GUIDE: A Guideline-Guided Dataset for Instructional Video Comprehension

Jiafeng Liang[1], **Shixin Jiang**[1], **Zekun Wang**[1], **Haojie Pan**[3], **Zerui Chen**[1], **Zheng Chu**[1], **Ming Liu**[1,2*], **Ruiji Fu**[3], **Zhongyuan Wang**[3] and **Bing Qin**[1,2]

[1]Harbin Institute of Technology, Harbin, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Kuaishou Technology, Beijing, China
{jfliang, sxjiang, zkwang, zrchen, zchu, mliu, qinb}@ir.hit.edu.cn

Abstract

There are substantial instructional videos on the Internet, which provide us tutorials for completing various tasks. Existing instructional video datasets only focus on specific steps at the video level, lacking experiential guidelines at the task level, which can lead to beginners struggling to learn new tasks due to the lack of relevant experience. Moreover, the specific steps without guidelines are trivial and unsystematic, making it difficult to provide a clear tutorial. To address these problems, we present the **GUIDE** (**G**uideline-G**uide**d) dataset, which contains 3.5K videos of 560 instructional tasks in 8 domains related to our daily life. Specifically, we annotate each instructional task with a guideline, representing a common pattern shared by all task-related videos. On this basis, we annotate systematic specific steps, including their associated guideline steps, specific step descriptions and timestamps. Our proposed benchmark consists of three sub-tasks to evaluate comprehension ability of models: (1) Step Captioning: models have to generate captions for specific steps from videos. (2) Guideline Summarization: models have to mine the common pattern in task-related videos and summarize a guideline from them. (3) Guideline-Guided Captioning: models have to generate captions for specific steps under the guide of guideline. We evaluate plenty of foundation models with GUIDE and perform in-depth analysis. Given the diversity and practicality of GUIDE, we believe that it can be used as a better benchmark for instructional video comprehension.

1 Introduction

Instructional videos guide learners how to accomplish multi-step tasks such as cooking, making up and embroidering, repairing, or creating new objects. Recently, numerous instructional video datasets have been proposed [Zala *et al.*, 2023; Tang *et al.*, 2019; Zhou *et al.*, 2018; Zhukov *et al.*, 2019]. As shown in Figure 1 (a), these datasets solely focus on fine-grained annotations, leading to trivial and unsystematic step captions, making it difficult to provide clear tutorial guidance. Moreover, while many instructional videos pertain to the same task, there are significant differences in the details and sequence of their steps, which increases the difficulty for beginners to learn. If there exists a model capable of analyzing various videos of the same task and organizing steps into a hierarchical structured tutorial, it will accelerate learning progression for beginners.

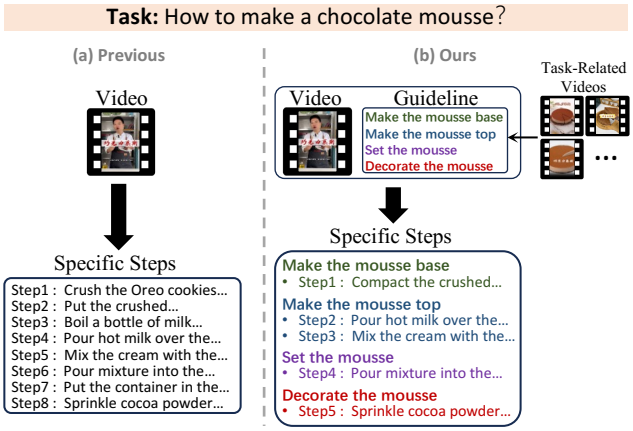

Figure 1: The steps in the previous dataset were very trivial and unsystematic, making it difficult for beginners to learn. In contrast, our dataset provides structured guideline-guided steps. Such guideline is a common pattern shared by all videos related to the same task.

Our inspiration comes from two aspects. First, according to educational psychology [Bolkan *et al.*, 2020], learners are always confused when learning an unfamiliar and challenging task because they lack relevant experience. Although previous datasets provide specific steps, their complexity would exceed learners' cognitive load. Thus, a clear guideline can help learners understand the task more efficiently. Second, planning a procedure from an instructional video is to complete a specific guideline (*i.e.*, a procedure matches a clear intention). Since a guideline usually involves specific steps, they can be used to support the procedure generating (shown in Figure 1 (b)).

To support these, we introduce **GUIDE**, a **G**uideline-G**uide**d dataset for instructional video comprehension. We propose a three-stage dataset construction pipeline on instructional video

* Corresponding Author.

