraged different information types to address specific needs in video navigation, which we list more in Table 2. We can see that the types in the *Method* category (i.e. Subgoal, Instruction, and Tool) are commonly used, while Goal, Status and Outcome are also used to some extent. At the same time, our investigation reveals that the other information types are underexplored by existing systems, such as Motivation or Context. We believe that future systems can establish important units based on the identified information types catered to user needs.

7 EXPLORATORY USER STUDY

From the preliminary analysis presented in Section 6, we demonstrate how our taxonomy could serve as an analytical framework for understanding existing video navigation systems. To further explore the potential of the taxonomy, we conducted an exploratory user study. Our study aimed to investigate how users would leverage the information types for navigating videos, by exposing information types to users and allowing them to navigate videos using the information types as a control mechanism. Through the study, we demonstrate the usefulness of the taxonomy both in accessing desired content and as a tool for observing and analyzing users' navigational behavior. We chose not to conduct a comparative study because the purpose was not to evaluate the video interface itself but rather to highlight the potential of the taxonomy in supporting video navigation, an aspect that has been underexplored in previous research. Below we explain the research probe used in the study, the study procedure, and the results.

7.1 Research Probe

As the apparatus of the study, we built a video interface that supports navigation based on information types (Figure 4). Users can see the video on the left (Figure 4a) and transcripts of the video on the right (Figure 4b). In the transcript panel, users can see each sentence of the transcript along with its timestamp and information type. The type label is color-coded based on the category of the taxonomy. The timeline also shows the same information below the video (Figure 4c). Each segment is color-coded based on its category and users can hover over each segment to see its type (Figure 4d). The type of the current segment is always shown right next to the

System	Type	Explanation
ToolScape [28], MixT [11], Fraser et al. [18]	Subgoal, Status	Presenting step-by-step information (Subgoal) with representative images for each step (Status)
Truong et al. [55]	Tool, Instruction, other types	Labeling segments as tool introductions (Tool), makeup application (Instruction), or commentary (other types)
VideoWhiz [39]	Tool, Subgoal, Status, Outcome	Presenting ingredients and equipment used in a recipe (Tool), visual milestones (Status, Subgoal), and the appearance of the final output (Outcome)
RubySlippers [6]	Tool, Instruction	Allowing users to refer to objects (Tool) and actions (Instruction) that appear in the video
Pause-and-Play [47], SoftVideo [61]	Instruction	Segmenting software tutorial videos into actionable steps (Instruction)
Weir et al. [57]	Goal, Subgoal, Instruction	A breakdown of a task into the goal (Goal), subgoals (Subgoal), and individual steps (Instruction)
Yang et al. [60]	Tool, Instruction	Segmenting recipe videos into actions (Instruction) and visualizing their dependencies as well as ingredients used (Tool) throughout the video.

Table 2: Example systems that support video navigation and information types associated with each system.

progress bar. Users can click either on the timeline or the script to navigate through the video. Finally, users can filter segments based on their type or category in the Filter panel (Figure 4e). Here, we grouped the categories into four high-level sections to help users better organize the types and categories: Intro, Procedure, Closing, and Miscellaneous[4]. We organized the categories based on their temporal positions reflecting our analysis in Section 5.2.3. Once users select certain types from the Filter panel, only the filtered segments are shown in the transcript panel and in the timeline. The video player automatically skips unselected portions.

7.2 Study Procedure

We recruited nine participants (6 male, 3 female, mean age=24.1, SD=2.26, min=22, max=29) through an online recruitment posting. All the participants watch how-to videos regularly, at least once a week. Participants performed three types of tasks: Search, Summarize, and Follow. These tasks represent real video-watching scenarios and are commonly used in evaluating video navigation systems [6, 27, 28, 55]. We chose three videos from HTM-Type that cover different tasks: Cooking[5], Slime[6], and Illustrator[7]. The Cooking video teaches how to make soft-boiled eggs. The Slime video explains how to make cloud slime. The Illustrator video demonstrates how to convert raster images to vector images. To minimize learning effects, different videos were used in each task. The videos used for each task were counterbalanced between the participants.

- **Search** task asked participants to find an answer to a given question from the video. For example, for the Illustrator video, the task asked: "*To make the image more cartoonish, which feature do you need to adjust?*" There were three search questions for a video, which we include in Appendix A.5.
- **Summarize** task asked participants to summarize the main points of the video while skimming through it. We asked participants to assume that they are making written instructions from the video content. We gave participants freedom in the content and format of the summary.
- **Follow** task asked participants to follow the task in the video. We prepared the tools used in each video. For the cooking video, we simulated the cooking environment with hand-made apparatus such as a stove made of paper.

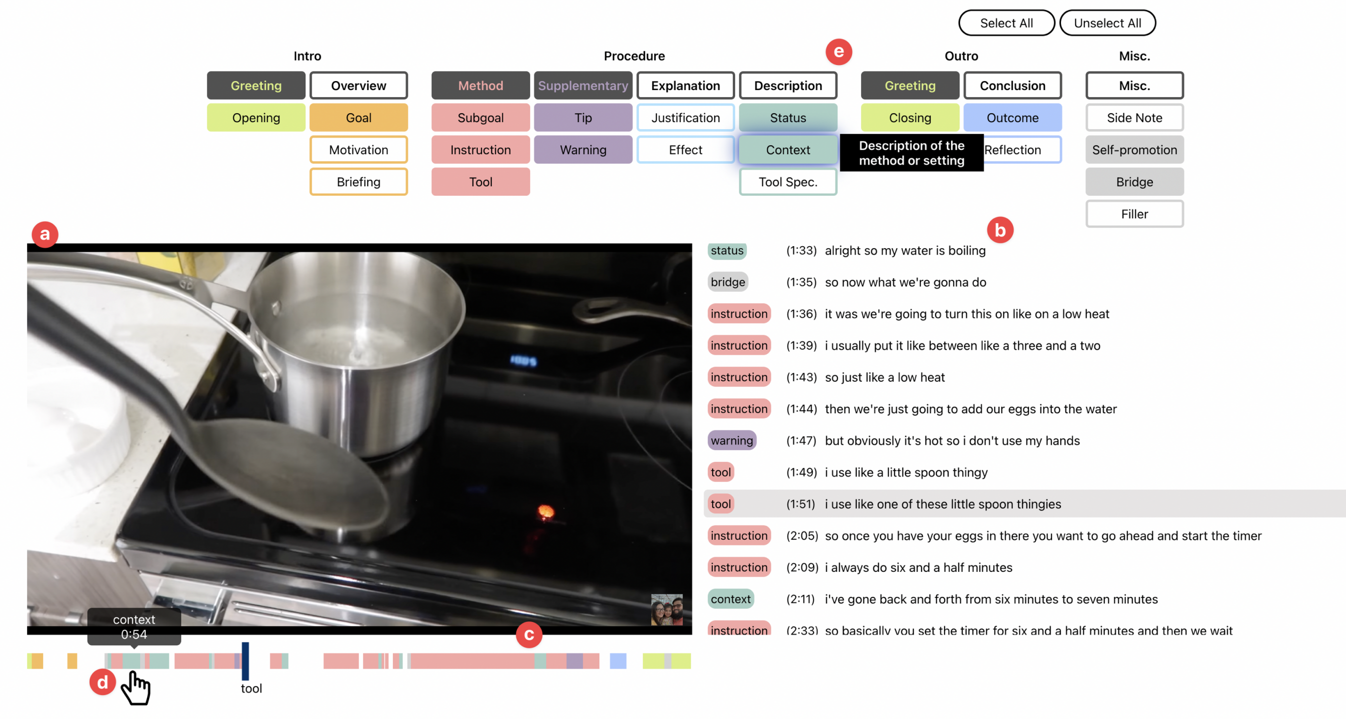
We first gave a tutorial on the system to the participants. After explaining its features, participants tried out the system with a video that was not used in the three tasks. Then, we explained the taxonomy presented in the system. After explaining the definitions and examples of each type, participants watched a video with our interface from beginning to end to get used to the taxonomy. Participants were subsequently asked to perform three tasks in the following order: Search, Summarize, and Follow. To accurately evaluate the role of information types in each task, participants were not allowed to use the browser’s native search function (i.e., Ctrl+F) in the transcript. After each task, we asked a few questions about their task strategy. After all the tasks were done, we conducted a semi-structured interview and survey, asking about their experience and perceptions of the taxonomy. Participants were compensated with 20,000 KRW (~15 USD) for a 1.5-hour-long study.

[4] In the process of grouping, Opening and Closing, which belong to the *Greeting* category, were divided into Intro and Outro, respectively.

[5] youtu.be/6CJryveLzvI

[6] youtu.be/Rcsy2HRuiyA

[7] youtu.be/_Yb6xLqvsf0



status (1:33) alright so my water is boiling

bridge (1:35) so now what we're gonna do

instruction (1:36) it was we're going to turn this on like on a low heat

instruction (1:39) i usually put it like between like a three and a two

instruction (1:43) so just like a low heat

instruction (1:44) then we're just going to add our eggs into the water

warning (1:47) but obviously it's hot so i don't use my hands

tool (1:49) i use like a little spoon thingy

tool (1:51) i use like one of these little spoon thingies

instruction (2:05) so once you have your eggs in there you want to go ahead and start the timer

instruction (2:09) i always do six and a half minutes

context (2:11) i've gone back and forth from six minutes to seven minutes

instruction (2:33) so basically you set the timer for six and a half minutes and then we wait

Figure 4: Our research probe used in the user study. (a) Users can see the video. (b) Each sentence of the script is shown with its timestamp and information type. Each type label is color-coded based on the category. (c) The same information is shown in the timeline. (d) When users hover over each segment, they can see the type and (e) its definition in the Filter panel. Users can filter segments based on their type or category in the Filter panel. Only the filtered segments are shown in the transcript panel and the timeline.

8 RESULTS

The participants were able to find and use appropriate types or categories of the taxonomy to complete the tasks. Below we explain how they used the taxonomy and the information types they perceived as important in detail. Then, we discuss how the participants perceive the prototype and the taxonomy.

8.1 How Taxonomy Was Used in Each Task

8.1.1 Search. The participants' strategy to search for the answer to questions was to relate a given question to a type and filter the video according to the type. For example, for a question asking about how the recipe is different from others (Slime), P3 thought it would be described when the instructor talked about the goal. Thus, he filtered the video to only see `Goal` and found the answer. For this task, participants looked for different information types depending on what each question asked. All the participants were able to match at least two questions out of three correctly to corresponding types (mean=2.44/3, SD=0.53), and thus found answers effectively. We include the list of questions and corresponding types in Appendix A.5.

8.1.2 Summarize. The participants actively used the information types and found them helpful when summarizing videos. In response to 5-point Likert scale questions about how helpful each category and type's existence was (including the removal of them), participants indicated that the existence of all of the categories (mean=4.61/5) and types (mean=4.66/5) were useful, when asked about each category and type individually (Figure 5-left).

When asked about the importance of each category in summarizing videos, they rated *Method* and *Overview* as the top two categories that contain the most important information (Figure 5-right, 4.89 and 4.11/5, respectively). Not surprisingly, all the participants looked for the Method category, as they are the main points of videos. Regarding *Overview*, P3 said, *"I looked for Overview because I felt it is necessary to include the purpose of the task when summarizing the video content."*

From the per-type evaluation, the participants rated `Instruction`, `Subgoal`, `Tool`, and `Goal` as the top four important types (4.89, 4.78, 4.78, and 3.89/5, respectively). Regarding `Instruction`, all the participants included instructions in their summaries (n=9) as they are the essential information in how-to videos. Interestingly, participants not only used the `Subgoal` information to organize their summary by subgoal unit (P7) but also to check and see if they have missed anything at the end (P3, P4). Participants also included the tools used in the video (n=5) and the goal of a video (n=6) in their summaries, along with a description of the goal (n=2) and warning (n=1). Additionally, some participants (P1, P6) looked for `Reflection`, expecting the part to provide a summary, although

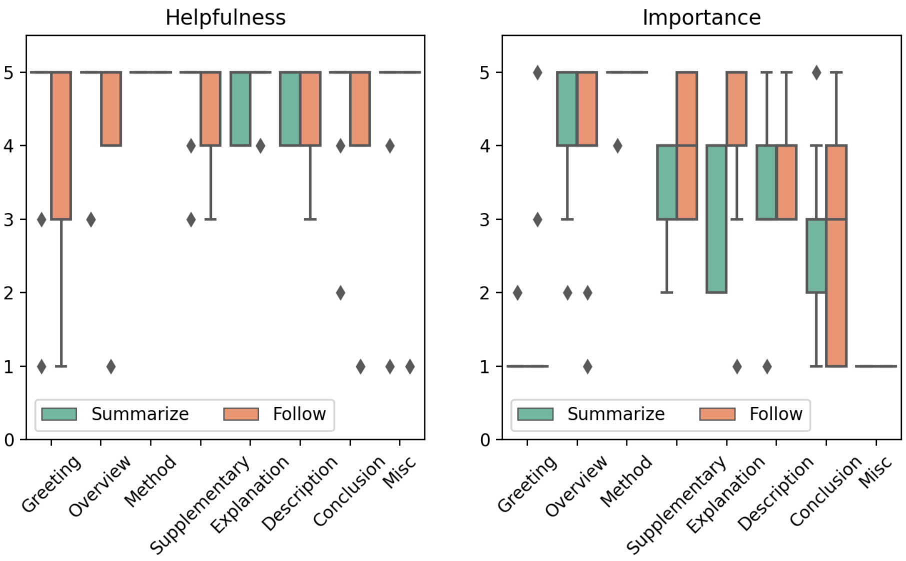


Figure 5: Helpfulness (left) and Importance score (right) of each category in the Summarize and Follow task.

the video did not include any summary information and thus rated low (2.67/5). All the types under the *Greeting* and *Miscellaneous* categories are rated the lowest (mean=1.61/5), as they do not include any task-relevant information.

8.1.3 Follow. In following the task performed in the videos, the participants perceived the information types to be helpful. In response to 5-point Likert scale questions about how helpful each category and type's existence was (including the removal of them), participants indicated that the existence of all of the categories (mean=4.35/5) and types (mean=4.32/5) were useful, when asked about each category and type individually (Figure 5-left).