Figure 2: Overview of the GUIDE dataset. The GUIDE consists of 560 task queries, each containing an average of 6.2 task-related videos. These instructional videos are divided into specific steps with timestamps and text descriptions (yellow area). Additionally, each task contains a set of guideline steps representing a common pattern shared by all task-related videos (purple area).

from video platform[1], collecting high-quality annotations. The GUIDE contains three annotations (shown in Figure 2): (1) 560 queries: each query represents an instructional task and contains an average of 6.2 task-related videos (total of 3.5K videos), (2) 15K step segments: each video is divided into an average of 4.3 specific step segments with corresponding timestamps and text description, and, (3) 560 guidelines: each instructional task contains a set of guideline steps that represent a common pattern of the task. Moreover, each specific step has its corresponding guideline step.

In GUIDE, we propose three challenging sub-tasks for instructional video analysis. (1) Step Captioning: models have to generate captions for specific steps from videos. (2) Guideline Summarization: models have to mine the common pattern in task-related videos and summarize a guideline from them. (3) Guideline-Guided Captioning: models have to generate captions for specific steps under the guide of guidelines. We benchmark various video foundation models and language foundation models (utilize video transcription instead of visual information), including VideoChat [Li *et al.*, 2023b], Video-LLaMA [Zhang *et al.*, 2023], mPLUG-Owl [Ye *et al.*, 2023], GPT-3.5-turbo [OpenAI, 2022], GPT-4 [OpenAI, 2023], Vicuna [Chiang *et al.*, 2023] and Flan-T5 [Chung *et al.*, 2022]. We also evaluate the performance of humans on the GUIDE for a better comparison.

The experimental results demonstrate that both video and language foundation models are struggle in all three sub-tasks. We initially observed that the specific steps generated under the ground-truth guideline guide are clearer and more accurate, indicating that the accurate guideline is helpful for generating instructional steps. Then, we explore the source of the ability to mine the guideline from multiple videos with different training settings. The results show that the single-video understanding ability is the basis of learning multiple videos, indicating that more pre-training and fine-tuning data is necessary. Subsequently, we investigate the bottleneck of video foundation models. The model demonstrates significant performance degradation compared to their text-only counterparts, indicating that more specialized visual encoders and visual-language bridges are needed to represent temporal procedures better. Finally, we perform a human evaluation demonstrating our dataset's promising applications in real-world scenarios. To summarize our contributions:

- We introduce a novel guideline-guided instructional video comprehension dataset GUIDE, containing task-level guideline annotations and video-level systematic specific step annotations.
- We design three challenging sub-tasks in GUIDE for instructional video analysis, namely step captioning, guideline summarization and guideline-guided captioning.
- We benchmark various foundation models and conduct extensive analyses to provide detailed insights.

2 Related Work

2.1 Instructional Video Comprehension

Understanding instructional videos presents a significant challenge, primarily due to their inherent procedural temporal structure. Recently, many researchers have studied the analysis

[1]Chinese short video platform: Kuaishou.

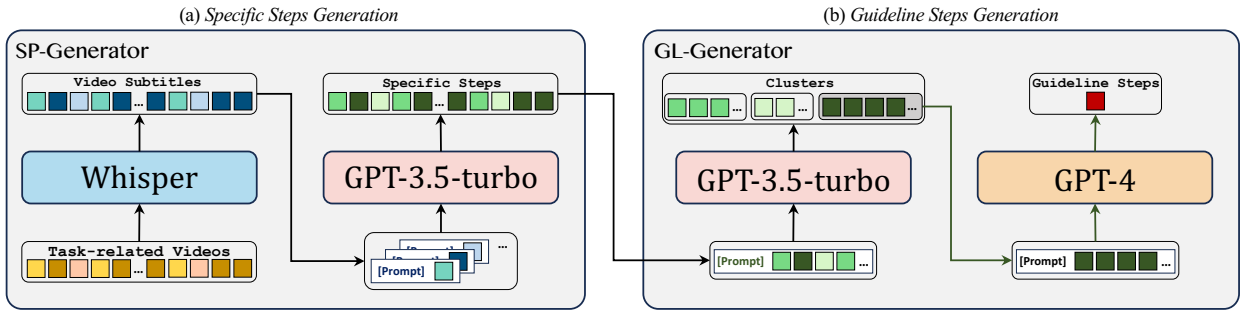

Figure 3: Overview of Automatic Annotation. (a) Transcribing the video into textual subtitles and generating specific steps based on subtitles. (b) Clustering the task-related videos and generating a set of guideline steps for the cluster with the highest number of videos.

of instructional videos [Dvornik *et al.*, 2023; Yang *et al.*, 2023; Huang *et al.*, 2018]. For instance, Yang *et al.* [2023] propose an end-to-end framework that augments a language model to predict timestamps and descriptions of steps seamlessly. Gu *et al.* [2023] propose a two-stream transformer, which constructs a video scene graph [Liang *et al.*, 2023] for video captioning by retrieving additional knowledge. However, the steps generated by these methods are trivial and unsystematic due to the lack of guidelines, making it difficult for people to learn. While some approaches [Han *et al.*, 2020; He *et al.*, 2023] extract the guideline by analyzing correlations between instructional videos, they do not explore the help of the guideline for generating step captions.