When asked about the importance of each category in following the videos, they rated *Method*, *Supplementary*, and *Explanation* to be the top categories that contain important information (Figure 5-right, 5, 4.11, 4.11/5, respectively). Not surprisingly, participants thought *Method* contained most of the information they should follow. After *Method*, the participants perceived *Supplementary* and *Explanation* to be important, which was different from the Summarize task. The participants thought the *Supplementary* category which includes `Tips` and `Warnings` to be important. P4 said, *"I thought tips and warnings are too detailed information for the Summarize task. However, they were necessary when following the video as they might contain important notes."* They also found the *Explanation* category which includes `Justification` and `Effect` to be helpful. P3 said, *"It was helpful to know the reasons behind instructions because then I can apply instructions to my context adaptively. For example, if I understand that the reason instructor boils eggs for six minutes is that it's the medium part of being too runny and firm, I can adjust the duration according to my taste."*

From the per-type evaluation, participants rated `Instruction`, `Subgoal`, and `Tool` as the top three important types (4.89, 4.78, and 4.45/5, respectively), followed by `Effect`, `Tip`, `Warning`, and `Justification`, and `Status` (4.11, 3.89, 3.89, 3.67, and 3.67/5, respectively). The participants used `Effect` and `Status` to make sure they are following correctly. P7 said, *"I considered Effect to be important because I wanted to check that the consequences of an action explained in the video are actually shown in my context."* Similarly, P8 said, *"I looked for Status to see if there is a desired state, and if so, I would have liked to refer to it when following."* We could see that the participants mainly focused on instructions while looking for additional information when following videos.

8.2 Effect of Taxonomy on Video-watching Experience

All the participants appreciated that the system enabled selective watching of videos. P8 said, *"When watching how-to videos, I usually watch the video at twice speed or skip parts because there is a lot of unrelated information. It was nice to be able to get rid of useless information."* Selective watching can also be helpful in repeated watches. P5 said, *"I think the system will be helpful especially when you watch a video again and again. For complex tasks like repairing, it is hard to perform the task at once. If you know where to watch repeatedly, it will be efficient."*

Some participants compared the selective watching feature to YouTube's Chapter where it segments a video into meaningful sections [63]. P2 and P4 appreciated that our system offers more details. P2 said, *"In YouTube, we can also skip some parts but it's based on topics. We still have to search within a topic by trial and error, to see the exact part I want."* However, other participants mentioned that the amount of higher-level information they could perceive for each section was limiting. P5 said, *"I could skip parts with the prototype, but YouTube chapters indicate subgoals better with a concise title, which makes it easier to access desired parts."*

The information type was helpful in grasping the overall content. P6 said, *"By looking at the timeline, I was able to quickly understand* *how the whole video is composed of. For example, from the timeline, I was able to figure out the style of the video, such as whether this video has a lot of intro or outro, or whether it has a lot of unrelated miscellaneous information."* It also allowed the participants to grasp the main points quickly. P8 said, *"I was able to understand the flow of the video quickly, by looking at the instructions only."* Participants also thought that it highlights important information for them. P5 said, *"Warnings are important information but they can be unnoticed easily. The prototype helped me identify them."*

8.3 Perception Toward Taxonomy

Overall, the participants were able to understand the meaning of each category and type well (Category mean=4.86, Type mean=4.75). They mentioned that the types were intuitive (P3), and they were able to see the reasoning behind the categorization (P9). All the participants mentioned that each sentence was well-matched with appropriate types, except for a few that were subjective. One feedback that many participants had in common was that the categories would be enough for filtering the video content (P1, P3, P4, P9). While types allowed for more precise control (P6), it was burdensome to recall the meaning of each type and click them one by one due to the large number of types (P9). In the same context, several participants also suggested indicating whether a type exists in the video so that they do not have to manually click to see if it is in the video. As such, when designing systems that display taxonomic information, we need to consider ways to reduce users' cognitive burden.

9 DISCUSSION

In this paper, we present a taxonomy of information types in how-to videos. We first demonstrated how our taxonomy can serve as an analytical framework for existing video navigation systems. We then investigated the utility of the taxonomy in video navigation through a user study. In this section, we first reflect on the user study and discuss findings. We then discuss how the taxonomy enables various video-related tasks and support the learning experience, and suggest opportunities for future work.

9.1 Information Type That Fits the User's Needs

While the essence of how-to videos is information that explains how to perform a step (i.e. `Instruction`), our taxonomy identifies a total of 21 information types that span instructions and beyond. From our user study, we could see that the participants used different information types for each task. In the Search task, they were able to actively match the corresponding information types to each question, finding answers effectively. In the Summarize task, *Method* and *Overview* were considered important – the participants used *Overview* to summarize the goal and overall approach. In the Follow task, in addition to *Method* that provides core information required to complete the task, the participants also considered *Supplementary* and *Explanation* important in getting additional information needed in following the video.

Just as important types vary depending on the task, our study also suggested that meaningful information types can depend on various factors such as the topic of the video or the user's level of expertise. P6 said, *"In videos teaching how to play tennis, justification or effect might be more important than just instructions. It is important to understand WHY a certain movement is needed to actually understand and follow the movement."* It also echoes Semeraro et al.'s finding on instructional videos for physical training, where having verbal cues helped users contextualize the movement [50]. Users' familiarity with the topic also affects which information types they focus on. For example, P8 was unfamiliar with Adobe Illustrator so she checked *Overview* for goal descriptions when following the video. She said, *"I would have skipped the part if I were familiar with the program."* Future work will need to investigate relevant information types depending on the topic and user context.

Moreover, some participants suggested further specification of instructions based on their importance. In how-to videos, there are optional or conditional instructions that users can choose to follow or not according to their preferences or environment. P6 mentioned that *"I thought all the instructions are necessary, but there were some instructions that I didn't need to follow. It would have been nice if it had been marked."* In fact, four participants additionally marked optional or conditional instructions in their summary when performing the Summarize task, which implies the importance of such information. As such, future work can specify the instruction types to support users' detailed needs.

In summary, our findings suggest that 1) information types other than *Method* can also play an important role in accessing desired information, which opens up opportunities for future systems to take into account a variety of information types. Our findings also suggest that 2) relevant information types can be different depending on the task, topic, and user context, which future work can investigate more in depth to support users' different needs. We hope that our taxonomy can serve as a starting point for such investigations.

9.2 Applications of Taxonomy in Video Tasks

The taxonomy can accelerate the design process of multiple applications if videos were labeled by information types. We examine possible applications in three of the most commonly performed video-related tasks: Authoring, Viewing, and Analysis. The creator first produces a video (Authoring), and then viewers watch it (Viewing). The creator can analyze the video content or viewership to improve the original video and make decisions about upcoming content (Analysis). We discuss how our taxonomy enables various applications in each of these tasks.

9.2.1 Authoring. Having a video labeled by the taxonomy can foster the video editing process. For example, instructors can find fillers or side notes that they have made, thus removing or fast-forwarding the parts if necessary. They can also add visual effects to parts that need extra attention, such as tips or warnings, or make transition effects when moving to the next step introduced by a subgoal. They can also add subtitles or textual descriptions and style them differently, depending on what and how much they want to emphasize [30].