2.2 Instructional Video Datasets

Existing instructional video datasets can be categorized into two types: action detection datasets [Zhukov *et al.*, 2019; Tang *et al.*, 2019; Kuehne *et al.*, 2014] and step caption datasets [Zhou *et al.*, 2018; Zala *et al.*, 2023; Damen *et al.*, 2018]. The former are predominantly employed for video segmentation and action recognition tasks, while the latter are used for video segmentation and step captioning tasks. For instance, COIN [Tang *et al.*, 2019] predefines many actions and assigns these actions to instructional videos to describe procedural processes. HIREST [Zala *et al.*, 2023] segments each video based on instructional steps and manually annotates captions for the steps. However, these datasets primarily focus on fine-grained annotations, leading to trivial and unsystematic step captions, making it difficult to provide clear tutorial guidance. Furthermore, many instructional videos related to the same tasks exhibit significant differences in specific procedures, increasing the difficulty for beginners to learn. In this paper, Our GUIDE dataset provides guideline annotations, representing the common pattern across multiple task-related videos. In addition, we annotate guideline-guided specific steps to improve the systematic nature of the data, which reduces the learning difficulty.

3 Dataset

3.1 Overview

In this section, we introduce our instructional video comprehension dataset, **GUIDE**. Initially, we describe the three-stage dataset construction pipeline. Subsequently, we provide an overview of the data statistics. Lastly, we introduce the three novel sub-tasks we proposed to comprehensively evaluate foundation models based on our dataset.

3.2 Dataset Construction Pipeline

GUIDE dataset construction pipeline contains three stages: video collection, automatic annotation, and manual annotation. In the Appendix A.1, we provide more details for each stage.

Video Collection In this stage, we aim to collect a large number of high-quality instructional videos. To ensure the widely-used of the GUIDE, we collect videos from 560 different instructional tasks across the 8 most common domains in our daily life. We require annotators to collect videos containing explicit instructional steps and clearly defined time boundaries between these steps. To further enhance the practicality of the dataset, we also require the collected videos that include detailed video subtitles, *i.e.*, each step accompanied by corresponding voice explanations. More details are provided in Appendix A.1.

Automatic Annotation As illustrated in Figure 3, the automatic annotation framework contains two stages: **Specific Steps Generation** and **Guideline Steps Generation**.

In the specific steps generation, we utilize the SP-GENERATOR module, comprising Whisper [Radford *et al.*, 2023] and GPT-3.5-turbo [OpenAI, 2022], to automatically generate the specific steps for each video. Given an instructional task query Q, which contains n related videos $\{v_1, v_2, ..., v_n\}$. We first use Whisper to generate video subtitles as $A_Q = \{a_1, a_2, ..., a_n\}$. Subsequently, we feed these subtitles along with their occurrence time and our carefully crafted prompt to GPT-3.5-turbo, enabling it to generate specific steps s and corresponding timestamps t for each video. Our SP-GENERATOR can be formulated as:

$$SPsteps = \text{SP-GENERATOR}(prompt, A_Q) \qquad (1)$$

where $SPsteps = \{[s,t]_1, [s,t]_2, ..., [s,t]_n\}$. More details are provided in Appendix A.1.

In the guideline steps generation, we aim to extract a set of guideline steps for each task. Actually, extracting a shared guideline from all task-related videos is challenging due to the coarse granularity of task queries in the video database. Specifically, despite many videos sharing the same query, they exhibit significant variations in specific content. For instance,

Dataset	Duration	Step caption	Guideline-Guided	#Videos / # Steps	# Words per Caption	# Steps per Video	# Guideline Steps per Task
COIN [2019]	477h	Predefined	✘	11.8K / 46K	4.8	3.9	-
CrossTask [2019]	376h	Predefined	✘	4.7K / 21K	2.4	7.4	-
YouCook2 [2018]	176h	Manually written	✘	2K / 14K	8.8	7.7	-
HIREST [2023]	476h	Manually written	✘	1.1K / 8.6K	4.4	7.6	-
GUIDE (Ours)	101h	Manually written (SP) + Predefined (GL)	✔	3.5K / 15K	6.5	4.3	3.7

Table 1: Comparison of GUIDE and other datasets with step annotations. Our dataset annotates the common pattern (Guideline Steps) across task-related videos. Moreover, GUIDE is the largest manually written caption dataset. 'SP' Specific Step and 'GL' denotes Guideline Step.

numerous videos fall under the task query '*Making Crayfish*', but variations in ingredients and procedures lead to diverse methods, such as '*Spicy and Numbing Crayfish*', '*Fragrant and Spicy Crayfish*' and '*Garlic Crayfish*'. We try to involve annotators in clustering videos based on the video content during the video collection stage, with the objective of identifying a single cluster that best represents the task query. However, manually clustering a large number of videos is challenging, and there are significant differences in subjective interpretations among annotators, making it difficult to establish clear clustering rules.