Our taxonomy also aligns with the components that facilitate video editing found in previous papers. DemoCut [12], a video editing system designed for how-to videos of physical demonstrations, supports five types of markers to assist in video editing: Step, Action, Closeup, Supply, and Cut-out. The system segments a video and applies editing effects based on the markers. Our taxonomy aligns

Application		Explanation	Example
Authoring	Editing	Removing or fast-forwarding parts of the video	Cut out irrelevant parts of the video (`Side Note`)
	Annotation	Adding visual effects or captions to the video	Highlight important parts of the video (`Tip`, `Warning`)
Viewing	Navigation	Supporting users to find relevant portions of the video	Repeat an instruction segment or jump to the next instruction (`Instruction`)
	Summarization	Providing a summary of the main points of the video	See an outline of how the goal is achieved (`Subgoal`, `Instruction`)
	Search and Selection	Supporting users to make a decision on which video to watch	See if one has required tools to follow the video (`Tool`)
Analysis	Feedback	Providing feedback to the author of the video about the content	Inform the author about how structured the video is (`Subgoal`)
	Comparison	Comparing content between multiple videos	Compare how approaches toward a same goal are different (`Instruction`)

Table 3: Possible applications of the taxonomy in video authoring, viewing, and analysis.

with several types of the markers, such as Step (`Subgoal`), Action (`Instruction`), Supply (`Tool`), or Cut-out (`Miscellaneous`).

9.2.2 Viewing. Our study revealed that the taxonomy can improve users' viewing experiences by enabling them to quickly find and skip irrelevant information based on the category and the type. Our findings echoes with Chang et al.'s finding on the types of jumping in how-to videos: Reference Jump (reminding users of past content), Replay Jump (re-watching a segment of the video), Skip Jump (skipping less interesting content), and Peek Jump (skipping ahead to see what to expect) [8]. Reference and Replay Jumps can happen around `Instruction`, to clarify any confusion and better understand the instruction. Skip Jump can happen around `Greeting` or `Side Note`, where a user wants to skip task-irrelevant parts. Lastly, Peek Jump can happen around `Status` or `Outcome`, where a user wants to see intermediate or final outcomes.

Our taxonomy can further support video navigation by segmenting a video into meaningful sections, by leveraging `Subgoal`, `Status`, or `Bridge` information. P9 said, *"If we have the Goal and Subgoal information, I think the video can be divided by each section like a table of contents. I would have liked it."* P8 mentioned the possibility of using `Status`. She said, *"If Subgoal remarks the start of a step, I thought Status remarks the end of a step. It showed intermediate outcomes."* One can also leverage `Bridge` as it may signal transition to next chapter. As such, we can leverage meaningful information types to make navigation easier.

The taxonomy can also be useful when summarizing a video. As observed from our user study, users could choose the relevant information such as `Goal`, `Tool`, or `Instruction` to summarize the main points. They can also see a succinct summary explained by the author with `Briefing` or `Reflection` or an outline of how the goal is achieved with `Subgoal`. We can also make the summary generation process interactive by allowing the users to choose the information type that they want to see in a summary. In this way, we can give users more control over the summarization process beyond the time budget [25].

Lastly, our taxonomy can help users make an informed decision when selecting videos to watch. Users can use certain information types to assist their decision. P3 said, *"I would check Overview, Tool, and Conclusion first when deciding on whether to watch the video or not. I would check Overview and Conclusion to see if I like the method and outcome, and I would check Tool to see if I have all the required tools."* They can also see the proportion of information types to make a decision. P8 said, *"I don't really like videos that have a lot of irrelevant information. I would filter out videos that have a high portion of Miscellaneous information."* The taxonomy can also be used to recommend videos, providing explanations of recommendations such as conciseness or required tools. As in Inel et al.'s work which provides explanations of a video summary [24], it will help users understand the video with transparency.

Different users can rely on different information types based on their navigational or learning needs. With our taxonomy, we believe that users will have more control and agency in navigating, summarizing, and selecting videos with more informed decisions.

9.2.3 Analysis. Our taxonomy can provide a systematic way to help instructors reflect on their videos by analyzing content, viewership, and watching patterns. Receiving feedback on a video is key for authors in improving their videos [44]. Researchers have proposed several systems for providing feedback on videos, such as a script-based review system [44] or a system that analyzes accessible factors of a video [32, 46]. By applying our taxonomy to their videos, the author can see how focused the video is (e.g., Do I have too many `Side Notes`?) or how structured the video is (e.g., Do I mention enough `Subgoals`?). It can also give feedback on its accessibility, by looking at how descriptive the video is (e.g., Are there an adequate number of `Descriptions`?) [32]. Authors can also see which information type received more attention from viewers, and make informed decisions about the content revision and production.

The taxonomy can also enable comparison between multiple videos. With an increasing number of videos, many systems have been proposed to enable the exploration and analysis of large collections of videos [14, 19, 33]. However, one of the challenges in comparing videos is the complexity of the size and items to be compared. Tharatipyakul et al. proposed video abstraction as a way to reduce such complexity [52]. Our taxonomy enables abstracting a video such as by taking `Instructions`, thereby enabling efficient comparison between videos. It will allow identifying commonalities and differences in approaches toward the same goal [5, 7] or classify workflows at scale [56].

9.3 Supporting the Learning Experience

Understanding the information types in videos can help users in organizing the information. Mayer's multimedia learning theory suggests that learning material should have an understandable structure and guide the learner in making a mental model (Active processing principle) [34]. He suggests that it is helpful to know how information models can be structured. We believe that our taxonomy can contribute to structuring information in videos by organizing the information based on their kind, and thereby help the learning process of users.

Our taxonomy also includes information types that are critical to effective instructional content. According to Morain and Swarts [38], successful tutorial videos begin with an overview of what is to be accomplished (`Goal`, `Briefing`), explain what is accomplished (`Subgoal`) and reasons for performing a step (`Justification`), and describe details such as the tool selection (`Tool`), the settings (`Context`), and the outcomes (`Outcome`). Identifying meaningful information types for learners can ultimately extend their learning experiences beyond following along.

Furthermore, our taxonomy shares several components with the taxonomy of information types in lecture videos. Although how-to videos and lecture videos differ in the type of knowledge they convey (e.g. procedural vs. declarative), they share the commonality of conveying instructional information. Comparing our taxonomy to Espino's investigation on the taxonomy of verbal information in MOOC videos, there are several common components: 'Opening/closing shot' (`Opening`, `Closing`), 'Overview of the contents' (`Briefing`), 'announce following section' (`Subgoal`), and 'Justify/motivate content' (`Justification`, `Motivation`) [16]. We can see that our taxonomy identifies major components that aid learners in their learning process.

9.4 Technical Pipeline

To foster leveraging our taxonomy and developing applications discussed in Section 9.2, it is essential to develop a technical pipeline that classifies segments of a video into the information types of the taxonomy. As one of the approaches, we can leverage the few-shot learning technique on transcripts of a video with large language models such as GPT-3 [4]. However, since our taxonomy is not only based on verbal information but verbal information that considers visual information, multimodal learning that takes visual information into account might yield better accuracy. The hierarchy of our taxonomy (Category and Type) enables Hierarchical Classification as well. We hope our dataset containing 9.9k sentences labeled according to the taxonomy can be served as a useful starting point to build such technical pipelines.

9.5 Limitations and Future Work

In our study, we chose verbal utterances as a primary source of information. This is because how-to videos usually have content creators explaining verbally how to perform a task [12], with an explicit intention of explaining the visual content [36]. They also give additional information that is difficult to be delivered visually. Due to the unique and extensive role of verbal information in how-to videos, we presumed that it would cover a wide range of information and thus chose it as our scope.

However, videos are multimodal and visual information also plays an important role [38]. Although we considered visual information when annotating each sentence to understand context, it does not cover information types that only visuals can convey. For example, visual information can describe instructions in more detail, sometimes accompanied with annotations that describe emphasis on objects or provide more detailed information of a tool used [12]. It would be interesting to investigate videos that deliver information only through a visual channel to understand the capacity of information types that visuals convey. Furthermore, verbal and visual information might not always align with each other [12, 21]. For example, an instructor can verbally share instructions first and then visually demonstrate them later. As such, future work can incorporate visual information in how-to videos for a more comprehensive taxonomy and analysis.

Also, while our taxonomy is based on diverse videos in terms of topics, styles, and production methods, they were YouTube videos whose lengths are between 5 minutes and 15 minutes. It may be that some types in the taxonomy are specific to YouTube videos (e.g., `Self-promotion`), and longer videos (e.g., live streams) or shorter videos (e.g., TikTok videos [53]) may have introduced additional types of information. Further research should explore a wider range of how-to videos, which could build upon our taxonomy.

10 CONCLUSION

We present a taxonomy of information types in How-to videos. Our taxonomy identifies 21 types of information under 8 categories: *Greeting*, *Overview*, *Method*, *Supplementary*, *Explanation*, *Description*, *Conclusion*, and *Miscellaneous*. We demonstrate the utility of the taxonomy in both analyzing users' navigational behavior and supporting their navigation in how-to videos. We first show how our taxonomy can serve as an analytical framework for understanding existing video navigation systems. Then, we further investigate how the information type can assist people watching how-to videos. An explorative user study with nine participants showed that type-based navigation enabled participants to find specific information and perform tasks effectively. We further discuss how the taxonomy enables multiple applications in video authoring, viewing, and analysis. Finally, we release a dataset, HTM-Type, which contains 120 videos containing 9.9k sentences with each sentence labeled according to the taxonomy. We hope that our work builds a foundation for understanding how-to videos in a more systematic way.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2020R1C1C1007587) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-01347, Video Interaction Technologies Using Object-Oriented Video Modeling).

REFERENCES

[1] D.G. Altman. 1990. Practical Statistics for Medical Research. (1990). https://doi.org/10.1201/9780429258589

[2] Microsoft Azure. 2022 (accessed Sep 14, 2022). *Speech to text.* https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text

[3] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for Placing Cuts and Transitions in Interview Video. *ACM Transactions on Graphics* 31, 4, Article 67 (jul 2012), 8 pages. https://doi.org/10.1145/2185520.2185563

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[5] Minsuk Chang, Leonore 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174025

[6] Minsuk Chang, Mina Huh, and Juho Kim. 2021. RubySlippers: Supporting Content-Based Voice Navigation for How-to Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 97, 14 pages. https://doi.org/10.1145/3411764.3445131

[7] 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 *Proceedings of Graphics Interface 2020* (University of Toronto) *(GI 2020)*. Canadian Human-Computer Communications Society / Société canadienne du dialogue humain-machie, 114 – 124. https://doi.org/10.20380/GI2020.13

[8] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to Design Voice Based Navigation for How-To Videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300931

[9] Hung-Tao Chen and Megan Thomas. 2020. Effects of lecture video styles on engagement and learning. *Educational Technology Research and Development* 68 (03 2020), 2147–2164. https://doi.org/10.1007/s11423-020-09757-6

[10] Peggy Chi, Nathan Frey, Katrina Panovich, and Irfan Essa. 2021. Automatic Instructional Video Creation from a Markdown-Formatted Tutorial. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 677–690. https://doi.org/10.1145/3472749.3474778

[11] Pei-Yu Chi, Sally Ahn, Amanda Ren, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2012. MixT: Automatic Generation of Step-by-Step Mixed Media Tutorials. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2380116.2380130

[12] Pei-Yu Chi, Joyce Liu, Jason Linder, Mira Dontcheva, Wilmot Li, and Bjoern Hartmann. 2013. DemoCut: Generating Concise Instructional Videos for Physical Demonstrations. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/2501988.2502052

[13] Konstantinos Chorianopoulos. 2018. A Taxonomy of Asynchronous Instructional Video Styles. *The International Review of Research in Open and Distributed Learning* 19, 1 (Feb. 2018). https://doi.org/10.19173/irrodl.v19i1.2920

[14] Maureen Daum, Enhao Zhang, Dong He, Magdalena Balazinska, Brandon Haynes, Ranjay Krishna, Apryle Craig, and Aaron Wirsingn. 2022. VOCAL: Video Organization and Interactive Compositional AnaLytics. In *The Conference on Innovative Data Systems Research (CIDR)*.

[15] Descript. 2022 (accessed Sep 6, 2022). Descript. https://www.descript.com/

[16] José Miguel Santos Espino. 2019. Anatomy of instructional videos: a systematic characterization of the structure of academic instructional videos.

[17] Logan Fiorella and Richard E. Mayer. 2018. What Works and Doesn't Work with Instructional Video. *Comput. Hum. Behav.* 89, C (dec 2018), 465–470. https://doi.org/10.1016/j.chb.2018.07.015

[18] C. Ailie Fraser, Joy O. Kim, Hijung Valentina Shin, Joel Brandt, and Mira Dontcheva. 2020. Temporal Segmentation of Creative Live Streams. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376437

[19] Daniel Fu, Will Crichton, James Hong, Xinwei Yao, Haotian Zhang, Anh Truong, Avanika Narayan, Maneesh Agrawala, Christopher Ré, and Kayvon Fatahalian. 2019. Rekall: Specifying Video Events using Compositions of Spatiotemporal Labels.

[20] Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference* (Atlanta, Georgia, USA) *(L@S '14)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/2556325.2566239

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. 2022. Temporal Alignment Network for long-term Video. In *CVPR*.

[22] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J. Mysore. 2019. B-Script: Transcript-Based B-Roll Video Editing with Recommendations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300311

[23] Christina Ilioudi, Michail Giannakos, and Konstantinos Chorianopoulos. 2013. Investigating Differences among the Commonly Used Video Lecture Styles. https://doi.org/10.13140/2.1.3524.9284

[24] Oana Inel, Nava Tintarev, and Lora Aroyo. 2020. Eliciting User Preferences for Personalized Explanations for Video Summaries. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 98–106. https://doi.org/10.1145/3340631.3394862

[25] Haojian Jin, Yale Song, and Koji Yatani. 2017. ElasticPlay: Interactive Video Summarization with Dynamic Time Budgets. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 1164–1172. https://doi.org/10.1145/3123266.3123393

[26] Kimia Kiani, George Cui, Andrea Bunt, Joanna McGrenere, and Parmit K. Chilana. 2019. Beyond "One-Size-Fits-All": Understanding the Diversity in How Software Newcomers Discover and Make Use of Help Resources. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300570

[27] Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-Driven Interaction Techniques for Improving Navigation of Educational Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 563–572. https://doi.org/10.1145/2642918.2647389

[28] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing Step-by-Step Information Extraction to Enhance Existing How-to Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4017–4026. https://doi.org/10.1145/2556288.2556986

[29] Yoonjoo Lee, John Joon Young Chung, Tae Soo Kim, Jean Y Song, and Juho Kim. 2022. Promptiverse: Scalable Generation of Scaffolding Prompts Through Human-AI Hybrid Knowledge Graph Annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 96, 18 pages. https://doi.org/10.1145/3491102.3502087

[30] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 17, 12 pages. https://doi.org/10.1145/3526113.3545702

[31] Ching Liu, Juho Kim, and Hao-Chuan Wang. 2018. ConceptScape: Collaborative Concept Mapping for Video Learning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173961

[32] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 272, 14 pages. https://doi.org/10.1145/3411764.3445233

[33] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2014. Video Lens: Rapid Playback and Exploration of Large Video Collections and Associated Metadata. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 541–550. https://doi.org/10.1145/2642918.2647366

[34] Richard E. Mayer. 2005. *Cognitive Theory of Multimedia Learning.* Cambridge University Press, 31–48. https://doi.org/10.1017/CBO9780511816819.004

[35] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 9876–9886.