Hence, we utilize the GL-GENERATOR module, comprising GPT-3.5-turbo [OpenAI, 2022] and GPT-4 [OpenAI, 2023], to cluster the task-related videos and generate corresponding guideline steps automatically. We feed $SPstep$ and crafted prompts into GPT-3.5-turbo to cluster the videos based on the content of the specific steps and their sequential order. Then, we retain only the cluster $SPstep^*$ with the maximum number, which includes m videos:

$$SPstep^* = \text{Max}\left(\text{Cluster}\left(SPstep\right)\right) \tag{2}$$

where $SPstep^* = \{[s,t]_1, [s,t]_2, ..., [s,t]_m\}$. Finally, we utilize GPT-4 to generate a set of guideline steps $GLstep$ for the current instructional task, as we find during the testing process that GPT-4 is capable of generating more comprehensive and common guideline steps compared to GPT-3.5-turbo. Our GL-GENERATOR can be formulated as:

$$GLstep = \text{GL-GENERATOR}(prompt, SPstep^*) \tag{3}$$

More details are provided in Appendix A.1.

Manual Annotation The results of automatic annotation cannot be regarded as the final annotations. GPT-3.5-turbo generates timestamps for each specific step based on the video subtitles' occurrence time. However, these timestamps are inaccurate because the steps in the video and the voice explanation may not coincide, and due to the lack of information in the subtitles, the specific steps may not be complete. In addition, despite providing detailed prompts for GPT-4, it still uncontrollably generates overly broad or excessively complex guideline steps. Thus, we employ manual annotation to solve these issues.

Initially, we employ an expert in each domain (*e.g.*, chef, dancer, etc.) to adjust all guideline steps, aiming to achieve consistent granularity across all of them. Then, we require the annotators to refine the specific steps generated by GPT-3.5-turbo and annotate the timestamps of steps by watching videos. It is essential that the refined specific steps contain explicit descriptions of the procedures. Furthermore, each specific step is also required to be annotated with its corresponding guideline step.

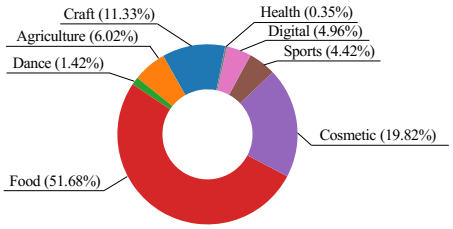

Figure 4: Task category distribution of GUIDE. There are a wide variety of categories for our videos. The most frequent categories are 'Food', 'Cosmetic', and 'Craft'.

3.3 Dataset Analysis

Task Category Distribution GUIDE dataset consists of 560 tasks from 8 common domains in daily life. As shown in Figure 4. The top three most frequent domains are 'Food', 'Cosmetic' and 'Craft'.

Dataset Statistics We collect a total of 3.5K instructional videos containing 560 different common tasks in daily life. The average video duration is 103 seconds, totalling 101 hours. Each task contains an average of 6.2 task-related videos and a predefined guideline shared across all task-related videos, resulting in 560 guidelines. On average, each guideline consists of 3.7 guideline steps, yielding a total of 2.1K guideline steps with an average length of 2.9 words per guideline step. Videos are split into multiple segments based on instructional steps, with an average of 4.3 specific steps per video, totalling 15K specific steps. Each specific step is annotated with a start-end timestamp and a step caption, with an average length of 6.5 words per caption.

Comparisons to Other Datasets Table 1 compares our GUIDE dataset to other instructional video datasets. GUIDE contains numerous open domain instructional tasks videos, each with annotated specific step captions written by annotators. While HIREST [Zala *et al.*, 2023] also provides manually written step captions for open-domain videos, its steps are trivial and unsystematic. In contrast, each instructional task in GUIDE is annotated with a guideline, representing a common pattern shared by all task-related videos. On this basis, we annotate systematic specific steps to provide a clear tutorial. Moreover, among all datasets containing handwritten step caption, GUIDE has the largest scale.

3.4 Task Definition

Step Captioning The step captioning task evaluates the models' capabilities to understand the procedural temporal knowledge of the instructional video. In this task, models have to generate a set of instructional step captions.

Guideline Summarization The guideline summarization task evaluates the models' capabilities to analyze correlations across videos. In this task, models have to mine the common pattern in task-related videos and summarize a guideline from them.

Guideline-Guided Captioning To explore the impact of guidelines on step captioning, we propose the guideline-guided captioning task. In this task, models have to generate specific step captions under the guide of guideline.

4 Experiments

4.1 Baselines

Video Foundation Models We evaluate three video foundation models on GUIDE: **VideoChat** [Li *et al.*, 2023b], **Video-LLaMA** [Zhang *et al.*, 2023] and **mPLUG-Owl** [Ye *et al.*, 2023]. VideoChat is instantiated using BLIP-2 [Li *et al.*, 2023a] and Vicuna-7B [Chiang *et al.*, 2023], and combines pre-trained ViT-G [Dosovitskiy *et al.*, 2021] and GMHRA [Wang *et al.*, 2022]. Video-LLaMA comprises a pre-trained ViT-G, an audio encoder, Imagebind [Girdhar *et al.*, 2023], and Vicuna-7B. mPLUG-Owl consists of a pre-trained ViT-L [Dosovitskiy *et al.*, 2021], a visual abstractor, and LLaMA-7B [Touvron *et al.*, 2023].

Language Foundation Models We evaluate four language foundation models on GUIDE: **GPT-3.5-turbo** [OpenAI, 2022], **GPT-4** [OpenAI, 2023], **Flan-T5-XXL** [Chung *et al.*, 2022] and **Vicuna-13B** [Chiang *et al.*, 2023]. GPT-3.5-turbo and GPT-4 are language models with powerful performance proposed by OpenAI. Viucuna is a decoder-only architecture and Flan-T5 is an encoder-decoder architecture.

Human Performance To evaluate the gap between the foundation models' comprehension and human understanding of videos, we ask three people (they are familiar with the program but not seen ground-truth annotations) to accomplish this.

4.2 Evaluation Metrics

Following previous work [Iashin and Rahtu, 2020; Zala *et al.*, 2023], we use METEOR [Banerjee and Lavie, 2005], CIDEr [Vedantam *et al.*, 2015] and SPICE [Anderson *et al.*, 2016] metrics for evaluating models. Additionally, we evaluate the models in two modes for step captioning: Entire Video Captioning (EVC): given an entire video, generate a set of step descriptions. Video Segment Captioning (VSC): given a ground-truth video segment of the step, generate a text description. More details are in the Appendix A.2.

4.3 Implementation Details

In video segment captioning (VSC), we divide the video into multiple segments based on ground-truth step timestamps. We uniformly sample 8 frames for each segment and feed them to models. In entire video captioning (EVC), we uniformly sample 32 frames for each video and feed them to models. In guideline summarization, we modify the input format of the video foundation model to enable simultaneous processing of multiple videos. We uniformly sample 32 frames from each video as input. More details are in the Appendix A.3.

4.4 Main Results

The main results are demonstrated in Table 2. We will summarize different findings in the following:

Step Captioning We observe from the experimental results that video foundation models demonstrate better performances on video segment captioning (VSC) than entire video captioning (EVC). This indicates that while the models can comprehend a specific step, they struggle to understand the entire instructional procedure. One possible explanation is that instructional videos are highly procedural, but the models' pre-training data mainly comprises videos that describe individual events. Conversely, language foundation models are not competent for VSC. This is primarily due to the absence of step-descriptive subtitles in step segments, which hinders models from generating step descriptions based on subtitles.

Guideline Summarization We observe from the experimental results that video foundation models are markedly trailing in this task. Moreover, even the strong GPT-4 demonstrates a substantial performance gap compared to human beings in this task. This indicates that these foundation models struggle to mine the correlation across multiple instructional videos.

Guideline-Guided Captioning By comparing the results of guideline-guided captioning and step captioning, we observe that both video and language foundation models perform much better with the guide of guidelines. This demonstrates the helpfulness of the guideline in generating specific steps.

4.5 Analysis

Importance of Accurate Guideline As shown in Table 3, we use the ground-truth and predicted guideline to guide the generation of specific steps respectively. The results show that there is significant improvement with the ground-truth guideline, indicating that the guideline is helpful to generate specific steps. Moreover, the results using the predicted guideline show a substantial decrease. This further emphasizes the importance of accurate guidelines.

Video Correlation Analysis Capability Mining a common guideline from multiple task-related videos depends on the model's capability to analyze correlations across multiple videos. To investigate the source of this ability, we fine-tune VideoChat (VideoChat has the best fine-tuning performance compared to Video-LLaMA and mPLUG-Owl) under three conditions: (1) single-video: fine-tuning model with single videos along with specific steps (VideoChat$_S$), (2) multiple-video: fine-tuning model with multiple task-relevant videos along with guideline steps (VideoChat$_M$), (3) single-video + multiple-video: fine-tuning model under multiple-video setting based on VideoChat$_S$ (VideoChat$_{S+M}$).