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

[37] Toni-Jan Keith Palma Monserrat, Shengdong Zhao, Kevin McGee, and Anshul Vikram Pandey. 2013. NoteVideo: Facilitating Navigation of Blackboard-Style Lecture Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1139–1148. https://doi.org/10.1145/2470654.2466147

[38] Matt Morain and Jason Swarts. 2012. YouTutorial: A Framework for Assessing Instructional Online Video. *Technical Communication Quarterly* 21, 1 (2012), 6–24. https://doi.org/10.1080/10572252.2012.626690 arXiv:https://doi.org/10.1080/10572252.2012.626690

[39] Megha Nawhal, Jacqueline B. Lang, Greg Mori, and Parmit K. Chilana. 2019. VideoWhiz: Non-Linear Interactive Overviews for Recipe Videos. In *Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019* (Kingston, Canada) *(GI'19)*. Canadian Human-Computer Communications Society, Waterloo, CAN, Article 15, 8 pages. https://doi.org/10.20380/GI2019.15

[40] Robert Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A Method for Taxonomy Development and its Application in Information Systems. *European Journal of Information Systems* 22 (05 2013). https://doi.org/10.1057/ejis.2012.26

[41] Daulet Nurmanbetov. 2021. *BERT-restore-punctuation model from huggingface*. https://huggingface.co/felflare/bert-restore-punctuation

[42] Ozlem Ozan and Yasin Ozarslan. 2016. Video lecture watching behaviors of learners in online courses. *Educational Media International* 53, 1 (2016), 27–41. https://doi.org/10.1080/09523987.2016.1189255 arXiv:https://doi.org/10.1080/09523987.2016.1189255

[43] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/2807442.2807502

[44] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2016. VidCrit: Video-Based Asynchronous Video Review. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 517–528. https://doi.org/10.1145/2984511.2984552

[45] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video Digests: A Browsable, Skimmable Format for Informational Lecture Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 573–582. https://doi.org/10.1145/2642918.2647400

[46] Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. 2021. Say It All: Feedback for Improving Non-Visual Presentation Accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 276, 12 pages. https://doi.org/10.1145/3411764.3445572

[47] Suporn Pongnumkul, Mira Dontcheva, Wilmot Li, Jue Wang, Lubomir Bourdev, Shai Avidan, and Michael F. Cohen. 2011. Pause-and-Play: Automatically Linking Screencast Video Tutorials with Applications. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/2047196.2047213

[48] Luca Ponzanelli, Gabriele Bavota, Andrea Mocci, Rocco Oliveto, Massimiliano Di Penta, Sonia Haiduc, Barbara Russo, and Michele Lanza. 2019. Automatic Identification and Classification of Software Development Video Tutorial Fragments. *IEEE Transactions on Software Engineering* 45, 5 (2019), 464–488. https://doi.org/10.1109/TSE.2017.2779479

[49] José Miguel Santos Espino, M.D. Afonso-Suárez, and Cayetano Guerra Artal. 2016. Speakers and boards: A survey of instructional video styles in MOOCs. 63 (05 2016), 101–115.

[50] Alessandra Semeraro and Laia Turmo Vidal. 2022. Visualizing Instructions for Physical Training: Exploring Visual Cues to Support Movement Learning from Instructional Videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 90, 16 pages. https://doi.org/10.1145/3491102.3517735

[51] William Sugar, Abbie Brown, and Kenneth Luterbach. 2010. Examining the anatomy of a screencast: Uncovering common elements and instructional strategies. *The International Review of Research in Open and Distributed Learning* 11, 3 (October 2010), 1–20. https://www.learntechlib.org/p/49134

[52] Atima Tharatipyakul and Hyowon Lee. 2018. Towards a Better Video Comparison: Comparison as a Way of Browsing the Video Contents. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (Melbourne, Australia) *(OzCHI '18)*. Association for Computing Machinery, New York, NY, USA, 349–353. https://doi.org/10.1145/3292147.3292183

[53] TikTok. 2022 (accessed Sep 14, 2022). TikTok. https://www.tiktok.com

[54] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. QuickCut: An Interactive Tool for Editing Narrated Video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 497–507. https://doi.org/10.1145/2984511.2984569

[55] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 108, 16 pages. https://doi.org/10.1145/3411764.3445721

[56] Xu Wang, Benjamin Lafreniere, and Tovi Grossman. 2018. Leveraging Community-Generated Videos and Command Logs to Classify and Recommend Software Workflows. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173859

[57] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. 2015. Learnersourcing Subgoal Labels for How-to Videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 405–416. https://doi.org/10.1145/2675133.2675219

[58] wikihow. 2022 (accessed Sep 14, 2022). *Wikihow*. https://www.wikihow.com/

[59] Kuldeep Yadav, Ankit Gandhi, Arijit Biswas, Kundan Shrivastava, Saurabh Srivastava, and Om Deshmukh. 2016. ViZig: Anchor Points Based Non-Linear Navigation and Summarization in Educational Videos. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 407–418. https://doi.org/10.1145/2856767.2856788

[60] Saelyne Yang, Sangkyung Kwak, Tae Soo Kim, and Juho Kim. 2022. Improving Video Interfaces by Presenting Informational Units of Videos. In *CHI'22 Extended Abstracts*. Association for Computing Machinery.

[61] Saelyne Yang, Jisu Yim, Aitolkyn Baigutanova, Seoyoung Kim, Minsuk Chang, and Juho Kim. 2022. SoftVideo: Improving the Learning Experience of Software Tutorial Videos with Collective Interaction Data. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 646–660. https://doi.org/10.1145/3490099.3511106

[62] Saelyne Yang, Jisu Yim, Juho Kim, and Hijung Valentina Shin. 2022. CatchLive: Real-Time Summarization of Live Streams with Stream Content and Interaction Data. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 500, 20 pages. https://doi.org/10.1145/3491102.3517461

[63] YouTube. 2022 (accessed Sep 14, 2022). Video Chapters. https://support.google.com/youtube/answer/9884579

[64] youtube dl. 2022 (accessed Sep 14, 2022). *youtube-dl*. https://www.npmjs.com/package/youtube-dl

A ANALYSIS RESULT OF HTM-TYPE

A.1 Information Distribution Statistics

	Greeting	Overview	Method	Supplementary	Explanation	Description	Conclusion	Misc.
Mean (%)	2.4	6.4	47.5	3.4	4.4	19.6	6.3	10.0
SD (%)	1.6	6.0	16.9	4.5	3.4	12.5	6.6	7.8
Min (%)	0.19	0.5	15.6	0.8	0.5	1.9	0.4	0.3
Max (%)	8.4	34.4	88.1	22.9	14.5	61.8	30.9	37.5

Table 4: The mean, standard deviation, and minimum and maximum proportion for each category distribution in a video. Videos that did not contain the category were excluded from the minimum value calculation for the corresponding category.