As shown in Table 4, we observe that the performance of VideoChat$_S$ and VideoChat$_M$ are superior to 5-shot VideoChat, and the performance of VideoChat$_M$ is better than VideoChat$_S$.

Methods	Step Captioning (EVC / VSC)			Guideline Summarization			Guideline-Guided Captioning		
	METEOR	CIDEr	SPICE	METEOR	CIDEr	SPICE	METEOR	CIDEr	SPICE
Human Performance	22.5 / 26.0	65.6 / 78.5	24.1 / 38.6	13.6	56.6	14.0	31.8	73.4	36.9
(a) *Video Foundation Models*									
VideoChat$_{\text{5-shot}}$	6.8 / 4.2	1.8 / 4.7	3.6 / 3.7	3.8	0.1	2.1	8.8	4.3	5.6
Video-LLaMA$_{\text{5-shot}}$	4.1 / 2.3	1.1 / 1.3	1.7 / 0.9	2.8	0.2	0.7	8.2	2.3	2.6
mPLUG-Owl$_{\text{5-shot}}$	7.9 / **5.8**	6.4 / **9.6**	5.9 / **6.8**	2.2	2.4	1.4	9.1	8.6	7.5
(b) *Language Foundation Models*									
Flan-T5$_{\text{5-shot}}$	4.7 / -	1.9 / -	5.2 / -	3.3	3.6	1.4	12.9	7.4	11.6
Vicuna$_{\text{5-shot}}$	9.5 / -	4.5 / -	7.4 / -	6.3	5.0	4.9	11.5	7.8	9.3
GPT-3.5-turbo$_{\text{5-shot}}$	14.9 / -	11.2 / -	12.4 / -	9.4	13.3	9.3	17.2	13.1	13.3
GPT-4$_{\text{zreo-shot}}$	16.7 / -	5.9 / -	12.1 / -	9.9	19.5	6.8	22.8	16.0	18.7
GPT-4$_{\text{5-shot}}$	**16.8** / -	**13.1** / -	**13.2** / -	**10.4**	**24.5**	**9.6**	**23.5**	**18.8**	**21.9**

Table 2: The results on three sub-tasks. We report the average results of three runs. 'EVC' denotes entire video captioning and 'VSC' denotes video segment captioning. Best results in each task are highlighted by **bold**. Better results between EVC and VSC are highlighted by underline.

Method	Guideline-Guided Captioning		
	METEOR	CIDEr	SPICE
mPLUG-Owl	7.9	6.4	5.9
- *Pred-guideline*	4.7	2.6	3.1
- *GT-guideline*	**9.1**	**8.6**	**7.5**

Table 3: The results of the mPLUG-Owl on guideline-guided captioning task with predicted and ground-truth guideline inputs. The first line is the result of the model without guideline. Best results are highlighted by **bold**.

Method	Guideline Summarization		
	METEOR	CIDEr	SPICE
VideoChat$_{\text{S+M}}$	**7.3**	**14.9**	**5.8**
VideoChat$_{\text{S}}$	7.2	10.8	3.4
VideoChat$_{\text{M}}$	3.4	2.5	2.9
VideoChat$_{\text{5-shot}}$	3.8	0.1	2.1

Table 4: The results for VideoChat on guideline summarization task under different fine-tuning settings. Best and second results are highlighted by **bold** and underline.

However, VideoChat$_{\text{S+M}}$ shows significant improvement compared to VideoChat$_{\text{M}}$. This indicates that the models' ability to analyze correlations across task-related videos is contingent upon their ability to comprehend single-video. Moreover, models only with single-video comprehension capabilities are incapable of multiple-video comprehension.

Bottleneck of Video Foundation Models To investigate the limitations of video foundation models on instructional video comprehension, we conduct experiments using the mPLUG-Owl on step captioning (EVC). We explore three different settings: (1) giving mPLUG-Owl both video and audio (by using video subtitles to simulate audio), (2) giving mPLUG-Owl only video, and (3) giving mPLUG-Owl only audio.

The experimental results are shown in Table 5. Surprisingly, mPLUG-Owl shows a significant improvement given only audio compared to when given only video. Additionally, we observe a substantial drop in the model's performance when adding video after only providing audio. We hypothesize that much irrelevant information is mixed in during the visual feature extraction process, which hinders the model's understanding of instructional videos. This indicates that more specialized visual encoders and visual-language bridges are needed to represent temporal procedures better.