	Opening	Closing	Goal	Motivation	Briefing	Subgoal	Instruction
Mean (%)	1.0	1.4	2.5	3.2	0.7	2.7	39.8
SD (%)	0.8	1.1	1.8	4.8	2.0	3.0	17.7
Min (%)	0.1	0.2	0.3	0.7	1.9	0.3	1.8
Max (%)	4.4	4.5	8.8	29.7	10.9	21.9	82.5
	Tool	Tip	Warning	Justification	Effect	Status	Context
Mean (%)	5.0	1.8	1.5	2.8	1.7	6.5	7.8
SD (%)	5.6	3.3	3.1	2.7	2.0	6.0	9.5
Min (%)	0.4	0.7	0.8	0.5	0.3	0.2	0.4
Max (%)	17.8	20.5	16.4	11.6	10.5	28.8	56.3
	Tool Spec.	Outcome	Reflection	Side Note	Self-promo	Bridge	Filler
Mean (%)	5.3	2.8	3.6	6.6	2.0	1.4	0.1
SD (%)	5.6	4.1	5.0	7.3	2.4	1.5	0.3
Min (%)	0.5	0.3	0.8	0.6	0.5	0.1	0.1
Max (%)	27.8	30.9	26.5	34.6	14.2	8.3	2.1

Table 5: The mean, standard deviation, and minimum and maximum proportion for each type distribution in a video. Videos that did not contain the type were excluded from the minimum value calculation for the corresponding type.

A.2 Information Distribution Based on Task Type

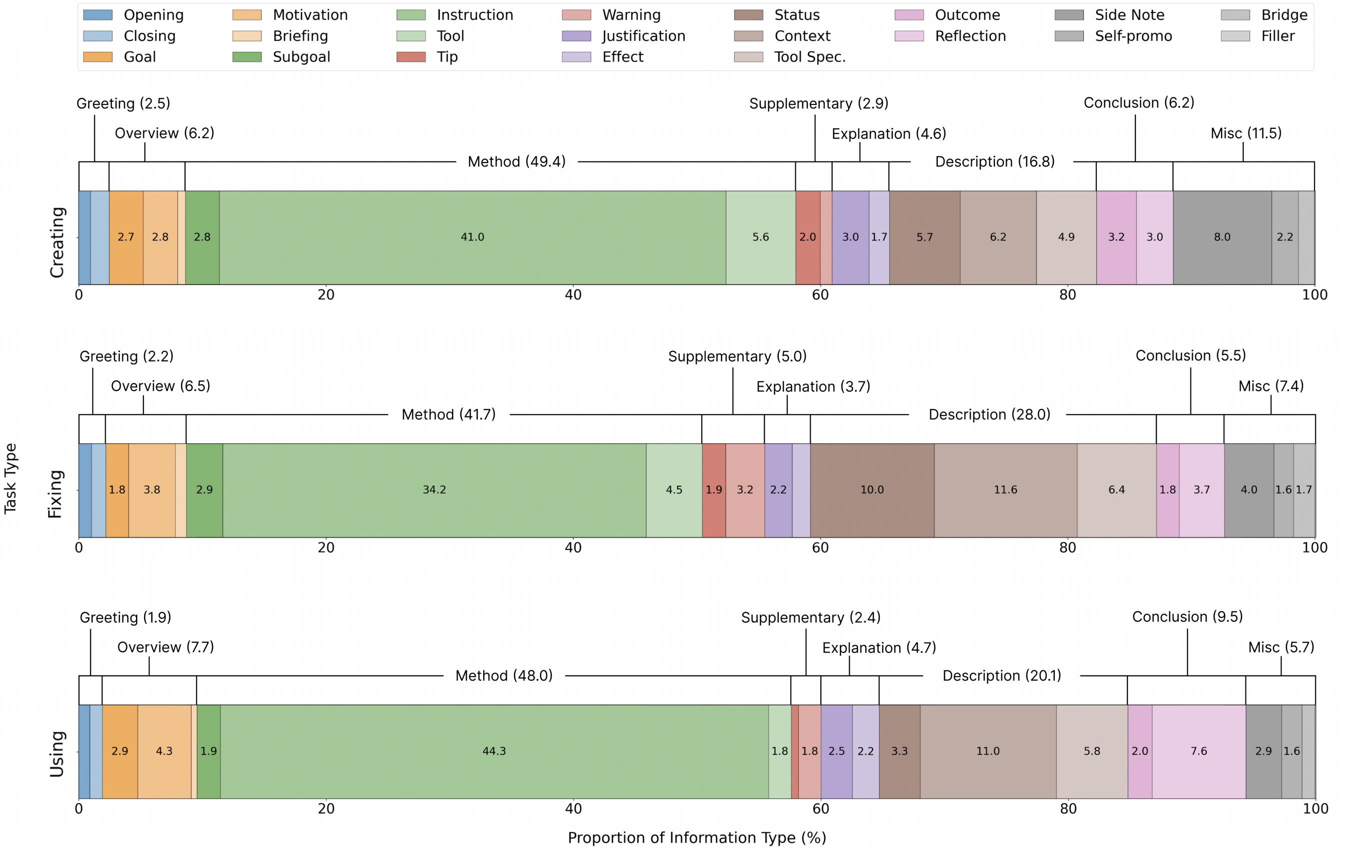


Figure 6: Distribution of Categories and Types by Task Type. Categories are denoted above the types using group brackets. Only proportions greater than 1.5% are written in text. The graph contains the distribution for Creating, Fixing, and Using, respectively. The Kruskal-Wallis test (post-hoc Dunn's test, p<0.05) showed significant differences between the task types for Creating-Fixing and Using-Fixing pairs for `Status` and Creating-Fixing for `Context`.