Method	Step Captioning (EVC)		
	METEOR	CIDEr	SPICE
mPLUG-Owl$_{\text{Video+Audio}}$	5.5	3.2	3.1
mPLUG-Owl$_{\text{Video}}$	3.6	2.9	2.1
mPLUG-Owl$_{\text{Audio}}$	**8.2**	**6.1**	**6.8**

Table 5: The results of the mPLUG-Owl$_{\text{zero-shot}}$ on step captioning (EVC) task with different modal information inputs. Best and second results are highlighted by **bold** and underline.

Human Evaluation of Foundation Models To better evaluate the applicability of the GUIDE in real-world scenarios, we follow the distribution of the dataset and randomly select 53 instructional tasks (286 videos) for huaman evaluation. We compare the results of VideoChat, mPLUG-owl, and GPT-4 on step captioning (EVC) and guideline-guided captioning. The human evaluators are required to rate the output based on whether the steps are clear and easy to learn. We implemented a three-level rating system to categorize the quality of outputs. A means 'steps are very easy to learn', B means 'steps are slightly hard to learn', and C means 'steps are very hard to learn'. As shown in Figure 6, guideline-guided captioning has better results compared to step captioning, indicating that the guideline helps models generate clearer and easier-to-learn instructional steps.

4.6 Case Study

In Figure 5, we list an example generated by foundation models and ground-truth annotation for step captioning (EVC),

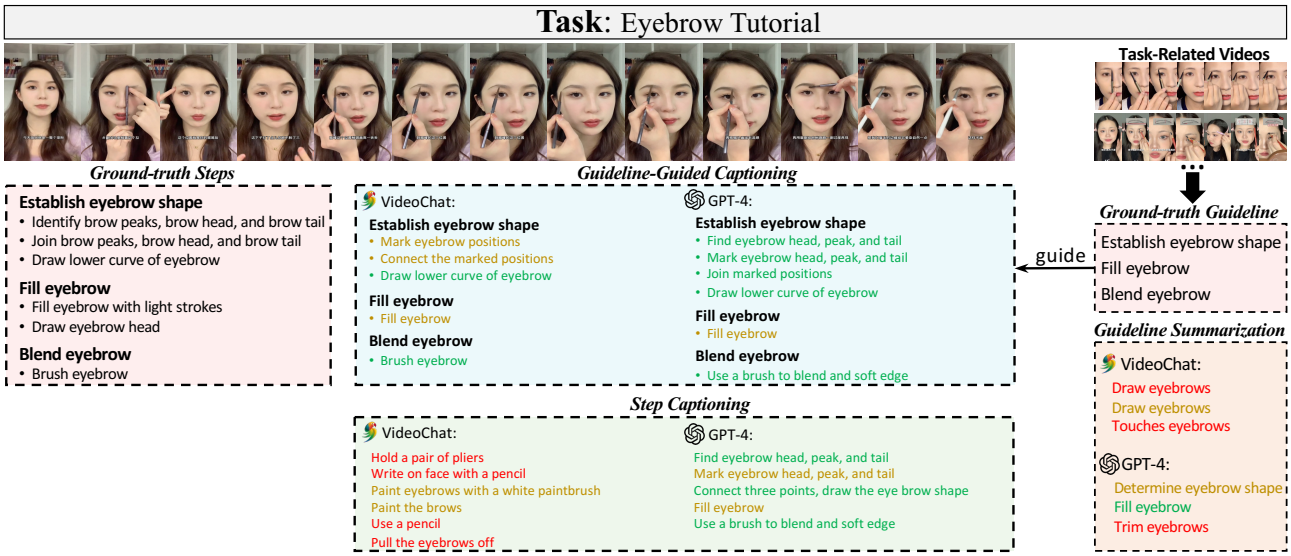

Figure 5: Comparison of foundation models and ground-truth annotation for step captioning, guideline summarization and guideline-guided captioning. Green, yellow, and red text denote 'correct', 'partially correct', and 'wrong' respectively.

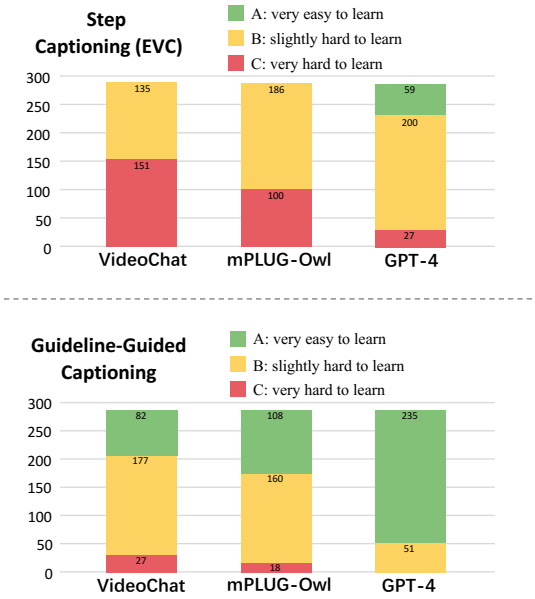

Figure 6: Human evaluation of foundation models on 53 instructional tasks (286 videos).

guideline summarization and guideline-guided captioning task on videos associated with the 'Eyebrow Tutorial'. In the step captioning (EVC), the specific steps generated by VideoChat are very inaccurate. The powerful GPT-4 also has some missing and redundant steps. In the guideline summarization, both GPT-4 and VideoChat are unsuccessful in summarizing the accurate guideline. In the guideline-guided captioning, the VideoChat results show a significant improvement in clarity and accuracy, and the specific steps generated by GPT-4 are essentially entirely correct.

5 Conclusion

In this paper, we introduce a guideline-guided dataset (GUIDE) for instructional video comprehension and propose three subtasks based on the GUIDE. We evaluate various foundation models on our dataset. Experimental results show that the accurate guideline is beneficial for generating clear, easier-to-learn and systemic specific steps. With different training settings, we found that the key to extracting an accurate guideline from multiple videos is single-video understanding ability, indicating more pre-training and fine-tuning data are necessary. Moreover, we observed in the ablation study that the bottleneck for video foundation models is the visual modality. We believe more specialized visual encoders and visual-language bridges are needed to to represent instructional procedures better. Finally, we perform a human evaluation demonstrating our dataset's promising applications in real-world scenarios.

Ethics Concerns

The videos are sourced from an open-source video platform that we cooperate with. Users have agreed to transfer the copyright to the platform when uploading videos, which does not involve privacy issues. Moreover, videos have been subject to strict ethical review (non-ethical content and sensitive information) by the platform.

Acknowledgements

We thank anonymous reviewers for their insightful feedback that helped improve the paper. The research in this article is supported by the National Key Research and Development Project (2021YFF0901602), the National Science Foundation of China (U22B2059, 62276083), and Shenzhen Foundational Research Funding (JCYJ20200109113441941), Major Key Project of PCL (PCL2021A06).

References

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer, 2016.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics, 2005.

[Bolkan *et al.*, 2020] San Bolkan, Goodboy, and Alan K. Instruction, example order, and student learning: Reducing extraneous cognitive load by providing structure for elaborated examples. *Communication Education*, 69(3):300–316, 2020.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[Chung *et al.*, 2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[Damen *et al.*, 2018] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The dataset. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 753–771. Springer, 2018.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Dvornik *et al.*, 2023] Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G. Derpanis, Richard P. Wildes, and Allan D. Jepson. Stepformer: Self-supervised step discovery and localization in instructional videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18952–18961. IEEE, 2023.

[Girdhar *et al.*, 2023] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE, 2023.

[Gu *et al.*, 2023] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18941–18951. IEEE, 2023.

[Han *et al.*, 2020] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[He *et al.*, 2023] Tianyao He, Huabin Liu, Yuxi Li, Xiao Ma, Cheng Zhong, Yang Zhang, and Weiyao Lin. Collaborative weakly supervised video correlation learning for procedure-aware instructional video analysis, 2023.

[Huang *et al.*, 2018] De-An Huang, Shyamal Buch, Lucio M. Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5948–5957. Computer Vision Foundation / IEEE Computer Society, 2018.

[Iashin and Rahtu, 2020] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[Kuehne *et al.*, 2014] Hilde Kuehne, Ali Bilgin Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 780–787. IEEE Computer Society, 2014.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun

Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2023b] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023.

[Liang *et al.*, 2023] Jiafeng Liang, Yuxin Wang, Zekun Wang, Ming Liu, Ruiji Fu, Zhongyuan Wang, and Bing Qin. GTR: A grafting-then-reassembling framework for dynamic scene graph generation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 1177–1185. ijcai.org, 2023.

[OpenAI, 2022] OpenAI. Introducing chatgpt, 2022.

[OpenAI, 2023] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[Tang *et al.*, 2019] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1207–1216. Computer Vision Foundation / IEEE, 2019.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[Wang *et al.*, 2022] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *CoRR*, abs/2212.03191, 2022.

[Yang *et al.*, 2023] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10714–10726. IEEE, 2023.

[Ye *et al.*, 2023] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023.

[Zala *et al.*, 2023] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23056–23065. IEEE, 2023.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics, 2023.

[Zhou *et al.*, 2018] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press, 2018.

[Zhukov *et al.*, 2019] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David F. Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3537–3545. Computer Vision Foundation / IEEE, 2019.