A.3 Information Distribution Based on Narration Style

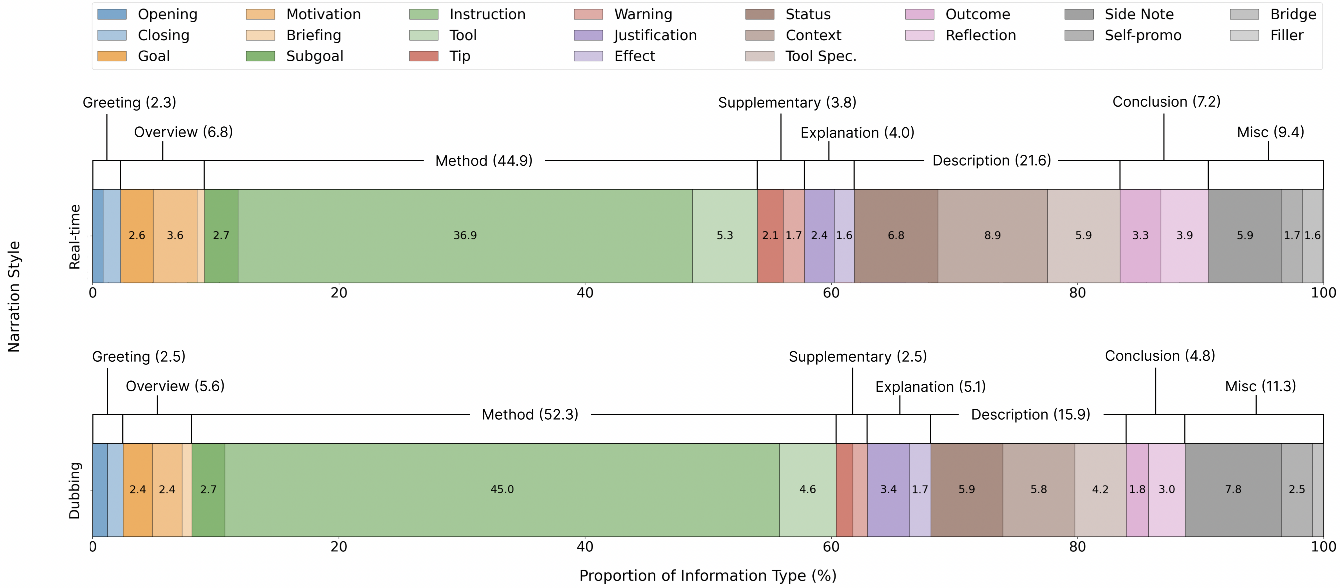


Figure 7: Distribution of Categories and Types by Narration Style. Categories are denoted above the types using group brackets. Only proportions greater than 1.5% are written in text. The graph contains the distribution for Real-time and Dubbed videos, respectively. The Kruskal-Wallis test (*p*<0.05) showed significant differences between the narration styles for `Instruction` and `Tool Specification`.

A.4 Information Distribution Based on Time

	Greeting	Overview	Method	Supplementary	Explanation	Description	Conclusion	Misc.
5th quantile (sec)	4	12	111	169	168	85	720	48
95th quantile (sec)	994	238	854	863	872	869	967	980
Mean (sec)	567	92	472	579	494	457	866	625
SD (sec)	458	96	233	223	224	248	85	332

Table 6: 5th quantile, 95th quantile, mean and standard deviations of time distribution for each category. The time has been normalized to 1000 seconds. The middle 90% of the category data lies between the 5th and 95th quantile.

A.5 Construction Dataset Validation

To verify that the dataset used to construct the taxonomy is representative, we compare the construction dataset (48 videos) and the entire dataset (120 videos) to ensure that the distributions of the videos are similar. First, we compare the video length and genre. For video length, the average video duration is 7 min 3 sec, SD=1 min 35 sec for the construction set and 7 min 8 sec, SD=1 min 23 sec for the entire dataset. For genre, both datasets contain an equal number of videos for each of the 12 genres provided by HowTo100M (4 and 10, respectively). Also, we confirm that the distributions of the video according to the task type and narration style are analogous as well. For task type, the proportions of Creating, Fixing, and Using videos are 70.8%, 18.8%, and 10.4%, respectively, for the construction dataset. For the entire dataset, the ratios are 68.3%, 22.5%, and 9.2%, respectively. For narration style, the real-time and dubbed videos were 62.5% and 37.5% for the construction set, and 65.0% and 35.0% for the entire dataset. We also analyze if the information type distribution differs between the two datasets (Figure 8). Comparing the two distributions reveals less than a 1.1% difference between the two datasets for each type and category.

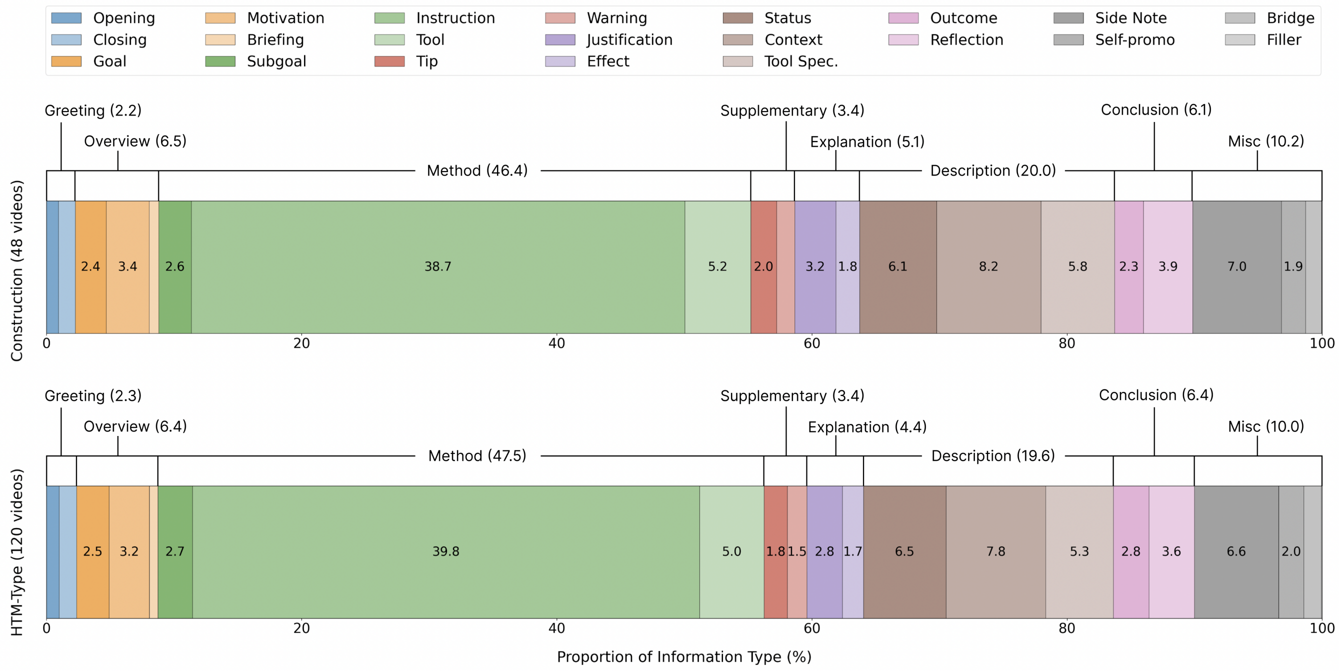


Figure 8: Distribution of Categories and Types for construction and HTM-Type datasets. Categories are denoted above the types using group brackets. Only proportions greater than 1.5% are written in text. The differences between the corresponding types are less than 1.1% for all types, showing a similar distribution across both datasets.

B QUESTIONS ASKED IN THE SEARCH TASK

Video	Question	Corresponding Type
Cooking	What are some things you should be aware of when putting the egg in the water?	Warning
	What were the ingredients used to season the egg?	Tool
	How long should you boil the egg?	Instruction
Slime	How is this recipe different from other recipes?	Goal
	In order to make the slime, in what order are activator, glue, and food coloring put into the mixture?	Instruction
	What should not go in the slime?	Warning
Illustrator	To make the image more cartoonish, which feature do you need to adjust?	Instruction
	After you adjust all the features, you click the Expand button. What does it do?	Effect
	Where did the author get the image they used?	Tool Specification

Table 7: Questions asked in the Search task and corresponding types that are related to each question.